NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Multivariate non-parametric quality control statistics

By
Kilani Ghoudi

Thesis submitted
to the School of Graduate Studies
in partial fulfillment of the requirements for the
degree of Master of Science in the subject of
Systems Science

at the
University of Ottawa

ISBN 0-315-60117-5

**To my parents and to
the memory of my sister**

Abstract

During the startup phase of a production process while statistics on the product quality are being collected it is useful to establish that the process is under control. Small samples $\{n_i\}_{i=1}^q$ are taken periodically for q periods. We shall assume each measurement is bivariate. A process is under control or on-target if all the observations are deemed to be independent and identically distributed and moreover the distribution of each observation is a product distribution. This would be the case if each coordinate of an observation is a nominal value plus noise. Let F^i represent the empirical distribution function of the i^{th} sample. Let $\bar{F}$ represent the empirical distribution function of all observations. Following Lehman (1951) we propose statistics of the form

$$\sum_{i=1}^{q} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F^i(s,t) - \bar{F}(s)\bar{F}(t))^2 d\bar{F}(s,t) \tag{1}$$

The asymptotics of nonparametric q-sample Cramer-Von Mises statistics were studied in Kiefer (1959). The emphasis there, however, is on the case where $n_i \to \infty$ while q stayed fixed. Here we study the following family of statistics

$$S_q = \sum_{i=1}^{q} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k_q(n, i, F^i(s,t), \bar{F}(s)\bar{F}(t)) n_i dF^i(s,t) \tag{2}$$

in the above quality control situation, where $q \to \infty$ while n_i stays fixed. In the univariate case, the above statistics were studied in McDonald (1989) where asymptotic normality was proved when the process is on-target.

Here we again establish asymptotic normality. For n times the statistic (1) we give expressions for the asymptotic mean and variance under the null hypothesis. Small sample studies show the statistic has good power for detecting an out-of-control situation where the two coordinates of the observations become correlated. The statistic is moderately powerful for detecting a change-point.

Acknowledgement

I wish to express my sincere thanks to Professor D. McDonald for suggesting the topic and for his constant encouragement throughout the development of this thesis.

I also wish to express my gratitude to the Tunisian government and to the Canadian International Development Agency for providing me with the financial support.

Contents

List of Figures

List of Tables

Introduction

Meeting a standard of quality is one of the main objectives of most industries. For this purpose the implementation of a quality control procedure becomes a necessity. This implementation involves at least one of the following situations.

1. testing if past observations were in control,

2. detecting as soon as possible a change in the distribution of the controlled observations.

In this thesis we mainly consider the first situation for bivariate observations. We suppose that small samples $\{n_i\}_{i=1}^q$ are taken periodically over q periods. We consider statistics of the form

$$S_q = \sum_{i=1}^{q} \int\int k_q(n, i, F^i(s,t), \bar{H}(s)\bar{G}(t)) n_i dF^i(s,t)$$

where

- $n = \sum_{i=1}^q n_i$

- F^i is the empirical distribution function of the i^{th} sample

- $\bar{H}$ is the first empirical marginal distribution function of all observations

- $\bar{G}$ is the second empirical marginal distribution function of all observations

These statistics can be classified with the Crámer-Von Mises family of statistics. The asymptotic properties of these statistics were studied by Kiefer (1959). He considered the case where $n_i \to \infty$ while q stayed fixed. McDonald (1989) considered the situation where $q \to \infty$ and n_i stays fixed for univariate observations. The asymptotic of these statistics for bivariate observations in the case where $q \to \infty$ while n_i stays fixed constitutes the main result of this thesis. Note that the same development carries over to a multidimensional situation.

A typical application of the above statistics would be the case of a startup phase in which we do not dispose of the nominal control values nor of sufficient information on the distribution of the measured qualities. Usually we assume that the process is in control during this phase and we use the first measurements to collect statistics on the product qualities. The non-parametric statistics presented here allow us to do retrospective tests on the past measurements to ensure that the process was always in control.

Chapter 1 presents a quick literature review that summarizes different approaches of solving the problems described above.

In Chapter 2 we consider the asymptotic behavior of the above statistics. Two dimensional L-statistics are introduced and following Stigler (1969) a projection argument is used in the proof of their asymptotic normality.

Chapter 3 concentrates on one particular statistic of this family.

$$S_q = \sum_{i=1}^{q} \int \int (F^i(s,t), \bar{H}(s)\bar{G}(t))^2 n_i dF^i(z)$$

This statistic is used to verify results of Chapter 2 and to carry out a simulation study comparing S_q to the Shewart $\bar{x}$ chart. This statistic proves very convenient especially if a correlation between coordinates of the random observations is present.

Chapter 1

Literature review

1.1 Introduction

Two common problems in quality control involve:

- Testing if past observations of a controlled process were in control and if not, estimating when the process went out of control.

- Detecting as soon as possible the change in distribution of controlled observations. This problem is usually referred as disruption or disorder problem.

An extensive literature on both problems exists in various scientific journals. Both problems have been studied under several assumptions and using different methodologies. In this chapter we present a quick review of the most important of these methodologies.

1.2 Posterior detection

This class of methods deal with the first problem mentioned above and can be divided into two major subclasses which we describe in the following sections.

1.2.1 Parametric posterior detection

Darkhovskii (1976) and Darkhovskii and Brodskii (1980,1987) posed the problem in the following way.

Let $X_1, \ldots, X_N$ be a sequence of random variables and it is known that $X_1, \ldots, X_{n_0}$ have known distribution function F_0 and the remaining variables have known distribution function F_1 The problem consists of giving an estimate of the change point n_0 based on the sequence $X_1, \ldots, X_N$.

1.2.2 Nonparametric Posterior detection

Sen and Srivastava (1975) presented a non-parametric method for testing whether the means of each variable in a sequence of independent random variables can be taken to be the same, against alternatives that a shift might have occurred after some point r. The same problem has been studied by Page (1954,1955),Chernoff and Zacks (1964),Kander and Zacks (1966),Battacharya and Johnson (1968) and Gardner (1969). Most of the papers on non-parametric posterior detection were focussed on a shift in the mean of the random variables.

1.3 Fastest detection of a change in distribution

In this section we will consider methods devoted to the detection of a change in distribution. A very wide literature exists on this subject and for more detail the reader may refer to Jandhyala (1985) who presented detailed literature review and a wide bibliography of this subject. As in the previous section we divide the methods into parametric and non-parametric and we start our description by the parametric ones.

1.3.1 Parametric Approach

The parametric approach itself can be divided into two subclasses. The first class uses a Bayesian approach and the second class groups the non-Bayesian approaches to the problem.

Bayesian setting

Using a Bayesian approach the problem was first presented and solved by Shiryayev (1963, 1978). He posed the problem as stated below. Suppose one is able to sequentially observe a series of independent observations $X_1, X_2, \ldots$ whose distributions may change from a known distribution F_0 to a known distribution F_1 at some unknown point in time ν. Formally $X_1, X_2, \ldots$ are independent random variables such that $X_1, X_2, \ldots, X_{\nu-1}$ are independent identically distributed with distribution function F_0 and $X_\nu, X_{\nu+1}, \ldots$, are independent identically distributed with distribution F_1 and where ν is referred as the disorder time and is such that $1 \leq \nu \leq \infty$. He introduced a cost structure in which he assumed that one loses one unit (i.e inspection cost) if one stops before the change (i.e $\tau < \nu$) and one loses $c(\tau - \nu)$ units if one stops after the change (i.e $\tau \geq \nu$).He also supposed that the disorder time ν has a geometric prior distribution. He proved that the optimal stopping rule is of the form $\tau^* = \inf\{n : \pi_n \geq A\}$. Where π_n is the conditional probability that the change occurred given $(X_1, \ldots, X_n)$ and where A is a constant.

Non-Bayesian setting

In this section we group most known non-Bayesian approaches dealing with the disorder problem. The first optimality results in this context were presented by Lorden (1971). His results are based on the restriction that the stopping rule τ must satisfy $\mathbf{E}(\tau/\nu = \infty) \geq B$ (i.e a very low false alarm rate). His criteria of selecting the best stopping rule is $\{ess \sup \mathbf{E}(\tau - \nu + 1)^+/X_1, \ldots, X_{\nu-1})\}$ which represents the speed with which the stopping time detects the change. He proved that a certain class of stopping rules is asymptotically (i.e $B \to \infty$) optimal and that Page's procedure (Page 1954) belongs to this class. A generalization of his results are presented by Moustakides (1986) .

Bojdecki (1979,1984) presented a probability maximizing approach for a somewhat generalized disorder problem. His aim was to find a stopping rule τ^* such that $P\{|\tau^* - \nu| \leq m\} = \sup\{P\{|\tau - \nu| \leq m\}; \tau \in \mathcal{T}\}$ where $\mathcal{T}$ is the set of all possible stopping rules. His generalization of the problem consists of considering the shift to an unknown distribution F_1 that belongs to a finite (or countable) family of distributions.

Pollak (1985) presented a non-Bayesian setting in which he derived a stopping rule τ that minimizes $\mathbf{E}(\tau/\tau \geq \nu)$ subject to the constraint $\mathbf{E}(\tau/\nu = \infty) \geq B$. He proved that this stopping time can be written as the limit of a sequence of Bayesian rules identical to those presented by Shiryayev (1963). He also presented an almost minimax stopping rule.

1.3.2 Non-parametric approach for the disruption problem

This set of methods is also divided in two subclasses; those using a Bayesian technique and those using non-Bayesian techniques.

Methods using a Bayesian technique

In this class we can cite the work of Zacks (1981) who used the same framework and the same cost structure as Shiryayev (1963) to solve a non-parametric version of the disorder problem. He considered the case of a change from an unknown binomial distribution with parameter θ to an unknown binomial distribution with parameter $\varphi > \theta$. He also assumed a prior $h(\theta, \varphi)$. He showed that the optimal stopping rule is of the form:

$$\tau^* = \inf\{n : \pi_n(\mathbf{X}_n) \geq b_n(\mathbf{X}_n)\},$$

where $\pi_n(\mathbf{X}_n)$ has the same definition as π_n above except that it depend on the past observations $\mathbf{X}_n = (X_1, \ldots, X_n)$ and where $b_n(\mathbf{X}_n)$ is a function of n and $\mathbf{X}_n$. Although the case considered is very simple the computation of $b_n(\mathbf{X}_n)$ is extremely difficult if not impossible.

Non-Bayesian methods

Battacharya and Frierson (1981) designed a non-parametric cumulative sums (Cusum) procedure to detect small disorders based on sequential ranks. The main result of their paper deals with the asymptotic behavior of such procedure. McDonald (1988a) presented a nonparametric cumulative sum procedure based also on sequential ranks designed to detect a change in the sampling distribution to a stochastically larger distribution. Hackl and Ledolter (1989) presented a non-parametric new technique using exponentially weighted moving averages(EWMA) on sequential ranks. Their sequential ranks were different than in McDonald (1988a) and Battacharya (1981). They ranked the last observation among the g last observations where g is a fixed parameter for the procedure. Darkhovskii and Brodskii (1988) discussed a non-parametric method for the fastest detection of a change in the mean of a random sequence. They didn't base their analysis on the whole past but they only consider the last N observations (N is called memory size). They established that for a special choice of N their method is asymptotically optimal for a sequence of independent random variables.

1.4 Multivariate case

Literature dealing with the multivariate case is mostly devoted to practical rules such as Cusums and T-charts. Crosier (1988) proposed two different techniques for constructing multivariate Cusum procedures. The first technique consists in reducing each multivariate observation to a scalar and then constructs the cusum on the scalars. The second procedure consists in forming a cusum vector directly from the observations. The same ideas are also discussed by Woodall and Ncube (1985). Hotelling (1950) and Jackson (1959,1979,1957) proposed some multidimensional T-charts to test if past observations are in control.

Chapter 2

Non-parametric approach

2.1 Introduction

In a very recent paper McDonald (1988b) presented a Crámer-Von Mises statistic based on ranks. He used this statistic to detect if a collection of data comes from independent replications of non-homogeneous Poisson processes. In this chapter we define statistics that belong to the same family of statistics for which we prove the asymptotic behavior under the null hypothesis. These statistics can be used to test if a collection of bivariate observations is the outcome of an on-target or in control process. We say that a process is in control if all the observations are deemed to be independent and identically distributed and moreover the distribution of each observation is the product distribution. This would be the case if each coordinate of an observation is a nominal value plus noise.

2.2 Definitions and results

Suppose we dispose of q samples each of size $\{n_i\}_{i=1}^q$. Let $n = \sum_{i=1}^q n_i$ and denote $Z_{ij} = (X_{ij}, Y_{ij})$;$i = 1, 2, \ldots, q$ and $j = 1, \ldots, n_i$ be the jth observation of the ith sample.

Definition 2.1 *We will define the null hypothesis to be the in control case or, in other words, the case where the Z_{ij} are bivariate iid and the coordinates themselves are independent.*

Let:

- $F(x, y)$ be the distribution function of Z.
- $\overline{F}(x, y)$ be the empirical distribution function of Z.
- $F^i(x, y)$ be the empirical distribution function of Z based only on sample i of size n_i.
- $H(x) = F(x, \infty)$ be the marginal distribution function of X.
- $G(y) = F(\infty, y)$ be the marginal distribution function of Y.
- $\overline{H}(x)$ be the empirical marginal distribution function of X.
- $\overline{G}(y)$ be the empirical marginal distribution function of Y
- $\overline{H}^* = \frac{n}{n+1}\bar{H}$, $\overline{G}^* = \frac{n}{n+1}\bar{G}$.

Definition 2.2 *Suppose we dispose of observations Z_{ij}; $i = 1, \ldots, q$ and $j = 1, \ldots, n_i$ and for each observation, say Z we are given the number of points a, b, c and d falling in each quadrant as shown in Figure 2.1. We will call the above information the structure and we will denote it by S. Note that the number of points in the first quadrant (a) includes the point Z itself.*

Let $F_S(x, y)$ be the conditional distribution of Z given S then:

$$F_S(x, y) = \frac{(n!)^2 H(x)^{a+c-1}(1 - H(x))^{b+d} G(y)^{a+d-1}(1 - G(y))^{b+c}}{(a + c - 1)!(a + d - 1)!(b + d)!(b + c)!} \tag{2.1}$$

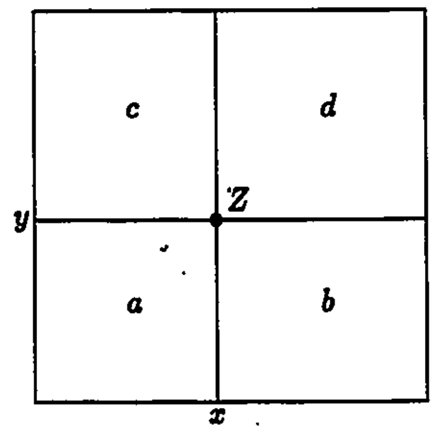

Figure 2.1: Structure $\mathcal{S}$

$$\mathbf{E}(F(x,y)|\mathcal{S}) = \frac{(a+c)(a+d)}{(n+1)^2} = \frac{n^2}{(n+1)^2}\overline{H}(x)\overline{G}(y) \tag{2.2}$$

In the rest of this chapter we will consider the following statistics for which we will study the asymptotic behavior under the in control hypothesis.

$$S_q = \sum_{i=1}^{q} \int k_q(n, i, F^i(x,y), \overline{H}^*(x)\overline{G}^*(y)) n_i dF^i(x,y) \tag{2.3}$$

which can be also written:

$$S_q = \sum_{i=1}^{q} \sum_{j=1}^{n_i} k_q(n, i, F^i(Z_{ij}), \overline{H}^*(X_{ij})\overline{G}^*(Y_{ij})) \tag{2.4}$$

Definition 2.3 *We say that S_q satisfies condition I if the following holds for any integers $\{n, q, i\}$ and any $(y, z) \in [0,1]^2$.*

$$|k_q(n, i, y, z)| \leq k_0 \tag{2.5}$$

$$|\frac{\partial k_q}{\partial y}(n, i, y, z)| \leq k_1, \tag{2.6}$$

$$|\frac{\partial k_q}{\partial z}(n, i, y, z)| \leq k_1, \tag{2.7}$$

and

$$|\frac{\partial^2 k_q}{\partial^2 z}(n,i,y,z)| \leq k_2. \tag{2.8}$$

Theorem 2.4 *Under the null hypothesis, if k_q satisfies condition I and if:*

$$\lim_{q\to\infty} \frac{1}{n} \max_{i\in\{1,\dots,q\}} n_i = 0,\ \limsup_{q\to\infty} \frac{n_q}{q} < \infty \text{ and } \liminf_{n\to\infty} \frac{\mathbf{Var}S_q}{n} > 0$$

Then:

$$\frac{S_q - \mathbf{E}S_q}{\sqrt{\mathbf{Var}S_q}} \Longrightarrow \mathcal{N}(0,1). \tag{2.9}$$

The proof of this theorem requires some preliminary propositions.

Define:

$$T_q = \sum_{i=1}^{q} \int k_q(n,i,F^i(x,y),H(x)G(y))n_i dF^i(x,y) \tag{2.10}$$

which is also equivalent to

$$T_q = \sum_{i=1}^{q}\sum_{j=1}^{n_i} k_q(n,i,F^i(Z_{ij}),F(Z_{ij})) \tag{2.11}$$

The one-dimensional Taylor expansion in the variable z of T_q around $\overline{H}^*(X_{ij})\overline{G}^*(Y_{ij})$ gives

$$T_q = S_q + \Delta_q + L_q \tag{2.12}$$

Where

$$\Delta_q = \sum_{i=1}^{q}\sum_{j=1}^{n_i} \frac{\partial k_q}{\partial z}(n,i,F^i(Z_{ij}),\overline{H}^*(X_{ij})\overline{G}^*(Y_{ij}))[F(Z_{ij}) - \overline{H}^*(X_{ij})\overline{G}^*(Y_{ij})] \tag{2.13}$$

And

$$\begin{aligned} L_q &= \frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{n_i} \frac{\partial^2 k_q}{\partial^2 z}(n,i,F^i(Z_{ij}),\theta H(X_{ij})G(Y_{ij}) + (1-\theta)\overline{H}^*(X_{ij})\overline{G}^*(Y_{ij})) \\ &\quad [F(Z_{ij}) - \overline{H}^*(X_{ij})\overline{G}^*(Y_{ij})]^2 \end{aligned} \tag{2.14}$$

where θ is a stochastic number in $[0,1]$.

Proposition 2.5 *If condition I holds then* $\mathbf{E}|L_q| \equiv O(1)$.

Proof :

$|L_q|$ is bounded by

$$k_2 \sum_{i=1}^{q} \sum_{j=1}^{n_i} [F(Z_{ij}) - \overline{H}^*(X_{ij})\overline{G}^*(Y_{ij})]^2$$

$$= k_2 \sum_{i=1}^{q} \sum_{j=1}^{n_i} [H(X_{ij})(G(Y_{ij}) - \overline{G}(Y_{ij}))$$

$$+ (\frac{n}{n+1})^2 \overline{G}(Y_{ij})(H(X_{ij}) - \overline{H}(X_{ij})) + \overline{G}(Y_{ij})H(X_{ij})(1 - (\frac{n}{n+1})^2)]^2$$

$$\leq k_2 \sum_{i=1}^{q} \sum_{j=1}^{n_i} 4(G(Y_{ij}) - \overline{G}(Y_j))^2 + 4(H(X_{ij}) - \overline{H}(X_{ij}))^2 + 4(\frac{2n+1}{(n+1)^2})^2$$

$$\leq 4k_2 \left[[\sqrt{n} \sup_y |G(y) - \overline{G}(y)|]^2 + [\sqrt{n} \sup_x |H(x) - \overline{H}(x)|]^2 + n(\frac{2n+1}{(n+1)^2})^2 \right]$$

From this and by the Kiefer's theorem (Serfling 1980) it follows that

$$\mathbf{E}|L_q| \leq C < \infty \tag{2.15}$$

Proposition 2.6 *The conditional distribution of* Δ_q *given the structure* S *has mean 0, variance* V_q *and is asymptotically normal. Moreover*

$$nV_q \sum_{i=1}^{q} \sum_{j=1}^{n_i} \sum_{l=1}^{q} \sum_{m=1}^{n_l} \frac{\partial k_q}{\partial z}(n, i, F^i(Z_{ij}), F(Z_{ij})) \frac{\partial k_q}{\partial z}(n, i, F^i(Z_{lm}), F(Z_{lm}))$$

$$\{\overline{H}^*(X_{ij})\overline{H}^*(X_{lm})K(G(Y_{ij}), G(Y_{lm})) + \overline{G}^*(Y_{ij})\overline{G}^*(Y_{lm})$$

$$K(H(X_{ij}), H(X_{lm}))\} = 1 + o(1). \tag{2.16}$$

Where $K(u, v) = \min(u, v) - uv$.

The proof of this proposition will be given after we present some properties of two dimensional L-statistics.

2.3 Two dimensional L-statistics

Later, we show that Δ_q belongs to the family of two dimensional L-statistics which we define in the following definition. The asymptotic normality of this family of statistics constitutes the main objective of this section.

Definition 2.7 *A statistic R_n is called a two dimensional L-statistic if it can be written as follows*

$$R_n = \sum_{i=1}^{n} C_{nr_is_i} U_{nr_i} V_{ns_i}. \tag{2.17}$$

where $\{(U_i, V_i); i = 1, \ldots, n\}$ are independent identically distributed; V_{ns_i} and U_{nr_i} are respectively the s_ith order statistic of V and the r_ith order statistic of U and where the $\{C_{njk} : j = 1, \ldots, n; k = 1, \ldots, n\}$ are constants.

In the following we will consider the case where U_i and V_i are mutually independent uniform (0,1).

Let

$$\hat{R}_n = \sum_{k=1}^{n} \mathbf{E}(R_n|V_k) - (n-1)\mathbf{E}R_n \tag{2.18}$$

and

$$\tilde{R}_n = \sum_{k=1}^{n} \mathbf{E}(R_n|U_k) - (n-1)\mathbf{E}R_n \tag{2.19}$$

Since U_i and V_i are independent we get

$$\mathbf{E}(U_{nr_i} V_{ns_i}|V_k) = \mathbf{E}U_{nr_i}\mathbf{E}(V_{ns_i}|V_k) \tag{2.20}$$

which implies

$$\hat{R}_n = \sum_{i=1}^{n} \hat{C}_{ni}\mathbf{E}(U_{nr_i})\hat{V}_{ns_i} - (n-1)\mathbf{E}R_n \tag{2.21}$$

Where $\hat{C}_{ni} = C_{nr_is_i}$ and where $\hat{V}_{ns_i}$ is given in Stigler (1969) via

$$\hat{V}_{ni} = n^{-1}\sum_{k=1}^{n}\int_0^{V_k} g_{ni}(t)dt \ + n\mathbf{E}V_{n-1,i-1} - (n-1)\mathbf{E}V_{ni} \tag{2.22}$$

where for any $t \in (0,1)$ $g_{ni}(t)$ is given by

$$g_{ni}(t) = n\begin{pmatrix} n-1 \\ i-1 \end{pmatrix} t^{i-1}(1-t)^{n-i} \tag{2.23}$$

The same argument applied to $\tilde{R}_n$ gives

$$\tilde{R}_n = \sum_{i=1}^{n}\hat{C}_{ni}\mathbf{E}(V_{ns_i})\hat{U}_{nr_i} - (n-1)\mathbf{E}R_n \tag{2.24}$$

where

$$\hat{U}_{ni} = n^{-1}\sum_{k=1}^{n}\int_0^{U_k} g_{ni}(t)dt \ + n\mathbf{E}U_{n-1,i-1} - (n-1)\mathbf{E}U_{ni} \tag{2.25}$$

Now let

$$\bar{R}_n = \hat{R}_n - \mathbf{E}\hat{R}_n + \tilde{R}_n - \mathbf{E}\tilde{R}_n \tag{2.26}$$

and let $p_i = \frac{i}{n+1}$

Proposition 2.8 (Stigler 1969 Theorem 1) *For any sequence of integers $\{b_n\}$ such that $b_n/\log(n) \to \infty$ and $b_n/n \to 0$,*

$$n\mathbf{Cov}(\hat{V}_{ni}, \hat{V}_{nj})/K(p_i, p_j) = 1 + o(1) \tag{2.27}$$

and

$$n\mathbf{Cov}(V_{ni}, V_{nj})/K(p_i, p_j) = 1 + o(1) \tag{2.28}$$

uniformly for p_i, p_j in $[b_n/n, 1 - b_n/n]$.

Let $A_n = [b_n/n, 1 - b_n/n]$

Proposition 2.9 *Assume that $\hat{C}_{ni} = 0$ for $\frac{r_i}{n+1} \notin A_n$ or $\frac{s_i}{n+1} \notin A_n$. If for some C independent of n we have*

$$\sum_{i,j=1}^{n} |\hat{C}_{ni}\hat{C}_{nj}|\{\frac{r_i r_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_i s_j}{(n+1)^2}K(p_{r_i}, p_{r_j})\}$$

$$\leq C \sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{r_i r_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_i s_j}{(n+1)^2}K(p_{r_i}, p_{r_j})\} \tag{2.29}$$

and

$$\sum_{i,j=1}^{n} |\hat{C}_{ni}\hat{C}_{nj}|\{K(p_{s_i}, p_{s_j})K(p_{r_i}, p_{r_j})\}$$

$$\leq C \sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{r_i r_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_i s_j}{(n+1)^2}K(p_{r_i}, p_{r_j})\} \tag{2.30}$$

for all n then

$$n\sigma^2(R_n)[\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{r_i r_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_i s_j}{(n+1)^2}K(p_{r_i}, p_{r_j})\}]^{-1}$$

$$= 1 + o(1) \tag{2.31}$$

and

$$n\sigma^2(\bar{R}_n)[\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{r_i r_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_i s_j}{(n+1)^2}K(p_{r_i}, p_{r_j})\}]^{-1}$$

$$= 1 + o(1). \tag{2.32}$$

Proof:

$$\sigma^2(R_n) = \sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\mathbf{Cov}(U_{nr_i}V_{ns_i}, U_{nr_j}V_{ns_j}) \tag{2.33}$$

Using the fact that U and V are independent gives

$$\begin{aligned} n\mathbf{Cov}(U_{nr_i}V_{ns_i}, U_{nr_j}V_{ns_j}) &= n\{\mathbf{E}(U_{nr_i}U_{nr_j})\mathbf{E}(V_{ns_i}V_{ns_j}) - \mathbf{E}U_{nr_i}\mathbf{E}U_{nr_j}\mathbf{E}V_{ns_i}\mathbf{E}V_{ns_j}\} \\ &= n\{\mathbf{Cov}(U_{nr_i}, U_{nr_j})\mathbf{Cov}(V_{ns_i}, V_{ns_j}) \end{aligned}$$

$$+\mathbf{E}U_{nr_i}\mathbf{E}U_{nr_j}\mathbf{Cov}(V_{ns_i}, V_{ns_j}) + \mathbf{E}V_{ns_i}\mathbf{E}V_{ns_j}\mathbf{Cov}(U_{nr_i}, U_{nr_j})\}$$

Immediate use of Proposition 2.8 implies

$$n\mathbf{Cov}(U_{nr_i}, U_{nr_j})\mathbf{Cov}(V_{ns_i}, V_{ns_j})[K(p_{s_i}, p_{s_j})K(p_{r_i}, p_{r_j})]^{-1} = O(\frac{1}{n}) \tag{2.34}$$

Since

$$n\sigma^2(R_n) = n\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\mathbf{Cov}(U_{nr_i}, U_{nr_j})\mathbf{Cov}(V_{ns_i}, V_{ns_j})$$
$$+\mathbf{E}U_{nr_i}\mathbf{E}U_{nr_j}\mathbf{Cov}(V_{ns_i}, V_{ns_j}) + \mathbf{E}V_{ns_i}\mathbf{E}V_{ns_j}\mathbf{Cov}(U_{nr_i}, U_{nr_j})\} \tag{2.35}$$

then

$$|n\sigma^2(R_n)[\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{r_ir_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_is_j}{(n+1)^2}K(p_{r_i}, p_{r_j})\}]^{-1} - 1|$$

$$\leq o(1)\frac{\sum_{i,j=1}^{n} |\hat{C}_{ni}\hat{C}_{nj}|\{K(p_{s_i}, p_{s_j})K(p_{r_i}, p_{r_j}) + \frac{r_ir_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_is_j}{(n+1)^2}K(p_{r_i}, p_{r_j})}{\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{r_ir_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_is_j}{(n+1)^2}K(p_{r_i}, p_{r_j})\}}$$

$$\leq Co(1) \qquad \text{by (2.29) and (2.30).}$$

$$= o(1).$$

$\hat{R}_n$ is $\sigma(V_1, \ldots, V_n)$ measurable and $\tilde{R}_n$ is $\sigma(U_1, \ldots, U_n)$ measurable therefore $\hat{R}_n$ and $\tilde{R}_n$ are independent. It follows that

$$\sigma^2(\bar{R}_n) = \sigma^2(\hat{R}_n) + \sigma^2(\tilde{R}_n) \tag{2.36}$$

which gives

$$n\sigma^2(\overline{R}_n) = \sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}n\{\mathbf{E}U_{nr_i}\mathbf{E}U_{nr_j}\mathbf{Cov}(\hat{V}_{ns_i}, \hat{V}_{ns_j}) + \mathbf{E}V_{ns_i}\mathbf{E}V_{ns_j}\mathbf{Cov}(\hat{U}_{nr_i}, \hat{U}_{nr_j})\}$$

Using the first part of Proposition 2.8 and following the same steps as above gives

$$|n\sigma^2(\bar{R}_n)[\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{r_ir_j}{(n+1)^2}K(p_{s_i}, p_{s_j}) + \frac{s_is_j}{(n+1)^2}K(p_{r_i}, p_{r_j})\}]^{-1} - 1| = o(1). \square$$

Proposition 2.10 *Under conditions of Proposition 2.9*

$$\frac{R_n - \mathbf{E}R_n}{\sqrt{\mathbf{Var}(R_n)}} \Longrightarrow \mathcal{N}(0,1) \tag{2.37}$$

Proof:

By Proposition 2.9 and Proposition 1 (Stigler 1969) we get

$$\frac{\mathbf{E}(R_n - \bar{R}_n)^2}{\sigma^2(\bar{R}_n)} = \frac{\sigma^2(R_n) - \sigma^2(\bar{R}_n)}{\sigma^2(\bar{R}_n)} \longrightarrow 0$$

This implies that $R_n - \mathbf{E}R_n$ and $\bar{R}_n$ are asymptotically equivalent. So in order to prove that $\frac{R_n - \mathbf{E}R_n}{\sqrt{\mathbf{Var}(R_n)}}$ converges weakly to a standard normal it is sufficient to show that $\bar{R}_n/\sqrt{\mathbf{Var}(\bar{R}_n)}$ converges in distribution to a standard normal.

$$\bar{R}_n = \sum_{i=1}^{n} \hat{C}_{ni}\mathbf{E}U_{nr_i}\hat{V}_{ns_i} + \sum_{i=1}^{n} \hat{C}_{ni}\mathbf{E}V_{ns_i}\hat{U}_{nr_i} + \delta_n, \tag{2.38}$$

where δ_n is non-random.

Replacing $\hat{V}_{ni}$ and $\hat{U}_{nr_i}$ by 2.22 and 2.25 gives:

$$\bar{R}_n = n^{-1}\sum_{i=1}^{n} \hat{C}_{ni}\mathbf{E}U_{nr_i}\sum_{k=1}^{n}\int_0^{V_k} g_{ni}(t)dt + \sum_{i=1}^{n} \hat{C}_{ni}\mathbf{E}V_{ns_i}\sum_{k=1}^{n}\int_0^{U_k} g_{nr_i}(t)dt + \delta_n$$

Interchanging the order of summation gives

$$\bar{R}_n = n^{-1}\sum_{k=1}^{n}\sum_{i=1}^{n}\left(\hat{C}_{ni}\mathbf{E}U_{nr_i}\int_0^{V_k} g_{ni}(t)dt + \hat{C}_{ni}\mathbf{E}V_{ns_i}\int_0^{U_k} g_{nr_i}(t)dt\right) + \delta_n$$

Let

$$Z_{nk} = n^{-1}\sum_{i=1}^{n}\left[\hat{C}_{ni}\mathbf{E}U_{nr_i}\int_0^{V_k} g_{ni}(t)dt + \hat{C}_{ni}\mathbf{E}V_{ns_i}\int_0^{U_k} g_{nr_i}(t)dt\right]$$

Then

$$\bar{R}_n = n^{-1}\sum_{k=1}^{n} Z_{nk} + \delta_n$$

Since $\mathbf{E}\bar{R}_n = 0$ then: $\bar{R}_n = n^{-1}\sum_{k=1}^{n}(Z_{nk} - \mathbf{E}Z_{nk})$. To prove the asymptotic normality of $\bar{R}_n$ we only need to check the Lindeberg condition (Billinsley 1985) for the $\sum_{k=1}^{n}(Z_{nk} - \mathbf{E}Z_{nk})$. A straightforward calculation gives

$$\mathbf{E}Z_{nk} = \sum_{i=1}^{n}\hat{C}_{ni}\mathbf{E}U_{nr_i}\int_0^1 (1-t)g_{ni}(t)dt + \hat{C}_{ni}\mathbf{E}V_{ns_i}\int_0^1 (1-t)g_{nr_i}(t)dt$$

and hence

$$Z_{nk} - \mathbf{E}Z_{nk} = \sum_{i=1}^{n} \hat{C}_{ni}\mathbf{E}U_{nr_i} \int_0^1 (t - I_{[V_k<t]})g_{ns_i}(t)dt + \hat{C}_{ni}\mathbf{E}V_{ns_i} \int_0^1 (t - I_{[U_k<t]})g_{nr_i}(t)dt$$

Let $B_n^2 = \mathrm{Var}(\sum_{k=1}^n (Z_{nk} - \mathbf{E}Z_{nk})) = n^2\sigma^2(\bar{R}_n)$.

Following Stigler (1969) it is sufficient to show that $P\{|Z_{nk} - \mathbf{E}Z_{nk}| \geq \varepsilon B_n\} = 0$ for large n to conclude that Lindeberg condition holds for $\sum_{k=1}^n (Z_{nk} - \mathbf{E}Z_{nk})$.

For this purpose let

$$H_{nk} = \sum_{i=1}^{n} \hat{C}_{ni}\mathbf{E}U_{nr_i} \int_0^{V_k} g_{ns_i}(t)dt,$$

$$L_{nk} = \sum_{i=1}^{n} \hat{C}_{ni}\mathbf{E}V_{ns_i} \int_0^{U_k} g_{nr_i}(t)dt,$$

$B_{1n}^2 = \mathrm{Var}(\sum_{k=1}^n (H_{nk} - \mathbf{E}H_{nk}))$ and $B_{2n}^2 = \mathrm{Var}(\sum_{k=1}^n (L_{nk} - \mathbf{E}L_{nk}))$.

We have $B_n^2 = B_{1n}^2 + B_{2n}^2$ and $Z_{nk} - \mathbf{E}Z_{nk} = H_{nk} - \mathbf{E}H_{nk} + L_{nk} - \mathbf{E}L_{nk}$ which implies

$$\begin{aligned} P\{|Z_{nk} - \mathbf{E}Z_{nk}| \geq \varepsilon B_n\} &\leq P\{|H_{nk} - \mathbf{E}H_{nk}| \geq \varepsilon \frac{B_n}{2}\} + P\{|L_{nk} - \mathbf{E}L_{nk}| \geq \varepsilon \frac{B_n}{2}\} \\ &\leq P\{|H_{nk} - \mathbf{E}H_{nk}| \geq \varepsilon \frac{B_{1n}}{2}\} + P\{|L_{nk} - \mathbf{E}L_{nk}| \geq \varepsilon \frac{B_{2n}}{2}\} \end{aligned}$$

The R.H.S of the above inequality is equal to 0 for large n by the same argument as in the proof of Theorem 2 in Stigler (1969) .

In the following proposition we present conditions under which (2.29) and (2.30) hold. This will allow us to give a more general statement of Proposition (2.10).

But first let $D_n = \{\text{integers } i \text{ for which } r_i \text{ or } s_i \text{ is not in } A_n\}$

Proposition 2.11 *If* $|\hat{C}_{ni}| \leq k_1$ *for any integers* n *and* i, *and if* $\liminf_{n\to\infty} \sigma^2(R_n)/n > 0$ *then (2.29) and (2.30) hold.*

Proof:

A straightforward calculation gives

$$\mathbf{Cov}(U_{nr_i}, U_{nr_j}) = \frac{r_i(n - r_j + 1)}{(n+1)^2(n+2)} \qquad \text{for } r_i < r_j \tag{2.39}$$

which implies

$$\mathbf{Cov}(U_{nr_i}, U_{nr_j}) = \frac{K(p_{r_i}, p_{r_j})}{(n+2)}$$

hence

$$\sigma^2(R_n) = \sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{K(p_{r_i}, p_{r_j})K(p_{s_i}, p_{s_j})}{(n+2)^2} + \frac{r_i r_j K(p_{s_i}, p_{s_j})}{(n+1)^2(n+2)} + \frac{s_i s_j K(p_{r_i}, p_{r_j})}{(n+1)^2(n+2)}\}$$

We claim that

$$\lim_{n\to\infty} \frac{\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\{\frac{r_i r_j K(p_{s_i}, p_{s_j}) + s_i s_j K(p_{r_i}, p_{r_j})}{(n+1)^2}\}}{n\sigma^2(R_n)} = 1$$

to.check this claim we note that $\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\frac{K(p_{r_i}, p_{r_j})K(p_{s_i}, p_{s_j})}{(n+2)^2} \le k_1^2$ hence

$$\lim_{n\to\infty} \frac{\sum_{i,j=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\frac{K(p_{r_i}, p_{r_j})K(p_{s_i}, p_{s_j})}{(n+2)^2}}{\sigma^2(R_n)} = 0$$

by hypothesis we get

$$\sum_{i,j=1}^{n} |\hat{C}_{ni}\hat{C}_{nj}|\{\frac{r_i r_j K(p_{s_i}, p_{s_j})}{(n+1)^2} + \frac{s_i s_j K(p_{r_i}, p_{r_j})}{(n+1)^2}\} \le 2k_1^2 n^2$$

Noting the expression of $n\sigma^2(R_n)$ and using the hypothesis $\liminf_{n\to\infty} \sigma^2(R_n)/n > 0$ we conclude that (2.29) holds. Noting also that

$$\sum_{i,j=1}^{n} |\hat{C}_{ni}\hat{C}_{nj}| K(p_{r_i}, p_{r_j})K(p_{s_i}, p_{s_j}) \le k_1^2 n^2$$

the same argument as above implies that (2.30) holds.

Theorem 2.12 *If $|\hat{C}_{ni}| \le k_1$ for any integers n and i and if $\liminf_{n\to\infty} \sigma^2(R_n)/n > 0$ then there exists a sequence b_n satisfying $\lim_{n\to\infty} \frac{b_n}{n} = 0$ and $\lim_{\to\infty} \frac{b_n}{\log n} = \infty$ for which we have*

$$\frac{\mathbf{Cov}(U_{nr_i}V_{ns_i}, U_{nr_j}V_{ns_j})\left(\sum_{i\in D_n}\sum_{j=1}^{n} \hat{C}_{ni}\hat{C}_{nj} + \sum_{j\in D_n}\sum_{i=1}^{n} \hat{C}_{ni}\hat{C}_{nj}\right)}{\sigma^2(R_n)} \to 0$$

and $(R_n - \mathbf{E}R_n)/\sqrt{\mathbf{Var}R_n}$ *is asymptotically a standard normal.*

Proof

$$\mathbf{Cov}(U_{nr_i}V_{ns_i}, U_{nr_j}V_{ns_j}) = \mathbf{Cov}(U_{nr_i}, U_{nr_j})\mathbf{Cov}(V_{ni}, V_{nj})$$

$$+\mathbf{E}U_{nr_i}\mathbf{E}U_{nr_j}\mathbf{Cov}(V_{ni}, V_{nj}) + \mathbf{E}V_{ns_i}\mathbf{E}V_{ns_j}\mathbf{Cov}(U_{nr_i}, U_{nr_j})$$

From (2.39) we get $\mathbf{Cov}(U_{nr_i}, U_{nr_j}) \leq \frac{1}{n+2}$.

Hence

$$\begin{aligned}\sum_{i\in D_n}\sum_{j=1}^{n}\mathbf{Cov}(U_{nr_i}V_{ns_i}, U_{nr_j}V_{ns_j}) &\leq \sum_{r_i=1}^{b_n}\sum_{r_j=1}^{n}\frac{2n+5}{(n+2)^2} + \sum_{r_i=n-b_n}^{n}\sum_{r_j=1}^{n}\frac{2n+5}{(n+2)^2} \\ &\leq \frac{2nb_n(2n+5)}{(n+2)^2} \leq 4b_n\end{aligned}$$

Therefore

$$\sum_{i\in D_n}\sum_{j=1}^{n}\hat{C}_{ni}\hat{C}_{nj}\mathbf{Cov}(U_{nr_i}V_{ns_i}, U_{nr_j}V_{ns_j}) \leq 4k_1^2 b_n$$

Now let us show that $|\sum_{i\in D_n}\hat{C}_{ni}U_{nr_i}V_{ns_i}|$ is asymptotically negligible with respect to $\sigma^2(R_n)$.

$$\mathbf{E}|\sum_{i\in D_n}\hat{C}_{ni}U_{nr_i}V_{ns_i}| \leq 4b_n k_1$$

Choose $b_n = (\log n)^{1+\epsilon}$ for some $\epsilon > 0$. Hence

$$\frac{\mathbf{Cov}(U_{nr_i}V_{ns_i}, U_{nr_j}V_{ns_j})\left(\sum_{i\in D_n}\sum_{j=1}^{n}\hat{C}_{ni}\hat{C}_{nj} + \sum_{j\in D_n}\sum_{i=1}^{n}\hat{C}_{ni}\hat{C}_{nj}\right)}{\sigma^2(R_n)} \to 0$$

and

$$\frac{\mathbf{E}|\sum_{i\in D_n}\hat{C}_{ni}U_{nr_i}V_{ns_i}|}{\sigma(R_n)} \to 0$$

From the above, Propositions (2.11) and (2.10) the asymptotic normality of $(R_n - \mathbf{E}R_n)/\sqrt{\mathbf{Var}R_n}$ follows.

Proof of proposition 2.6

Conditioning on the structure $\mathcal{S}$, Δ_q can be written

$$\Delta_q = \sum_{i=1}^{n} \hat{C}_{ni}[U_{nr_i} V_{ns_i}] \; + \Lambda_q \tag{2.40}$$

where $\hat{C}_{ni} = \frac{\partial k_q}{\partial z}(n, i, F^i(Z_{ij}, \overline{H}^*(X_{ij})\overline{G}^*(Y_{ij}))$, which is a constant for a given structure $\mathcal{S}$. $U_{nr_i} = H(X_{ij})$ is the r_ith order statistic of a uniform (0,1) and $V_{ns_i} = G(Y_{ij})$ is the ith order statistic of a uniform (0,1). Proposition 2.9 gives the conditional asymptotic variance of Δ_q and Theorem 2.12 implies its conditional asymptotic normality.

Proposition 2.13 *$\frac{1}{n}V_q$ is asymptotic to A_q where*

$$A_q = \frac{1}{n^2} \sum_{i=1}^{q} \sum_{j=1}^{n_i} \sum_{l=1}^{q} \sum_{m=1}^{n_l} \frac{\partial k_q}{\partial z}(n, i, F^i(Z_{ij}), F(Z_{ij})) \frac{\partial k_q}{\partial z}(n, i, F^i(Z_{ij}), F(Z_{ij}))$$

$$[H(X_{ij})H(X_{lm})K(G(Y_{ij}), G(Y_{lm})) + G(Y_{ij})G(Y_{lm})K(H(X_{ij}), H(X_{lm}))]$$

Proof:

The asymptotics of $\frac{1}{n}V_q$ are given in Proposition 2.6. Next $\sup_{(x,y)} |\overline{H}^*(x)\overline{G}^*(y) - H(x)G(y)| \longrightarrow 0$ as $q \longrightarrow \infty$ (see proof of Proposition 2.5).

Since $\frac{\partial k_q}{\partial z}(n, i, y, z)$ is uniformly bounded for any integers n and i and any $(y, z) \in [0, 1]^2$

hence

$$|\frac{\partial k_q}{\partial z}(n, i, F^i(Z_{ij}), \overline{H}^*(X_{ij})\overline{G}^*(Y_{ij})) - \frac{\partial k_q}{\partial z}(n, i, F^i(Z_{ij}), F(Z_{ij}))|$$

$$\leq K_2|\overline{H}^*(X_{ij})\overline{G}^*(Y_{ij}) - H(X_{ij})G(Y_{ij})| \to 0 \text{ uniformly as } q \to \infty$$

$$|K(\overline{H}^*(X_{ij}), \overline{H}^*(X_{lm})) - K(H(X_{ij}), H(X_{lm})|$$

$$\leq 2|\overline{H}^*(X_{ij}) - H(X_{ij})| + 2|\overline{H}^*(X_{lm}) - H(X_{lm})| \to 0$$

uniformly for all i, j, l and m as $q \to \infty$. The same holds for G and gives

$$|K(\overline{G}^*(Y_{ij}), \overline{G}^*(Y_{lm})) - K(G(Y_{ij}), G(Y_{lm})| \to 0 \text{ uniformly as } q \to \infty.$$

Using the same argument as above it is easy to see that
$|\overline{H}^*(X_{ij})\overline{H}^*(X_{lm}) - H(X_{ij})H(X_{lm})| \to 0$ uniformly as $q \to \infty$.
This implies that each of the n^2 terms in $\frac{1}{n}V_q - A_q$ is going uniformly to 0 and since we have $\frac{1}{n^2}$ in front of the summation, the proof is complete.□

Proposition 2.14 *If* $\liminf_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{n} > 0$ *then* $\frac{V_q}{\mathbf{Var}\Delta_q} \Longrightarrow 1$ *and* $\frac{\Delta_q}{\sqrt{\mathbf{Var}\Delta_q}} \Longrightarrow \mathcal{N}(0,1)$ *independently of* S_q.

Proof:

Proposition 2.13 shows that A_q is the asymptotic equivalent to $\frac{1}{n}V_q$. Since the Z_{ij} are independent and for different i's $F^i(Z_{ij})$ are independent then in the computation $\mathbf{E}(A_q - \mathbf{E}A_q)^4$ each term involving six or more different rows is equal to 0. Hence $(\mathbf{E}A_q - \mathbf{E}A_q)^4$ contains at most kn^5 non-null bounded terms; where k is a constant independent of n. The $\frac{1}{n^8}$ in front of this expectation implies that $(\mathbf{E}A_q - \mathbf{E}A_q)^4 \leq C/n^3$ where C is a constant independent of n. Hence

$$\sum_{q=1}^{\infty}(\mathbf{E}A_q - \mathbf{E}A_q)^4 \leq \infty$$

Theorem 1.3.5 (Serfling 1980) implies $A_q - \mathbf{E}A_q$ and hence $\frac{1}{n}V_q - \frac{1}{n}\mathbf{E}V_q$ converges *a.s.* to 0. Since $\mathbf{E}V_q = \mathbf{Var}\Delta_q$ and $\liminf_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{n} > 0$ It follows that $\frac{V_q}{\mathbf{Var}\Delta_q}$ converges *a.s.* to 1. From Proposition 2.6 it follows that $\frac{\Delta_q}{\sqrt{\mathbf{Var}\Delta_q}} \Longrightarrow \mathcal{N}(0,1)$ independently of the Structure S. Since S_q is a function of the structure then the limiting distribution of Δ_q is independent of S_q.□

Proposition 2.15 *If k_q satisfies condition I then there exists a universal constant C such that*

$$\mathbf{E}\left[\sum_{j=1}^{n_i} k_q(n,i,F^i(Z_{ij}),F(Z_{ij}))\right]^4 \leq Cn_i^2 \tag{2.41}$$

Proof:

$$\sum_{j=1}^{n_i} k_q(n,i,F^i(Z_{ij}),F(Z_{ij})) = \int_0^1 k_q(n,i,F^i(z),F(z))n_i dF^i(z)$$

Therefore the L.H.S of (2.41) is bounded by

$$C_1 n_i^2 \mathbf{E}\left[\int_0^1 k_q(n,i,F^i(z),F(z))\sqrt{n_i}\, d(F^i(z)-F(z))\right]^4 \tag{2.42}$$

$$+ C_1 n_i^2 \mathbf{E}\left[\sqrt{n_i}\int_0^1 \{k_q(n,i,F^i(z),F(z)) - k_q(n,i,F(z),F(z))\}\, dF(z)\right]^4 \tag{2.43}$$

$$+ C_1 n_i^2 \mathbf{E}\left[\sqrt{n_i}\mathbf{E}\left(\int_0^1 \{k_q(n,i,F(z),F(z)) - k_q(n,i,F^i(z),F(z))\}\, dF^i(z)\right)\right]^4 \tag{2.44}$$

Expression (2.43) is bounded by $C_1 K_1 n_i^2 \mathbf{E}\left\{\sqrt{n_i}\sup_z |F(z) - F(z)|\right\}^4$ (2.45)

By Kiefer's theorem (Serfling 1980) it follows that:

$$\mathbf{E}\left\{\sqrt{n_i}\sup_z |F^i(z) - F(z)|\right\}^4 \leq M < \infty \tag{2.46}$$

and hence (2.43) is bounded by Cn_i^2.

Next expression (2.44) is bounded by

$$\begin{aligned} & C_1 n_i^2 \mathbf{E}\left[\sqrt{n_i}\mathbf{E}\left(\int_0^1 \left\{|\frac{\partial k_q}{\partial y}(n,i,F(z),F(z))||F^i(z)-F(z)|\right\} dF^i(z)\right)\right]^4 \\ \leq\ & C_1 K_1^4 n_i^2 \mathbf{E}\left\{\sqrt{n_i}\sup_z |F^i(z)-F(z)|\right\}^4 \\ \leq\ & Cn_i^2. \text{ by the above} \end{aligned}$$

Also expression (2.42) is bounded by

$$\begin{aligned} & C_1 n_i^2 \mathbf{E}[\int_0^1 \left\{k_q(n,i,F^i(z),F(z)) - k_q(n,i,F(z),F(z))\right\} \\ & \quad \sqrt{n_i}\, d(F^i(z)-F(z))]^4 \end{aligned} \tag{2.47}$$

$$+ C_1 n_i^2 \mathbf{E}\left[\int_0^1 k_q(n,i,F(z),F(z))\sqrt{n_i}\, d(F^i(z)-F(z))\right]^4 \tag{2.48}$$

Expression (2.47) is bounded by

$$
\begin{aligned}
& C_1 n_i^2 \mathbf{E}\left[\int |k_q(n,i,F(z),F(z))| \sqrt{n_i}|F^i(z)-F(z)|\, d(F^i(z)-F(z))\right]^4 \\
\leq\ & 2C_1K_1^4 n_i^2 \mathbf{E}\left\{\sqrt{n_i}\sup_z |F^i(z)-F(z)|\right\}^4 \\
\leq\ & Cn_i^2.
\end{aligned}
$$

After integrating by parts (2.48) is bounded by:

$$
\begin{aligned}
& C_1 n_i^2 \mathbf{E}[\int \left\{\frac{\partial k_q}{\partial y}(n,i,F(z),F(z)) + \frac{\partial k_q}{\partial z}(n,i,F(z),F(z))\right\} \\
& \qquad \sqrt{n_i}(F^i(z)-F(z))\, dF(z)]^4 \\
\leq\ & 2C_1K_1^4 n_i^2 \mathbf{E}\left\{\sqrt{n_i}\sup_z |F^i(z)-F(z)|\right\}^4 \\
\leq\ & Cn_i^2.
\end{aligned}
$$

Adding (2.42), (2.43) and (2.44) implies that expression (2.41) is bounded by Cn_i^2 □

Proposition 2.16 *If*

$$
\liminf_{q\to\infty} \frac{1}{n}(\max_{i=1,\ldots,q} n_i) = 0 \text{ and } \liminf_{q\to\infty} \frac{\mathbf{Var}T_q}{n} > 0
$$

Then

$$
\frac{T_q - \mathbf{E}T_q}{\sqrt{\mathbf{Var}T_q}} \Longrightarrow \mathcal{N}(0,1). \tag{2.49}
$$

Proof:

$$
T_q = \sum_{i=1}^{q}\sum_{j=1}^{n_i} k_q(n,i,F^i(Z_{ij}),F(Z_{ij}))
$$

For different i's the $k_q(n,i,F^i(Z_{ij}),F(Z_{ij}))$ are independent therefore we only need to verify the Liapounov condition for the central limit theorem for $\sum_{j=1}^{n_i} k_q(n,i,F^i(Z_{ij}),F(Z_{ij}))$. Proposition 2.15 implies the Liapounov condition for the fourth moment if

$$\lim_{q\to\infty} \frac{\sum_{i=1}^{q} Cn_i^2}{(\mathbf{Var}T_q)^2} = 0 \tag{2.50}$$

but by hypothesis $\liminf_{q\to\infty} \frac{\mathbf{Var}T_q}{n} > 0$ which implies that (2.50) holds if

$$\frac{1}{n}(\max_{i=1,\dots,q} n_i) \to 0.\square$$

The following lemma will be used in the proof of Theorem 2.4.

Lemma 2.17 (McDonald 1989) *Let $\{\tilde{T}_q, \tilde{S}_q, \Delta_q, \tilde{L}_q\}_{q=1}^{\infty}$ be such that*

- *a) $\tilde{T}_q = \tilde{S}_q + \Delta_q + \tilde{L}_q$ where $\tilde{T}_q, \tilde{S}_q, \Delta_q$ and $\tilde{L}_q$ have expectation 0.*
- *b) Along every weakly convergent subsequence the joint law of $(\tilde{S}_q/\sqrt{\mathbf{Var}\tilde{S}_q}, \Delta/\sqrt{\mathbf{Var}\Delta_q})$ converges weakly to the joint distribution of two independent random variables S, Z where Z is a standard normal (S may, in principle, depend on the subsequence).*
- *c)*

  $$0 < \liminf_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{\mathbf{Var}\tilde{S}_q} \le \limsup_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{\mathbf{Var}\tilde{S}_q} < \infty$$

  and

  $$\limsup_{q\to\infty} \frac{\mathbf{Var}\tilde{L}_q}{\mathbf{Var}\tilde{S}_q} = 0$$
- *d) $\tilde{T}_q/\sqrt{\mathbf{Var}\tilde{T}_q}$ converges weakly to a standard normal,*

then S is a standard normal distribution.

Proof of theorem 2.4

We have $T_q = S_q + \Delta_q + L_q$. Proposition 2.5 gives $\mathbf{E}|L_q| \le O(1)$ and since $\mathbf{Var}S_q \to \infty$ as $q \to \infty$ this implies L_q is asymptotically negligible.

First assume $\liminf_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{n} > 0$ Asymptotically we have: $\frac{\mathbf{Var}T_q}{n} = \frac{\mathbf{Var}(S_q+\Delta_q)}{n} \geq \frac{\mathbf{Var}S_q}{n}$ since Δ_q is asymptotically independent of S_q by Proposition (2.13).

Hence

$$\liminf_{q\to\infty} \frac{\mathbf{Var}T_q}{n} \geq \liminf_{q\to\infty} \frac{\mathbf{Var}S_q}{n} > 0.$$

The same argument as in proposition 2.15 gives $\mathbf{Var}T_q \leq C \sum_{i=1}^{q} n_i$

$$\limsup_{q\to\infty} \frac{\mathbf{Var}T_q}{q} \leq C \lim_{q\to\infty} \frac{n}{q} < \infty$$

Which implies $\limsup_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{n} < \infty$

Hence $\limsup_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{\mathbf{Var}S_q} < \infty$. If $\liminf_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{n} > 0$ then by letting $\tilde{T}_q = T_q - \mathbf{E}T_q, \tilde{S}_q = S_q - \mathbf{E}S_q$ and $\tilde{L}_q = L_q - \mathbf{E}L_q$ in Lemma (2.17) gives the asymptotic normality of S_q

If $\liminf_{q\to\infty} \frac{\mathbf{Var}\Delta_q}{n} = 0$ then there exists a subsequence q_k along which $\frac{\mathbf{Var}\Delta_q}{n}$ is going to 0. Along this subsequence Δ_q is asymptotically negligeable and S_q is asymptotically equivalent to T_q and hence it is asymptotically normal.□

Chapter 3

Applications

3.1 Introduction

In this chapter we present a particular statistic which we will define in the next section and which belongs to the family of statistics studied in chapter 2. Using this statistic we first verify the properties and results developed in the previous chapter, then we carry out simulations of its behavior under some alternative hypotheses.

3.2 Definition

Consider the following statistic:

$$S_q = \sum_{i=1}^{q} \int (F^i(z) - \overline{H}^*(z)\overline{G}^*(z))^2 n_i dF^i(z) \tag{3.1}$$

where the notation of the previous chapter is conserved. The kernel associated with the above statistic is $k_q(n, i, F^i(z), \overline{H}^*(z)\overline{G}^*(z)) = (F^i(z) - \overline{H}^*(z)\overline{G}^*(z))^2$. It is easy to verify that S_q belongs to the family of statistics defined in chapter 2 and that k_q satisfies condition I defined previously.

The following propositions will prove very useful and they give an estimate of the mean and the variance of the above statistic S_q.

Proposition 3.1 *Let T_q (associated with the kernel k_q) be as defined in the previous chapter. Then*

$$\mathbf{E}T_q = \sum_{i=1}^{q} \frac{5n_i + 17}{36n_i} \tag{3.2}$$

and

$$\mathbf{Var}T_q = \sum_{i=1}^{q} \frac{921n_i^3 + 7856n_i^2 + 10945n_i - 4926}{32400n_i^3} \tag{3.3}$$

Proof:

The proof involves tedious calculations which we carried out using a Maple program. Here we present the technique used in the proof and for more detail we refer the reader to the Maple program presented in appendix A.

The technique used borrows the idea of the structure defined in Chapter 2. We define the distribution function for each structure and then we sum over all possible structures. Since the computation of the variance is more complex and more general than that of the expected value we omit the latter.

Note that $T_q = \sum_{i=1}^{q} \sum_{j=1}^{n_i} (F^i(Z_{ij}) - F(Z_{ij}))^2$, hence

$$\mathbf{Var}T_q = \sum_{i=1}^{q} \mathbf{Var}\left[\sum_{j=1}^{n_i} (F^i(Z_{ij}) - F(Z_{ij}))^2\right];$$

this because the F^i's are independent for different i's.

$$\begin{aligned}\mathbf{Var}\left[\sum_{j=1}^{n_i}(F^i(Z_{ij}) - F(Z_{ij}))^2\right] &= \mathbf{E}\left[\sum_{j=1}^{n_i}(F^i(Z_{ij}) - F(Z_{ij}))^2\right]^2 \\ &\quad - \left(\mathbf{E}\left[\sum_{j=1}^{n_i}(F^i(Z_{ij}) - F(Z_{ij}))^2\right]\right)^2\end{aligned}$$

Let us first compute

$\mathbf{E}\left[\sum_{j=1}^{n_i}(F^i(Z_{ij}) - F(Z_{ij}))^2\right]^2 = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} \mathbf{E}[(F^i(Z_{ij}) - F(Z_{ij}))^2 (F^i(Z_{ik}) - F(Z_{ik}))^2]$

First we consider the case $j \neq k$. Figures 3.1a and 3.1b show the two possible configurations of the two points Z_{ij} and Z_{ik} (note that strictly speaking we have

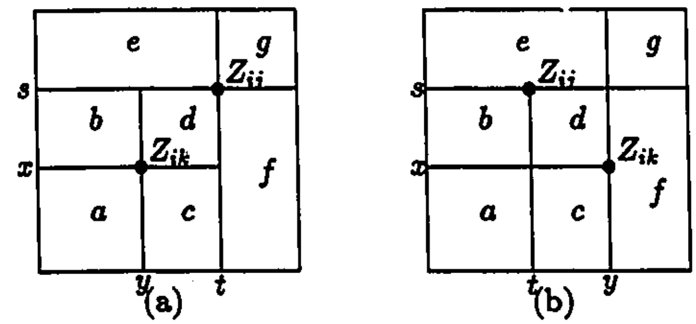

Figure 3.1: Possible configurations

four possible configurations but it is easily seen that these configurations are two by two equivalent).

It can be seen that we can have either one of the two configurations with probability 1/2. Now we restrict our attention to the configuration of Figure 3.1a. In this case the joint density of Z_{ij} and Z_{ik} can be written as

$$f(x,y,s,t,a,b,c,d,m,e,f,g) = \binom{n_i - 2}{m, e, f, g}\binom{m}{a,b,c,d}(xy)^a\,((s-x)y)^b$$

$$((t-y)x)^c\,((s-x)(t-y))^d\,((1-s)t)^e\,((1-t)s)^f\,((1-s)(1-t))^g$$

where $m = a + b + c + d$.

A straightforward calculation gives

$$\mathbf{E}(F^i(Z_{ij}) - F(Z_{ij}))^2(F^i(Z_{ik}) - F(Z_{ik}))^2$$

$$= \frac{1}{n_i^4}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(a+1-n_i xy)^2(m+2-n_i st)^2$$

$$f(x,y,s,t,a,b,c,d,m,e,f,g)\,dx\,dy\,ds\,dt$$

We developed the right hand side into nine terms which we labeled from A to I. Note that for ease of notation, in the following development we denote $f(x,y,s,t,a,b,c,d,m,e,f,g)$ simply by $f(.)$.

$$A = \frac{1}{n_i^4}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(a+1)^2(m+2)^2 f(.)\, dxdydsdt$$

$$B = \frac{-2}{n_i^3}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(xy)(a+1)(m+2)^2 f(.)\, dxdydsdt$$

$$C = \frac{1}{n_i^2}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(xy)^2(m+2)^2 f(.)\, dx\, dy\, ds\, dt$$

$$D = \frac{-2}{n_i^3}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(a+1)^2(st)(m+2) f(.)\, dx\, dy\, ds\, dt$$

$$E = \frac{4}{n_i^2}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(xy)(a+1)(st)(m+2) f(.)\, dx\, dy\, ds\, dt$$

$$F = \frac{-2}{n_i}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(xy)^2(st)(m+2) f(.)\, dx\, dy\, ds\, dt$$

$$G = \frac{1}{n_i^2}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(a+1)^2(st)^2 f(.)\, dx\, dy\, ds\, dt$$

$$H = \frac{-2}{n_i}\int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(xy)(a+1)(st)^2 f(.)\, dx\, dy\, ds\, dt$$

$$I = \int_0^1\int_0^1\int_0^t\int_0^s \sum_{m,e,f,g}\sum_{a,b,c,d}(xy)^2(st)^2 f(.)\, dx\, dy\, ds\, dt$$

We choose the first term A to describe the typical calculations carried out by the Maple program.

First we write $(a+1)^2 = a(a-1)+3a+1$ and replace this in the expression of A. Then we use the multinomial theorem after integrating with respect to x and y, developing any power of m into term of the form $m(m-1)(m-2), \ldots$.then we use the multinomial theorem a second time and integrate with respect to s and t

$$\begin{aligned} A &= \frac{1}{n_i^4}[\frac{(n_i-2)(n_i-3)(n_i-4)(n_i-5)}{324} + \frac{7(n_i-2)(n_i-3)(n_i-4)}{100} \\ &\quad + \frac{289(n_i-2)(n_i-3)}{576} + \frac{47(n_i-2)}{36} + 1] \end{aligned}$$

We use the same technique for $B, C, \ldots, I$. For the second configuration the only things that change are the distribution and the expression of $(F^i(Z_{ij}) - F(Z_{ij}))^2(F^i(Z_{ik}) - F(Z_{ik}))^2$ which become

$$f(x,y,s,t,a,b,c,d,m,e,f,g) = \binom{n_i - 2}{m,e,f,g}\binom{m}{a,b,c,d}(xt)^a((s-x)t)^b$$
$$((y-t)x)^c((s-x)(y-t))^d((1-s)y)^e((1-y)s)^f((1-s)(1-y))^g$$
$$(F^i(Z_{ij}) - F(Z_{ij}))^2(F^i(Z_{ik}) - F(Z_{ik}))^2 = \frac{1}{n_i^4}(a+b+1-n_ist)^2(a+c+1-n_ixy)^2$$

Note that the same technique is also used for the case $j = l$ and for the computation of the expected value of T_q.

Proposition 3.2 *Let A_q be associated with the above kernel k_q defined as in Proposition (2.13) then*

$$\mathbf{E}A_q = 4\sum_{i=1}^{q}\left[\frac{(n_i-1)(25285+12983n_i)}{3175200n_i} + \frac{13n_i}{600} + \frac{47}{1800}\right] + \frac{61q(q-1)}{3200}$$

Proof:

The same technique as in the proof of Proposition (3.1) is used. The only difference is the presence of one more case that corresponds to the situation of Figure 3.2 where we have two different rows i and l.

The joint distribution in this case is just the product and can be written as

$$f(x,y,s,t,a,b,c,d,e,f,g,h) = \binom{n_i-2}{a,b,c,d}(xy)^a((1-x)y)^b((1-y)x)^c$$
$$((1-x)(1-y))^d\binom{n_l-2}{e,f,g,h}(st)^e((1-s)t)^f((1-t)s)^g((1-s)(1-t))^h$$

The rest of the proof uses the same ideas as above and for more informati n see appendix A.

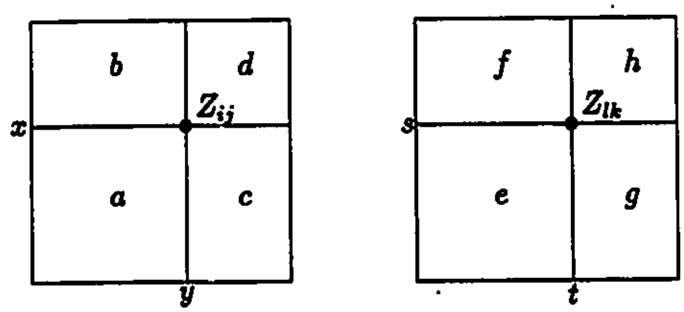

Figure 3.2: Configuration 3 ($i \neq l$)

3.3 Verification of the results of Chapter 2

A simulation study was carried out to verify the results of Chapter 2.

The first set of simulations verifies the asymptotic normality of S_q and is carried out using bivariate independent uniform random numbers generated using the IMSL library random number generator. These simulation results are presented in Figure 3.3 a,b and c. Figure 3.3b represents a simulation result based on 1000 replications of 100 samples (i.e $q = 100$) each of size 5. The histogram represented in figure 3.3b shows the asymptotic normality of S_q. Figures 3.3a and 3.3c were also based on a 1000 simulations. Figure 3.3a corresponds to 30 samples each of size 10 and figure 3.3c corresponds to 200 samples each of size 5. These figures shows that the asymptotic normality of S_q is valid even for a small number of samples ($q = 30$).

All simulation results presented in the second set (Figures 3.4a ,b and c) were based on a 1000 replications. They all use the same number of samples ($q = 100$) and the same sample size ($n_i = 5$ for all $i = 1, \dots, q$) but they use different distributions. Figure 3.4a corresponds to the same case as Figure 3.3b, Figure 3.4b corresponds to bivariate independent normal and Figure 3.4c corresponds to

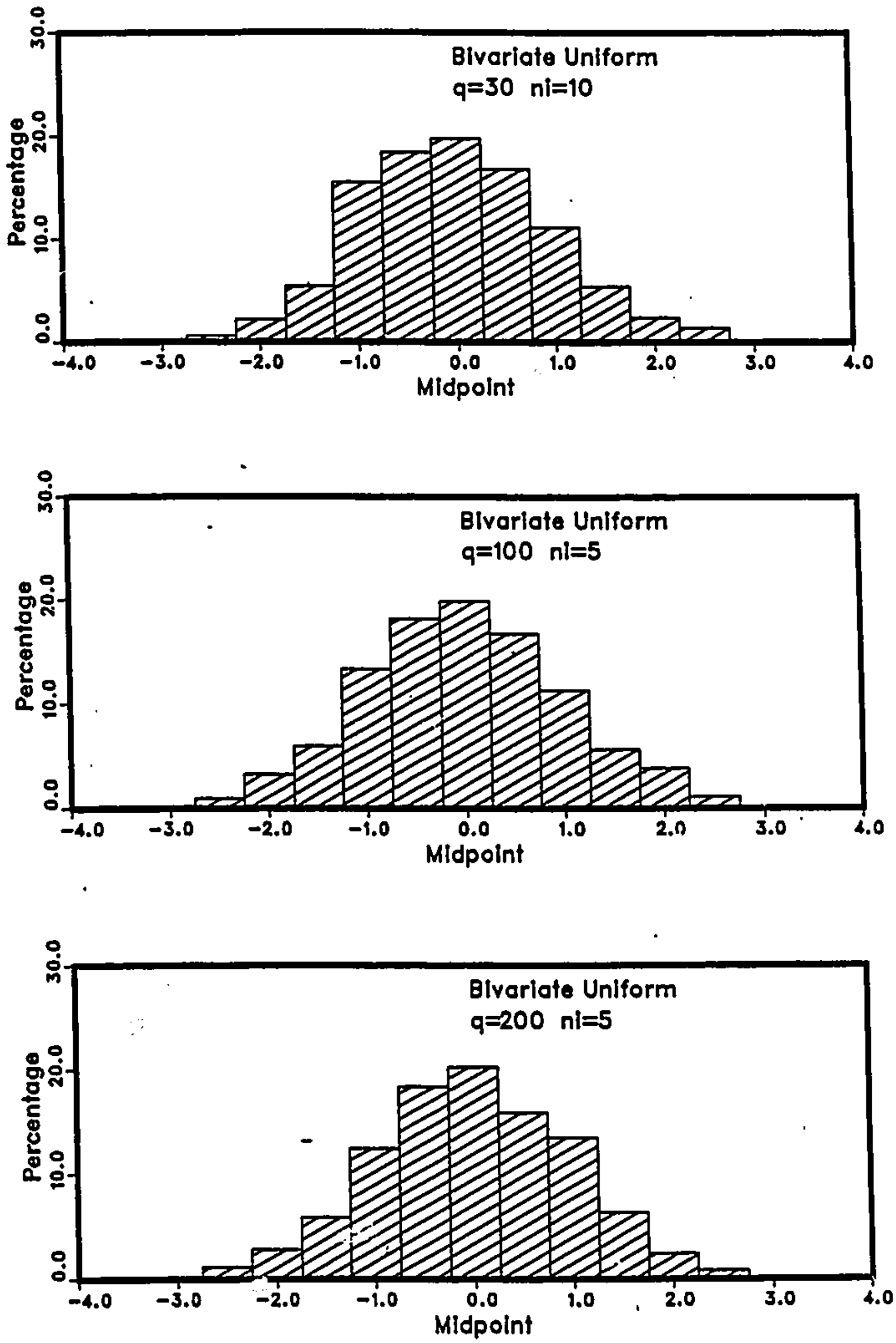

Figure 3.3: Histograms of standardized S_q for different q's and n_i's

bivariate independent exponentials These results emphasize the non-parametric property of S_q. These figures confirm the asymptotic normality and Table 3.1 shows that the expected value and the variance of S_q are independent of the distribution of Z_{ij} when the null hypothesis is satisfied.

Distribution	$\mathbf{E}S_q$	$\mathbf{Var}S_q$
Bivariate uniform	23.2574	1.9115
Bivariate normal	23.2573	1.9117
Bivariate exponential	23.2574	1.9115

Table 3.1: Non-parametric property of S_q

In practice it is often hard to do simulations to get the expected value and the variance of S_q. Propositions 3.1 and 3.2 provide the following estimates of these values.

$$\mathbf{E}S_q \simeq \mathbf{E}T_q \tag{3.4}$$

and

$$\mathbf{Var}S_q \simeq \mathbf{Var}T_q - n\mathbf{E}A_q \tag{3.5}$$

To verify the precision of this estimate, we present estimates (in Table 3.2) of both the variance and the expected value using 1000 simulations and estimates are given using equations 3.4 and 3.5. These results show the precision of the estimates (3.4 and 3.5) even for a small number of samples ($q = 30$). Moreover the following proposition implies that the use of equation 3.4 to estimate $\mathbf{E}S_q$ will only make a test using this estimate more conservative. In other words a

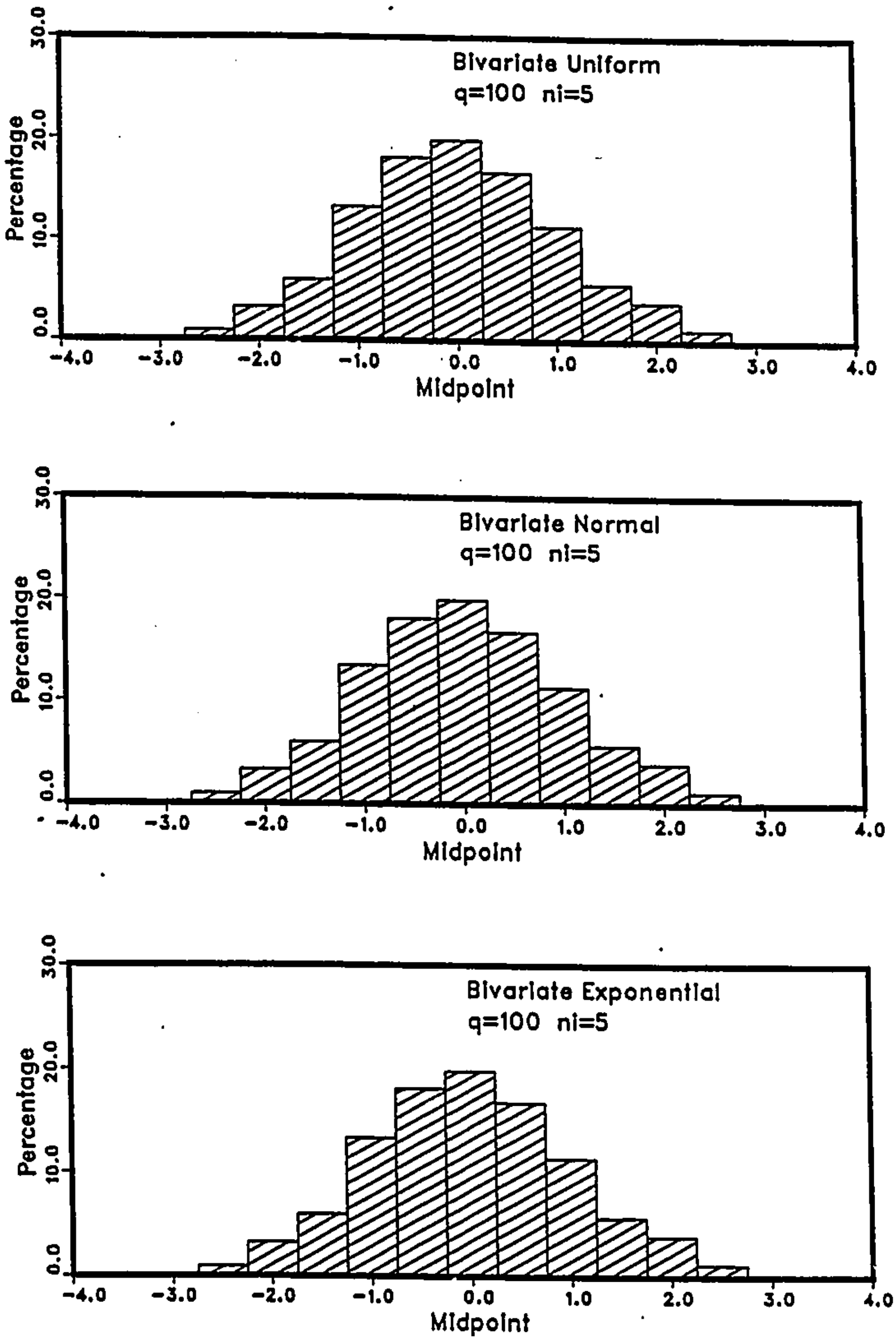

Figure 3.4: Histograms of standardized S_q for different distributions

test using this estimate will erroneously reject the null hypothesis less often then of test using the exact value.

Number of samples	Sample size	Simulated $\mathbf{E}S_q$	Estimated $\mathbf{E}S_q$	Simulated $\mathbf{Var}S_q$	Estimated $\mathbf{Var}S_q$
100	5	23.2574	23.3333	1.9117	1.9678
200	5	46.5792	46.6667	4.0873	3.9388
30	10	5.4588	5.5833	0.4058	0.4018

Table 3.2: Quality of the estimates of $\mathbf{E}S_q$ and $\mathbf{Var}S_q$

Proposition 3.3 *For the kernel k_q defined in the beginning of this chapter, $\mathbf{E}T_q$ is an overestimate of $\mathbf{E}S_q$ (i.e $\mathbf{E}T_q \geq \mathbf{E}S_q$)*

Proof:

From Chapter 2 we know that

$$T_q = S_q + \Delta_q + L_q$$

hence $\mathbf{E}T_q = \mathbf{E}S_q + \mathbf{E}\Delta_q + \mathbf{E}L_q$. However $\mathbf{E}\Delta_q = 0$ and by the definition of k_q we get

$$L_q = \sum_{i=1}^{q}\sum_{j=1}^{n_i} 4(F^i(Z_{ij}) - F(Z_{ij}))^2 \geq 0$$

Therefore $\mathbf{E}L_q \geq 0$ and this completes the proof $\square$.

3.4 Behavior of S_q under some alternatives

In this section we will use the statistic S_q,defined above, and study its behavior under some alternatives that will be defined in later sections. In order to compare this statistics with well known competitors we introduce the Shewart $\bar{x}$ chart which we will discuss in the next section.

3.4.1 Shewart $\overline{x}$ chart

Suppose groups (or samples) of, say, five items are drawn from a process at regular intervals, and measurements of quality are made. The simple Shewart $\overline{x}$ chart plots the mean of each sample of size, say, five against time and if this mean goes over or below certain action limits the process is declared out of control (see Figure 3.5). Usually it assumes that the measured variables have normal distribution under the in control hypothesis. It also assumes that an out of control situation means a change in the mean of the initial normal distribution which without loss of generality is assumed to be a standard normal.

To determine the action limits we use the following technique:
We suppose that from past observations of the process, we dispose of good estimates of both the mean μ and the standard deviation σ. First we compute the standard error of $\overline{x}$, $\sigma/\sqrt{n}$ where n is the sample size (five for example). The second step uses normal tables to define the limits. For example if the process is in control, only one point in a thousand would be above $\mu + 3.09\sigma/\sqrt{n}$, and only one point in a thousand would be below $\mu - 3.09\sigma/\sqrt{n}$. These define the upper and lower limits for the Shewart chart. Therefore if a point falls outside the limit action we declare the process out of control.

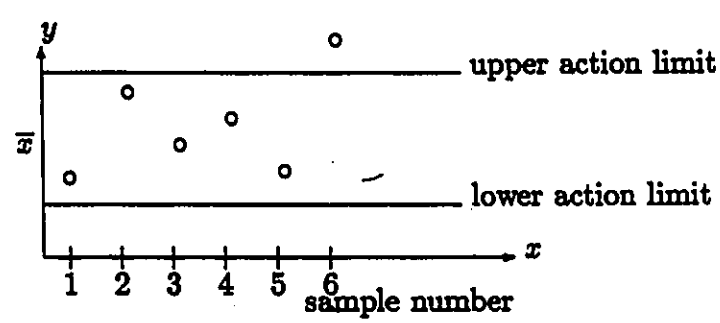

Figure 3.5: Shewart $\overline{x}$ chart

3.4.2 Alternatives

Four major alternatives will be considered. First we consider a shift in one coordinate then we consider a shift in both coordinates, after that we consider a change to correlated random variables and finally we consider a combination of a shift and correlation.

Shift in the mean of one coordinate

It is well documented in the literature that the Shewart $\bar{x}$ chart detects a shift in the mean of normal observations very well. In this section we compare the behavior of the statistic S_q to that of the Shewart chart.

First let us define precisely the alternative hypothesis. Suppose we dispose of q samples (say, $q = 100$) each of size n_i (say, five). We also suppose that each observation is a bivariate normal random variable and that the first, say, 80 samples are from independent *iid* bivariate normal (0,1) distribution and that the rest of the samples are also from independent bivariate normal but the distribution of one of the coordinate, say, X is shifted in mean (*i.e* X has a normal (μ,1) distribution). Table (3.3) summarizes the simulation results obtained for different shifts in the x coordinate. A shift is defined and referred to by its ratio to the standard deviation.

Shift	2 std	1.5 std	1 std	0.5 std
S_q	936	730	347	111
Shewart	1000	1000	826	142

Table 3.3: Shift in one direction

Note:

Note that all the results of this section and all other following sections are based on 1000 replications of a simulation study using a 100 samples ($q = 100$) each of size 5 ($n_i = 5; \forall i$). The numbers presented in this section tables represent the number of times out of a thousand in which the statistic S_q or the Shewart rejects the null hypothesis (in control hypothesis). Note also that the actions limits of the Shewart were fixed so that under the null hypothesis the Shewart and the statistic S_q reject with the same probability (in this case they both reject a portion of 50 out of a 1000).

Shift in two directions

For this alternative we adopt the same conditions as in the preceding section except that both coordinates will shift in mean after, say, 80 samples. Results of simulations related to this alternatives are presented in Table (3.4). These results prove that S_q beats the Shewart chart when the two coordinates are shifted. This can be explained by the fact that the Shewart treats each shift alone, while S_q combines both shifts.

Shift	2 std	1.5 std	1 std	0.5 std
S_q	1000	1000	981	297
Shewart	1000	1000	953	219

Table 3.4: Shift in two directions

Correlation between the two coordinates

We define this alternative as follows.

In the in control case X and Y are independent normal (0,1) random variables.

An out of control situation can induce a correlation between the two coordinates while at the same time conserving the same marginals of X and Y. This alternative is called correlation of the two coordinates. The simulation study done to evaluate the quality of the statistic S_q compared to that of the Shewart chart is summarized in table (3.5). These results prove that for this alternative the statistic S_q has much better power than the Shewart statistic. This conclusion is strengthened by another simulation in which we considered that the two coordinates were correlated right from the beginning. For a correlation of 0.3, the statistic S_q rejected the null hypothesis in 804 cases out of a thousand while the Shewart chart rejected in only 52 cases and for a correlation of 0.5 S_q rejects in 999 cases while the Shewart chart reject in 43 cases.

Cor(x, y)	0.9	0.7	0.5	0.3
S_q	757	431	244	138
Shewart	49	48	50	49

Table 3.5: Correlation of the two coordinates

Small shift and correlation

In this alternative we grouped all the above out-of-control situations. In other words when the change occurs the coordinates may shift in mean and/or may become correlated. Table (3.6) presents some combinations of shifts and correlations that we use to decide on the power of S_q compared to that of the well known Shewart chart. It can be seen from table (3.6) that S_q performs better than the Shewart and it also show that S_q collects the effects of all changes. In other words any change results in an increase in S_q. For example, the change resulting from a shift together with a correlation is greater than the one resulting from a shift alone. The same thing holds for a shift in two directions and a shift in one direction.

$\mathbf{Cor}(x, y)$	0.7	0.5	0.2	0.7	0.5	0.2
Shift	1.0	1.0	1.0	0.5	0.5	0.5
S_q	840	731	500	438	309	164
Shewart	843	832	828	157	148	140

Table 3.6: Shift and correlation

Conclusion

The asymptotic normality of the family of bivariate randomness statistics is derived. For a particular statistic of this family, defined in Chapter 3, we give an expression for the asymptotic mean and variance. This statistic is also used to verify the results of Chapter 2 and then to carry out a simulation study comparing it with the Shewart chart. These simulations show that this statistic has good power if a correlation is present between the two coordinates or if one or more change-points are involved.

A proof of the asymptotics of this family of statistics under different or more general hypotheses would be of great importance since it would allow the computation of the power of these statistics. In the proof of the asymptotic normality of these statistics we introduce two dimensional L-statistics. We only prove their asymptotic normality under the null hypothesis (bivariate uniforms). A general proof, which will take care of different distributions and of different hypotheses such as correlation of the components, would be very useful.

A practical issue, which requires more work, is a comparison with different tests such as Shewart chart based on generalized T-test, non-parametric CUSUM and multivariate CUSUMs.

Appendix A

Proof of Propositions 3.1 and 3.2

In this appendix we present three maple programs that carry the calculation involved in the proof of Propositions 3.1 and 3.2. The first Program computes the expected value of T_q, while the second program computes the variance of T_q. The third program computes the expected values of the estimate A_q.

```
#This maple program computes the expected value of Tq
#It uses the multinomial theorem in the computation of the involved
#integration.

A:=m+1;
s7:=expand(A**2);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s7);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s2);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V1:=s11/n**2;

s7:=expand(A*s*t);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s7);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s2);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V2:=-2*s11/n;

s7:=(s*t)**2;
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s7);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s2);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V3:=s11;


n1:=n-1;
n2:=n1*(n-2);
n3:=n2*(n-3);
n4:=n3*(n-4);
E0:=normal(V1+V2+V3);
E1:=subs(n=n(i),E0);
Exptq:=sum(n(i)*E1,i=1..q);
quit;
```

```
#This program uses the same technique as the first one and it computes
#the variance of Tq. It carry calculation for all different configuration
#presented in chapter 3.

# We start by computing the terms that does not depend on the
# configuration, in other words the variance for a particular i
A:=m+1;
s7:=expand(A**4);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s7);
s9:=subs(m4=nk4*(s*t)**4,m3=nk3*(s*t)**3, m2=nk2*(s*t)**2,m1=nk1*s*t,s2);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V1:=s11/n**4;

s7:=expand(A**3*s*t);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s7);
s9:=subs(m4=nk4*(s*t)**4,m3=nk3*(s*t)**3, m2=nk2*(s*t)**2,m1=nk1*s*t,s2);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V2:=-4*s11/n**3;

s7:=expand((s*t)**2*A**2);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s7);
s9:=subs(m4=nk4*(s*t)**4,m3=nk3*(s*t)**3, m2=nk2*(s*t)**2,m1=nk1*s*t,s2);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V3:=6*s11/n**2;

s7:=expand((s*t)**3*A);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s7);
s9:=subs(m4=nk4*(s*t)**4,m3=nk3*(s*t)**3, m2=nk2*(s*t)**2,m1=nk1*s*t,s2);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V4:=-4*s11/n;

s7:=(s*t)**4;
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s7);
s9:=subs(m4=nk4*(s*t)**4,m3=nk3*(s*t)**3, m2=nk2*(s*t)**2,m1=nk1*s*t,s2);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V5:=s11;

nk1:=n-1;
nk2:=nk1*(n-2);
nk3:=nk2*(n-3);
nk4:=nk3*(n-4);
E2:=normal(V1+V2+V3+V4+V5);

# now we compute covariance term for the first configuration
# Note that we compute each of the nine terms defined in chapter 3 alone

B:=a+1;
A:=m+2;
s1:=expand(B**2);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
```

```
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=A**2*s6;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V1:=s11/n**4;
# Above is the first term (A)
# Below starts the computation of the second term. Note that the program
# is almost repeating itself.
s1:=expand(B);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=A**2*s6*x*y;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V2:=-2*s11/n**3;
# Above is the second term (B)

# and so on...

s1:=1;
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=A**2*s6*(x*y)**2;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V3:=s11/n**2;

s1:=expand(B**2);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
```

```
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=A*s*t*s6;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V4:=-2*s11/n**3;

s1:=expand(B);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=A*s*t*s6*x*y;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V5:=4*s11/n**2;

s1:=1;
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=A*s*t*s6*(x*y)**2;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V6:=-2*s11/n;

s1:=expand(B**2);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=(s*t)**2*s6;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
```

```
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V7:=s11/n**2;

s1:=expand(B);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=(s*t)**2*s6*x*y;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V8:=-2*s11/n;

s1:=1;
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=(s*t)**2*s6*(x*y)**2;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V9:=s11;

V:=normal(V1+V2+V3+V4+V5+V6+V7+V8+V9);
# End of the first configuration and Beginning of the second one.
# Note that the same steps will be repeated starting from the first
# up to the nineth term.
B:=a+b+1;
A:=a+c+1;
s1:=expand(A**2*B**2);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
```

```
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s10:=int(s9,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V1:=s11/n**4;
s1:=expand(A*B**2);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*x*y;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V2:=-2*s11/n**3;
s1:=expand(B**2);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
```

```
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*(x*y)**2;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V3:=s11/n**2;

s1:=expand(A**2*B);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*s*t;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V4:=-2*s11/n**3;
s1:=expand(A*B);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
```

```
   a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
   a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
   a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
   a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
   b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
   a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
   a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
   c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*x*y*s*t;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V5:=4*s11/n**2;
s1:=expand(B);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
   a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
   a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
   b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
   a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
   a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
   a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
   a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
   a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
   b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
   a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
   a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
   c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*(x*y)**2*s*t;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V6:=-2*s11/n;

s1:=expand(A**2);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
   a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
   a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
```

```
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*(s*t)**2;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V7:=s11/n**2;
s1:=expand(A);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*x*y*(s*t)**2;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V8:=-2*s11/n;
s1:=1;
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
```

```
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*(x*y)**2*(s*t)**2;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V9:=s11;
V0:=normal(V1+V2+V3+V4+V5+V6+V7+V8+V9);
# end of the second configuration.

# we put the contribution of each configuration plus the term not
# depending on any configuration and we get the following.

n1:=n-2;
n2:=n1*(n-3);
n3:=n2*(n-4);
n4:=n3*(n-5);
var:=normal(2*V+2*V0);
# the coefficient to is because of the symetric role of Zij and Zik
vi:=normal(n*(n-1)*var+n*E2);
vii:=subs(n=n(i), vi);
vartq:=sum(vii,i=1..q);
quit;
```

```
# This program computes an estimate of the variance of Delta_q
# defined in Chapter 2 It performs all computation involved in the
# proof of proposition 3.2.
printlevel:=-1;
B:=a+1-nl*x*y;
A:=e+1-nl*s*t;
C:=x*s*(t-y*t);
D:=x*s*(y-t*y);
F:=y*t*(s-s*x);
G:=y*t*(x-x*s);
nl:=n(l);
# The following computation are related to the first configuration.
s1:=expand(A*C);
s2:=subs(e=(nl-1)*s*t,s1);
s3:=int(s2,s=0..1);
s4:=int(s3,t=0..y);
s5:=B*s4;
s6:=subs(a=(nl-1)*x*y,s5);
s7:=int(s6,x=0..1);
s8:=int(s7,y=0..1);
V1:=s8;
s1:=expand(A*D);
s2:=subs(e=(nl-1)*s*t,s1);
s3:=int(s2,s=0..1);
s4:=int(s3,t=y..1);
s5:=B*s4;
s6:=subs(a=(nl-1)*x*y,s5);
s7:=int(s6,x=0..1);
s8:=int(s7,y=0..1);
V2:=s8;
s1:=expand(A*F);
s2:=subs(e=(nl-1)*s*t,s1);
s3:=int(s2,s=0..x);
s4:=int(s3,t=0..1);
s5:=B*s4;
s6:=subs(a=(nl-1)*x*y,s5);
s7:=int(s6,x=0..1);
s8:=int(s7,y=0..1);
V3:=s8;
s1:=expand(A*G);
s2:=subs(e=(nl-1)*s*t,s1);
s3:=int(s2,s=x..1);
s4:=int(s3,t=0..1);
s5:=B*s4;
s6:=subs(a=(nl-1)*x*y,s5);
s7:=int(s6,x=0..1);
s8:=int(s7,y=0..1);
V4:=s8;
var0:=normal(V1+V2+V3+V4);
# end of first configuration

# beginning of the second configuration.
B:=a+1-nl*x*y;
A:=m+2-nl*s*t;
C:=x*s*(y-t*y);
```

```
D:=y*t*(x-s*x);
s1:=expand(A*B*C);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=s6;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V1:=s11/ni*(ni-1);

s1:=expand(A*B*D);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s6:=subs(a4=m4*(x*y)**4/(s*t)**4,a3=m3*(x*y)**3/(s*t)**3,
    a2=m2*(x*y)**2/(s*t)**2,a1=m1*x*y/(s*t),s5);
s7:=s6;
s8:=subs(m4=m*(m-1)*(m-2)*(m-3),m3=m*(m-1)*(m-2),m2=m*(m-1),m1=m,s7);
s9:=expand(s8);
s2:=subs(m**4=m4+6*m3+7*m2+m1,m**3=m3+3*m2+m1,m**2=m2+m1,m=m1,s9);
s10:=int(s2,x=0..s);
s11:=int(s10,y=0..t);
s9:=subs(m4=n4*(s*t)**4,m3=n3*(s*t)**3, m2=n2*(s*t)**2,m1=n1*s*t,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V2:=s11/ni*(ni-1);

V:=2*normal(V1+V2);

B:=a+c+1-ni*x*y;
A:=a+b+1-ni*s*t;
C:=x*s*(t-t*y);
D:=y*t*(x-s*x);
s1:=expand(A*B*C);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
```

```
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s10:=int(s9,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V1:=s11/ni*(ni-1);
s1:=expand(A*B*D);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s3:=subs(b**4=b4+6*b3+7*b2+b1,b**3=b3+3*b2+b1,b**2=b2+b1,b=b1,s2);
s4:=subs(c**4=c4+6*c3+7*c2+c1,c**3=c3+3*c2+c1,c**2=c2+c1,c=c1,s3);
s5:=expand(s4);
s6:=subs(a4=m4*(x*t)**4,a3*b1=m4*(x*t)**3*(t*(s-x)),
    a3*c1=m4*(x*t)**3*(x*(y-t)),a2*b2=m4*(x*t)**2*(t*(s-x))**2,
    a2*c2=m4*(x*t)**2*(x*(y-t))**2,b2*c2=m4*(t*(s-x))**2*(x*(y-t))**2,
    b3*c1=m4*(t*(s-x))**3*x*(y-t),b1*c3=m4*(t*(s-x))*(x*(y-t))**3,
    a2*b1*c1=m4*(t*x)**2*(t*(s-x))*x*(y-t),
    a1*b2*c1=m4*(t*x)*(t*(s-x))**2*x*(y-t),
    a1*b1*c2=m4*(t*x)*(t*(s-x))*(x*(y-t))**2,m4=m4/(s*y)**4,s5);
s7:=subs(a3=m3*(x*t)**3,a2*b1=m3*(x*t)**2*(t*(s-x)),
    a2*c1=m3*(x*t)**2*(x*(y-t)),a1*b2=m3*(x*t)*(t*(s-x))**2,
    a1*c2=m3*(x*t)*(x*(y-t))**2,b1*c2 =m3*(t*(s-x))*(x*(y-t))**2,
    b2*c1=m3*(t*(s-x))**2*(x*(y-t)),
    a1*b1*c1=m3*(t*x)*(t*(s-x))*(x*(y-t)),m3=m3/(s*y)**3,s6);
s8:=subs(a2=m2*(x*t)**2,a1*b1=m2*(x*t)*(t*(s-x)),
    a1*c1=m2*(x*t)*(x*(y-t)),b2=m2*(t*(s-x))**2,
    c2 =m2*(x*(y-t))**2,b1*c1=m2*(t*(s-x))*(x*(y-t)),m2=m2/(s*y)**2,s7);
s9:=subs(a1=m1*(x*t),b1=m1*(t*(s-x)), c1=m1*(x*(y-t)),m1=m1/(s*y),s8);
s19:=s9*x*y;
s10:=int(s19,x=0..s);
s11:=int(s10,t=0..y);
s9:=subs(m4=n4*(s*y)**4,m3=n3*(s*y)**3, m2=n2*(s*y)**2,m1=n1*s*y,s11);
s10:=int(s9,s=0..1);
s11:=int(s10,y=0..1);
V2:=s11/ni*(ni-1);
V0:=2*normal(V1+V2);
n1:=ni-2;
n2:=n1*(ni-3);
n3:=n2*(ni-4);
n4:=n3*(ni-5);
var1:=normal(V+V0);
# end of the second configuration

# Third configuration
B:=a+1-ni*s*t;
A:=a+1-ni*s*t;
C:=s*s*(t-t*t);
D:=t*t*(s-s*s);
s1:=expand(A*B*C);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
```

```
s9:=subs(a4=k4*(s*t)**4,a3=k3*(s*t)**3, a2=k2*(s*t)**2,a1=k1*s*t,s5);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V1:=s11;
s1:=expand(A*B*D);
s2:=subs(a**4=a4+6*a3+7*a2+a1,a**3=a3+3*a2+a1,a**2=a2+a1,a=a1,s1);
s5:=expand(s2);
s9:=subs(a4=k4*(s*t)**4,a3=k3*(s*t)**3, a2=k2*(s*t)**2,a1=k1*s*t,s5);
s10:=int(s9,s=0..1);
s11:=int(s10,t=0..1);
V2:=s11;
k1:=ni-1;
k2:=k1*(ni-2);
k3:=k2*(ni-3);
k4:=k3*(ni-1);
printlevel:=1;
var2:=normal(V1+V2);
var1:=var1;
var0:=var0;
expAq:=1/n*(sum(sum(var0,i=1..q),i=1..q)+sum(var1+var2,i=1..q));
quit;
```

Appendix B

Simulation program

In this appendix we give an example of simulation program used in the simulation study presented in Chapter 3.

```
Program simul(input,output,file1,file2);
(* This program simulates the behavior of the statistic Sq of chapter 3
At some point in time (tau) it introduces a change in the distribution
of the observed random variables x and y. This change could be a shift
in one or both directions or it could be a correlation between the two
coordinates. *)

(*   Declaration part *)
const
     q=100;
     n_dim=2000;n_max=15;
type
     vector=array(.1..n_dim.) of real;
     vector1=array(.1..q.) of integer;
     vector0=array(.1..1000.) of shortreal;
     matrix=array(.1..q,1..n_max.) of real;
     information=record
                    a,b,c,d,fl:integer;
                 end;
     matrix1=array(.1..q,1..n_max.) of information;
     matrix2=array(.1..q.) of real;

var
     Sq,Aq,Tq:vector; (* Sq is the statistic and Aq is the variance of
                        Delta_q given the structure *)
     n:vector1;
     var_Sq,var_Tq,exp_tq,var_Dq,exp_Sq,var_Sqe,pi,theta,radius,
                     phi0,sin_phi,cos_phi,var_Vq,exp_Vq,k1,k0,
                     reject_level1,reject_level2,
                     shift,sh_level1,sh_level2:real;
     x,y:matrix;
     x0:vector0;
     dseed:real;
     nr,i,j,m,l,k,n_replication,n_point,tau,sq_reject,sh_reject:integer;
     struct:matrix1;
     x_bar,y_bar:matrix2;
     file1,file2:text;
procedure ggnml(var dseed:real; var nr:integer;var x:vector0);external;
(* the above procedure is the IMSL random number generato, in this case
it generates normal (0,1) random numbers. Some others simulations were
carried using uniform (0,1) and others distribution such as exponential.

 end of the declaration part *)

procedure mean_var(x:vector;nb:integer;var exp_x,var_x:real);
   (* this procedure compute the variance and the expected value of
      the vector x *)
   Var
     i:integer;
   Begin (* of procedure mean_var *)
     exp_x:=0; var_x:=0;
     For i:=1 to nb do
       exp_x:=exp_x+x(.i.);
     exp_x:=exp_x/nb;
     For i:=1 to nb do
```

```
       var_x:=var_x+sqr(x(.i.)-exp_x);
     var_x:=var_x/(nb-1);
  End; (* of procedure mean_var *)

Procedure structure(x,y:matrix;q:integer;n:vector1;var struct:matrix1);
(* In this procedure we constract the structure S defined in chapter 2
*)
     Var
       i,j,l,m:integer;
     Begin (* structure *)
         (* The following procedure constuct the structure associated
            with the observation given in x and y . the result is put
            in the matrix struct. *)

         For i:=1 to q do
             For j:=1 to n(.i.)-1 do
              For m:=j+1 to n(.i.) do
                Begin
                  if x(.i,j.) <= x(.i,m.)
                  then
                    if y(.i,j.) <= y(.i,m.)
                    then
                      begin
                        struct(.i,m.).a:=struct(.i,m.).a+1;
                        struct(.i,m.).fi:=struct(.i,m.).fi+1;
                        struct(.i,j.).d:=struct(.i,j.).d+1;
                      end
                    else
                      begin
                        struct(.i,j.).b:=struct(.i,j.).b+1;
                        struct(.i,m.).c:=struct(.i,m.).c+1;
                      end
                  else
                    if y(.i,m.) <= y(.i,j.)
                    then
                      begin
                        struct(.i,j.).a:=struct(.i,j.).a+1;
                        struct(.i,j.).fi:=struct(.i,j.).fi+1;
                        struct(.i,m.).d:=struct(.i,m.).d+1;
                      end
                    else
                      begin
                        struct(.i,m.).b:=struct(.i,m.).b+1;
                        struct(.i,j.).c:=struct(.i,j.).c+1;
                      end
                end;

         For i:=1 to q-1 do
          For j:=1 to n(.i.) do
           For l:=i+1 to q do
            For m:=1 to n(.l.) do
                Begin
                  if x(.i,j.) <= x(.l,m.)
                  then
                    if y(.i,j.) <= y(.l,m.)
```

```
                    then
                      begin
                        struct(.l,m.).a:=struct(.l,m.).a+1;
                        struct(.i,j.).d:=struct(.i,j.).d+1;
                      end
                    else
                      begin
                        struct(.i,j.).b:=struct(.i,j.).b+1;
                        struct(.l,m.).c:=struct(.l,m.).c+1;
                      end
                  else
                    if y(.l,m.) <= y(.i,j.)
                    then
                      begin
                        struct(.i,j.).a:=struct(.i,j.).a+1;
                        struct(.l,m.).d:=struct(.l,m.).d+1;
                      end
                    else
                      begin
                        struct(.l,m.).b:=struct(.l,m.).b+1;
                        struct(.i,j.).c:=struct(.i,j.).c+1;
                      end;
                end;
   end; (* end of procedure structure *)

Procedure statistic(k:integer; struct:matrix1;q:integer;n:vector1;
                    var Sq,Aq:vector);
(* In this procedure we conpute the statistic Sq using the structure S
defined in previous procedure. *)
var
  i,j,l,m:integer;
  a1,a2,a31,a32,a41,a42:real;
Begin (* of the procedure statistic *)
  Sq(.k.):=0;
  For i:=1 to q do
   begin
     a2:=0;
     For j:=1 to n(.i.) do
     begin
       a1:=(struct(.i,j.).fi/n(.i.)-k0*(struct(.i,j.).a
       +struct(.i,j.).b)*(struct(.i,j.).a+struct(.i,j.).c));
       a2:=a2+sqr(a1);
     end;
     Sq(.k.):=Sq(.k.)+a2;
   end;

End; (* of procedure statistic *)

Procedure variance( var exp_Tq,var_Tq:real);
(* here we compute the variance and the expected value of Tq using
proposition 3.1 *)
 var
   i,j:integer;
 Begin (* variance *)
   var_Tq:=0; exp_Tq:=0;
```

```
  For i:= 1 to q do
   begin
     var_Tq:=var_Tq+(148*n(.i.)*n(.i.)*n(.i.)+1803*n(.i.)*n(.i.)
     +1860*n(.i.)-2463)/(16200*n(.i.)*n(.i.)*n(.i.));
     exp_Tq:=exp_Tq+(5*n(.i.)+17)/(36*n(.i.));
   end;
End; (* of procedure variance *)

begin (* main program *)
    dseed:=2147493647.0;
    readln(file2,n_replication); (* number of simulation *)
    readln(file2,tau);           (* time of the change   *)
    reject_level1:=1.645*sqrt(1.9117)+23.2573;
    reject_level2:=-1.645*sqrt(1.9117)+23.2573;
    sh_level1:=3.65/sqrt(5.0);
    sh_level2:=-3.65/sqrt(5.0);
(* above are defined the rejection levels for both Sq and the Shewart
   chart (sh) *)
    sq_reject:=0;
    sh_reject:=0;
    For i:=1 to q do
      n(.i.):=5;
    n_point:=0;
    For i:=1 to q do
      n_point:=n_point+n(.i.);
    k0:=1/(n_point+1)/(n_point+1);
    k1:=1/(n_point+1);
    pi:=3.141592654;
    phi0:=pi/2;
    readln(file2,cos_phi);   (* value of the correlation *)
    readln(file2,shift);     (* value of the shift       *)
    sin_phi:=sqrt(1-sqr(cos_phi));
    writeln ('q=',q:3,'  n_replica',n_replication:7);
    writeln ('tau=',tau:3,'  Cos(phi)=',cos_phi:7:4,'  Shift =',
    shift:7:4);
    nr:=1000;
For k:=1 to n_replication do
 Begin
    For i:= 1 to q do
     For j:= 1 to n(.i.) do
      begin
      struct(.i,j.).a:=1;
      struct(.i,j.).b:=0;
      struct(.i,j.).c:=0;
      struct(.i,j.).d:=0;
      struct(.i,j.).fi:=1;
      end;
    ggnml(dseed,nr,x0); (* generates  random numbers *)


    For i:=1 to q do
      for j:=1 to n(.i.) do
      begin
         x(.i,j.):=x0(. 5*(i-1)+j.);
         y(.i,j.):=x0(. 500+5*(i-1)+j.);
```

```
      end;
    For i:=tau+1 to q do
      for j:=1 to n(.i.) do
        Begin
          x(.i,j.):=shift+x(.i,j.);
          y(.i,j.):=sin_phi*y(.i,j.)+cos_phi*x(.i,j.);
        end;
    structure(x,y,q,n,struct);
    statistic(k,struct,q,n,Sq,Aq);
    writeln(file1,sq(.k.));
    i:=1;x_bar(.i.):=0;y_bar(.i.):=0;
     for j:=1 to n(.i.) do
     begin
       x_bar(.i.):=x_bar(.i.)+x(.i,j.);
       y_bar(.i.):=y_bar(.i.)+y(.i,j.);
     end;
     x_bar(.i.):=x_bar(.i.)/n(.i.);
     y_bar(.i.):=y_bar(.i.)/n(.i.);
    while((i< q) and (x_bar(.i.)< sh_level1)
    and (x_bar(.i.) > sh_level2) and (y_bar(.i.)<sh_level1)
    and (y_bar(.i.) > sh_level2)) Do
     Begin
     i:=i+1;
     x_bar(.i.):=0;
     y_bar(.i.):=0;
     for j:=1 to n(.i.) do
     begin
       x_bar(.i.):=x_bar(.i.)+x(.i,j.);
       y_bar(.i.):=y_bar(.i.)+y(.i,j.);
     end;
     x_bar(.i.):=x_bar(.i.)/n(.i.);
     y_bar(.i.):=y_bar(.i.)/n(.i.);
    end;
    (* tests for rejection *)
    If(i < q) then sh_reject:=sh_reject+1
              else if((x_bar(.i.) >=sh_level1) or
                 (x_bar(.i.) <= sh_level2) or (y_bar(.i.) >= sh_level1)
             or (y_bar(.i.) <= sh_level2)) then sh_reject:=sh_reject+1;
    If((sq(.k.) >= reject_level1))
     then  sq_reject:=sq_reject+1;
 end;
     mean_var(sq,n_replication,exp_sq,var_sq);
     writeln ('reject_level1=',reject_level1:9:4);
     writeln (file1,'exp_Sq=',exp_Sq:9:4);
     writeln (file1,'Var_Sq=',var_Sq:9:4);
     writeln ('reject_level2=',reject_level2:9:4);
     writeln ('sh_level1=',sh_level1:9:4);
     writeln ('sh_level2=',sh_level2:9:4);
     writeln ('sh_reject=',sh_reject:7);
     writeln ('sq_reject=',sq_reject:7);
end. (* end of the main program *)
```

Bibliography

[1] P.K Battacharya and D. Frierson. A nonparametric chart for detecting small disorders. *Annals of Statistics*, 9:544–554, 1981.

[2] P.K Battacharya and R.A Johnson. A nonparametric control for shift at an unknown time point. *Annals of Mathematics and Statistics*, 39(5):1731–1743, 1968.

[3] Patrick Billingsley. *Probability and Measure*. Jhon Wiley, New York, 1985.

[4] T. Bojdecki. Probability maximizing approach to optimal stopping and its application to a disorder problem. *Stochastics*, 3:61–71, 1979.

[5] T. Bojdecki and J. Hosza. On a generalized disorder problem. *Stochastics*, 18:349–359, 1984.

[6] H. Chernoff and S. Zacks. Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematics and Statistics*, 37:1196–1210, 1966.

[7] Ronald B. Crosier. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30(3):291–303, 1988.

[8] B.S. Darkhovskii. A nonparametric method for the posteriori detection of the disorder time of a sequence of independent random variables. *Theory of probability and its application*, 21:178–183, 1976.

[9] B.S. Darkhovskii and B.E.Brodskii. A nonparametric method for fastest detection of a change in the mean of a random sequence. *Theory of probability and applications*, 32(4):640–648, 1987.

[10] B.S. Darkhovskii and B.E. Brodskii. A posteriori detection of the disorder time of a random sequence. *Theory of probability and its application*, 25:624–628, 1980.

[11] L.A Gardner. On detecting changes in the mean of normal variates. *Annals of Mathematics and Statistics*, 40:116–126, 1969.

[12] Peter Hackl and Johannes Ledolter. *A new nonparemetric quality control technique*. Technical Report 160, Department of Statistics and Actuarial Science University of Iowa, March 1989.

[13] Harold Hotelling. A generalized t test and measure of multivariate dispersion. *Proceeding of the second Berkeley synposium in mathematical statistics and probability*, 23–41, 1950.

[14] J.Edward Jackson. Quality control methods for several related variables. *Technometrics*, 1(4):359–377, 1959.

[15] J.Edward Jackson and R.H. Morris. An application of multivariate quality control to photographic processing. *Journal of the American Statistical Association*, 52:186–199, 1957.

[16] J.Edward Jackson and Govind S. Mudholkar. Control procedure for residuals associated with principal component analysis. *Technometrics*, 21(3):341–349, 1979.

[17] V.K. Jandhyala. *Residual Processes For Regression Models with applications to detection of parameter change at unknown times*. PhD thesis, University of Western Ontario., London, Ontario Canada, October 1985.

[18] Z. Kander and S. Zacks. Test procedures for possible changes in parameters of statistical distributions occuring at unknown time points. *Annals of Mathematics and Statistics*, 35:999–1018, 1964.

[19] J. Kiefer. K-sample analogues of the kolmogorov-smirnov and cramer-von mises tests. *Annals of Mathematics and Statistics*, 30:420–447, 1959.

[20] E.L Lehman. Consistency an unbiasedness of certain nonparametric tests. *Annals of Mathematics and Statistics*, 22:165–179, 1951.

[21] G. Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematics and Statistics*, 42:1897–1908, 1971.

[22] D. McDonald. A cusum procedure based on sequential ranks. 1988a. To be published.

[23] D. McDonald. On non-homogeneous, spatial poisson processes. 1988b. To be published.

[24] D. McDonald. On the asymptotics of randomness statistics. 1989. To be published.

[25] George Moustakides. On optimal stopping times for changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, 1986.

[26] E.S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

[27] E.S. Page. A test for change in a parameter occuring at unknown time point. *Biometrika*, 42:523–526, 1955.

[28] Moshe Pollak. Optimal detection of a change in distribution. *The Annals of Statistics*, 13(1):206–227, 1985.

[29] A. Sen and M. Srivastava. On tests for detecting change in mean. *The Annals of Statistics*, 3(1):98–108, 1975.

[30] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics.* Jhon Wiley, New York, 1980.

[31] A.N. Shiryayev. On optimum methods in quickest detection problems. *Theory of probability and its applications*, 8:22–46, 1963.

[32] A.N. Shiryayev. *Optimal stopping rules.* Springer Verlag, New York, Heidelberg, Berlin, 1978.

[33] Stephen Mack Stigler. Linear functions of order statistics. *The Annals of Mathematical Statistics*, 40(3):770–788, 1969.

[34] William H. Woodall and Matoteng M. Ncube. Multivariate cusum quality control procedures. *Technometrics*, 27(3):285–292, 1985.

[35] S. Zacks and Z. Barzily. Bayes procedures for detecting a shift in the probability of success in a series of bernoulli trials. *Journal of statistical planning and inference.* 5:107–119, 1981.