SOME NONPARAMETRIC PROCEDURES FOR GENERAL RIGHT CENSORED DATA

by

Igusti Ngurah Agung

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1347

SOME NONPARAMETRIC PROCEDURES

FOR GENERAL RIGHT CENSORED DATA

by

Igusti Ngurah Agung

A Dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics

Chapel Hill

1981

Approved by:

Advisor

Reader

Reader

© 1981
Igusti Ngurah Agung
ALL RIGHTS RESERVED

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The author wishes to thank the members of the doctoral committee, Dr. Pranab Kumar Sen, Dr. Dana Quade, Dr. Chirayath M. Suchindran, Dr. James E. Grizzle, and Dr. Moye W. Freyman for their support during this research, as well as their review of the manuscript. The author is especially grateful to Dr. Sen, the author's dissertation advisor, for his continuing encouragement, assistance and helpful suggestions; also to Dr. Quade for his review in detail, as well as his suggestions for the improvement of the manuscript. Dr. Suchindran, the author's course advisor, is also due special thanks for his time, patience and encouragement during the author's doctoral program.

A special recognition is due to the author's wife, Dra. Alit M. Agung, for her patience, encouragement and many sacrifices during the long course of study, and also the author's children, Ningsih, Ratna and Dharma, for their patience and understanding.

The author thanks Mrs. Beryl Glover for her cooperation and excellent typing of the manuscript.

Finally, the author is particularly indebted to the Department of Education and Culture, the Republic of Indonesia; the National Family Planning Coordination Board (BKKBN), Jakarta; the United States Agency for International Development; and the Ford Foundation for their financial support during the author's training program in the United States.

July 10, 1981
Igusti N. Agung

ABSTRACT

IGUSTI NGURAH AGUNG. Some Nonparametric Procedures for General Right Censored Data (Under the Direction of DR. PRANAB KUMAR SEN).

This study covers three main topics. First, the pair chart of Quade (1973) will be extended to singly censored two-sample problems and uncensored three-sample problems. These extensions are considered as descriptive statistics for the censored K-sample problem in testing the null hypothesis that each sample has the same distribution function. Furthermore, based on these charts, we develop a maximum distance statistic for a statistical test.

An extension of Kendall's (1938) tau to censored bivariate and multivariate data is the second topic of this study. Considering a general right censored bivariate data, we propose an unconditional and a conditional generalized Kendall's tau (GKT) as measures of association between the two components or variables. For right censored multivariate data, we will consider a vector statistic as a generalization of Simon's (1977) Kendall's tau. All these extensions can be represented as U-statistics of Hoeffding (1948). Hence, they have limiting normal distributions. Moreover, we also study their null and nonnull distributions under a sequence of alternative hypotheses.

Finally, for m-variates, we introduce a type of symmetry, which is considered as sector symmetry. A chi-squared statistic and an index of symmetry are developed for testing the null hypothesis

that the m-variates are symmetric dependent.

Applications of the pair chart and the GKT for demographic data also are given.

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

The present work deals with the development of some nonparametric statistical procedures associated with K-sample problems and the m-variate variables, which are possibly censored on the right. This chapter presents a review of some statistical procedures and the outlines of this work.

The literature review is given in the following four sections. Section 1.2 describes various types of censoring. In section 1.3, we review some plotting methods, such as the pair chart of Quade (1973) for 2-sample problems and the hazard plot of Nelson (1972) for multiply censored life data. In Section 1.4, a number of measures of association such as the product moment correlation, Kendall's (1938) tau, Spearman's rho, and the Goodman-Kruskal (1954) index gamma are briefly presented. In Section 1.5, some conditional statistics for censored data are given. This section covers the conditional W statistic of Gehan (1965), the k-th order statistic proposed by Chatterjee and Sen (1973), and other statistics which are related to the Mann-Whitney U statistic.

Finally, the outline of this work will be presented in Section 1.6.

1.2 Types of Censoring

Data are said to be right (or left) censored if the values of the observations may not go beyond (or below) some point(s). If all values of the observations may not go beyond (or below) a fixed point, then we have singly right (or left) censored data. Otherwise, we have multiply right (or left) censored data; in which the values of the censored (incomplete) observations and those of the uncensored (complete) one are intermixed. Life test data are frequently right censored; that is, the failure times of unfailed units are known only to be beyond their current running times. Data are said to be doubly censored if there are singly, right and left, censored observations. And data are said to be censored if they are right or left censored. Finally, data are said to be general (randomly) right censored if they are right censored, and each unit in the sample can be considered as having both a random censoring time and a failure time which are statistically independent.

Now, considering a possible random variable relating to the censoring procedure we can differentiate between Type I and Type II censoring. Type I censoring occurs in the right censoring case if the censoring point(s) is pre-chosen. Hence, the number of uncensored observations is a random variable. In Type II censoring, the number of uncensored observations is pre-chosen, and the censoring time is random. Type I and Type II censoring are the two basic types of planned censoring, in which the researcher has planned to stop the observations at certain point(s). Unplanned censoring may

occur in clinical trials; in which the researcher has to stop the study because of the unexpected results which may occur.

Previous paragraphs clearly show the characteristics of censored or incomplete data. Otherwise, if the data are not censored then they are said to be complete.

1.3 Plotting Methods

Several types of plots have been used for presenting the data and for statistical testing. As a descriptive statistic, a plot, such as the residual plot or the pair chart, could show us the relations between the corresponding variables or samples. However, the results of each analysis tend to be very subjective, because one may see a specific deviation which can not be detected by the others.

Since we are proposing an extension of the pair chart of Quade (1973) for right-censored data, first we will consider Quade's paper.

Let $X_{i1}, X_{i2}, \ldots, X_{in_i}$ be a random sample of size n_i on a variable X_i with unknown distribution F_i, i=1,2. A pair chart can be used as a descriptive statistic for testing the null hypothesis $H_0: F_1 = F_2$. The pair chart also can be used as an aid in computing and interpreting various nonparametric procedures for the two-sample problem; such as the Kolmogorov-Smirnov test, the Wilcoxon-Mann-Whitney test, and Mood's squared-rank test.

The construction of the pair chart is as follows. Draw a rectangle of width n_1 units and height n_2 units. If the smallest observation in the combined sample is an X_1, draw a line from the

lower left corner of this rectangle one unit to the right; if it is an X_2, draw the line one unit up instead. From the end of the first line draw a second line, one unit to the right if the second smallest observation is an X_1, and one unit up if it is an X_2. Continue in the same manner for all (n_1+n_2) observations. Then the (n_1+n_2) line segments form a path from the left hand corner of the rectangle to the upper right corner.

If ties occur between X_1 and X_2, then we have boxes, corresponding to the ties, along the path of the pair chart.

This path divides the $n_1 x n_2$-rectangle into two polygons with areas $U(X_1)$ and $U(X_2)$, below and above the path respectively. The boxes corresponding to the ties are equally divided between $U(X_1)$ and $U(X_2)$. These areas represent the familiar Mann-Whitney (1947) U-statistic. As a descriptive statistic, a large difference between $U(X_1)$ and $U(X_2)$ would indicate that we may reject the null hypothesis $H_0: F_1 = F_2$.

It is clear that the quantities $U(X_i)$, i=1,2 are easily calculated with the aid of the pair chart.

Another type of test statistic, the triplets $N(X_1,X_2,X_1)$ and $N(X_2,X_1,X_2)$ will be reviewed for later extension. If there are no ties, Quade noted that $N(X_i,X_j,X_i)$, $i\neq j=1,2$ is the number of ways in which it is possible to choose from the data 2 X_i's and 1 X_j such that the X_j lies between the X_i's. With the aid of the pair chart, the quantities $N(X_i,X_j,X_i)$, $i\neq j=1,2$ are calculated as follows.

In the r-th row of the $n_1 x n_2$-rectangle, let L_r be the number of squares to the left of the path, and R_r the number to the right of

it; if some squares, say Q_r of them lie within a box, count these as equally divided between L_r and R_r.

$$N(X_1,X_2,X_1) = \sum_{r=1}^{n_2} L_r R_r - \sum_{r=1}^{n_2} Q_r(Q_r + 2)/12 \qquad (1.3.1)$$

Similarly,

$$N(X_2,X_1,X_2) = \sum_{c=1}^{n_1} A_c B_c - \sum_{c=1}^{n_1} H_c(H_c + 2)/12 \qquad (1.3.2)$$

where A_c, B_c, and H_c are the numbers of squares above the path, below it, and boxed respectively, within the c-th columns.

The limitations of this pair chart is that it may not be used as such for censored data, in particular for right-censored data. Because a censored observation and an uncensored observation are not always comparable, and any pair of censored observations are not comparable either, it is necessary to extend this pair chart for censored data which we shall discuss later on.

For comparing two one-dimensional samples, Wilk and Gnanadesikan (1968) introduced two kinds of probability plots, namely quantile versus quantile plots (Q-Q plots) and percent versus percent plots (P-P plots). If the two variables are both uniform on (0,1), then the two plots are identical.

If X_i,i=1,2 are identically distributed then the Q-Q plot and P-P plot of the X_i's are straight lines with slope 1 through the origin. If X_1 is a linear function of X_2 then the corresponding Q-Q plot will still be linear but with possibly changed location and slope but the P-P plot will not remain linear.

Wilk and Gnanandesikan noted the use of the probability plot,

as an internal comparison, in regression analysis. For example, the Q-Q plot of the ordered residuals and a normal distribution could show whether the sample came from normal population.

As the pair charts, the probability plots are limited to uncensored data.

Finally, for multiply censored life data, Nelson (1972) introduced the hazard plot for graphical analysis. As a descriptive statistic, the hazard plot would show whether the assumed distribution of the times to failure of the units or subjects adequately fits the data. On hazard plotting paper for a theoretical distribution, the data and the cumulative hazard scales are chosen so any such cumulative hazard function is a straight line on the paper. Papers have been developed for the exponential, normal, lognormal, extreme value and Weibull distributions.

As usual, the hazard density function h(x), for a distribution of time x to failure is defined as

$$h(x) = f(x)/\{1 - F(x)\} \qquad (1.3.3)$$

where F(x) is the cumulative distribution function (c.d.f), which is assumed to be differentiable with $F'(x) = f(x)$. And the cumulative hazard function H(x) satisfies the equation

$$F(x) = 1 - \exp.\{-H(x)\}$$

For the hazard plotting method, it is assumed that the censoring is random, that is, if the unfailed units were to run to failure, their failure times would be statistically independent of their censoring times. (The life distribution of the units censored at a particular age must be the same as the conditional life

distribution of the units that run beyond that age.)

1.4 Measure of Association

A classic measure of association between two variables X and Y is the parametric coefficient of correlation, r. Suppose we have a bivariate random sample, (X_i,Y_i), i=1,...,n, of size n from a two-dimensional variable (X,Y) then it is defined by

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2\}^{\frac{1}{2}}} \qquad (1.4.1)$$

where $\bar{X}$ and $\bar{Y}$ are the average of the X_i's and Y_i's, respectively. This coefficient is known as the product-moment correlation.

Based on this correlation, Hotelling and Pabst (1936) studied the rank correlation r_s, which was invented by C. Spearman. By taking X_i as the rank of the i-th individual with respect to the first component, and Y_i as his rank with respect to the other component, Spearman's r_s can be written as

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (1.4.2)$$

with variance

$$\sigma^2 = 1/(n-1) \qquad (1.4.3)$$

where d_i is the difference between the two ranks of the i-th individual (=observation).

Another index of rank correlation was introduced by Kendall (1938), which is known as Kendall's tau, t. The index was defined by

$$t = \frac{2(C - D)}{n(n - 1)} \tag{1.4.4}$$

where C denotes the total number of concordant pairs of observations (X_i,Y_i) i=1,...,n, and D denotes the total number of discordant pairs. An alternative formula given by Kendall (1970) is

$$t = 1 - \frac{4s}{n(n-1)} \tag{1.4.5}$$

where s is the minimum number of interchanges between neighbors required to transform one ranking to the other. Here, it was assumed that X and Y have continuous marginal distribution functions, so there are no ties among the X_i's or the Y_i's.

Furthermore, if ties occur, Kendall (1970) proposed an index tau-B.

$$t_B = \frac{C - D}{\sqrt{\{\tfrac{1}{2}n(n-1) - T\}}\sqrt{\{\tfrac{1}{2}n(n-1) - U\}}} \tag{1.4.6}$$

where T denotes the total number of tied pairs in one ranking, and U for ties in the other, as a modification of the previous tau, which is considered as tau-A.

Considering the occurrence of ties, Goodman and Kruskal (1954) proposed an index γ(gamma) for the rank correlation coefficient. The index was defined by

$$\gamma = \frac{C - D}{C + D} = \frac{C - D}{\tfrac{1}{2}n(n-1) - T_{12}} \tag{1.4.7}$$

where T_{12} denotes the total number of tied pairs among (X_i,Y_i), i=1, ...,n. In this case, (X_i,Y_i) and (X_j,Y_j) are considered as tied if $X_i = X_j$ or $Y_i = Y_j$ or both.

A more general measure of the correlation between two sets of observations, ranked or otherwise, was introduced by Daniel (1948). He introduced the quantity

$$\Gamma = \frac{\Sigma a_{ij} b_{ij}}{\sqrt{(\Sigma a_{ij}^2 \; \Sigma b_{ij}^2)}} \tag{1.4.8}$$

as a correlation coefficient, where a_{ij}, b_{ij} are scores assigned to corresponding pairs i,j of the two variables X, Y; and $a_{ij} = -a_{ji}$, $b_{ij} = -b_{ji}$. Special cases of Γ are Kendall tau, Spearman's rho and the product-moment correlation.

Daniels and Kendall (1947) have shown that the sample correlation Kendall's tau, t, is an unbiased estimator of the population value τ, i.e.,

$$E(t) = \tau \tag{1.4.9}$$

Using a U-statistic, Hoeffding (1948) showed that t is an unbiased estimator of τ, a regular functional of F of degree 2.

However, the sample correlation Spearman's rho is not an unbiased estimator of the population value ρ. Hoeffding showed that

$$E(r_s) = \{(n-2)\ \rho + 3\tau\}/(n+1) \tag{1.4.10}$$

Moran (1948), and Durgin and Stuart (1951) presented different formulas in showing this bias. The correlation between Kendall's tau and Spearman's rho over all permutations of the sample values was found by Daniels (1944). He obtained

$$r_{r_s t} = \frac{2(n+1)}{\sqrt{\{2n(2n+5)\}}} \tag{1.4.11}$$

All indexes considered above are limited to complete (= uncensored) data. Hence, it is necessary to propose measures of association, which are applicable for censored data. Of course, those measures should be applicable to complete data as well.

1.5 Conditional Statistics for Censored Data

We assume that we have two samples of sizes n_1 and n_2 from populations having (discrete or continuous) cumulative distribution F and G respectively. And let r_1, r_2 be the number of censored elements in the two samples. We denote the samples as follows:

Sample 1:	$X_{11}, X_{12}, \ldots, X_{1,n_1-r_1}$	uncensored
	$X^*_{11}, X^*_{12}, \ldots, X^*_{1,r_1}$	censored
Sample 2:	$X_{21}, X_{22}, \ldots, X_{2,n_2-r_2}$	uncensored
	$X^*_{21}, X^*_{22}, \ldots, X^*_{2,r_2}$	censored

Halperin (1960) proposed a nonparametric two sample test, U_c, which is an extension of the Wilcoxon-Mann-Whitney test to two samples censored at the same fixed point T. The null hypothesis to be tested was taken to be $F(x) = G(x)$, $-\infty < x \leq T$ against the alternative $F(x) > G(x)$. He defined U_c by $P_r\{U_c \leq U_c\} = \alpha$, where the probability was computed over the conditional universe for $r_1 + r_2 = r$ (the observed total number of censored elements). To compute U_c, for any value of r_1, we use

$$U_c = U(n_1-r_1, n_2-r_2) + r_1(n_2-r_2) \qquad (1.5.1)$$

where $U(n_1-r_1, n_2-r_2)$ is the Mann-Whitney U statistic for the uncensored elements of the two samples. Here, we should note that U_c is a

conditional statistic, even under the null hypothesis.

A more general conditional statistic was proposed by Gehan (1965). He gave a distribution-free two-sample test, an extension of the Wilcoxon test to samples with arbitrary censoring on the right. The null hypothesis was

H_0: $F(t) = G(t)$, $(t \leq T)$, against either

H_1: $F(t) < G(t)$, $(t \leq T)$ or

H_2: $F(t) < G(t)$ or $F(t) > G(t)$, $(t \leq T)$

He defined

$$U_{1j} = \begin{cases} +1 & X_{1i} < X_{2j} \text{ or } X^*_{1i} \leq X_{2j} \\ -1 & X_{1i} > X_{2j} \text{ or } X_{1i} \geq X^*_{2j} \\ 0 & \text{otherwise} \end{cases} \qquad (1.5.2)$$

and calculated the statistic $W = \sum_{i,j} U_{ij}$, where the sum is over all $n_1 n_2$ comparisons. Gehan considered the conditional mean and variance of W under H_0, denoted by $E(W \mid P, H_0)$ and $Var(W \mid P, H_0)$, where P was the pattern of observations. To test H_0 against either H_1 or H_2, a value of

$$Z = \frac{W}{\sqrt{\{var(W \mid P, H_0)\}}} \qquad (1.5.3)$$

was taken as asymptotically normal with zero mean and unit variance. The conditional variance of W was given in Gehan's paper, including several particular cases. One of them is

$$var(W \mid P, H_0) = \frac{N(N-n_1)(N-r)}{N(N-1)} \left[Nr + 1/3\{(N-r)^2 - 1\}\right] \qquad (1.5.4)$$

with $N = n_1 + n_2$ and $r = r_1 + r_2$, if there are no ties and all censored observations occur after the (N-r)-th uncensored observation, which was considered by Halperin. In this special case, the censoring is at the k (=N-r)-th order statistic of the combined sample of size $N=n_1+n_2$. This case is considered also by Sobel (1966) and Basu (1967).

Sobel (1966) proposed a statistic

$$V_k^{n_1,n_2} = V_k^{(N)} = \sum_{j=1}^{k} (n_{1j}n_2 - n_1 n_{2j}) \tag{1.5.5}$$

where n_{ij} is the number of uncensored observations on X_i among the first j ordered observations of the combined sample, so that

$$n_{1j} + n_{2j} = j, \quad j=1,2,\ldots,k \tag{1.5.6}$$

Another statistic, $T_k^{(N)}$, was introduced by Basu (1967). He defined

$$T_k^{(N)} = \sum_{j=1}^{k} e_j z_j \tag{1.5.7}$$

where

$$z_j = \begin{cases} 1 & \text{if the j-th ordered observation is an } X_1 \\ 0 & \text{otherwise} \end{cases}$$

$$e_j = \begin{cases} (j-k-1)/N + (k+1)^2/2N^2 & \text{if } 1 \le j \le k \\ (k+1)^2/2N^2 & \text{if } k < j \le N \end{cases} \tag{1.5.7}$$

Basu showed that the statistics $V_k^{(N)}$ and $T_k^{(N)}$ are equivalent for testing H_0: F = G against one-sided (or two-sided) alternative H_1: $F \neq G$. By putting k = N, he obtained the relationship of $T_k^{(N)}$ with the Wilcoxon statistic and the Mann-Whitney statistic, U, that is

$$NT_N^{(N)} = U + n_1(n_1 + 1)/2 - n_1(N^2 - 1)/2N \tag{1.5.8}$$

Censoring at the k-th order statistic was considered also by Chatterjee and Sen (1973). They, for a fixed-plan truncation scheme, proposed a conditional test for H_0. Assuming continuity of F and G, under H_0, let

$$Z_{N(1)} < Z_{N(2)} \cdots < Z_{N(N)}$$

$$\text{or } Z_1 < Z_2 \cdots < Z_n$$

be the ordered values of the combined sample and $R_{Ni}(=R_i)$ be the rank of $z_{N(i)}$ among $Z_{N(1)},\ldots Z_{N(N)}$, then $(R_1,R_2,\ldots,R_N) = (1,2,\ldots,N)$. If the experimentation stops at a pre-fixed time point, then the number of (completed) observations is

$$k(z) = \sum_{i=1}^{N} U(z - Z_1) \tag{1.5.9}$$

where u(t) is 1 or 0 according to $t \geq 0$ or $t < 0$. Under H_0

$$\underline{P}\{k(z) = k\} = \binom{N}{k}\{F(z)\}^k\{1 - F(z)\}^{N-k},$$

$$k = 0,1,\ldots,N \tag{1.5.10}$$

and

$$N^{-1}k(z) \longrightarrow F(z) \text{ a.s., as } N \longrightarrow \infty$$

Furthermore, Chatterjee and Sen proposed the conditional statistic $T_{Nk(z)}$ $(= T_n(z))$ below

$$T_N(z) = \begin{cases} 0, & k(z) = 0 \\ \sum_{i=1}^{k(z)} \dfrac{c_i - \bar{c}_N}{C_N}\{a_N(i) - a_N^*(k(z))\}, & 1 \leq k(z) \leq N-2 \\ T_N, & k(z) = N-1,N \end{cases} \tag{1.5.11}$$

with

$$C_N^2 = \sum_{i=1}^{N} (c_i - \bar{c}_N)^2, \quad \bar{c}_N = N^{-1} \sum_{i=1}^{N} c_i$$

and

$$a_N^*(k) = (N - k)^{-1} \sum_{j=k+1}^{N} a_N(j)$$

where $c_1,\ldots,c_N$ are given numbers and $\{a_N(1),\ldots,a_N(N)\}$ represents a set of scores to be suitably chosen, such that

$$\sum_{i=1}^{N} a_N(k) = 0 \quad \text{and} \quad \sum_{i=1}^{N} a_N^2(i) = N-1 \tag{1.5.12}$$

For the two sample problem, we have c_i is 1 or 0 according as z_i comes from the F or G distribution. They also derived, for every $N(\geq 1)$ and $k(z) \geq 0$.

$$E(T_N(z) \mid H_0) = 0 \quad \text{and} \tag{1.5.13}$$

$$\text{Var}(T_N(z) \mid H^0) = V_N(z) \tag{1.5.14}$$

with

$$V_N(z) = \begin{cases} 0, & k(z) = 0 \\ 1 - (N-1)^{-1}\left[\sum_{i=k(z)+1}^{N} a_N^2(i) - (N - k(z))^{-1} \times \right. & \\ \left. \left\{\sum_{i=k(z)+1}^{N} a_N(i)\right\}^2\right], & 1 \leq k(z) \leq N-2 \\ 1, & k(z) = N-1, N \end{cases} \tag{1.5.15}$$

and as $N \to \infty$

$$\{V_{Nn}^*\}^{-\frac{1}{2}} T_N(z) \xrightarrow{L_0} N(0,1) \tag{1.5.16}$$

where $n^* = [NF(z)]$, $0<F(z)\leq 1$; and L_0 stands for the convergence in

law under H_0. A special case of this general formulation was presented by Davis (1978). Taking $a_N(i) = i$, Davis proposed a two sample Wilcoxon type statistic for analyzing the data for which $pN(0 < p \leq 1)$ smallest observations are to be observed sequentially.

1.6 Outline of the Present Work

Considering a two-sample life testing problem, pair charts for right censored data are developed, as an extension of the pair chart of Quade (1973). This extension will be called the Censored Pair Chart (CPC). And two types of CPC's are proposed, which are considered as CPC-I and CPC-II. The construction of the CPC-I is based on the original data having complete and incomplete observations. However, the CPC-II is constructed based on a transformed data in which the data are treated as a complete data. The main purpose of the CPC-I is to present Gehan's (1965) W statistic as a chart, based on either grouped or ungrouped data. Some other statistics such as triplet statistics and the maximum distance, D, statistic, can be calculated based on a CPC. This topic is discussed in Chapter II. Moreover, this chapter also presents the distribution and a large sample approximation to the D statistic. However, the large sample approximation is limited to the case of equal sample sizes.

The derivation of the distribution of the D statistic can be considered as an application of the lattice path counting as treated in Mohanty (1979). Also we consider using the Hodges' (1958) method in calculating the level of significance for the D statistic. This

method is associated with the well known Pascal Triangle, which is important especially for unequal sample sizes.

Chapter III presents an extension of Kendall's (1938) tau to right censored bivariate variable. This extension will be called a generalized Kendall's Tau (GKT), which can be represented as the U statistic of Hoeffding (1948). Hence, the GKT can be approximated by a normal distribution function. We propose two types of GKT's, the unconditional GKT (UGKT) and the conditional GKT (CGKT).

By extending the definition of concordant and discordant pairs of observations to right censored bivariate data, the GKT's can be written as functions of concordant and discordant pairs.

The distribution of the UGKT is discussed in detail under a null hypothesis of independence and a sequence of alternative hypothesis. Since censored data are associated with two variables, true variable $\underline{X}$ and censoring variable $\underline{Y}$, we define a variable $\underline{Z} = \text{Min}(\underline{X},\underline{Y})$ and an indicator variable $\underline{\delta}$ presented in Section 3.8. Then we study the distribution function of $\underline{Z}$ related to that of $\underline{X}$ based on a dependence function, introduced by Sibuya (1960). Examples are given under the assumption that $\underline{X}$ has a bivariate normal distribution function.

Now, for the CGKT, its conditional distribution is studied for given general patterns of observations on both components of bivariate data. This general pattern was introduced by Gehan (1965). And the distribution function of the CGKT is considered under the assumption that all possible orderings of observations which lead to

the fixed given patterns are equally likely. A detailed discussion is given for a particular case, in which all observations on one component are uncensored, and they have a natural ordering. Examples are given to illustrate the variabilities in calculating the d.f. of the CGKT. Even in this special case, it seems impossible to obtain an explicit expression for the d.f. of the CGKT, even though the sample size is not very large. The complexity of deriving this d.f. is outlined.

In Chapter IV, we propose two statistical procedures. The first procedure is an extension of the GKT's to right censored multivariate data, and the second procedure is for sector symmetry or interchangeability of an m-variate variable.

The extension of the GKT's to the censored m-variate variable will be presented as a vector statistic, which can be considered as an extension of the vector statistic of Simon (1977). The components of this vector statistic are the GKT's of all possible pairs of components and the Simon's indexes for groups of more than two components. The latter will be called the Generalized Simon's Statistic (GSS).

Based on a value-case diagram, which is an extension of the sign-case diagram of Simon (1977), a matrix equation is developed for the vector statistic. Considering its components, in particular the GSS, we propose concordant and discordant properties for m-variate pairs. Then, the GSS can be represented as a function of concordant and discordant pairs.

As an application of the UGKT of the bivariate variable, in this chapter we consider pairwise independence among the m variates. Finally, based on the m-variate dependence function, proposed by Sen (1967), we study the non-null distribution of the GSS.

Now, for the second procedure, we introduce a necessary and sufficient condition for the sector symmetry. First, we develop a chi-squared statistic and an index of symmetry for testing the null hypothesis that the m-variate variable is sector symmetric, based on complete data. Then, this is extended to incomplete (censored) data sets. As in Chapter II, that is the CPC, the analysis of censored data can be done based on the transformed data, in which the data are treated as a complete data set, or the original data.

The first analysis is appropriate for testing that the true variable is symmetric, by introducing additional restrictions. For the second kind of analysis, we propose a conditional chi-squared statistic. And for a particular case, in which only one component may be censored we propose other statistical tests, in particular a chi-squared test based on a weighted linear combination method. The distribution of this chi-square is studied under a null hypothesis and a sequence of alternative hypotheses, including its large sample approximation.

Chapter V presents the use of the CPC for K-sample problems in testing the null hypothesis that the K samples have the same d.f. Here, we consider two possible approaches: (i) using CPC's in pairwise comparisons, and (ii) using CPC's to present the Breslow's (1970) statistic.

For K = 3, we develop a three dimensional chart as a descriptive statistic for three-sample problems. This chart is called a triplet chart. The limitation of this triplet chart is that it should be constructed based on a complete data set or a censored data set, which could be treated as a complete data set. Examples are given for small sample sizes to illustrate the construction of the triplet chart and the use of the triplet chart as a descriptive statistic.

In studying the properties of a triplet chart, we consider the two kinds of orthogonal projections of its path: (i) its projections on the three coordinate planes, and (ii) its projection on the plane $X_1+X_2+X_3=0$ where X_i is the i-th coordinate axis.

For equal sample sizes, we introduce a maximum distance, D, statistic, for three sample problems. The distribution of this D statistic is studied using the second kind of the projection of the corresponding triplet charts. Here, the d.f. of the D statistic will be represented by the number of paths which lead to a certain value of D. Moreover, we also develop some recursive formulas.

In Chapter VI, we have the applications of the CPC and the UGKT to censored data in demography.

Finally, in Chapter VII, we have concluding remarks and suggestions for further research.

CHAPTER II

THE PAIR CHART FOR GENERAL SINGLY CENSORED TWO SAMPLE PROBLEMS

2.1 Introduction

This chapter deals with an extension of the pair chart, as represented by Quade (1973), to right censored data. The extension will be called a censored pair chart (CPC). For the case of left censoring, the data can be transformed into the right censoring problem, if one multiplies the observed values by -1. Hence, the discussion is limited to the general right censored two sample problems.

Let X_{ij}, $j=1,\ldots, n_i-r_i$; X^*_{ik}, $k=1,\ldots,r_i$; be a random sample of size n_i on variable X_i with unknown distribution function (d.f.) F_i, $i=1, 2$. So, each sample has two types of observations: (i) the complete (uncensored) observation X_{ij}, that is the true value for the j-th individual or subject in the i-th sample; and (ii) the incomplete (censored) observation X^*_{ik}, that is the value or point at which the observation is terminated or censored for the k-th individual in the i-th sample.

Two types of CPC's will be introduced. First, a CPC which is directly applicable for computing the W statistic of Gehan (1965). This W provides a conditional test for

$$H_0: F_1(x)=F_2(x) \quad (x \leq T) \text{ against}$$
$$H_1: F_1(x)\neq F_2(x) \quad (x \leq T) \qquad (2.1.1)$$

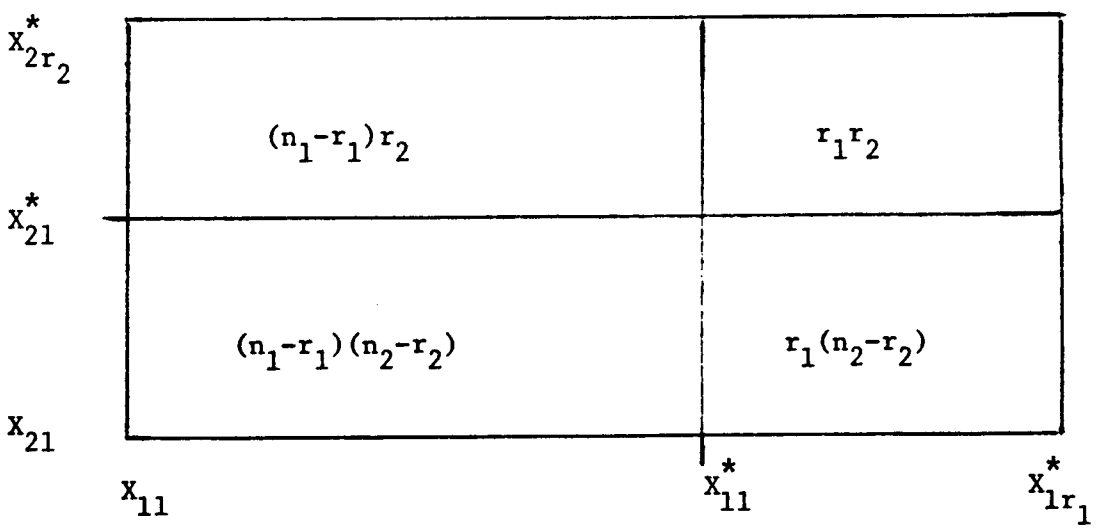
where T is the upper bound of the observed values. Gehan defined

$$U_{ij} = \begin{cases} -1 & X_{1i} < X_{2j} \text{ or } X_{1i} \leq X^*_{2j} \\ 0 & X_{1i} = X_{2j} \text{ or } (X^*_{1i}, X^*_{2j}) \text{ or } X^*_{1i} < X_{2j} \text{ or } X^*_{2j} < X_{1i} \\ +1 & X_{1i} > X_{2j} \text{ or } X^*_{1i} \geq X_{2j} \end{cases} \qquad (2.1.2)$$

and calculated the generalized Wilcoxon statistic $W = \Sigma\Sigma U_{ij}$, where the sum is over $n_1 n_2$ comparisons. To construct the CPC, without loss of generality, we may assume that X_{ij}, $j=1,\ldots,n_i-r_i$, is increasing for each $i=1,2$; but X^*_{ik}, $i=1,\ldots,r_i$ is decreasing for each i. Different orderings between X_{ij}'s and X^*_{ik}'s within each sample are needed in order to obtain two polygons on the X_i-side, $i=1,2$ with areas $A(X_i)$, which are the statistics considered later.

Furthermore, as in Quade (1973), we draw a rectangle of width n_1 units, and height n_2 units. So we have a rectangle of size $n_1 n_2$. This rectangle is divided into four sub-rectangles of sizes $(n_1-r_1)(n_2-r_2)$, $(n_1-r_1)r_2$, $r_1(n_2-r_2)$, and $r_1 r_2$ as shown in Figure 2.1.

Figure 2.1. The Format of the Censored Pair Charts



Each sub-rectangle is further subdivided into unit squares. Hence we have $n_1 n_2$ squares in total. Only the first three sub-rectangles are needed for the construction of the censored pair chart, which will be called CPC of the first type (CPC-I). In each sub-rectangle, we can construct a pair chart as the pair chart of Quade, as follows.

In the (n_1-r_1) (n_2-r_2) sub-rectangle, if the smallest observation in the combined uncensored samples is an X_1, draw a line from the lower left corner of this rectangle one unit to the right; if it is X_2, draw the line one unit up instead. From the end of the first line draw a second line, one unit to the right if the smallest remaining observation is an X_1, one unit up if it is an X_2. Continue in the same manner for all $(n_1-r_1)+(n_2-r_2)$ uncensored observations.

Next, in the $(n_1-r_1)r_2$ sub-rectangle, if the smallest observation in the combined subsamples; X_{1j},j-1,...,n_1-r_1, and X^*_{2k}, k=1,..., r_2; is an X_1, draw a line one unit to the right from the upper left corner, otherwise draw a line one unit down. From the end of the first line draw a second line, one unit to the right if the smallest remaining observation is an X_1, and one unit down otherwise. Continue in the same manner for all $(n_1-r_1)+r_2$ observations.

Finally, in the $r_1(n_2-r_2)$ sub-rectangle, if the smallest observation in the combined sub-samples: X^*_{1k}, k=1,...,r_1, and X_{2j}, j=1..., n_2-r_2 is an X_2, draw a line one unit up from the lower right corner; otherwise, draw the line one unit to the left. From the end of the first line draw a second line, one unit up if the smallest remaining observation is an X_2, and one unit to the left otherwise. Continue in the same manner for all $r_1+(n_2-r_2)$ observations.

The paths of the three pair charts and the left and the lower sides of the $n_1 n_2$ rectangle form two polygons with areas $A(X_1)$ and $A(X_2)$ unit squares. As a descriptive statistic, a large difference between $A(X_1)$ and $A(X_2)$ would indicate that X_1 and X_2 do not have the same distribution functions. Thus, we reject the null hypothesis.

With the aid of the CPC, we can calculate the W statistic of Gehan as $A(X_1) - A(X_2)$. This will be discussed in section 2.2 in more detail.

The other type of CPC, which will be called CPC-II, is constructed as follows. Let X_{ij} be the true observation for the j-th individual in the i-th sample. ($j=1,\ldots,n_i$; $i=1,2$) from the random variable X_i with c.d.f. F_i. Since this observation may be censored by a varaible T_{ij}, it cannot always be observed. But we observe

$$X'_{ij} = \min(X_{ij}, T_{ij}) \tag{2.1.3}$$

along with the indicator variable

$$\delta_{ij} = \begin{cases} 0 & (X'_{ij} = X_{ij}) \\ 1 & (X'_{ij} = T_{ij} < X_{ij}) \end{cases} \tag{2.1.4}$$

which shows whether or not X'_{ij} is in fact uncensored. Assuming X_i and T_i are independent random variables, we would have

$$\begin{aligned} P(X'_i \geq x) &= P(\min(X_i, T_i) \geq x) \\ &= P(X_i \geq x).P(T_i \geq x) \end{aligned} \tag{2.1.5}$$

Hence

$$1-F'_i(x) = (1-F_i(x))(1-G_i(x)) \tag{2.1.6}$$

where F'_i and G_i are the c.d.f.'s of X'_i and T_i respectively. This

shows, in general, $F_1'(x) \neq F_2'(x)$ under the null hypothesis H_0 in (2.1.1). However, if we accept the assumption that $G_1(x) = G_2(x)$, then $F_1'(x) = F_2'(x)$ under H_0. In this special case, we would test the null hypothesis

$$H_0': F_1'(x) = F_2'(x), \text{ against}$$

$$H_1': F_1'(x) \neq F_2'(x) \qquad (2.1.7)$$

using the usual statistics, such as Kolmogorov-Smirnov statistic and pair chart. Hence, based on X_{ij}', $j=1,\ldots,n_i$, $i=1,2$, we can construct the second type of the CPC (CPC-II), which is the same as that of Quade. However, indicator variables δ_{ij} may be considered in addition to this pair chart. This and the comparison between CPC-I and CPC-II will be discussed in section 2.3.

2.2. Censored Pair Chart of Type-I

2.2.1 Discussion

For discussion, the data of Freireich et al. (1963) considered by Gehan (1965) and Breslow (1970) will be re-analyzed using a censored pair chart. The data are given in Table 2.1.

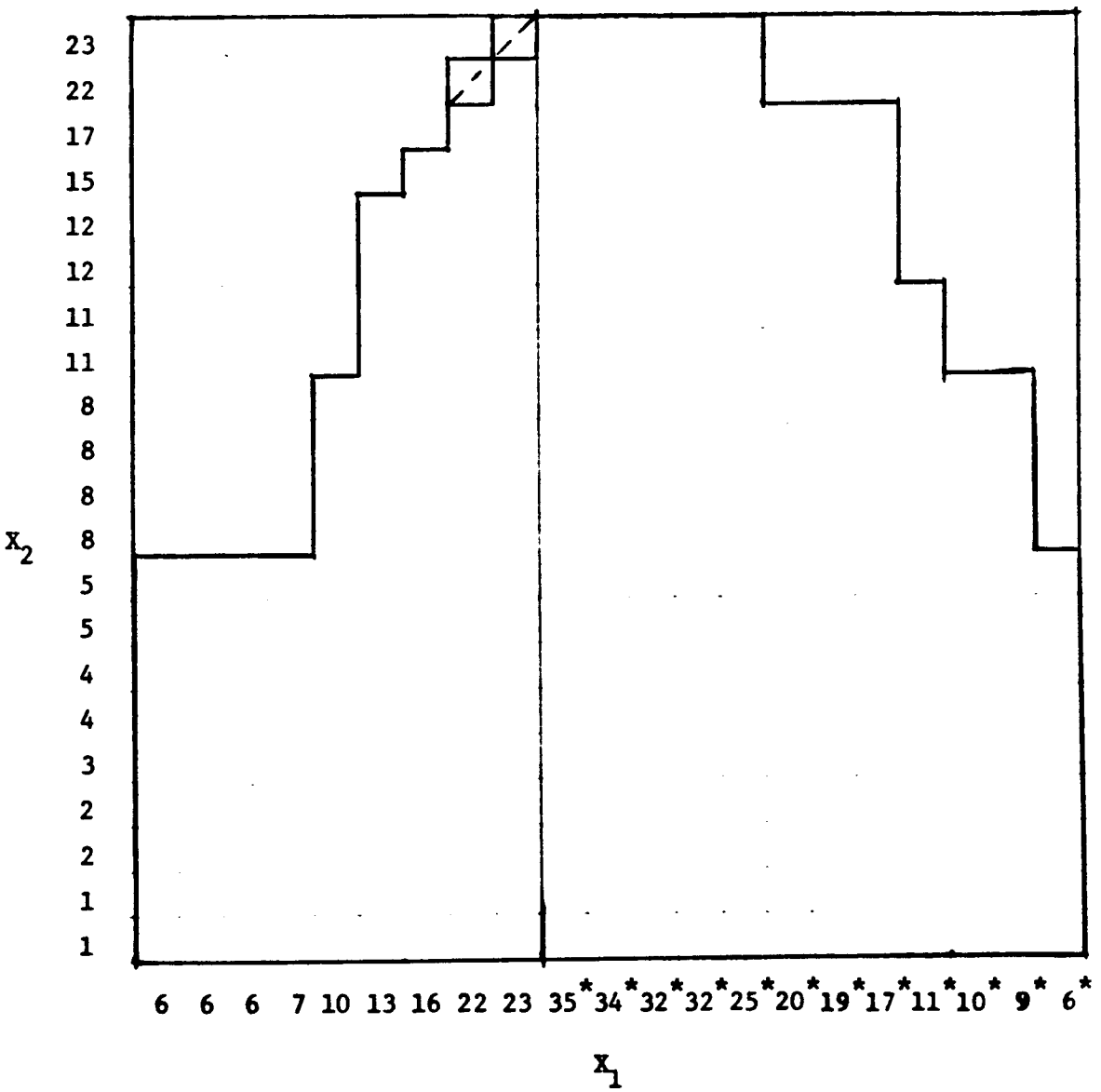
Table 2.1. The data of Freireich et al. (1963)

Sample 1: $6,6,6,7,10,13,16,22,23,35^*,34^*,32^*,32^*,25^*,20^*,19^*,17^*,11^*,10^*,9^*,6^*$.

Sample 2: 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23.

The first sample consists of lengths of remission in weeks for 21 patients with acute leukemia, maintained on the drug 6-MP, in which 12 are censored on the right. The second sample consists of the

Figure 2.2 The CPC-I for the data of Freireich et al. (1963)



lengths of unmaintained remission for 21 patients having complete observations. The CPC-I of these data is given in Figure 2.2. This figure shows a large difference between $A(X_1)$ and $A(X_2)$, which indicates that the treatment group (maintained patients) and the control group (unmaintained patients) do not have the same distribution functions.

From Figure 2.2, we can calculate $A(X_1)=336$ and $A(X_2)=65$. Hence, $W=336 - 65 = 271$, which is in fact the same as the result of Gehan. However, Gehan calculated the statistic as $W = 335 - 64 = 271$. The difference in calculation is the effect of the ties between uncensored observations, i.e., 22 and 23, in the two samples. As defined by Quade (1973), in a pair chart the square(s) corresponding to the ties are equally divided between $A(X_1)$ and $A(X_2)$. If there are no ties, $A(X_1)$ (or $A(X_2)$) is the total number of times the observed values on X_1 (or X_2) are larger than those on X_2 (or X_1).

Here, we should note that ties cannot occur between a censored observation and an uncensored one.

2.2.2 Alternative expression for W

Observing the CPC-I, $A(X_1)$ and $A(X_2)$ can be written as

$$A(X_1) = U(X_1) + C(X_1)$$

$$A(X_2) = U(X_2) + C(X_2) \qquad (2.2.1)$$

where $U(X_1)$ and $U(X_2)$ are the Mann-Whitney U-statistics corresponding to the uncensored observations with $U(X_1) + U(X_2) = (n_1-r_1)(n_2-r_2)$, and $C(X_1)$ (or $C(X_2)$) is the total number of times X^*_{1i}'s (or X^*_{2j}'s) are larger than X_{2j} (or X_{1i}). In Figure 2.2, we have $U(X_1)=124$,

$U(X_2)=65$, $C(X_1)=212$, and $C(X_2)=0$.

Thence the W-statistic of Gehan can be written as

$$W = (U(X_1) - U(X_2)) + (C(X_1) - C(X_2))$$
$$= 2U(X_1) + (C(X_1) - C(X_2)) - (n_1-r_1)(n_2-r_2) \quad (2.2.2)$$

2.2.3. The triplets statistic

Let $N(X_1,X_2,X_1)$ (or $N(X_2,X_1,X_2)$) be the number of ways in which it is possible to choose from the data 2 of X_1's and 1 of X_2's (or 2 of X_2's and 1 of X_1's) such that X_2 lies between the X_1's (or X_1 lies between the X_2's). Using Quade's formulas, with the aid of a CPC-I, we can calculate the values of the triplets $N(X_1,X_2,X_1)$ and $N(X_2,X_1,X_2)$. Note that the two observed polygons in a CPC-I have a common boundary, that is the path of the pair chart corresponding to the complete (uncensored) observations. In the j-th row, j=1,..., n_2-r_2, let L_j be the number of unit squares to the left of this path, and R_j^* to the right of it which belong to $A(X_1)$: if some unit-squares, say Q_j of them lie within a box corresponding to the ties between the X_1's and X_2's, count these as equally divided between L_j and R_j^*. Then

$$N(X_1,X_2,X_1) = \sum_{j=1}^{k} L_j R_j^* - \sum_{j=1}^{k} Q_j(Q_j+2)/12, \quad k=n_2-r_2 \quad (2.2.3)$$

Similarly,

$$N(X_2,X_1,X_2) = \sum_{i=1}^{m} B_i A_i^* - \sum_{i=1}^{m} H_i(H_i+2)/12, \quad m=n_1-r_1 \quad (2.2.4)$$

where B_i, A_i^*, and H_i are the unit squares below the path, above it belonging to $A(X_2)$, and boxed respectively, within the i-th column, i=1,...,n_1-r_1.

Extending the statistic of Crouse and Steffens (1969), as noted

by Quade, let

$$V = N(X_1,X_2,X_1) - N(X_2,X_1,X_2);$$

then V is a conditional statistic for testing H_0: $F_1(x)=F_2(x)$ against suspected differences in scale, assuming that X_1 and X_2 do not differ in location.

Using Figure 2.2, for the data of Freireich et al., we obtain:

$$\begin{aligned} N(X_1,X_2,X_1) &= 9(0)(21)+4(4)(10)+2(5)(13)+2(5)(12)+6(11)+(7)(10) \\ &\quad +(7.5)(6.5)+(8.4)(5.5) - 2(1(1+2)/12) \\ &= 641 \\ N(X_2,X_1,X_2) &= 4(9)(12)+(13)(8)+(17)(4)+(18)(3)+(19.5)(1.5)+ \\ &\quad +(20.5)(0.5) - 2(1(1+2)/12) \\ &= 694 \end{aligned}$$

Hence,

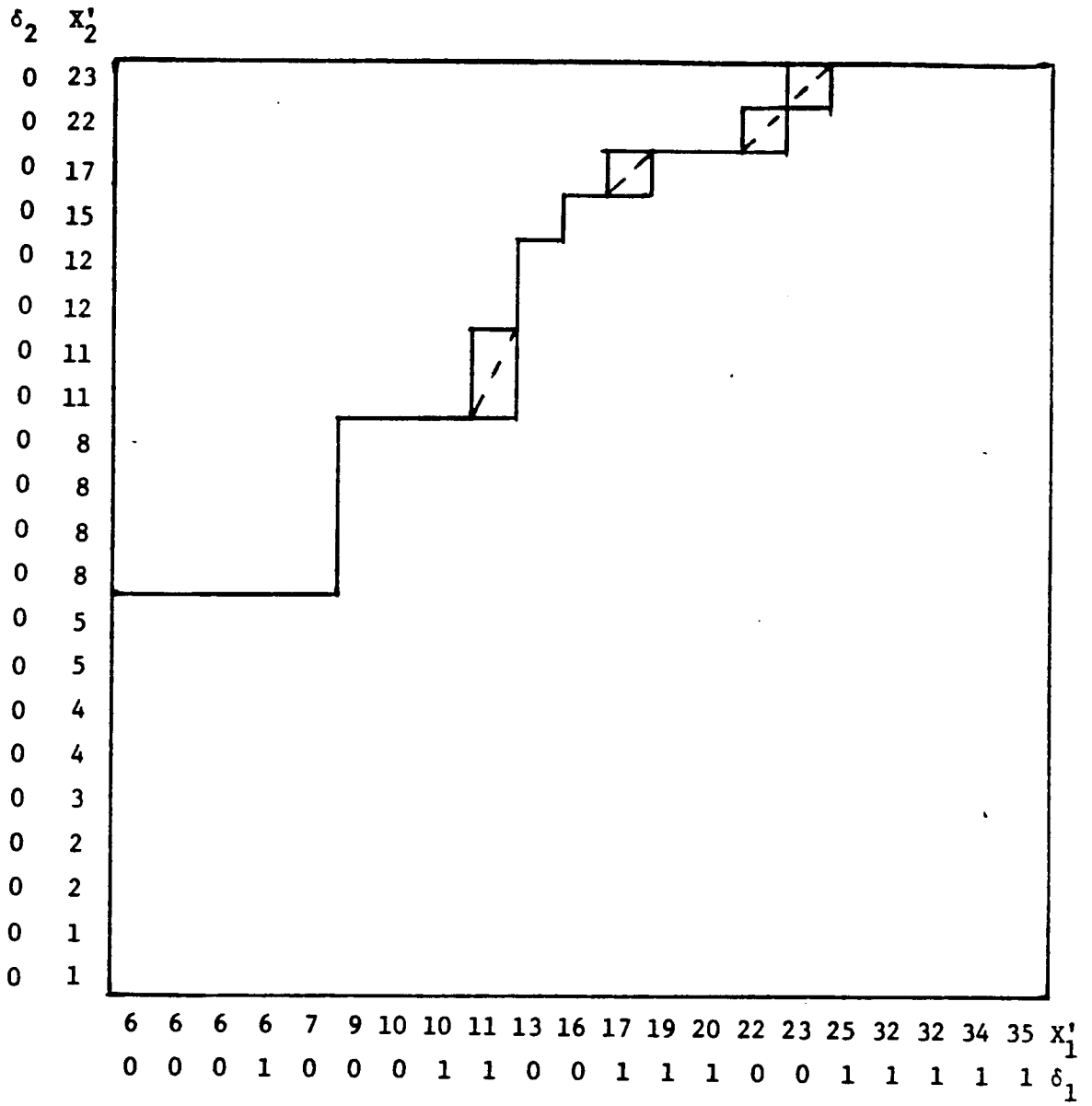
$$V = 641 - 694 = -53$$

2.3. Censored Pair Chart of Type-II

2.3.1. Discussion

As noted in section 2.1 a CPC-II is in fact a pair chart corresponding to the observed values on the random variables $X_i' = \min(X_i,T_i)$, i=1,2. However, the interpretation of the chart may depend on the indicator variables δ_i (or δ_{ij}, $j=1,\ldots,n_i$, i=1,2). The construction of the chart is the same as that of Quade, since we consider all observations on X_i' as uncensored observations. However, in addition to these observed values, on the X_i'-axis we would have the

Figure 2.3 The CPC-II for the data of Freireich et al. (1963)



corresponding values of δ_i. Figure 2.3 shows the CPC-II of the Freireich et al., data.

On a CPC-II, we observe the Mann-Whitney type U-statistic $U(X_1')$ (or $U(X_2')$) with $U(X_1') + U(X_2') = n_1 n_2$, where n_i is the sample size on X_i', i=1,2. If there are no ties between X_{1j}' and X_{2k}', then $U(X_1')$ (or $U(X_2')$) is the total number of the X_{1j}''s (or X_{2k}''s) are larger than X_{2k}' (or X_{1j}'). This U-statistic is the usual Mann-Whitney statistic for testing the null hypothesis H_0': $F_1'(x) = F_2'(x)$, under the assumption that $G_1(x) = G_2(x)$, i.e., the two samples have the same censoring function. In computing this statistic, we ignore complete the indicator variables δ_i, i=1,2.

Considering the data of Freireich et al., Figure 2.3 shows a 'large' value of $(U(X_1') - U(X_2'))$. As a descriptive statistic, this figure suggests the rejection of H_0'. The assumption of equal censoring for these data was noted by Breslow (1970). With the aid of Figure 2.3, we can calculate $U(X_1') = 334.5$, and then $U(X_2') = 441 - 334.5 = 106.5$. Note that the boxes corresponding to the ties between X_{1j}' and X_{2k}' at the observed values 11, 17, 22 and 23 are equally divided by $U(X_1')$ and $U(X_2')$. Thence, based on this value of $U(X_1') = 333.5$, we may use the usual Mann-Whitney U-statistic to test H_0': $F_1'(x) = F_2'(x)$ against the alternative $F_1'(x) \neq F_2'(x)$.

2.3.2. The conditional Mann-Whitney U-statistic

In section 3.1, we proposed the (unconditional) statistic $U(X_1')$ (or $U(X_2')$), which is independent of the indicator variables δ_i, i=1,2, for testing the hypothesis H_0: $F_1(x) = F_2(x)$ under the assumption $G_1(x) = G_2(x)$, i.e., the two samples have the same

censoring function. Without this assumption the statistic $U(X_1')$ would not be appropriate for testing H_0. However, the CPC-II can be used as an aid for computing the conditional Mann-Witney U-statistics:

$$\begin{aligned}
&\text{(i).} \quad U(X_i' \mid \delta_1,\delta_2) \\
&\text{(ii).} \quad U(X_i' \mid \delta_i) = \sum_{\delta_{i'}} U(X_i' \mid \delta_i,\delta_{i'}), \quad i\neq i'=1,2 \\
&\text{(iii).} \quad U(X_1' \mid \delta_i) = \sum_{\delta_i} U(X_i' \mid \delta_i,\delta_{i'}), \quad i\neq i'=1,2
\end{aligned} \tag{2.3.1}$$

This is accomplished by considering the ordered pairs (X_{ij}',δ_{ij}), $j=1,\ldots,n_i$, $i=1,2$ as the n_1,n_2 observations with $\delta_{ij} = 0$ or 1 according to whether X_{ij}' is uncensored or censored. In this case, for the ties between X_1' and X_2', we define

$$\begin{aligned}
&(X_i'=x,\delta_i=0) < (X_{i'}=x,\delta_{i'}=1), \text{ and} \\
&(X_i'=x,\delta_i) = (X_{i'}=x,\delta_{i'}) \text{ if } \delta_i = \delta_{i'}
\end{aligned} \tag{2.3.2}$$

Thence, if ties occur, the box(es) corresponding to the ties in a CPC-II should be adjusted accordingly. If there are no ties between X_1' and X_2' with $(X_i'=x,\delta_i=0) < (X_{i'}=x,\delta_{i'}=1)$, then

$$U(X_i') = U(X_i' \mid \delta_i=0) + U(X_i' \mid \delta_i=1), \quad i=1,2$$

Now, considering the statistics $U(X_i)$, $C(X_i)$, and $A(X_i)$, $i=1,2$, as proposed in Section 2.2, it is easy to verify that

$$\begin{aligned}
&U(X_i' \mid \delta_1=\delta_2=0) = U(X_i) \\
&U(X_i' \mid \delta_i=1,\delta_{i'}=0) = C(X_i) \ , \ i\neq i' \\
&U(X_i' \mid \delta_{i'}=0) = A(X_i) \ , \ i\neq i'
\end{aligned} \tag{2.3.4}$$

Hence the W-statistic of Gehan can be written as

$$W = U(X_1' \mid \delta_2=0) - U(X_2' \mid \delta_1=0) \tag{2.3.5}$$

This shows that W is in fact a conditional statistic, that is a difference between two conditional U-statistics corresponding to the variables $X_i' = \min(X_i, T_i)$, i=1,2.

2.4 The Pair Chart for Categorical Data

Censored categorical data in two-sample life testing problems would have the following format

	Sample 1		Sample 2	
Class	un-cen.	cen.	un-cen.	cen.
j	f_{1j}	c_{1j}	f_{2j}	c_{2j}

where f_{ij} = the number of uncensored observations in the i-th sample belonging to the j-th class (time interval)

c_{ij} = the number of censored observations in the i-th sample belonging to the j-th class

$i = 1, 2; j = 1,\ldots,k.$

In order to construct a censored pair chart for these data, we consider all uncensored observations in the j-th class as having the same value for each j. Without loss of generality, we may define j as the value of the uncensored observations in the j-th time interval, provided that the time intervals are increasing with j. For the censored observations in the j-th time interval, we define c^* as their value. The value of c^* will be determined by the method of counting chosen for the occurrence of the censored observations. Here, we will consider the following four possible methods:

(1) Gehan's method.

He considered the censored observations as tied in the j-th class, and counted them as occurring after class (j-1) but before j. Using our notation, $(j-1)^*$ is defined to be the values of the censored observations in the j-th class.

(2) The censored observations in the j-th class are considered as tied with observed values j^*. The reason for this assumption is that a right-censored observation in the j-th class is considered as occurring after uncensored observed values in the j-th class.

(3) We define the time intervals corresponding to all uncensored observations. The censored observations are divided as follows. If X_i^* is a censored observation, and m_j is the mid-point of the j-th time interval such that $m_j < X_i^* < m_{j+1}$, then j^* is defined to be its value, or X_1^* is counted as occurring after the j-th interval but before the (j+1)-th.

(4) The censored observations in the j-th interval are considered as tied having the same values as those of the uncensored observations, that is j, but the number of censored observations, c_{ij}, is reduced to $c_{ij}/2$ if c_{ij} is even, and $(c_{ij}+1)/2$ if it is odd. In this case, the censored data are in fact transformed into uncensored data.

Considering fixed time-intervals, or the same time-intervals

for the four methods, let

$$W_k = (U_k(X_1) - U_k(X_2)) + (C_k(X_1) - C_k(X_2)) \qquad (2.4.1)$$

be the value of the W statistic computed using the k-th method of counting. Since in the fourth method, the censored observations are transformed into uncensored observations, $C_4(X_1) - C_4(X_2)$ will be defined as the value contributed by those observations in the transformed data. It is clear that $U_k(X_1) - U_k(X_2)$ would have the same values for all k. But $C_k(X_1) - C_k(X_2)$ vary. It is easy to verify that $C_k(X_i)$, i=1,2; k=1,2,3,4 satisfy the following inequalities

$$\begin{aligned} &\text{(i)} \quad C_1(X_i) \leq C_3(X_i) \leq C_2(X_i) \\ &\text{(ii)} \quad C_4(X_i) < C_3(X_i) \end{aligned} \qquad (2.4.2)$$

In (i), $C_1(X_i) = C_3(X_i)$ (or $C_3(X_i) = C_2(X_i)$) is attainable if the censored observations occur within the first (or second) half of the j-th time interval, for all j. However, $C_1(X_i)$ is strictly less than $C_2(X_i)$. So, we may consider the third method as a 'compromise' between the first two methods. Now, how about the values of W_k? Based on these inequalities we could not derive any general relation between the W_k's, except for the following particular case.

If all observations on X_2 are uncensored, hence $C_k(X_2)=0$, using the same time-intervals, then W_k is a function of $C_k(X_1)$ for each k. Thence

$$\begin{aligned} &(1) \quad W_1 \leq W_3 \leq W_2 \\ &(2) \quad W_4 < W_3 \end{aligned} \qquad (2.4.3)$$

As an illustration, we will consider again the data of Freireich et al. in Table 2.1, which were analyzed as grouped data by Gehan (1965). Following Gehan's procedure, we would have the grouped data

as in Table 2.2,

Table 2.2. Grouped data of Freireich et al. (1963)

Class j	Interval (weeks)	Sample 1 (6-MP) un-cen. f_{1j}	Sample 1 (6-MP) un-cen. v_{1j}	Sample 1 (6-MP) cen. c_{1j}	Sample 1 (6-MP) cen. v^*_{1j}	Sample 2 (Placebo) un-cen. f_{2j}	Sample 2 (Placebo) un-cen. v_{2j}	Sample 2 (Placebo) cen. c_{2j}	Sample 2 (Placebo) cen. v^*_{2j}
1	0 - 4	0	-	0	-	7	1	0	-
2	5 - 9	4	2	2	1^*	6	2	0	-
3	10 - 14	2	3	2	2^*	4	3	0	-
4	15 - 19	1	4	2	3^*	2	4	0	-
5	20 - 24	2	5	1	4^*	2	5	0	-
6	25+	0	-	5	5^*	0	-	0	-

where

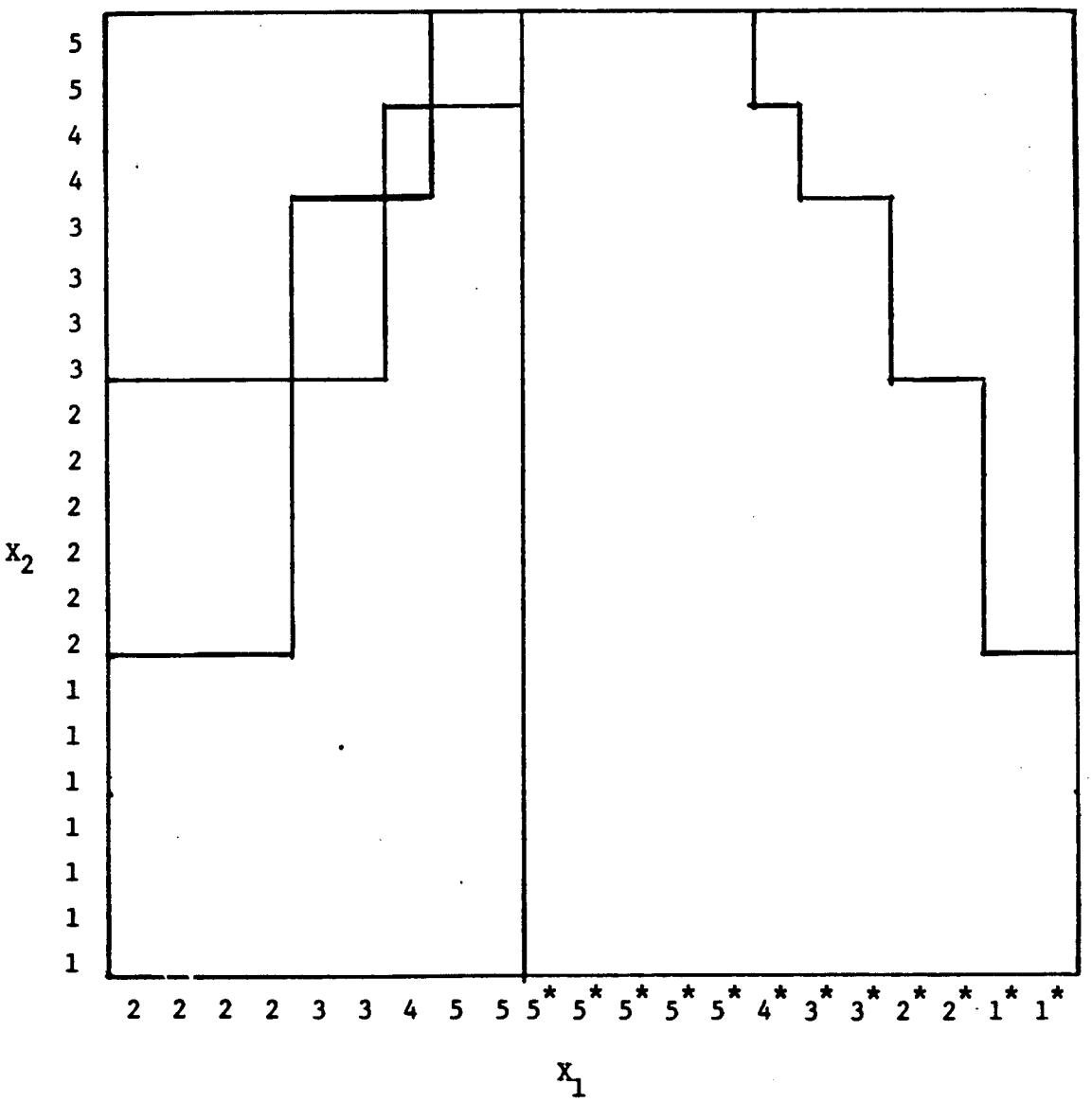
$v_{1j} = v_{2j}$ is the value defined for uncensored observations in the j-th interval.

$v^*_{1j} = v^*_{2j}$ is the value defined for censored observations in the j-th interval.

The CPC-I of these data is given in Figure 2.4. As a descriptive statistic, Figure 2.4 shows a large difference between $A(X_1)$ and $A(X_2)$, hence we may reject H_0. With the aid of this CPC, we calculate $A(X_1) = 326$ and $A(X_2) = 61$. Thence $W_1 = 326 - 61 = 265$, as computed by Gehan.

The construction of the CPC for grouped data based on the other methods is straightforward. We would obtain $W_2 = 289$, $W_3 = 273$, and $W_4 = 264$. Note that the values of W_k satisfy the inequalities (2.4.3). Comparing with $W = 271$ for ungrouped data, the third method gives a value which is very close to W, and the fourth method gives

Figure 2.4 The CPC-I for grouped data of Freireich et al. (1963)



the smallest value of 264, which is smaller than that of Gehan.

Remark.

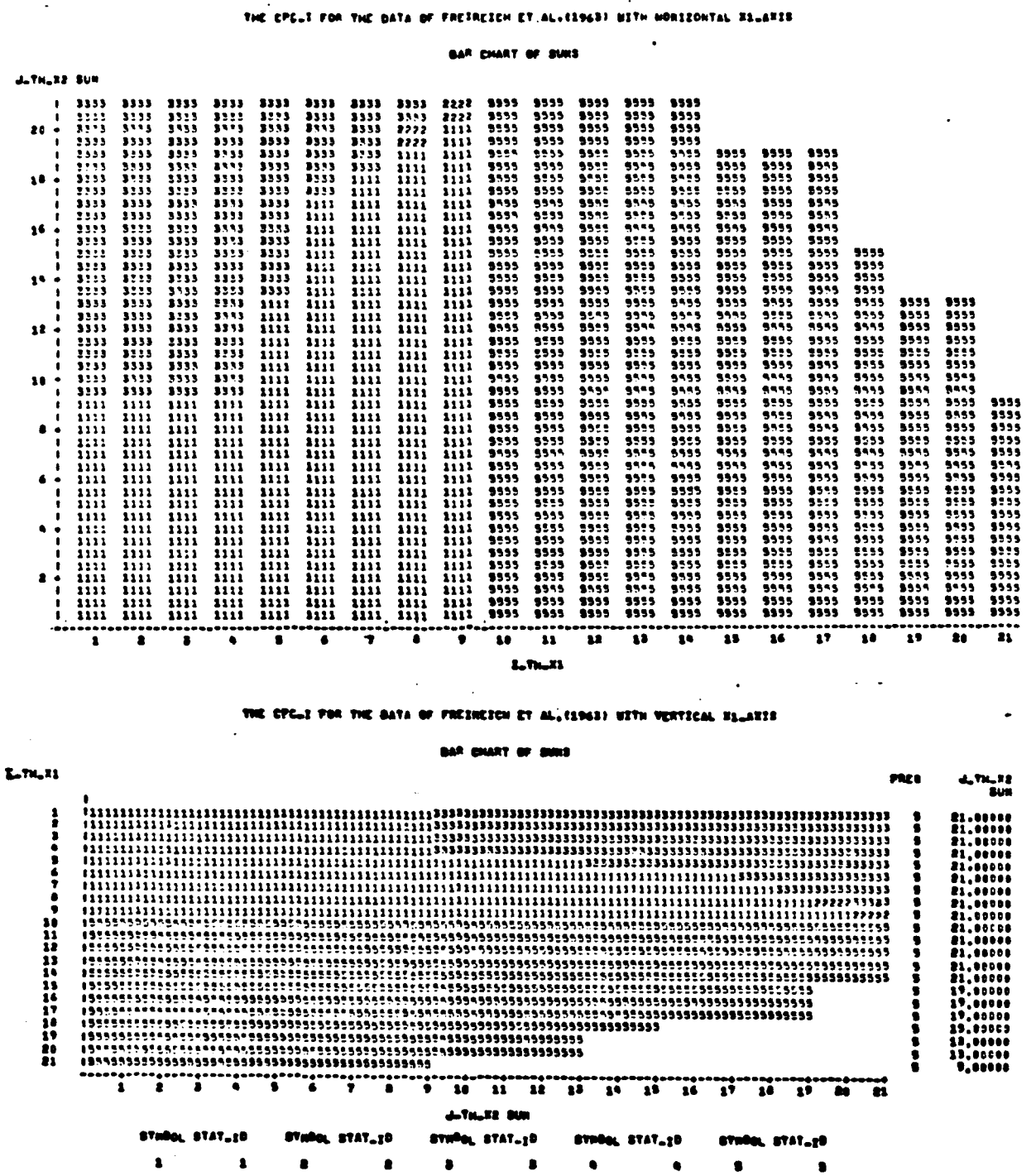
Discrete data can be considered as a special case of categorical data with all intervals of length zero. In this case, we assume that the i-th population, i-1,2, has discrete distribution function F_i, and the F_i's are purely discontinuous with the same finite set of discontinuity points. Hence, the construction of the CPC for discrete data is the same as that for categorical data.

2.5 Computer plotting of the pair chart

With the help of a SAS program, the CPC can be presented as a bar chart, either vertical or horizontal. Here, we should note that the chart procedure in SAS could not give the printout of the observations with special ordering as defined in Section 2.1 of the X_i-axis. Thence, in the bar chart, the special-ordered n_i-observations on X_i will be presented by the integers $1,2,\ldots,n_i$ along the X_i-axis. This computer plotting is necessary, especially for large sample sizes.

For an illustration, we will consider the CPC-I for the data of Freireich et al. Using the SAS program, given as an appendix, we obtain the vertical and horizontal bar charts in Figure 2.5. Note that the X_1-axis is horizontal in the vertical bar chart, and it is vertical in the horizontal bar chart. Furthermore, the bars in each chart are divided in general into five kinds of sub-bars (=sub-regions) having symbols (=STAT_ID) 1, 2, 3, 4, or 5. The sub-regions having symbols 1, 2, and 3 correspond to un-censored pairs of observations (X_1,X_2) such

Figure 2.5. A Copy of the Bar Charts as an Alternative Presentation of the CPC-I



that $X_1 > X_2$, $X_1 = X_2$, $X_1 < X_2$, respectively. The area of the sub-region with STAT_ID = 4 (or 5) denotes the statistic $C(X_2)$ (or $C(X_1)$). For the data of Freireich et al., $C(X_2) = 0$.

If there are no ties, then the sub-regions having symbols 1 and 3 will correspond to $U(X_1)$ and $U(X_2)$, the Mann-Whitney U-statistics for the uncensored observations. Thence the statistic $A(X_1)$ (or $A(X_2)$) is presented by the total area of the sub-regions having symbols 1 and 5 (or, 3 and 4).

The program would print also the corresponding values of the statistics $U(X_i)$, $C(X_i) = A(X_i)$, i=1,2; and W.

2.6 The Maximum Distance, D, Statistic

2.6.1 Discussion

Quade (1973) introduced the use of the pair chart for computing the Kolmogorov-Smirnov statistic. He considered the lattice points of the path which are the farthest below or above the diagonal of the $n_1 x n_2$-rectangle. Using these points he computed the maximum vertical distance between the corresponding two empirical distribution functions. Here, we will consider the point on the path which is the farthest from the diagonal of the rectangle. This leads to the maximum distance, D, statistic.

This D statistic is defined corresponding to the CPC-II or the pair chart of Quade (1973). Considering the CPC-II of the two samples of sizes n_1 and n_2, let $D(X_1', X_2')$ be the distance from the lattice point $(X_1', X_2'))$ on the path of a CPC-II to the line $n_1X_2' = n_2X_1'$. Then we define

$$D = \text{Max}\{D(X_1', X_2')\} \tag{2.6.1}$$

For $X_i' = 1,\ldots,n_i$; i=1,2, it is easy to verify that

$$\min_i \{n_i/\sqrt{n_1^2 + n_2^2}\} \leq D \leq n_1 n_2/\sqrt{n_1^2 + n_2^2} \tag{2.6.2}$$

and

$$D(X_1', X_2') = \frac{n_1 n_2}{\sqrt{n_1^2 + n_2^2}} \left| X_1'/n_1 - X_2'/n_2 \right| \tag{2.6.3}$$

Thence

$$D = \frac{n_1 n_2}{\sqrt{n_1^2+n_2^2}} \quad \text{Max} \left| X_1'/n_1 - X_2'/n_2 \right| \tag{2.6.4}$$

for fixed sample sizes. Note that the maximum value of $\left|X_1'/n_1 - X_2'/n_2\right|$ for lattice points (X_1', X_2') on the path determines the Kolmogorov-Smirnov statistic for testing H_0': $F_1'(x) = F_2'(x)$, as noted by Quade (1973).

So, under the assumption that the two samples have the same censoring function (see Section 2.3), this D-statistic is appropriate for testing the null hypothesis H_0: $F_1(X) = F_2(X)$. Large value of D suggests the rejection of H_0.

2.6.2 The Distribution of the D-Statistic

2.6.2.1 For Equal Sample Sizes

If $n_1=n_2=n$, then

$$D_{(i)} = i/\sqrt{2} \tag{2.6.5}$$

where $D_{(i)}$ is the i-th ordered possible value of the D-statistic, for i-1,...,n. Using Theorem 2, Chapter 1 of Mohanty (1979), under the null hpyothesis H_0', we obtain:

$$P(D<i\sqrt{2}) = \begin{cases} 0 & ; \ i \leq 1 \\ L\ (n;i)\ /\ \binom{2n}{n} & ; \ 1 < i \leq n. \\ 1 & ; \ i > n. \end{cases} \qquad (2.6.6)$$

with

$$L(n;i) = \sum_k \{\binom{2n}{n-2ik}_+ - \binom{2n}{n+i+2ik}_+\} \qquad (2.6.7)$$

where

$$\binom{y}{z}_+ = \begin{cases} \binom{y}{z} & \text{when } y \geq z \\ 0 & \text{when } y < 0 \text{ or } y < z \\ 1 & \text{when } z = 0 \end{cases} \qquad (2.6.8)$$

and the summation is over all integer values of k: positive, negative or zero.

Note that L(n;i), which is the $|L(n,n;i,i)|$ of Mohanty (1979), denotes the number of paths from the origin to (n,n) that do not touch the lines $X_2' = X_1' + i$ and $X_2' = X_1' - i$. Thence,

$$P(D=i/\sqrt{2}) = P(D<(i+1)/\sqrt{2}) - P(D<i/\sqrt{2}) \qquad (2.6.9)$$

for i=1,2,...,n.

For i>n/2, we may consider applying formula (4.9) of Mohanty. We obtain that the number of paths from (0,0) to(n,n) that touch the line $X_2' = X_1' + i$ <u>exactly</u> r (>0) times is given by

$$\frac{2i+r-1}{2n-r+1} \times \binom{2n-r+1}{n-i-r+1} \qquad (2.6.10)$$

Note that, in this case, i>n/2, a path which touches the line $X_2' = X_1' + i$ will not touch or cross the line $X_2' = X_1' - i$ for r = 1, 2,..., (n+1-i). Now, let

$$L_{(i)} = \sum_{r=1}^{n+1-i} \frac{2i+r-1}{2n-r+1} \binom{2n-r+1}{n-i-r+1} \qquad (2.6.11.a)$$

which can be simplified as

$$L_{(i)} = \begin{cases} 1 + \sum_{k=o}^{n-i-1} \{\binom{2n-k-1}{n+i-1} - \binom{2n-k-1}{n+i}\}, & \frac{n}{2} < i \leq n-1 \\ 1 \quad , \; i=n \end{cases} \qquad (2.6.11.b)$$

Then

$$P(D=i/\sqrt{2};\ n/2 < i \leq n) = 2L_{(i)} / \binom{2n}{n} \qquad (2.6.12)$$

Furthermore, let

$$L_{(j\geq i)} = \sum_{j=i}^{n} L_{(i)} \qquad (2.6.13.a)$$

Then we obtain

$$L_{(j\geq i)} = \sum_{k=0}^{n-i-1} \binom{n+i+k}{k+1} \qquad (2.6.13.b)$$

Hence

$$P(D\geq i/\sqrt{2};\ n/2 < i \leq n) = 2L_{(j\geq i)} \; / \binom{2n}{n}. \qquad (2.6.14)$$

2.6.2.2 For Unequal Sample Sizes

In this case, we would consider in general the level of significance for the D-test, instead of the probability function of the D-statistic. Using Hodges' (1958) method, presented by Quade (1973), we can calculate the level of significance for either the one-sided or the two-sided D test. Let $L(X_1', X_2')$ be the number of possible paths or routes from the origin to the point (X_1', X_2') then we have the recursion formula

$$L(X_1', X_2') = L(X_1'-1, X_2') + L(X_1', X_2'-1). \qquad (2.6.15)$$

with initial condition $L(0,0) = 1$. This formula is presented also by Monahty (1979).

A particular path of the CPC-II gives rise to a value $D \geq c/\sqrt{n_1^2+n_2^2}$ if and only if it reaches the line $n_2X_1' - n_1X_2' = c$ or the line $n_2X_1' - n_1X_2' = -c$. Under H_0', the $(n_1+n_2)!/(n_1!n_2!)$ possible paths are equally likely; thence the significance level associated with $D = c/\sqrt{n_1^2+n_2^2}$ is

$$P = 1 - \frac{L(n_1,n_2)\, n_1!n_2!}{(n_1+n_2)!} \tag{2.6.16}$$

the value of $L(n_1,n_2)$ can be computed using the recursion formula (2.6.15), with boundary conditions that

(i) $L(X_1',X_2') = 0$ for $X_i' < 0$; $X_i' > n_i$;
or $|n_2X_1' - n_1X_2'| \geq c$ if two sided D test;
that is to test $H_0':F_2' = F_2'$ against $H_1': F_1' \neq F_2'$

(ii) $L(X_1',X_2') = 0$ for $X_i' < 0$; $X_i' > n_i$, or
$n_2X_1' - n_1X_2' \leq -c$ to test H_0' against H_1':
$F_1' > F_2'$; that is the one-sided D test.

Since $L(n_1-X_1',n_2-X_2')$ denotes also the paths from (X_1',X_2') to (n_1,n_2), the recursion formula need not be carried out beyond the line $X_1'+x_2' = (n_1+n_2+1)/2$.

Considering all possible paths from (0,0) to (X_1',X_2') on the line $X_1'+X_2' = c$. For $c > 0$, we would have the well known Pascal triangle. It is generated by the recursion formula (2.6.15). However, in computing the level of significance for the D test corresponding to any observed values we would have an '<u>incomplete</u>' Pascal triangle, which depends on the sample sizes and the boundary conditions: (i) or (ii).

By observing all possible lines $n_2X_1' - n_1X_2' = c$ for $0 < c \leq n_1n_2$ containing some lattice point(s), we may construct the corresponding incomplete Pascal triangle. Using this triangle, we can compute the probability function of the D statistic. The points on the boundary lines are associated with a value of the D statistic. Let c_k be the value of c associated with the k-th ordered value of D, that is $D_{(k)}$, for $k = 1,2,\ldots,N(n_1,n_2)$; then

$$D_{(N)} = n_1n_2/\sqrt{n_1^2+n_2^2} \tag{2.6.17}$$

is the largest value of D. Let $L_k(n_1,n_2)$ be the number of paths from (0,0) to (n_1,n_2) between the lines $|n_2X_1' - n_1X_2'| = c_k$, including the paths giving the value of $D = D_{(k)}$; then

$$P(D=D_{(k)}) = \{L_k(n_1,n_2) - L_{k-1}(n_1,n_2)\}\frac{n_1!n_2!}{(n_1+n_2)!} \tag{2.6.18}$$

and

$$\begin{aligned} L_N(n_1,n_2) &= (n_1+n_2)!/(n_1!n_2!) \\ L_N(n_1,n_2) - L_{N-1}(n_1,n_2) &= 2L(n_1,0)L(0,n_2). \\ &= 2. \end{aligned} \tag{2.6.19}$$

where $L_N(n_1,n_2) = L(n_1,n_2)$ is the number of all possible paths from (0,0) to (n_1,n_2), and $L(n_1,0)$ is the number of paths from (0,0) to $(n_1,0)$ or from $(0,n_2)$ to (n_1,n_2). Similarly, $L(0,n_2)$ is the number of paths from (0,0) to $(0,n_2)$ or from $(n_1,0)$ to (n_1,n_2). The values of $L_k(n_1,n_2)$, for $k=1,\ldots,N-2$, could be computed using the incomplete Pascal triangles.

In fact, the boundary lines corresponding to each incomplete Pascal triangle can be written as

$$D_{(k)} = | X_1' \sin\alpha - X_2' \cos\alpha | \qquad (2.6.20)$$

with $\tan\alpha = n_2/n_1$.

These equations are called the normal equations of parallel lines; and $D_{(k)}$ denotes the distances of the lines from the origin. And, it is clear that each boundary line should contain at least one lattice point. If $n_1=n_2=n$, then $N = \max(k) = n$ and each boundary line has $(n-k+1)$ lattice points, for $k=1,\ldots,n$.

Another method for computing the probability function of the D-statistic is based on the vector representation of Mohanty (1979).

2.7. Vector Representation for D-Statistic

Following Mohanty's notation, the sample sizes will be written as $n_1=m$ and $n_2=n$. Then a path in a CPC-II can be represented as a vector $(a_1,a_2,\ldots,a_n) = \underset{\sim}{a}$, where $a_i (i=1,\ldots,n)$ is the minimal distance, measured parallel to the X_1'-axis, of the points $(m,n-i)$ from the path, such that

(i) a_i is an integer, $i=1,\ldots,n$

(ii) $0 \leq a_1 \leq \ldots \leq a_n$

provided there are no tied observations. In this case, $a_1 + a_2 + \ldots a_n = U(X_1')$, the Mann-Whitney statistic.

For further discussion, the following definition and theorem of Mohanty (1979) are needed.

Definition 2.7.1.

A path $(x_1,\ldots,x_n)$ dominates the path $(y_1,\ldots,y_n)$ if and only if $y_i \leq x_i$ for all i.

Theorem 2.7.1

$$|(\underset{\sim}{a})| = \underset{n\times n}{\mathrm{Det}}(d_{ij}) \tag{2.7.1}$$

where

$$d_{ij} = \begin{cases} \binom{a_n+n}{n} & i=j=1 \\ \binom{a_{n-1}+n-i}{n-i}_+ & i\neq 1, j=1 \\ \binom{a_n-a_{n-j+1}+j-1}{j}_+ & i=1, j\neq 1 \\ \binom{a_{n-1}-a_{n-j+1}+j-i+1}{j-i} & i\neq 1, j\neq 1 \end{cases} \tag{2.7.2}$$

with $\binom{y}{2}_+$ defined by (2.6.8).

In this theorem $|(\underset{\sim}{a})|$ denotes the number of paths dominated by the vector $\underset{\sim}{a}$.

Associated with the boundary line $X_1'\sin\alpha - X_2'\cos\alpha + D_{(k)} = 0$, as given in section 2.6, let $\underset{\sim}{a}_{(k)}$ be the path dominates all paths from (0,0) to (m,n) laying in the positive side of this boundary line, then

$$C|\underset{\sim}{a}_k| = \binom{m+n}{n} - |(\underset{\sim}{a}_{(k)})| \tag{2.7.3}$$

denotes the number of paths crossing this boundary line. These paths may or may not cross the other boundary line $X_1'\sin\alpha - X_2'\cos\alpha - D_{(k)} = 0$. Hence, we obtain

$$P\left(D > D_{(k)};\ \tfrac{1}{2} \leq \frac{D_{(k)}\sqrt{m^2+n^2}}{mn} \leq 1\right)$$
$$= 2\left\{\binom{m+n}{n} - |(\underset{\sim}{a}_{(k)})|\right\} m!n!/(m+n)! \tag{2.7.4}$$

Since associated with $\frac{1}{2} \leq D_{(k)}\sqrt{m^2+n^2}/mn \leq 1$ the paths $C(a_k)$ will not cross the last boundary line. For other values of D, that is $0 < D_{(k)} < mn/2\sqrt{m^2+n^2}$, we should consider another result of Mohanty:

$$|(\underset{\sim}{b},\underset{\sim}{a})| = |(\underset{\sim}{a})| + |(a_n-b_n, a_n-b_{n-1}, \ldots, a_n-b_1)| - \binom{a_n+n}{n} \tag{2.7.5}$$

representing the number of paths which dominate $\underset{\sim}{b}$ and are dominated by $\underset{\sim}{a}$, when no path can cross both boundaries. Now, let $\underset{\sim}{b}_{(k)}$ be a path on the negative side of the line $X_1'\sin\alpha - X_2'\cos\alpha - D_{(k)} = 0$, which is dominated by all other paths on this side, then

$$C(\underset{\sim}{b}_k, \underset{\sim}{a}_k) = \binom{m+n}{n} - |(\underset{\sim}{b}_{(k)}, \underset{\sim}{a}_{(k)}| \tag{2.7.6}$$

denotes the number of possible paths which cross the boundary lines $|X_1'\sin\alpha - X_2'\cos\alpha| = D_{(k)}$. Hence,

$$P(D>D_{(k)};\ 0< \frac{D_{(k)}\sqrt{m^2+n^2}}{mn} < \tfrac{1}{2})$$

$$= \{\binom{m+n}{n} - |(\underset{\sim}{b}_{(k)}, \underset{\sim}{a}_{(k)})|\}m!n!/(m+n)! \tag{2.7.7}$$

In fact, this formula holds for all values of D. However, formula (2.7.4) is simpler to use for large values of $D_{(k)}$. So, we have

$$P(D\geq D_{(k)}) = \frac{\{\binom{m+n}{n} - |(\underset{\sim}{b}_{(k)}, \underset{\sim}{a}_{(k)})|\}m!n!}{(m+n)!} \tag{2.7.8}$$

and

$$P(D=D_{(k)}) = P(D>D_{(k-1)}) - P(D>D_{(k)}). \tag{2.7.9}$$

2.8 Large Approximation for Equal Sample Sizes

For large n, so i is also large, using Sterling's formula,

(2.6.13) can be written as

$$L_{(j\geq i)} \simeq \sum_{k=0}^{n-i-1} (1- \frac{k+1}{n+i+k})^{-n-i-k-\frac{1}{2}} (n+i-1)^{-k-1} \frac{e^{-k-1}}{(k+1)!} \tag{2.8.1}$$

Since

$$\lim_{\substack{n\to\infty \\ i\to\infty}} (1- \frac{k+1}{n+i+k})^{-n-i-k-\frac{1}{2}} = e^{k+1} \tag{2.8.2}$$

For each k, then (2.8.1) can be approximated as

$$L_{(j\geq i)} \simeq \sum_{k=0}^{n-i-1} \frac{(n+i-1)^{-k-1}}{(k+1)!}$$

$$= (1 + \frac{1}{n+i-1})^{n-i} - 1 \tag{2.8.3}$$

Furthermore, for $n/2 < i < n$, we could have

$$\lim_{n\to\infty} i/n = p; \ \tfrac{1}{2}<p<1. \tag{2.8.4}$$

with

$$L_{(j\geq i)} \simeq \exp(\frac{1-p}{1+p}) - 1 \tag{2.8.5}$$

Thence, for (2.6.14), we have two possible approximations for large n, that is

$$P(D\geq i/\sqrt{2};\ n/2<i<n) \simeq \{(1+ \frac{1}{n+i-1})^{n-i} - 1\}\ 2^{-2n+1}\sqrt{\pi n} \tag{2.8.6}$$

if (n-i) is small, and

$$P(D\geq i/\sqrt{2};\ n/2<i<n) \simeq \{\exp(\frac{1-p}{1+p}) - 1\}\ 2^{-2n+1}\sqrt{\pi n} \tag{2.8.7}$$

otherwise.

2.9. Large Unequal Sample Sizes

In contrast with the previous section, here we will not propose an approximation formula for the level of significance for the D test, that is (2.7.8). But we will illustrate some facts, which cause the approximation is so complex to derive.

First, let D (m,n; x,y) be the value of D corresponding to a fixed point (x,y) in the path of the pair chart of the samples of sizes $m \neq n$. Then in general

$$D(m_1,n_1;x,y) \neq D(m_2,n_2;x,y) \tag{2.9.1}$$

except if $m_1 n_2 = m_2 n_1$. Figure 2.6 illustrates the cases for $m_1=m_2=8$; $n_1=6$, $n_2=5$, and (x,y) = A(2,4). This figure shows

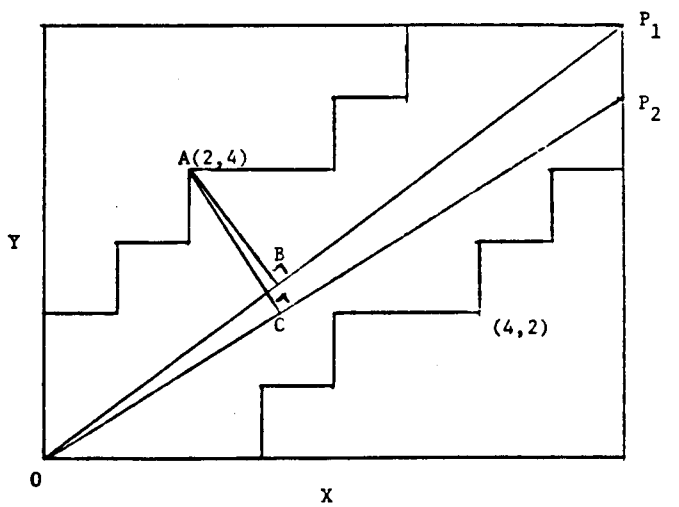
$$D(8,6;2,4) = AB < AC = D(8,5;2,4). \tag{2.9.2}$$

On the other side

$$D(8,6;4,2) > D(8,5;4,2). \tag{2.9.3}$$

Using (2.6.20), we obtain AB = 2 and AC = 22 $\sqrt{89}$.

Figure 2.6 Illustrative CPC-II for (m,n) = (8,6) and (8,5)



This affects the corresponding boundary lines (2.6.20) at each point (x,y). Furthermore, we could note that

$$|D(m_1,n_1;x,y) - D(m_2,n_2;x,y)| \qquad (2.9.4)$$

is a function of (x,y) beside (m_i,n_i), i=1,2; and it is not a constant, even for fixed values of (m_1,n_1) and (m_2,n_2). Figure 2.7 shows that (2.9.4) increases if $x^2 + y^2$, or the distance of (x,y) to (0,0), increases.

Finally, associated with formula (2.7.9) we should consider the paths represented by the vectors $\underset{\sim}{a}_k$ and $\underset{\sim}{b}_k$. Figure 2.7 shows the paths

$$\underset{\sim}{a}_k(8,6) = (3,4,6,7,8,8).$$
$$\underset{\sim}{b}_k(8,6) = (0,0,1,2,4,5). \qquad (2.9.5)$$

corresponding to the value of $D_k(8,6) = 2$, where (8,6) on the left hand sides indicate the sample sizes m,n. It is clear, these vectors are affected by the boundary lines (2.6.20), and (2.9.4) as well.

This leads to the complexity of computing the values of $|\underset{\sim}{a}_k,\underset{\sim}{b}_k|$ for different large values of m,n. So, this problem is still open.

CHAPTER III

MEASURES OF ASSOCIATION FOR GENERAL RIGHT-CENSORED BIVARIATE SAMPLES

3.1 Introduction

A nonparametric measure of association between the components of a bivariate random variable $(X_1,X_2) = \underline{X}$ is proposed for general right-censored samples, as an extension of Kendall's (1938) tau. Such a measure is called a '<u>correlation</u>,' written $t_c(X_1,X_2) = t_c$ where (t) indicates Kendall's tau, and the subscript (c) indicates '<u>censored</u>' samples.

Considering a censored bivariate sample of size n; (X_{1i}, X_{2i}), i=1,2,...,n; we define an indicator variable

$$\delta_{ri} = \begin{cases} 0 \text{ if } X_{ri} \text{ uncensored} \\ 1 \text{ if } X_{ri} \text{ censored} \end{cases} \qquad (3.1.1)$$

for r = 1 or 2. Then, the i-th bivariate sample observation will be written as

$$\{(X_{1i},\delta_{1i}),(X_{2i},\delta_{2i})\} = \underline{X}_i,\underline{\delta}_i = \underline{X}_i(\underline{\delta}_i). \qquad (3.1.2)$$

Using this notation, we are defining that an observation on X_r is an ordered pair (X_{ri},δ_{ri}). Hence, a censored observation $(X_{ri},1)$ cannot be smaller than an uncensored observation $(X_{rj},0)$, since we are considering only right-censored samples. Note that, if X_{ri} is

censored, then X_{ri} is the point (= value) at which it was censored.

For simplicity, whenever there is no confusion, the i-th observation will be written as $\underline{X}_i = (X_{1i}, X_{2i})$. Without loss of generality, we may assume the n observations have the following ordering.

$$\underline{X}_i = \begin{cases} \underline{X}_i(0,0), & i=1,\ldots,k. \\ \underline{X}_i(0,1), & i=k+1,\ldots,\ell. \\ \underline{X}_i(1,0), & i=\ell+1,\ldots,m. \\ \underline{X}_i(1,1), & i=m+1,\ldots,n. \end{cases} \qquad (3.1.3)$$

In this paper, the general right-censored sample will be called sample, in short. And the sample has an uncensored (bivariate) sub-sample having observations $\underline{X}_i = \underline{X}_i(0,0)$, i=1,...,k., and a censored (bivariate) sub-sample having observations $\underline{X}_i$, i=k+1,...,n.

If m=n, then the censoring occurs only on either one component of the bivariate observation $\underline{X}_i$, that is if X_{ai} is censored, then X_{bi}, $a \neq b=1,2$ is uncensored. This kind of sample will be called an A-sample. If m=k, so we only have observations $\underline{X}_i(0,0)$, i=1,...,k, and $X_i(1,1)$, i=k+1,...,n, then the sample will be called a B-sample; both components are either censored or uncensored.

Furthermore, the censoring may occur after the k_r-th order uncensored (complete) observations on the component X_r, r=1,2. In this case, we would have $(X_{ri},0) < (X_{rj},1)$, for all $i=1,\ldots,k_r$; $j=k_r+1,\ldots,n$.

Finally, a modification of the generalized Kendall's tau, t_c, will be proposed in section 3.6. This modified index of correlation will be called the index Alpha, α. It is expected that Alpha could

reach the values -1 and +1 for any general right-censored data.

3.2 The Generalized Kendall's Tau

3.2.1 Definitions

Considering a censored bivariate sample of size n on variable $\underline{X} = (X_1, X_2)$ given in Section 3.1, we may consider the following measures of association between X_1 and X_2, depending on the treatment of the censored observations and the proposed assumptions. The corresponding nonparametric correlation coefficients will be written as $t_{c,1}$, and $t_{c,2}$, which are called the generalized Kendall's tau.

3.2.1.1 Unconditional Generalized Kendall's Tau

Using the notations in Section 3.1 the nonparametric correlation, $t_{c,1} = t_{c,1}(X_1, X_2)$ between the components of $\underline{X}$, is defined as

$$t_{c,1} = \binom{n}{2}^{-1} \Sigma_n' \, U(X_i, X_j) = \binom{n}{2}^{-1} \Sigma_n' U_{ij} \qquad (3.2.1)$$

where

$$U_{ij} = u(X_{1i} - X_{1j}, \delta_{1i} - \delta_{1j}) \; u(X_{2i} - X_{2j}, \delta_{2i} - \delta_{2j}) \qquad (3.2.2)$$

$$= u_{1ij} u_{2ij}$$

with

$$u(a,b) = \begin{cases} 1 \text{ if } a>0, b \geq 0 \text{ or } a=0, b>0 \\ -1 \text{ if } a<0, b \leq 0 \text{ or } a=0, b<0 \\ 0 \text{ otherwise} \end{cases} \qquad (3.2.3)$$

as a function of ordered pair (a,b).

The Σ_n' stands for summation over all possible pairs $(\underline{X}_i, \underline{X}_j)$,

$i<j$; $i,j=1,\ldots,n$.

Considering the functions u_{rij}; $r=1,2$, we could note that $u_{rij} = -u_{rji}$. Hence,

$$\Sigma_n'' \; u_{rij} = 0 \tag{3.2.4}$$

where Σ_n'' stands for the summation over all permutations (i,j) such that $i\neq j, i,j=1,\ldots,n$. Furthermore, it is easy to verify that

$$\max(\Sigma_n' u_{rij}) = \binom{n}{2} \;, \; r=1,2 \tag{3.2.5}$$

If $u_{rij} \neq 0$ for all i, j and r, then

$$\Sigma_n' u_{rij}^2 = n(n-1)/2 \tag{3.2.6}$$

and tau-c,1 can be written as

$$t_{c,1} = \frac{\Sigma_n' u_{1ij} u_{2ij}}{\sqrt{\Sigma_n' u_{1ij}^2 \; \Sigma_n' u_{2ij}^2}} \tag{3.2.7}$$

This shows that tau-c,1 is a type of the general correlation coefficient, introduced by Daniels (1948), and Kendall (1970), provide that $u_{rij} \neq 0$ for all r,i,j.

Following Hoeffding's (1948) notation tau-c,1 can be presented as

$$t_{c,1} = \frac{1}{n(n-1)} \; \Sigma_n'' \, U_{ij} \tag{3.2.8}$$

where Σ_n'' stands for summation over all permutation (i,j), $i\neq j$. This formula shows that tau-c,1 is symmetric in $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$. Thence, tau-c,1 is in fact a U-statistic.

This will be discussed further in Section 3.3 and Section 3.4.

It is well known that the Kendall's tau for uncensored data is presented as a difference between the number of concordant pairs,

C, and discordant pairs, D, that is $t = (C - D)/N$, where N is the number of all possible pairs. This idea of concordance and discordance is extended for censored data by defining concordant and discordant censored pairs.

A pair of observations (X_{1j}, δ_{1i}), (X_{2i}, δ_{2i}) , i=1,2..., is called noncomparable if

$$X_{ri} < X_{rj} \text{ and } \delta_{ri} > \delta_{rj}, \quad i \neq j = 1,2$$

for at least one value of r=1,2. Otherwise the pair is called comparable. Furthermore, a comparable pair can be classified into 3 possible categories, that is concordant, discordant, or tied.

The pair is called concordant if

$$(X_{r1}, \delta_{r1}) > (X_{r2}, \delta_{r2}) \text{ or } (X_{r1}, \delta_{r1}) < (X_{r2}, \delta_{r2})$$

for r=1,2; and the pair is discordant if

$$(X_{11}, \delta_{11}) > (X_{12}, \delta_{12}) \text{ and } (X_{21}, \delta_{21}) < (X_{22}, \delta_{22})$$

or

$$(X_{11}, \delta_{11}) < (X_{12}, \delta_{12}) \text{ and } (X_{21}, \delta_{21}) > (X_{22}, \delta_{22})$$

Here, we should note that (a,b) > (c,d) if and only if $a>c$; $b \geq d$ or $a \geq c$; $b>d$. Finally, the pair is called tied if it is comparable and $(X_{r1}, \delta_{r1}) = (X_{r2}, \delta_{r2})$ for at least one value of r.

Let C, D, T and I be the number of concordant, discordant, tied and incomparable pairs out of $N = n(n-1)/2$ possible pairs, then the unconditional generalized Kendall's tau; $t_{c,1}$; can be written as

$$t_{c,1} = \frac{C - D}{C + D + T + I} = \frac{C - D}{N} \qquad (3.2.9)$$

Considering a random pair of observations let p, q, r and s (=1-p-q-r) be the probability of being concordant, discordant, tied

and incomparable pair respectively; then we have the estimators:

$$\hat{p} = C/n, \quad \hat{q} = D/N$$

$$\hat{r} = T/N, \quad \hat{s} = I/N \tag{3.2.10}$$

Hence

$$t_{c,1} = E(t_{c,1}) = p - q \tag{3.2.11}$$

The variance of $t_{c,1}$ will be computed as follows. Let $F(\underline{X};\underline{\delta}_i) = F_{\underline{\delta}_i}(\underline{X})$, and $F_{r:\underline{\delta}_i}(X_r)$, r=1,2 are its marginal distribution functions for i=1,2. Then, we have at most four distribution functions, associated with the value $\delta = (0,0),(0,1),(1,0)$ and $(1,1)$. Under the assumption that $F(\underline{\delta}_1)$'s are continuous, the Φ_1 function of Hoeffding (1948) can be generalized for the right censored data. Let

$$m_2(\underline{X};\underline{\delta}_1|\underline{\delta}_2) = \begin{cases} 1 - 2F_{1,\underline{\delta}_2} - 2F_{2,\underline{\delta}_2} + 4F_{\underline{\delta}_2} & ; \underline{\delta}_1 = \underline{\delta}_2 \\ 1 - 2F_{1,\underline{\delta}_2} - F_{2,\underline{\delta}_2} + 3F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (0,0),\ \underline{\delta}_2 = (0,1) \\ 1 - F_{1,\underline{\delta}_2} - 2F_{2,\underline{\delta}_2} + 3F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (0,0),\ \underline{\delta}_2 = (1,0) \\ 1 - F_{1,\underline{\delta}_2} - F_{2,\underline{\delta}_2} + 2F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (0,0),\ \underline{\delta}_2 = (1,1) \\ -F_{2,\underline{\delta}_2} + 2F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (0,1),\ \underline{\delta}_2 = (0,0) \\ F_{2,\underline{\delta}_2} = F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (0,1),\ \underline{\delta}_2 = (1,0) \\ 1 - F_{1,\underline{\delta}_2} - 2F_{2,\underline{\delta}_2} + 3F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (0,1),\ \underline{\delta}_2 (1,1) \\ -F_{2,\underline{\delta}_2} + F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (1,0),\ \underline{\delta}_2 = (0,0) \\ F_{1,\underline{\delta}_2} - F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (1,0),\ \underline{\delta}_2 = (0,1) \end{cases}$$

$$\begin{cases} 1 - 2F_{1,\underline{\delta}_2} - F_{2,\underline{\delta}_2} + 3F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (1,0),\ \underline{\delta}_2 = (1,1) \\ F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (1,1),\ \underline{\delta}_2 = (0,0) \\ -F_{1,\underline{\delta}_2} + 2F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (1,1),\ \underline{\delta}_2 = (0,1) \\ -F_{2,\underline{\delta}_2} + 2F_{\underline{\delta}_2} & ; \underline{\delta}_1 = (1,1),\ \underline{\delta}_2 = (1,0) \end{cases} \tag{3.2.12}$$

and

$$\Phi_1(\underline{X};\underline{\delta}_1) = \sum_{\underline{\delta}} m_1(\underline{X};\underline{\delta}_1 | \underline{\delta}_2=\underline{\delta})\ P(\underline{\delta}_2=\underline{\delta}) \tag{3.2.13}$$

Considering the possible values of $\underline{\delta}_k$, we may write

$$P(\underline{\delta}_k=(a,b)) = p_1^2(1-p_1)^{1-a}\ p_2^b(1-p_2)^{1-b} \tag{3.2.14}$$

where

$$P(\delta_{rk}=1) = p_r;\ P(\delta_{rk}=0) = 1 - p_r \tag{3.2.15}$$

for r=1,2.

Thence we may write

$$\begin{aligned} \tau_{c,1} &= E\{\Phi_1(X;\underline{\delta}_1)\} \\ &= \sum_{\underline{\delta}} P(\underline{\delta}_1=\underline{\delta}) \int\int \Phi_1(\underline{X};\underline{\delta})\ dF(\underline{X};\underline{\delta}) \end{aligned} \tag{3.2.16}$$

The variance of $t_{c,1}$ is, by Hoeffding (1948),

$$\sigma^2(t_{c,1}) = \frac{2}{n(n-1)}\{2(n-2)\ \zeta_1(\tau_{c,1}) + \zeta_2(\tau_{c,2})\} \tag{3.2.17}$$

where

$$\zeta_1(\tau_{c,1}) = E\{\Phi_1^2(\underline{X};\underline{\delta}_1)\} - \tau_{c,1}^2 \tag{3.2.18}$$

$$\begin{aligned} \zeta_2(\tau_{c,1}) &= E(U_{12}^2) - \tau_{c,1}^2 \\ &= (p + q) - \tau_{c,1}^2 \end{aligned} \tag{3.2.19}$$

In this case, we are, in fact considering non-random causes for censoring. In other words, we are not taking into consideration the censoring variable. So, another situation, in which the causes of censoring are random needs to be proposed. In this situation, we should consider the minimal functions between the studied variables and the censoring variable. Then, the data having new variables corresponding to the minimal functions are uncensored data. Hence, Hoeffding's results are directly applicable. This will be discussed in Section 3.8.

3.2.1.2 Conditional Generalized Kendall's Tau

Corresponding to the function U_{ij} above, we define a function $V_{ij} = v_{1ij}v_{2ij}$ such that

$$V_{ij} = V(\underline{X}_i(\underline{\delta}_i),\underline{X}_j(\underline{\delta}_j)) = \begin{cases} U_{ij} \text{ if } \delta_{ri} = \delta_{rj} = 0 \\ \quad \text{or } \delta_{ri} \neq \delta_{rj} \\ 0 \text{ otherwise.} \end{cases} \tag{3.2.20}$$

The difference between the functions U_{ij} and V_{ij} is in transforming the censored observations, that is if $\delta_{ri} = \delta_{rj} = 1$, then $v_{rij} = 0$ but u_{rij} may take a non-zero value -1 or +1, if $X_{ri} \neq X_{rj}$.

Thence, another nonparametric correlation, Kendall's tau-c,2 will be defined as

$$t_{c,2} = \{\binom{n}{2} - N^*\}^{-1} \Sigma_n' V_{ij} \tag{3.2.21}$$

where N^* is the total number of all possible pairs $(\underline{X}_i,\underline{X}_j)$ in which $\delta_{ri} = \delta_{rj} = 1$ for at least one value of r=1,2.

Using the general ordering of observations given in Section 3.1,

we have

$$N^* = \binom{n-k}{2} - (\ell-k)(m-\ell). \tag{3.2.22}$$

Thence, if $\ell = k$ or $m = \ell$, we have

$$t_{c,2} = \frac{2}{2k(n-k) + k(k-1)} \Sigma'_n V_{ij} \tag{3.2.23}$$

If C^* and D^* are the number of concordant and discordant pairs out of the N^* observed pairs, then $t_{c,2}$ can be written as

$$t_{c,2} = \frac{(C - C^*) - (D - D^*)}{N - N^*} \tag{3.2.24}$$

As the index $t_{c,1}$; $t_{c,2}$ is also symmetric in $\underline{X}_1,\ldots,\underline{X}_n$: and hence it is a U statistic of Hoeffding. This $t_{c,2}$ will be considered as a conditional generalized Kendall tau. For a given pattern of observations, as proposed by Gehan (1965) we will observe the distribution of $t_{c,2}$ under the permutation of the observations. This will be discussed in Section 3.6.

3.3 Characteristics of the Generalized Kendall Tau

In this section, the nonparametric correlation coefficients $t_{c,1}$ and $t_{c,2}$ will be written as one formula:

$$t_{c,s} = N_s^{-1} \Sigma'_n A_{s,ij} = N_s^{-1} \Sigma'_n a_{s,1ij} a_{s,2ij} \tag{3.3.1}$$

with

$$N_s = \begin{cases} n(n-1)/2 & , s = 1 \\ n(n-1)/2 - N^*, & s = 2 \end{cases} \tag{3.3.2}$$

and

$$a_{s,rij} = \begin{cases} u_{rij} & , s = 1 \\ v_{rij} & , s = 2 \end{cases} \tag{3.3.3}$$

because both correlations have some common characteristics.

3.3.1 General Properties

Since $A_{s,ij} = A_{s,ji}$, then $t_{c,s}$ is '<u>symmetric under interchange</u>' of its variables, so that

$$t_{c,s}(X_1,X_2) = t_{c,s}(X_2,X_1) \tag{3.3.4}$$

And it is '<u>positively invariant</u>,' that is invariant under monotonic increasing transformations. So, if $g_i(\cdot)$, i=1,2 are monotone increasing functions, then

$$t_{c,s}\ (g_1(X_1),\ g_2(X_2)) = t_{c,s}(X_1,X_2) \tag{3.3.5}$$

In particular, for $g_i(x) = a_i + b_i x$, $b_i > 0$, we have linear transformation with positive slope.

It is easy to verify that

$$-1 < t_{c,s} < +1 \tag{3.3.6}$$

In general, for general right-censored data, the limits -1 or +1 can not be reached, even though there are no ties, because $a_{s,rij}$ may take a value zero even $x_{ri} \neq x_{rj}$. This deficiency suggests some modifications of $t_{c,s}$, which will be considered later.

3.3.2 Some Special Cases

If all n-observations are uncensored, then $N_1 = N_2$ and $u_{rij} = v_{rij}$. Consequently, $t_{c,1} = t_{c,2}$, that is

$$t_{c,s} = \binom{n}{2}^{-1} \Sigma'_n u(X_{1i}-X_{1j},0)\ u(X_{2i} - X_{2j},\ 0) \tag{3.3.7}$$

which is the same as that of Kendall (1970).

Another special case is a censored sample, in which the uncensored and censored observations on X_2 have the following orderings.

$$(X_{21},0) < (X_{22},0) \text{ ------ } < (X_{2m},0); \text{ and}$$

$$(X_{2,m+1},1) < \text{ ---------- } < (X_{2n},1) \qquad (3.3.8)$$

where X_{2i}, $i > m$ are the values at which they were censored. In this case, we have

$$t_{c,s} = N_s^{-1}\{ \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} a_{s,1ij} + (2-s) \sum_{i=m+1}^{n-1} \sum_{j=i+1}^{n} a_{s,1ij}$$

$$+ \sum_{i=1}^{m} \sum_{j=m+1}^{n} a_{s,1ij} a_{s,2ij} \}; \; s=1,2. \qquad (3.3.9)$$

since $a_{2,2ij} = v_{2ij} = 0$ for all $i=m+1,\ldots n$; $j=i+1$. If the censoring on X_2 occurs after the m-th order completed observations, such that

$$(X_{21},0)<\text{---}<(X_{2m},0) < (X_{2,m+1},1)<\text{---}<(X_{2n},1) \qquad (3.3.10)$$

then

$$t_{c,s} = N_s^{-1}\{ \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{s,1ij} - (s-1) \sum_{i-m+1}^{n-1} \sum_{j=i+1}^{n} a_{s,1ij} \}$$

(3.3.11)

that is a function of the observations on X_1 only.

In the previous cases, if ties occur on X_2, then the number of terms on the right hand side of each formula will be reduced. For example, if the censoring on X_2 occurs at '<u>one point</u>' after the m-th order completed observations, such that

$$(X_{21},0)<\text{---}<(X_{2m},0) < (X_{2,m+1},1) =\text{---}=(X_{2n},1) \qquad (3.3.12)$$

then

$$t_{c,s} = N_s^{-1} \{ \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} a_{s,1ij} + \sum_{i=1}^{m} \sum_{j=m+1}^{n} a_{s,1ij} \}$$

(3.3.13)

3.4 Alternative Expression of the Generalized Kendall's Tau

Considering the ordering of the n-observations assumed in Section 3.1, we would have uncensored sub-sample $\underline{X}_i$, i=1,...,k; and censored sub-sample $\underline{X}_i$, i=k+1,...,n. Thence, the Kendall's tau-c,x can be presented as

$$t_{c,s} = N_s^{-1} \{ \binom{k}{2} t_{00} + k(n-k) t_{01} + N_s^* t_{11,s} \} \qquad (3.4.1)$$

where

$$N_s^* = \begin{cases} \binom{n-k}{2}, & s=1 \\ \binom{n-k}{2} - N^*, & s=2 \end{cases}$$

$$t_{00} = \binom{k}{2}^{-1} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} A_{s,ij} \qquad (3.4.2)$$

$$t_{01} = (k(n-k))^{-1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} A_{s,ij} \qquad (3.4.3)$$

$$t_{11,s} = N_s^{*-1} \sum_{i=k+1}^{n-1} \sum_{j=i+1}^{n} A_{s,ij} \qquad (3.4.4)$$

This shows that $t_{c,s}$ is a weighted average of t_{00}, t_{01} and $t_{11,s}$, since

$$N_s = \binom{k}{2} + k(n-k) + N_s^*$$

It is clear that t_{00} is the Kendall tau for the uncensored sub-sample. Hence, t_{00} could have a value from -1 to +1. However, the limits -1 and +1 cannot be reached if ties occur as noted also by Quade (1974).

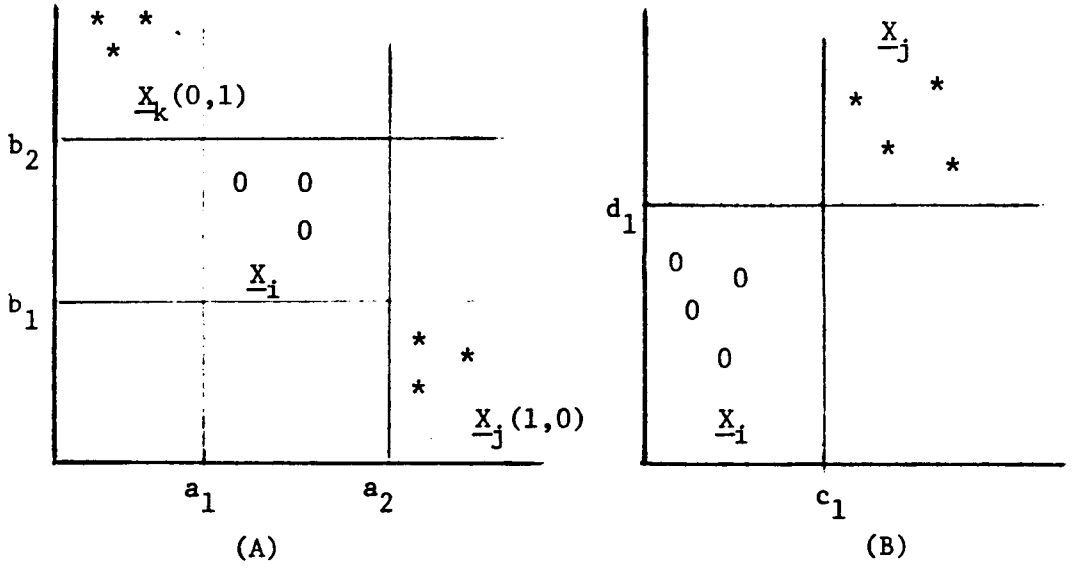
The value of t_{01} depends on

$$A_{s,ij} = a_s(X_{1i}-X_{1j},-\delta_{1j})\ a_s(X_{2i}-X_{2j},-\delta_{2j})$$

$$i = 1,2,\ldots,k;\ j=k+1,\ldots,n. \qquad (3.4.5)$$

in which at least one of δ_{1j}, δ_{2j} is equal to +1, because $\underline{X}_j$'s are censored. If $\delta_{rj} = 1$, then $a_{s,rij} = -1$ or 0 according to whether $X_{ri} = X_{rj}$ or not. If $\delta_{rj} = 0$ for at most one r, then $a_{s,rij} = -1$, 0 or +1 according to whether $X_{ri} < X_{rj}$, $X_{ri} = X_{rj}$ or $X_{ri} > X_{rj}$. Hence, $A_{s,ij}$ may take a value -1, 0, or +1. This shows that t_{01} could have a value from -1 to +1.

Figure 3.1 Illustrative Graphs for (A). $t_{01} = -1$ and (B). $t_{01} = +1$.



In Figure 3.1, the (0)'s indicate uncensored observations and the (*)'s indicate censored observations. In Figure 3.1(A), uncensored observations, $\underline{X}_i$, satisfy $(a_1,b_1) < \underline{X}_i < (a_2,b_2)$; and the censored observations, $\underline{X}_j(1,0)$ and $\underline{X}_k(0,1)$, should satisfy

$$X_{1j} > a_2; X_{2j} < b_1$$

$$X_{1k} < a_1; X_{2k} > b_2$$

This implies $U_{ij} = U_{ik} = -1$ for all i,j and k. Thence $t_{01} = -1$. In this case, we have $X_{1k} < X_{1i} < X_{1j}$ and $X_{2k} > X_{2i} > X_{2j}$.

In Figure 3.1(B), uncensored observations, $\underline{X}_i$, satisfy $\underline{X}_i < (c_1,d_1)$; and the censored observations, $\underline{X}_j(\underline{\delta}_j)$; $\underline{\delta}$, = (0,1), (1,0) or (1,1), should satisfy $\underline{X}_j > (c_1,d_1)$. This implies $U_{ij} = +1$ for all i and j. Hence $t_{01} = +1$. In fact, here we have $\underline{X}_i(0,0) < \underline{X}_j(\underline{\delta}_j)$, with $\underline{\delta}_j \neq (0,0)$ for all i,j. This index, t_{01}, can be considered as a measure of association between the censored and uncensored sub-samples.

Finally, $t_{11,s}$ is the Kendall tau for the censored sub-sample. The most general form of this sub-sample is $\underline{X}_i(0,1)$, i = k+1,...,ℓ; $\underline{X}_i(1,0)$, i=ℓ+1,...,m and $\underline{X}_i(1,1)$, i=m+1,...,n. This index would show the difference between tau-c,1 and tau-c,2. Considering the possible values of $\underline{\delta}$ for each pair of observations $(\underline{X}_i,\underline{X}_j)$, and writing $A_{s,ij}(\underline{\delta}_i,\underline{\delta}_j) = A_{s,ij}$, then we have

$$A_{1,ij}(\underline{\delta}_i,\underline{\delta}_j) = U_{ij}(\underline{\delta}_i,\underline{\delta}_j) \tag{3.4.6}$$

for all $\underline{\delta}_i,\underline{\delta}_j$ = (0,1), (1,0) and (1,1), so there are six possible functions for each (i,j); and

$$A_{2,ij}(\underline{\delta}_i,\underline{\delta}_j) = \begin{cases} V_{ij}(\underline{\delta}_i,\underline{\delta}_j), & \underline{\delta}_i = (0,1),\ \underline{\delta}_j = (1,0) \\ 0 \text{ otherwise} \end{cases} \tag{3.4.7}$$

Note that $V_{ij}(0,1;1,0) = U_{ij}(0,1;1,0)$. Hence $t_{11,s}$ for s = 2 can

be simplified as

$$t_{11,2} = N_2^* \sum_{i=k+1}^{\ell} \sum_{j=\ell+1}^{m} A_{2,ij} \qquad (3.4.8)$$

or

$$t_{11,2} = \frac{1}{(\ell-k)(m-\ell)} \sum_{i=k+1}^{\ell} \sum_{j=\ell+1}^{m} U_{ij}(0,1;1,0) \qquad (3.4.9)$$

Thence, in general, $t_{11,s}$ can be written as

$$t_{11,s} = \frac{2(2-s)}{(n-k)(n-k-1)} \sum_{i=k+1}^{n-1} \sum_{j=i+1}^{n} U_{ij} + \frac{(s-1)}{(\ell-k)(m-\ell)} \sum_{i=k+1}^{\ell} \sum_{j=\ell+1}^{m} U_{ij}(0,1;1,0) \qquad (3.4.10)$$

that is a function of U_{ij}.

Now we will observe the possible values of $A_{s,ij}$, that is U_{ij} or V_{ij}. Since $V_{ij}(0,1;1,0) = U_{ij}(0,1;1,0)$, we may consider only the values of $U_{ij}(\underline{\delta}_1, \underline{\delta}_j)$. The table below, Table 3.1, shows all possible values of U_{ij} by $\underline{\delta}_i$, by $\underline{\delta}_j$.

Table 3.1 Values of the Function U_{ij} by $\underline{\delta}_i$ and $\underline{\delta}_j$

		$\underline{\delta}_j$		
	U_{ij}	(0,1)	(1,0)	(1,1)
$\underline{\delta}_i$	(0,1)	-1,0,1	-1,0	-1,0,1
	(1,0)	-1,0	-1,0,1	-1,0,1
	(1,1)	-1,0,1	-1,0,1	-1,0,1

Based on this table, we conclude that $t_{11,1}$ could reach the limits -1 and +1, but $t_{11,2}$ could not. In fact,

$$-1 \leq t_{11,2} \leq 0. \qquad (3.4.11)$$

Note that in computing $t_{11,2}$, we are not concerned with the relations between $\underline{X}_i$ and $\underline{X}_j$, if both X_{ri} and X_{rj} are censored for at least one r. Figure 3.1(A) shows the case, where $t_{11,2} = -1$. The value $t_{11,2} = 0$ is attainable if we have censored observations $\underline{X}_j(\underline{\delta}_j), \delta_j = (0,1), (1,1)$; or $\underline{\delta}_j = (1,0), (1,1)$. It is clear that $t_{11,2} = 0$ for a B-sample, but in general $t_{11,2} \neq 0$ for an A-sample.

If $t_{11,2} = 0$, then $N^* = \binom{n-k}{2}$. And hence

$$t_{c,2} = \left(\binom{k}{2} + k(n-k)\right)^{-1} \left\{\binom{k}{2} t_{00} + k(n-k) t_{01}\right\} \qquad (3.4.12)$$

that is a weighted average of t_{00} and t_{01}. It is clear that, in this case, $t_{c,2}$ could reach the limits -1 and +1.

3.5 Asymptotic Distribution of Tau-c,1

In Section 3.2, we have shown that the statistic $t_{c,1}$ is the unbiased estimator of $t_{c,1}$, the Kendall's tau-c,1 of the parent population. Here, we would consider its asymptotic distribution function. Theorem 8.1 of Hoeffding (1948) can be modified as

Theorem 3.1

Let $\underline{X}_{i,\underline{\delta}_i}$ i=1,...,n be n independent bivariate random vectors, $\underline{X}_{i,\underline{\delta}_i}$ having the distribution function $F_{\underline{\delta}_i}(\underline{X})$; $\underline{\delta}_i = (0,0)$, $(0,1)$, $(1,0)$ or $(1,1)$. Let $U(\underline{X}_{1,\underline{\delta}_1}, \underline{X}_{2,\underline{\delta}_2}) = U_{12}$ be a symmetric function in its vector argument $\underline{X}_{i,\underline{\delta}_i}$ which does not involve n.

If there exist two positive numbers A and B such that

$$\int\cdots\int | U_{12}^3 | \, dF_{\underline{\delta}_1}(\underline{X}_1) \, dF_{\underline{\delta}_2}(\underline{X}_2) < A. \qquad (3.5.1)$$

for $\underline{\delta}_i=(0,0),(0,1),(1,0),(1,1)$; $i=1,2,\ldots,n$; and

$$\zeta_1(\tau_{c,1}) > B. \qquad (3.5.2)$$

with $\zeta_1(\tau_{c,1})$ given by (3.2.18), then, as $n\to\infty$, the d.f. of $(t_{c,1} - \tau_{c,1})/\sigma(t_{c,1})$ tends to the normal d.f. with mean 0 and variance 1.

Now, we need to show that the statistic $t_{c,1}$ satisfies the conditions of this theorem. From (3.2.2) and (3.2.3), we have

$$|U_{12}^3| \leq 1. \qquad (3.5.3)$$

So there exists a positive number A such that (3.5.1) holds. For (3.5.2), we should consider that

$$\zeta_1(\tau_{c,1}) = \mathrm{VAR}\{\Phi_1(\underline{X};\underline{\delta}_1)\} \qquad (3.5.4)$$

from (3.2.16) and (3.2.18). In general, this is a positive quantity, except if Φ_1 is a constant function of $\underline{X}$ and $\underline{\delta}_1$.

Thence, if Φ_1 is not a constant, as $n\to\infty$, the d.f. of $t_{c,1}$ tends to a normal d.f.

3.6 Some Characteristics of the Conditional Generalized Kendall's Tau

3.6.1 Alternative Expression of Tau-c,2

Following Gehan's (1965) notation, let

m_{ij} = number of uncensored observations on the i-th component of rank j, in the rank ordering of uncensored observations

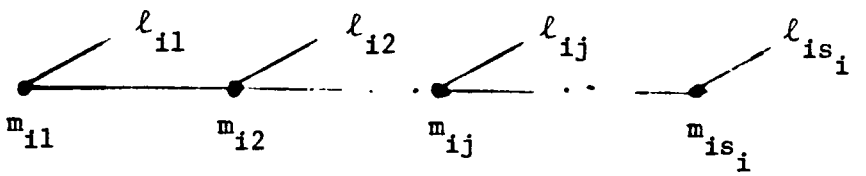
on the i-th component.

ℓ_{ij} = number of right-censored observations on the i-th component with values greater than observations at rank j but less than observations at rank (j+1).

with

$$\Sigma_j(m_{ij} + \ell_{ij}) = n,$$

be the general pattern, P_i, of the i-th component of the bivariate n observations on $\underline{X} = (X_1, X_2)$. This may be illustrated as shown below:



Note that $m_{ij} > 0$ for all i and all $j=1,2,\ldots,s_i$, but ℓ_{ij} can be zero for some j. Hence, the i-th component has at most $2s_i$ distinct values.

We will consider the general pattern for X_1, P_1, as fixed and the pattern for X_2, P_2, as coming from (n!) possible allocations of X_2. Corresponding to each m_{1j}, we have uncensored observations $X_{1jk}, k=1,\ldots,m_1$ on X_1 with $R(X_{1jk}) = j$ for all k. And for each ℓ_{1j}, we have censored observations X^*_{1jk}, $k = 1,\ldots,\ell_{ij}$ on X_1. We define $R(X^*_{1jk}) = j^*$ with j^* strictly larger than j. Using the notations as in Section 3.2, an observation on X_1 will be written as (j,0) if it is uncensored of rank j, and (j,1) if it is censored. Corresponding to (j,0) or (j,1), observations on X_2 may take values $(k_j,0)$ and $(k'_j,1)$, for some $k_j, k'_j = 1,\ldots,s_2$. Hence the n bivariate observations

would have the following general forms

1. $\{(i,0), (k_i,\delta_{2i})\}$,

where $(i,0)$, $i=1,\ldots,s_1$ indicates an uncensored observation on X_1, and (k_i,δ_{2i}) is the associated observation on X_2 with $\delta_{2i} = 0$ or 1 corresponding to whether it is uncensored or censored. Here, we have s_1 distinct ranks (or values) for the observation on X_1, with m_{1i} of rank i. So, k_i may take more than one value of $1,2,\ldots,s_2$. In fact, for each i, k_i has m_{1i} possible values.

2. $\{(j,1), (k_j,\delta_{2j})\}$,

where $(j,1)$ indicates a censored observation on X_1, and (k_j,δ_{2j}) has the same meaning as that of the previous type of observation. For the censored observation $(j,1)$ on X_1; j may take only some distinct value(s) of $1,\ldots,s_1$ corresponding to whether ℓ_{1j}'s is not zero. Hence, we would have at most s_1 distinct values for the censored observations on X_1. For each j with $\ell_{1j} \neq 0, k_j$ has ℓ_{1j} possible values.

Considering the observations on X_1, we have that $(i,0)$ and $(j,1)$ are increasing on i and j respectively. And the v_{1ij} function would satisfy

$$v_{1ij} = v\{(i,\delta_{1i}) - (j,\delta_{1j})\} = \begin{cases} 1 \ (i > j, \ \delta_{1j} = 0) \text{ or} \\ \quad (i = j, \ \delta_{1i} = 1, \delta_{1j} = 0) \\ 0 \text{ otherwise} \end{cases} \tag{3.6.1}$$

The index $t_{c,2}$ can be presented as a function of the second component of $\underline{X}$:

$$t_{c,2} = \frac{2}{k(2n-k-1)} \left[\sum_{i=1}^{s_1-1} \sum_{l=i+1}^{s_1} \sum_{k_i,k_j}' v((k_j,\delta_{2j}) - (k_i,\delta_{2i})) \right.$$

$$\left. + \sum_{i=1}^{s_1-1} \sum_{j,1}^{*} \sum_{k_i,k_j}' v((k_j,\delta_{2j}) - (k_i,\delta_{2i})) \right]$$

where

$\sum_a$ means summation over (all possible) a such that $(a,\delta_{1a}) = (a,0)$, that is uncensored on X_1.

$\sum_a^*$ means summation over all possible a such that $(a,\delta_{1a}) = (a,1)$, that is censored on X_1.

$\sum_{k_a}'$ means summation over all possible k_a for each a.

Note that, if $m_{1a} > 1$, k_a may take either multiple values of 1,2,..., or s_2, or several distinct values, or a combination of both for each a.

If $\ell_{1j} = 0$ for $j=1,2,\ldots,s_1-1$, then the censored observations on X_1 have one value, that is $(s_1,1)$, hence

$$t_{c,2} = \frac{2}{k(2n-k-1)} \left[\sum_{i=1}^{s_1-1} \sum_{j=i+1}^{s_1} \sum_{k_i,k_j}' v((k_j,\delta_{2j}) - (k_i,\delta_{2i})) \right.$$

$$\left. + \sum_{i=1}^{s_1-1} \sum_{k_i,k_{s_1}}' v((k_{s_1},\delta_{2s_1}) - (k_i,\delta_{2i})) \right] \qquad (3.6.3)$$

If $\ell_{1j} = 0$ for all j, that is all observations on X_1 are uncensored, then we have

$$t_{c,2} = \frac{2}{k(2n-k-1)} \sum_{i=1}^{s_1-1} \sum_{j=i+1}^{s_i} \sum_{k_i,k_j}' v((k_j,\delta_{2j}) - (k_i,\delta_{2i})) \qquad (3.6.4)$$

If $m_{1j} = 1$ and $\ell_{1j} = 0$ for all j, then

$$t_{c,2} = \frac{2}{k(2n-k-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} v((k_j,\delta_{2j}) - (k_i,\delta_{2i})) \quad (3.6.5)$$

In this case k_a, for each a, can take only one value out of $1,2,\ldots,s_2$.

Now, considering the observations on X_2, if $\ell_{2j} = 0$ for $j = 1,2,\ldots,s_2-1$, then the censored observations on X_2 are equal to $(s_2,1)$ with a fixed value of s_2. And, in general, an observation on X_2 can be written as

$$\{k_a + (s_2-k_a)\delta_{2a}, \delta_{2a}\}$$

where a is the rank of the corresponding observation on the first component of $\underline{X}$. Associated with a type of censoring on X_1, a formula for $t_{c,2}$ can be obtained from previous formulas by replacing (k_a,δ_{2a}) a=i,j. Of course, the formula obtained can be simplified easily, according to censoring types on X_1.

3.6.2 Conditional Distribution of Tau-c,2

Considering P_1 as a fixed observation pattern on X_1, and assuming the observation pattern on X_2, P_2, comes from (n!) equally likely allocations of X_2, then we have (n!) possible values of the index $t_{c,2}$ for given P_1 and P_2. Hence, under this assumption, the conditional probability of $t_{c,2}$ given P_1 and P_2 is

$$P_r(t_{c,2} = r_i \mid P_1,P_2) = 1/(n!) \quad (3.6.6)$$

for i=1,2,...,(n!). The corresponding values of $t_{c,2} = r_i$, can be computed easily for small sample sizes. For large sample size,

however, the use of a computer should be more practical.

Furthermore, having all values of r_i's we can compute the conditional k-th moment of $t_{c,2}$, $E(t_{c,2}^k \mid P_1,P_2)$, and the conditional variance of $t_{c,2}$, $VAR\,(t_{c,2} \mid P_1,P_2)$, as follows.

$$E(t_{c,2}^k \mid P_1,P_2) = (n!)^{-1} \sum_{i=1}^{n!} r_i^k \tag{3.6.7}$$

$$VAR(t_{c,2} \mid P_1,P_2) = E(t_{c,2}^2 \mid P_1,P_2) - E^2(t_{c,2} \mid P_1,P_2) \tag{3.6.8}$$

For further discussion, we will consider a special case, in which all observations on X_1 are uncensored such that $m_{1j} = 1$ for all $j=1,2,\ldots,n$. In this case, we are in fact assuming that X_1 has a continuous (marginal) distribution function. In this case, the general pattern, P_2, of the observations on X_2 will be written as

$$m_j,\ \ell_j,\ j=1,2,\ldots,k \text{ with } \sum_{j-1}^{k} (\ell_j + m_j) = n$$

So, these observations will be represented by

$$\underset{m_1}{1,\ldots,1};\ \underset{\ell_1}{1^*,\ldots,1^*};\ \ldots;\ \underset{m_j}{j,\ldots,j};\ \underset{\ell_j}{j^*,\ldots,j^*};\ \ldots;$$

$$\underset{m_k}{j,\ldots,k};\ \underset{\ell_k}{k^*,\ldots,k^*}. \tag{3.6.9}$$

where $j,\ldots,j$ are the ranks of the m_j uncensored observations on X_2; and $j^*,\ldots,j^*$ are the ranks or defined values of the ℓ_j censored observations on X_2. This ordering will be considered as the <u>natural ordering</u> for the general right-censored sample. Now,

considering the n! permutations of these ordered observations, let S_i be the set of permutations having i total number of inversions each. In this case, we define that an inversion will occur in a permutation if a j^* lies in front of an integer $j' \leq j$, or if a j lies in front of an integer $j' < j$. Note that $j^*, j=1,\ldots,k$ are not considered as integers. Hence, we would have the sets.

$$S_i,\ i=0,1,\ldots,I \tag{3.6.10}$$

with

$$I = \sum_{j=1}^{k} m_j(L_j + M_j) \tag{3.6.11}$$

where

$$L_j = \sum_{i=j}^{k} \ell_i \text{ and } M_j = \sum_{i=j+1}^{k} m_j \tag{3.6.12}$$

The value of i = I is attainable if the permutation is the inverse of the natural ordering, that is

$$\underbrace{k^*,\ldots,k^*}_{\ell_k};\ \underbrace{k,\ldots,k}_{m_k};\ \ldots;\ \underbrace{1^*,\ldots,1^*}_{\ell_1};\ \underbrace{1,\ldots,1}_{m_1}.$$

Let $n(S_i)$ be the number of elements or permutations in S_i, then we have

$$n(S_i) = n(S_{I-i}),\ i=0,1,\ldots,I \tag{3.6.13}$$

since each permutation in S_i is an inverse ordering of each permutation in S_{I-i}. The method for computing $n(S_i)$ will be proposed using the illustrative examples below.

Let $t_{c,2;i}$ be the conditional Kendall's tau between X_1 and the elements of S_i, then $t_{c,2;i}$ can be easily derived from $t_{c,2;i-1}$ by considering that an inversion contributes a value of -1 to the

numerator of formula 3.6.2 in computing its value. This implies

$$N_2 \, (t_{c,2;i} - t_{c,2;i-1}) = 2 \qquad (3.6.14)$$

Thence,

$$t_{c,2;i} = \frac{I - 2i}{N_2} \, , \; i=0,1,\ldots,I \qquad (3.6.15)$$

where

$N_2 = n(n-1)/2 - L_1(L_1-1)/2$, and L_1 is the number of the censored observations.

Assuming that the $n!$ permutations are equally likely, we obtain

$$P_r(t_{c,2} = t_{c,2;i}) = n(S_i)/(n!) \qquad (3.6.16)$$

Consequently,

$$E(t_{c,2}^{2r+1}) = E(t_{c,2}^{2r+1} \mid P_1,P_2) = 0; \; r=0,1,\ldots \qquad (3.6.17)$$

and

$$\sigma^2(t_{c,2}) = VAR(t_{c,2} \mid P_1,P_2) = (n!)^{-1} \sum_{i=1}^{I} n(S_i) \, t_{c,2;i}^2 \qquad (3.6.18)$$

This symmetric distribution suggests that the distribution of $t_{c,2}$ can be approximated by a normal distribution with mean 0 and variance $\sigma^2(t_{c,2})$, provided the sample size is large.

3.6.3 Illustrative Examples

(1) Let X_1: 1, 2, 3, 4, 5; and

X_2: 1, 2, 3, 4^*, 5^*

be the observations on X_1 and X_2 respectively, then we have

$$S_0 = (1,2,3,4^*,5^*); \; (1,2,3,5^*,4^*)$$

$$S_1 = (1,2,4^*,3,5^*);\ (1,2,5^*,3,4^*)$$
$$(1,3,2,4^*,5^*);\ (1,3,2,5^*,4^*)$$
$$(2,1,3,4^*,5^*);\ (2,1,3,5^*,4^*)$$

Note that in the second element of S_0 the ordering $5^*,4^*$ does not count as an inversion based on the previous definition. The reason for this is the censored observations are considered as not comparable in computing the conditional Kendall tau. Hence, $n(S_0) = 2$. The elements (or permutations) of S_1 are constructed from the elements of S_0 by permuting two adjacent components which could give an inversion. Three elements on the left side are constructed from $(1,2,3,4^*,5^*)$ by permuting $3,4^*$; 2,3 and 1,2 respectively. And the other three are constructed from $(1,2,3,5^*,4^*)$ by permuting $3,5^*$; 2,3 and 1,2, respectively. So, each permutation in S_1 has only one inversion. Here, we are in fact observing all possible adjacent components in each permutation of S_0, which could increase the number of inversion by +1.

Furthermore, using the elements of S_1, we can obtain the permutations having two inversions each; which become the elements of S_2. For example, from $(1,2,4^*,3,5^*) \in S_1$ we obtain the elements of S_2:

$(1,2,4^*,5^*,3)$,
$(1,4^*,2,3,5^*)$, and
$(2,1,4^*,3,5^*)$

from $(2,1,3,4^*,5^*)$, we obtain two permutations $(2,1,4^*,3,5^*)$ and $(2,3,1,4^*,5^*)$. However, only the second can be counted as a new

element of S_2, because the first permutation is the same as one of the previous three permutations.

Hence, in general, the elements of S_{i+1} can be constructed from those of S_i,i=0,...,I-1. Since $n(S_i) = n(S_{I-i})$, we only need to compute S_i,i=0,...,k $\leq$ I/2. As shown in the previous paragraph, for i $\geq$ 1, we should note the possibility of getting the same permutations for S_{i+1} from two elements of S_i.

For this example, finally we obtain the following distribution of $t_{c,2}$.

Table 3.2 The Distribution of $t_{c,2}$ for data in Example 1.

i	$n(S_i)$	$t_{c,2;i}$	Prob.
0	2	1	2/120
1	6	7/9	6/120
2	12	5/9	12/120
3	18	3/9	18/120
4	22	1/9	22/120

The table shows only the values for i=0,1,...,4; since the others are symmetric except for $t_{c,2;i}$, which have negative values for i=5,6,...,9.

Note that the pattern P_2 can be written as 1, 2, 3, 3^*, 3^*.

(2) Let X_1: 1, 2, 3, 4, 5; and

X_2: 1, 2, 3^*, 4, 5^*

be the observations on X_1 and X_2 respectively, then we have

$S_0 = (1,2,3^*,4,5^*); (1,2,4,3^*,5^*); (1,2,4,5^*,3^*)$

Comparing with the set S_0 in the first example, here we have two adjacent components $(4,3^*)$, which is not counted as an inversion, because we cannot say that the uncensored observation, 4, is larger than the censored observation, 3^*. Hence, here we have $n(S_0) = 3$. Using the same steps as in the previous example, we obtain

Table 3.3 The Distribution of $t_{c,2}$ for Data in Example 2

i	$N(S_i)$	$t_{c,2;i}$	Prob.
0	3	8/9	3/120
1	9	6/9	9/120
2	15	4/9	15/120
3	21	2/9	21/120
4	24	0	24/120
5	21	-2/9	21/120
6	15	-4/9	15/120
7	9	-6/9	9/120
8	3	-8/9	3/120

Note that, here we have $I = 9$, and the values of $t_{c,2}$ range from -8/9 to +8/9. This situation suggests a modification for the generalized Kendall's tau, which will be considered later.

The pattern for X_2 also can be written as $(1,2,2^*,3,3^*)$, because the uncensored observations have only three ranks.

Previous examples show all the cases which can be found in more general patterns. Considering the most general pattern, P_2, given in Section 3.6, it seems impossible to find a general formula

for $n(S_i)$. For an illustration, let us consider S_0.

The natural ordering of censored observations is a permutation having zero inversions, so it belongs to S_0. However, there are many other permutations having zero inversions, because the $(\ell_{j-1} + m_j)$ observations (= components):

$$|(j-1)^*,\ldots,(j-1)^*;\ j,\ldots,j\ |$$

$$\ell_{j-1} \qquad\qquad m_j$$

$$j=1,\ldots,k+1;\ \ell_0 = m_{k+1} = 0$$

can be permuted as many as $(\ell_{j-1} + m_j)!$ ways without affecting the number of inversions. Note that the orderings $(j-1)^*,j$ and $j,(j-1)^*$ are not considered as inversions. So far, we may conclude that

$$n(S_0) \geq \prod_{j=1}^{k+1} (\ell_{j-1} + m_j)! = K$$

with

$\ell_0 = m_{k+1} = 0$ are defined.

In Example 1, we have $m_1 = m_2 = m_3 = 1$ and $\ell_1 = \ell_2 = 0$; $\ell_3 = 2$; since the pattern P_2 is 1, 2, 3, 3^*, 3^*, hence $K = 2$. Here, we obtain $n(S_0) = K = 2$. In Example 2, the pattern is 1, 2, 2^*, 3, 3^*, so $m_1 = m_2 = m_3 = 1$; $\ell_1 = 0$, $\ell_2 = \ell_3 = 1$. Hence we have again $K = 2$, but $n(S_0) = 3$.

Now, considering the group of these K permutations, we have a sub-group G, in which each permutation has ordered components

$$|(j-1)^*,\ldots,(j-1)^*;\ j^*,\ldots,j^*|$$

$$c_{j-i} \qquad\qquad c_j$$

for $0 \leq c_j \leq l_j$; $j=2,\ldots,k$; with $c_{j-1} \neq 0$ and $c_j \neq 0$ for at least one j. These $(c_{j-1} + c_j)$ components also can be permuted without affecting the number of inversions, because all censored observations are considered as uncomparable. So, there are $(\ell_{j-1}\ell_j)$ combinations of $(c_{j-1} + c_j)$, and for each combination we would have $(c_{j-1} + c_j)!$ permutations. However, $(c_{j-1}!)(c_j!)$ of these permutations are counted in the group of K permutations. Thence, for each permutation having combinations $c_{j-1} + c_j$ with $c_{j-1} \neq 0$ and $c_j \neq 0$, there are

$$C_{j-1,j} = [(c_{j-1} + c_j)! - (c_{j-1}!)(c_j!)]$$

additional permutations from the $c_{j-1} + c_j$ which should be counted in finding the elements of S_0. In Example 2, we have $c_2 = c_3 = 1$, since the pattern P_2 is 1, 2, 2^*, 3, 3^*; hence $C_{2,3} = 1$. Thus, we have $n(S_0) = 2 + 1 = 3$.

If the censoring on X_2 occurs after the k-th ordered completed observations, then we will not have any $C_{j-1,j}$, as shown by Example 1. In this case

$$n(S_0) = \prod_{j=1}^{k+1} (\ell_{j-1} + m_j)! = K$$

$$\ell_0 = m_{k+1} = 0$$

Otherwise $n(S_0)$ is strictly larger than K.

The contribution of $C_{j-1,j}$ in calculating $n(S_0)$ depends on $\ell_j, m_j, j=1,\ldots,k$. Furthermore, $n(S_i)$, $i \neq 0$ would be affected also by $C_{j',j}$ for $j' \neq j$, which will be so complex. Hence, in computing the value of $n(S_i)$, we may consider the following method:

(i) One may find a formula for $n(S_i)$ for a particular pattern,

that is the pattern of his data, or he may calculate $n(S_i)$ starting from i=0 using the steps shown above.

(ii) One may consider all permutations and compute the n! values of $t_{c,2}$ for the bivariate sample of size-n; then the groups are formed with respect to the same values of $t_{c,2}$.

3.7 An Index Alpha as a Modification of the Generalized Kendall's Tau

It has been noted, in previous sections, that the generalized Kendall tau, t_c, in general could not reach the limits -1 and +1. This situation is shown by Example 2. It is known that tied observations also cause this deficiency. So, in this section, we will propose an index of correlation Alpha, α_s, as a modification of the generalized Kendall's tau, $t_{c,s}$.

First, we will consider the index α_1 as a modified 'unconditional' generalized Kendall's tau, $t_{c,1}$. Considering the functions U_{ij}, as defined in Section 3.2, let

$$Z_u = \Sigma'_n \, (U_{ij} = 0) \tag{3.7.1}$$

be the total number of pairs $(\underline{X}_i, \underline{X}_j)$, $i < j$; $i, j=1,\ldots,n$ such that $U_{ij} = 0$. Then the index $\alpha_1(X_1, X_2)$ is given by

$$\alpha_1 = \frac{\binom{n}{2}}{\binom{n}{2} - Z_u} \, t_{c,1} \tag{3.7.2}$$

or

$$\alpha_1 = \{\binom{n}{2} - Z_u\}^{-1} \, \Sigma'_n \, U_{ij} \tag{3.7.3}$$

If C, D, T and I are the number of concordant, discordant, tied and incomparable observed pairs, respectively, then

$$\alpha_1 = \frac{C - D}{N - T - I} = \frac{C - D}{C + D} \qquad (3.7.4)$$

This shows that $-1 \leq \alpha_1 \leq 1$. And it has the same formula as the Goodman-Kurskal (1965) index gamma, for uncensored data. So, this index can be considered as the generalized G-K index.

Similarly, we define an index α_2 as a modification of $t_{c,2}$ by

$$\alpha_2 = \frac{N - N^*}{N - Z_v} \; t_{c,2} \qquad (3.7.5)$$

or

$$\alpha_2 = \{N - Z_v\}^{-1} \; \Sigma_n' V_{ij} \qquad (3.7.6)$$

where

$$Z_v = \Sigma_n'(V_{ij} = 0) \qquad (3.7.7)$$

is the total number of pairs $(\underline{X}_i, \underline{X}_j)$ $i < j$, $i, j = 1, \ldots, n$ such that $V_{ij} = 0$.

If C^* and D^* are the numbers of concordant and discordant pairs out of the N^* pairs, as given in Section 2.1.2, then α_2 can be represented as

$$\alpha_2 = \frac{(C - C^*) - (D - D^*)}{(C - C^*) + (D - D^*)} \qquad (3.7.8)$$

It is clear that the index α_2 can take a value from -1 to +1. This is considered as the conditional index alpha. If all observations are uncensored, then $\alpha_2 = (C - D)/(C + D)$, that is the G-K index gamma.

3.8 Test of Independence Using $t_{c,1}$-Statistic

3.8.1 For Uncensored Data

Let $\underline{X}_i = (X_{1i}, X_{2i})$ be the true observation on the i-th subject or individual which may be censored by a variable $\underline{Y}_i = (Y_{1i}, Y_{2i})$. So, we observe

$$Z_{ri} = \text{Min}(X_{ri}, Y_{ri}) \tag{3.8.1}$$

for r=1,2 along with indicator variables

$$\delta_{ri} = \begin{cases} 0 \text{ of } Z_{ri} = X_{ri} \\ 1 \text{ if } Z_{ri} = Y_{ri} < X_{ri} \end{cases} \tag{3.8.2}$$

Associated with known d.f.'s of $\underline{X}$ and $\underline{Y}$, we may consider $\underline{Z}_i = (Z_{1i}, Z_{2i})$ as uncensored observations for all i=1,2,...,n. Thence the generalized Kendall's tau, $t_{c,i}$, between Z_1 and Z_2 is in fact the usual Kendall's tau-a:

$$t_{c,1}(Z_1, Z_2) = t_a(Z_1, Z_2) \tag{3.8.3}$$

If $F(\cdot)$, $G(\cdot)$ and $H(\cdot)$ are the d.f.'s $\underline{X}$, $\underline{Y}$ and $\underline{Z}$, respectively, then

$$\begin{aligned} P(Z_1 \geq u, Z_2 \geq v) &= P(\text{Min}(X_1, Y_1) \geq u, \text{Min}(X_2, Y_2) \geq v) \\ &= P(X_1 \geq u, Y_1 \geq u, X_2 \geq v, Y_2 \geq v) \end{aligned} \tag{3.8.4}$$

Assuming that $\underline{X}$ and $\underline{Y}$ are independent random variables, we obtain

$$\Delta H(u,v) = \Delta F(u,v) \cdot \Delta G(u,v) \tag{3.8.5}$$

where

$$\Delta W(u,v) = 1 - W(u,\infty) - W(\infty,v) + W(u,v)$$

which will be written as

$$\Delta W = 1 - W_1 - W_2 + W \tag{3.8.6}$$

If $W = W_1 W_2$ then $\Delta W = (1-W_1)(1-W_2)$. So, if $\underline{Y}$ has independent components then (3.8.5) becomes

$$\Delta H = \Delta F \cdot (1 - G_1)(1 - G_2) \tag{3.8.7}$$

In this case, independence of X_1 and X_2 would clearly imply the independence of Z_1 and Z_2. So, under the null hypothesis, H_0: $t_{c,1}(X_1, X_2) = 0$, Z_1 and Z_2 are independent. Thence, using Hoeffding's results, if $\underline{Z}$ has continuous distribution, we have

$$E_{H_0} t_a(Z_1, Z_2) = \tau_a = 0 \tag{3.8.8}$$

$$\sigma^2_{H_0} t_a(Z_1, Z_2) = \frac{2(2n + 5)}{9n(n - 1)} \tag{3.8.9}$$

For large n, the distribution function of $\sqrt{n}(t_a - \tau_a)$ tends to the normal form, by theorem 7.1 of Hoeffding (1948).

If H_0 is true, the critical region of size ε of the t_a-test may be defined by $|t_a| \geq c_n$ where c_n is the smallest number satisfying.

$$P(|t_a| > c_n \mid H_0) \leq \varepsilon \tag{3.8.10}$$

We may write $c_n = C'_n \sigma(t_a) = 2b_n/3\sqrt{n}$, as given by Hoeffding. And, the power function of the test is

$$P_n(K_n) = P\{|t_a| > 2b_n/3\sqrt{n} \mid K_n\}$$

Since $\sigma(t_a) \longrightarrow 0$, as $n \longrightarrow \infty$, $P_n(K_n) \longrightarrow 1$ for any alternative hypothesis K_n with $\tau_a \neq 0$.

Now we shall study the distribution of $t_a(\underline{Z})$ assuming a sequence of alternative hypotheses.

$$K_n^*: \quad \Omega^*(H_1,H_2) = 1 + n^{-\frac{1}{2}} \lambda^*(H_1,H_2) \tag{3.8.12}$$

where

$$H_1 = H(Z_1,\infty);\ H_2 = H(\infty,Z_2);\ \Omega^*(H_1,H_2) = H(\underline{Z})/(H_1H_2) \tag{3.8.13}$$

and λ is some function of H_1, H_2, the marginal distributions of $\underline{Z}$; with $\lambda \neq 0$ for n=1,2,... . This Ω^* is considered as a dependence function of $\underline{Z}$, introduced by Sibuya (1960).

Writing $t_a(\underline{Z})$ as

$$t_a = \binom{n}{2}^{-1} \Sigma_n' U_{ij} \tag{3.8.14}$$

we obtain

$$E_{K_n^*}(t_a) = \binom{n}{2}^{-1} \Sigma_n' E_{K_n^*}(U_{12}) = E_{K_n^*}(U_{12}) \tag{3.8.15}$$

where

$$E_{K_n^*}(U_{12}) = \int\int\int\int U_{12}\,dA_1\,dA_2 \tag{3.8.16}$$

with

$$A_j = H(\underline{Z}_j) = H_{1j}H_{2j}\,\Omega^*(H_{1j},H_{2j}) \tag{3.8.17}$$

From (3.8.8), (3.8.12) and (3.8.16), we would have

$$E_{K_n^*}(t_a) = n^{-\frac{1}{2}}\beta^* + n^{-1}\gamma + E_{H_0}(t_a) \approx n^{-\frac{1}{2}}\beta^* \tag{3.8.18}$$

for n large, with

$$\begin{aligned}\beta^* &= 2\int\int\int\int U_{12}\,d(H_{11}H_{21})\,d(H_{12}H_{22}\lambda^*(H_{12},H_{22})) \\ &= 8\int\int H_1H_2\lambda^*(H_1,H_2)\,d(H_1H_2),\end{aligned} \tag{3.8.19}$$

since

$$\gamma = \int\int\int\int U_{12}\,d(H_{11}H_{21}\lambda^*(H_{11},H_{21}))\,d(H_{12},H_{22}\lambda^*(H_{12},H_{22})) \tag{3.8.21}$$

If $\lambda^*(H_1,H_2) = f(H_1)g(H_2)$ then

$$\beta^* = 8\int_0^1 xf(x)dx\int_0^1 yg(y)dy \qquad (3.8.21)$$

Here, we should note that

$$U(\underline{Z}_1,\underline{Z}_2) = u_1(H_{11} - H_{12})\, u_2(H_{21} - H_{22}) \qquad (3.8.22)$$

under the assumption that $\underline{Z}_j$ has continuous density functions, with marginal d.f.'s H_{1j} and H_{2j}. And hence

$$U_{12} = \begin{cases} +1 \ , \ H_{i1} > H_{i2} \text{ or } J_{i1} > H_{i2} \text{ for } i=1,2 \\ -1 \ , \ H_{11} > H_{12}, H_{21} < H_{22} \text{ or } H_{11} < H_{12}, H_{21} > H_2 \end{cases} \qquad (3.8.23)$$

The variance of t_a under K_n^* is

$$\sigma^2_{K_n^*}(t_a) = \binom{n}{2}^{-1} \int\int\int\int U_{12}^2 dA_i dA_2 + \binom{n}{2}^{-2} \sum'_{(i,j)\neq(r,s)} \sum' \int\ldots\int U_{ij}U_{rs}dA_i dA_j dA_r dA_s - E^2_{K_n^*}(t_a) \qquad (3.8.24)$$

where the multiple integrals under the double summations can be 6 or 8 integrals according to whether ($i = r$, $j \neq s$), ($i \neq r$, $j = s$) or ($i \neq r$, $j \neq s$). So, we obtain

$$\sigma^2_{K_n^*}(t_a) = \binom{n}{2}^{-1}\{1 + 2(n-2)\gamma_0 + (n-2)(n-3)E^2_{K_n^*}(t_a)\} - E^2_{K_n^*}(t_a) \qquad (3.8.25)$$

where

$$\gamma_0 = \int\ldots\int U_{12}U_{13}dA_1 dA_2 dA_3 = \int\int(1 - 2H_1 - 2H_2 + 4H)^2\, dH \qquad (3.8.26)$$

If $\lambda(H_1,H_2) = f(H_1)\, g(H_2)$ then we obtain

$$\gamma_0 = 1/9 + n^{-\frac{1}{2}}\gamma_1 + n^{-1}\gamma_2 \qquad (3.8.27)$$

where

$$\gamma_1 = 8\int_0^1 (2x - 1)x\, f(x)\, dx \int_0^1 (2y - 1)y\, g(y)\, dy$$

$$+ 4\int_0^1 x\, f(x)\, dx \int_0^1 yg(y)dy$$

$$\gamma_2 = 2y\int_0^1 x^2 f^2(x)dx\int_0^1 y^2 g^2(y)dy$$

And hence

$$\sigma^2_{K_n^*}(t_a) = \sigma^2_{H_0}(t_a) + R_n \qquad (3.8.28)$$

with

$$R_n = \binom{n}{2}^{-1}(n - 2)(2n^{-\frac{1}{2}}\gamma_1 + 2n^{-1}\gamma_2 + (n - 3)n^{-1}\beta^{*2}) - n^{-1}\beta^*$$

So, for large n, $\sigma^2_{K_n^*}(t_a) \simeq \sigma^2_{H_0}(t_a) = 2(2n + 5)/9n(n-1)$

Thence, as $n \longrightarrow \infty$, $(t_a - n^{-\frac{1}{2}}\beta^*)/\sigma_{H_0}(t_a)$ has a normal distribution having zero mean and unit variance.

Finally, we would study the dependence of X_1 and X_2 associated the dependence of Z_1 and Z_2. From (3.8.4-7), we obtain

$$H_i = F_i + G_i(1 - F_i);\ i=1,2$$

$$H - H_1 \cdot H_2 = \{F - F_1 \cdot F_2\}\{1 - G_1\}\{1 - G_2\} \qquad (3.8.29)$$

where $W = W(x,y)$, $W_1 = W(x,\infty)$ and $W_2 = W(\infty,y)$. This implies, since $o \leq (1 - W_i) \leq 1$.

Theorem 3.8.1

The marginal d.f. of $\underline{Z}$, H_i, is larger than F_i, the marginal d.f. of $\underline{X}$. That is

$$H_i \geq F_i \tag{3.8.31}$$

and

$$| H - H_1H_2 | \leq | F - F_1F_2 | \tag{3.8.32}$$

Remark 3.8.1

Associated with (3.8.12), let

$$\Omega(F_1,F_2) = 1 + n^{-\frac{1}{2}} \lambda(F_1,F_2) \tag{3.8.33}$$

with

$$\Omega(F_1,F_2) = F/(F_1F_2)$$

be the dependent function of $\underline{X}$, then

$$|H_1H_2 \lambda^*(H_1,H_2)| \leq |F_1 F_2 \lambda(F_1,F_2)| \tag{3.8.34}$$

and

$$|\lambda^*(H_1,H_2)| \leq |\lambda(F_1,F_2)| \tag{3.8.35}$$

Remark 3.8.2

Considering n uncensored observations on the $\underline{X}$ variable, under the alternative hypothesis

$$K_n: \Omega(F_1,F_2) = 1 + n^{-\frac{1}{2}} \lambda(F_1,F_2) \tag{3.8.36}$$

we would have, for large n,

$$E_{K_n}(t_a) = n^{-\frac{1}{2}}\beta \tag{3.8.37}$$

with

$$\beta = \delta \iint F_1F_2\lambda(F_1,F_2)dF_1F_2 \tag{3.8.39}$$

Then

$$L = (|\beta| - |\beta|^*)/|\beta| \tag{3.8.40}$$

can be considered as the loss of the power of the test, because of censoring.

Definition 3.8.1

Two univariate variables U and V are called positively (or negatively) quadrant dependent if, and only if,

$$F(u,v) - F(u,\infty)\, F(\infty,v) \geq 0 \ (\text{or} \leq 0) \qquad (3.8.41)$$

where $F(u,v)$ is the joint d.f. of (u,v).

Theorem 3.8.2

X_1 and X_2 are positively (or negatively quadrant) dependent if, and only if, Z_1 and Z_2 are.

Theorem 3.8.3

If X_1 and X_2 are either positively or negatively quadrant dependent, then

$$|\beta^*| \leq |\beta| \qquad (3.8.42)$$

provided any one of the following conditions holds:

(i)

$$\left[\frac{1 - F_i}{f_i} + \frac{1 - G_i}{g_i}\right] g_i \leq 1 \ , \ i=1,2 \qquad (3.8.43)$$

(ii)

If X_i and Y_i are non-negative variables, let

$$\lambda_{x,i}(t) = f_i(t)/\{1 - F_i(t)\}$$

$$\lambda_{y,i}(t) = g_i(t)/\{1 - G_i(t)\} \qquad (3.8.44)$$

be the hazard functions corresponding to X_i and Y_i respectively.

Then the conditions would be

$$\lambda_{x,i} + \lambda_{y,i} \leq \lambda_{x,i} \cdot \lambda_{y,i} \tag{3.8.45}$$

for i=1.2.

(iii)

If X_i and Y_i have standard normal distribution functions for i=1,2.

Proof: (Theorem 3.8.3)

Considering formulae (3.8.19) and (3.8.39), we need to show that

$$|H_1H_2\ \lambda^*(H_1,H_2)\ h_1h_2| \leq |F_1F_2\ \lambda(F_1,F_2)\ f_1f_2| \tag{3.8.46}$$

It is easy to verify that

$$H_1H_2\ \lambda^*(H_1,H_2)\ h_1h_2 = F_1F_2\ \lambda(F_1,F_2)\ f_1f_2 \times M \tag{3.8.47}$$

with

$$M = \left[\frac{1-F_1}{f_1} + \frac{1-G_1}{g_1}\right]\left[\frac{1-F_2}{f_2} + \frac{1-G_2}{g_2}\right]\{1 - G_1\}\{1-G_2\}g_1g_2 \tag{3.8.48}$$

Then, we should show that

$$0 < M \leq 1 \tag{3.8.49}$$

Conditions (i) and (ii) clearly give this inequality.

For (iii), M can be written as

$$M = \left[\frac{1-F_1}{f_1} + \frac{1-G_1}{g_1}\right]\left[\frac{1-F_2}{f_2} + \frac{1-G_2}{g_2}\right]\left[\frac{1-G_1}{g_1}\right]\left[\frac{1-G_2}{g_2}\right]g_1^2g_2^2 \tag{3.8.50}$$

Since F_i, G_i are standard normal, then $(1 - F_i)/f_i$ and $(1 - G_i)/g_i$, i=1,2 are the Miller Ratios (see Johnson and Kotz (1970)), which are non-negative. Using their upper bounds and the normal density functions

for g_i we obtain (3.8.49).

Example 3.8.1

Let $\underline{X}$ be normal having d.f.

$$f(u,v;\rho) = \frac{1}{2\pi(1-\rho^2)} \exp\left[\frac{1}{2(1-\rho^2)} (u^2 + v^2 - 2\rho uv)\right] \quad (3.8.51)$$

with

$$F(u,v;\rho) = \int_{-\infty}^{u} \int_{-\infty}^{v} f(s,t;\rho)\, ds\, dt$$

we have, see Sibuya (1960)

$$\frac{df(u,v;\rho)}{d\rho} = f(u,v;\rho) \quad (3.8.52)$$

We may consider $f(u,v;\rho)$ as a function of ρ, then using a Taylor series expansion we have

$$f(u,v;\rho) = \sum_{k=0}^{\infty} \frac{\rho^k}{k!} \left[\frac{d^k f(u,v;\rho)}{d\rho^k}\right] \rho = 0 \quad (3.8.53)$$

For large n, with $\rho = n^{-\frac{1}{2}}r$, this can be approximated by

$$f(u,v;n^{-\frac{1}{2}}r) \simeq f_1(u)f_2(v)\{1 + n^{-\frac{1}{2}}uvr\} \quad (3.8.54)$$

where f_i, i=1,2 are the marginal density functions, and hence

$$\frac{dF(u,v;n^{-\frac{1}{2}}r)}{dr} = n^{-\frac{1}{2}} f_1(u)f_2(v) \quad (3.8.55)$$

Now, we would find the value of β under the alternative hypothesis K_n, which can be written as

$$K_n : \rho = n^{-\frac{1}{2}}r \quad (3.8.56)$$

From (3.8.33), (3.8.39) and (3.8.55), we obtain

$$\frac{d\beta}{dr} = 8 \iint f_1^2(u)f_2^2(v)\, du\, dv \quad (3.8.57)$$

Thence, we have a differential equation

$$\frac{d\rho}{dr} = 2/\pi \tag{3.8.58}$$

with boundary condition $\beta=0$, for $\rho=0$. Thus

$$\beta = 2r/\pi \tag{3.8.59}$$

This value also can be obtained based on Greiner's relation

$$\tau_a = E(t_a) = \frac{2}{\pi} \sin^{-1}\rho \tag{3.8.60}$$

(see Kendall (1949)), for $\rho = n^{-\frac{1}{2}}r$, by taking the first term of its Taylor series expansion.

Example 3.8.2

If $\underline{X}$ has d.f. (3.8.51), then (3.8.29) implies

$$\frac{dH(u,v;\rho)}{d\rho} = \{1 - G_1(u)\}\{1 - G_2(v)\}f(u,v;\rho) \tag{3.8.61}$$

Hence

$$n^{-\frac{1}{2}} \frac{d\beta}{d\rho}^* = 8\iint\{1 - G_1(u)\}\{1 - G_2(v)\}f(u,v;\rho)\, dA \tag{3.8.62}$$

with

$$\begin{aligned} A &= \{1 - H_1(u)\}\{1 - H_2(v)\} \\ &= \{1 - F_1(u)\}\{1 - G_1(u)\}\{1 - F_2(v)\}\{1 - G_2(v)\} \end{aligned} \tag{3.8.63}$$

If G_i's are standard normal d.f.'s, then

$$G_i(x) = F_i(x),\ i=1,2 \tag{3.8.64}$$

Hence

$$n^{-\frac{1}{2}} \frac{d\beta}{d\rho}^* = \frac{32}{9} \iint f(u,v;\rho)\, d\{1 - F_1(u)\}^3\{1 - F_2(v)\}^3 \tag{3.8.65}$$

Under K_n in (3.8.56), (3.8.65) reduces to

$$n^{-\frac{1}{2}} \frac{d\beta^*}{d\rho} = 32n^{-\frac{1}{2}}\left[\int\{1 - F_1(u)\}^2 f_1^2(u)\, du\right]^2 \tag{3.8.66}$$

Hence

$$\beta^* = 32r\left[\int\{1 - F_1(u)\}^2 f_1^2(u)\, du\right]^2 \tag{3.8.67}$$

To compute the value of β^*, we should consider using a Taylor series expansion, that is

$$1 - F_1(u) = \tfrac{1}{2} - \frac{1}{\sqrt{2\pi}}\left(u - \frac{u^3}{u!} + \frac{3u^5}{5!} - \ldots\right) \tag{3.8.68}$$

and the recursive formula

$$\int_{-\infty}^{\infty} u^k e^{-u^2} du = \frac{k-1}{2} \cdot \frac{k-3}{2} \cdot \ldots \tfrac{1}{2} \cdot \sqrt{\pi} \tag{3.8.69}$$

for k is an even integer, and it is zero for k odd. Thence, we have

$$\beta^* = 32r\left\{\frac{1}{8\sqrt{\pi}} + \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \mathrm{Pol}(u^2)\, 3^{-u^2} du\right\}^2 \tag{3.8.70}$$

where

$$\mathrm{Pol}(u^2) = \left(u - \frac{u^3}{3!} + \frac{3u^5}{5!} - \ldots\right)^2 = u^2 - u^4/3 + 7u^6/90 - \ldots \tag{3.8.71}$$

Based on the previous example, we obtain

Theorem 3.8.4

If $\underline{X}$ has a bivariate normal distribution, then under the alternative hypothesis (3.8.56)

$$\beta^* = 8r\, I_1 * I_2 \tag{3.8.72}$$

with

$$I_i = \int\{1 - G_i\} f_i\, d\{1 - F_i\}\{1 - G_i\},\ i=1,2$$

Proof

.By substituting (3.8.55) and (3.8.29) in (3.8.19), we easily obtain (3.8.72).

3.8.2 For Censored Data

In this case, we should consider the observations

$$\underline{Z}_i(\underline{\delta}_i) = (Z_{1i}, \delta_{1i}; Z_{2i}, \delta_{2i})$$

with $\delta_{ri} = 0$ or 1 according to whether the Z_{ri} is uncensored or censored. For i=1,2,...,n, let us define

$$p_{ij} = P\{U_{ij} = +1\} \tag{3.8.73}$$

$$q_{ij} = P\{U_{ij} = -1\} \tag{3.8.74}$$

where

$$U_{ij} = u_{1ij}u_{2ij}$$

with

$$u_{rij} = u(Z_{ri} - Z_{rj},\ \delta_{ri} - \delta_{rj})$$

Since, there are in general four types of observations, we may have ten different values of p_{ij}'s or q_{ij}'s. And let

$$\binom{n}{2}p = \Sigma'_n p_{ij} \tag{3.8.75}$$

$$\binom{n}{2}q = \Sigma'_n q_{ij} \tag{3.8.76}$$

For illustration, we consider a particular case, that is a data set having k uncensored observations and (n-k) censored observations only on the second component. So, we have $\underline{Z}_i(0,0)$; i=1, ...,k; and $\underline{Z}_i(0,1)$, i=k-1,...,n. In this case, we have only three distinct values of p_{ij}'s (or q_{ij}'s). Let

$$p_{00} = P\{U_{12} = +1, \text{ the two observations are uncensored}\}$$

$$p_{11} = P\{U_{12} = +1, \text{ the two observations are censored}\}$$

$$p_{01} = P\{U_{12} = +1, \text{ only one observation is censored}\} \tag{3.8.77}$$

Similarly, q_{00}, q_{11} and q_{01} are the corresponding probabilities of $U_{12} = -1$. Having or assuming the d.f. of $\underline{Z}$, as in the previous subsection, we have

$$p_{kk} = \int_0^1 \int_0^1 \left(\int_0^{H_{11}} \int_0^{H_{21}} + \int_{H_{11}}^1 \int_{H_{21}}^1 \right) dA_2 \, dA_1, \quad k=0,1$$

$$p_{01} = 2 \int_0^1 \int_0^1 \int_{H_{11}}^1 \int_{H_{21}}^1 dA_2 \, dA_1 \tag{3.8.78}$$

and

$$q_{kk} = \int_0^1 \int_0^1 \left(\int_0^{H_{11}} \int_{H_{21}}^1 + \int_{H_{11}}^1 \int_0^{H_{21}} \right) dA_2 \, dA_1, \quad k=0,1$$

$$q_{01} = 2 \int_0^1 \int_0^1 \int_0^{H_{11}} \int_{H_{21}}^1 dA_2 \, dA_1 \tag{3.8.79}$$

Note the reduction of the integration regions, such as p_{01} and q_{01}, whenever we have two different types of observations. In this case, we obtain $p_{00} = p_{11}$, and $q_{00} = q_{11}$.

Considering the statistics t_{00}, t_{11} and t_{01} in Section 3.4, we in fact have

$$E(t_{00}) = p_{00} - q_{00}$$

$$E(t_{11}) = p_{11} - q_{11}$$

$$E(t_{01}) = p_{01} - q_{01} \tag{3.8.80}$$

and then

$$E(t_{c,1}) = \binom{n}{2}^{-1} \{\binom{k}{2}(p_{00} - q_{00}) + k(n-k)(p_{01} - q_{01}) + \binom{n-k}{2}(p_{11} - q_{11})\} \tag{3.8.81}$$

$$E(t_{c,1}) = p - q \tag{3.8.82}$$

If $A = H(\underline{Z}) = H_1 * H_2$ or $\Omega(H_1, H_2) = 1$, then we obtain

$$p_{kk} = q_{kk} = \tfrac{1}{2}\ , \ k=0,1$$
$$p_{01} = q_{01} = 2(\tfrac{1}{2}) = \tfrac{1}{2} \tag{3.8.83}$$

And then

$$p = q = \tfrac{1}{2}$$

This shows

$$E_{H_0}(t_{c,1}) = 0 \tag{3.8.84}$$

and the variance of $t_{c,1}$ can be written as

$$\sigma^2(t_{c,1}) = \binom{n}{2}^{-2} \sum_{i,j}{}' \sum_{r,s}{}' \int\cdots\int U_{ij}U_{rs}\, dA_i\, dA_j\, dA_r\, dA_s \tag{3.8.85}$$

Under H_0, we have

$$dA_k = dH_{1k}\, dH_{2k} \tag{3.8.86}$$

Using similar steps to those in the previous subsection and taking all possible pairs of observations, we obtain

$$\sigma^2_{H_0}(t_{c,1}) = 2(2n+5)/(9n(n-1)) \tag{3.8.87}$$

Finally, we shall study the distribution of $t_{c,1}$ under the alternative hypothesis

$$K_n^*: \Omega(H_1, H_2) = 1 + n^{-\frac{1}{2}} f(H_1)g(H_2) \tag{3.8.88}$$

In this case, we substitute

$$dA_k = dH_{1k}H_{2k}\{1 + n^{-\frac{1}{2}} f(H_{1k})\, g(H_{2k})\} \qquad (3.8.89)$$

in (3.8.75-79) and (3.8.85), then we obtain

$$E_{K_n}^{*}(t_{c,1}) = n^{-\frac{1}{2}}\beta^{*} \qquad (3.8.90)$$

where

$$\beta^{*} = 8\int_0^1 xf(x)dx \int_0^1 yg(y)dy$$

and

$$\sigma^2_{K_n^{*}}(t_{c,1}) = \sigma^2_{H_0}(t_{c,1}) + R_n \qquad (3.8.91)$$

As in the previous subsection $R_n \longrightarrow 0$ as $n \longrightarrow \infty$. Hence for large n,

$$\{t_{c,1} - n^{-\frac{1}{2}}\beta^{*}\}/\sigma_{H_0}(t_{c,1}) \qquad (3.8.91)$$

could be approximated by the normal distribution N(0,1).

CHAPTER IV

KENDALL'S TAU AND SECTOR SYMMETRY FOR RIGHT CENSORED MULTIVARIATE DATA

4.1 Introduction

This chapter extends the generalized Kendall's Taus (= GKT's) to right-censored multivariate data sets. This proposal can also be considered as an extension of Simon's (1977) Kendall's Tau for uncensored multivariate data.

For the discussion, we would write the GKT's in the two-dimensional sample, in Chapter III, as

$$t_{12} = N^{-1} \Sigma'_n A_{2,i,j} = N^{-1} \Sigma'_n a_{1ij} a_{2ij} \qquad (4.1)$$

By substituting appropriate values for N and $A_{2,ij}$, we can obtain the indexes $t_{c,1}$, $t_{c,2}$, $^2_{ij}$, α_1 or α_2 as given in the preceding chapter.

In an m-dimensional case, there is no real valued statistic which can provide a complete description of the association between the m components. So, we should consider a vector valued statistic having as its components GKT's of pairs of components and GKT's of higher order. The latter can be considered as the generalized Simon's Statistics for right-censored multivariate data sets.

This chapter will also introduce a type of symmetry for m-variate variables. This type of symmetry, called sector symmetry, is proposed in Section 4.5. And in Section 4.6, the corresponding statistical tests are presented.

4.2 GKT in Multivariate Problems

Let $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$ be the sample of m-vectors, having uncensored sub-sample $\underline{X}_i$, i=1,...,k; and censored sub-sample $\underline{X}_i$, i=k+1, ...,n. And let

$$(X_{ri}, \delta_{ri}),\ r=1,2,\ldots,m \tag{4.2.1}$$

denote the i-th observation on the r-th component, with $\delta_{ri} = 0$ or 1 according as the observation is in fact uncensored or censored.

In studying multivariate data via GKT's, we examine each of the $\frac{1}{2}n(n-1)$ point pairs and classify each according to the value of

$$a_{rij} = a(X_{ri}-X_{rj}, \delta_{ri}-\delta_{rj}); \tag{4.2.2}$$

for r=1,...,m. In this case Simon's "sign-case diagram" cannot be used as such. So, we define a new variable

$$(X_{0i}, \delta_{0i}) \tag{4.2.3}$$

such that

$$(X_{01}, \delta_{01}) < \ldots < (X_{0n}, \delta_{0n})$$

with $\delta_{0i} = 0$ at least for i=1,2,...,k; and we should take $\delta_{0i} = 0$ for all i whenever we are studying the conditional generalized Kendall's tau. And we will observe the GKT's

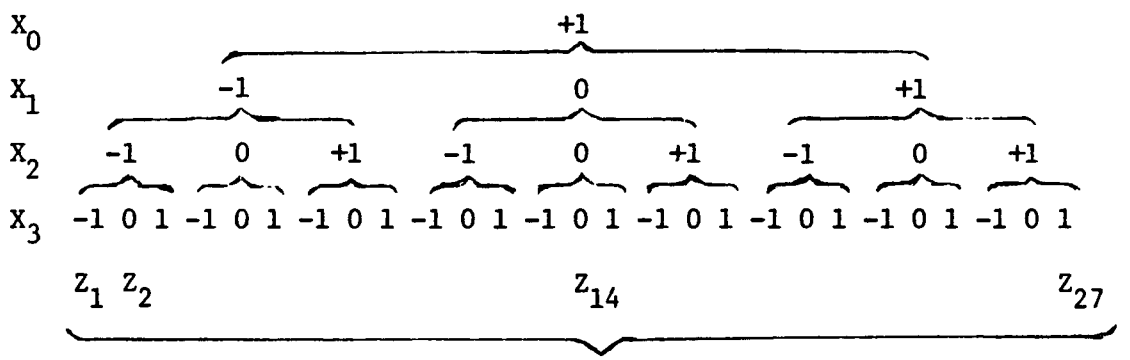
$$t_{0i} = N^{-1} \sum_{i>j} a_{0ij} a_{rij} \tag{4.2.4a}$$

for r=1,2,...,m. It is clear that $a_{0ij}>0$ for all i>j, and hence t_{0r} can be written as

$$t_{0r} = N^{-1} \sum_{i>j} a_{rij} \qquad (4.2.4b)$$

The function a_{rij} takes the values -1, 0, +1, as given in Chapter II. These three possibilities for a_{rij} give rise to 3^m different value categories for a point pair. Table 4.1 shows a "value-case diagram" for m = 3. Note that, without the zero values this diagram is the same as the Simon's diagram in four dimensions.

Table 4.1 Value-Case Diagram for m=3



The fractions of ½n(n-1) pairs.

Now, let Z denote the 3^mx1 vector giving the fractions of the ½n(n-1) point pairs falling into the various categories. For illustration, first we consider the case m=2. In this case we have

$$M_2 * Z = T_2 \qquad (4.2.5)$$

with

$$M_2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix} \qquad (4.2.6)$$

and

$$T_2 = (1, t_{01}, t_{02}, t_{12})' \qquad (4.2.7)$$

where t_{rs} is the GKT between X_r and X_s. If in the sample, there are (X_{ri},δ_{ri}), i=1,...,n which satisfy the conditions on the X_0's, might be after reordering, for some value of r, then we could take that component as X_0. Thence $\underline{Z}$ becomes a 3^{m-1}x1 vector.

Otherwise, we observe t_{0r},r=1,...,m that is the GKT's between X_r and a new variable X_0 as defined above. All observations on X_0 can be considered as having natural ordering. However, we may not be interested in these t_{0r}'s, then we could delete the corresponding rows of the matrix M. Thence, for m=2 we obtain

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix} Z = \begin{bmatrix} 1 \\ t_{12} \end{bmatrix} \qquad (4.2.8)$$

In the following discussion, the vector T has t_{0r}'s as its components.

Now, for m = 3, we have

$$M_3 * Z = T_3 \qquad (4.2.9)$$

where

$$M_3 = \{M_{3(-1)} \mid M_{3(0)} \mid M_{3(1)}\} \qquad (4.2.10)$$

with

$$M_{3(-1)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

$$M_{3(0)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$M_{3(1)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

and

$$T_3 = (1, t_{01}, t_{02}, t_{03}, t_{12}, t_{13}, t_{23}, t_{123})'$$

The index t_{123} can be written as

$$t_{123} = N^{-1} \sum_{i>j} a_{1ij} a_{2ij} a_{3ij} \qquad (4.2.11)$$

Using Simon's notation, $t_{123} = t_{0123}$.

Thus, in general, we would have

$$t_{r_1 \ldots r_s} = N^{-1} \sum_{i>j} A_{s,ij} \; ; \; 2 \leq s \leq m \qquad (4.2.12)$$

where

$$A_{s,ij} = a_{r_1 ij} a_{r_2 ij} \cdots a_{r_s ij} \qquad (4.2.13)$$

for $r_1 < r_2 \ldots < r_s$; $r_k = 1, 2, \ldots, m, k = 1, \ldots, s$.

And there are $m!/(s!(m-s)!)$ combinations leading to indexes with s-components. All these indexes will be considered as the possible GKT's in the m-dimensional data sets. These indexes and the natural indexes $t_{0r}, r = 1, \ldots, m$ can be presented as a matrix equation

$$M_m * Z = T_m \qquad (4.2.14)$$

where Z is the $3^m * 1$ vector of the fractions, as noted above.

M_m is a $2^m x 3^m$ matrix. The first (m+1) rows can be written easily, following the value case diagram as given in Table 4.1. The remaining rows are the element products of combinations of s-rows, $r_1 < r_2 \ldots < r_s$ out of the 2nd to the m+1-st rows of M_m.

T_m is the $2^m x 1$ vector of the GKT's with components

$$1, t_{01}, \ldots, t_{0m}, t_{12}, \ldots, t_{123}, \ldots, t_{12\ldots m}$$

This vector will be considered as a generalized vector valued statistic of Simon's (1977) or GSS (-Generalized Simon's Statistic) for right-censored m-dimensional data sets.

Furthermore, we note that the two matrices:

$$M_m, \text{ and } \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix} \otimes M_{m-1}$$

wnere $\otimes$ denotes the direct or Kronecker product, have the same row-vectors; that is, the first matrix can be obtained from the second by permuting its rows.

4.3 Alternative Presentation of Simon's Statistic

Let $\underline{X}_i, i=1,\ldots,k$ and $\underline{X}_i^*, i=k+1,\ldots,n$ be the uncensored and censored sub-samples of the m-variate variables, then the Simon's statistic (4.2.12) can be written as

$$t_{r_1\ldots r_s} = N^{-1}\{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} A_{s,ij} + \sum_{i=1}^{k}\sum_{j=k+1}^{n} A_{s,ij} + \sum_{i=k+1}^{n-1}\sum_{j=i+1}^{n} A_{s,ij}\} \tag{4.3.1}$$

for $s \leq m$.

This can be considered as an extension of (3.4.1) in Chapter III. Hence, (4.2.12) can be written as

$$t_{r_1\ldots r_s} = N^{-1}\{\binom{k}{2} t_{00;r_1\ldots r_s} + k(n-k)t_{01;r_1\ldots r_s} + N^* t_{11;r_1\ldots r_s}\} \tag{4.3.2}$$

with

$$N = \binom{k}{2} + k(n-k) + N^* \tag{4.3.3}$$

So, $t_{r_1 \ldots r_s}$ is a weighted average of $t_{00;r_1 \ldots r_s}$, the Simon's statistic of the uncensored sub-sample; $t_{01;r_1 \ldots r_2}$, an index of association of Simon's type between uncensored and censored sub-samples; and $t_{11;r_1 \ldots r_s}$, the Simon's statistic of the censored sub-sample.

As in Chapter II, in censored data, there are at least two distinct distributions for $\underline{X}_i$, i=1,2,...,n. If $\underline{X}_i$, i=1,...,k have the same distributions $F(\underline{X})$; and $\underline{X}_i^*$, i=k+1,...,n have d.f.'s $G(\underline{X})$, then

$$\begin{aligned} E(t_{r_1 \ldots r_s}) = N^{-1}\{&\binom{k}{2} \int \ldots \int A_{s,12} dF(\underline{X}_1) dF(\underline{X}_2) \\ &+ k(n-k) \int \ldots \int A_{s,12} dF(\underline{X}_1) dG(\underline{X}_2) \\ &+ N^* \int \ldots \int A_{s,12} dG(\underline{X}_1) dG(\underline{X}_2)\} \end{aligned} \tag{4.3.4}$$

In this case, in fact, we have shown also that

$$E(t_{00;r_1 \ldots r_s}) = \int \ldots \int A_{s,12} dF(\underline{X}_1) dF(\underline{X}_2) \tag{4.3.5}$$

$$E(t_{01;r_1 \ldots r_s}) = \int \ldots \int A_{s,12} dF(\underline{X}_1) dG(\underline{X}_2) \tag{4.3.6}$$

$$E(t_{11;r_1 \ldots r_s}) = \int \ldots \int A_{s,12} dG(\underline{X}_1) dG(\underline{X}_2) \tag{4.3.7}$$

Now, we will show that the index $t_{r_1 \ldots r_s}$ in (4.3.4) can be expressed as a function of the number of concordant and discordant pairs. For two lxs vectors $\underline{X}_i(\underline{\delta}_i)$, i=1,2 taken at random from the population, define a vector $\underline{W} = (w_1, w_2, \ldots, w_s)$ by

$$w_r = \begin{cases} +1 & \text{if } (X_{r1},\delta_{r1}) > (X_{r2},\delta_{r2}) \\ -1 & \text{if } (X_{r1},\delta_{r1}) < (X_{r2},\delta_{r2}) \\ 0 & \text{otherwise} \end{cases} \qquad (4.3.8)$$

And the two vectors are called concordant or discordant if the product $w_1 \cdot w_2 \cdots w_s = +1$ or -1, respectively. Let

$$p_s = P\left(\prod_{r=1}^{s} w_r = +1\right)$$

$$q_s = P\left(\prod_{r=1}^{s} w_r = -1\right) \qquad (4.3.9)$$

Then

$$\hat{p}_s = N_1^{-1} C_s \text{ and } \hat{q}_s = N_1^{-1} D_s \qquad (4.3.10)$$

where C_s and D_s are the numbers of concordant and discordant pairs, respectively, out of the $N_1 = n(n-1)/2$ pairs of observations. Thence, the Simon's statistic (4.2.12) can be written as

$$t_{r_1 \ldots r_s} = (C_s - D_s)/N_1 \qquad (4.3.11)$$

with

$$E(t_{r_1 \ldots r_s}) = p_s - q_s \qquad (4.3.12)$$

If $p_s = q_s$ then $E(t_{r_1 \ldots r_s}) = 0$ and

$$\sigma^2(t_{r_1 \ldots r_s}) = (p_s + q_s)/N_1 \qquad (4.3.13)$$

For all possible values of s ($\geq$m), the index $t_{r_1 \ldots r_s}$ given in (4.3.1) can be presented as a matrix equation. Similar to Section 4.1, it is easy to verify a matrix equation

$$M^* \begin{bmatrix} \underline{Z}_{00} & \underline{0} & \underline{0} \\ \underline{0} & \underline{Z}_{01} & \underline{0} \\ \underline{0} & \underline{0} & \underline{Z}_{11} \end{bmatrix} = \underline{T}_w \tag{4.3.14}$$

where

$$\underline{T}_w = \begin{bmatrix} 1 & 1 & 1 \\ t_{00;01} & t_{01;01} & t_{11;01} \\ . & . & . \\ t_{00;12\ldots m} & t_{01;12\ldots m} & t_{11;12\ldots m} \end{bmatrix} \tag{4.3.15}$$

$\underline{Z}_{00}$, $\underline{Z}_{01}$ and $\underline{Z}_{11}$ are 3^mx1 vectors giving the fractions of $k(k-1)/2$, $k(n-k)$ and N^* point pairs falling into various categories. The subscript (00) indicates pairs within the uncensored sub-sample; (11) indicates pairs within the censored sub-sample; and (01) indicates pairs between the uncensored and censored sub-samples. And the corresponding matrix M may be written as

$$\underline{M} = \underline{J} \otimes \underline{M}_{00} \tag{4.3.16}$$

with $\underline{J} = (1\ 1\ 1)$; $\underline{M}_{00}$ is in fact the matrix $\underline{M}_m$ of (4.3.11) corresponding to the uncensored sub-sample, and $\otimes$ denotes a Kronecher product.

From (4.2.14), (4.3.2) and (4.3.15) we obtain

$$\underline{T}_m = \underline{T}_w * \underline{W} \tag{4.3.17}$$

where

$$\underline{W} = N^{-1}(k(k-1)/2,\ k(n-k),\ N^*)'. \tag{4.3.18}$$

4.4 Dependence Functions and Tests of Independence

4.4.1 Pairwise Independence

Let $\underline{X}_i = (X_{1i},\ldots,X_{mi})$ be the true observation on the i-th subject which may be censored by a variable $\underline{Y}_i = (Y_{1i},\ldots,Y_{mi})$. So, we observe

$$Z_{ri} = \text{Min}(X_{ri},Y_{ri}) \tag{4.4.1}$$

for i=1,...,n; r=1,...,m; along with indicator variables

$$\delta_{ri} = \begin{cases} 0 & \text{if } Z_{ri} = X_{ri} \\ 1 & \text{if } Z_{ri} = Y_{ri} < X_{ri} \end{cases} \tag{4.4.2}$$

As in Chapter III, let $H(\underline{z})$ be the d.f. of $\underline{Z}$ and $H_r(z_r)$, $H_{rs}(z_r,z_s)\ldots$ be its marginal d.f.'s.

Let us define the pairwise dependence function.

$$\Omega_{rs} = \Omega_{rs}(H_r,H_s) = \frac{H_{rs}(z_r,z_s)}{H_r(z_r)H_s(z_s)} \tag{4.4.3}$$

for all pairs (r,s). Then we will consider the problem of testing

$$H_o: \ \Omega_{rs} = 1 \text{ for all pair } (r,s) \tag{4.4.4}$$

against the sequence of alternative

$$K_n^*: \ \Omega_{rs}(H_r,H_s) = 1 + n^{-\frac{1}{2}}\lambda_{rs}(H_r,H_s) \tag{4.4.5}$$

for n-1,2,..., where λ_{rs} is not identically equal to zero (a.e), for at least one pair (r,s), $1 \leq r < s \leq m$.

Using the results in Section 3.8 for each pair of (r,s), the statistic

$$(t_{rs} - n^{-\frac{1}{2}}\beta_{rs}^*)/\sigma_{H_0} \tag{4.4.6}$$

where t_{rs} is the unconditional GKT, and

$$\sigma^2_{H_0} = 2(2n+5)/9n(n+1)$$

$$\beta^*_{rs} = 8\int\int H_r H_s \lambda_{rs}(H_r,H_s)\,dH_r\,dH_s \qquad (4.4.7)$$

has a normal distribution having zero mean and unit variance, as $n \to \infty$.

If we have uncensored data, and if F_{rs}, the joint d.f. of X_r and X_s, is a bivariate normal with correlation ρ_{rs}. Then under

$$K_n: \rho_{rs} = n^{-\frac{1}{2}}\rho_{o,rs} \qquad (4.4.8)$$

we have, from (3.8.53-56),

$$\beta_{rs} = 2\rho_{o,rs}\pi^{-1} \qquad (4.4.9)$$

4.4.2 Non-null Distribution of the Generalized Simon's Statistic

Considering the d.f. of the m-variate variable $\underline{Z}$, let

$$\Omega(H_1,\ldots,H_m) = H(\underline{z}) / \prod_{r=1}^{m} H_r(z_r) \qquad (4.4.10)$$

be the m-variate dependence function, proposed by Puri and Sen (1971). Here, we shall study the distribution of the unconditional GKT:

$$t_{12\ldots m} = \binom{n}{2}^{-1} \Sigma_n' A_{m,ij} \qquad (4.4.11)$$

under a sequence of alternative hypotheses.

$$K_n^{**}: \quad \Omega = 1 + n^{-\frac{1}{2}}\lambda(H_1,\ldots,H_m) \qquad (4.4.12)$$

for n=1,2,... . Using the same method as in the bivariate case, in which $\underline{Z}_i$, i=1,...,n are considered as uncensored, we have

$$E_{K_n^{**}}(t_{12\ldots m}) = n^{-\frac{1}{2}}\beta^{**}_{(m)} \tag{4.4.13}$$

where

$$\beta^{**}_{(m)} = 2\int\ldots\int A_{m,12}\, d\Big(\prod_{r=1}^{m} H_{r1}\Big)\, d\Big\{\lambda(H_{12},\ldots,H_{m2})\prod_{r=1}^{m} H_{r2}\Big\} \tag{4.4.14}$$

And for large n,

$$\sigma^2_{K_n^{**}}(t_{12\ldots m}) \simeq \sigma^2_{H_0}(t_{12\ldots m}) \tag{4.4.15}$$

with

$$\sigma^2_{H_0}(t_{12\ldots m}) = \binom{n}{2}^{-1}\Big\{\int\ldots\int A^2_{m,12}\prod_{r=1}^{m}(dH_{r1}dH_{r2}) + 2(n-2)\int\ldots\int A_{m,12}A_{m,13}\prod_{r=1}^{m} dH_{r1}H_{r2}H_{r3}\Big\} \tag{4.4.16}$$

We obtain

$$\sigma^2_{H_0}(t_{12\ldots m}) = \binom{n}{2}^{-1}\{1 + 2(n-2)\cdot 3^{-m}\} \tag{4.4.17}$$

Based on Hoeffding's theorems, as $n\to\infty$, $t_{12\ldots m}$ has a normal distribution with mean $n^{-\frac{1}{2}}\beta^{**}_{(m)}$ and variance $\sigma^2_{H_0}(t_{12\ldots m})$ under K_n^{**}.

If $\lambda(H_1,\ldots,H_m) = \prod_{r=1}^{m}\lambda_r(H_r)$, then

$$\beta^{**}_{(m)} = 2\prod_{r=1}^{m}\int\int a_{r12}\, dH_{r1}dH_{r2}\lambda_r(H_{r2}) = 2^{m+1}\prod_{r=1}^{m}\int_0^1 u\lambda_r(u)\,du \tag{4.4.18}$$

4.5 A Type of Symmetry

Considering a m-variate variable $\underline{X}$, we propose a type of

symmetry. Observing the ordered values of the components of $\underline{X}$, that is X_r, r=1,...,m; one may be interested in calculating the values of

$$P\{X_{r_1} < X_{r_2} \ldots < X_{r_m}\} = p(r_1, r_2, \ldots r_m) \tag{4.5.1}$$

for all possible $r_i \neq r_j = 1,2,\ldots,m$. Here, however, we are concerned with the equality of those values, instead of the values themselves.

If $F(\underline{X})$ is the d.f. of the variable $\underline{X}$, then

$$p(\underline{r}) = p(r_1,\ldots,r_m) = \int_{S(\underline{r})} \ldots \int dF(\underline{x}). \tag{4.5.2}$$

where

$$S(\underline{r}) = \{\underline{x}: x_{r_1} < x_{r_2} \ldots < x_{r_m}\} \tag{4.5.3}$$

$S(\underline{r})$ may be called the $\underline{r}$-th sector of the m-dimensional space.

Definition 4.5.1

A m-variate variable $\underline{X}$ is said to be sector symmetry if and only if

$$p(r_1, r_2, \ldots, r_m) = p(1,2,\ldots,m) \tag{4.5.4}$$

for all possible orderings of r_i's, that is (m!) orderings.

It is clear that

$$p_{m!} = \Sigma_{\underline{r}} p(\underline{r}) \leq 1 \tag{4.5.5}$$

where the summation is over all m! possible values of $\underline{r}$. The value $p_{m!} = 1$ is attainable if the equalities among any number of the components of $\underline{X}$ have probability zero, or $F(\underline{x})$ is a continuous d.f. Let

$$q = 1 - p_{m!} \tag{4.5.6}$$

then $q = 0$ if and only if $p_{m!} = 1$.

Furthermore, we may consider testing the null hypotheses

$$H_0: p(r_1,\ldots,r_m) = p(1,\ldots,m),\ \underline{r} \in P_m \qquad (4.5.7)$$

where

$$P_m = \{\underline{r}: X_{r_1} < X_{r_2} \ldots < X_{r_m}\} \qquad (4.5.8)$$

is the set of all possible permutations of 1,2,...,m; which will be discussed in the following section. This is a sector symmetry test. This H_0 implies that $F(\underline{x})$ is a symmetric function of its arguments. So, the problem is in fact the same as to test the interchangeability of the m variates $x_1,x_2,\ldots,x_m$ in $F(\underline{x})$. If H_0 is true, then

$$p(\underline{r}) = p \leq 1/m! \qquad (4.5.9)$$

and the equality is attainable if $F(\underline{x})$ is a continuous d.f.

For m=2, a positive bivariate variable, (X_1,X_2) has a unique characteristic. Writing

$$X_1 = R\cos\theta;\ X_2 = R\sin\theta \qquad (4.5.10)$$

for $0 \leq \theta \leq \pi/2$, then H_0 in (4.5.7) reduces to

$$H_{2,0}: \text{Median } (\theta) = \pi/4 \qquad (4.5.11)$$

4.6 Tests for Sector Symmetry

4.6.1 Complete Data Sets

4.6.1.1 Chi-Squared Tests

Considering a random sample of size n on an m-variate variable $\underline{X}$ having d.f. $F(\underline{x})$, let $n(\underline{r})$ be the number of observations belonging to the domain $S(\underline{r})$ in (4.5.3), then the probability that exactly $n(\underline{r})$ observations fall into the sector

$S(\underline{r})$ for all $\underline{r} \in P_m$ is

$$\frac{n!q^{n_0}}{n_0!} \Pi_{\underline{r}} \{n(\underline{r})!\}^{-1}\{p(\underline{r})\}^{n(\underline{r})} \tag{4.6.1}$$

with

$$n_0 = n - \Sigma_{\underline{r}} n(\underline{r}) \tag{4.6.2}$$

where $p(\underline{r})$ is given by (4.5.2) and q is given by (4.5.7). This is the multinomial p.d.f. of (m!) variables $n(\underline{r})$'s.

If $F(\underline{x})$ is continuous, then $q = 0$ and (4.6.1) becomes the multinomial p.d.f. of (m! - 1) variables, that is

$$n!\Pi_{\underline{r}} \{n(\underline{r})!\}^{-1}\{p(\underline{r})\}^{n(\underline{r})} \tag{4.6.3}$$

with

$$\Sigma_{\underline{r}} p(\underline{r}) = 1, \text{ and } \Sigma_{\underline{r}} n(\underline{r}) = n \tag{4.6.4}$$

From this point, it is assumed that the d.f. $F(\underline{x})$ is continuous. In this case, the unbiased estimator of $p(\underline{r})$ is

$$\hat{p}(\underline{r}) = n^{-1} \cdot n(\underline{r}) \tag{4.6.5}$$

Define

$$\chi^2_{m!-1} = \Sigma_{\underline{r}} \frac{\{n(\underline{r}) - np(\underline{r})\}^2}{np(\underline{r})} \tag{4.6.6}$$

Then, as $n \to \infty$, $\chi^2_{m!-1}$ has a limiting chi-squared distribution with (m!-1) degrees of freedom. Hence, $\chi^2_{m!-1}$ is an appropriate statistic for testing the null hypothesis

$$H_0^*: p(\underline{r}) = p_0(\underline{r}) \ , \ \underline{r} \in P_m \tag{4.6.7}$$

where $p_0(\underline{r})$ is a specified number for each $\underline{r}$. And the null hypothesis H_0 in (4.5.8) is a particular case of H_0^*.

If H_0 is true, the statistic

$$\chi^2_{H_0} = \Sigma_{\underline{r}} \frac{\{n(\underline{r}) - n/m!\}^2}{n/m!} \tag{4.6.8}$$

has an approximate chi-square distribution with (m!-1) degrees of freedom. Hence, this is a test for sector symmetry.

4.5.1.2 Coefficient of Sector Symmetry

Let

$$S^2 = \Sigma_{\underline{r}}\{n(\underline{r}) - n/m!\}^2 \tag{4.6.9}$$

be the corrected sum of squares of $n_{\underline{r}}$'s or the observed sum of squares of deviations from the average value, then

$$\text{Max}(S^2) = n^2(1 - 1/m!) \tag{4.6.10}$$

Define

$$R_m^2 = S^2/\text{Max}(S^2) \tag{4.6.11}$$

and call R_m^2 the coefficient of sector symmetry or the index of symmetry. It is clear that, as R_m^2 increases from 0 to +1 the deviations become more different or larger. Hence, this index is an appropriate statistic for testing the null hypothesis H_0 in (4.5.8). Large values of R_m^2 suggest the rejection of H_0. It is easy to verify that

$$\chi^2_{H_0} = (m!-1)nR_m^2 \tag{4.6.12}$$

and

$$R_m^2 = \Sigma_{\underline{r}} \frac{\{n(\underline{r})/n - 1/m!\}^2}{1 - 1/m!}$$

or

$$R_m^2 = \frac{m!}{m!-1} \{\Sigma_{\underline{r}} (n(\underline{r})/n)^2 - 1/m!\} \tag{4.6.13}$$

This implies

$$R_m^2 \leq \frac{m!}{m!-1} \{\Sigma_{\underline{r}} n(\underline{r})/n - 1/m!\} \tag{4.6.14}$$

with the equality is attainable if $n(\underline{r}) = 0$ for all except one value of $\underline{r} \in P_m$. Hence

$$0 \leq R_m^2 \leq 1 \tag{4.6.15}$$

This R_m^2 can be considered as the sample index of symmetry, which is an estimator of the population index of symmetry

$$I_m^2 = \Sigma_{\underline{r}} \frac{(p(\underline{r}) - 1/m!)^2}{1-1/m!} \tag{4.6.16}$$

with $p(\underline{r})$ given by (4.5.2). This index I_m^2 can be considered as a measure of the agreement between $p(\underline{r})$ and $1/m!$.

4.6.2 Censored Data Sets

Considering the m-variate variables $\underline{X}$, $\underline{Y}$, and $\underline{Z}$ discussed in Section 4.4, we have

$$P(\underline{Z} \geq \underline{z}_0) = P(\underline{X} \geq \underline{z}_0)\ P(\underline{Y} \geq \underline{z}_0) \tag{4.6.17}$$

provided the variable $\underline{X}$ and the censoring variable Y are independent. let z_{r0} be the r-th component of $\underline{z}_0$, define

$$\underline{z}(\underline{r}) = (z_{r_1 0}, z_{r_2 0}, \ldots, z_{r_m 0}) \tag{4.6.18}$$

Then the components of $\underline{z}(\underline{r})$ are a permutation of the $\underline{z}_0$'s, for each $\underline{r} \in P_m$.

Now, if the d.f.'s of $\underline{X}$ and $\underline{Y}$, $F(\underline{x})$ and $G(\underline{x})$, are symmetric functions of their arguments, then

$$P(\underline{X} \geq \underline{z}_0) = P(\underline{X} \geq \underline{z}(\underline{r})) \tag{4.6.19}$$

$$P(\underline{Y} \geq \underline{z}_0) = P(\underline{Y} \geq \underline{z}(\underline{r})). \tag{4.6.20}$$

for all $\underline{r} \in P_m$. This implies

$$\begin{aligned} P(Mm(\underline{X},\underline{Y}) \geq \underline{z}_0) &= P(\underline{X} \geq \underline{z}_0)P(\underline{Y} \geq \underline{z}_0) \\ &= P(\underline{X} \geq \underline{z}(\underline{r}))P(\underline{Y} \geq \underline{z}(\underline{r})) \\ &= P(Mm(\underline{X},\underline{Y}) \geq \underline{z}(\underline{r})) \end{aligned} \tag{4.6.21}$$

Hence

$$P(\underline{Z} \geq \underline{z}_0) = P(\underline{Z} \geq \underline{z}(\underline{r})) \tag{4.6.22}$$

for all $\underline{r} \in P_m$, or the d.f. $\underline{Z}$, $H(\underline{z})$, is a symmetric function of its arguments. Thus, it has been shown

Theorem 4.6.1

If C_{os} is the class of symmetric functions, and if

(i) $\underline{X}$ and $\underline{Y}$ are independent random variables

(ii) $F(\underline{x}) \in C_{os}$ and $G(\underline{y}) \in C_{os}$, and

(iii) $\underline{Z} = \text{Min}(\underline{X}, \underline{Y})$,

then $H(\underline{z}) \in C_{os}$.

Under the conditions of Theorem 4.5.1, the null hypothesis H_0 in (4.5.8) for right-censored data on $\underline{X}$ with censoring variable $\underline{Y}$, can be tested using uncensored (complete) data on $\underline{Z}$. Thence, the chi-squared test in (4.6.6) and the index of symmetry in (4.6.13) are applicable for the m-variate variable $\underline{Z}$.

On the other hand, without these conditions, we should consider the component (X_{ri}, δ_{ri}) of $\underline{X}_i$ in (4.2.1). And the classification of $\underline{X}_i$ into the r-th sector $S(\underline{r})$ in (4.5.3) is defined as

$$(X_{r_1 i},\delta_{r_1 i}) < (X_{r_2 i},\delta_{r_2 i}) \ldots < (X_{r_m i},\delta_{r_m i}) \qquad (4.6.23)$$

Otherwise $\underline{X}_i$ is a nonclassified observation. This implies that $\underline{X}_i \in S(\underline{r})$ if and only if $\underline{X}_i$ has components

$$(X_{r_1 i},0) < \ldots < (X_{r_j i},0) < (X_{r_{j+1} i},1) < \ldots < (X_{r_m i},1) \qquad (4.6.24)$$

for $j=0,1,\ldots,m$. The value $j=0$ indicates all components are censored, and $j=m$ indicates uncensored $\underline{X}_i$.

A better approach to classification may be considered, following the Gehan (1965) general pattern, given in Section 3.6. In this case, it is defined that $\underline{X}_i \in S(\underline{r})$ if and only if

$$(X_{r_1 i},0) < \ldots < (X_{r_{m-1} i},0) < (X_{r_m i},\delta_{r_m i}) \qquad (4.6.25)$$

where $\delta_{r_m i} = 0$ or 1. This implies that observations having more than one censored component are considered as non-classified observations. We consider this classification to be more realistic than that in (4.6.23), because if $\delta_{r_m i} = 1$, then the true observation is certainly larger than $X_{r_m i}$ and it belongs to $S(\underline{r})$ anyway.

Now, let n_c be the number of classified observations out of n, then we have

$$\hat{p}(\underline{r} \mid n_c) = n_c^{-1}\, n(\underline{r}) \qquad (4.6.26)$$

the estimator of the conditional probability $p(\underline{r}|n_c)$, instead of (4.5.16). Define a conditional chi-squared statistic

$$\chi_c^2 = \Sigma_{\underline{r}} \frac{\{n(\underline{r}) - n_c/m!\}^2}{n_c/m!} \qquad (4.6.27)$$

For testing the null hypothesis

$$H_{o,c}: p(\underline{r} \mid n_c) = p(1,2,\ldots,m \mid n_c) \qquad (4.6.28)$$

Under the assumption that if $H_{o,c}$ is true, then $F(\underline{x})$ is symmetric, this χ_c^2 becomes an appropriate test for testing the null hypothesis H_0 in (4.5.7).

4.7 Other Statistical Tests

Here, we will consider a special case of the m-variate variable $\underline{X}$ in which the censoring may occur only on a certain component. Based on (4.4.1), without loss of generality, we can write

$$\underline{Z} = (\min(X_1, Y_1), \underline{X}_0)$$
$$\neq \underline{X} \text{ in distribution} \qquad (4.7.1)$$

where $\underline{X}_0$ is a $1x(m-1)!$ vector of the uncensored components of $\underline{X}$, and Y_1 is the censoring variable associated with X_1, the first component of $\underline{X}$. And this special case does not satisfy the conditions of the Theorem 4.6.1.

It is clear that a necessary condition for variable $\underline{X}$ to be interchangeable is that $\underline{X}_0$ should be interchangeable. So, first, we need to test the null hypothesis

$$H_{01}: \underline{X}_0 \text{ is interchangeable} \qquad (4.7.2)$$

by applying the chi-squared test in (4.6.8) or the index of symmetry in (4.6.13). A rejection of H_{01} would clearly indicate the rejection of H_0 in (4.5.7), that is $\underline{X}$ has a symmetric d.f.

On the other hand, the acceptance of H_{01} in (4.7.2) suggests further consideration, that is whether or not the H_0 in (4.5.7) can

be tested. Under a certain assumption, the conditional chi-squared in (4.6.27) is applicable. In general, however, $\text{Min}(X_1,Y_1) = Z_1$ does not have the same d.f. as the marginal d.f.'s of $\underline{X}_0$. This implies the variable $\underline{Z}$ can not be interchangeable. But, this does not imply that $\underline{X}$ is not interchangeable. Hence, in general, H_0 in (4.5.7) can not be tested based on censored data.

Under the assumption that equality among any components of $\underline{X}$ has zero probability, the i-th observations has the following ordered components

$$(X_{r,i},0)<\ldots<(X_{r_{k-1}i},0)\nearrow^{(X_{1i},\delta_{1i})}<(X_{r_{k+1}i},0)<\ldots \tag{4.7.3}$$

where (X_{1i},δ_{1i}) is a possibly censored observation on X_1, which is larger than (k-1) observed components, for k=1,2,...,m. Note that (X_{1i},δ_{1i}) and $(X_{r_{k+j}i},0)$, j=1,2,...,(m-k) are non comparable if $\delta_{1i}=1$. For a given value of k, there are

$$\binom{m-1}{k-1}\times(k-1)!\times(m-k)! = (m-1)! \tag{4.7.4}$$

possible orderings of the m components of $\underline{X}$. Let S_k be the set of all these (m-1)! orderings. Accepting the fact that $\underline{X}_0$ is symmetric or interchangeable, and assuming that all possible orderings are equally likely, then

$$P(\underline{X}_i \in S_k : \underline{X}_0 \text{ symmetric}) = (m-1)!/m! = 1/m \tag{4.7.5}$$

for k=1,2,...,m.

In this case, we may be interested in studying the d.f. of $\underline{X}$ within the set S_k for a given k, under the null hypothesis H_{01}

in (4.7.2). Or we may be considering the conditional d.f. of $\underline{X}_0$ for a given value of $Z_1=\min(X_1, Y_1)$, that is $F(\underline{X}_0 \mid z_k)$ with $(z_k, \underline{X}_0) \in S_k$. Consider the null hypothesis

$$H_{02,k}: F(\underline{X}_0 \mid z_k) = F(\underline{X}_0' \mid z_k) \tag{4.7.6}$$

where $\underline{X}_0'$ is any permutation of the components of $\underline{X}_0$, then the conditional statistic

$$\chi_k^2 = \Sigma \frac{\{n(\underline{r} \mid r_k=1) - n_k\, p(\underline{r} \mid r_k=1)\}}{n_k \cdot p(\underline{r} \mid r_k=1)} \tag{4.7.8}$$

where

n_k = The number of observations belong to S_k

$n(\underline{r} \mid r_k=1)$ = The number of observations belong to $S(\underline{r} \mid r_k=1)$ in (4.5.3)

$p(\underline{r} \mid r_k=1)$ = The conditional probability of $p(\underline{r})$ in (4.5.2) given $r_k=1$

and the summation is over all $\underline{r}$ with $r_k=1$, has an approximate chi-squareddistribution with $((m-1)! - 1)$ degrees of freedom for large values of n_k.

Under $H_{02,k}$, we obtain

$$\chi_k^2 = \Sigma \frac{\{n(\underline{r} \mid r_k=1) - n_k/(m-1)!\}^2}{n_k/(m-1)!} \tag{4.7.9}$$

for $k=0,1,\ldots,(m-1)$.

This shows that we have m conditional chi-squared tests. So, we may combine these tests to obtain an overall test. The method is usually referred to as the summation of chi-squareds procedure. This would lead to a chi-squared having a large number of degrees

of freedom. Hence, it is suggested to consider the following weighted linear combination method.

Let $\underline{r}_0$ be a 1x(m-1) vector of the permutation of (2,3,...,m) or (1,2,...,m-1) denoting an ordering of the components of $\underline{X}_0$ in (4.7.1). Let $\underline{r}_{0i}$ be the i-th possible value of $\underline{r}_0$, then each $\underline{r}_{0i}$ would associate with m elements of $S(\underline{r})$ in (4.5.3). Considering all possible values of i=1,2,...,(m-1)!, and k=1,2,...,m, within each category, we have

n_{ik} = The number of observations in the (i,k)-th cell or category out of the n observations

p_{ik} = The probability that a random observation falls into the (i,k)-th cell

Now let $\underline{p}'$ be the 1x(m-1)! vector of the probabilities with the i-th m components: $p_{i1}, p_{i2}, \ldots, p_{im}$, and let $\underline{N}' = \{(n_{ik})\}$ be the corresponding number of observation vector, then

$$\hat{\underline{p}} = n^{-1}\underline{N} \tag{4.7.10}$$

is the maximum likelihood estimator of $\underline{p}$, or the vector of sample proportions, since $\underline{N}$ has a multinomial density function

$$\Pr\{(n_{ik})\} = n! \prod_{i,k} (n_{ik}!)^{-1}(p_{ik})^{n_{ik}} \tag{4.7.11}$$

Set

$$\underline{U}_n = \sqrt{n}\,(\hat{\underline{p}} - \underline{p}) \tag{4.7.12}$$

then

$$E(\underline{U}_n) = \underline{0}$$

$$\mathrm{Cov}(\underline{U}_n) = \mathrm{Diag}(\underline{p}) - \underline{p}\,\underline{p}' \tag{4.7.13}$$

where Diag($\underline{p}$) is a diagonal matrix having $\underline{p}$ as diagonal elements.

Theorem 4.7.1

As $n \to \infty$, $\underline{U}_n$ has a multivariate normal distribution function with mean vector $\underline{0}$ and covariance matrix (Diag($\underline{p}$) - $\underline{p}\,\underline{p}'$). For a proof see Bishop, Fienberg and Holland (1975), p. 470.

Define weighted linear combinations

$$L_i = n^{-1} \sum_{k=1}^{m} w_k n_{ik} \tag{4.7.14}$$

where

$$s_k = 1/\text{s.e.}\ (\bar{p}_{.k}),\ k=1,2,\ldots,m \tag{4.7.15}$$

is independent of n_{ik} for $i=1,2,\ldots,(m-1)!$. Here we should note that defining the values of w_k in (4.7.15) corresponding to the n_{ik}'s makes them independent with respect to k. However, since the n_{ik}'s are not independent, for correction, the dependency of the n_{ik}'s will be taken into account in computing the variance-covariance matrix of $\underline{L} = (L_1, L_2, \ldots)'$. From (4.7.11), we obtain

$$\mu_i = E(L_i) = \Sigma w_k p_{ik} \tag{4.7.16}$$

$$\sigma_{ij} = \begin{cases} n^{-1}\{\Sigma w_k^2 p_{ik} - (\Sigma w_k p_{ik})^2\}; & i=j \\ -n^{-1} \sum_{k,\ell} w_k w_\ell p_{ik} p_{j\ell} & ; i \neq j \end{cases} \tag{4.7.17}$$

Set $\underline{Q}_n = \sqrt{n}\ (\underline{L} - E(L)) = \sqrt{n}(\underline{L} - \underline{\mu})$, then

$$\underline{Q}_n = (\underline{I} \otimes \underline{w}')\underline{U}_n = \underline{W}'\underline{U}_n \tag{4.7.18}$$

where $\underline{w} = (w_1, \ldots, w_m)'$, $\underline{I}$ is an identity matrix and $\otimes$ denotes the Kronecker's product. Theorem 4.7.1 implies that $\underline{Q}_n$ has a

multivariate normal distribution function, as $n \to \infty$, with mean vector $\underline{0}$ and covariance matrix

$$\underline{V} = \underline{W}'\{\text{Diag}(\underline{p}) - \underline{p}\,\underline{p}'\}\underline{W} = \{(n\sigma_{ij})\} \tag{4.7.19}$$

with σ_{ij} as given in (4.7.17).

Theorem 4.7.2

Let $\underline{Q} \sim N(\underline{\mu}, \underline{V})$. A necessary and sufficient condition that $(\underline{Q}-\underline{\mu})'\underline{A}(\underline{Q}-\underline{\mu})$ have a chi-square distribution is

$$\underline{V}\,\underline{A}\,\underline{V}\,\underline{A}\,\underline{V} = \underline{V}\,\underline{A}\,\underline{V} \tag{4.7.20}$$

in which case the degrees of freedom are Rank $(\underline{A}\,\underline{V})$.

If $|\underline{V}| \neq 0$, then the condition reduces to

$$\underline{A}\,\underline{V}\,\underline{A} = \underline{A} \tag{4.7.21}$$

For a proof see Rao (1973), p. 188.

This theorem implies that the statistic

$$\chi^2 = \underline{Q}_n'\,\underline{V}^{+}\underline{Q}_n \tag{4.7.22}$$

where $\underline{V}^{+}$ is the Moore-Penrose inverse of $\underline{V}$, has an approximate chi-squared distribution function with degrees of freedom rank $(\underline{V}^{+}\,\underline{V})$. If $|\underline{V}| \neq 0$ then $\underline{V}^{+} = \underline{V}^{-1}$ and the chi-squared has $(m-1)!$ degrees of freedom.

Considering the null hypothesis

$$H_0: \underline{p} = \underline{p}_0, (\underline{p}_0 \text{ a fixed probability vector}) \tag{4.7.23}$$

we obtain $\underline{\mu} = \underline{\mu}_0$, if H_0 is true, and the statistic (4.7.21) reduces to

$$\chi^2_{H_0} = n(\underline{L}-\underline{\mu}_0)'\underline{V}_0^{+}(\underline{L} - \underline{\mu}_0) \tag{4.7.24}$$

where $\underline{V}_0$ is the value of $\underline{V}$ in (4.7.19) for $\underline{p}=\underline{p}_0$, which has an approximate chi-squared distribution function with Rank $(\underline{V}_0^+\underline{V}_0)$ degrees of freedom.

Finally, we consider a sequence of alternative hypotheses

$$K_n: \quad \underline{p} = \underline{p} + n^{-\frac{1}{2}}\underline{\lambda} \tag{4.7.25}$$

Under K_n, instead of (4.7.11), set

$$\underline{U}_n = \sqrt{n}(\underline{p} - \underline{p}_0) \tag{4.7.26}$$

so that

$$E_{K_n}(\underline{U}_n) = \underline{\lambda} \tag{4.7.27}$$

and

$$\begin{aligned} \mathrm{Cov}_{K_n}(\underline{U}_n) = {} & \{\mathrm{Diag}(\underline{p}_0) - \underline{p}_0\underline{p}_0'\} + n^{-\frac{1}{2}}\,\mathrm{Diag}(\underline{\lambda}) - 2\underline{p}_0\underline{\lambda}' \\ & + n^{-1}\underline{\lambda}\,\underline{\lambda}', \end{aligned} \tag{4.7.28}$$

see Bishop, Fienberg and Holland (1975). For large n,

$$\mathrm{Cov}_{K_n}(\underline{U}_n) \simeq \{\mathrm{Diag}\,(\underline{p}_0) - \underline{p}_0\underline{p}_0'\} + n^{-\frac{1}{2}}\{\mathrm{Diag}\,(\underline{\lambda}) - 2\underline{p}_0\underline{\lambda}'\} \tag{4.7.29}$$

From (4.7.18) we obtain

$$E_{K_n}(\underline{Q}_n) = \underline{W}'\underline{\lambda} \tag{4.7.30}$$

$$\mathrm{Cov}_{K_n}(\underline{Q}_n) \simeq \underline{W}'\mathrm{Cov}_{K_n}(\underline{U}_n)\underline{W} = \underline{V}_{K_n} \tag{4.7.31}$$

Hence we have

$$\chi^2_{K_n} = \underline{Q}_n'\underline{V}^+_{K_n}\,\underline{Q}_n \tag{4.7.32}$$

which is a noncentral chi-square with noncentrality parameter

$$V = \underline{\lambda}'\,\underline{W}\,\underline{W}'\,\underline{\lambda} \tag{4.7.33}$$

Here, we should note that we may consider using another

approximation for $\mathrm{Cov}_{K_n}(\underline{U}_n)$, instead of (4.7.29), that is

$$\mathrm{Cov}_{K_n}(\underline{U}_n) \simeq \{\mathrm{Diag}(\underline{p}_0) - \underline{p}_0\underline{p}_0'\} \tag{4.7.34}$$

provided $n^{-\frac{1}{2}}$ is sufficiently large. Using this approximation, (4.7.32) reduces to

$$\chi^2_{K_n} = \underline{Q}_n'\underline{V}_0^+\underline{Q}_n \tag{4.7.35}$$

CHAPTER V

CHARTS FOR K-INDEPENDENT SAMPLES

5.1 Introduction

In this chapter we propose two subjects. First, we consider the use of the censored pair chart for K-independent samples, as a descriptive statistic and to test the null hypothesis that the K parent populations have the same distribution functions. Second, we introduce a chart in 3-dimensional space based on 3-independent samples, as an extension of the (2-dimensional) pair chart of Quade (1973), for uncensored data.

5.2 Equality Test for K-Samples

Let X_{ij} be the j-th observation within the i-th sample from a variable X_i with distribution function (d.f.) F_i, for i=1,...,k; j=1,...,n_i. Then we consider the null hypothesis

$$H_0: F_1 = F_2 \ldots = F_K \quad \text{against}$$

$$H_1: F_{i_1} \neq F_{i_2} \quad \text{for at least one pair of } (i_1, i_2)$$

By observing the corresponding pair charts we could detect whether or not we may reject H_0. For this purpose, we could use the following two methods:

(i) Pairwise Comparisons. In this case we could have

$K(K-1)/2$ comparisons or tests whether the two corresponding variables have the same distribution functions. So, we need to construct as many pair charts. The construction of these pair charts or censored pair charts is straightforward, as given in Chapter II. However, the number of the pair charts needed may cause a problem. We should have as many as $K(K-1)/2$ in order to detect that the K populations tend to have the same distribution functions. For rejecting H_0, we may not need so many pair charts as for the acceptance of H_0. But, this problem could be overcome with the help of a computer program.

(ii) Pair Chart based on Breslow (1970) statistic. Breslow defined a vector score statistic

$$\underline{W} = (W_1,\ldots,W_K)'$$

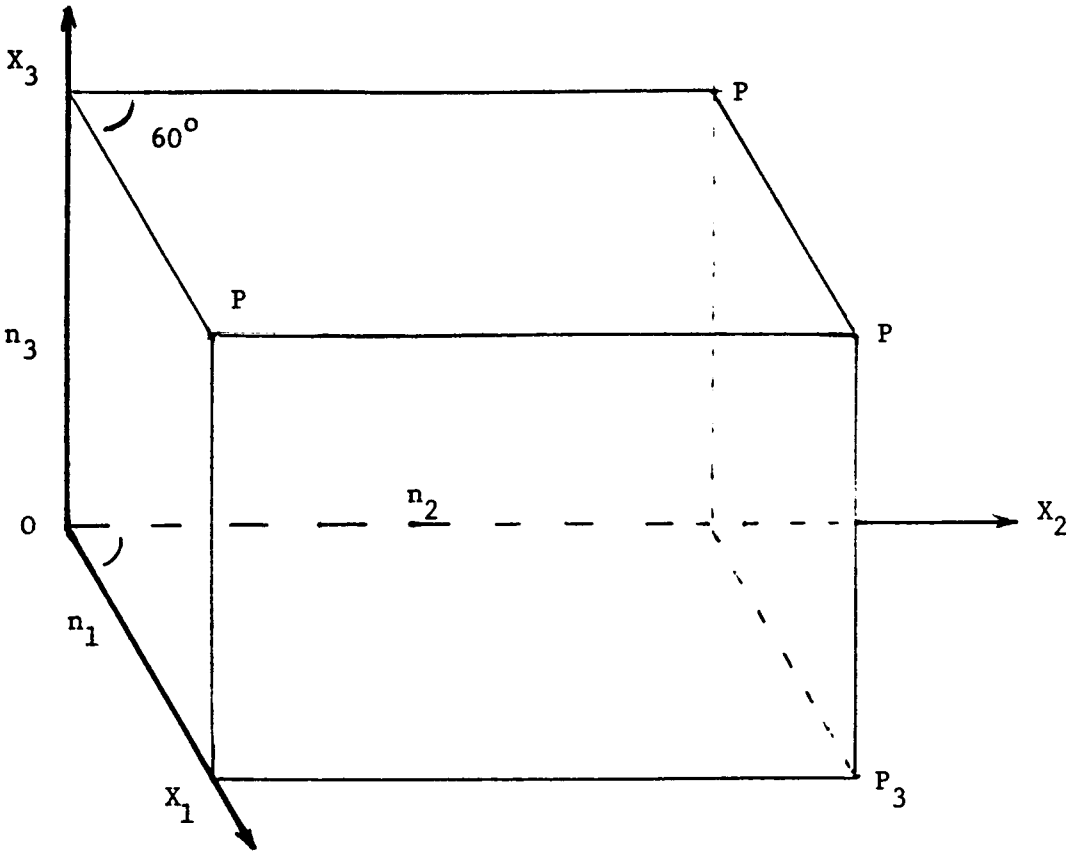
where W_i is in fact the conditional Gehan (1965) W statistic, as given in Section 2.1, comparing the i-th sample with the remaining (K-1) samples. Thence, the use of the censored pair chart becomes obvious. Here, we need at most K pair charts.

5.3 Triplet Chart for Uncensored Data

For K=3, we will consider using a three-dimension chart as a descriptive statistic for making a decision for the rejection of H_0. The construction of the 3-dimensional or triplet chart is as follows.

Considering the uncensored 3-samples of sizes n_1,n_2,n_3 we construct a rectangular parallelepiped of size n_1-units x n_2-units x n_3-units (or $n_1 \times n_2 \times n_3$) as shown in Figure 5.1.

Figure 5.1. The Perspective of the Triplet Chart



This figure shows the perspective of triplet charts with angle $0_1=60^{\circ}$. The rectangular parallelepiped is subdivided into $n_1 \times n_2 \times n_3$ cubes having sides of unit length. Note that in the perspective, a unit length on the X_1-axis should be taken as half a unit length on the X_2-axis or X_3-axis. The path of the triplet chart will start at the point 0 (or the origin) and end at the point P. The construction of this path follows the steps below.

If the smallest observation in the combined samples is an X_i, draw a unit line (= a segment of one unit length) from 0 to the positive direction of X_i, that is from 0 parallel with X_i axis.

From the end of this first unit-line draw a second unit-line parallel with the X_i axis if the second smallest observation is an X_i say. Continue in this manner for all $(n_1 + n_2 + n_3)$ observations. The $(n_1 + n_2 + n_3)$ unit-lines then form the path of a triplet chart from 0 to P which can be considered as the origin (0,0,0) to the point (n_1,n_2,n_3).

For illustration, we consider four data sets: (A), (B), (C), and (D), in Table 5.1. And the corresponding triplet charts are given in Figure 5.2.

Table 5.1 Illustrative Data

(A) X_1: 2, 6, 7, 10, 15, 16

X_2: 4, 5, 11, 13, 17

X_3: 1, 3, 8, 9, 12, 14

(B) X_1: 2, 13, 14, 15

X_2: 1, 6, 8, 9, 10, 11

X_3: 3, 4, 5, 7, 12

(C) X_1: 1, 4, 7, 8

X_2: 2, 4, 4, 8, 9, 10

X_3: 1, 3, 5, 6, 8, 10

(D) X_1: 1, 3, 5, 8

X_2: 2, 4, 6, 9, 11, 13, 15

X_3: 7, 10, 12, 14

Figure 5.2. Triplet Charts for Data in Table 5.1

Chart A

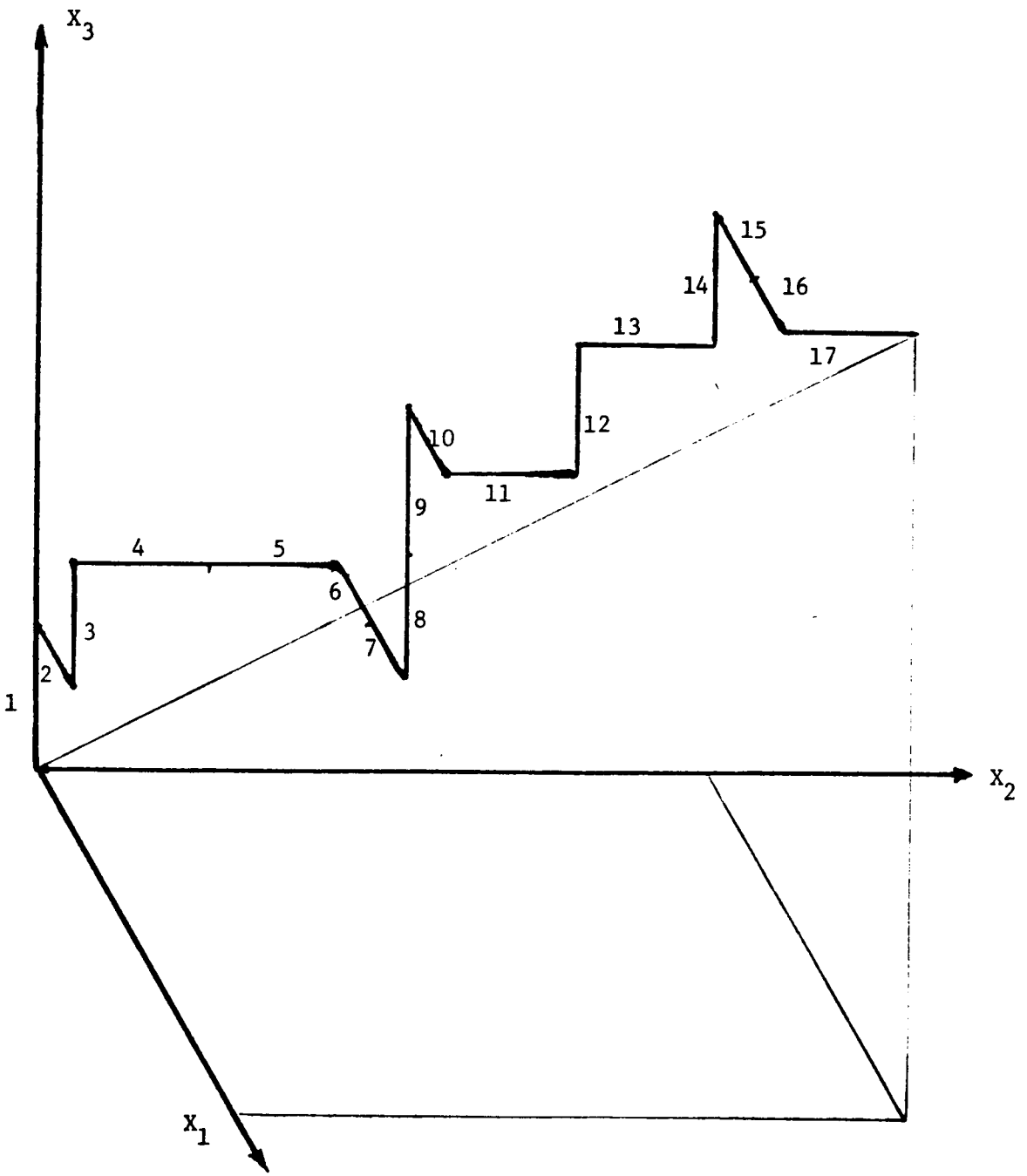


Figure 5.2. Triplet Charts for Data in Table 5.1

Chart B

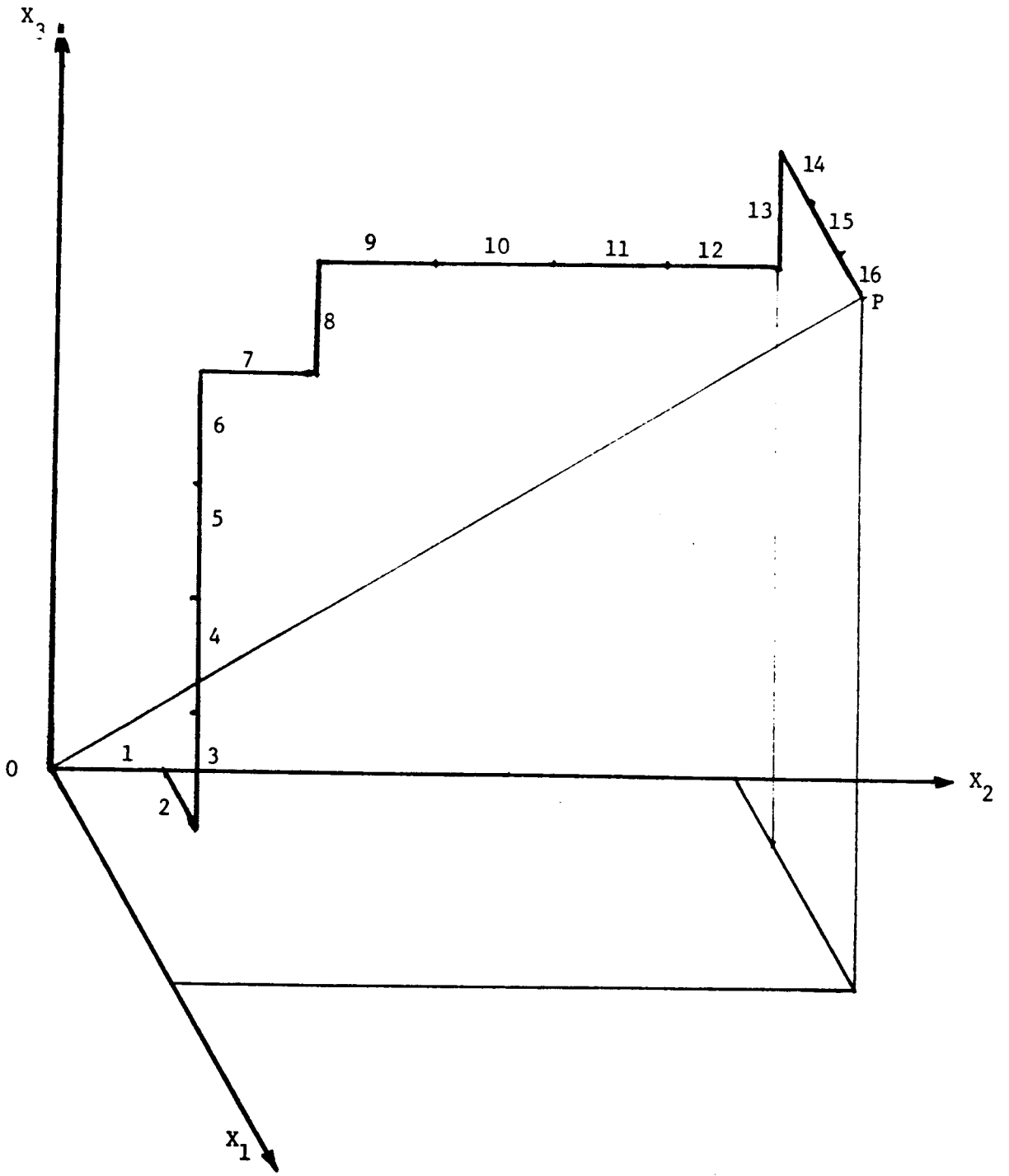


Figure 5.2. Triplet Charts for Data in Table 5.1

Chart C

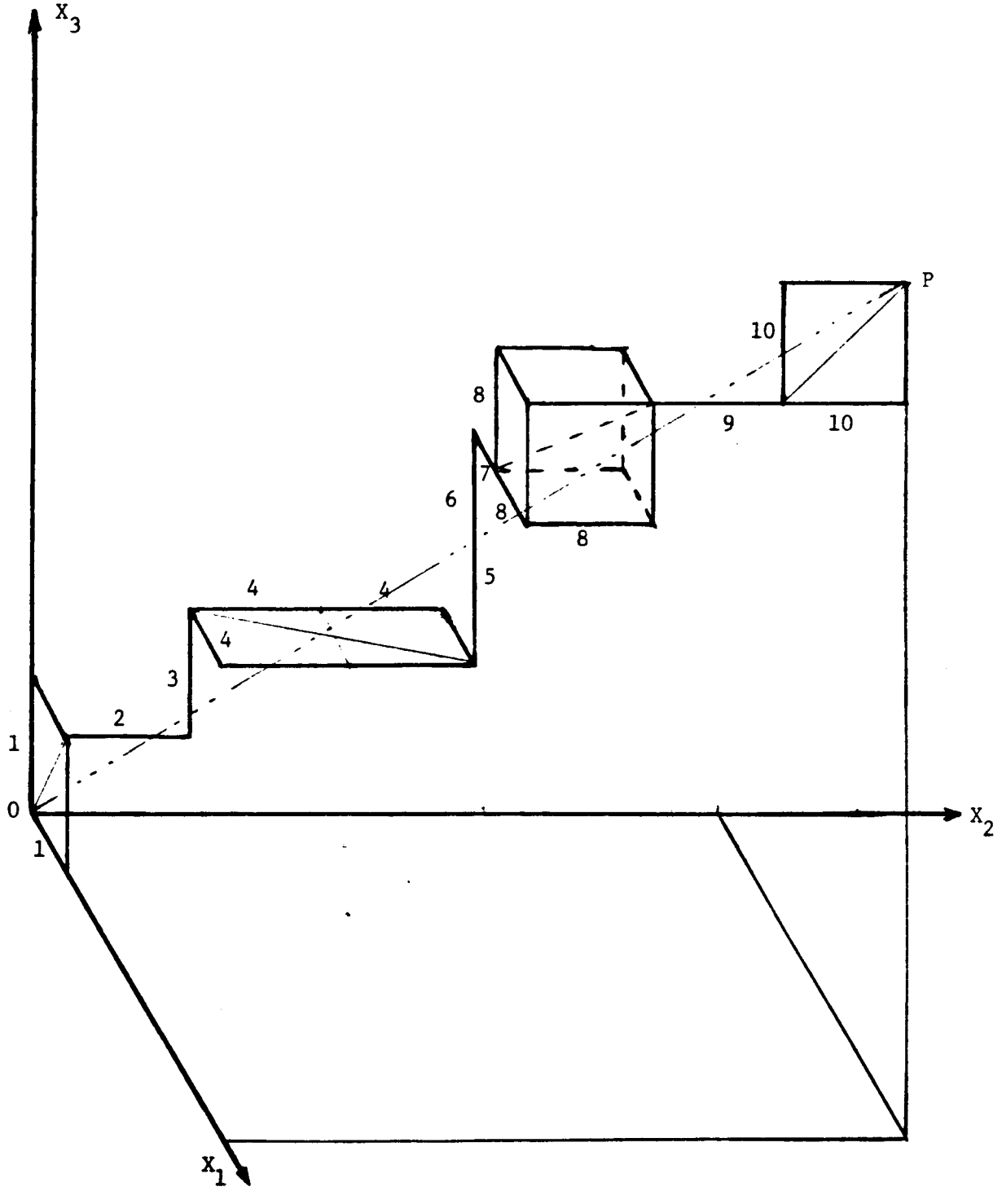
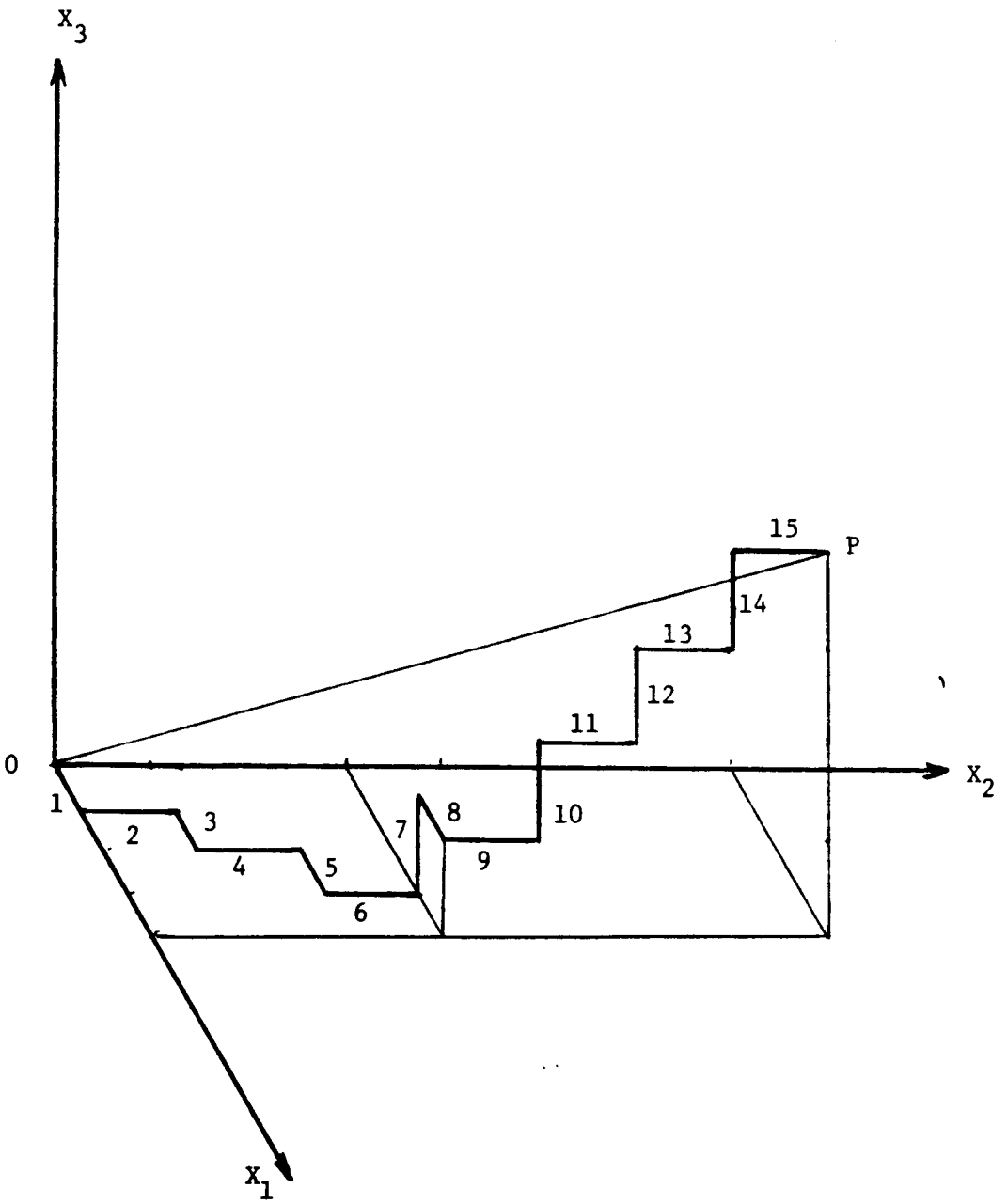


Figure 5.2. Triplet Charts for Data in Table 5.1

Chart D



It is easy to see that the projection of the triplet chart onto the X_i,X_j-plane is the pair chart between X_i and X_j. Note Figure (C) shows how tied observations should be treated. If we have $X_i=X_j\neq X_k$ then we have a rectangle parallel with the X_i,X_j-plane. For data set (C) we have $X_1 = X_3 = 1$; $X_1 = X_2 = 4$ and $X_2 = X_3 = 10$; these observations give the three rectangles along the path of the triplet chart (C). In fact, we have two unit-squares and one rectangle of size 1x2 units. If $X_1 = X_2 = X_3$ then we have a rectangular parallelepiped. For data (C) we have a cube corresponding to $X_1 = X_2 = X_3 = 8$.

5.4 Statistical Tests Based on the Triplet Chart

Considering the usage of a triplet chart as a descriptive statistic, we should observe the distributions of the corresponding unit lines, parallel with X_1,X_2 or X_3 which will be considered as orthogonal, horizontal or vertical unit-lines, respectively, along the path of the chart. The $(n_1+n_2+n_3)$ unit-lines would form broken lines or segments along the path. Let S be the number of segments along a path of a triplet chart, then

$$3 \leq S \leq (n_1 + n_2 + n_3) \qquad (5.4.1)$$

The value $S = 3$ is attainable if all observations on X_i are larger than all observations on X_j, and they are smaller than all observations on X_k, for $i \neq j \neq k = 1,2,3$. The corresponding triplet chart suggests the rejection of H_0, in particular $F_j > F_i > F_k$. However, the value $S = n_1 + n_2 + n_3 = N$ does not directly imply

to the acceptance of H_0. For example, for data (D) in Table 5.1, we have $S = N = 14$ and its triplet chart (D) in Figure 5.2. This suggests us to observe the distributions of the orthogonal segments, the horizontal segments and the vertical segments, which are parallel with X_1-axis, X_2-axis, and X_3-axis respectively. In chart (D), three out of four orthogonal segments lie on the left side (or at the beginning of the path) and all four of the vertical segments lie on the right side of the three orthogonal segments. This does not indicate that $F_1 = F_2 = F_3$. In fact chart (D) suggests that $F_1 > F_3$. Note that at the beginning of the path we have orthogonal and horizontal unit-segments only, and near the end of the path we observe only the horizontal and vertical unit-segments. So, the statistic S is beneficial to show a rejection of H_0. Thence, the triplet chart is beneficial in showing a rejection of H_0. For instance, chart (A) shows that we could not reject H_0, and chart (B) shows a rejection of H_0. In fact, in chart (A) we have $S = 13$ with $n = 17$, and the three types of segments are 'well' distributed along the path. So, based on this chart (A), we may accept the null hypothesis. We define

$$I = (S - 3)/(n_1 + n_2 + n_3 - 3) \qquad (5.4.2)$$

and call I an inequality index or rejection index. It has values from 0 to 1. If I closes to zero then we reject H_0, or we accept the inequality of F_1, F_2, and F_3.

A problem which arises is: if tied observations occur, how should S be counted? First, we consider a cube along the path. It

is clear that the cube shows or indicates a better chance for having $F_1 = F_2 = F_3$, so a cube should be considered as (s x 3) segments, where s is the length of the edge of the cube. Similarly for a rectangular parallelepiped of size $s_1 \times s_2 \times s_3$ units should be counted as $s_1 s_2 s_3$-segments. On the other hand, a rectangle along the path would indicate less chance for having $F_1 = F_2 = F_3$. As the rectangle size increases, the differentiation between F_{i_1} (or F_{i_2}) and F_{i_3} increases, provided X_{i_1} and X_{i_2} generate the rectangle. So, a rectangle will be considered as one segment for S. This would lead to a lower value of I. In chart (C), we have S = 11 and N = 16.

In order to differentiate large values of I or S, we need to observe the maximum value, D, of the distances from the corner points of a path to the diagonal line, OP, of the $n_1 \times n_2 \times n_3$-rectangular parallelepiped. Large value of D would suggest a rejection for the null hypothesis:

$$H_0: \quad F_1 = F_2 = F_3 \tag{5.4.3}$$

whether the value of I is small or large. So, statistic D is worthwhile to study in more detail, instead of S or I.

Let λ_i be the angle between OP, with $P(n_1,n_2,n_3)$, and the X_i-axis then

$$\cos\lambda_i = n_i/\sqrt{(n_1^2 + n_2^2 + n_3^2)}, \; i=1,2,3 \tag{5.4.4}$$

And it is easy to verify that

$$0 < D(n_1,n_2,n_3) \leq \max_i (n_i \sin\lambda_i) \tag{5.4.5}$$

If the vector $\underline{x} = (x_1,x_2,x_3)$ denotes a point on the path of a triplet chart then

$$D(n_1,n_2,n_3) = \text{Sup.}\sqrt{\{\Sigma x_i^2 - (\Sigma x_i \cos\lambda_i)^2} \tag{5.4.6}$$

But the lattice points of the path constitute the locus of points

$$(n_1 F_{n_1}(x_1), n_2 F_{n_2}(x_2), n_3 F_{n_3}(x_3)) \tag{5.4.7}$$

for $-\infty < x_i < \infty$, where

$$F_{n_i}(x_i) = \{\text{number of observations } X_i \text{ such that } X_i \leq x_i\}/n_i \tag{5.4.8}$$

is the empirical distribution function for i=1,2 and 3. Hence (5.4.6) becomes

$$D(n_1,n_2,n_3) = \text{Sup.}\sqrt{\{\Sigma n_i^2 F_{n_i}^2(x_i) - (\Sigma n_i F_{n_i}(x_i)\cos\lambda_i)^2\}} \tag{5.4.9}$$

By substituting (5.4.4), we obtain

$$D(n_1,n_2,n_3) = \text{Sup.}\sqrt{\{\Sigma n_i^2 F_{n_1}^2(x_i) - (\Sigma n_i^2)^{-1}(\Sigma n_i^2 F_{n_i}(x_i))^2\}} \tag{5.4.10}$$

If we have equal sample sizes, then we obtain, by writing $n_i = n$ and $D(n,n,n) = D$;

$$D = \text{Sup.}\left|n\sqrt{\{\Sigma F_{n_i}^2(x_i) - (\Sigma F_{n_i}(x_i))^2/3\}}\right| \tag{5.4.11}$$

Based on the triplet chart, it is easy to verify that

$$0 \leq 3(D-1)/n\sqrt{2} \leq 1 \tag{5.4.12}$$

The distribution of this variable D will be discussed in Section 5.6.

As a descriptive statistic, chart (C) with S = 11 shows a small value of $D(n_1,n_2,n_3)$. Hence we could not reject the null hypotheses H_0 in (5.4.3).

Here, we note that the D statistic is presented as a function of n_i, i=1,2,3. So, we are not considering using standardization.

The idea of standardization was noted by Quade (1973) on the pair chart.

Other statistical tests associated with 3-sample problems were introduced by David (1958) and Conover (1965).

Furthermore, the usage of triplet may be extended for uncensored K-samples. Instead of using pairwise comparisons, as noted in Section 5.2, we would use triple comparison method. However, since $K(K-1)(K-2)/6 > K(K-1)/2$ for $K > 5$, then this method becomes worse than the previous method if K>5 in the sense of the number of triplet charts needed.

5.5 Orthogonal Projections of the Triplet Chart

Two kinds of projections will be considered in this section; these are the projections on the coordinate-planes generated by positive axes X_1, X_2 and X_3 and the projection on the plane

$$X_1 + X_2 + X_3 = 0 \qquad (5.5.1)$$

that is the plane perpendicular to the line $X_1 = X_2 = X_3$.

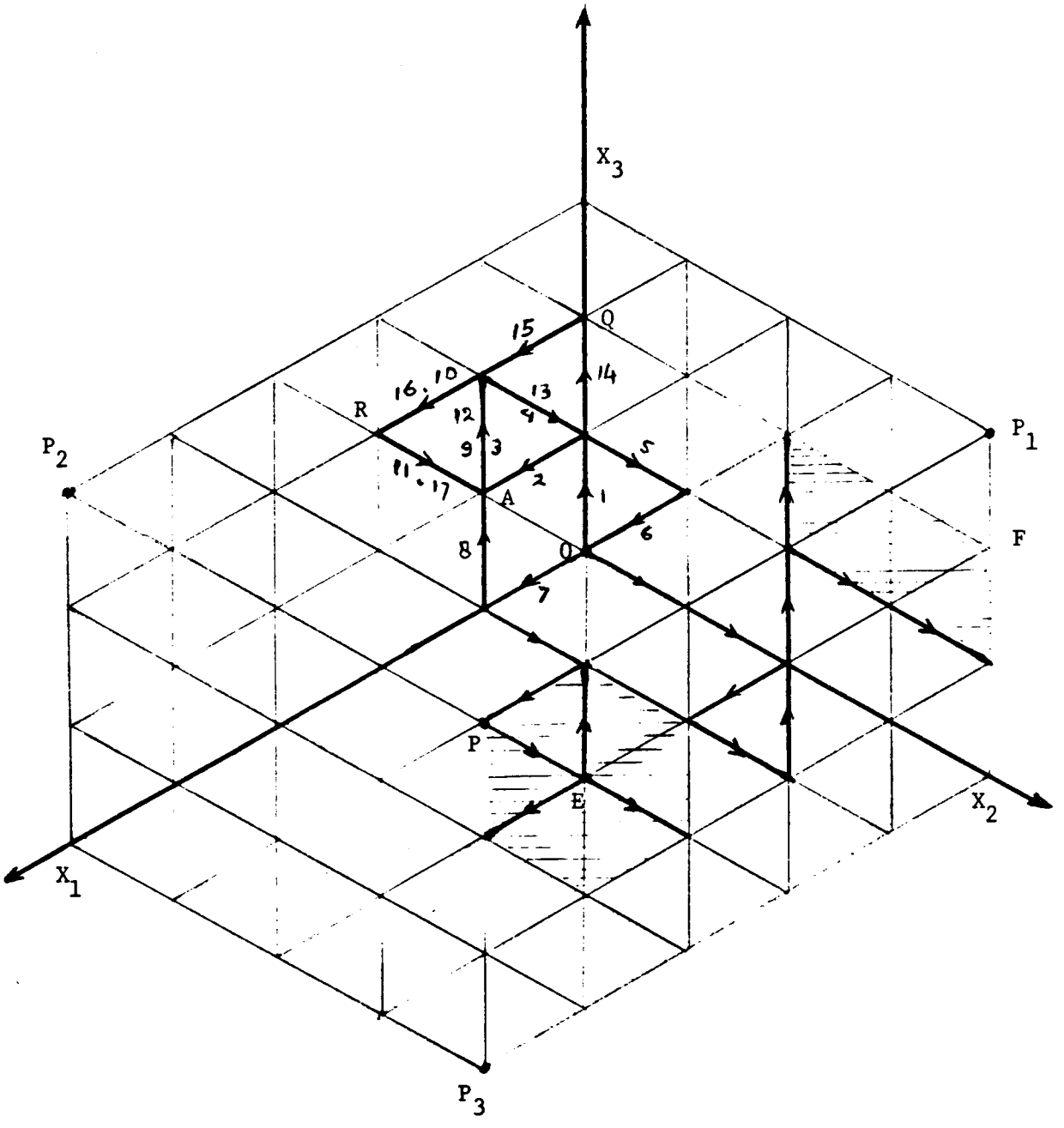
As noted in the previous section, the 3 projections of the first kind are the usual pair charts. Since the pair chart has been discussed in detail, in Chapter II, here we consider only the characteristics of the triplet chart related to its projections. First, considering a segment, in the path of a triplet chart which is parallel with X_i, as the segment increases in length, the differentiation between X_i and X_j, $i \neq j$ increases, because the pair chart between X_i and X_j would have a segment with the same length. And this

segment does not have any effect on the third pair chart, on the X_j, X_k-plane, for $i \neq j \neq k$. Second, we consider a rectangle of size $s_i \times s_j$ parallel with the X_i, X_j-plane, $i \neq j$. Then the pair chart between X_i and X_k (or, X_j and X_k) has a segment of length s_i (or s_j). Hence, the length of s_i (or s_j), clearly determines the differentiation between X_i and X_k (or, X_j and X_k). The third projection, that is the pair chart between X_i and X_j, would have a rectangle of the same size $s_i \times s_j$. Likewise, if we have a rectangular parallelepiped of size $s_i \times s_j \times s_k$-units in the path of the triplet chart. If $s_i = s_j$, and as s_i increases, the differentiation between X_i and X_j decreases. However, a large difference between s_i and s_j may have different effects; for instance the effect on the value of the Mann-Whitney statistic $U(X_i)$. These characteristics would lead us to make a better judgment in observing the triplet chart as a descriptive statistic.

Now, we will start with the second kind of projection. It is well known, that the projection of a unit-cube in the X_1, X_2, X_3-space on the plane (5.5.1) forms a hexagon that is a polygon having six equal sides of length $\sqrt{6}/3$. Figure 5.3 shows the projection of a 5x4x3-rectangular parallelepiped with all its unit-cubes. This projection will be considered as <u>hexagonal projection</u>.

In Figure 5.3, any two adjacent points have a distance of $\sqrt{6}/3$ units. The point 0 is the origin and P is the projection of the point (5,4,3). P_1, P_2 and P_3 are the projections of the points (0,4,3), (5,0,3) and (5,4,0), respectively. Using this type of diagram, the projection of a triplet chart can be constructed as follows.

Figure 5.3. The Projection of a 5x4x3-Unit Cubes



First, we would note that drawing a line in the diagram means drawing a line from one point to the adjacent point along the positive direction of X_i-axis. Then, we observe the ordered values of the combined 3-samples. If the first value is an X_i, draw a line from 0 parallel with X_i to an adjacent point, then from this point draw a line parallel with X_j if the second value is an X_j. Continue in this manner for all observations. For illustration, Figure 5.2 shows the (projection of the) triplet chart of data set (A). The lines are numbered from 1 to 17, from 0 to A, the projection of the point (6,5,6). It is not difficult to count the number of segments, S, along the triplet chart, by following the broken path (projection) from 0 to A.

This hexagonal projection has a special interest if $n_i = n$, because it shows the value of variable D.

Figure 5.3 also shows the path of the triplet charts for the data

	X_1	X_2	X_3
(E)	1,3,5	2,4,5	5
(F)	2	1,1,3,6,6	4,5,6

in order to illustrate the paths having tied observations. For data set (E) we have a cube, and for (F) we have a rectangle of size 2x1.

5.6 The Maximum Distance, D, Statistic

5.6.1 Values of D Statistic

For the distribution of the D statistic as given in (5.4.7), we only consider the case $n_i = n$; that is equal sample sizes. In this case, the hexagonal projection of the triplet chart has a special characteristic, that is the projections of the end points of the path coincide at O. This implies all possible values of D can be seen easily on the hexagonal projection.

As a descriptive statistic, the projection of the triplet chart would show us: (i) the length or value of D; and (ii) the location or distribution of the path around the origin O. By observing these two characteristics, we may or may not reject the null hypothesis, without knowing the distribution of D.

It is easy to see that, if V_n is the set of distinct values of D for samples of sizes n, then V_{n-1} is a subset of V_n. Applying the Pythagorean formula, we can compute easily the values of D. We obtain

$$V_n = V_{n-1} \cup A_n \cup M_n \tag{5.6.1}$$

where

$$A_n = \{(n^2/2 + (n-2k)^2/6)^{\frac{1}{2}} : n-2k > 0\} \tag{5.6.2}$$

$$M_n = \begin{cases} \{n/\sqrt{2}\} & \text{if } n \text{ is even} \\ \phi = \text{empty set}, & \text{otherwise} \end{cases} \tag{5.6.3}$$

which are the sets of the int((n+2)/2) largest values of D corresponding to the 3-samples of size n. For illustration, Table 5.2 shows the values of D for n = 2,3 and 4.

Table 5.2. The Values of D for n = 2,3 and 4

n	D
2	$\sqrt{6/3}$, $\sqrt{2}$, $2\sqrt{6/3}$
3	$\sqrt{6/3}$, $\sqrt{2}$, $2\sqrt{6/3}$, $\sqrt{42/3}$, $\sqrt{6}$
4	$\sqrt{6/3}$, $\sqrt{2}$, $2\sqrt{6/3}$, $\sqrt{42/3}$, $\sqrt{6}$, $\sqrt{2/2}$, $\sqrt{78/3}$, $\sqrt{4\ 6/3}$.

It is clear that if $D_{n,i}$ denotes the i-th ordered values of D, such that $D_{n,1} < D_{n,2} < \ldots < D_{n,k}$, then

$$D_{n,i} = D_{n-1,i} \tag{5.6.4}$$

for $i=1,2,\ldots,\#(V_{n-1})$, where $\#(V_{n-1})$ is the number of elements of the set V_{n-1} with

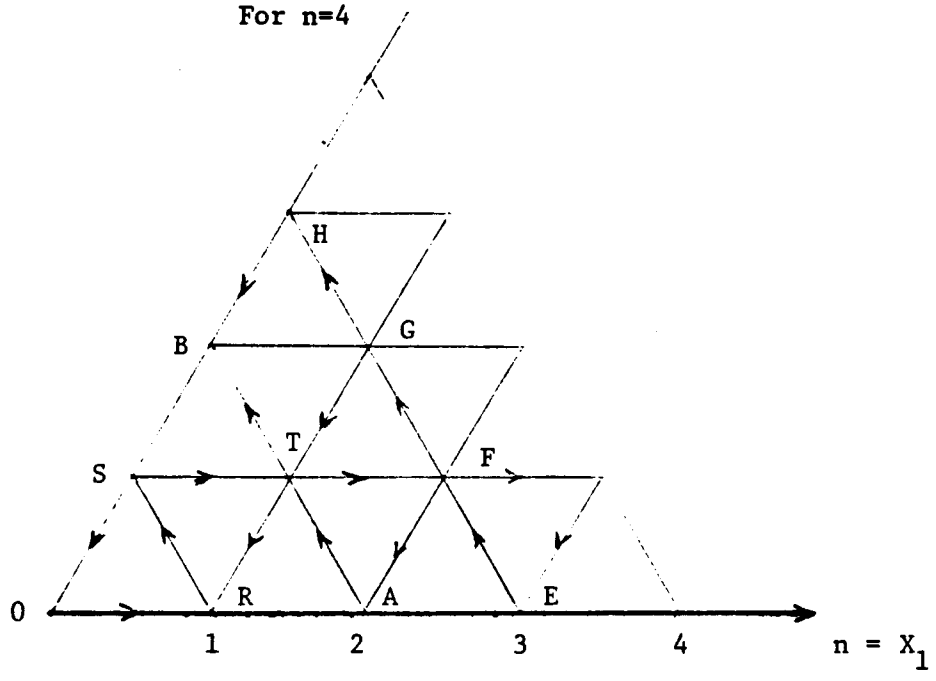
$$\#(V_n) - \#(V_{n-1}) = [(n+2)/2] = \text{INT}((n+2)/2) \tag{5.6.5}$$

The probability function of D will be discussed in the next sub-section under the assumption the parent population has a continuous distribution function. So, there are no tied observations. However, we do not write the probability itself, but the number of paths or triplet charts out of $(3n)!/(n!)^3$, which give a certain value of D.

5.6.2. Recursive Formulas for D Statistic

Here we consider all possible paths which lead to a certain value of D, using a tree diagram. The tree diagram can be constructed by following the paths in the hexagonal projection. For this purpose, we need only one sixth of the projection as given in Figure 5.3.

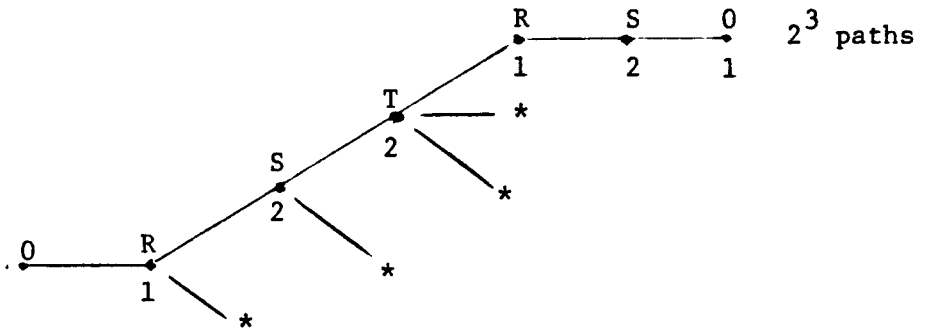
Figure 5.4. One Sixth of the Hexagonal Projection For n=4



In Figure 5.4, we have the n-axis which is parallel with X_i-axis, for one i. Each path should start from 0 following the arrows; the directions of the positive X_i-axis; and returns to 0. Note that there is only one choice from 0 to the point n=1, then from this point there are two choices, one along the n-axis or the '<u>boundary</u>' and the other '<u>inside</u>' of the sector. At this point, the tree diagram has two branches having values 1 along the boundary, and 2 for the inside path. This value 2 is obtained, because of two symmetric choices with respect to the boundary. Furthermore, from a point inside the sector, T say, there are 3 choices. For a certain pair of values D,n; we may or may not use all 3 choices or paths from a certain point. For illustration, first we consider the case $D = \sqrt{2}$, $n \geq 2$. Figure 5.5 shows their corresponding tree diagrams.

Figure 5.5 Tree Diagrams for (a) $D = \sqrt{2}$, $n = 2$ and (b) $D = \sqrt{2}$, $n = 3$

(a) n = 2



(b) n = 3

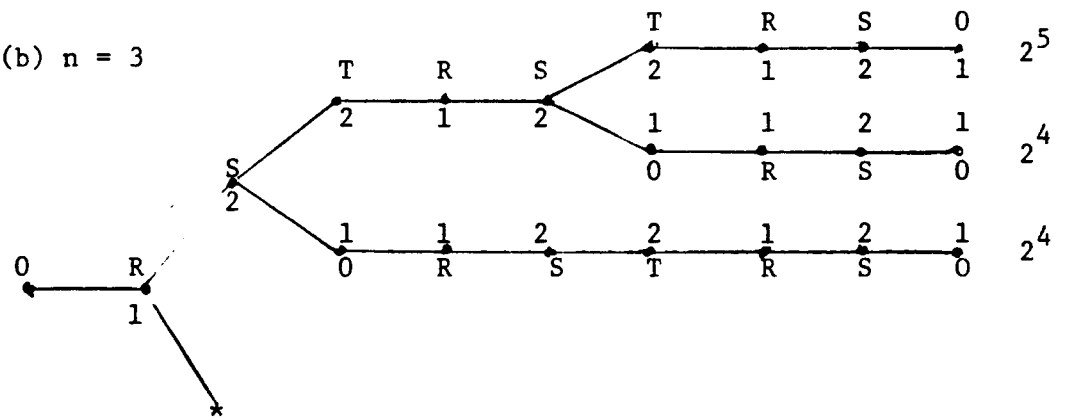


Figure 5.5(a) shows that from each point, R, S and T, we can take only one path or direction, because the others (*)'s lead to larger or smaller values of D. For instance, R-* gives $D = 2\sqrt{6}/3$ and S-* gives $D = \sqrt{6}/3$. So, the tree diagram has only one branch with 6 segments corresponding to the 3x2 observations. And we

obtain 1.2.2.1.2.1 = 8 paths. Thence, if $P_{n,i}$ is the number of paths leading to $D_{n,i}$, the i-th ordered value of D, we obtain

$$P_{2,2} = P_2(D = \sqrt{2}) = 3.2^3 \tag{5.6.6}$$

It is easy to see that

$$P_{n,1} = P_n(D = \sqrt{6}/3) = 6^n \tag{5.6.7}$$

for all n. The value of $P_{n,2}$, n>2 can be calculated based on the tree diagram (b). This tree diagram has 3 branches - the first corresponds to the path which does not enter 0 within the path. This type of path will be written as $O_{n,k}$, so we have

$$O_{3,2} = O_3(D = D_{3,2} = \sqrt{2}) = 3.2^5$$

This path circles the triangular RST twice. As n increases, the number of this recircling increases. Then we obtain

$$O_{n,2} = 3.2^{2n-1} \tag{5.6.8}$$

for all n > 1.

Now, $P_{3,2}$ can be written as

$$P_{3,2} = P_{1,1}P_{2,2} + P_{2,2}P_{1,1} + O_{3,2} = 12P_{2,2} + 3.2^5 \tag{5.6.9}$$

instead of using a tree diagram. The first term denotes the paths, which enter the point 0 once with $D = \sqrt{2}$. Thence, we have $P_{3,2} = 3.2^7 = 384$. Considering all possible partitions of 4, we can compute

$$P_{4,2} = P_{1,1}P_{1,1}P_{2,2} + P_{1,1}P_{2,2}P_{1,1} + P_{2,2}P_{1,1}P_{1,1} + P_{2,2}^2 + P_{1,1}O_{3,2} + O_{3,2}P_{1,1} + O_{4,2}$$

Keeping $D = \sqrt{2}$, the first 4 terms in the right indicate all possible

paths (in the projection) which reach 0 twice before the end of the paths, or the paths reentered 0 three times. The second 2 terms indicates the paths reentered 0 two times. This can be written as

$$P_{4,2} = 3P_{1,1}^2P_{2,2} + P_{2,2}^2 + 2P_{1,1}0_{3,2} + 0_{4,2} \qquad (5.6.10)$$

In general, we can write

$$P_{n,2} = F(P_{1,1},P_{2,2},0_{3,2},\ldots,)_{n,2}) \qquad (5.6.11)$$

that is a function of $P_{1,1},P_{2,2},0_{3,2},\ldots,0_{n,2}$ for n>3. This function has the following form.

$$P_{n,2} = \sum_{k=1}^{[n/2]*} c_k P_{1,1}^{n-2k}P_{2,2}^k + \sum_{k=3}^{n} \sum_{r+s+kt=n}^{*} c_k' P_{1,1}^r P_{2,2}^s 0_{k,2}^t \qquad (5.6.12)$$

where Σ^* denotes the summation over all possible choices, by keeping in mind the number of reentering of the path to 0; that is 1,2,...,n, c_k and c_k' depend on k. For instance,

$$P_{5,2} = 4P_{1,1}^3P_{2,2} + 3P_{1,1}P_{2,2}^2 + 3P_{1,1}^2 0_{3,2} + 2P_{2,2}0_{3,2} + 2P_{1,1}0_{4,2} + 0_{5,2} \qquad (5.6.13)$$

where the coefficients indicate the number of distinct orderings of the corresponding factors.

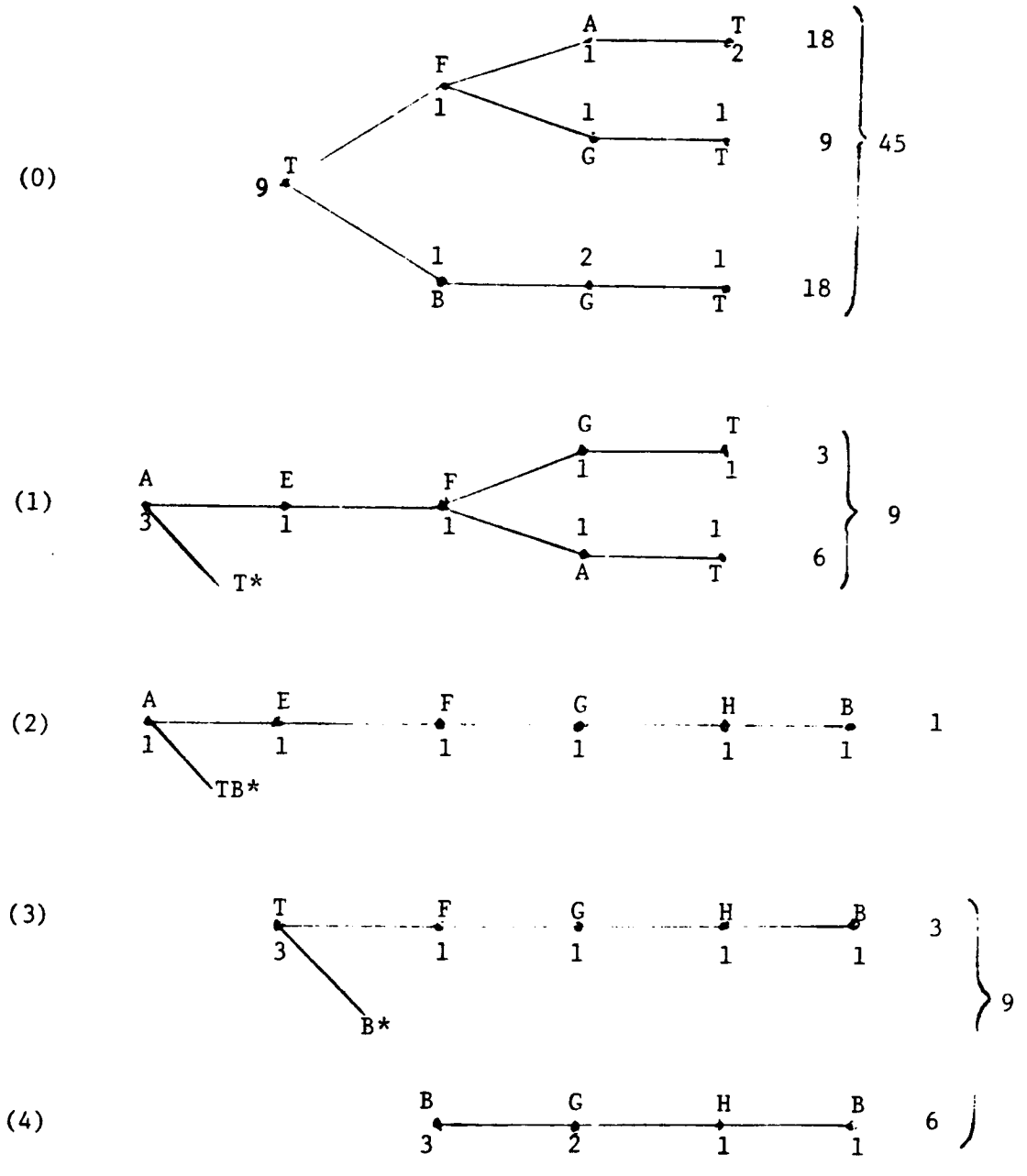
Similarly, we obtain

$$P_{2,3} = P_2(D = 2/3\sqrt{6}) = 3(2 + 2^3) = 30 \qquad (5.6.14)$$

$$0_{n,3} = 3(2 + 2^3)2^{3(n-2)} - 3.2^{2n-1} \qquad (5.6.15)$$

$$P_{n,3} = \sum_{k=1}^{[n/2]*} c_k P_{1,1}^{n-2k} P_{2,3}^k - \sum_{k=3}^{n} \sum_{r+s+kt=n}^{*} c_k' P_{1,1}^r P_{2,2}^s 0_{k,2}^t \qquad (5.6.16)$$

Figure 5.6. Tree Diagrams for Computing $P_3(D_{3,i})$ from $P_2(D_{2,j})$.



for $n \geq 3$. Note that the first product in $O_{n,3}$ indicates the number of paths through the points T and A or B, see Figure 5.4; and the second product indicates the paths through the point T only, or $O_{n,2}$. And the formula of $P_{n,3}$ can be obtained from that of $P_{n,2}$ by replacing the subscript 2.

This tree diagram method and the previous general formula can be easily extended for higher ordered values of D. For instance, we obtain $P_{3,4} = P_3(D = \frac{1}{2}\sqrt{42}) = 270$ and $P_{3,5} = P_3(D = \sqrt{6}) = 114$ using tree diagram. It is clear that, as n increases, the tree diagram becomes more complex. So, we would like to compute these values from $P_{2,2} = P_2(D = \sqrt{2}) = 24$ and $P_{2,3} = P_2(D = 2/\sqrt{3\sqrt{6}}) = 30$ as follows.

In Figure 5.4, $P_{3,4}$ denotes the number of paths through the point F or G, but not E and H. These paths can be constructed from $P_{2,2}$ and $P_{2,3}$, the paths through A, B or T. Again, we use tree diagram as given in Figure 5.6 - (0).

The diagram (0) starts at T, where the (24 + 30)/6 = 9 paths, giving $D = \sqrt{2}$ or $2\sqrt{6}/3$, go through. This gives $P_3(D = OF = OG) = 6x45 = 270 = P_{3,4}$. In order to compute $P_{3,5} = P_3(D = OE = OH)$ we need diagrams (1) - (4). In diagram (1) the segment AT* is replaced by AEFGT (=4 segments). Here segment AT is counted as the segment of 3 paths out of $P_{2,3}$. In diagram (2) segment ATB*, which is the part of OATBO triplet chart, is replaced by two branches. Similarly, TB* is replaced by TFGHB, and T is a point of three paths. Finally, start from B, a point of 3 paths, we put the path BGHB. Hence, we obtain

$$P_{3,5} = (3+6+1+3+6)x6 = 114$$

In these diagrams, we should note that we really use one side of $n=X_1$-axis, because we use one sixth of $(P_{2,2} + P_{2,3})$. So, segment EF is not counted as twice, but segment BG is.

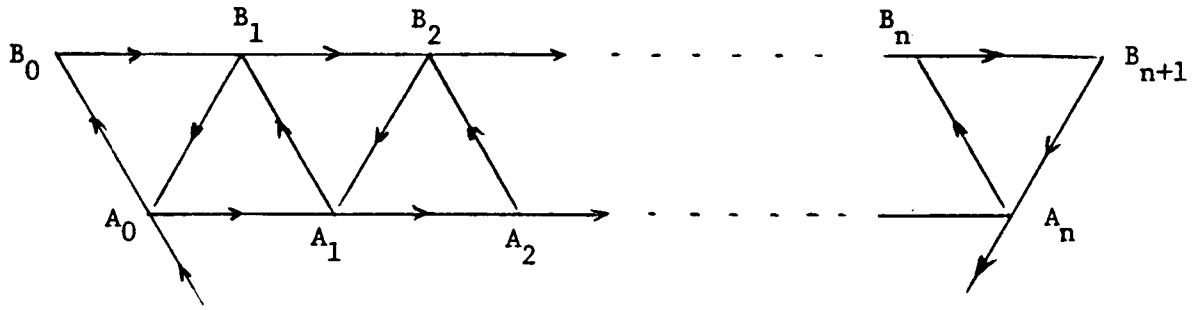
For further discussion, we need to compute the number of paths passing each point E, F, G and H. These can be taken from the five previous tree diagrams. Table 5.3 shows the *F = 24 denotes the number of paths not containing any segments FG or FGH; FG = 16 denotes the number of paths containing FG, including the FGH = 4. Likewise for the other boundary points or segments.

Table 5.3 Paths Containing the Boundary Points or Segments for n=3

Points	Segments	# of Paths
E	EF	10
	EFG	4
	EFGH	1
F	*F	18+6 = 24
	FG	9+3+1+3 = 16
	FGH	1+3 = 4
G	*G	9+18+3 = 30
	GH	1+3+6 = 10
H	*H	1+3+6 = 10

Note that for point E, there are no *E segment, because all paths (=10) contain the segment EF. Based on this table we could compute the paths for n=4. In general we would have the following diagram

Figure 5.7. The Paths Between Two Sample Sizes n and (n+1)



where A_i; i=1,...,n and B_j; j=0,...,n+1 are the points on the side of one sixth of the hexagonal projection corresponding to 3-samples of sizes n and (n+1), respectively. Having the number of paths containing each point A_i, we can compute the number of paths through point B_j. In fact, we need to find the number of paths having points B_j or B_{n-j+1}, and those paths do not have any point outside the segment $B_jB_{j+1}\cdots B_{n-j+1}$, which will be written as

$$P_{n+1}(B_j \text{ or } B_{n-j+1}) \tag{5.6.17}$$

for j=0,.,...,[(n+1)/2}. Then, if (n+1) = 2m, we have

$$P_{n+1}(D = OB_j) = P_{n+1}\{D = (2m^2 + \tfrac{2}{3}j^2)^{\frac{1}{2}}\} = 6\ P_{n+1}(B_j \text{ or } B_{n-j+1}) \tag{5.6.18}$$

However, if (n+1) = 2m+1, that is for odd sample sizes, we have

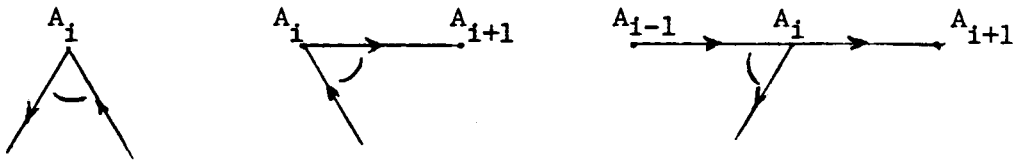
$$P_{n+1}(D = OB_j) = P_{n+1}\left[D = \{(2m+1)^2 + j^2/9\}^{\frac{1}{2}}\right] = 6\ P_{n+1}(B_j \text{ or } B_{n-j+1}) \tag{5.6.19}$$

for $j \neq 0$.

Now, we will proceed with computing the value of $P_{n+1}(B_j \text{ or } B_{n-j+1})$. For this purpose we should distinguish between the cases

$j \leq 1$ and $j > 1$. First, we should note that there are three types of paths through each point A_i, as shown in Figure 5.8.

Figure 5.8. Types of Path at Each Point



The number of paths of these types will be written as A_{ik} for k = 0, 1 or 2 according to whether the paths having 0 segment in A_0A_n,(a); 1 segment A_iA_{i+1},(b), or two segments $A_{i-1}A_i$ or A_iA_{i+1},(c), respectively. So

$$P_n(A_i) = A_{i0} + A_{i1} + A_{i2} \tag{5.6.20}$$

would indicate the total number of paths containing the point A_i. Then for $j \leq 1$, using tree diagram, we obtain

$$P_{n+1}(B_j) = \sum_{i=1}^{n} (1 + 2\delta_{1i})A_{i2} \tag{5.6.21}$$

where $\delta_{11} = 1$ and $\delta_{1i} = 0$ for $i \neq 1$. Hence

$$P_{n+1}(B_n \text{ or } B_{n+1}) = 1 + 2 \sum_{i=1}^{n-1} (1 + 2\delta_{1i})A_{i2} \tag{5.6.22}$$

Thence

$$P_{n+1}(D = \frac{n+1}{2}\sqrt{2}) = 6\{1 + 2 \sum_{i=1}^{n-1} (1 + 2\delta_{1i})A_{i2}\} \tag{5.6.23}$$

or

$$P_{n+1}(D = \frac{n+1}{2}\sqrt{2}) = 6\{-1 + 2 \sum_{i=1}^{n} (1 + 2\delta_{1i})A_{i2}\} \tag{5.6.24}$$

Also

$$P_{n+1}\left[D = \{n+1)^2/2 + (n-1)^2/6\}^{\frac{1}{2}}\right] = 6\{-1 + \sum_{i=1}^{n}(1 + 2\delta_{1i})A_{i2}\} \tag{5.6.25}$$

Next, for $j > 1$, the $P_{n+1}(B_j \text{ or } B_{n-j+1})$ will be written as

$$2P_{n+1}(B_j \text{ or } B_k, j < k < n-j+1) + P_{n+1}(B_j \text{ and } B_{n-j+1}) \tag{5.6.26}$$

where the second term indicates the number of paths containing only the segment B_jB_{n-j+1} of the boundary B_0B_{n+1}. We obtain

$$P_{n+1}(B_j \text{ or } B_{n-j+1}) = 2\{P_n(A_j) + \sum_{i=j}^{n-j-1} A_{i2}\} + A_{n-j,2} \tag{5.6.27}$$

for $j=2,3,\ldots,\{(n+1)/2\}$, and $j \neq n-j+1$.

If $(n+1) = 2m$, then for $j = m$, we would have

$$B_{m0} = P_{2m-1}(A_m) \tag{5.6.28}$$

So, we may write

$$P_{n+1}(B_j \text{ or } B_{n-j+1}) = \begin{cases} P_n(A_j) \text{ if } 2j = n+1 \\ 2\{P_n(A_j) + \sum_{i=j}^{n-j-1} A_{i2}\} + A_{n-j,2} \\ \text{otherwise} \end{cases} \tag{5.6.29}$$

Using this result we can easily obtain the probabilities of D having values corresponding to the boundary points for any sample size.

Finally, we need to consider computing

$$P_{n+k}(A_i \text{ or } A_{n-i}) \tag{5.6.30}$$

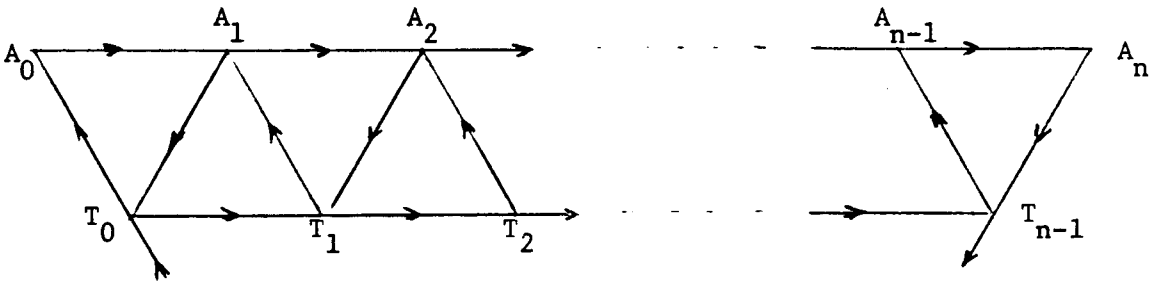
from $P_n(A_i)$, $i=0,1,\ldots,n$; for $k > 0$. Formulas (5.6.12) and (5.6.16) show the results for $n=2$. In contrast with the previous paragraphs, in which we observe the paths and the corresponding tree diagrams between A_0A_n and B_0B_{n+1} or outside A_0A_n; here we should observe the

path inside A_0A_n. So, (5.6.28) clearly indicates the paths which can not go outside of A_0A_n, because these paths should lead to $D = OA_i$. That is,

$$P_{n+k}(D = OA_i) = 6P_{n+k}(A_i \text{ or } A_{n-i}) \tag{5.6.31}$$

as shown in Figure 5.9.

Figure 5.9. The Paths for Sample Size (n+k) in the Hexagonal of Size n.



At this point, it is understood that the number of paths through each point inside the hexagonal of size-n has been computed. Hence, we have the values of

$$P_n(A_i \text{ or } A_{n-i}) \tag{5.6.32}$$

and

$$P_n(T_i \text{ or } T_{n-i-1}) \tag{5.6.33}$$

for $i \leq n/2$. Having these values, including the number of paths having a certain type through each point A_i or T_i, as given in Figure 5.9; we can compute (5.6.30), as follows.

First, we consider the case $k = 1$, then we use this result, as a recursive formula, to obtain $P_{n+k}(A_i \text{ or } A_{n-i})$ for $k > 1$. Having the values of (5.6.32-33), we can compute

$$P^*_{n+1,i} = P^*_{n+1}(A_i \text{ or } A_{ni}) \tag{5.6.34}$$

for $i < n/2$ by applying formula (5.6.27-29). Then we obtain:

$$P^{*}_{n+1,i} = \begin{cases} 1 + 2\sum_{j=1}^{n-2}(k + 2\delta_{1j})T_{j2} & \text{if } i = 0,1 \\ 2\{P_n(T_i) + \sum_{j=1}^{n-i-2} T_{j2}\} + T_{n-i-1,2} & \text{if } 1<i<n/2 \\ P_n(T_i) & \text{if } i = n/2 \text{ exist} \end{cases} \tag{5.6.35}$$

This value indicates the number of paths having $D = OA_i$, which pass along T_iA_i or A_iT_{i-1} only once, corresponding to the 3-samples of sizes (n+1). These paths would pass twice along a segment inside the hexagonal of size (n-1). In fact, they circle a certain triangular once along each path. So, in addition to $P^{*}_{n+1,i}$, we should find the number of paths, inside the hexagonal of size n, which circle either the triangular $T_iA_iA_{i+1}$ or $T_iT_{i+1}A_{i+1}$. This will be written as

$$P^{**}_{n+1,i} = P^{**}_{n+1}(A_i \text{ or } A_{n-i}) \tag{5.3.36}$$

Thence

$$P_{n+1}(D = OA_i) = 6(P^{*}_{n+1,i} + P^{**}_{n+1,i}) \tag{5.6.37}$$

Now, we still need to find the value of (5.6.36). This value can be computed from the previous computed values of (5.6.32). Having these values, we should have also the values of

$$P_{n,i,1} = P_n(A_i \text{ or } A_j, i < j < n-i) \tag{5.6.38}$$

and

$$P_{n,i,2} = P_n(A_i \text{ and } A_{n-i}) \tag{5.6.39}$$

as noted in formula (5.6.26). These values could be computed from

$T_{n-1,j}$, $j=0,1,\ldots,(n-1)$ using formulas (5.6.27-28).

By observing the possible circling once around either the triangular $A_iA_{i+1}T_i$ or $A_iT_{i-1}T_i$, we obtain

$$P^{**}_{n+1,i} = \begin{cases} 2P_{n,i,1} + 1, & \text{if } i = 0 \\ 2\{(2 + \delta_{1i})P_{n,i,1} + P_{n,i+1,1}\} + & \\ +\{(2 + \delta_{1i})P_{n,i,2} + P_{n,i+1,2}\} & \text{if } 0 < i < n/2 \\ A_{i0}, & \text{if } i = n/2 \text{ exist} \end{cases} \tag{5.6.40}$$

where the Kronecker $\delta_{1i} = 1$ if $i=1$ and $\delta_{1i} = 0$ otherwise, and A_{i0} is the number of paths, which touch the line A_0A_n at point A_i, see Figure 5.9.

Thence, substituting (5.6.35) and (5.6.40) in (5.6.37), we obtain $P_{n+1}(D = DA_i)$ for $i \leq n/2$. So, we have shown the existence of a recursive formula of the form

$$P_{n+1}(D = OA_i) = F\{P_n(A_j), P_n(T_{j'}); \; i < j \leq n/2, \; j' = 0, \ldots,(n-1)\} \tag{5.6.41}$$

Thence, using this formula k times we can compute (5.6.31) for $k > 1$. Formulas (5.6.12) and (5.6.14) show a result for n=2, which can be written as

$$P_{2+k}(OA_i) = \sum_{j=1}^{(k+2)/2}{}^{*} c_j P_1^{k+2-2j}(\sqrt{6}/3)P_2^j(OA_i) + \sum_{j=3}^{k+2} \sum_{r+s+jt=k+2}{}^{*} c'_j P_1^r(\sqrt{6}/3)P_2^s(OA_i)O_j^t(OA_i) \tag{5.6.42}$$

for i=0 or 1 according to whether $OA_i = 2\sqrt{6}/3$ or $\sqrt{2}$.

5.7 Triplet Chart for Censored Data

Triplet chart may be used for censored data if we can transform the combined 3-samples into uncensored data. In this case we are to use the estimated values, instead of the observed values. This transformation could be done, for instance, in life testing problems using Nelson's (1972) hazard plotting method. However, this method assumes that the parent populations have a known distribution function, such as the exponential, Weibull, extreme value, normal or log-normal distribution.

So, this method is a combination of parametric and nonparametric procedures.

5.8 Alternative Presentation of the Triplet Chart

Associated with the samples of sizes n_i, i=1,2,3, we observe the hexagonal projection as a polar coordinate system having 0 as the origin and X_1 as the axis, see Figure 5.2. And let

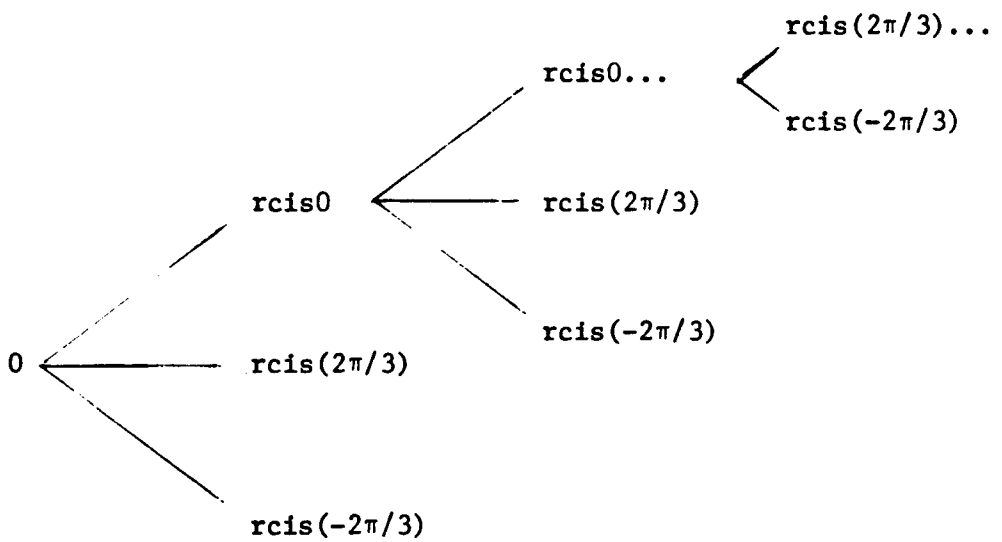
$$\text{rcis}\alpha = r(\cos\alpha + i \sin\alpha), \qquad (5.8.1)$$

where $i = \sqrt{-1}$, be a vector from 0 to a lattice point, for certain values of r and α. For example for $r = \sqrt{6}/3$ and $\alpha = 0, 2\pi/3$ and $-2\pi/3$ denote the points on the X_1, X_2 and X_3 axes, respectively, next to the origin. The projection of the point $P(n_1,n_2,n_3)$ will be written as $r(n_1,n_2,n_3)\ \text{cis}\alpha(n_1,n_2,n_3) = r^*\ \text{cis}\alpha^*$. If $n_1 = n_2 = n_3$ then $r^* = 0$.

Assuming there are no tied observations between the X_i's, the path from 0 to $r^*\text{cis}\alpha^*$ would satisfy the equation

$$\sqrt{6}/3 \sum_{j=1}^{N} \text{cis}\alpha_j = r^*\text{cis}\alpha^* \qquad (5.8.2)$$

where $N = n_1 + n_2 + n_3$ and $\alpha_j = 0, +2\pi/3$ or $-2\pi/3$ with n_1 of $\alpha_j = 0$; n_2 of $\alpha_j = +2\pi/3$, and n_3 of $\alpha_j = -2\pi/3$. The j-th value of α is associated with the j-th lattice point in the path. The $N!/(n_1!n_2!n_3!)$ orderings of the values of α_j's corresponds to the number of paths from (0,0,0) to (n_1,n_2,n_3). The paths from one lattice point to the next point can be presented as a tree diagram:



From each point there are three possible paths or branches. Until we observe n_1 of zeros, n_2 of $2\pi/3$'s, or n_3 of $-2\pi/3$'s. Then the number of branches from each point decreases to 2; and finally to 1; until we have N branches along each path. Thus, a path can be presented by a Nxl vector $\underline{\alpha} = (\alpha_1,\alpha_2,\ldots,\alpha_N)'$ having n_1 components of zeros, n_2

components of $2\pi/3$'s and n_3 components of $-2\pi/3$'s. It is clear that a path from the origin to the next k-th lattice point can be written as

$$\underline{\alpha}_{(k)} = (\alpha_1, \alpha_2, \ldots, \alpha_k)' \tag{5.8.3}$$

for $1 \leq k < N$. And the k-th lattice point has polar coordinate

$$\sqrt{6}/3 \cdot \sum_{j=1}^{k} \text{cis}\alpha_j \tag{5.8.4}$$

If the k-th lattice point is the origin then

$$\sum_{j=1}^{k} \cos\alpha_j = 0 \tag{5.8.5}$$

and

$$\sum_{j=1}^{k} \sin\alpha_j = 0 \tag{5.8.6}$$

This implies $k = 3m$ for some m, and $\underline{\alpha}_{(k)}$ would have m of each possible components; 0, $2\pi/3$ and $-2\pi/3$.

Now, we may consider the ordered observations of the combined samples associated with its vector $\underline{\alpha}$. Let $X_{i(j)}$ be the j-th ordered observation coming from the i-th sample, then the j-th component of $\underline{\alpha}$ is

$$\alpha_j = \begin{cases} 0 & \text{if } X_{i(j)} = X_{1(j)} \\ 2\pi/3 & \text{if } X_{i(j)} = X_{2(j)} \\ -2\pi/3 & \text{if } X_{i(j)} = X_{3(j)} \end{cases} \tag{5.8.7}$$

for $j=1,2,\ldots,N$. Now, let

$$\phi_j = 3\alpha_j/2 \tag{5.8.8}$$

be the j-th component of the vector $\underline{\phi}$, then ϕ_j can take a value -1, 0, or $+1$. And it is clear that $\underline{\phi}$ can be considered as an alternative

presentation of the triplet chart. For illustrations, based on data sets (A), (B) and (D) in Table 5.1, we obtain the following $\underline{\phi}$ vectors, as the alternative presentation of their triplet charts given in Figure 5.2.

$$\underline{\phi}_A = (-1,0,-1,1,1,0,0,-1,-1,0,1,-1,1,-1,0,0,1),$$

$$\underline{\phi}_B = (1,0,-1,-1,-1,-1,1,-1,1,1,1,1,1,-1,0,0,0),$$

and

$$\underline{\phi}_D = (0,1,0,1,0,1,-1,0,1,-1,1,-1,1,-1,1).$$

In the case of tied observations, in particular between samples, we may use "upper line" above the corresponding elements of $\underline{\phi}$. For data set (C) in Table 5.1, we have

$$\underline{\phi}_C = (\overline{-1,0},1,-1,\overline{0,1,1},-1,-1,0,\overline{-1,0,1},1,\overline{-1,1}).$$

If we are concerned only with the rejection index I given in (5.4.2), then we should consider using this vector $\underline{\phi}$, instead of the triplet chart. Because it is much easier to obtain the elements of $\underline{\phi}$ than to construct the triplet chart.

Finally, we consider computing the value of the D statistic based on this vector representation, if we have equal sample sizes, n say. It is easy to verify that

$$D = \underset{1<k<3n}{\text{Max.}} \left| \frac{\sqrt{6}}{3} \sum_{j=1}^{k} \text{cis}\alpha_j \right|, \qquad (5.8.8)$$

provided $\underline{\alpha}$ is a fixed vector for a given data set.

CHAPTER VI

APPLICATIONS TO DEMOGRAPHIC DATA

6.1 Applications of the Censored Pair Chart

To illustrate application of the CPC, we use a subsample of data from the National Fertility Survey of the United States, 1970 (Ryder and Westoff, 1977). In this study, we compare the experiences of white women and black women on their first birth interval (FBI) and the time to separation of their first marriage (TSFM). However, this study is limited to women who married for the first time at 18 years of age.

Table 6.1 shows a statistical analysis, based on "Proc Means" of SAS, of the FBI data for the women under consideration, after deleting all births occurring before 7 months after marriage. Figure 6.1 shows their CPC of Type-I, which is represented as a vertical bar chart. Let X_1 and X_2 denote the blacks' and the whites' FBI, respectively; in Figure 6.1 we have statistics $A(X_i)$,$C(X_i)$, $Q(X_i)$ and $U(X_i)$ for i=1,2. This figure shows that

$$U(X_1) < U(X_2) \tag{6.1.1}$$

$$A(X_1) < A(X_2) \tag{6.1.2}$$

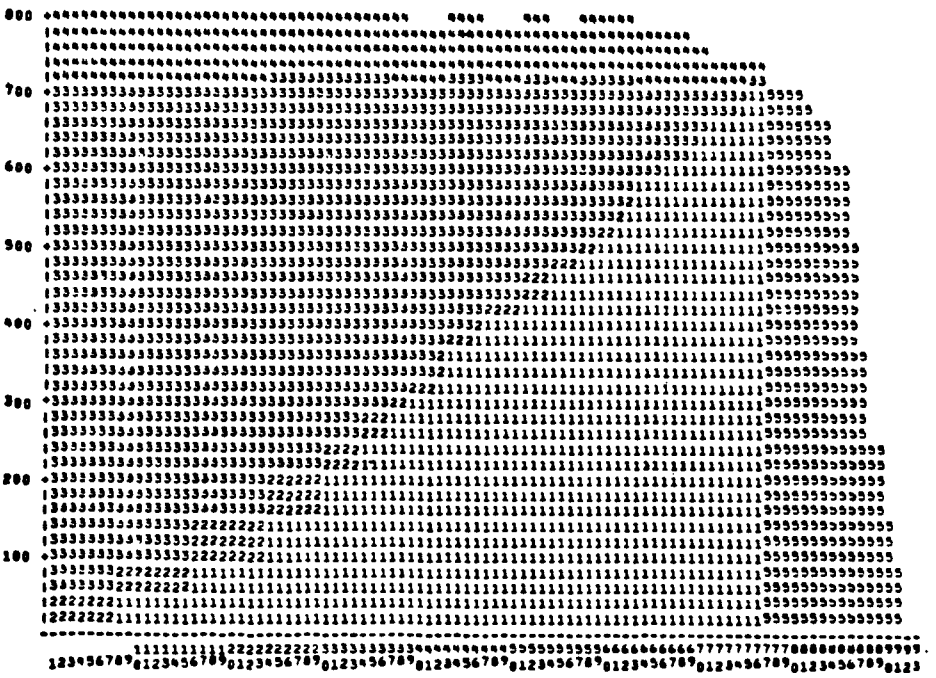
where $U(X_1)$ is the Mann-Whitney U statistic for the complete observations, and $W=A(X_1)-A(X_2)$ is Gehan's W statistic for the censored

Table 6.1. First Birth Interval for Women Marrying for the First Time at Age 18, Based on the U.S.-N.F.S., 1970

Statistic	Race			
	White		Black	
	Complete	Incomplete	Complete	Incomplete
Sample Size	707	113	76	17
Mean FBI	22.08	67.22	19.70	67.65
S.D.	20.94	73.62	17.38	92.35
Minimum	7.00	0.	7.00	1.00
Maximum	218.00	329.00	93.00	339.00
S.E. of Mean	0.79	6.93	1.99	22.40
C.V.	94.84	109.53	88.26	136.51

Figure 6.1. The CPC-I of the First Birth Interval Between Black and White Women Marrying for the First Time at Age 18, Based on the U.S.-N.F.S., 1970

X_2 = White



X_1 = Black

Symbols:	1	2	3	4	5
Statis.	$\{U(X_1)-Q/2\}$	Q	$\{U(X_2)-Q/2\}$	$C(X_2)$	$C(X_1)$

data. The inequalities (6.1.1) and (6.1.2) suggest that X_1 and X_2 do not have the same distribution functions for their first birth intervals. In fact, they suggest that the white women tend to have longer first birth interval than the black women.

Furthermore, we obtain $U(X_1) = 23821.50$, $U(X_2) = 29910.50$, $A(X_1) = 31396.50$ and $A(X_2) = 36298.50$. Hence $W = -4902$.

Looking at the path of the pair chart based on the complete observations, we may wonder whether the two groups have a significant difference with respect to their FBI, because the path is relatively close to the diagonal line. Here, however, we are not interested in doing further statistical tests.

Now, we will consider their first marriage experiences. Table 6.2 shows a statistical analysis, based on "Proc Means" of SAS, of the TSFM data for the women under study. In this case, right censoring occurs, because they are still married at the time of survey. So, it is not surprising that we have a large number of incomplete observations in both samples, that is 690 out of 816 observations for the whites and 58 out of 92 observations for the blacks. Let Y_1 and Y_2 denote the blacks' and the whites' TSFM, respectively. Figure 6.2 shows

$$U(Y_1) > U(Y_2) \tag{6.1.3}$$

$$A(Y_1) < A(Y_2) \tag{6.1.4}$$

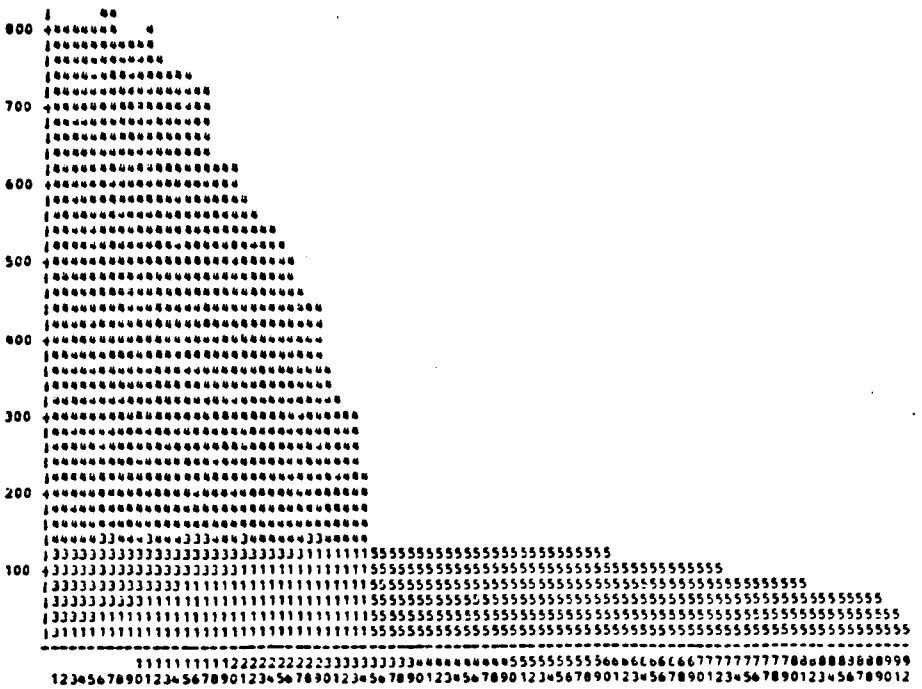
Considering only the complete observations, (6.1.3) suggests that the white women tend to have shorter time to separation than the blacks. However, based on the whole data set, complete

Table 6.2. Time to Separation From First Marriage for Women Marrying for the First Time at Age 18, Based on the U.S.-N.F.S., 1970

Statistic	Race			
	White		Black	
	Complete	Incomplete	Complete	Incomplete
Sample Size	126	690	34	58
Mean	84.77	178.04	109.26	168.57
S.D.	69.29	89.94	78.54	95.45
Minimum	0	5.00	11.00	23.00
Maximum	307.00	360.00	292.00	339.00
S.E. of Mean	6.17	3.42	13.47	12.53
C.V.	81.74	50.51	71.88	56.63

Figure 6.2. The CPC-I of the Time of Separation from the First Marriage Between Black and White Women Marrying for the First Time at Age 18, based on the U.S.-N.F.S., 1970

Y_2 = White



Y_1 = Black

Symbols:	1	2	3	4	5
Statis.	$\{U(Y_1)-Q/2\}$	Q	$\{U(Y_2)-Q/2\}$	$C(Y_2)$	$C(Y_1)$

and incomplete observations, (6.1.4) suggests that the whites have longer time to separation than the blacks. Anyway, both inequalities suggest the rejection of the null hypothesis that Y_1 and Y_2 have the same distribution functions.

Furthermore, we obtain $U(Y_1) = 2563.50$, $U(Y_2) = 1720.50$, $A(Y_1) = 8191.50$ and $A(Y_2) = 18667.50$. Hence $W = A(Y_1)-A(Y_2) = 10476$.

Remarks:

(1) Note that, in Figure 6.1 and Figure 6.2, a unit length on the vertical axis is much shorter than on the horizontal axis. This situation should be taken into consideration in comparing the corresponding regions in the bar charts. For example, Figure 6.1 suggests $C(X_1) < C(X_2)$; but in fact $C(X_1) > C(X_2)$.

(2) The previous two applications show us in general how the result based on only the complete observations may be affected by the censored observations. In the second application, as shown by (6.1.3) and (6.1.4) as well as Figure 6.2, the result based on the whole data set contradicts the result only based on the complete data set.

(3) Such a large number of censored observations, in the second application, arouses some questions about the distributions of the censoring variables of the white and the black groups. In this case, the censorings occur at a fixed point of time, that is, the time of survey. However, the subjects entered the studied groups at random. So, by considering the subjects as having entered at a point of time, the censoring can

be considered as random.

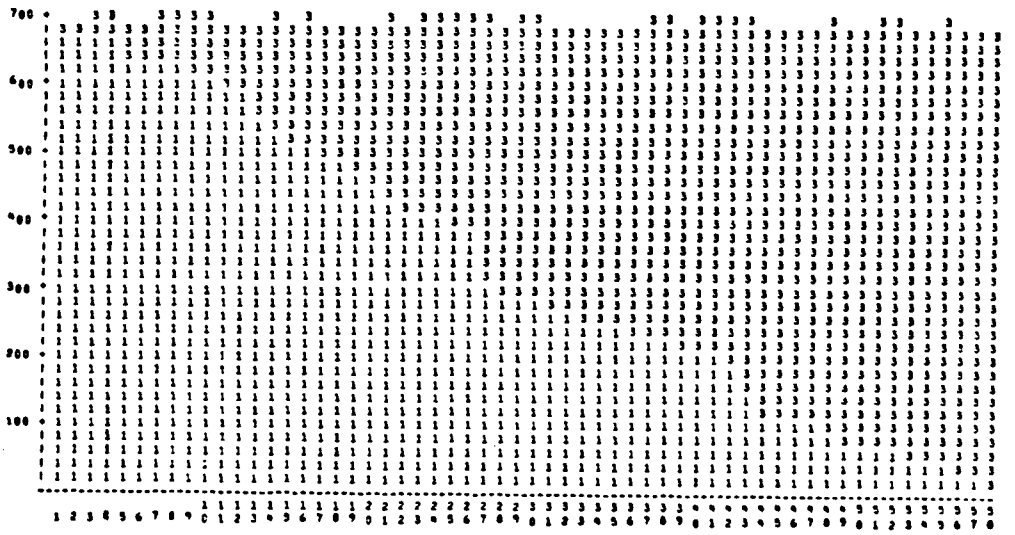
Furthermore, because we have large number of censored observations in both samples, we may consider testing whether the censoring variables have the same distribution functions. For this purpose, we can use the pair chart. Keeping in mind that the censored observations were assumed to have decreasing ordering, we obtain the pair chart as given in Figure 6.3 with diagonal line from upper left corner to lower right corner. This pair chart suggests that the censoring variables have the same distribution functions, since its path, that is the boundary between the area having symbol (1) and that having symbol (3), is relatively close to the diagonal line. As noted in Chapter V, this pair chart can be standardized following Quade's result, such that we would have a square, instead of a rectangle.

The values of the corresponding Mann-Whitney U statistic and Gehan's W statistic are $U(Y_1^*) = 18653.5$; $U(Y_2^*) = 21366.5$, and $W = -2713$.

Hence, the CPC-I in Figure 6.2 is a "valid" or "good" descriptive statistic for testing the null hypothesis that the times for separation from first marraiges of black and white women have the same distribution functions. And this CPC-I suggests rejection of the null hypothesis.

Figure 6.3. The Pair Chart of the Censoring Variables or Entering Times Between Black and White Women Marrying for the First Time at Age 18, Based on the U.S.-N.F.S., 1970

Y_2^* = White



Y_1^* = Black

Symbol:	1	3
Statis:	$U(Y_1^*)$	$U(Y_2^*)$

6.2 Application of the Generalized Kendall's Tau

Here, we will consider a bivariate vector $\underline{X} = (X_1, X_2)$ where X_1 is the second birth interval and X_2 is the length of breastfeeding after the first birth. Censoring may occur on both variables, X_1 and X_2. Censoring on X_1 may be caused by divorce, separation, widowhood after the first birth, and the time point of survey. Censoring on X_2 occurs if the mothers are still breastfeeding at the time of survey. We may note that termination of breastfeeding is either voluntary or involuntary, caused by child death. In both cases, the length of breastfeeding, X_2, is considered as uncensored and we are interested in comparing measures of association between X_1 and X_2 based on the GKT in Chapter III and other possible indexes, such as the product moment correlation, the rank correlation and Kendall's tau.

For illustration, we take a group of mothers having certain characteristics from the data of SriLanka based on the World Fertility Survey (WFS), 1975. The characteristics of the group are:

(i) 1 or 2 children ever born

(ii) neither husband nor wife sterilized

(iii) ever used contraceptive

(iv) 18 years old at the first birth

Table 6.3 shows a statistical analysis of the corresponding data.

Table 6.3. The Statistical Analysis of the Illustrative Bivariate Data

Statistics	Indicator Variable		
	(0,0)	(1,0)	(1,1)
N	57	17	27
Mean	(38.42, 15.59)	(91.59, 11.53)	(10.11, 10.11)
S.D.	(23.31, 11.18)	(99.34, 11.43)	(7.09, 7.09)
Minimum	(12, 0)	(0, 0)	(2, 2)
Maximum	(123, 48)	(356, 36)	(26, 26)
S.E. of Mean	(3.09, 1.48)	(24.09, 2.77)	(1.36, 1.36)
C.V.	(60.68, 71.78)	(108.46, 99.14)	(70.11, 70.11)

From this table, we can note that the statistics associated with $\underline{\delta}$ = (1,1) have the same values for X_1 and X_2. This indicates that the observed values on X_1 and X_2 are the same. Within this subgroup, all mothers are still breastfeeding at the time of survey, and they have not had a second birth.

Table 6.4 shows some correlation coefficients: (1) the GKT's and the index Alpha associated with the censored data with n = 101; (ii) the product moment correlation (Pearson), the rank correlation (Spearman) and Kendall's Tau-B associated with the complete sub-data with n = 57; and (iii) under the assumption that the whole data

Table 6.4. Correlation Coefficients and Prob.>|r| Under H_0: $\rho = 0$ for the Illustrative Data

Correlation Coefficients	r	P(|p|>r)
Censored Data: N = 101		
Tau-C,1	0.132	0.0000
Tau-C,2	0.098	0.0000(*)
Alpha-1	0.247	0.0000
Alpha-2	0.200	0.0000(*)
Uncensored Sub-Data: N = 57		
Pearson	0.351	0.0074
Spearman	0.306	0.0205
Tau-B (>Tau-A)	0.234	0.0144
Assuming the Data are Complete: N = 101		
Pearson	0.258	0.0092
Spearman	0.404	0.0001
Tau-B	0.328	0.0000
Tau-C,1 (=Tau-A)	0.315	0.0000
Alpha-1 (=Gamma)	0.340	0.0000

(*) Is computed using normal approximation of the U-statistic.

set (N = 101) are complete. This table suggests that the group of mothers under study in Sri Lanka has positive correlation between their second birth interval and the length of their breastfeeding.

Considering the original data having complete and incomplete observations the index Tau-C,1 of 0.132 as well as the index Alpha-1 of 0.247 clearly suggests the rejection of the null hypothesis with p = 0.0000. This probability is computed using normal approximation. The computation of the probabilities $P(|\rho|>r)$ for Tau-C,2 and Alpha-2 are still open, because of the large number of permutations involved, that is 101.

Under the assumption that the data is complete, we are considering the variable $\underline{Z} = (Z_1, Z_2) = \text{Min}(\underline{X}, \underline{Y})$ given in (3.8.1). Table 6.4 shows that Z_1 and Z_2 have a positive correlation $t_{c,1} = t_a = .315$ with p = 0.0000. This method has advantanges if the dependence of Z_1 and Z_2 implies the dependence of X_1 and X_2. Otherwise, this method is not appropriate for testing the null hypothesis that there is association between the second birth interval and the length of breastfeeding.

For comparison, Table 6.4 also shows the correlation coefficients between X_1 and X_2 for the complete sub-sample. The table shows the Pearson correlation coefficient r = 0.351 with p = 0.0074.

Table 6.5 shows the correlation coefficients of bivariate data sets for selected groups at first birth of mothers in Sri Lanka, that is 20, 25 and 30 or above. This table and Table 6.4 show

Table 6.5. Correlation Coefficients and Prob.>|r| Under H_0: $\rho=0$ for Some Selected Age Groups at First Birth of Mothers in Sri Lanka

	Age Groups at First Birth		
	20	25	≥ 30
Censored Data			
Tau-C,1	.165/	.161/	.197/
Tau-C,2	.130/ -	.107/	.251/
Alpha-1	.301/	.313/	.298/
Alpha-2	.261/ -	.248/	.420/
Complete Assumption			
Pearson	.263/.013	.465/.000	.394/.000
Spearman	.474/.000	.438/.000	.459/.000
Tau-B	.373/.000	.356/.000	.351/.000
Tau-C,1 (=Tau-A)	.457/ -	.345/	.338/
Alpha-1 (=Gamma)	.384/	.367/	.364/
Complete Obs.	47	33	32
Incomp. Obs.	41	41	60

that the number of complete observations decreases as the group age increases. Note that the inequalities

$$\tau_{c,1}(X_1, X_2) < \alpha_1(X_1, X_2),$$

$$\tau_{c,2}(X_1, X_2) < \alpha_2(X_1, X_2),$$

$$\tau_{c,1}(X_1, X_2) < \tau_{c,1}(Z_1, Z_2), \text{ and}$$

$$\alpha_1(X_1, X_2) < \alpha_1(Z_1, Z_2).$$

which are easily derived from Chapter III, are satisfied for all age groups.

Considering the value of $\alpha_1(X_1, X_2)$, as well as the Pearson $r(Z_1, Z_2)$, its value increases from age group 18 to 25 and then decreases to age group 30 or over. The maximum value of $\alpha_1(X_1, X_2)$ is .313 and its minimum value is .247 at age group 18. However, $\alpha_1(Z_1, Z_2)$ has maximum value of .384 at age group 20 and minimum value of .364 at age 30 or above.

Finally all correlation coefficients are positive with $P(|\rho|>r) < .10$. So, the second birth interval and the length of breastfeeding for Sri Lanka mothers tend to have a positive correlation.

CHAPTER VII

CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

7.1 Charts

It has been shown that the censored pair chart (CPC), as well as the triplet chart, can be drawn easily for small sample sizes. Hence, the charts should be considered as descriptive statistics, especially whenever the sample sizes are small.

For large sample sizes, we proposed a bar chart as an alternative presentation of a CPC. The bar chart can be constructed using a SAS program. As shown in Figure 2.6 and Figure 6.1-2, this bar chart has a drawback in comparing the two areas corresponding to the statistics $A(X_i)$, $i=1,2$, because the horizontal and vertical axes have different unit-lengths. So, a better computer plot needs to be developed for the construction of the CPC, in which both axes have the same units.

A computer plot of the triplet chart for a large sample size needs to be considered. Based on our study, however, the construction of the triplet chart is not worth doing, whenever the sample sizes are large. And it is suggested to consider the pairwise comparisons using the CPC. Moreover, it is easier to compute the inequality index in (5.3.2), instead.

Based on the chart for equal sample sizes, we proposed a

maximum distance, D, statistic, including its distribution. For further study, we may suggest constructing a table of the critical values of the D-statistic.

7.2 Generalized Kendall's tau

For bivariate data, the null and non-null distributions of the unconditional generalized Kendall's tau (UGKT) have been discussed in detail. For the conditional generalized Kendall's tau (CGKT), however, it is impossible to obtain an explicit general formula for its distribution function due to the variabilities of the pattern of the observations, even though the sample size is small. As the sample size n increases the number of permutations, n!, which leads to a fixed pattern of observations, would increase very rapidly. So, for large n, it seems impossible to take into account the n! permutations for our study. This situation suggests that we consider taking a sufficiently large random sample from the n! permutations for further research on the conditional GKT. This kind of randomization was suggested by Chung and Fraser (1958) for a multivariate two-sample problem. And Boyett and Shuster (1977) consider using randomization for some nonparametric one-sided tests in multivariate analysis.

The extension of these GKT's to censored multivariate data is presented as a vector statistic. However, we only study the null and non-null distribution of the UGKT, which is considered as the generalized Simon's (1977) statistic. The previous paragraph implies that the complexity of the distribution of the CGKT would increase

with the number of component variables. This problem is open for further study.

The relationship between the d.f. of the true variable, $\underline{X}$, and the d.f. of the variable $\underline{Z} = \text{Min}(\underline{X},\underline{Y})$ in (4.4.1), where $\underline{Y}$ is the censoring variable, has been studied, under the null hypothesis of dependence and a sequence of the alternative hypotheses, for general distribution functions. The results are illustrated under normality assumptions. This may be extended to other type of distribution functions.

Considering the vector statistic T_m in (4.2.14) and the non-null distribution of $t_{12...m}$ discussed in sub-section 4.4.2, we may consider for future study the correlation matrix of T_m. Under the total independence hypothesis, Simon (1977) proved that the covariance of any two components of T_m is zero, for uncensored data.

7.3 Sector Symmetry

We have introduced the idea of sector symmetry, and developed a chi-squared statistic and an index of symmetry for a statistical test. However, this test is valid under the assumption that equalities among any number of the m variates, $m \geq 2$, have zero probabilities. So, research on how to treat tied components or variates is still open.

A problem arises, because an observation having tied variates could be classified into more than one particular sector. This would violate a property of categorical data.

Compared with other nonparametric tests, the Wilcoxon and

normal score statistics are usually more efficient than the Kolmogorov-Smirnov (K-S) statistic for location or scale differences or other parametric alternatives. On the other hand, the K-S statistic is valid for general alternatives, and the other rank statistics may not be so. A similar situation occurs in the test of symmetry.

7.4 Applications

It is undoubted that the CPC and the GKT can be used in the study of life testing problems. Examples have been illustrated in the previous chapters, for a clinical trial/experiment and, for demographic data. As noted in Section 7.1, a better program needs to be developed for the CPC for large sample sizes. The availability of such a program would lead to a better or a wider usage of the CPC.

The difficulties in interpreting a path or a curve in 3-dimensional space suggest a limitation on the application of the triplet chart. So, we suggest using the triplet chart only for small sample sizes. In the case of equal sample sizes, the hexagonal projection of the triplet chart seems to give a good picture for making a decision whether the corresponding three present populations tend not to have the same distribution functions.

The application of the generalized Simon's statistic (GSS) is straightforward. But presenting the value of the GSS, that is $t_{12\ldots m}$ for $m>2$, is insufficient for describing the real life situations. A certain value of GSS could be associated with a postive and a negative association of a certain pair of variates. So,

it is suggested to use the vector statistic to present or describe the associations between the m-variates.

APPENDIX

SAS PROGRAM FOR THE CPC-I

```
// EXEC SAS,REGION=250K
//SYSIN DD *
DATA WSTAT;
*************************************************;
*                                                *;
*   NOTE FOR UNEQUAL SAMPLE SIZES:               *;
*   PUT ADDITIONAL OBSERVATIONS HAVING           *;
*   MISSING VALUES                               *;
*                                                *;
*************************************************;
INPUT DX X DY Y;
IF X=. AND Y=. THEN DELETE;
IF X=. THEN DO;
  DX=9; X=0;
  END;
IF Y=. THEN DO;
  DY=9; Y=0;
  END;
IF DX=0 THEN X1=X;
ELSE X1=-1*X;
CARDS;
PROC SORT OUT=WSTAT; BY DX X1;
PROC MATRIX;
FETCH ZDX DATA=WSTAT(KEEP=DX);
FETCH ZX  DATA=WSTAT(KEEP= X);
FETCH ZDY DATA=WSTAT(KEEP=DY);
FETCH ZY  DATA=WSTAT(KEEP= Y);
Z=ZDX||ZX||ZDY||ZY;
N=NROW(Z);
 YLX=J(N,1,0);* THE NUMBER OF Y'S LESS THAN EACH X;
  NY=J(N,1,0);* THE NUMBER OF UNCENSURED OBS. ON Y;
YCGX=J(N,1,0);* THE NUMBER OF Y_CENSURED GREATER THAN EACH X;
YLXC=J(N,1,0);* THE NUMBER OF Y'S LESS THAN EACH X_CENSORED;
   Q=J(N,1,0);* THE NUMBER OF TIMES X=Y;
I=0;
LOOP: I=I+1;
   DA=Z(I,1);  A=Z(I,2);
AA=0; AB=0; AC=0; AD=0; AE=0;
K=0;
LOOPA: K=K+1;
  DB=Z(K,3); B=Z(K,4);
  IF DA=0 THEN DO;
    IF DB=0 THEN DO;
    AB=AB+1;
     IF B<A THEN AA=AA+1;
    IF B=A THEN AE=AE+1;
    END;
      IF DB=1 THEN DO;
      IF B>=A THEN AC=AC+1;
      END;
  END;
  IF DA=1 THEN DO;
    IF DB=0 THEN DO;
    IF A>=B THEN AD=AD+1;
    END;
  END;
```

```
 IF K<N THEN GO TO LOOPA;
  YLX(I,1)=AA;
    NY(I,1)=AB;
 YCGX(I,1)=AC;
 YLYC(I,1)=AD;
    Q(I,1)=AE;
IF I<N THEN GO TO LOOP;
XLY=NY-YLX-Q; * THE NUMBER OF X'S LESS THAN EACH Y;
* START COMPUTING THE U C A AND W STATISTICS;
    UX=SUM(YLX) + .5*SUM(Q);
    UY=SUM(NY) - UX;
    CX=SUM(YLXC);
    CY=SUM(YCGX);
    AX=UX+CX;  AY=UY+CY;
    W=AX-AY; * THE W STATISTIC OF GEHAN;
PRINT UX UY CX CY AX AY W;
I=1:N;  I=I';
V=(I||YLX||J(N,1,1))//(I||Q||J(N,1,2))//(I||XLY||J(N,1,3))//
   (I||YCGX||J(N,1,4))//(I||YLXC||J(N,1,5)) ;
V=V||(Z(,1)//Z(,1)//Z(,1)//Z(,1)//Z(,1));
OUTPUT V OUT=CPC;
DATA PC; SET CPC; IF COL4~=9;
I_TH_X1 = COL1;
J_TH_X2 = COL2;
STAT_ID=COL3;
*****************************************************;
*                                                   *;
*   NOTE FOR LARGE SAMPLE SIZES;                    *;
*   USE ONLY HBAR (HORIZONTAL BAR)                  *;
*   OR CHART WITH VERTICAL X_AXIS                   *;
*                                                   *;
*****************************************************;
PROC CHART;
VBAR  I_TH_X1 / SUBGROUP=STAT_ID SUMVAR=J_TH_X2  DISCRETE NOSPACE;
TITLE THE CPC_I FOR THE DATA OF FREIREICH ET AL.(1963) WITH HORIZONTAL X1_AXIS;
PROC CHART;
HBAR  I_TH_X1 / SUBGROUP=STAT_ID SUMVAR=J_TH_X2  DISCRETE NOSPACE;
TITLE THE CPC_I FOR THE DATA OF FREIREICH ET AL.(1963) WITH VERTICAL X1_AXIS;
//
```

BIBLIOGRAPHY

Barr, Donald R. and Davidson, Teddy (1973). "A Kolmogorov-Smirnov Test for Censored Samples." Technometrics, Vol. 15, #4, 739-757.

Basu, A.P. (1967). "On the Large Sample Properties of a Generalized Wilcoxon-Mann-Whitney Statistic," Ann. Math. Statist. 36, 905-915.

Basu, A.P. (1967). "On Two K-Sample Rank Tests for Censored Data," Ann. Math. Statist. 36, 1520-1535.

Bishop, Y.M.M.; Feinberg, S.E. and Holland, P.W. (1975). Discrete Multivariate Analysis, The MIT Press, Cambridge, Mass.

Blomquist, Nils (1950). "On a Measure of Dependence Between Two Random Variables," Ann. Math. Statist. 21, 593-600.

Boyett, J.M., and Shuster, J.J. (1977). "Nonparametric One-Sided Tests in Multivariate Analysis with Medical Applications," J. Amer. Statist. Assoc. 72, 665-668.

Breslow, N. (1970). "A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship," Biometrika 57, 3, 579-594.

Brown, Jr., B.W.M.; Hollander, M. and Korwar, R.M. (1974). "Nonparametric Tests of Independence for Censored Data, with Applications to Heart Transplant Studies." Reliability and Biometry 327-354.

Chatterjee, S.K. and Sen, P.K. (1973). "Nonparametric Testing Under Progressive Censoring," Calcutta Statist. Assoc. Bull. 22 (1), 13-50.

Chung, J.H., and Fraser, D.A.S. (1958). "Randomization Tests for a Multivariate Two Sample Probme." Ann. Math. Statist. 53, 729-735.

Conover, W.J. (1965). "Several K-Sample Kolmogorov-Smirnov Tests," Ann. Math. Statist. 36, Part I, 1019-1026.

Conover, W.J. (1967). "The Distribution Functions of Tsao's Truncated Smirnov Statistics," Ann. Math. Statist. 38, 1208-1215.

Conover, W.J. (1971). Practical Nonparametric Statistics, John Wiley & Sons, Inc., New York.

Creason, J.P. (1978). "The Theory and Application of a General Iterative Maximum Likelihood Procedure to Randomly Censored Univariate and Bivariate Normal Linear Models." Ph.D. Dissertation, Dept. of Biostatistics, University of North Carolina at Chapel Hill.

Crouse, C.F. and Steffens, F.E. (1969). "A Distribution-Free Two Sample Test for Dispersion for Symmetrical Distribution." South African Statistical Journal 3, 55-67.

Daniels, H.E. (1944). "The Relation Between Measures of Correlation in the Universe of Sample Permutations." Biometrika 33, 129-135.

Daniels, H.E. and Kendall, M.G. (1947). "The Significance of Rank Correlations Where Parental Correlation Exists." Biometrika 34, 197-208.

Daniels, H.E. (1948). "A Property of Rank Correlations." Biometrika 35, 416-417.

Daniels, H.E. (1950). "Rank Correlation and Population Models." J. Roy. Statist. Soc. B 12, 171-181.

Daniels, H.E. (1952). "Note on Durbin and Stuart's Formula for $E(r_s)$, J. Roy. Statist. Soc. B. 13, 310.

David, H.T. (1958). "A Three Sample Kolmogorov-Smirnov Test," Ann. Math. Statist. 29, 842-851.

Davis, C.E. (1978), "A Two-Sample Wilcoxon Test for Progressively Censored Data," Commun. Statist. Theor. Meth., A7(4), 389-398.

Davis, J.A. (1967). "A Partial Coefficient for Goodman and Kruskal's Gamma," J. Amer. Statist. Assoc. 62, 189-193.

Durbin, J. and A. Stuart (1951). "Inversions and Rank Correlation Coefficients." J. Roy. Statist. Soc. B. 13, 303-309.

Flies, J.L. (1973). Statistical Methods for Rates and Proportions, John Wiley & Sons, Inc., New York.

Freireich, E., et al. (1963). "The Effect of 6-mercaptopurine on the Duration of Steroid-Induced Remissions in Acute Leukemia," Blood 21, 699-716.

Gehan, E.A. (1965). "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," Biometrika 52, 203-218.

Gelberg, M.G. (1974). "The Relation Between Mann-Whitney's Statistic and Kendall's Correlation Coefficient Tau," Theory Probability and Applications 19, 205-207.

Goodman, Leo A., and Kruskal, W.H. (1954). "Measures of Association for Cross Classifications," J. Amer. Statist. Assoc. 49, 732-764.

Goodman, Leo A. (1959). "Partial Test for Partial Taus," Biometrika 46, 425-432.

Halperin, Max (1960. "Extension of the Wilcoxon-Mann-Whitney Test to Samples Censored at the Same Fixed Point," J. Amer. Statist. Assoc. Vol. 55, 125-138.

Halperin, Max, and Ware, James (1974). "Early Decisions in a Censored Wilcoxon Two-Sample Test for Accumulating Survival Data," J. Amer. Statist. Assoc. 69, 414-422.

Hodges, Jr., J.L. (1958). "The Significance Probability of the Smirnov Two-Sample Test," Arkiv for Matematik 3, 469-486.

Hoeffding, W. (1948). "A Class of Statistics with Asymptotically Normal Distributions, Ann. Math. Statist. 19, 293-325.

Hoffding, W. (1947). "On the Distribution of the Rank Correlation Coefficient-t When the Variates are not Independent," Biometrika 34, 183-196.

Hotteling, H. and Pabst, M.R. (1936). "Rank Correlation and Tests of Significance Involving no Assumption of Normality," Ann. Math. Statist. 7, 29-43.

Johnson, N.I. and Kotz, S. (1970). Continuous Univariate Distributions - 2. Houghton Mifflin Company, New York.

Johnson, Richard A., and Mehrotra, K.G. (1972). "Locally Most Powerful Rank Tests for the Two-Sample Problem with Censored Data," Ann. Math. Statist. 43, 823-831.

Kendall, M.G. (1938). "A New Measure of Rank Correlation," Biometrika 30, 81-93.

Kendall, M.G. (1942). "Partial Rank Correlation," Biometrika 32, 277-283.

Kendall, M.G. (1949). "Rank and Product-Moment Correlation," Biometrika 36, 177-193.

Kendall, M.G. (1970). Rank Correlation Methods, Griffin, London.

Koziol, James A. and Byar, David P. (1975). "Percentage Points of the Asymptotic Distribution of One and Two Sample K-S Statistics for Truncated or Censored Data," Technometrics, Vol. 17, #4, 507-510.

Kruskal, W.H. (1958). "Ordinal Measure of Association," J. Amer. Statist. Assoc. 53, 814-867.

Mohanty, S.G. (1979). Lattice Path Counting and Applications. Academic Press, New York.

Moran, P.A.P. (1951), "Partial and Multiple Rank Correlation," Biometrika 38, 26-32.

Nelson, W.B. and Hahn, G.J. (1971). "Regression Analysis of Censored Data-Linear Estimation Using Ordered Observations," General Electric Corporate Research and Development TIS Report No. 71-C-122.

Nelson, W. (1972). "Theory and Applications of Hazard Plotting for Censored Failure Data," Technometrics Vol. 14, No. 4, 945-966.

Quade, Dana (1973). "The Pair Chart," Statistica Neerlandica 27 Nr. 1, 29-45.

Quade, Dana (1974). "Nonparametric Partial Correlation." From Measurement in the Social Sciences: Theories and Strategies. Edited by H.M Blalock, Jr., Aldine Publishing Company, Chapter 13, 369-398.

Rao, C.R. (1973). Linear Statistical Inference and its Applications. John Wiley & Sons, Inc., New York.

Rao, U.V.R., Savage, I.R. and Sobel, M. (1960). "Contributions to the Theory of Rank Order Statistics: The Two-Sample Censored Case," Ann. Math. Statist. 31, 415-426.

Ryder, N.B., and Westoff, C.W. (1977). The Contraceptive Revolution, Princeton University Press, Princeton, N.J.

Sen, P.K. (1960). "On Some Convergence Properties of U-Statistics." Calcultta Statistical Association Bulletin, Vol. 10, Nos. 37 & 38, 1-18.

Sen, P.K. (1967). "On Some Nonparametric Generalization of Wilk's Test for H_M, H_{VC}, and H_{MVC}, I." Annals of the Inst. of Statistical Mathematics, Vol. 19, 451-471.

Shirahata, S. (1975). "Locally Most Powerful Rank Tests for Independence with Censored Data." The Annals of Statistics 3, 241-245.

Sibuya, M. (1960). "Bivariate Extreme Statistic, I," Annals of the Institute of Statistical Mathematics, Tokyo, 11, 195-210.

Simon, Gary (1977). "A Nonparametric Test of Total Independence Based on Kendall's Tau." Biometrika 64, 2, 277-82.

Sobel, M. (1966). "On a Generalization of Wilcoxon's Rank Sum Test for Censored Data." Technical Report No. 69 (Revised), University of Minnesota.

Tjostheim, Dag (1978). "A Measure of Association for Spatial Variables," Biometrika 65, 109-114.

Weier, D.R. and Basu, A.P. (1980). "An Investigation of Kendall's τ Modified for Censored Data with Application," Journal of Statistical Planning and Inference 4, 381-390.

Wilk, M.B. and Gnanadesikan, R. (1968). "Probability Plotting Methods for the Analysis of Data," Biometrika 55, 1-17.

"World Fertility Survey: Sri Lanka, 1975, First Report." Department of Census and Statistics, Ministry of Plan Implementation, Sri Lanka.