A NONPARAMETRIC TEST BASED ON RUNS FOR A SINGLE SAMPLE LOCATION PROBLEM

Dissertation
zur Erlangung des akademischen Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Statistik
der Universität Konstanz

vorgelegt von
Milton Januario Rueda Varon

Tag der mündlichen Prüfung: 22.04.2010
Referent: Prof. Dr. Siegfried Heiler
Referent: Prof. Dr. Jan Beran

Acknowledgements

I would like to thank my tutor Prof. Dr. Siegfried Heiler for his unending support, extremely constructive feedback, excellent supervision, and all his encouragement during all stages of this work. I thank Prof. Dr. Jan Beran for supporting this dissertation as second supervisor. I am grateful to Prof. Dr. Jimmy Corzo for his comments and suggestions. I would also like to thank Dr. Marc Handlery for his help during the writing of this dissertation.

I am very grateful to my wife Norma Celis, who was an immense source of strength and motivation, for her support, for her love, for her understanding and for keeping me focussed and determined in my purpose. I am truly thankful for the trust she deposited in me, for her understanding every time difficulties arose and for never allowing me to consider stepping back and not finishing my PhD degree.

I am grateful to the Deutscher Akademischer Austausch Dienst (DAAD) for supporting the work on this Dissertation.

Finally, I would also like to thank Toni Stocker and Susanne Schneider for their moral support and encouragement during the last years.

Contents

List of Tables

List of Figures

Summary

The Runsstatistic C belongs to the the well known group of nonparametric methods which require no assumptions about the population probability distributions. As nonparametric methods make fewer assumptions, their applicability is much wider than the corresponding parametric methods. In particular, they may be applied in situations where less is known about the application in question. Many statistical methods require assumptions to be made about the format of the data to be analyzed. For example, the t-test requires that the distribution of the variable be Normal. When the normality assumption is questionable a nonparametric test should be applied. Also, due to the reliance on fewer assumptions, nonparametric methods are more robust. Specifically, nonparametric methods were developed to be used in cases when the researcher knows nothing about the parameters of the variable of interest in the population. In more technical terms, nonparametric methods do not rely on the estimation of parameters (such as the mean or the standard deviation) describing the distribution of the variable of interest in the population. In this dissertation, a statistic based on runs is designed to test a hypothesis about the location (median) of a population.

The Runsstatistic C is a good alternative for a single sample location problem. The main point of this work is the determination of the distribution of the statistic C under the null hypothesis.

The distribution of the statistic under the null hypothesis is determined and converges to the normal distribution.

The test is compared with traditional nonparametric tests. The results are very promising. The power function of the test C is as good as the power function of the other tests and sometimes even better.

Zusammenfassung

Der Runsstatistik C gehört zu den sog. nichtparametrischen Verfahren. Der größte Vorteil solcher Methoden liegt darin, dass diese auch Stichproben aus nicht-normalverteilten Grundgesamtheiten vergleichen können. Obwohl es Tests gibt, die auf Abweichungen von der Normalverteilung reagieren, ist es kaum möglich, statistisch nachzuweisen, dass eine Stichprobe tatsächlich einer normalverteilten Grundgesamtheit entspringt. Viele der häufig verwendeten Testverfahren, wie z.B. der t-Test, nehmen normalverteilte Grundgesamtheiten an. Wenn diese Annahme erfüllt ist, sind diese Tests trennschärfer als die entsprechenden nichtparametrischen Verfahren. Wenn aber keine normalverteilte Grundgesamtheit vorliegt, sollten diese Tests nicht verwendet werden. Ihre Trennschärfe (oder Power), d.h. ihre Fähigkeit, die Nullhypothese abzulehnen, wenn die Alternativhypothese tatsächlich zutrifft, ist dann nur gering. In solchen Situationen sollte man dann entweder zuerst "normalisierende" Transformationen durchführen (z. B. logarithmische Transformation) oder aber zu nichtparametrischen Verfahren greifen. Bei Abweichungen von der Normalverteilung sind auf jeden Fall nichtparametrische Tests trennschärfer.
In der vorliegenden Dissertation wurde eine Teststatistik, die sich auf Runs basiert, für das Einstichproben Lageproblem dargestellt.

Die Runsstatistik stellt nämlich eine gute Alternative für das Einstichproben-Lageproblem dar. Für die formale Behandlung in einem mathematisch-statistischen Modell existiert zugleich noch keine Lösung. Der Schwerpunkt liegt auf der Bestimmung der Verteilung der Statistik C unter der Nullhypothese.

Die Verteilung der Statistik wird bestimmt und es wird gezeigt, dass sie gegen die Normalverteilung konvergiert.

Der Test wird mit traditionellen nichtparametrischen Tests verglichen. Die Ergebnisse sind sehr verheißungsvoll. Die Gütefunktion des Runstests ist genauso gut wie die Gütefunktion der anderen Tests und gelegentlich sogar besser.

Chapter 1
Introduction

Nonparametric[1] statistics are one of the most important branches of statistics and are widely used in many areas of science, engineering, economics and medicine. The importance of nonparametric statistical methods lies in the fact that only very few assumptions are made about the underlying population from which the data are collected. This is in contradistinction to many classical statistical methods which usually assume that the underlying populations are normal. Nonparametric tests may be, and often are, more powerful in detecting population differences when certain assumptions are not satisfied.

For the one-sample situation, the prime concern in research is examining a measure of central tendency (location) for the population of interest. The best-known measures of location are the mean and median. The median has three advantages relative to the mean. First, when the distribution is skewed, the median is less sensitive to outliers than the mean. The second is that median always exists and, finally, the median can be used even when the data are measured on an ordinal scale. When the appropriate assumptions are satisfied, standard parametric and nonparametric tests can be used to test for the mean and median. The t-test assumes normality (which of course implies symmetry) while the Wilcoxon signed rank test assumes just symmetry. The sign test for example can be used to test for the median of asymmetric data. There are few techniques available for testing when certain assumptions about the underlying population are questionable.

Runs are important in applied probability and statistical inference. They are used in many areas, such as hypothesis testing, system reliability, quality control, data mining and genetics. There have been various publications dealing with the distribution theory of runs. Early discussions for runs appeared in the works of Mood (1940), Levene and Wolfowitz (1944), Wolfowitz (1944), and Dobrushin

[1]The term nonparametric was first used by Wolfowitz, 1942

(1953). New results on runs have been derived by many authors including Ortiz (1983), Ortiz and Corzo (1983), Fernandez and Ortiz (1986), Philippou and Makri (1986), Corzo (1989), Corzo (1990), Fu and Koutras (1994), Koutras and Alexandrou (1995), Han and Aki (1999). Recent investigations are due to Stefanov (2000), Chadjiconstantinidis and Koutras (2001), Kong (2001), Fu and Lou (2003), Balakrisnan and Kountras (2002), Kong(2006).

Various test procedures have been proposed for testing the null hypothesis that the median of a distribution is equal to a specific value θ_0. Some combinations of the Sign test and the Wilcoxon signed rank test were proposed in the literature for obtaining a reasonable power while maintaining the nominal significance level. A test, which combines the sign and the signed rank test, has been proposed by Lemmer (1987) and was shown to work very well, except when the user has no idea whether the distribution is skewed to the left or to the right. Lemmer (1993) proposed an adaptive procedure for this case, which determines whether the sign or signed rank test should be used after calculating a measure of skewness. Baklizi (2005) used the P-value from the triples test to obtain modified Wilcoxon scores and developed an adaptive rank test. Bandyopadhyay and Dutta (2007) purposed two adaptive test procedures for testing $H_0 : \theta = 0$. The first adaptive procedure has a probabilistic approach which uses the P-value from the triples test for symmetry given in Randles et al. (1980) and the second adaptive test has a deterministic approach like a the statistics used by Baklizi (2005).

In this dissertation a statistic based on runs is designed to test a hypothesis about the location (median) of a population and has the only one requirement: that the scale of measurement[2] should be ordinal, interval or ratio. This is the only restriction. This makes the statistic C as widely applicable as possible.

The statistic C has different applications in many fields of applied statistics. One application is in high-fidelity or escape models in data mining. Another application from C is a paired comparisons, where a test based on C is used to determine whether there is a significant difference between the values of the same measurement made under two different conditions. Both measurements are made on each unit in a sample, and the test is based on the paired differences between these two values. The usual null hypothesis is that the difference in the mean values is zero. In this dissertation, these applications will not be analyzed, but practical exercises have been conducted with great success. The different applications of C offer different opportunities for research.

[2]The statistic C could be applied too to nominal variables if they are ordered in some sense

The outline of this dissertation is a follows. In Chapter 2 we introduce the basic ideas and the various statistics used for a single sample location problem. This chapter presents a brief comparison between these statistics with their respective advantages and disadvantages. Chapter 3 presents the statistics proposed and some of its properties. In Chapter 4, the distribution of the statistic is determined using ordinary generating functions. Based on this distribution, the probability function under the null hypothesis and the corresponding power function of C are calculated and presented in chapters 5 and 6 respectively. Furthermore, the critical values for sample sizes between 4 and 30 are shown in the appendix as well as algorithms that facilitate the calculations. Chapter 7 contains the approximation of the distribution of the statistic C to the normal distribution. Chapter 8 proposes suggestions for further research.

Chapter 2
The One Sample Problem

In order to use the well known parametric t-test, the data must either have been sampled from a population that is normally distributed, or the sample size must be sufficiently large so that asymptotic normality of the sample mean can be assumed. If these assumptions cannot be made, then non parametric procedures such as the sign test, or Wilcoxon's signed rank test should be employed. A nonparametric procedure is specifically designed where only very general characteristics of the relevant populations are postulated or hypothesized, for example, that a distribution is symmetric about some specified point. These nonparametric tests have less restrictive assumptions about the shape of the parent population than the t-test. Wilcoxon's signed rank test assumes that the sample is drawn from a continuous, symmetric population while the sign test only requires that the population is continuous around the vicinity of the median. Nonparametric tests are also usually easier to apply and understand than the corresponding parametric tests and are generally insensitive to outliers. Some combinations of these tests are proposed in the literature for obtaining a reasonable power while maintaining the nominal significance level. This chapter presents the tests previously named and some of the procedures mostly used in this situation.

Let $X_1, ..., X_n$ be a random sample with a common continuous cumulative distribution function $F(X - \theta)$, where θ (unspecified) is the median and by definition $P[X_i < \theta] = P[X_i > \theta] = 1/2$. Hence without loss of generality we may set $\theta = 0$. The Hypothesis of interest here is:

$$H_0 : \theta = 0 \qquad versus \qquad H_1 : \theta > 0. \tag{2.1}$$

The best known test for this problem when F is normally distributed is the t-test. In this case θ coincides with μ, the mean of the distribution, and the Hypothesis can be formulated as: $H_0 : \mu \leq 0$ $v.s.$ $H_1 : \mu > 0$, where H_0 is rejected when,

$$T = \sqrt{n}(\overline{X} - \mu)/\sigma > Z_{(1-\alpha)}, \tag{2.2}$$

where $Z_{(1-\alpha)}$ is the upper $(1-\alpha^{th})$ percentile for the normal distribution[1] with parameters μ and σ^2. If σ^2 is unknown, t-distribution is used with $n-1$ d.f.

If the underlying distribution is normal with mean μ and variance σ^2, then the power function of the test based on T becomes

$$P_w(\theta/\sigma) = 1 - \Phi^*(t(\alpha, n-1|n-1)|n-1, \sqrt{n}\theta/\sigma)), \tag{2.3}$$

where $\Phi^*(x|r,c)$ is the c.d.f. of a noncentral t-distribution with r degrees of freedom and noncentrality parameter c.

When the normality assumption is questionable a nonparametric test should be applied. Two of the well known nonparametric tests are presented below.

2.1 The Sign Test

This test is one of the oldest statistical procedures and one of the most widely used because of its simplicity and easy implementation. The sign test is an alternative that can be applied when distributional assumptions are suspect. However, it is not as powerful as the t-test when the distributional assumptions are in fact valid (see chapter 6).

Let

$$S = \sum_{i=1}^{n} s\left(X_i\right), \tag{2.4}$$

where $s(X_i) = 1$ if $x > 0$ and 0 otherwise.

The rule is to reject $H_0 : \theta = 0$ in favor of $H_1 : \theta > 0$ if $S \geq k$. The critical value k is determined so that $P_{H_0}\left(S \geq k\right) = \alpha$.

Under $H_0 : \theta = 0$, $s(X_1), ..., s(X_n)$ are i.i.d. Binomial with parameters l and $p = P(X > 0) = 1/2$. Accordingly, the statistic S is the sum of n binomial random variables and has a binomial distribution with parameters n and $p = 1/2$. Then, the critical value can be found through this distribution. We could easily generate tables to apply the exact sign test for any sample size n. However, we know that the normal approximation to the binomial is especially good when $p = 1/2$. Therefore, the normal approximation to the binomial can be used to determine the rejection regions. This convergence in distribution may be denoted by

[1] σ^2 known

$$\frac{S - E(S)}{\sqrt{Var(S)}} \quad \xrightarrow{d} \quad Z \sim N(0,1), \tag{2.5}$$

where $E(S) = n/2$ and $Var(S) = n/4$.

Under $H_1 : \theta > 0$, S still has a binomial distribution but now $p = P(X > 0)$ which depends on F, and the binomial distribution can be well approximated by the normal distribution. We can derive expressions to approximate the power of the sign test based on the normal approximation. The power for this alternative can be evaluated with a continuity correction as

$$\begin{aligned} P_w[\theta_1] &= Pr[K \geq k_\alpha | H_1 : \theta_1 > \theta_0] && (2.6) \\ &= 1 - \Phi\left(\frac{k_\alpha - n\theta - 0.5}{\sqrt{n\theta(1-\theta))}}\right), && (2.7) \end{aligned}$$

where $\theta = Pr[X > \theta_1 | \theta_1 > \theta_0]$ and k_α is such that

$$\alpha = Pr[K \geq k_\alpha | H_0] = 1 - \Phi\left(\frac{2k_\alpha - n - 1}{\sqrt{n}}\right). \tag{2.8}$$

The equation [2.8] implies that $k_\alpha = [n + 1 + \sqrt{n}\Phi^{-1}(1-\alpha)]/2$.

Substituting k_α into [2.7] and simplifying gives

$$P_w[\theta_1] = 1 - \Phi\left[\frac{n(0.5 - \theta) + 0.5\sqrt{nz_\alpha}}{\sqrt{n\theta(1-\theta))}}\right], \tag{2.9}$$

where $\Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$th quantile of the standard normal distribution.

The Sign test is not very powerful on small samples. This is because the test uses only information about the element positions relative to the assumed median: to the left or to the right. The test does not use information about their values.

2.2 The Wilcoxon Signed Rank Test

The other nonparametric procedure to be covered in this chapter is the Wilcoxon signed rank test. This test is based on a special case of what are called rank order statistics and uses only information in the sign of the observations. No metric information on how far the observation is from zero is incorporated into the test.

The magnitude of any observation is used only in determining its relative position in the sample array.

Let $X_1, ..., X_n$ a random sample from $F(X-\theta)$, $F \in \Omega_s$[2]. The Wilcoxon signed rank test is based on the statistic

$$W = \sum_{i=1}^{n} iW_i = \sum_{i=1}^{n} R_i^+ s(X_i), \tag{2.10}$$

where $W_i = 1$ if $|X|_{(i)}$ corresponds to a positive observation and 0 otherwise, R_i^+ is the rank of $\left|X_i\right|$ from $\left|X\right|_{(1)} < ... < \left|X\right|_{(n)}$ the ordered absolute values and $s(X_i)$ is calculated as in the sign test.

The rule is to reject $H_0 : \theta = 0$ in favor of $H_1 : \theta > 0$ if $W = \sum_{i=1}^{n} R_i^+ s(X_i) \geq w$, where the critical value w is determined from the distribution of W under H_0 such as $P_{H_0}(W \geq w) \leq \alpha$.

Under the null hypothesis $H_0 : \theta = 0$, $W_1, \ldots, W_n$ are independent, identically distributed Binomial random variables with parameters $n = 1$ and $p = Pr[W_i = 0] = Pr[W_i = 1] = 1/2$. Hence $W = \sum_{i=1}^{n} iW_i$ is a linear combination of these variables, its mean and variance can be determined by

$$E(W|H_0) = \frac{n(n+1)}{4} \tag{2.11}$$

and

$$Var(W|H_0) = \frac{n(n+1)(2n+1)}{24}, \tag{2.12}$$

and from a generalization of the central limit theorem, the asymptotic distribution of W is the normal distribution.

Calculating the power of the signed rank test, even using the normal approximation, requires a considerable amount of work (see Hettmansperger (1984)). For a fixed alternative, the power is approximated by

$$P_w[W \geq w] \doteq 1 - \Phi\left(\frac{w - E(W)}{\sqrt{Var(W)}}\right), \tag{2.13}$$

where

[2] $\Omega_s = \left\{F : F \in \Omega_0 \, and \, F(x) = 1 - F(-x)\right\}$, the subclass of Ω_s, of symmetric distributions centered at 0.

$$E(W) = np_1 + \frac{n(n-1)}{2}p_2, \tag{2.14}$$

$$Var(W) = np_1(1-p_1) + \frac{n(n-1)}{2}p_2(1-p_2), \tag{2.15}$$

$p_1 = Pr[X_1 > 0]$ and $p_2 = Pr[X_1 + X_2 > 0]$.

If F is a $N(0, \sigma^2)$ distribution, then it is easy to see that

$$P_w[W \geq w] = 1 - \Phi\left[Z_\alpha - \frac{\left(\frac{n(n-1)}{2} + \frac{n}{\sqrt{2}}\right)}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\frac{\theta}{\sigma\sqrt{\pi}}\right], \tag{2.16}$$

where Z_α is the upper α percentile of the standard normal distribution.

The Wilcoxon Signed Rank test does not require the assumption that the population is normally distributed. Unfortunately, the scope of this test is limited to distributions which are symmetric relative to the median. With non-symmetric distributions the test does not work correctly.

These tests are two of the most important nonparametric tests, however each one presents limitations in its application. In the next chapter, a test based on runs is proposed and provides a good alternative to the mentioned limitations.

2.3 Adaptive Tests

Various adaptive test procedures have been suggested in the literature. These tests are normally based on intuitive grounds and simple calculations, and they originated in principle from earlier descriptive statistics. They are mainly based on a preliminary test or measure of asymmetry, and then choosing between the sign or the Wilcoxon signed rank tests accordingly. The idea is to improve the power of the sign test using the rank test. The procedure is not complicated, first identify the degree of symmetry and then according to this indicator, choose the test to use. This section presents three procedures used in this context.

2.3.1 Adaptive Test for the Median (Lemmer 1993)

Lemmer (1987) has proposed a test, which combines the sign and the Wilcoxon signed rank test. It has been shown to work well, except when the user has no idea

whether the distribution is skewed to the left or to the right. Lemmer (1993) proposed an adaptive procedure for this case, which determines whether the sign or the Wilcoxon signed rank test must be used after calculating a measure of skewness.

Lemmer uses, as measure of symmetry, the statistic given by

$$Q_3 = \frac{\overline{U}_\gamma - \overline{M}_{0.5}}{\overline{M}_{0.5} - \overline{L}_\gamma}, \tag{2.17}$$

where $\overline{U}_\gamma$, $\overline{M}_\gamma$ and $\overline{L}_\gamma$ denotes the mean of the γn (largest, middle, smallest, respectively) combined order statistics (Randles and Wolfe (1979), p 389).

Then, the first adaptive test statistic is given by

$$A = WI_{(Q_3 \notin J)} + SI_{(Q_3 \in J)}, \tag{2.18}$$

where J is an interval to be specified, S and W are the sign and the Wilcoxon signed rank statistics respectively and $I_{(x)}$ the well-known indicator function.

The second adaptive test statistic is given by

$$R = WI_{(R^* \leq r_0)} + SI_{(R^* > r_0)}, \tag{2.19}$$

where R^*=number of runs in the $\{S_i\}$ sequence. R^* can also be expressed as

$$R^* = 1 + I_2 + \cdots + I_n, \tag{2.20}$$

where

$$\begin{aligned} I_k &= 0 \text{ if } S_k = S_{k-1} && (2.21)\\ &= 1 \text{ if } S_k \neq S_{k-1}, && (2.22) \end{aligned}$$

and $S_1, S_2, \ldots, S_n$ denote the indicator variables designating the signs of the $X_{(i)}$ values (S_i is 1 if $X_{(i)}$ is nonnegative, 0 otherwise).

The first test (A) is based on calculating the runs test statistic of symmetry (McWilliams (1990)) and using it as a basis for choosing between the sign test and the Wilcoxon signed rank test. The second procedure (R) is based on calculating a measure of symmetry and using the Wilcoxon signed rank test if this measure falls in the region indicating large asymmetry and the Sign test otherwise.

A disadvantage with the first procedure is that the runs test may give highly significant values, not because the distribution is asymmetric but because it is

symmetric about the true value of the median, which may be different from the one specified by the null hypothesis. Thus, this procedure would inappropriately choose the Sign test and therefore be less powerful. The second procedure has a disadvantage, also shared by (A), in the discontinuous nature of the test selection method. It is not difficult to imagine a situation where a very small change in one observation value in the data may result in a different choice of the test statistic. This could give a conflicting decision compared to the decision obtained with the other test (O'Gorman (1996)).

2.3.2 A Continuosly Adaptive Rank Test for Shift in Location (Baklizi (2005))

Baklizi used the P-value from the triples test (Randles et al. (1980)) to obtain modified Wilcoxon scores and developed an adaptive rank test with the assumption of symmetry combining the sign and signed rank tests.

Before introducing the statistic we present a review of the triples test proposed by Randles et al. (1980). The null hypothesis for the triples test is that the underlying population is symmetric about θ against the alternative that it is asymmetric.

Let $X_1, \ldots, X_n$ denote a random sample from a continuous distribution with median θ. We take all possible triples from the sample (i.e., $\binom{n}{3}$ combinations). A triple of observations is skewed to the right if the middle observation is closer to the smaller observation than it is to the larger. Let

$$f^*(X_i, X_j, X_k) = \frac{1}{3}\Big[sign(X_i+X_j-2X_k)+sign(X_i+X_k-2X_j)+sign(X_j+X_k-2X_i)\Big], \tag{2.23}$$

where $sign(x) = 1, 0, -1$ according as $x >, =, < 0$. Thus, the range of the function f^* is $\{-1/3, 0, 1/3\}$. The triples test is then based on the U-statistic

$$\hat{\eta} = \frac{1}{\binom{n}{3}} \sum_{i<j<k} f^*(X_i, X_j, X_k). \tag{2.24}$$

Reject the null hypothesis of symmetry if $|V| > \tau_{\alpha/2}$, where $\tau_{\alpha/2}$ is the upper $\alpha/2$th quantile of the standard normal distribution, and

$$V = \frac{\sqrt{n}\hat{\eta}}{\hat{\sigma}_{\hat{\eta}}}. \tag{2.25}$$

In order to compute $\hat{\sigma}^2_{\hat{\eta}}$, i.e. the estimated variance of $\hat{\eta}$, a rather complex expression is used, which can be found in Randles et al. (1980).

Let P_R denote the P-value of the Randles test. Consider the Wilcoxon scores, $a^*(j)$ as follows:

$$a^*_{(j)} = \begin{cases} j \text{ if } Y^{(j)} > 0, \\ 0 \text{ if } Y^{(j)} \leq 0, \end{cases} \tag{2.26}$$

where $Y^{(j)}$ is the observation corresponding to $|Y|_{(j)}$, the jth largest Y in magnitude. Given that $P_R = p$, the scores of the proposed test for Baklizi are as follows:

$$a_{(j)} = \begin{cases} j^p \text{ if } Y^{(j)} > 0, \\ 0 \text{ if } Y^{(j)} \leq 0. \end{cases} \tag{2.27}$$

The reason for this choice of scores is that the P-value can be considered as the amount of evidence against symmetry of the distribution present in the data. Small values of p are evidence of asymmetry. Therefore, the scores of the proposed statistic tend towards those of the Sign test given by

$$s(j) = \begin{cases} 1 \text{ if } Y^{(j)} > 0, \\ 0 \text{ if } Y^{(j)} \leq 0. \end{cases} \tag{2.28}$$

Otherwise, if the data do not present evidence of asymmetry, the P-value is large. As the P-value approaches 1, the scores of the proposed statistic by Baklizi approaches the scores of the Wilcoxon signed rank test. The advantage of this statistic is that it adapts its scores smoothly and continuously according to the "amount" of asymmetry in the distribution as indicated by the magnitude of the P-value of the preliminary symmetry test. A disadvantage of this procedure is that the symmetry is an important factor in the construction of statistics and their impact can not be fully measured.

2.3.3 Adaptive Nonparametric Tests for a Single Sample Location Problem (Bandyopadhyay and Dutta (2007))

Bandyopadhyay and Dutta suggest two adaptive test procedures, one is a probabilistic approach while the other is a deterministic approach. The deterministic approach is based on calculating a measure of symmetry and using it as a basis for choosing between the sign test and the Wilcoxon signed rank test. As in the procedure proposed by Baklizi, the probabilistic approach is also a combination of the Sign test and the Wilcoxon signed rank test according to evidence of asymmetry provided by the P-value from the triples test for symmetry given in Randles et al (1980).

The probabilistic approach is given by the following rule. Let p denote the P-value corresponding to an observed $\hat{\eta}$ (see [2.24]) . The P-value can be considered as the amount of evidence against symmetry of the distribution present in the data, as in the previous procedure. Whenever p is observed, perform a Bernoullian trial with probability of success p. If success occurs, use the Wilcoxon signed rank test; otherwise, use the Sign test. The adaptive test rule is: Reject H_0 with probability p if $W > w$ and with probability $(1 - p)$ if $S > s$, where w and s are the upper α-critical values for the Wilcoxon signed rank and Sign tests respectively.

For the deterministic approach a simple measure of symmetry is introduced on which the preliminary test would be based. The proposed measure of symmetry has limits -1 and 1, and is given by

$$Q = \frac{X_{(n)} - 2\tilde{X} + X_{(1)}}{X_{(n)} - X_{(1)}}, \tag{2.29}$$

where $\tilde{X}$, $X_{(1)}$ and $X_{(n)}$ denotes the median, minimum and maximum of the distribution respectively.

For a symmetric distribution the median is expected to be equidistant from both extremes, while for a positively skewed distribution the median will be closer to the minimum and for a negatively skewed distribution it will be closer to the maximum. Then the quantity is divided by rank to express it as a pure number.

The proposed adaptive test statistic is then given by

$$T = SI_{(|Q|>c)} + WI_{(|Q|\leq c)}, \tag{2.30}$$

where $I_{(x)}$ is an indicator function assuming the values 1 or 0 according as x is true or false.

The authors examined different values and $c = 0.075$ is found to be the best choice in terms of robustness of the test.

In this section different adaptive tests for the one sample problem were described, however none of these procedures presented any new statistics. In all cases the authors worked with the well-known statistics of the Sign and Wilcoxon. The following chapter presents a new statistic to the problem in question.

Chapter 3
The Runs Statistic

In developing the sign test only the signs of the observations are used. The Wilcoxon signed rank test takes information of the magnitudes of the X_i (i.e., of the $\left|X\right|_i$) into account as well. The run statistic C takes additional information about the size, position and the distribution of X_i into account. Furthermore, C is more sensitive to small changes in the parameter of location and offers more levels of significance for small samples than the sign test and the Wilcoxon signed rank test. This will be discussed below.

Let $X_1, ..., X_n$ be a random sample with a common continuous cumulative distribution function $F(X-\theta)$. A test for the hypothesis $H_0 : \theta = 0$ $v.s.$ $H_1 : \theta > 0$, based on runs has been proposed by Corzo (1989). A run is defined to be a succession of two or more identical symbols which are followed and preceded by different symbols or no symbol at all Gibbons (1992).

The test algorithm is simple. The first point is the construction of the runs, for this purpose the variable η_j is defined as

$$\eta_j = S(X_{D_j}) = \begin{cases} 1 & \text{if } X_{D_j} > 0, \\ 0 & \text{otherwise,} \end{cases} \qquad j = 1, 2, \ldots, n, \tag{3.1}$$

where D_j is the antirank of $\left|X\right|_{(j)}$ such that $\left|X_{D_j}\right|$=$\left|X\right|_{(j)}$. Hence D$_j$ labels the X which corresponds to the jth ordered absolute value.

Then $\eta_1, \ldots, \eta_n$ is a dichotomized sequence and may be represented by

$$\begin{aligned} \eta_1 = \ldots = \eta_{L_1} \neq \eta_{L_1+1} = \ldots = \eta_{L_1+L_2} \neq \\ \eta_{L_1+L_2+1} = \ldots \neq \ldots = \eta_{L_1+\ldots+L_{u-1}} \neq \ldots = \eta_n. \end{aligned} \tag{3.2}$$

In this structure $U \geq 1$ different groups of identical symbols can be identified, each one of which defines a run and such that the ith group has the length L_i.

This sequence is usually denoted as $\vec{\eta}$ where $\vec{\eta_1}, \ldots, \vec{\eta_k}$ indicate the k runs of the dichotomized sequence. Obviously $\eta_1, \ldots, \eta_n$ are independent rv's bernoulli with parameter $p = Pr[\eta_j = 1], j = 1, \ldots, n$.

Example 1 *Consider the observations: 3.1, -4.2, -2.4, 4, 5. The ordered absolute values are:* $|-2.4| < |3.1| < |4| < |-4.2| < |5|$. *Then we have that* η_1*=0,* η_2*=1,* η_3*=1,* η_4*=0,* η_5*=1 or* $\vec{\eta} = (0, 1, 1, 0, 1)$.

The changes in the dichotomized succession are identified with the following indicators:

$$I_1 = 1, \tag{3.3}$$

$$I_j = \begin{cases} 1 \text{ if } \eta_{j-1} \neq \eta_j, \\ 0 \text{ if } \eta_{j-1} = \eta_j, \end{cases} \qquad j = 2, \ldots, n. \tag{3.4}$$

To capture the relevant information in the succession, the number of runs until the jth element of the dichotomized succession is obtained through the following partial sums:

$$r_i = \sum_{j=1}^{i} I_j, i = 1, \ldots, n. \tag{3.5}$$

Naturally $r_i \leq r_j$ for $i < j$, and r_n is the total number of runs in the sequence.

For Example to determine I_j and r_j of the observations: 3.1, -4.2, -2.4, 4, 5, we note that η_1=0, η_2=1, η_3=1, η_4=0, η_5=1, then the indicators I_j are I_1=1, I_2=1, I_3=0, I_4=1 and I_5=1, and the partial sums: r_1=1, r_2=2, r_3=2, r_4=3 and r_5=4. Obviously, there are 4 runs in this sequence.

The test statistic proposed is

$$C = \frac{1}{r_n} \sum_{j=1}^{n} \delta_j r_j, j = 1, \ldots, n, \tag{3.6}$$

where,

$$\delta_j = \begin{cases} 1 & \text{if } \eta_j = 1, \\ -1 & \text{if } \eta_j = 0, \end{cases} \quad j = 1, 2, \ldots, n. \tag{3.7}$$

For the observations in the Example 1: 3.1, -4.2, -2.4, 4, 5, we have that r_1=1, r_2=2, r_3=2, r_4=3, r_5=4 and δ_1=-1, δ_2=1, δ_3=1, δ_4=-1, δ_5=1. Then the statistic is

$$C = \frac{1}{r_n} \sum_{j=1}^{n} \delta_j r_j = \frac{(-1)(1) + (1)(2) + (1)(2) + (-1)(3) + (1)(4)}{4} = 1. \tag{3.8}$$

It is easy to notice that C includes the number of runs until every element of the dichotomized succession, increasing their value when $\eta_j = 1$ ($\delta_j = 1$, runs of ones) and decreasing when $\eta_j = 0$ (δ_j =-1, runs of zeros). Obviously, great values of C indicate greater number of runs of ones, and it's an indication that $\theta > 0$. Additionally the inverse of the total number of runs $\frac{1}{r_n}$ is used as a factor of standardization. Some properties of the statistic C are discussed below.

Theorem 1 *The statistics C takes values between -n and n.*

Proof. From [3.5], for any $r_j, j = 1, ..., n$ we have that $r_j \leq r_n$ and hence that $\frac{1}{r_j} \geq \frac{1}{r_n}$. Then

$$C = \sum_{j=1}^{n} \frac{1}{r_n} \delta_j r_j \leq \sum_{j=1}^{n} \frac{r_n}{r_n} \delta_j = \sum_{j=1}^{n} \delta_j, \leq n, \tag{3.9}$$

and similarly,

$$C \geq \sum_{j=1}^{n} \delta_j \geq -n, \tag{3.10}$$

because $\delta_j = 1$ or -1.
This completes the proof.

■

Theorem 2 *The statistics C may be rewritten as*

$$C = \frac{1}{k} \sum_{j=1}^{k} \delta_j^* j L_j = \frac{1}{k} \big[\delta_1^* L_1 + \delta_2^* 2L_2 + \ldots + \delta_j^* j L_j\big], \tag{3.11}$$

where

$$\delta_j^* = \begin{cases} 1 & \textit{if } \vec{\eta_j} = 1, \\ -1 & \textit{if } \vec{\eta_j} = 0, \end{cases} \quad j = 1, 2, \ldots, k, \tag{3.12}$$

and $\vec{\eta_j}$ is the jth run, L_j is the length of the jth run and k is the number of runs.

Proof. For a same run with length L_i, we have that $r_i = r_{i+1} = \ldots = r_{i+L_i}$ and $\delta_i = \delta_{i+1} = \ldots = \delta_{i+L_i}$ accordingly $r_i\delta_i = r_{i+1}\delta_{i+1} = \ldots = r_{i+L_i}\delta_{i+L_i}$. Any sequence with k runs will take the following form: $r_1\delta_1 = \ldots = r_{L_1}\delta_{L_1} \neq r_{L_1+1}\delta_{L_1+1} = \ldots = r_{L_1+L_2}\delta_{L_1+L_2} \neq r_{L_1+L_2+1}\delta_{L_1+L_2+1} = \ldots \neq \ldots = r_{L_1+\ldots+L_{k-1}}\delta_{L_1+\ldots+L_{k-1}} \neq \ldots = r_n\delta_n$, where L_i is the length of the run i and k is the number of runs.

Also, from [3.5] $r_1 \leq r_2 \leq \ldots \leq r_n$ and for the first run $r_1 = \ldots = r_{L_1} = 1$ and for the second run $r_{L_1+1} = \ldots = r_{L_1+L_2} = 2$,..., etc. then $r_1\delta_1 = \ldots = r_{L_1}\delta_{L_1} = L_1\delta_1^*$, $r_{L_1+1}\delta_{L_1+1} = \ldots = r_{L_1+L_2}\delta_{L_1+L_2} = L_2\delta_2^*$, ... and it follows $C = \frac{1}{k}\big[\delta_1^* L_1 + \delta_2^* 2L_2 + \ldots + \delta_k^* k L_k\big]$.

This completes the proof.

■

As a result of the previous theorem, the blocks of objects of ones and zeros must alternate, if the sequence begins with a run of ones the statistics C may be rewritten as

$$C = \frac{1}{k}\sum_{j=1}^{k}(-1)^{j+1} j L_j = \frac{1}{k}\big[L_1 - 2L_2 + 3L_3 - \ldots \pm kL_k\big], \tag{3.13}$$

and if the sequence begins with a run of zeros

$$C = \frac{1}{k}\sum_{j=1}^{k}(-1)^{j} j L_j = \frac{1}{k}\big[-L_1 + 2L_2 - 3L_3 + \ldots \pm kL_k\big], \tag{3.14}$$

where L_j is the length of the jth run and k is the number of runs.

On the other hand, without loss of generality, it is possible to make any analysis with arrangements that start with zeros or arrangements that start with ones. Also, the statistic is symmetric around zero (see the next Theorem).

Theorem 3 *The values of the statistic C are symmetrical around 0.*

$Proof$. Consider first the set υ of binary arrangements, which are sequences of elements taken from the symbols $\eta = \{0, 1\}$,

$$\upsilon := \{0, 1, 00, 10, 01, 11, 000, 100, 010, 001, \ldots, 111001100, \ldots\}.$$

Then, for any $n > 0$, there are 2^n different arrangements of zeros and ones, representing all possible configurations of zeros and ones in an arrangement of size n.

Let $\vec{\eta_k} = \{\vec{\eta_1}, \vec{\eta_2}, \ldots, \vec{\eta_k}\}$ be an arrangement in υ_n with k runs , where $L_1, L_2, \ldots, L_k$ are respectively the lengths of each run. Clearly, $L_1 + L_2 + \ldots + L_k = n$.

Then, for a specific n there are two symmetrical sequences $\vec{\eta}_{k0}$ and $\vec{\eta}_{k1}$ with lengths of runs $L_1, L_2, \ldots, L_k$, in which the only difference is that one of them, begins with zeros ($\delta_1^* = -1$), and another begins with ones ($\delta_1^* = 1$). In other words, from [3.14] and [3.13] the statistics C for the sequences $\vec{\eta}_{k0}$ and $\vec{\eta}_{k1}$ is given by

$$C(\vec{\eta}_{k0}) = (1/k)(-L_1 + 2L_2 - 3L_3 + \ldots \pm kL_k),$$

and

$$C(\vec{\eta}_{k1}) = (1/k)(L_1 - 2L_2 + 3L_3 - \ldots \pm kL_k),$$

respectively.

We see that $C(\vec{\eta}_{k1}) = -C(\vec{\eta}_{k0})$ and $-C(\vec{\eta}_{k1}) = C(\vec{\eta}_{k0})$. Hence, it is easy to notice that to each value of the statistic (positive or negative) belongs a reciprocal value (positive or negative respectively), making the statistic symmetrical around zero. This completes the proof.

■

Likewise, it can be seen that positive values of the statistic are an indicator of the predominance of runs of ones. This indicates a large number of observations larger than the median. We have the most extreme case, when all the values are higher than the median, then there is only one run and the statistics takes the maximum value $C = n$. On the other hand, when the number of ones and zeros is similar, the values of statistics fluctuate around 0, this is an indicator that the median of the distribution sampled is zero. If, however, the number of zeros or ones increases, the values of statistics differ from zero in both positive and negative directions indicating departure from the null hypothesis (median non-zero).
For $\theta > 0$, it is expected that C takes "large" positive values. Accordingly, we reject the null hypothesis H_0 for large values of C, e.i., if $C \geq c$, where with a level of significance α,

$$Pr_{H_0}[C \geq c] = \alpha. \tag{3.15}$$

In order to apply the statistic C for testing [2.1], the rule is to reject $H_0 : \theta = 0$ in favor of $H_1 : \theta > 0$ if $C \geq c_{1-\alpha/2}$. The critical value $c_{1-\alpha/2}$ is determined so that $P_{H_0}\left(C \geq c_{1-\alpha/2}\right) = \alpha$. Hence, we must first determine the distribution of C under H_0. A method for the calculation of the distribution of C is presented in the following chapter.

Chapter 4
The Probability Distribution of C

In this chapter, the probability distribution function of the statistic C is determined using ordinary generating functions and an algorithm [1] on Mathematica [2] is developed to facilitate the calculations. Generating functions are particularly useful for solving counting problems. In particular, problems involving choosing items from a set often lead to nice generating functions by letting the coefficient of t^n be the number of ways to choose n items. Moreover, an alternative approach based on partitions is presented in this chapter.

Many authors have studied the distribution of the total number of runs. Results concerning to the distribution of runs of $n = n_1 + n_2$ elements, when the numbers n_1, n_2 of 1's and 0's in the sequence are fixed were published by Ising (1925) and Stevens (1939). Later, Wald and Wolfowitz (1940) rederived those results and used to test whether two samples are from the same population. Barton and David (1957) gave a recursion formula to calculate the number of ways to have the total number of runs as k of a system with m objects. Whitworth (1959) investigated the distribution of the total number of runs of two kinds of objects. Gordon et al. (1986), Schilling (1990), determined the number of possible arrangements with u runs, given n_0 0's. Shaughnessy (1981) and Schuster and Gu (1997) gave more recursive formulas for multiple objects. Those recurrences were an improvement over that of Barton and David (1957) in that they did not involve composition of k in the summation. Kong (2006) developed one explicit formula and one new recursion formula for the distribution of the total number of runs of multiple objects using ordinary generating functions. Macwilliams (1990) showed that under $H_0 : \theta = 0$, the total number of runs is binomial, but the literature contains no explicit formula for the distribution of the total number of runs for a one sample problem (i.e. for a fixed n and the number of runs variable). In this chapter the distribution of the total number of runs for the binary case and the probabil-

[1] see Appendix

[2] Mathematica, is a computational mathematical software program used in varied mathematical fields and other areas of technical computing.

ity distribution of C are determined using some of the results mentioned above. Likewise, an alternative method based on the partition function is developed to determine the distribution of C.

The probability function of the statistic C given in this chapter, is of great importance, because it is the basis to determine the probability function under the null hypothesis and the power function of the statistic.

Assume an ordered sequence $\{\vec{\eta}_j\}$ of n elements of two types, n_1 of the first type (1´s) and n_2 of the second type (0´s), where $n_1 + n_2 = n$ and η_j defined as in [3.1]. Accordingly, the total number of runs (r_n) in the sequence $\{\vec{\eta}_j\}$ should be between 1 and n, see [3.5]. Then, the sets $r_n^{(i)} : \{r_n = i\}, i = 1, 2 \ldots, n$ (the set of possible sequences with i runs) are disjoint and exhaustive[3]. This information may be used to determine the distribution of C.

Theorem 4 *Let r_n be the total number of runs in a sequence $\{\vec{\eta}_j\}$ of n elements, then*

$$Pr[C = c] = Pr\Big[\frac{1}{r_n}\sum_{j=1}^{n}\delta_j r_j = c\Big] = \sum_{i=1}^{n} Pr\Big[\sum_{j=1}^{i}\delta_j r_j = ci\Big] Pr[r_n = i], \quad (4.1)$$

where r_n is the total number of runs.

$Proof$. From [3.6] and apply the theorem of total probability, we have

$$Pr[C = c] = \sum_{i=1}^{n} Pr\Big[\frac{1}{r_n}\sum_{j=1}^{n}\delta_j r_j = c | r_n = i\Big] Pr[r_n = i], \quad (4.2)$$

and for $r_n = i$ (i runs), we obtain that

$$Pr\Big[\frac{1}{r_n}\sum_{j=1}^{n}\delta_j r_j = c | r_n = i\Big] = Pr\Big[\sum_{j=1}^{i}\delta_j r_j = ci\Big]. \quad (4.3)$$

This completes the proof.

■

The probability distribution of r_n, the total number of runs, is presented in the following theorem.

[3] $\bigcup_{i=1}^{n} r_n^{(i)} = \Omega$ or simply $Pr[\bigcup_{i=1}^{n} r_n^{(i)}] = 1$

Theorem 5 *The probability distribution of r_n, the total number of runs in a random sample of size $n = n_1 + n_2$, n_1 of type 1 and n_2 of type 2, is given by*

$$Pr[r_n = i] = \begin{cases} p^n + (1-p)^n & \text{if } i = 1, \\ \sum_{n_1=u}^{n-u} f_{r_n}[i|n_1] f_B[n_1] & \text{if } 1 < i \leq n, \end{cases} \tag{4.4}$$

where

$$u = \begin{cases} \frac{i}{2} & \text{if } i \text{ is even}, \\ \\ \frac{i-1}{2} & \text{if } i \text{ is odd}, \end{cases} \tag{4.5}$$

$$f_{r_n}[i|n_1] = \begin{cases} \frac{2\binom{n_1-1}{\frac{i}{2}-1}\binom{n-n_1-1}{\frac{i}{2}-1}}{\binom{n}{n_1}} & \text{if } i > 1 \text{ and is even}, \\ \\ \frac{\binom{n_1-1}{\frac{i-1}{2}}\binom{n-n_1-1}{\frac{i-3}{2}}+\binom{n_1-1}{\frac{i-3}{2}}\binom{n-n_1-1}{\frac{i-1}{2}}}{\binom{n}{n_1}} & \text{if } i > 1 \text{ and is odd}, \end{cases} \tag{4.6}$$

if either $n_1 = 0$ or $n_2 = n - n_1 = 0$ then $i = 1$ with probability one, and

$$f_B[n_1] = \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1}, \tag{4.7}$$

with $p = Pr[\eta_j = 1]$, $1 - p = Pr[\eta_j = 0]$ and η_j is defined as in [3.1].

$Proof$. The conditional probability distribution of r_n when n_1 and $n_2 = n - n_1$ are positive integers is given by Gibbons and Chakraborti, 1992. They showed that the probability distribution of r_n, the total number of $n = n_1 + n_2$ objects, n_1 of Type 1 and n_2 of type 2, in a random sample is given by [4.6]. Although the distribution in [4.6] can be used to calculate probabilities for a fixed n_1 and $n_2 = n - n_1$, these probabilities must be calculated for all possible values of n_1 in the sequence $\{\vec{\eta}\}$, $0 \leq n_1 \leq n$. Then we can sum over all values of n_1 to determine $Pr[r_n = i]$. Using the theorem of total probability, we have

$$Pr[r_n = i] = \sum_{n_1=u_1}^{n-u_1} f_{r_n}[i|B = n_1] Pr[B = n_1], \tag{4.8}$$

where $f_{r_n}[i|B = n_1]$ is the probability of obtaining i runs with n_1 and n_2 fixed. Note that $B = \sum_{j=1}^{n} \eta_j$ is the number of ones in the sequence $\{\vec{\eta_i}\}$, and η_j, $j = 1, \ldots, n$ are iid Bernoulli random variables with $p = Pr[\eta_j = 1]$. Then, it

follows directly that B, is binomial with parameters n and p.

It is easy to see that if $r_n = 1$ (one run), we have the sequence $\eta_1 = \eta_2 = \cdots = \eta_n = 1$ or the sequence $\eta_1 = \eta_2 = \cdots = \eta_n = 0$, then $Pr[r_n = 1] = p^n + (1-p)^n$, because $Pr[\eta_j = 1] = p$ and $Pr[\eta_j = 0] = 1 - p$. This completes the proof.

■

Example 2 *Suppose that $n = 6$. Using the result of theorem [5], we have*

$$Pr[r_n = i] = \begin{cases} p^6 + (1-p)^6 & \text{if } i = 1, \\ \sum_{n_1=u}^{6-u} f_{r_n}[i|n_1] f_B[n_1] & \text{if } 1 < i \leq 6. \end{cases}$$

The probability will be illustrated for $i = 2$ with $p = 1/4$, then from theorem [5], $u = 1$ and $f_B[n_1] = \binom{6}{n_1}(1/4)^{n_1}(3/4)^{6-n_1}$ and

$$\begin{aligned} Pr[r_n = 2] &= \sum_{n_1=u}^{6-u} f_{r_n}[i|n_1] f_B[n_1] \\ &= f_{r_n}[2|1] f_B[1] + f_{r_n}[2|2] f_B[2] + \ldots + f_{r_n}[2|5] f_B[5] \\ &= \frac{2\binom{0}{0}\binom{4}{0}}{\binom{6}{1}} \binom{6}{1} \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^5 + \ldots + \frac{2\binom{4}{0}\binom{0}{0}}{\binom{6}{5}} \binom{6}{5} \left(\frac{1}{4}\right)^5 \left(\frac{3}{4}\right)^1 \\ &= 0.177246. \end{aligned}$$

This result can be easily verified by constructing all possible sequences for $n = 6$ with two runs. Table [4.1] lists these sequences and their probabilities $\Big($Note that $\sum p(\vec{\eta}) = 0.177246\Big)$.

The distribution in theorem [4] can be rewritten using theorem [2], then

$$Pr[C = c] = Pr\Big[\frac{1}{r_n}\sum_{j=1}^{n} \delta_j r_j = c\Big] = \sum_{i=1}^{n} Pr\Big[\sum_{j=i}^{i} \delta_j^* j L_j = ci\Big] Pr[r_n = i], \tag{4.9}$$

where L_j is the length of the jth run, $L_1 + L_2 + \cdots + L_i = n$, and δ_j^* is defined as in theorem [2].

The probability $Pr[r_n = i]$ in [4.9] was discussed in theorem [5]. Then, to determine $Pr[C = c]$, we need only the probabilities

$$Pr\Big[\sum_{j=i}^{i} \delta_j^* j L_j = ci\Big]. \tag{4.10}$$

η_1	η_2	η_3	η_4	η_5	η_6	$p(\vec{\eta})$
1	0	0	0	0	0	0,000732
1	1	0	0	0	0	0,002197
1	1	1	0	0	0	0,006592
1	1	1	1	0	0	0,019775
1	1	1	1	1	0	0,059326
0	0	0	0	0	1	0,000732
0	0	0	0	1	1	0,002197
0	0	0	1	1	1	0,006592
0	0	1	1	1	1	0,019775
0	1	1	1	1	1	0,059326

Table 4.1: Possible sequences of $\vec{\eta}$ for $n = 6$ and $i = 2$

In Chapter 3 we have shown that, if $\eta_1 = 1$ (the sequence begins with a one), we can write that $\sum_{j=i}^{i} \delta_j^* j L_j = \sum_{j=1}^{i} (-1)^{j+1} j L_j = L_1 - 2L_2 + 3L_3 - \ldots \pm iL_i$ and if $\eta_1 = 0$ (the sequence begins with a zero), $\sum_{j=i}^{i} \delta_j^* j L_j = \sum_{j=1}^{i} (-1)^{j} j L_j = -L_1 + 2L_2 - 3L_3 + \ldots \pm iL_i$, then

$$Pr\Big[\sum_{j=i}^{i} \delta_j^* j L_j = ci\Big] = \begin{cases} Pr\Big[\sum_{j=1}^{i} (-1)^{j+1} j L_j = ci\Big] & \text{if } \eta_1 = 1, \\ \\ Pr\Big[\sum_{j=1}^{i} (-1)^{j} j L_j = ci\Big] & \text{if } \eta_1 = 0. \end{cases} \tag{4.11}$$

The next theorem facilitates the calculation of the above probabilities.

Theorem 6 *Let $L_1 + L_2 + \cdots + L_i = n$. Then*

$$Pr\Big[\sum_{j=1}^{i} (-1)^{j+1} j L_j = ci\Big] = \frac{G_1(i,c)}{\binom{n-1}{i-1}} \quad \textit{if } \eta_1 = 1, \textit{ and} \tag{4.12}$$

$$Pr\Big[\sum_{j=1}^{i} (-1)^{j} j L_j = ci\Big] = \frac{G_2(i,c)}{\binom{n-1}{i-1}} \quad \textit{if } \eta_1 = 0, \tag{4.13}$$

where $G_1(i,c)$ is the number of solutions of $\sum_{j=1}^{i} (-1)^{j+1} j L_j = ci$ and $G_2(i,c)$ is the number of solutions of $\sum_{j=1}^{i} (-1)^{j} j L_j = ci$.

Proof. If $\eta_1 = 1$, $P\Big[\sum_{j=1}^{i} (-1)^{j+1} j L_j = ci\Big]$ can be obtained as the positive

integer combination[4] that solves the system $\sum_{j=1}^{i}(-1)^{j+1}jL_j = ci$, then

$$P\Big[\sum_{j=1}^{i}(-1)^{j+1}jL_j = ci\Big] = \left[\frac{\text{number of solutions of } \sum_{j=1}^{i}(-1)^{j+1}jL_j = ic}{\text{number of arrangements of } \sum_{j=1}^{i}(-1)^{j+1}jL_j}\right], \tag{4.14}$$

where $n = L_1 + L_2 + \ldots + L_i$.

The total number of arrangements of $\sum_{j=1}^{i}(-1)^{j+1}jL_j$ with $n = L_1 + L_2 + \ldots + L_i$ and $L_i \geq 1$, denotes the number of compositions [5] $C(n,i)$ of n with exactly i parts. The number of compositions of n into i parts (where 0 is not allowed as a part) is given by

$$C(n,i) = \binom{n-1}{i-1}, \tag{4.15}$$

and replacing [4.15] in [4.14] we have theorem [6].

If $\eta_1 = 0$, the proof is analogous. This completes the proof.

■

In theorem [6], $G_1(i,c)$ is the number of solutions of $\sum_{j=1}^{i}(-1)^{j+1}jL_j = ci$, then if $i = 1$, we have that

$$G_1(1,c) = I_1(c) = \begin{cases} 1 \text{ if } c = n, \\ 0 \text{ otherwise}, \end{cases} \tag{4.16}$$

and for $i > 1$, $G_1(i,c)$ may be rewritten as[6]

$$G_1(i,c) = \text{number of solutions of}\Big[\sum_{j=1}^{i}(-1)^{j+1}jL_j = ci\Big], \tag{4.17}$$

$$= \text{number of solutions of}\Big[\sum_{j=1(2)}^{A} jL_j - 2\sum_{j=1}^{B} jL_{2j} = ci\Big], \tag{4.18}$$

[4] The equations that have integer coefficients and for which integer solutions are required, are called "Diophantic equations".

[5] A combinatorial composition is defined as an ordered arrangement of i nonnegative integers which sum to n. Compositions are merely partitions in which the order of the summands is considered.

[6] The index $j = 1(2)$ denotes $j = 1, 3, 5, \ldots$.

where $L_1 + L_2 + \cdots + L_i = n$,

$$A = \begin{cases} \text{i-1} & \text{if } i \text{ is even,} \\ \text{i} & \text{if } i \text{ is odd,} \end{cases} \tag{4.19}$$

and

$$B = \begin{cases} \frac{i}{2} & \text{if } i \text{ is even,} \\ \frac{i-1}{2} & \text{if } i \text{ is odd.} \end{cases} \tag{4.20}$$

It's important to note that the number of terms of $\sum_{j=1(2)}^{A} jL_j$ is given by

$$A^* = \begin{cases} \frac{i}{2} & \text{if } i \text{ is even,} \\ \frac{i+1}{2} & \text{if } i \text{ is odd,} \end{cases} \tag{4.21}$$

and the number of terms of $\sum_{j=1}^{B} jL_{2j}$ is B. Naturally $A^* + B = i$.

Example 3 *Suppose that $i = 10$, then $A = 9$, $B = 5$, $A^* = 5$,*

$$G_1(10, c) = \textit{number of solutions of} \Big[\sum_{j=1}^{10} (-1)^{j+1} jL_j = 10c \Big],$$

and

$$\begin{aligned} \sum_{j=1}^{10} (-1)^{j+1} jL_j &= L_1 - 2L_2 + 3L_3 - \ldots - 10L_{10} \\ &= L_1 + 3L_3 + \ldots - 2(L_2 + 2L_4 + \ldots + 5L_{10}) \\ &= \sum_{j=1(2)}^{9} jL_j - 2\sum_{j=1}^{5} jL_{2j}. \end{aligned}$$

Hence,

$$G_1(10, c) = \textit{number of solutions of} \Big[\sum_{j=1(2)}^{9} jL_j - 2\sum_{j=1}^{5} jL_{2j} = 10c \Big].$$

As can be seen in theorem [6], $\sum_{j=1}^{i} L_j = n$. Then, $\sum_{j=1(2)}^{A} L_j = n_1$ (the lengths of ones) and $\sum_{j=1}^{B} L_{2j} = n_2$ (the lengths of zeros), because $n_1 + n_2 = n$. We

call $C_1 = \sum_{j=1(2)}^{A} jL_j$ and $C_2 = \sum_{j=1}^{B} jL_{2j}$, then C_1 denote the possible positive integral solutions of the system

$$\sum_{j=1(2)}^{A} jL_j = L_1 + 3L_3 + 5L_5 + \ldots + AL_A = C_1, \tag{4.22}$$

with

$$L_1 + L_3 + L_5 + \ldots + L_A = n_1, \tag{4.23}$$

and C_2 denote the possible positive integral solutions of the system

$$\sum_{j=1}^{B} jL_{2j} = L_2 + 2L_4 + 3L_6 + \ldots + BL_{2B} = C_2 \tag{4.24}$$

with

$$L_2 + L_4 + L_6 + \ldots + L_{2B} = n_2, \tag{4.25}$$

where by [4.18]

$$C_1 - 2C_2 = ci. \tag{4.26}$$

The previous systems of equations can be solved using ordinary generating functions [7] (or simply generating functions). The ordinary generating functions for each system are determined below.
In the following theorems, $[x^n]$ will denote the nth coefficient of the sequence $\{C(x)\}$, and $[x^n y^k]$ will denote the coefficient (n, k) of the sequence $\{C(x, y)\}$.

Theorem 7 *The generating function of the number $C_1(x)$ of positive integral solutions of the equation [4.22] is*

$$C_1(x) = \sum_{i=1} C_1(i)x^i = \prod_{i=1(2)}^{A} \frac{x^i}{1 - x^i}, \tag{4.27}$$

where A is defined as in [4.19].

$Proof$. Appendix **C** shows that $\frac{1}{1-x}$ is the closed-form generating function for the sequence $L_1(x)$

$$\langle 1, 1, 1, \ldots, 1\rangle \longleftrightarrow 1 + x + x^2 + x^3 + \cdots = \sum_{i\geq 0} x^i. \tag{4.28}$$

The first coefficient in the above sequence corresponds to x^0, i.e. $i = 0$. But by theorem [4.18], $L_i > 0$. This means that the coefficients i should be positive

[7] see Appendix.

($i > 0$). Then, applying the operator left-shifting (see Appendix), we can generate a new generating function for $L_1^*(x)$ with only coefficients for $i > 0$, whose coefficients are given by

$$[x^p]L_1^*(x) = [x^p]xL_1(x) = \left[x^{1+p}\right]L_1(x). \tag{4.29}$$

Hence, we have that $\frac{x}{1-x}$ is the closed-form generating function for the sequence

$$\langle 1, 1, 1, \ldots, 1\rangle \longleftrightarrow x + x^2 + x^3 + \cdots = \sum_{i\geq 1} x^i. \tag{4.30}$$

Following a similar procedure we can determine the closed-forms of

$$\begin{aligned}
\langle 3, 3, 3, \ldots, 3\rangle &\longleftrightarrow x^3 + x^6 + \ldots \longleftrightarrow \frac{x^3}{1-x^3},\\
\langle 5, 5, 5, \ldots, 5\rangle &\longleftrightarrow x^5 + x^{10} + \ldots \longleftrightarrow \frac{x^5}{1-x^5},\\
&\vdots
\end{aligned}$$

Now, let $C_1(x)$ be the possible positive integral solutions of $L_1 + 3L_3 + 5L_5 + \ldots + AL_A$. Then, the generating function $C_1(x)$ for the sequence $\{C_1(x)\}$ is given by

$$\begin{aligned}
C_1(x) &= (x + x^2 + \ldots)(x^3 + x^6 + \ldots)\ldots(x^A + x^{2A} + \ldots)\\
&= \left(\sum_{L_1\geq 1} x^{L_1}\right)\left(\sum_{L_3\geq 1} x^{3L_3}\right)\ldots\left(\sum_{L_A\geq 1} x^{AL_A}\right)\\
&= \frac{x}{1-x}\frac{x^3}{1-x^3}\cdots\frac{x^A}{1-x^A}\\
&= \prod_{i=1(2)}^{A}\frac{x^i}{1-x^i},
\end{aligned}$$

where A is defined as in [4.19].
Hence indeed the coefficient of x^{C_1} is equal to the number of solutions of [4.22]. This completes the proof.

■

To see an interpretation to theorem [7], notice that the right-hand side of the above equations can be written as

$$x^{1+3+5+\cdots+A} + x^{1(2)+3+\cdots+A} + \cdots = \sum_{L_1, L_3, \ldots, L_A \geq 1} x^{L_1+3L_3+\cdots+AL_A}, \tag{4.31}$$

where the sum is taken over all finite sequences $L_1, L_3, L_5, \ldots$ such that $L_i \geq 1$. We note that the $C_1 th$ coefficient $[x^{C_1}]$ denotes the number of solutions of [4.22] and C_1 denotes the possible positive integral solutions of [4.22].

Theorem 8 *The generating function of the number $C_2(x)$ of positive integral solutions of the equation [4.24] is*

$$C_2(x) = \sum_{i=1} C_2(i)x^i = \prod_{i=1}^{B} \frac{x^i}{1-x^i}, \tag{4.32}$$

where B is defined as in [4.20].

$Proof$. As in the preceding proof, we have

$$\begin{aligned} C_2(x) &= (x + x^2 + \ldots)(x^2 + x^4 + \ldots)\ldots(x^B + x^{2B} + \ldots) \\ &= \left(\sum_{L_2 \geq 1} x^{L_1}\right)\left(\sum_{L_4 \geq 1} x^{2L_4}\right)\ldots\left(\sum_{L_B \geq 1} x^{BL_B}\right) \\ &= \frac{x}{1-x}\frac{x^2}{1-x^2}\cdots\frac{x^B}{1-x^B}. \\ &= \prod_{i=1}^{B} \frac{x^i}{1-x^i} \end{aligned}$$

Consequently, the coefficient of x^{C_2} equals the number of solutions of [4.24]. This completes the proof.

■

According to theorems [7] and [8], the coefficients of x^{C_1} and x^{C_2} in $\{C_1(x)\}$ and $\{C_2(x)\}$, are respectively the number of solutions of [4.22] and [4.24] with $L_i \geq 1$.

Now, by [4.26], we have that

$$ci = C_1 - 2C_2,$$

and then

$$C_1 = ci + 2C_2. \tag{4.33}$$

In [4.33] we observe that for a given value ci, C_1 can be written as a function of C_2. We will examine the range of values of C_2.

It is easy to see, that C_2 takes values between $[p_1, p_2]$ given by

$$p_1 = \begin{cases} \frac{\left(\frac{i}{2}\right)\left(\frac{i}{2}+1\right)}{2} & \text{if } i \text{ is even,} \\ \frac{\left(\frac{i-1}{2}\right)\left(\frac{i-1}{2}+1\right)}{2} & \text{if } i \text{ is odd,} \end{cases} \tag{4.34}$$

and

$$p_2 = \begin{cases} \frac{\left(\frac{i}{2}\right)\left(\frac{i}{2}-1\right)}{2} + \left(\frac{i}{2}\right)(n-i+1) & \text{if } i \text{ is even,} \\ \frac{\left(\frac{i-1}{2}\right)\left(\frac{i-1}{2}-1\right)}{2} + \left(\frac{i-1}{2}\right)(n-i+1) & \text{if } i \text{ is odd.} \end{cases} \tag{4.35}$$

For example, if i is even, from [4.20] we have that $L_2 + 2L_4 + 3L_6 + \ldots + (i/2)L_{(i/2)} = C_2$ and because $L_i \geq 1$, we have the minimum value of C_2 if $L_i = 1$, for $i = 1, 2 \ldots (i/2)$, given by

$$p_1 = 1 + 2 + 3 + \ldots + (i/2) = \frac{\left(\frac{i}{2}\right)\left(\frac{i}{2}+1\right)}{2}, \tag{4.36}$$

and we have the maximum value, if $L_i = 1$ for $i = 1, 2, \ldots, (i/2-1)$ and $L_{(i/2)} = n - (i-1)$ [8]. Because $L_2 + 2L_4 + 3L_6 + \ldots + (i/2)L_{(i/2)} = C_2$, then

$$p_2 = 1 + 2 + 3 + \ldots + (i/2-1) + (i/2)(n-(i-1)) \tag{4.37}$$

$$= \frac{\left(\frac{i}{2}\right)\left(\frac{i}{2}-1\right)}{2} + (i/2)(n-i+1). \tag{4.38}$$

Hence, the above information shows that the product of the coefficients $[x^{ci+2C_2}]$ and $[x^{C_2}]$ in $\{C_1(x)\}$ and $\{C_2(x)\}$ respectively provides the joint solution of equations [4.22,4.24] for a specific value of C_2. Through the summation over all possible values of C_2, it's possible to find all of these solutions in $[p_1, p_2]$, then

$$\sum_{C_2=p_1}^{p_2} [x^{ci+2C_2}]C_1(x)[x^{C_2}]C_2(x), \tag{4.39}$$

provides the possible solutions of equations [4.22] and [4.24] for ci and $L_i \geq 1$. Note that restrictions for L_i given in [4.23] and [4.25] have not been analyzed.

Now, we need to solve the complete system of equations [4.22,4.23,4.24,4.25]. This system can be solved using ordinary generating functions as it was discussed previously. The following theorems provide this calculation.

[8] If we have i runs, the maximum value of C_2 is obtained if the first $i-1$ runs have one element and the last run the rest, because $1 < 2 < \cdots < (i/2)$.

Theorem 9 *The generating function of the number $T_1(x,y)$ of positive integral solutions of the equations [4.22 , 4.23] is*

$$T_1(x,y) = \sum_{m,n} T_1(m,k)x^m y^k = \prod_{i=1(2)}^{A} \frac{x^i y}{1-x^i y}, \tag{4.40}$$

where A is defined as in [4.19].

$Proof$. The proof is analogous to the ones used in theorems [7,8].

Let $T_1(C_1, n_1)$ be the number of positive integral solutions of $L_1+3L_3+5L_5+\ldots+AL_A = C_1$ with $L_1+L_3+L_5+\ldots+L_A = n_1$. Then, the generating function $T_1(x,y)$ for the sequence $\{T_1(x,y)\}$ is given by

$$\begin{aligned}
T_1(x,y) &= (xy + x^2y^2 + \ldots)(x^3y + x^6y^2 + \ldots)\ldots(x^Ay + x^{2A}y^2 + \ldots) \\
&= \left(\sum_{L_1\geq 1} x^{L_1}y^{L_1}\right)\left(\sum_{L_3\geq 1} x^{3L_3}y^{L_3}\right)\ldots\left(\sum_{L_A\geq 1} x^{AL_A}y^{L_A}\right) \\
&= \prod_{i\geq 1(2)}\left(\sum_{L_i\geq 1}^{A} x^{iL_i}y^{L_i}\right) \\
&= \sum_{L_1,L_2,\ldots,L_A\geq 1}^{A} x^{L_1+3L_3+\cdots+AL_A}y^{L_1+L_3+\cdots+L_A} \\
&= \prod_{i=1(2)}^{A} \frac{x^i y}{1-x^i y}.
\end{aligned}$$

Therefore, the coefficient of $x^m y^k$ in [4.40] equals the number of solutions of [4.22,4.23]. This completes the proof.

■

Theorem 10 *The generating function of the number $T_2(x,y)$ of positive integral solutions of the equations [4.24,4.25] is*

$$T_2(x,y) = \sum_{m,n} T_2(m,k)x^m y^k = \prod_{i=1}^{B} \frac{x^i y}{1-x^i y}, \tag{4.41}$$

where B is defined as in [4.20].

Proof. The proof is analogous to the one used in the above theorems. Let $T_2(C_2, n_2)$ be the number of positive integral solutions of $L_2 + 2L_4 + 3L_6 + \ldots + BL_{2B} = C_2$ with $L_2 + L_4 + L_6 + \ldots + L_{2B} = n_2$. Then, the generating function $T_2(x, y)$ for the sequence $\{T_2(x, y)\}$ is given by

$$\begin{aligned} T_2(x,y) &= (xy + x^2y^2 + \ldots)(x^2y + x^4y^2 + \ldots)\ldots(x^By + x^{2B}y^2 + \ldots) \\ &= \left(\sum_{L_2\geq 1} x^{L_2}y^{L_2}\right)\left(\sum_{L_4\geq 1} x^{2L_4}y^{L_4}\right)\ldots\left(\sum_{L_{2B}\geq 1} x^{BL_{2B}}y^{L_{2B}}\right) \\ &= \prod_{i\geq 1}\left(\sum_{L_{2i}\geq 1} x^{iL_{2i}}y^{L_{2i}}\right) \\ &= \sum_{L_2,L_4,\ldots,L_{2B}\geq 1}^{B} x^{L_2+2L_4+\cdots+BL_{2B}}y^{L_2+L_4+\cdots+L_{2B}} \\ &= \prod_{i=1}^{B}\frac{x^iy}{1-x^iy}. \end{aligned}$$

Hence, the coefficient of x^my^k in [4.41] indeed equals the number of solutions of [4.24,4.25].This completes the proof.

■

Relationships between the generating function and the number of solutions of [4.22,4.23] were seen in [4.33]. In view of our construction, and because $n_1 = n - n_2$ and $C_1 = ci + 2C_2$, multiplying the coefficients of $[x^{ci+2C_2}y^{n-n_2}]$ and $[x^{C_2}y^{n_2}]$ in $\{T_1(x, y)\}$ and $\{T_2(x, y)\}$ respectively, we obtain precisely the solution of equations [4.22,4.23,4.24,4.25] for specific values of C_2 and n_2.

From [4.25] we have that $\sum_{j=1}^{B} L_{2j} = n_2$, then the smallest value of n_2 is B, if $L_{2j} = 1, j = 1, \ldots, B$. Using the same principle, the smallest value of n_1 is $i-B$, where i denotes the number of runs. Hence, the maximum value of $n_2 = n - n_1$ is $n - (i - B)$, obviously n minus the minimum value of n_1. Clearly, the limits for n_2 are given by

$$B \leq n_2 \leq n - i + B. \tag{4.42}$$

Through the summation over all possible values of C_2 and n_2 it is possible to determine $G_1(i, c)$ in [4.18]. Then we get that

$$G_1(i,c) = \sum_{C_2=p_1}^{p_2}\sum_{n_2=B}^{n-i+B}[x^{ci+2C_2}y^{n-n_2}]T_1(x,y)[x^{C_2}y^{n_2}]T_2(x,y), \tag{4.43}$$

denotes the total number of solutions of equations [4.22,4.23,4.24,4.25] if $\eta_1 = 1$.

Similarly, from Theorem [6], the number of solutions of $\sum_{j=1}^{i}(-1)^j jL_j = ic$ is $G_2(i,c)$. If $i = 1$, we have that

$$G_2(1,c) = I_2(c) = \begin{cases} 1 \text{ if } c = -n, \\ 0 \text{ otherwise,} \end{cases} \tag{4.44}$$

and for $i > 1$, $G_2(i,c)$ may be rewritten as

$$\begin{aligned} G_2(i,c) &= \text{number of solutions of} \Big[\sum_{j=1}^{i}(-1)^j jL_j = ic \Big] && (4.45) \\ &= \text{number of solutions of} \Big[-\sum_{j=1(2)}^{A} jL_j + 2\sum_{j=1}^{B} jL_{2j} = ic \Big]. && (4.46) \end{aligned}$$

Using the definitions for C_1 and C_2 in [4.26], we have that

$$C_1 = 2C_2 - ci. \tag{4.47}$$

By Theorems [9, 10], multiplying the coefficients of $[x^{2C_2-ci}y^{n-n_2}]$ and $[x^{C_2}y^{n_2}]$ in $\{T_1(x,y)\}$ and $\{T_2(x,y)\}$ respectively, we obtain precisely the solution of equations [4.22, 4.23, 4.24, 4.25] for specific values of C_2 and n_2. Through the summation over all possible values of C_2 and n_2 it is possible to determine $G_2(i,c)$ in [4.18]. Then we get that

$$G_2(i,c) = \sum_{C_2=p_1}^{p_2} \sum_{n_2=B}^{n-i+B} [x^{2C_2-ci}y^{n-n_2}]T_1(x,y)[x^{C_2}y^{n_2}]T_2(x,y), \tag{4.48}$$

which denotes the total number of solutions of equations [4.22,4.23,4.24,4.25] if $\eta_1 = 0$.

In conclusion, based on [4.43] and [4.48], we can determine the distribution of C. The following theorem shows this result.

Theorem 11 *The probability distribution function of the statistic C is given by*

$$Pr[C = c] = \sum_{i=1}^{n} Pr[r_n = i] \left[\frac{G_1(i,c)}{\binom{n-1}{i-1}} Pr[\eta_1 = 1] + \frac{G_2(i,c)}{\binom{n-1}{i-1}} Pr[\eta_1 = 0] \right], \tag{4.49}$$

$-n \leq c \leq n$.
where $Pr[r_n = i]$ are defined by theorem [5], $G_1(i,c)$ and $G_2(i,c)$ by [4.43] and [4.48] respectively.

$Proof$. By [3.1], we have that

$$Pr[\eta_1 = 1] + Pr[\eta_1 = 0] = 1, \tag{4.50}$$

and using the theorem of total probability and theorem [6], we then have

$$\begin{aligned} Pr\Big[\sum_{j=i}^{i} \delta_j^* j L_j = ci\Big] &= Pr\Big[\sum_{j=1}^{i} (-1)^{j+1} j L_j = ci\Big] Pr[\eta_1 = 1] \\ &\quad + Pr\Big[\sum_{j=1}^{i} (-1)^{j} j L_j = ci\Big] Pr[\eta_1 = 0], \end{aligned} \tag{4.51}$$

and then

$$Pr\Big[\sum_{j=i}^{i} \delta_j^* j L_j = ci\Big] = \left[\frac{G_1(i,c)}{\binom{n-1}{i-1}} Pr[\eta_1 = 1] + \frac{G_2(i,c)}{\binom{n-1}{i-1}} Pr[\eta_1 = 0]\right] \tag{4.52}$$

Finally, by replacing [4.52] in [4.9] we obtain the probability distribution function of the statistic C. This completes the proof.

■

4.1 An Additional Approach

The theory of partitions has an interesting history. Certain problems in partitions date back from early middle ages and its study has fascinated a number of great mathematicians: Euler, Ramanujan, Hardy, Legendre, Rademacher, Sylvester and Dyson to name a few. They have all contributed to the development of an advanced theory in this field. There are many papers, books chapters and complete books devoted entirely to partitions. This comprehensive literature and its application to different areas makes partitions a powerful tool.

Many of the mathematical sciences have seen applications of partitions recently. For example in nonparametric statistic, restricted partitions are required in the formulation of statistics and their distributions. Some authors such as Richmond and Knopfmacher(1995), Voinov and Nikulin (1997), Nikulin et al. (1999), Nikulin, Smirnov and Voinov (2002), Kong (2006) and Wagner (2009) have made

great advances in nonparametric statistics using the theory of partitions.

In this section the distribution of the statistic C using the theory of partitions is presented. This approach is useful, because it facilitates the calculations and the approximations for large sizes.

Definition 1 *A partition of a positive integer n is a finite nonincreasing sequence of positive integers $\lambda_1, \ldots, \lambda_k$ such that $\sum_{i=1}^{k} \lambda_i = n$. The λ_i are called the parts of the partition.*

Definition 2 *The number of (unordered) ways of writing the number n as the sum of exactly m positive integers is called $p(n, m)$, or simply the partition of n into m parts. More rigorously,*

$$p(n, m) = \lambda_1 + \lambda_2 + \ldots + \lambda_m = n, \tag{4.53}$$

with $\lambda_1, \lambda_2, \ldots, \lambda_m$ positive integers and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 1$.

We shall consider $p(n, m)$, the number of partitions of n that lie in a set of partitions S and have m parts. This immediately leads to the two variable generating function (see Andrews 1976)

$$D(x, y) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} p(S, m, n) z^m p^n, \tag{4.54}$$

and its respective closed-form is given by

$$\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} p(S, m, n) z^m p^n = \prod_{j=0}^{\infty} \frac{1}{1 - zq^j}. \tag{4.55}$$

Using the generating function above, various closed-forms are easily determined. For example, the generating function for $p_{od}(n, m)$, the number of partitions of n that lie in a set of partitions S and have m odd parts, is given by

$$D_{od}(x, y) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} p_{od}(S, m, n) z^m p^n, \tag{4.56}$$

and its respective closed-form is

$$\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} p_{od}(S, m, n) z^m p^n = \prod_{j_{odd}}^{\infty} \frac{1}{1 - zq^j}. \tag{4.57}$$

In partition identities, we are often interested in the number of partitions that satisfy some condition. We denote such a number by $p(n|[condition])$. For example $p(n|[\text{odd parts}])$. In this regard Euler's theorem is of great importance, which is presented below.

Theorem 12 *Euler's Theorem. The number of partitions of n using exactly m odds parts equals the number of partitions of n into m distinct parts.*

$Proof$. Constantine (1987) gives an ingenious demonstration based on Ferrer's diagrams.

■

Definition 3 *Let n and m be positive integers. Then we can define the following functions:*

- $p(n,m)$*, the number of partitions of n into exactly m positive integral parts*
- $P(n,m) = \sum_{r\leq m} p(n,r)$*, the number of partitions of n into at most m integer parts, or what is the same, into parts not exceeding m.*
- $q(n,m)$*, the number of partitions of n into exactly m positive distinct parts.*

We do not need to consider each one of the above functions separately in virtue of elementary relations

$$P(n,m) = p(n+m,m) = q\Big[n + \binom{m+1}{2}, m\Big]. \tag{4.58}$$

These functions have been studied in detail by Szekeres in two remarkable papers Szekeres (1951) and Szekeres (1953).

We are interested in restricted partitions, that is, partitions in which the largest part is, say, $\leq N$ and the number of parts is $\leq m$. This task will naturally lead us to the Gaussian polynomials [9] (see Andrews (1976)). The q-binomial numbers are generating functions for certain partitions. The complete method can be studied in Andrews (2004). The following identities show this relation between q-binomial

[9]The q-binomial coefficient is a q-analog for the binomial coefficient, also called a Gaussian coefficient or a Gaussian polynomial. A q-binomial coefficient is given by

$$\binom{n}{m}_q = \frac{(q)_n}{(q)_m (q)_{n-m}} = \prod_{i=0}^{m-1} \frac{1-q^{n-i}}{1-q^{i+1}} \tag{4.59}$$

where

$$(q)_k \equiv \prod_{i=1}^{\infty} \frac{1-q^m}{1-q^{k+m}}. \tag{4.60}$$

numbers and partitions and in turn allow the determination of the respective generating functions in both cases:

$$\prod_{j=1}^{N}(1+zq^j) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} p(n|\text{m distinct parts each} \leq N\ )z^m q^n, \quad (4.61)$$

$$\prod_{j=1}^{N}\frac{1}{1-zq^j} = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} p(n|\text{m parts each} \leq N\ )z^m q^n. \quad (4.62)$$

Theorems [9] and [10] present the generating functions of $T_1(C_1, n_1)$ and $T_2(C_2, n_2)$ respectively. We can rewrite these generating functions as

$$T_1(x,y) = \prod_{i=1(2)}^{A}\frac{x^i y}{1-x^i y} = y^{A^*}x^{\sum_{i=1(2)}^{A} i}\prod_{i=1(2)}^{A}\frac{1}{1-x^i y} = y^{A^*}x^{\sum_{i=1(2)}^{A} i}D_1(x,y), \quad (4.63)$$

and

$$T_2(x,y) = \prod_{i=1}^{B}\frac{x^i y}{1-x^i y} = y^{B}x^{\sum_{i=1}^{B} i}\prod_{i=1}^{B}\frac{1}{1-x^i y} = y^{B}x^{\sum_{i=1}^{B} i}D_2(x,y), \quad (4.64)$$

where A^* and B are defined as in [4.21] and [4.20] respectively.

Clearly, $D_1(x,y)$ in [4.63] is the generating function of $p(n|[\text{odd parts}])$, and using Euler's theorem, we note that the coefficients of $D_1(x,y)$ correspond to partitions of n into m distinct parts each $\leq A$. Similarly, $D_2(x,y)$ in [4.64] is the generating function of the number of partitions of n into m parts each $\leq B$. Hence, applying the operator left-shifting (see Appendix), the coefficients of $T_1(x,y)$ and $T_2(x,y)$ can be expressed as

$$\begin{aligned}
[x^a y^b]T_1(x,y) &= [x^a y^b]y^{A^*}x^{\sum_{i=1(2)}^{A} i}D_1(x,y) && (4.65)\\
&= [x^{a-\sum_{i=1(2)}^{A} i}y^{b-A^*}]D_1(x,y) && (4.66)\\
&= q(a-\sum_{i=1(2)}^{A} i, b-A^*), && (4.67)
\end{aligned}$$

and,

$$\begin{aligned}
[x^a y^b]T_2(x,y) &= [x^a y^b]y^{B}x^{\sum_{i=1}^{B} i}D_2(x,y) && (4.68)\\
&= [x^{a-\sum_{i=1}^{B} i}y^{b-B}]D_2(x,y) && (4.69)\\
&= p(a-\sum_{i=1}^{B} i, b-B), && (4.70)
\end{aligned}$$

where A^*, B, $p(a,b)$ and $q(a,b)$ are defined as in [4.21], [4.20] and definition [3] respectively.

Thus, the functions $G_1(i,c)$ and $G_2(i,c)$ in [4.43] and [4.48] respectively, can be written as partition functions given by

$$\begin{aligned} G_1(i,c) &= \sum_{C_2=p_1}^{p_2} \sum_{n_2=B}^{n-i+B} [x^{ci+2C_2} y^{n-n_2}] T_1(x,y) [x^{C_2} y^{n_2}] T_2(x,y) \\ &= \sum_{C_2=p_1}^{p_2} \sum_{n_2=B}^{n-i+B} q(ci + 2C_2 - \sum_{i=1(2)}^{A} i, n - n_2 - A^*) p(C_2 - \sum_{i=1}^{B} i, n_2 - B), \end{aligned}$$

and

$$\begin{aligned} G_2(i,c) &= \sum_{C_2=p_1}^{p_2} \sum_{n_2=B}^{n-i+B} [x^{2C_2-ci} y^{n-n_2}] T_1(x,y) [x^{C_2} y^{n_2}] T_2(x,y) \\ &= \sum_{C_2=p_1}^{p_2} \sum_{n_2=B}^{n-i+B} q(2C_2 - ci - \sum_{i=1(2)}^{A} i, n - n_2 - A^*) p(C_2 - \sum_{i=1}^{B} i, n_2 - B), \end{aligned}$$

where $p(a,b)$ and $q(a,b)$ are defined as in Definition [3].

Therefore, the probability distribution function of the statistic C in Theorem [11] can be calculated as a function of partitions and, as mentioned above, the theory of partitions has been extensively studied and some mathematical software include modules that facilitate the calculation of these functions. Likewise, elementary properties of partitions can be easily implemented to facilitate the calculation of $Pr[C = c]$.

Chapter 5
The Distribution of C under the Null Hypothesis

Statistics with a distribution-free property (Randles and Wolfe (1979)) provide assurance that the testing procedure maintains the designated α-level over a wide variety of distributional assumptions. In this chapter, the probability distribution function of the statistic C under the null hypothesis is determined using the generating functions seen in the previous chapter. The statistic C has the distribution-free property, this means, among other things, that the level of significance, say, α, for such a test is constant over some collection φ of possible joint distributions where $F \in \varphi$; that is, the probability of a type I error is α for any underlying joint distribution that belong to φ. As an example, the function is analyzed for some sample sizes in order to observe their behavior. The critical values for sample sizes between 4 and 30 have been calculated and are presented in the Appendix. In the same way, an algorithm that facilitates the calculations is presented in the Appendix. Moreover, using a class of discrete probability distributions built by Voinov and Nikulin (1997), we present an alternative approach to the determination of the distribution C under the null hypothesis.

Under the null hypothesis $H_0 : \theta = 0$, every sequence of the $n_1 + n_2 = n$ objects is equiprobable because

$$Pr[\eta_i = 1] = Pr[\eta_i = 0] = \frac{1}{2},\ i = 1, \ldots, n. \tag{5.1}$$

In other words, $Pr[X > 0] = Pr[X \leq 0] = 1/2$.

Mcwilliams (1990)[1]. showed that under the null hypothesis $H_0 : \theta = 0$, the variable $r_n - 1$, where r_n is the number of runs in the sequence $\{\eta_i\}$, is binomial with parameters $n - 1$ and $1/2$. This result is easily testable using the probability

[1]MacWilliams introduced the probability distribution of r_n under H_0, but in theorem [5] in the previous chapter, the result is more general, since it shows the probability distribution of r_n for any $p = Pr[X > 0]$

distribution in Theorem [5] with $p = 1/2$. The following theorem shows this result.

Theorem 13 *The probability distribution of r_n, the total number of runs of a specific $n = n_1 + n_2$ objects, n_1 of type 1 and n_2 of type 2 under the null hypothesis, in a random sample is given by*

$$Pr_{H_0}[r_n = i] = \binom{n-1}{i-1}\left(\frac{1}{2}\right)^{n-1},\ i = 1, \ldots, n. \tag{5.2}$$

$Proof$. Mcwilliams (1990) or Theorem [5] with $p = 1/2$.

■

Now, using the theorems [13] and [4], we can easily define the probability distribution function of the statistic C under the null hypothesis $H_0 : \theta = 0$.

Theorem 14 *The probability distribution function of the statistic C under the null hypothesis $H_0 : \theta = 0$, is given by*

$$Pr_{H_0}[C = c] = \frac{1}{2^n}\left[\sum_{i=1}^{n} G_1(i,c) + G_2(i,c)\right], -n \leq c \leq n, \tag{5.3}$$

where $G_1(i,c)$ and $G_2(i,c)$ are defined by [4.43] and [4.48] respectively.

$Proof$. Replacing [5.1] and [5.2] in theorem [11], it follows directly that

$$\begin{aligned} Pr_{H_0}[C = c] &= \sum_{i=1}^{n}\left[\binom{n-1}{i-1}\left(\frac{1}{2}\right)^{n-1}\right]\left[\frac{G_1(i,c)}{\binom{n-1}{i-1}}\left(\frac{1}{2}\right) + \frac{G_2(i,c)}{\binom{n-1}{i-1}}\left(\frac{1}{2}\right)\right] \\ &= \frac{1}{2^n}\left[\sum_{i=1}^{n} G_1(i,c) + G_2(i,c)\right], -n \leq c \leq n. \end{aligned}$$

This completes the proof.

■

Note that the Distribution of C under the null hypothesis $H_0 : \theta = 0$, does not depend on which $F \in \Omega_0$ we are sampling from, and the critical values can be found without knowing F. It is in this sense that we say C is distribution free under $H_0 : \theta = 0$.

An algorithm is presented in the Appendix to facilitate the calculations. Through this algorithm, we determine the critical values $P_{H_0}\left(C \geq c_{1-\alpha/2}\right) = \alpha$, for sample

sizes between 4 and 30, the respective tables with critical values are presented in the Appendix. Similarly, we provide some numerical results for the distribution of C under the null hypothesis described in this chapter in order to illustrate the theoretical results. Table [5.1] gives the probability distribution function and the exact distribution function of Statistic C under the null hypothesis $H_0 : \theta = 0$ for $n = 5$ and a histogram, a graphical representation of Table [5.1], can be seen in Figure [5.1].

Values of C	$Pr[C = c]$	$F(c)$
-5,000	0,0313	0,0313
-3,500	0,0313	0,0625
-2,667	0,0313	0,0938
-2,000	0,0625	0,1563
-1,500	0,0313	0,1875
-1,333	0,0313	0,2188
-1,000	0,0938	0,3125
-0,667	0,0313	0,3438
-0,600	0,0313	0,3750
-0,500	0,0313	0,4063
-0,333	0,0313	0,4375
-0,250	0,0625	0,5000
0,250	0,0625	0,5625
0,333	0,0313	0,5938
0,500	0,0313	0,6250
0,600	0,0313	0,6563
0,667	0,0313	0,6875
1,000	0,0938	0,7813
1,333	0,0313	0,8125
1,500	0,0313	0,8438
2,000	0,0625	0,9063
2,667	0,0313	0,9375
3,500	0,0313	0,9688
5,000	0,0313	1,0000

Table 5.1: The Probability Distribution Function and the Distribution of Statistic C under the null hypothesis $H_0 : \theta = 0$ for $n = 5$.

For small sample sizes, the probability obtained in [5.3] is easily verifiable. It is sufficient to generate the 2^n possible arrangements of ones and zeros and calculate the statistic in each. These 2^n arrangements are equiprobable under the

null hypothesis, then the probability of C can be calculated as

$$Pr_{H_0}[C = c] = \frac{\text{Number of arrangements where the statistic is } c}{2^n}, -n \leq c \leq n. \tag{5.4}$$

For example, if $n = 5$, we generate the $2^5 = 32$ possible arrangements of ones and zeros, and for each of them we calculate the value of the statistic C. The results can be seen in the Appendix and using these probabilities we can verify the values obtained in Table [5.1].

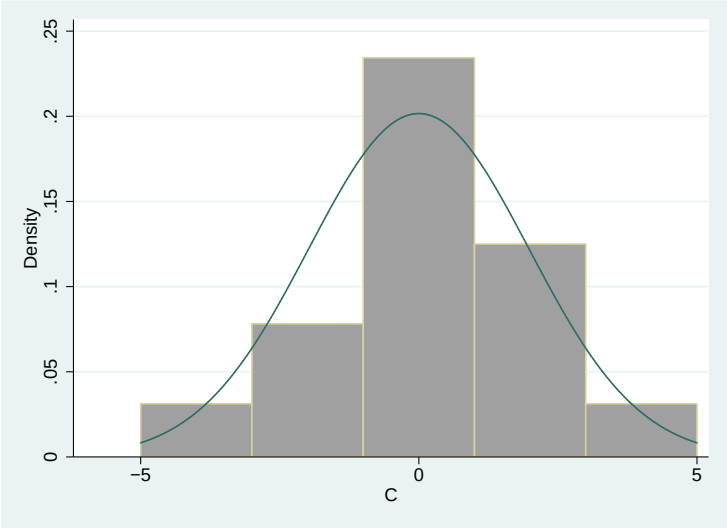

Figure 5.1: Histogram for C and Normal Curve ($n = 5$)

Mean	0,00
Median	-
Std. Deviation	1,98
Variance	3,91
Skewness	-
Kurtosis	1,13
Percentiles	
1	- 5,00
5	- 4,03
10	- 2,47
25	- 1,00

Table 5.2: Basic Statistics for the Statistic C, $n = 5$

Table [5.2] also presents some basic statistics for the statistic C. Some of the properties discussed in Chapter 3 can be easily observed in in these tables,

as that the expected value is zero and that the statistic is symmetric about zero to mention a few. Similarly, statistics for sample sizes 10, 15, 20, 25 and 30 are presented as an Appendix, with the respective histograms that show the behavior of the statistic C. Here it is visually observed that the asymptotic distribution of the the statistic C is the normal distribution, this approximation will be discussed in the next chapter.

5.1 An Additional Approach

Combinatorics and applications of combinatorial methods in probability and statistics has become a very active and fertile area of research in the recent past. Many authors currently use combinatorial models as a solution to problems in various areas. Models based on urns and partitions are very important in cryptography, programming and genetics because these models are similar to probability models in these areas.

Problems relating to partitions of integers have been considered by Voinov and Nikulin (1994, 1995, 1996). In Chapter 11 of Balakrishnan (1997), an algorithm has been developed for constructing partitions of an integer by arbitrary positive integers. This algorithm helps in introducing a class of discrete probability distributions which are useful in determining the distribution of C under the null hypothesis. Below we present the class of discrete probability distributions built by Voinov and Nikulin (1997).

Suppose that an urn contains balls. The balls bear fixed positive numbers $a_1, a_2, \ldots, a_l$, $l \in Z^+$. Let p_i be the probability that a ball bearing the number a_i will be drawn $(i = 1, 2, \ldots, l)$ with $\sum_{i=1}^{l} p_i = 1$. Let the random variable X take the value r if, of n balls drawn with replacement, r_1 bear the number a_1, r_2 bear the number a_2, and so on, and $\sum_{i=1}^{l} a_i r_i = r$ with $\sum_{i=1}^{l} r_i = n$. The probability that the summation of numbers on ball drawn r is given in the following theorem.

Theorem 15 *The probability that* $X = \sum_{i=1}^{l} a_i r_i = r$ *with* $\sum_{i=1}^{l} r_i = n$ *is given by*

$$Pr[X = r] = \sum_{\sum_{i=1}^{l} a_i r_i = r} \binom{n}{r_1, r_2, \ldots, r_l} \prod_{i=1}^{l} p_i^{r_i}, \tag{5.5}$$

where

$$\binom{n}{r_1, r_2, \ldots, r_l} = \frac{n!}{r_1! \ldots r_{l-1}! (n - \sum_{i=1}^{l-1} r_i)!}, \tag{5.6}$$

and is zero if $\sum_{i=1}^{l} r_i > n$, $(n\tilde{a}_1 \leq r \leq n\tilde{a}_2)$, $\tilde{a}_1 = min_{1\leq i\leq l}\{a_i\}$, *and* $\tilde{a}_2 = max_{1\leq i\leq l}\{a_i\}$.

$Proof$. See Balakrishnan (1997).

■

This class of discrete probability distributions can be used in the determination of the probability distribution function of the statistic C under the null hypothesis. In this section, we will investigate the connection between the distribution of C and X.
From [4.51] and theorems [11] and [13] we have

$$\begin{aligned} Pr_{H_0}[C=c] &= \sum_{i=1}^{n} Pr[r_n=i]\Bigg[Pr\Big[\sum_{j=1}^{i}(-1)^{j+1}jL_j=ci\Big]Pr[\eta_1=1] \\ &\quad +Pr\Big[\sum_{j=1}^{i}(-1)^{j}jL_j=ci\Big]Pr[\eta_1=0]\Bigg] \\ &= \sum_{i=1}^{n}\binom{n-1}{i-1}\left(\frac{1}{2}\right)^{n-1}\Bigg[Pr\Big[\sum_{j=1}^{i}(-1)^{j+1}jL_j=ci\Big]\left(\frac{1}{2}\right) \\ &\quad +Pr\Big[\sum_{j=1}^{i}(-1)^{j}jL_j=ci\Big]\left(\frac{1}{2}\right)\Bigg] \\ &= \left(\frac{1}{2}\right)^{n}\sum_{i=1}^{n}\binom{n-1}{i-1}\Bigg[Pr\Big[\sum_{j=1}^{i}(-1)^{j+1}jL_j=ci\Big] \\ &\quad +Pr\Big[\sum_{j=1}^{i}(-1)^{j}jL_j=ci\Big]\Bigg]. \end{aligned}$$

Using [4.18] and [4.46], we get that

$$Pr_{H_0}\Big[\sum_{j=1}^{i}(-1)^{j+1}jL_j=ci\Big]=Pr_{H_0}\Big[\sum_{j=1(2)}^{A}jL_j-2\sum_{j=1}^{B}jL_{2j}=ci\Big], \quad (5.7)$$

and

$$Pr_{H_0}\Big[\sum_{j=1}^{i}(-1)^{j}jL_j=ci\Big]=Pr_{H_0}\Big[-\sum_{j=1(2)}^{A}jL_j+2\sum_{j=1}^{B}jL_{2j}=ci\Big]. \quad (5.8)$$

The probability in [5.7][2] may be written using the random variables C_1 and C_2 from [4.22- 4.25] as

[2]For [5.8], the analysis is similar

$$Pr_{H_0}\Big[\sum_{j=1(2)}^{A} jL_j - 2\sum_{j=1}^{B} jL_{2j} = ci\Big] = Pr_{H_0}[C_1 - 2C_2 = ci], \qquad (5.9)$$

where $\sum_{j=1(2)}^{A} jL_j = C_1$ with $\sum_{j=1(2)}^{A} L_j = n_1 = n - n_2$, and $\sum_{j=1}^{B} jL_{2j} = C_2$ with $\sum_{j=1}^{B} L_{2j} = n_2$.

Evidently the variables C_1 and C_2 depend on n_2, where n_2 denotes the number of zeros in a sequence with i runs and from [4.42] we know that $B \leq n_2 \leq n - i + B$. By definition of n_2 and applying the same argument used in [4.7], under the null hypothesis $n_2 - B$ is binomial with parameters $n - i$, $p = 1/2$. From this, and using the theorem of total probability, we can write [5.9] as

$$\begin{aligned} Pr_{H_0}[C_1 - 2C_2 = ci] &= \sum_{n_2=B}^{n-i+B} Pr_{H_0}[C_1 - 2C_2 = ci|n_2]Pr[n_2] \\ &= \sum_{n_2=B}^{n-i+B} Pr_{H_0}[C_1 - 2C_2 = ci|n_2]\binom{n-i}{n_2-B}(1/2)^{n-i} \\ &= (1/2)^{n-i}\sum_{n_2=B}^{n-i+B} Pr_{H_0}[C_{1,n_2} - 2C_{2,n_2} = ci]\binom{n-i}{n_2-B}, \end{aligned} \qquad (5.10)$$

where $\sum_{j=1(2)}^{A} jL_j = C_{1,n_2}$ and $\sum_{j=1}^{B} jL_{2j} = C_{2,n_2}$ with n_2 fixed.

Hence, the probability of the statistic C may be written using [5.10] in the form

$$\begin{aligned} Pr_{H_0}[C = c] &= \left(\frac{1}{2}\right)^n \sum_{i=1}^{n}\binom{n-1}{i-1}\left(\frac{1}{2}\right)^{n-i}\Bigg[\sum_{n_2=B}^{n-i+B}\binom{n-i}{n_2-B} \\ &\quad Pr_{H_0}[C_{1,n_2} - 2C_{2,n_2} = ci] + Pr_{H_0}[-C_{1,n_2} + 2C_{2,n_2} = ci]\Bigg]. \end{aligned} \qquad (5.11)$$

For n_2 fixed, the variables C_{1,n_2} and C_{2,n_2} follow the distribution given in the theorem [15]. The following theorems show these distributions.

Theorem 16 *Let the random variable $C_{1,n_2} = \sum_{j=1(2)}^{A} jL_j$ with $\sum_{j=1(2)}^{A} L_j = n - n_2$ and L_j, $j = 1, \ldots, A$, be positive integers. The probability distribution of*

C_{1,n_2} is given by

$$Pr[C_{1,n_2} = r] = \sum_{\sum_{j=1(2)}^{A} jL_j = r - \sum_{j=1(2)}^{A} j} \binom{n - n_2 - A^*}{L_1, L_3, \ldots, L_A} \prod_{j=1}^{A} p_i^{L_i} \tag{5.12}$$

$$= \sum_{\sum_{j=1(2)}^{A} jL_j = r - \sum_{j=1(2)}^{A} j} \binom{n - n_2 - A^*}{L_1, L_3, \ldots, L_A} \left(\frac{1}{A^*}\right)^{\sum_{j=1(2)}^{A} L_j}, \tag{5.13}$$

where

$$\binom{n - n_2 - A^*}{L_1, L_3, \ldots, L_A} = \frac{(n - n_2 - A^*)!}{L_1! \ldots (n - \sum_{j=1(2)}^{A-1} L_j)!}, \tag{5.14}$$

A^* and A are defined as in [4.21] and [4.19] respectively.

Proof. According to the definition of C_{1,n_2}, we consider the problem of representing a positive integer r as a sum of the positive integers $\{1, 3, \ldots, A\}$. In other words, we would like to consider all integral representations of $n - n_2$ as

$$L_1 + 3L_3 + \ldots + AL_A = r, \tag{5.15}$$

where $L_1 + L_3 + \ldots + L_A = n - n_2$ and L_j, $j = 1, \ldots, A$, are positive integers.

Thus, making some considerations, the probability model in theorem [15] applies to describe this problem. The model can be used if we ensure the following considerations:

- In theorem [15] we assumed that $L_i \geq 0$, but we need that $L_i \geq 1$ (Theorem [16]). Then if we assume that each L_i [3] must appear at least once, the problem reduces to distribute the $n - n_2 - A^*$ remaining elements, i.e. we require $L_1^*, L_3^*, \ldots, L_A^*$ such that

$$r = L_1 + 3L_3 + \ldots + AL_A \tag{5.16}$$

$$= (L_1^* + 3L_3^* + \ldots + AL_A^*) + (L_1 + L_3 + \ldots + L_A) \tag{5.17}$$

$$= (L_1^* + 3L_3^* + \ldots + AL_A^*) + \sum_{j=1}^{A} L_j, \tag{5.18}$$

  where $L_1^* + L_3^* + \ldots + L_A^* = n - n_2 - A^*$, and L_j^*, $j = 1, \ldots, A$, are non-negative integers.

[3] From [4.21], we know that the number of terms of $\sum_{j=1(2)}^{A} jL_j$ is A^*.

Evidently, using theorem [15] we have

$$\begin{aligned} Pr[C_{1,n_2} = r] &= Pr[C^*_{1,n_2} = r - \sum_{j=1(2)}^{A} j] \\ &= \sum_{\sum_{j=1(2)}^{A} jL_j = r - \sum_{j=1(2)}^{A} j} \binom{n - n_2 - A^*}{L_1^*, L_3^*, \ldots, L_A^*} \prod_{j=1}^{A} p_i^{L_i}, \end{aligned}$$

where $C^*_{1,n_2} = \sum_{j=1(2)}^{A} jL_j^*$ with $L_1^* + L_3^* + \ldots + L_A^* = n - n_2 - A^*$, L_j^*, $j = 1, \ldots, A$, are non-negative integers and A^* is defined as in [4.21].

- In theorem [15], p_i is defined as the probability that a ball bearing the number i, which in our case means that there is a zero in a certain position in the sequence dichotomized $\{\vec{\eta}\}$. Under the null hypothesis $H_0 : \theta = 0$, every sequence of the $n_1 + n_2 = n$ objects is equiprobable because $Pr[\eta_i = 1] = Pr[\eta_i = 0] = 1/2$, then it is easy to verify that the p_i are also equiprobable with $p_i = 1/A^*$, where A^* denotes the number of terms of C_{1,n_2}.

This completes the proof.

■

Theorem 17 *Let the random variable $C_{2,n_2} = \sum_{j=1}^{B} jL_{2j}$ with $\sum_{j=1}^{B} L_{2j} = n_2$ and L_{2j}, $j = 1, \ldots, B$, positive integers. The probability distribution of C_{2,n_2} is given by*

$$Pr[C_{2,n_2} = r] = \sum_{\sum_{j=1}^{B} jL_{2j} = r - \sum_{j=1}^{B} j} \binom{n_2 - B}{L_2, L_4, \ldots, L_{2B}} \prod_{i=1}^{B} p_i^{L_{2i}} \tag{5.19}$$

$$= \sum_{\sum_{j=1}^{B} jL_{2j} = r - \sum_{j=1}^{B} j} \binom{n_2 - B}{L_2, L_4, \ldots, L_{2B}} \left(\frac{1}{B}\right)^{\sum_{j=1}^{B} L_{2j}}, \tag{5.20}$$

where

$$\binom{n_2 - B}{L_2, L_4, \ldots, L_{2B}} = \frac{(n_2 - B)!}{L_2! \ldots (n - \sum_{j=1}^{B-1} L_{2j})!}, \tag{5.21}$$

and B is defined as in [4.20].

$Proof.$ As in the preceding proof, using theorem [15] we have

$$\begin{aligned} Pr[C_{2,n_2} = r] &= Pr[C^*_{2,n_2} = r - \sum_{j=1}^{B} j] \\ &= \sum_{\sum_{j=1}^{B} jL_{2B} = r - \sum_{j=1}^{B} j} \binom{n_2 - B}{L^*_2, L^*_4, \ldots, L^*_{2B}} \prod_{j=1}^{B} p_i^{L_{2B}}, \end{aligned}$$

where $C^*_{2,n_2} = \sum_{j=1}^{B} jL^*_{2B}$ with $L^*_2 + L^*_4 + \ldots + L^*_{2B} = n_2 - B$, L^*_j, $j = 1, \ldots, B$, are non-negative integers, $p = 1/B$ and B is defined as in [4.20]. This completes the proof.

■

Thus, combining the distribution of C_{1,n_2} and C_{2,n_2} in theorems [16] and [17] respectively, we can determine [5.11].

Chapter 6
The Power Function of C

As discussed in the previous chapter, the distribution-free property provides assurance that the testing procedure maintains the designated α-level over a wide variety of distributional assumptions. A second and equally important consideration for testing procedures is their effectiveness in detecting alternative hypotheses. This property is naturally analyzed through the power function. Certain comparisons among the tests can be made using this function. The power function definition is given below (see Randles and Wolfe (1979)).

Definition 4 *Suppose that a model is indexed by parameter(s) ξ. The power function of a test of hypothesis relevant to this model is given by*

$$Pw[\xi] = P_{\xi}[\textit{the null hypothesis is rejected}]$$

Under the null hypothesis $H_0 : \theta = 0$, we know that $p = Pr[\eta_i = 1] = Pr[\eta_i = 0] = \frac{1}{2}$, or simply $p = Pr[X > 0] = Pr[X \leq 0] = 1/2$. More generally, for any $\theta > 0$, the probability distribution function of C in Theorem [11] depends on the underlying distribution $F(x)$, because $p = Pr[\eta_i = 1] \neq \frac{1}{2}$. Also, if the rule rejects whenever $\theta > 0$, the power function of the statistic C is given in the following theorem.

Theorem 18 *The power function of the statistic C is given by*

$$Pw[\theta, p] = \sum_{i=1}^{n} Pr[r_n = i] \left[\frac{G_1(i,c)}{\binom{n-1}{i-1}}[p] + \frac{G_2(i,c)}{\binom{n-1}{i-1}}[1-p] \right], \qquad (6.1)$$

where $Pr[r_n = i]$ are defined by theorem [5], $G_1(i,c)$ and $G_2(i,c)$ by [4.43] and [4.48] respectively.

$Proof$. From Theorem [11] and define $p = Pr[\eta_i = 1] = Pr[X > 0]$ follows directly [6.1].
This completes the proof.

■

Now we compare the power of the test proposed C and its direct competitors, i.e. the sign test (S) and the Wilcoxon signed rank test (W)[1]. The powers of the sign test and the Wilcoxon signed rank test were presented in Chapter 2, however the power comparisons between these tests are not easily made. For example for the Wilcoxon signed rank test the form of the power function for nonnormal distributions is even more complex. A representation of each rank configuration under alternatives must be done, but the power function simplifies in only a few special cases (see Randles and Wolfe (1979)). In order to examine and relate their power properties, we use a simulation study.

In the simulation study we have analyzed the performance of the proposed test C compared with the sign test (S) and the Wilcoxon signed rank test (W). The hypothesis under consideration is given by (see 2.1)

$$H_0 : \theta = 0 \qquad versus \qquad H_1 : \theta > 0.$$

The study describes attainment of the nominal size and the power of the tests. We take sample sizes 10, 15 20, 25 and 30, and the nominal significance level of the test α, was 0.05. Randomization was used for the tests so that all three tests have a nominal $\alpha = 0.05$ level. We used the standard normal (N(0,1)), the Cauchy distribution (C(0,1)), the Uniform distribution (U(0,1)), the Laplace distribution (L(0,1)) and the logistic distribution (Log(0,1)) to generate random samples for this study [2]. We analyzed the power of the tests at $\theta = q_{0.5}, q_{0.55}, q_{0.6}, q_{0.65}, q_{0.7}$, where q_i is the ith quantile of the distribution. This range was chosen, because in this interval the power of the test is more sensitive to variations and differences can be seen with more clarity. Moreover, the results are easily comparable with the results obtained by the authors presented in Chapter 2.

The method is not complicated, we simulate the selection of random variables using a SAS-Macro (see appendix) and the power of the test is computed as the relative frequency with which a particular test rejects the null hypothesis H_0. The results of the simulation study involving 5000 replications of the random sampling process for $n = 10$ are presented in Table [6.1]. These results are graphed in Figure [6.1].

[1]The adaptive approaches in chapter 2, are a combination of the sign test and the Wilcoxon signed rank test, for this reason the comparison is done directly between these tests.

[2]Randles and Wolfe (1979) show a similar Monte Carlo study, they used the same distributions here analyzed to compare the powers of the t, S and Wilcoxon tests.

Distribution	q	C	S	W
N(0,1)	0,50	0,0554	0,0487	0,0510
	0,55	0,1036	0,0641	0,0988
	0,60	0,1828	0,0911	0,1704
	0,65	0,3150	0,1305	0,2828
	0,70	0,4512	0,1907	0,4154
C(0,1)	0,50	0,0522	0,0489	0,0502
	0,55	0,0848	0,0609	0,0794
	0,60	0,1374	0,0833	0,1286
	0,65	0,2096	0,1239	0,1978
	0,70	0,2936	0,1873	0,2932
U(0,1)	0,50	0,0532	0,0495	0,0486
	0,55	0,1244	0,0623	0,1076
	0,60	0,2506	0,0873	0,2110
	0,65	0,4132	0,1269	0,3580
	0,70	0,6048	0,1859	0,5440
L(0,1)	0.50	0,0490	0,0507	0,0490
	0.55	0,0838	0,0615	0,0758
	0.60	0,1496	0,0841	0,1316
	0.65	0,2352	0,1243	0,2210
	0.70	0,3452	0,1937	0,3290
Log(0,1)	0.50	0,0532	0,0495	0,0486
	0.55	0,0982	0,0623	0,0892
	0.60	0,1778	0,0873	0,1594
	0.65	0,2894	0,1269	0,2614
	0.70	0,4154	0,1859	0,3882

Table 6.1: Power of the proposed test (C), the sign test (S) and the Wilcoxon signed rank test (W), for $n = 10$

Certain comparisons among the tests can be made using Tables [6.1], [G.1], [G.2], [G.3] and [G.4], where we present the powers of the tests for the sample sizes and distributions selected. For the normal distribution and the logistic distribution, we observed that the proposed test C is better than the W test and the S test for $n \leq 15$, for $n > 15$ the differences between the C test and the W test are minimal (≈ 0.02), although the W test is slightly better than the C test. The S test remains in third place.

When the underlying distribution is Cauchy, it can be seen that the proposed test C is better than the W test and the S test for $n = 10$. For $n = 15$ and $n = 20$

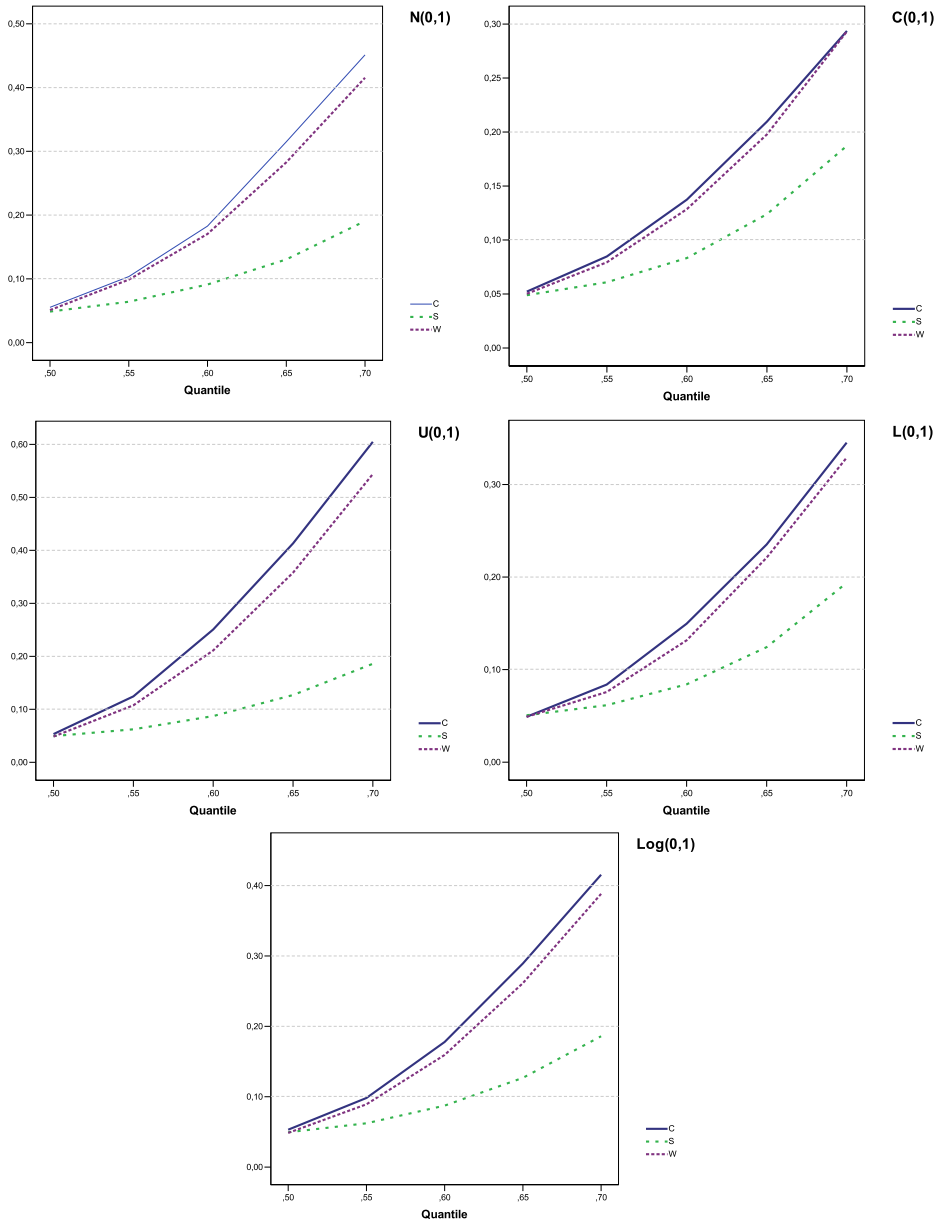

Figure 6.1: Power of the proposed test (C), the sign test (S) and the Wilcoxon signed rank test (W), for $n = 10$

the differences between the C test and the W test are minimal, although the W test is slightly better than the C test. The S test remains in third place. However, for $n = 25$, the three tests have similar values, although the C test and W test are only slightly better than the S test, but for $n = 30$ the S is better than the C test

and W test.

For the uniform distribution, the proposed test C is better than the W test and the S test. Both the C and W tests stay ahead at the S test, which remains in third place.

We observe that if the underlying distribution is Laplace, the proposed test C is better than the W test and the S test for $n < 15$. For $15 < n < 30$ the differences between the C test and the W test are minimal, although the W test is slightly better than the C test, while the S test remains in third place. However, for $n = 30$, the three tests have similar values, although the S test is slightly better than the C test and the W test.

In conclusion and in agreement with the simulation study, for $n < 20$ the proposed test C has high power compared to the other competitors referred in this dissertation. Generally, for $n > 20$ the three tests have similar values, although the C test and the W test have are more powerful than the S test, which remains in third place.

Chapter 7
Asymptotic Normality

The distribution of the statistic C under both null and alternative hypotheses is given in chapters 5 and 6. Under $H_0 : \theta = 0$, the distribution of the statistic C does not depend on the distribution sampled, that is, it is distribution free. We have several alternatives to determine the limiting distribution of statistics C. In Chapter 4 we present a way to describe the distribution of the statistic C using functions of partitions. We can use this alternative approach by applying different limit theorems for such functions, for example Richmond and Knopfmacher (1995) showed that the variable $C(n,k)/C(n)$, where $C(n,k)$ is the number of compositions of a positive integer n into k distinct parts and $C(n) = \sum_k C(n,k)$, tend to a normal distribution. It is clear that $C(n,k) = k!q(n,k)$ where $q(n,k)$ is the number of partitions of n into k distinct parts (see definition [3]). Thus, the random variable $q(n,k)$ has asymptotically normal distribution with parameters $[k_1/k!, C(n)/k!\sqrt{2m/logm}]$ (see Richmond and Knopfmacher (1995) and Szekeres (1953)).

Another result of great importance was presented by Goh and Schmutz (1995). They proved a central limit theorem for the number of different part sizes in a random integer partition. Then, if λ is one of the $P(n)$ partitions of the integer n, let $D_n(\lambda)$ be the number of distinct part sizes that λ has, we have that

$$\frac{\#\{\lambda : D_n(\lambda) \leq A_n + xB_n\}}{P(n)} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{7.1}$$

Thus, using the relations given in [4.58], the asymptotic distribution of the partition functions in chapter 4 can be determined.

In Chapter 5 we present an additional approach for the distribution of C under H_0 using a class of discrete probability distribution built by Voinov and Nikulin (1997). Nikulin, Smirnov and Voinov (2002) analyzed also a class of multivariate discrete distributions induced by an urn containing balls marked by vectors with arbitrary integer components. They propose an algorithm for probabilities

evaluation based on a solution of a system of two linear diophantine equations in nonnegative integers. Though our interest lies in the univariate case, we present the theorems for the multivariate case to show a more general problem. Naturally, the univariate case is a direct consequence of the multivariate case. Below we present the model proposed by these authors, see (Nikulin, Smirnov and Voinov (1999)).

Let a random vector $X = (X_1, X_2, \ldots, X_m)^T$ take the value $r = (r_1, r_2, \ldots, r_m)^T$ if sums of numbers of jth components of vector on n balls drawn with replacement are

$$\sum_{i=1}^{K} a_{ij} l_{a_i} = r_j,$$
$$n \min_{1 \le i \le k_j} \{a_{ji}\} \le r_j \le n \max_{1 \le i \le k_j} \{a_{ji}\}, j = 1, 2 \ldots, m,$$

where l_{a_i}, $i = 1, \ldots, K$, are nonnegative integers and denote the number of balls in the sample bearing the vector a_i such that

$$\sum_{i=1}^{K} l_{a_i} = n.$$

Evidently,

$$Pr[X = r, p] = \sum \frac{n!}{\prod_{i=1}^{K} l_{a_i}} \prod_{i=1}^{K} p_{a_i}^{l_{a_i}}, \tag{7.2}$$

where $p = (p_{a_1}, \ldots, p_{a_K})^T$ is the vector of parameters and the summation is performed over all sets of nonnegative solutions l_{a_i}, $i = 1, \ldots, K$ of the system of linear diophantine equations

$$\begin{cases} \sum_{i=1}^{K} a_{ij} l_{a_i} = r_j, j = 1, 2 \ldots, m, \\ \sum_{i=1}^{K} l_{a_i} = n. \end{cases} \tag{7.3}$$

We can clearly observe the relationship between [7.2] and the probability function given in Theorem [15].

From [7.2] we see that in order to evaluate probabilities, we have to solve the system of equations [7.3] in nonnegative integers or to construct all vector partitions of r with exactly n parts. Many algorithms can be used for a solution of system [7.3]. One of these algorithms was developed by Nikulin, Smirnov and Voinov. The algorithm and the full development of the model can be seen in (Nikulin, Smirnov and Voinov (2002)). We will use this result in the case of the

random variable defined in Theorem [15] and thus determine its limit distribution.

The following theorem shows the limit distribution for a class of discrete probability distributions built by Voinov and Nikulin (1997), the one presented in Chapter 5 (see Theorem [15]).

Theorem 19 *For $n \to \infty$, the random variable X with probabilities p_i, $\sum_{i=1}^{l} p_i = 1$ has asymptotically the normal distribution with parameters*

$$\mu = n \sum_{i=1}^{l} a_i p_i, \tag{7.4}$$

$$\sigma^2 = \sum_{i=1}^{l} a_i^2 (p_i - p_i^2) - 2 \sum_{i \neq j}^{l} \sum a_i a_j p_i p_j. \tag{7.5}$$

$Proof$. see Nikulin, Smirnov and Voinov (2002)

■

In Chapter 5, we use the distribution of X (Theorem [15]) to determine the probability distributions of $C_{1,n_2} = \sum_{j=1(2)}^{A} jL_j$ and $C_{2,n_2} = \sum_{j=1}^{B} jL_{2j}$ (Theorems [16] and [17]). Now, using Theorem [19], we can determine the limit distribution for these variables. The following theorems are a direct consequence of Theorem [19].

Theorem 20 *For $n \to \infty$, the random variable $C_{1,n_2} = \sum_{j=1(2)}^{A} jL_j$ with $\sum_{j=1(2)}^{A} L_j = n - n_2 = n_1$ and L_j, $j = 1, \ldots, A$, positive integers, has an asymptotic normal distribution with parameters*

$$\mu_{C_{1,n_2}} = (n - n_2 - A^*) A^* \sum_{j=1(2)}^{A} j, \tag{7.6}$$

$$\sigma^2_{C_{1,n_2}} = \frac{n - n_2 - A^*}{A^{*2}} \Big[(A^* - 1) \sum_{j=1(2)}^{A} j - 2 \sum_{\substack{i \neq j \\ i=1(2) \\ j=1(2)}}^{A} \sum ij \Big], \tag{7.7}$$

where A^* and A are defined in [4.21] and [4.19] respectively .

$Proof$. Using Theorem [16] we see that, through the distribution of X, we can determine the distribution of C_{1,n_2}. Taking into account Theorem [19], we obtain directly the limit distribution given in Theorem [20]. This completes the proof.

■

Theorem 21 *For $n \to \infty$, $C_{2,n_2} = \sum_{j=1}^{B} jL_{2j}$ with $\sum_{j=1}^{B} L_{2j} = n_2$ and L_{2j}, $j = 1, \ldots, B$, positive integers, has an asymptotic normal distribution with parameters*

$$\mu_{C_{2,n_2}} = (n_2 - B)B \sum_{j=1(2)}^{B} j, \tag{7.8}$$

$$\sigma^2_{C_{2,n_2}} = \frac{n_2 - B}{B^2}\Big[(B-1)\sum_{j=1}^{B} j - 2\sum_{\substack{i\neq j \\ i=1 \\ j=1}}^{B}\sum ij\Big], \tag{7.9}$$

where B is defined as in [4.20].

$Proof.$ As in the preceding proof, using Theorem [17] we see that, through the distribution of X, we can determine the distribution of C_{2,n_2}, and by Theorem [19], we obtain directly Theorem [21].

This completes the proof.

■

Now, from [5.10] we have a linear combination of random variables C_{1,n_2} and C_{2,n_2} given by

$$Pr_{H_0}[C_1 - 2C_2 = ci] = (1/2)^{n-i} \sum_{n_2=B}^{n-i+B} Pr_{H_0}[C_{1,n_2} - 2C_{2,n_2} = ci]\binom{n-i}{n_2 - B},$$

and

$$Pr_{H_0}[-C_1 + 2C_2 = ci] = (1/2)^{n-i} \sum_{n_2=B}^{n-i+B} Pr_{H_0}[-C_{1,n_2} + 2C_{2,n_2} = ci]\binom{n-i}{n_2 - B},$$

then, we can define the random variables $Y^{(1)}_{i,n_2}$ and $Y^{(2)}_{i,n_2}$ as

$$\begin{aligned} Y^{(1)}_{i,n_2} &= (1/i)(C_{1,n_2} - 2C_{2,n_2}), \\ Y^{(2)}_{i,n_2} &= (1/i)(-C_{1,n_2} + 2C_{2,n_2}), \end{aligned} \tag{7.10}$$

and using Theorem [21] and the fact that the weighted sum of two independent normally distributed random variables has a normal distribution, we have that the limit distributions for $Y^{(1)}_{i,n_2}$ and Y^{2}_{i,n_2} are given by

$$Y^{(1)}_{i,n_2} \to N\big[(1/i)(\mu_{C_{1,n_2}} - 2\mu_{C_{2,n_2}}), (1/i^2)(\sigma^2_{C_{1,n_2}} + 4\sigma^2_{C_{2,n_2}})\big], \tag{7.11}$$

and

$$Y_{i,n_2}^{(2)} \rightarrow N\big[(1/i)(-\mu_{C_{1,n_2}} + 2\mu_{C_{2,n_2}}), (1/i^2)(\sigma^2_{C_{1,n_2}} + 4\sigma^2_{C_{2,n_2}})\big]. \tag{7.12}$$

Using [7.10], we can make use of the fact that $Y_{i,n_2}^{(1)} = -Y_{i,n_2}^{(2)}$ to derive the expression for the covariance between $Y_{i,n_2}^{(1)}$ and $Y_{i,n_2}^{(2)}$, getting thus

$$\begin{aligned} Cov(Y_{i,n_2}^{(1)}, Y_{i,n_2}^{(2)}) &= E\big[(Y_{i,n_2}^{(1)} - E(Y_{i,n_2}^{(1)}))(Y_{i,n_2}^{(2)} - E(Y_{i,n_2}^{(2)}))\big] \\ &= -E\big[(Y_{i,n_2}^{(1)} - E(Y_{i,n_2}^{(1)}))(Y_{i,n_2}^{(1)} - E(Y_{i,n_2}^{(1)}))\big] \\ &= -E\big[(Y_{i,n_2}^{(1)} - E(Y_{i,n_2}^{(1)}))^2\big] \\ &= -\sigma^2_{Y_{i,n_2}^{(1)}} = (1/i^2)(\sigma^2_{C_{1,n_2}} + 4\sigma^2_{C_{2,n_2}}). \end{aligned} \tag{7.13}$$

Thus, we define the random variable $Y_{i,n_2} = Y_{i,n_2}^{(1)} + Y_{i,n_2}^{(2)}$, whose limit distribution is given by

$$Y_{i,n_2} \rightarrow N\big[0, (4/i^2)(\sigma^2_{C_{1,n_2}} + 4\sigma^2_{C_{2,n_2}})\big]. \tag{7.14}$$

Therefore, the limit distribution of the statistic C can be determined as linear combination of the random variables Y_{i,n_2}. The following theorem shows this distribution.

Theorem 22 *For $n \rightarrow \infty$, $n_2 \rightarrow \infty$ and $n_2/n \rightarrow 1/2$, the random variable C under the null hypothesis $H_0 = \theta = 0$, has an asymptotic normal distribution with parameters*

$$\mu_C = 0, \tag{7.15}$$

$$\begin{aligned} \sigma_C^2 &= \sum_{i=1}^{n} \sum_{n_2=B}^{n-i+B} a^2_{i,n_2} Var(Y_{i,n_2}) \\ &\quad +2 \sum \sum_{i<j} a_{i,n_2} a_{j,n_2} Cov(Y_{i,n_2}, Y_{j,n_2}), \end{aligned} \tag{7.16}$$

where

$$a_{i,n_2} = \binom{n-i}{i-1} \binom{n-i}{n_2 - B} \left(\frac{1}{2}\right)^{2n-i}, \tag{7.17}$$

and Y_{i,n_2} is defined as in [7.14].

Proof. From [5.11] and using [7.10], we obtain an expression for the distribution of C under the null hypothesis as a linear combination of random variables

Y_{i,n_2}, which in this case becomes

$$\begin{aligned} Pr_{H_0}[C=c] &= \left(\frac{1}{2}\right)^n \sum_{i=1}^{n} \binom{n-1}{i-1} \left(\frac{1}{2}\right)^{n-i} \left[\sum_{n_2=B}^{n-i+B} \binom{n-i}{n_2-B} Y_{i,n_2} \right] \\ &= \sum_{i=1}^{n} \sum_{n_2=B}^{n-i+B} \binom{n-1}{i-1} \binom{n-i}{n_2-B} \left(\frac{1}{2}\right)^{2n-i} Y_{i,n_2}, \end{aligned} \tag{7.18}$$

and from [7.14], the limit distribution of C under the null hypothesis is a combination of Normal random variables with mean μ_C and variance σ_C^2 given by

$$\begin{aligned} \mu_C &= E\left[\sum_{i=1}^{n} \sum_{n_2=B}^{n-i+B} \binom{n-1}{i-1} \binom{n-i}{n_2-B} \left(\frac{1}{2}\right)^{2n-i} Y_{i,n_2} \right] \\ &= \sum_{i=1}^{n} \sum_{n_2=B}^{n-i+B} \binom{n-1}{i-1} \binom{n-i}{n_2-B} \left(\frac{1}{2}\right)^{2n-i} E[Y_{i,n_2}] \\ &= 0, \end{aligned} \tag{7.19}$$

and

$$\begin{aligned} \sigma_C^2 &= Var\left[\sum_{i=1}^{n} \sum_{n_2=B}^{n-i+B} \binom{n-1}{i-1} \binom{n-i}{n_2-B} \left(\frac{1}{2}\right)^{2n-i} Y_{i,n_2} \right] \\ &= \sum_{i=1}^{n} \sum_{n_2=B}^{n-i+B} \left[\binom{n-1}{i-1} \binom{n-i}{n_2-B} \left(\frac{1}{2}\right)^{2n-i} \right]^2 Var[Y_{i,n_2}] + \\ &\quad 2 \sum \sum_{i<j} a_{i,n_2} a_{j,n_2} Cov(Y_{i,n_2}, Y_{j,n_2}), \end{aligned} \tag{7.20}$$

where a_{i,n_2} is defined as in [7.17].

This completes the proof.

■

Chapter 8
Suggestions for Further Research

Runs Statistics have important applications in various felds. The procedures based on runs are widely used due to their simplicity and applicability under fairly general assumptions. Runs also play a critical role in testing, where many applications have been used quite successfully. In this dissertation, the statistic C has been only used for a single sample location problem, but different variants of this statistic have been implemented in other problems, such as symmetry and randomization, with excellent results. However, the application of statistics C has not been developed extensively due to the lack of its distribution function.

In this dissertation, the distribution of the statistic C was determined using ordinary generating functions, and based on this distribution, the probability function under the null hypothesis and the corresponding power function of the test C were calculated. This distribution opens new horizons to many applications and provides a solid foundation to future uses. The distribution function of C can be used to determine theoretical distributions of different variants of this statistic. These variants allow the application of C to other problems. Therefore, this result is useful and revealing to study the distribution associated with runs in various applied problems. This is precisely the aim and purpose of this dissertation.

Most traditional applications of runs test have been developed for bivatiare runs distribution. New applications or extensions to more general runs tests have been hindered by the lack of exact critical values for distributions with more than two kinds of elements. The intent with this dissertation is to provide an alternative to partially remedy this situation by presenting a new approach which can be used to generate statistics on multiple runs distributions.

The statistic C has different applications in various fields of applied statistics such as data mining, paired comparisons and segmentation to name a few. These applications of C offer many opportunities for research.

Appendix A
Algorithm: Probability Distribution of Statistic C

In this appendix an Algorithm in *Mathematica* is presented. Through this algorithm extensive manipulations of power series can be made to determine the probability function of the statistics C given by (see theorem [11])

$$Pr[C=c]=\sum_{i=1}^{n} Pr[r_n=i]\left[\frac{G_1(i,c)}{\binom{n-1}{i-1}}Pr[\eta_1=1]+\frac{G_2(i,c)}{\binom{n-1}{i-1}}Pr[\eta_1=0]\right], \quad \text{(A.1)}$$

$-n \leq c \leq n$.

Some of the instructions used in the algorithm will be briefly explained.

- The instruction Series(h(x, y), (x, 0, p)) will display the first p terms of the power series expansion of a function f about $x=0$. For example, to see the first 10 terms of the series for $\frac{x}{1-2x}$, about the origin, you would enter

  Series((2x/(1 - 2x)), (x, 0, 10)).

  and *Mathematica* would respond

  $$2x+4x^2+8x^3+16x^4+32x^5+64x^6+128x^7+256x^8+512x^9+1024x^{10}+0[x]^{11}.$$

- If you want to obtain the list of coefficients of the terms of this series, then ask for

  CoefficientList(Series((2x/(1 - 2x)), x, 0, 10),x),

  to obtain

$$0, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024.$$

- If you want to see only the coefficient of x^5 then you would enter

  Coefficient[Series((2x/(1 - 2x)), x, 0, 10),x,5],

  instead, and the 32 would appear.

- Through the instructions Function(n, c, h(n,c)), Sum(f(i), (i, 1, 7)) and Product((y*x*i), (i, 1, p)) functions, sums and products can be defined in several variables. The way the ranges of variables are specified in Sum and Product is an example of the rather general iterator notation that Mathematica uses.

Variations of the above instructions can be used to determine [A.1]. To facilitate this calculation, [A.1] has been divided into 4 sums[1] as follows:

$$\begin{aligned} Pr[C = c] &= \sum_{i=1}^{n} Pr[r_n = i]\frac{G_1(i,c)}{\binom{n-1}{i-1}}p + \sum_{i=1}^{n} Pr[r_n = i]\frac{G_2(i,c)}{\binom{n-1}{i-1}}(1-p) \\ &= \sum_{k=1(2)}^{n} Pr[r_n = i]\frac{G_{11}(i,c)}{\binom{n-1}{i-1}}p \\ &+ \sum_{k=2(2)}^{n} Pr[r_n = i]\frac{G_{12}(i,c)}{\binom{n-1}{i-1}}p \\ &= \sum_{k=1(2)}^{n} Pr[r_n = i]\frac{G_{21}(i,c)}{\binom{n-1}{i-1}}(1-p) \\ &+ \sum_{k=2(2)}^{n} Pr[r_n = i]\frac{G_{22}(i,c)}{\binom{n-1}{i-1}}(1-p), \end{aligned} \tag{A.2}$$

where $p = Pr[\eta_1 = 1]$.

Thus, the first part of the algorithm consists of 4 groups of commands (see Figure A.1). Each group is formed by a set of instructions that provide solutions of each of the previous equations. For specific values n and c, the functions provided solutions of each of the parts in [A.2] as follows:

[1] The function has been divided according to whether i is even or odd.

1. Solutions for $\sum_{k=1(2)}^{n} G_{11}(n, c)$ through $h1a[n, r]$.
2. Solutions for $\sum_{k=2(2)}^{n} G_{12}(n, c)$ through $h1b[n, r]$.
3. Solutions for $\sum_{k=1(2)}^{n} G_{21}(n, c)$ through $h2a[n, r]$.
4. Solutions for $\sum_{k=2(2)}^{n} G_{22}(n, c)$ through $h2b[n, r]$.

These instructions must be executed separately and the order of execution does not matter.

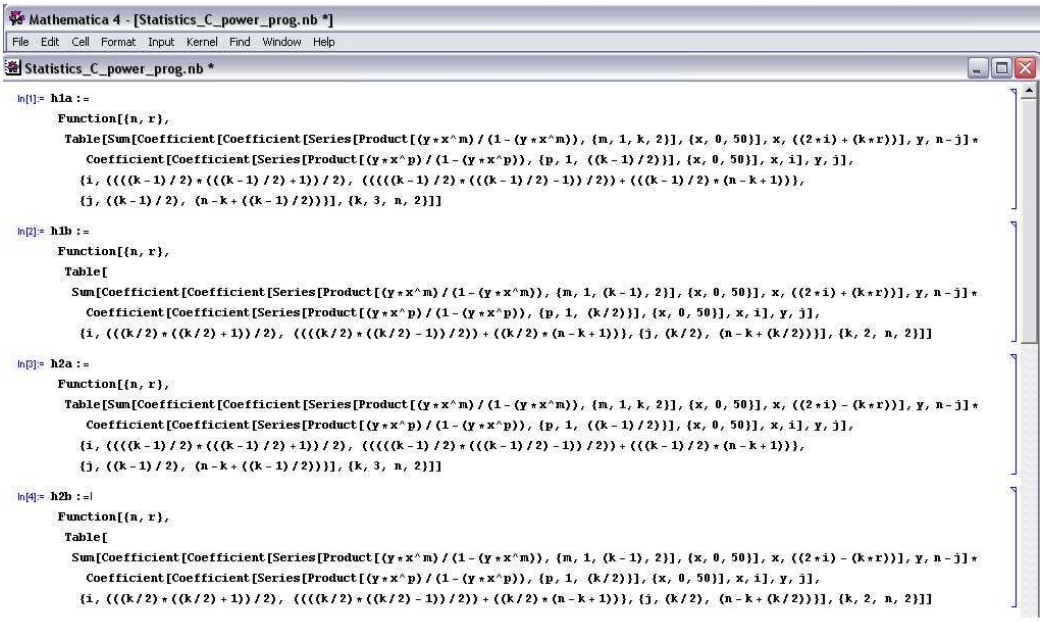

Figure A.1: Algorithm: Probability Distribution (Part 1)

Finally, in the second part of the algorithm we calculate the probabilities for $Pr[r_n = i]$ and add the functions used in the first part (see Figure A.2). To determine the probability of C we use the function,

PowC := Function[n, c, p, (Apply[Plus, Join[U1a[n, r, p], U1b[n, r, p], U2a[n, r, p], U2b[n, r, p]]])].

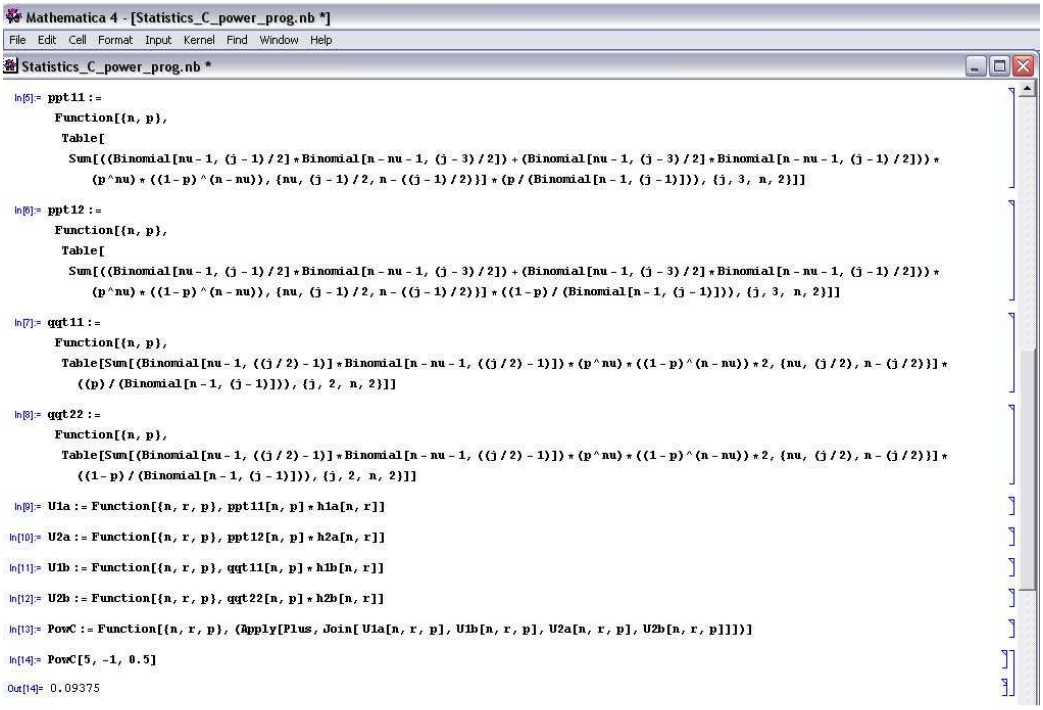

Figure A.2: Algorithm: Probability Distribution (Part 2)

For example $PowC[5, -1, 0.5]$[2],becomes

$$Pr[C = -1] = 0.09375.$$

This result can be compared with the distribution given in Table [5.1].

[2]In $PowC[n, c, p]$, n is the sample size, c corresponds to the desired probability and $p = Pr[\eta_1 = 1]$.

Appendix B

Possible Arrangements of Ones and Zeros and Values of the Statistic C for $n = 5$

η_1	η_2	η_3	η_4	η_5	δ_1	δ_2	δ_3	δ_4	δ_5	I_1	I_2	I_3	I_4	I_5	r_1	r_2	r_3	r_4	r_5	C
0	0	0	0	0	-1	-1	-1	-1	-1	1	0	0	0	0	1	1	1	1	1	**-5,00**
1	0	0	0	0	1	-1	-1	-1	-1	1	1	0	0	0	1	2	2	2	2	**-3,50**
0	1	0	0	0	-1	1	-1	-1	-1	1	1	1	0	0	1	2	3	3	3	**-2,67**
1	1	0	0	0	1	1	-1	-1	-1	1	0	1	0	0	1	1	2	2	2	**-2,00**
0	0	1	0	0	-1	-1	1	-1	-1	1	0	1	1	0	1	1	2	3	3	**-2,00**
1	0	1	0	0	1	-1	1	-1	-1	1	1	1	1	0	1	2	3	4	4	**-1,50**
0	1	1	0	0	-1	1	1	-1	-1	1	1	0	1	0	1	2	2	3	3	**-1,00**
1	1	1	0	0	1	1	1	-1	-1	1	0	0	1	0	1	1	1	2	2	**-0,50**
0	0	0	1	0	-1	-1	-1	1	-1	1	0	0	1	1	1	1	1	2	3	**-1,33**
1	0	0	1	0	1	-1	-1	1	-1	1	1	0	1	1	1	2	2	3	4	**-1,00**
0	1	0	1	0	-1	1	-1	1	-1	1	1	1	1	1	1	2	3	4	5	**-0,60**
1	1	0	1	0	1	1	-1	1	-1	1	0	1	1	1	1	1	2	3	4	**-0,25**
0	0	1	1	0	-1	-1	1	1	-1	1	0	1	0	1	1	1	2	2	3	**-0,33**
1	0	1	1	0	1	-1	1	1	-1	1	1	1	0	1	1	2	3	3	4	**0,25**
0	1	1	1	0	-1	1	1	1	-1	1	1	0	0	1	1	2	2	2	3	**0,67**
1	1	1	1	0	1	1	1	1	-1	1	0	0	0	1	1	1	1	1	2	**1,00**
0	0	0	0	1	-1	-1	-1	-1	1	1	0	0	0	1	1	1	1	1	2	**-1,00**
1	0	0	0	1	1	-1	-1	-1	1	1	1	0	0	1	1	2	2	2	3	**-0,67**
0	1	0	0	1	-1	1	-1	-1	1	1	1	1	0	1	1	2	3	3	4	**-0,25**
1	1	0	0	1	1	1	-1	-1	1	1	0	1	0	1	1	1	2	2	3	**0,33**
0	0	1	0	1	-1	-1	1	-1	1	1	0	1	1	1	1	1	2	3	4	**0,25**
1	0	1	0	1	1	-1	1	-1	1	1	1	1	1	1	1	2	3	4	5	**0,60**
0	1	1	0	1	-1	1	1	-1	1	1	1	0	1	1	1	2	2	3	4	**1,00**
1	1	1	0	1	1	1	1	-1	1	1	0	0	1	1	1	1	1	2	3	**1,33**
0	0	0	1	1	-1	-1	-1	1	1	1	0	0	1	0	1	1	1	2	2	**0,50**
1	0	0	1	1	1	-1	-1	1	1	1	1	0	1	0	1	2	2	3	3	**1,00**
0	1	0	1	1	-1	1	-1	1	1	1	1	1	1	0	1	2	3	4	4	**1,50**
1	1	0	1	1	1	1	-1	1	1	1	0	1	1	0	1	1	2	3	3	**2,00**
0	0	1	1	1	-1	-1	1	1	1	1	0	1	0	0	1	1	2	2	2	**2,00**
1	0	1	1	1	1	-1	1	1	1	1	1	1	0	0	1	2	3	3	3	**2,67**
0	1	1	1	1	-1	1	1	1	1	1	1	0	0	0	1	2	2	2	2	**3,50**
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	**5,00**

Table B.1: Possible Arrangements of Ones and Zeros and Values of the Statistic C for $n = 5$

Appendix C
Ordinary Generating Function

Definition 5 *Ordinary generating function (OGF). Suppose we are given a sequence $l_1, l_2, \ldots$ The ordinary generating function (also called OGF) associated with this sequence is the function whose value at x is $L(x) = \sum_{i=0}^{\infty} l_i x^i$. The sequence $l_1, l_2, \ldots$ denotes the coefficients of the generating function.*

It may have noticed that this definition is incomplete because we spoke of a function but did not specify its domain. The domain will depend on where the power series converges; however, for combinatorial applications, there is usually no need to be concerned with the convergence of the power series. As a result of this, we will often ignore the issue of convergence. In fact, we can treat the power series like a polynomial with an infinite number of terms.

If we have a doubly indexed sequence $l_{i,j}$, we can extend the definition of a generating function:

$$L(x, y) = \sum_{j \geq 0} \sum_{i \geq 0} l_{i,j} x^i y^j = \sum_{i,j=0}^{\infty} l_{i,j} x^i y^j.$$

Clearly, we can extend this idea to any number of indices, we are not limited to just one or two.

For example, here are some sequences and their respective generating functions:

$$\begin{aligned}
\langle 0, 0, 0, \ldots, 0 \rangle &\longleftrightarrow 0 + 0x + 0x^2 + 0x^3 + \cdots = 0, \\
\langle 1, 0, 0, \ldots, 0 \rangle &\longleftrightarrow 1 + 0x + 0x^2 + 0x^3 + \cdots = 1, \\
\langle 4, 2, 0, \ldots, 1 \rangle &\longleftrightarrow 4 + 2x + 0x^2 + 0x^3 + \cdots = 4 + 2x.
\end{aligned}$$

The pattern in the above functions is simple, the nth term in the sequence (here, indexing from 0) is the coefficient of x^n in the generating function.

Now, we use the sum of an infinite geometric series given by:

$$1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}, \text{ for } |x| < 1.$$

This formula gives closed-form generating functions for a whole range of sequences. For example:

$$\begin{aligned}
\langle 1, 1, 1, 1 \ldots \rangle &\longleftrightarrow 1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}, \\
\langle 1, -1, 1, -1 \ldots \rangle &\longleftrightarrow 1 - x + x^2 - x^3 + \cdots = \frac{1}{1+x}, \\
\langle 1, c, c^2, c^3 \ldots \rangle &\longleftrightarrow 1 + cx + c^2x^2 + c^3x^3 + \cdots = \frac{1}{1-cx}, \\
\langle 1, 0, 1, 0 \ldots \rangle &\longleftrightarrow 1 + x^2 + x^4 + \cdots = \frac{1}{1-x^2}, \\
\langle 1, 0, 0, 1 \ldots \rangle &\longleftrightarrow 1 + x^3 + \cdots = \frac{1}{1-x^3}.
\end{aligned}$$

The magic of generating functions is that we can carry out all sorts of manipulations on sequences by performing mathematical operations on their associated generating functions.

Now, we present various operations and theorems very useful, and we show their effects in terms of sequences.

Definition 6 $[x^n]$ *Given a generating function* $L(x)$ *we use* $[x^n]L(x)$ *to denote* l_n*, the coefficient of* x^n*. For a generating function in more variables, the coefficient may be another generating function. For example* $[x^n y^k]L(x, y) = l_{n,k}$ *and* $[x^n]L(x, y) = \sum_{i \geq 0} l_{n,i} y^i$*.*

In the previous definition can be seen that the generating function uniquely determines its coefficients. In other words, given a generating function there is just one sequence that gives rise to it. Without this uniqueness, generating functions would be of little use since we would not be able to recover the coefficients from the function alone.

Theorem 23 *Taylor's Theorem If* $L(x)$ *is the generating function for a sequence* $l_1, l_2, \ldots$*, then* $l_n = \frac{L^{(n)}(0)}{n!}$*, where* $L^{(0)}$ *is the* n*th derivative of* L *and* $0! = 1$*.*

There are two important differences in the study of generating functions here (in this sense) and in calculus. We have already noted one: convergence is usually not an issue. The second is that our interest is in the reverse direction: We study

generating functions to learn about their coefficients but in calculus one studies the coefficients to learn about the functions.

Theorem 24 *Convolution Formula Let $A(x)$, $B(x)$, and $C(x)$ be generating functions. Then $C(x) = A(x)B(x)$ if and only if*

$$c_n = \sum_{k=0}^{n} a_k b_{n-k} \text{ for all } n \geq 0.$$

The sum can also be written $\sum_{k\geq 0} a_{n-k}b_k$ and also as the sum $a_i b_j$ over all i, j such that $i + j = n$.

There are other operations which we can perform on formal power series. Although they will look much like the typical operations on functions of a complex variable, these are all operations within the ring of power series: they are defined "axiomatically" on the symbols without the need to take any limits. Start with a formal power series $L(x) = \sum_n l_i x^i$.

1. *Right-shifting*
$$[x^n]x^k L(x) = [x^{n-k}]L(x).$$
That is, if $H(x) = x^k L(x) = \sum_n h_n x^n$, then $h_n = l_{n-k}$, the shifted sequence.

2. *Left-shifting.* For one-sided series, we can create a new series by
$$H(x) = \frac{L(x) - \sum_{n=0}^{m-1} l_n x^n}{x^m} = \sum_{n\geq 0} l_{n+m} x^n.$$
That is, $[x^n]H(x) = [x^{n+m}]L(x)$ for $n \geq 0$. This is a, truncated left shift, and the sum above cannot in general be extended over all integers. A two-sided left-shift is obtained by $[x^n]L(x)/x^m = [x^{n+m}]L(x)$.

3. *Derivatives.* We define the derivative operator on formal power series by $H'(x) = \sum_n n l_n x^{n-1}$. Hence,
$$[x^{n-1}]L'(x) = n[x^n]L(x),$$
$$[x^{n-k}]L^{(k)}(x) = n^{\underline{k}}[x^n]L(x).$$

4. *Shift-Derivative Combinations.* This operation is even more useful when we combine shift and derivative operations. Let S be the right-shift operator

$SG(x) = zG(x)$ and D be the derivative operator $DG(x) = G'(x)$. Then, we find that

$$[x^n]S^k D^k L(x) = n^{\underline{k}}[x^n]L(x),$$
$$[x^n](SD)^k L(x) = n^k[x^n]L(x).$$

Both of these are useful operators in their own right. The second identity generalizes easily to a useful form. If $Q(x)$ is some polynomial, then $Q(SD)$ is an operator consisting of a linear combination of $(SD^)ks$. Using linearity, the second equality above thus shows that

$$[x^n]Q(SD)L(x) = Q(n)[x^n]L(x).$$

5. *Partial Summation.* By convolution and the fact that $[x^n]1/(1-x) = 1_{n\geq 0}$,

$$\frac{L(x)}{1-x} = \sum_n \left(\sum_{k\leq n} l(x) \right) x^n.$$

Multiplying by $1/(1-z)$ converts a sequence to the sequence of partial sums.

6. *Integration.* We can define cumulative definite integrals along the real line as follows:

$$\int_0^x L(t)dt = \sum_{n\geq 1} \frac{1}{n} l_{n-1} x^n.$$

Hence, $[x^n]\int_0^x L(t)dt = (1/n)[x^{n-1}]L(x)$, for $n \geq 1$.

7. *Scaling.* If c is a constant,

$$[x^n]L(cx) = c^n[x^n]L(x).$$

8. *Extraction of Sub-sequences* (one-sided only). Let m be a positive integer. Consider the roots of unity $e^{2\pi ij/m}$ for $j = 0, \ldots, m-1$. We have the following for integer $n \geq 0$

$$\sum_{j=0}^{m-1} e^{2\pi inj/m} = \begin{cases} m & \text{if } n = dm \text{ for positive integer } d, \\ \frac{e^{2\pi in}-1}{e^{2\pi in/m}-1} = 0 & \text{otherwise,} \end{cases}$$

where the second case is just an ordinary finite geometric sum. Given a one-sided formal power series, consider the average over the roots of unity:

$$\begin{aligned} H(x) &= \frac{1}{m}\sum_{j=0}^{m-1} L(xe^{2\pi ij/m}) \\ &= \sum_{n\geq 0} l_n x^n \frac{1}{m}\sum_{j=0}^{m-1} e^{2\pi inj/m} \\ &= \sum_{n\geq 0} l_m.nx^{m.n} \\ &= F(x^m), \end{aligned}$$

for a power series $F(x)$ with $[x^n]F(x) = [x^{mn}]G(x)$. Denote the mapping from G to F by the operator M_m.

9. *Composition.* If $L(x)$ and $F(x)$ are two generating functions, we might wish to compute the composition of the two functions.

$$L(F(x)) = \sum_{n\geq 0} l_n F^n(x) = \sum_{n\geq 0} l_n (\sum_m f_m x^m)^n.$$

Notice that if $f_0 \neq 0$, then the contribution of each term in the sum to $[x^n]L(F(x))$ could be non-zero if $f_n \neq 0$ infinitely often. The algebraic construction of formal power series does not support this infinite sum, so the composition is not well-defined as a formal power series in this case. If $f_0 = 0$, however, then $[x^n]L(F(x))$ depends on at most n terms in the sum, which is well defined. The composition operation is thus well defined for two formal power series only if $f_0 = 0$ or if $L(x)$ has only finitely many non-zero coefficients (i.e., it is a polynomial).

Appendix D

Algorithm: Probability Distribution of Statistic C under the Null Hypothesis

In this appendix an Algorithm in *Mathematica* is presented. Through this algorithm extensive manipulations of power series can be made to determine the probability function of the statistics C under the null hypothesis given by (see theorem [5.3])

$$Pr_{H_0}[C = c] = \frac{1}{2^n}\left[\sum_{i=1}^{n} G_1(i,c) + G_2(i,c)\right], -n \leq c \leq n \qquad \text{(D.1)}$$

This algorithm is similar to one presented in Appendix A. To facilitate the calculation of the probability function, [D.1] has been divided into 4 sums[1] as follows:

$$\begin{aligned} Pr_{H_0}[C = c] &= \frac{1}{2^n}\left[\sum_{i=1}^{n} G_1(i,c) + G_2(i,c)\right] \\ &= \frac{1}{2^n}\left[\sum_{k=1(2)} G_{11}(i,c) + \sum_{k=2(2)} G_{12}(i,c)\right. \\ &+ \left.\sum_{k=1(2)} G_{21}(i,c) + \sum_{k=2(2)} G_{22}(i,c)\right]. \end{aligned} \qquad \text{(D.2)}$$

Thus, the first part of the algorithm consists of 4 groups of commands (see Figure A.1). Each group is formed by a set of instructions that provide solutions of each of the previous equations. For specific values n and c, the functions provided solutions of each of the parts in [D.2] as follows:

[1]The function has been divided according to whether i is even or odd.

1. Solutions for $\sum_{k=1(2)}^{n} G_{11}(n,c)$ through $h1a[n,r]$.
2. Solutions for $\sum_{k=2(2)}^{n} G_{12}(n,c)$ through $h1b[n,r]$.
3. Solutions for $\sum_{k=1(2)}^{n} G_{21}(n,c)$ through $h2a[n,r]$.
4. Solutions for $\sum_{k=2(2)}^{n} G_{22}(n,c)$ through $h2b[n,r]$.

These instructions must be executed separately and the order of execution does not matter.

Finally, to determine the probability function of the statistics C under the null hypothesis(see Figure D.1), we use the function

EstR:= Function[n,r, (Apply[Plus, Join[h1a[n,r], h1b[n,r], h2a[n,r], h2b[n,r]]])/(2^n)]

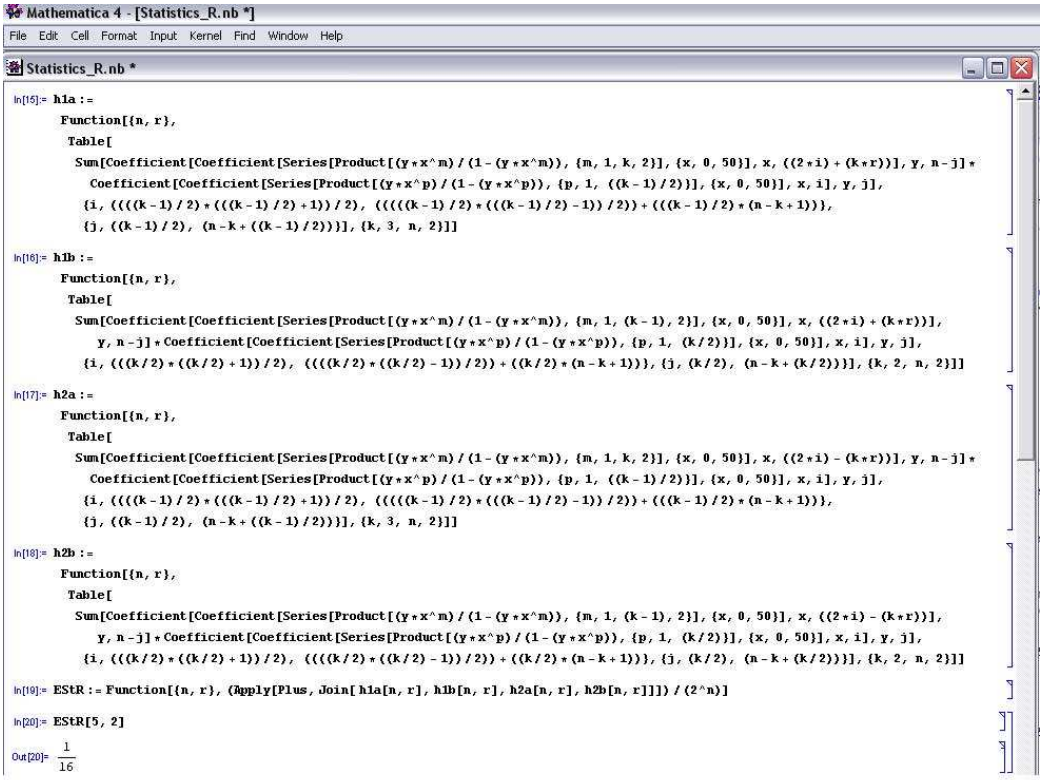

Figure D.1: Algorithm: Probability Distribution under the Null Hypothesis

For example $EStR[5,2]$[2], becomes

[2]In $EStR[n,c]$, n is the sample size and c corresponds to the desired probability.

$$Pr_{H_0}[C = 2] = 1/16 = 0.0625.$$

This result can be compared with the distribution given in Table [5.1].

Appendix E
Critical Values of Statistic C

The table entry is the cumulative probability of c or less (left-tail) under the null distribution of C.

n	0,005		0,01		0,025		0,05		0,1		0,2		0,3	
	C	α^*	C	α^*	C	α^*	C	α^*	C	α^*	C	α^*	C	α^*
4							2,5000	0,06250	1,6667	0,12500	1,0000	0,18750	0,5000	0,31250
5					3,5000	0,03125	2,6667	0,06250	2,0000	0,09375	1,0000	0,21875	0,6667	0,31250
											1,3333	0,18750		
6			4,5000	0,01563	3,6667	0,03125	3,0000	0,04688	2,0000	0,10938	1,3333	0,21875	0,7500	0,31250
									2,3333	0,09375	1,5000	0,17188	0,8000	0,29688
													1,0000	0,28125
7	5,5000	0,00781	4,6667	0,01563	4,0000	0,02344	3,0000	0,05469	2,3333	0,10156	1,5000	0,20313	0,8333	0,31250
							3,3333	0,04688	2,5000	0,08594	1,6667	0,19531	1,0000	0,26563
8	6,5000	0,00391	5,0000	0,01172	4,0000	0,02734	3,3333	0,05078	2,4000	0,10938	1,5714	0,20313	0,8333	0,32422
					4,3333	0,02344	3,5000	0,04297	2,5000	0,09375	1,6000	0,19922	1,0000	0,26953
											1,6667	0,19141	1,1667	0,26172
9	6,0000	0,00586	5,3333	0,01172	4,3333	0,02539	3,3333	0,05273	2,6000	0,10156	1,6000	0,20703	0,8571	0,32422
	6,6667	0,00391	5,5000	0,00977	4,5000	0,02148	3,4000	0,04883	2,6667	0,09766	1,6667	0,19141	1,0000	0,28711
									2,7500	0,09180	1,7143	0,18750	1,1429	0,28320
10	6,3333	0,00586	5,5000	0,01074	4,3333	0,02637	3,5714	0,05176	2,6667	0,10254	1,7143	0,20215	0,8889	0,32617
	6,5000	0,00488	5,6000	0,00977	4,4000	0,02441	3,6000	0,04883	2,7143	0,09961	1,7500	0,19531	1,0000	0,29883
			5,6667	0,00879	4,5000	0,02246	3,6667	0,04688	2,7500	0,09570	1,8000	0,19043	1,1111	0,29785
11	6,5000	0,00537	5,6667	0,01074	4,5714	0,02588	3,7143	0,05029	2,8000	0,10010	1,8000	0,20117	1,1250	0,30029
	6,6000	0,00488	5,7500	0,00977	4,6000	0,02441	3,7500	0,04834	2,8333	0,09814	1,8333	0,19434	1,1429	0,29541
	6,6667	0,00439	5,8000	0,00879	4,6667	0,02295	3,8000	0,04590	2,8571	0,09521	1,8571	0,18945	1,1667	0,28711
12	6,6667	0,00537	5,8000	0,01025	4,6667	0,02539	3,8000	0,05103	2,8889	0,10010	1,8333	0,20313	1,1429	0,30127
	6,7500	0,00488	5,8333	0,00977	4,7143	0,02466	3,8333	0,04932	3,0000	0,08911	1,8571	0,19824	1,1667	0,29614
	6,8000	0,00439	6,0000	0,00806	4,7500	0,02393	3,8571	0,04785	3,1111	0,08862	1,8750	0,19507	1,1818	0,29590
13	6,8000	0,00513	5,8333	0,01074	4,8333	0,02502	3,8889	0,05212	2,9000	0,10571	1,9000	0,20129	1,1818	0,30054
	6,8333	0,00488	6,0000	0,00891	4,8571	0,02417	4,0000	0,04529	3,0000	0,09521	2,0000	0,18286	1,2000	0,29602
	7,0000	0,00403	6,1667	0,00854	5,0000	0,02051	4,1111	0,04504	3,1000	0,09497	2,1000	0,18237	1,2222	0,29382
14	6,8333	0,00531	6,0000	0,01019	4,8889	0,02625	3,9000	0,05579	3,1000	0,10065	1,9091	0,20691	1,2000	0,30157
	7,0000	0,00439	6,1429	0,00989	5,0000	0,02295	4,0000	0,04956	3,1111	0,09991	2,0000	0,19049	1,2222	0,29938
	7,1667	0,00421	6,1667	0,00952	5,1111	0,02283	4,1000	0,04944	3,1250	0,09814	2,0909	0,19012	1,2308	0,29932
15	7,0000	0,00504	6,2000	0,01022	5,1111	0,02530	4,1429	0,05057	3,1429	0,10123	1,9231	0,21213	1,2308	0,30429
	7,1429	0,00488	6,2500	0,00986	5,1250	0,02481	4,1667	0,04944	3,1667	0,09961	2,0000	0,19809	1,2500	0,29968
	7,1667	0,00473	6,2857	0,00952	5,1429	0,02408	4,1818	0,04938	3,1818	0,09955	2,0769	0,19803	1,2727	0,29883
16	7,2000	0,00502	6,3750	0,01009	5,2000	0,02504	4,2727	0,05029	3,2308	0,10181	2,1000	0,20108	1,2857	0,30045
	7,2500	0,00484	6,4000	0,00981	5,2222	0,02477	4,2857	0,04919	3,2500	0,09976	2,1111	0,19855	1,3000	0,29811
	7,2857	0,00468	6,4286	0,00952	5,2500	0,02412	4,3000	0,04889	3,2727	0,09932	2,1250	0,19582	1,3077	0,29776
17	7,3750	0,00502	6,4444	0,01046	5,3000	0,02541	4,3636	0,05065	3,3077	0,10184	2,1333	0,20157	1,3077	0,30249
	7,4000	0,00489	6,5000	0,00989	5,3333	0,02453	4,3750	0,04970	3,3333	0,09936	2,1429	0,19981	1,3333	0,29709
	7,4286	0,00472	6,5455	0,00988	5,3636	0,02452	4,3846	0,04969	3,3571	0,09935	2,1538	0,19955	1,3571	0,29706

Table E.1: Critical Values of Statistic C, $P_{H_0} = C \geq c_{1-\alpha/2} \approx \alpha$, for $4 \leq n \leq 17$.

n	0,005		0,01		0,025		0,05		0,1		0,2		0,3	
	C	α^*	C	α^*	C	α^*	C	α^*	C	α^*	C	α^*	C	α^*
18	7,4444	0,00525	6,6364	0,01003	5,4286	0,02514	4,4615	0,05029	3,4167	0,10026	2,1818	0,20220	1,3571	0,30158
	7,5000	0,00494	6,6667	0,00968	5,4444	0,02469	4,5000	0,04823	3,4286	0,09929	2,2000	0,19982	1,3636	0,29976
	7,5455	0,00493	6,7000	0,00961	5,4545	0,02461	4,5455	0,04788	3,4444	0,09782	2,2143	0,19973	1,3750	0,29757
19	7,6364	0,00505	6,7500	0,01005	5,5455	0,02535	4,5556	0,05030	3,4667	0,10162	2,2308	0,20242	1,3889	0,30110
	7,6667	0,00486	6,7778	0,00987	5,5556	0,02490	4,5714	0,04976	3,5000	0,09835	2,2500	0,19982	1,4000	0,29841
	7,7000	0,00483	6,8000	0,00970	5,5714	0,02452	4,5833	0,04953	3,5385	0,09810	2,2667	0,19976	1,4118	0,29841
20	7,7500	0,00509	6,8750	0,01010	5,6429	0,02539	4,6429	0,05072	3,5556	0,10051	2,3000	0,20033	1,4211	0,30081
	7,7778	0,00499	6,8889	0,00990	5,6667	0,02470	4,6667	0,04953	3,5714	0,09985	2,3077	0,19949	1,4286	0,29955
	7,8000	0,00491	6,9000	0,00978	5,6923	0,02466	4,6923	0,04940	3,5833	0,09922	2,3125	0,19945	1,4375	0,29946
21	7,8889	0,00503	6,9286	0,01064	5,7692	0,02519	4,7500	0,05007	3,6429	0,10001	2,3571	0,20066	1,4444	0,30178
	7,9000	0,00498	7,0000	0,00974	5,7778	0,02480	4,7692	0,04987	3,6667	0,09811	2,3636	0,19904	1,4545	0,29979
	7,9091	0,00495	7,0714	0,00974	5,7857	0,02478	4,7778	0,04925	3,6875	0,09809	2,3750	0,19818	1,4615	0,29852
22	7,9286	0,00546	7,1333	0,01002	5,8750	0,02521	4,8462	0,05017	3,7222	0,10028	2,4118	0,20012	1,4737	0,30170
	8,0000	0,00498	7,1429	0,00992	5,8889	0,02488	4,8571	0,04985	3,7273	0,09931	2,4167	0,19872	1,5000	0,29636
	8,0714	0,00498	7,1538	0,00991	5,9000	0,02454	4,8667	0,04982	3,7333	0,09915	2,4211	0,19872	1,5238	0,29636
23	8,1667	0,00505	7,2667	0,01004	5,9412	0,02570	4,9231	0,05008	3,7895	0,10027	2,4500	0,20123	1,5263	0,30010
	8,1818	0,00499	7,2727	0,00991	6,0000	0,02399	4,9286	0,04989	3,8000	0,09927	2,4545	0,19992	1,5294	0,29995
	8,2000	0,00490	7,2857	0,00983	6,0588	0,02399	4,9333	0,04982	3,8125	0,09918	2,4615	0,19875	1,5333	0,29921
24	8,3000	0,00501	7,3750	0,01004	6,0833	0,02525	4,9444	0,05257	3,8571	0,10005	2,4762	0,20244	1,5385	0,30101
	8,3077	0,00499	7,3846	0,00998	6,0909	0,02495	5,0000	0,04972	3,8667	0,09975	2,5000	0,19864	1,5455	0,29970
	8,3333	0,00486	7,4000	0,00983	6,1000	0,02468	5,0556	0,04972	3,8750	0,09938	2,5263	0,19862	1,5500	0,29969
25	8,4375	0,00503	7,4706	0,01021	6,1875	0,02509	5,1176	0,05003	3,9231	0,10003	2,5385	0,20075	1,5789	0,30037
	8,4444	0,00496	7,5000	0,00987	6,2000	0,02479	5,1250	0,04983	3,9286	0,09945	2,5455	0,19982	1,5833	0,29904
	8,4545	0,00490	7,5333	0,00986	6,2105	0,02479	5,1333	0,04966	3,9333	0,09907	2,5500	0,19981	1,5882	0,29868
26	8,5455	0,00505	7,6154	0,01001	6,2941	0,02508	5,2105	0,05005	3,9545	0,10236	2,5909	0,20047	1,6087	0,30020
	8,5556	0,00500	7,6250	0,00996	6,3000	0,02490	5,2143	0,04973	4,0000	0,09810	2,6000	0,19912	1,6111	0,29995
	8,5625	0,00500	7,6364	0,00983	6,3077	0,02468	5,2222	0,04954	4,0455	0,09810	2,6111	0,19896	1,6154	0,29866
27	8,6875	0,00503	7,7273	0,01001	6,3889	0,02508	5,2857	0,05008	4,0667	0,10056	2,6364	0,20041	1,6316	0,30065
	8,6923	0,00499	7,7333	0,00998	6,4000	0,02481	5,2941	0,05000	4,0714	0,09994	2,6429	0,19941	1,6364	0,29999
	8,7000	0,00493	7,7500	0,00983	6,4118	0,02479	5,3000	0,04979	4,0769	0,09929	2,6471	0,19898	1,6400	0,29999
28	8,8125	0,00501	7,8421	0,01006	6,4762	0,02523	5,3684	0,05022	4,1364	0,10005	2,6842	0,20052	1,6538	0,30108
	8,8182	0,00496	7,8462	0,00997	6,5000	0,02460	5,3750	0,04999	4,1429	0,09942	2,6875	0,19979	1,6667	0,29855
	8,8235	0,00496	7,8571	0,00989	6,5263	0,02459	5,3810	0,04999	4,1500	0,09940	2,6923	0,19894	1,6800	0,29855
29	8,9167	0,00504	7,9524	0,01001	6,5833	0,02504	5,4545	0,05019	4,1905	0,10050	2,7308	0,20010	1,6875	0,30018
	8,9231	0,00499	8,0000	0,00947	6,5882	0,02498	5,4615	0,04986	4,2000	0,09977	2,7333	0,19924	1,6923	0,29935
	8,9286	0,00496	8,0476	0,00947	6,5909	0,02498	5,4667	0,04956	4,2083	0,09977	2,7368	0,19901	1,6957	0,29934
30	9,0667	0,00500	8,0714	0,01001	6,6842	0,02510	5,5333	0,05015	4,2632	0,10028	2,7647	0,20046	1,7083	0,30098
	9,0714	0,00496	8,0769	0,00992	6,6875	0,02498	5,5385	0,04987	4,2667	0,09975	2,7692	0,19987	1,7143	0,29996
	9,0769	0,00491	8,0833	0,00984	6,6923	0,02480	5,5455	0,04972	4,2692	0,09975	2,7727	0,19984	1,7200	0,29996

Table E.2: Critical Values of Statistic C, $P_{H_0} = C \geq c_{1-\alpha/2} \approx \alpha$, for $18 \leq n \leq 30$.

Appendix F

Statistics for Sample Sizes 10, 15, 20, 25 and 30

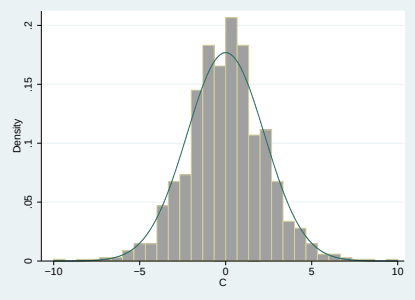

Figure F.1: Histogram for C and Normal curve ($n = 10$)

Mean	0,00
Median	-
Std. Deviation	2,26
Variance	5,09
Skewness	-
Kurtosis	0,419
Percentiles	
1	- 5,60
5	- 3,60
10	- 2,71
25	- 1,40

Table F.1: Basic Statistics for the Statistic C ($n = 10$)

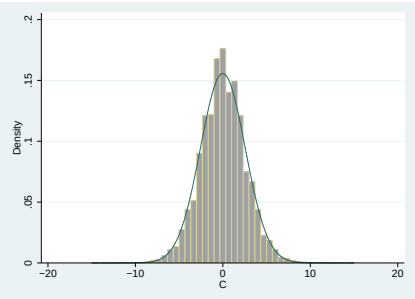

Figure F.2: Histogram for C and Normal curve ($n = 15$)

Mean	0,00
Median	-
Std. Deviation	2,56
Variance	6,56
Skewness	-
Kurtosis	0,369
Percentiles	
1	- 6,25
5	- 4,17
10	- 3,17
25	- 1,63

Table F.2: Basic Statistics for the Statistic C ($n = 15$)

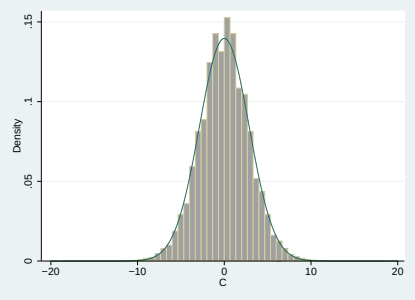

Figure F.3: Histogram for C and Normal curve ($n = 20$)

Mean	0,00
Median	-
Std. Deviation	2,86
Variance	8,16
Skewness	-
Kurtosis	0,347
Percentiles	
1	- 6,89
5	- 4,67
10	- 3,57
25	- 1,85

Table F.3: Basic Statistics for the Statistic C ($n = 20$)

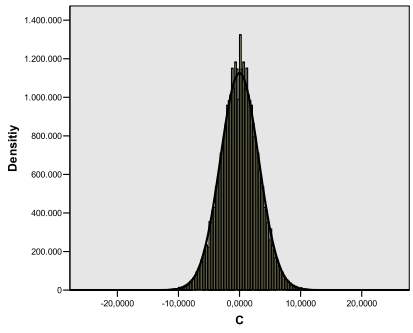

Figure F.4: Histogram for C and Normal curve ($n = 25$)

Mean	0,00
Median	-
Std. Deviation	3,13
Variance	9,79
Skewness	-
Kurtosis	0,36
Percentiles	
1	- 7,50
5	- 5,13
10	- 3,93
25	- 2,00

Table F.4: Basic Statistics for the Statistic C ($n = 25$)

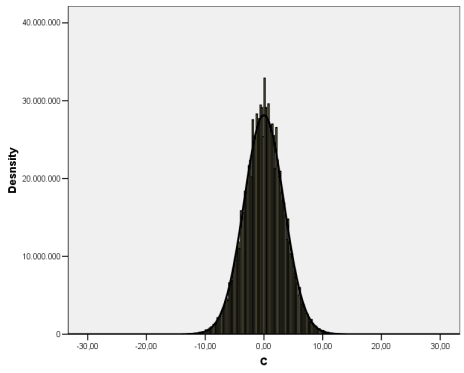

Figure F.5: Histogram for C and Normal curve ($n = 30$)

Mean	0,00
Median	-
Std. Deviation	3,38
Variance	11,43
Skewness	-
Kurtosis	0,29
Percentiles	
1	- 8,07
5	- 5,54
10	- 4,27
25	- 2,21

Table F.5: Basic Statistics for the Statistic C ($n = 30$)

Appendix G

Power of the C test, the sign test (S) and the Wilcoxon signed rank test (W) for $n = 15, 20, 25, 30$

Distribution	q	C	S	W
N(0,1)	0,50	0,0514	0,0528	0,0520
	0,55	0,1152	0,0776	0,1124
	0,60	0,2280	0,1284	0,2256
	0,65	0,3932	0,2104	0,3892
	0,70	0,5870	0,3366	0,5870
C(0,1)	0,50	0,0490	0,0506	0,0500
	0,55	0,0962	0,0754	0,0938
	0,60	0,1658	0,1276	0,1696
	0,65	0,2654	0,2090	0,2684
	0,70	0,3768	0,3348	0,3996
U(0,1)	0,50	0,0450	0,0478	0,0450
	0,55	0,1508	0,0756	0,1404
	0,60	0,3424	0,1244	0,3172
	0,65	0,5678	0,2012	0,5342
	0,70	0,7726	0,3280	0,7468
L(0,1)	0,50	0,0472	0,0498	0,0472
	0,55	0,0912	0,0734	0,0928
	0,60	0,1726	0,1232	0,1718
	0,65	0,2856	0,2072	0,2960
	0,70	0,4420	0,3258	0,4588
Log(0,1)	0,50	0,0450	0,0478	0,0450
	0,55	0,1072	0,0756	0,1040
	0,60	0,2164	0,1244	0,2162
	0,65	0,3678	0,2012	0,3704
	0,70	0,5542	0,3280	0,5550

Table G.1: Power of the C test, the sign test (S) and the Wilcoxon signed rank test (W) for $n = 15$

Distribution	q	C	S	W
N(0,1)	0,50	0,0505	0,0499	0,0496
	0,55	0,1319	0,0819	0,1390
	0,60	0,2827	0,1501	0,2914
	0,65	0,4867	0,2739	0,4966
	0,70	0,7095	0,4413	0,7236
C(0,1)	0,50	0,0527	0,0497	0,0540
	0,55	0,1053	0,0905	0,1132
	0,60	0,1925	0,1517	0,1966
	0,65	0,3153	0,2715	0,3344
	0,70	0,4539	0,4451	0,4884
U(0,1)	0,50	0,0471	0,0513	0,0482
	0,55	0,1865	0,0857	0,1820
	0,60	0,4287	0,1585	0,4096
	0,65	0,6813	0,2737	0,6662
	0,70	0,8713	0,4435	0,8668
L(0,1)	0,50	0,0527	0,0517	0,0544
	0,55	0,1057	0,0849	0,1100
	0,60	0,2011	0,1633	0,2150
	0,65	0,3471	0,2723	0,3654
	0,70	0,5439	0,4403	0,5802
Log(0,1)	0,50	0,0471	0,0513	0,0482
	0,55	0,1229	0,0857	0,1250
	0,60	0,2659	0,1585	0,2762
	0,65	0,4545	0,2737	0,4734
	0,70	0,6681	0,4435	0,6802

Table G.2: Power of the C test, the sign test (S) and the Wilcoxon signed rank test (W) for $n = 20$

Distribution	q	C	S	W
N(0,1)	0,50	0,0436	0,0502	0,0440
	0,55	0,1432	0,0928	0,1462
	0,60	0,3310	0,1790	0,3348
	0,65	0,5668	0,3332	0,5772
	0,70	0,7970	0,5430	0,8028
C(0,1)	0,50	0,0508	0,0490	0,0510
	0,55	0,1170	0,0936	0,1166
	0,60	0,2138	0,1764	0,2172
	0,65	0,3486	0,3294	0,3664
	0,70	0,5132	0,5378	0,5556
U(0,1)	0,50	0,0476	0,0486	0,0472
	0,55	0,2036	0,0908	0,1934
	0,60	0,5016	0,1816	0,4786
	0,65	0,7658	0,3352	0,7488
	0,70	0,9272	0,5402	0,9236
L(0,1)	0,50	0,0484	0,0502	0,0490
	0,55	0,1136	0,0912	0,1146
	0,60	0,2324	0,1880	0,2392
	0,65	0,4162	0,3278	0,4302
	0,70	0,6310	0,5322	0,6566
Log(0,1)	0,50	0,0476	0,0486	0,0472
	0,55	0,1350	0,0908	0,1386
	0,60	0,3030	0,1816	0,3098
	0,65	0,5270	0,3352	0,5416
	0,70	0,7602	0,5402	0,7654

Table G.3: Power of the C test, the sign test (S) and the Wilcoxon signed rank test (W) for $n = 25$

Distribution	q	C	S	W
N(0,1)	0,50	0,0503	0,0482	0,0506
	0,55	0,1545	0,1308	0,1556
	0,60	0,3721	0,2884	0,3776
	0,65	0,6437	0,5144	0,6496
	0,70	0,8619	0,7320	0,8636
C(0,1)	0,50	0,0493	0,0474	0,0484
	0,55	0,1175	0,1286	0,1202
	0,60	0,2345	0,2242	0,2468
	0,65	0,3931	0,3644	0,4124
	0,70	0,5763	0,5342	0,6120
U(0,1)	0,50	0,0493	0,0454	0,0490
	0,55	0,2307	0,1372	0,2190
	0,60	0,5517	0,2938	0,5344
	0,65	0,8243	0,5040	0,8140
	0,70	0,9575	0,7242	0,9566
L(0,1)	0,50	0,0489	0,0510	0,0498
	0,55	0,1187	0,1332	0,1212
	0,60	0,2557	0,2928	0,2664
	0,65	0,4697	0,5026	0,4866
	0,70	0,7035	0,7336	0,7282
Log(0,1)	0,50	0,0447	0,0470	0,0466
	0,55	0,1507	0,1286	0,1532
	0,60	0,3429	0,2918	0,3550
	0,65	0,5905	0,5110	0,6022
	0,70	0,8209	0,7322	0,8328

Table G.4: Power of the C test, the sign test (S) and the Wilcoxon signed rank test (W) for $n = 30$

Appendix H
SIMULA: A SAS Macro for the Power of the Statistics C, S and W.

To facilitate the calculation of the power function, a SAS macro that simulates a specific number of samples for a predetermined sample size was developed. The macro provides a simple way to use the power of Monte Carlo simulation techniques to investigate the statistical power of the tests C, S and W. The macro may also be used to compare the powers of the tests in question.

MACRO SIM

In order to simulate the data, two parameters are required (n,s). The variables are supplied as arguments when the macro is called:

- The sample size n (In this case 10, 15, 20, 25, 30)
- The number of simulations s (In this case 5000)

The window generated by the macro for the entering of the parameters is shown in Figure H.1 and the complete macro for the simulation [1] is shown in Figure H.2.

The complete macro is presented below.

```
%window SIM color=gray irow=2 rows=99 icolumn=5
columns=70
#1 @18 "POWER ESTIMATION" attr=highlight color=yellow
#2 @22 "OF STATISTICS C, S AND W" attr=highlight
color=yellow
#3 @8 "NUMBER OF SAMPLES:"@40 SIZE 4 ATTR=UNDERLINE
```

[1] We present the macro for the normal distribution, because for the other distributions the macro works the same way.

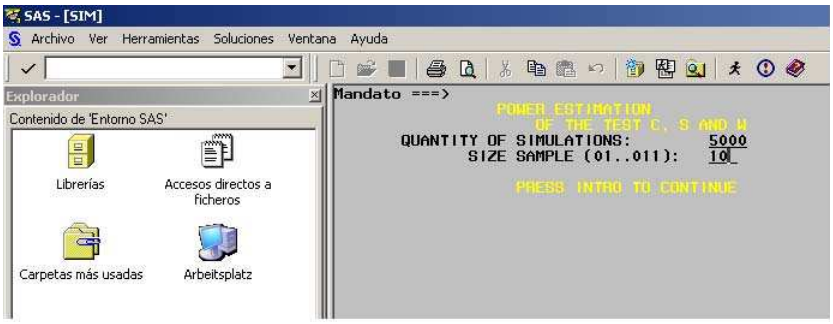

Figure H.1: Window SAS Macro SIM

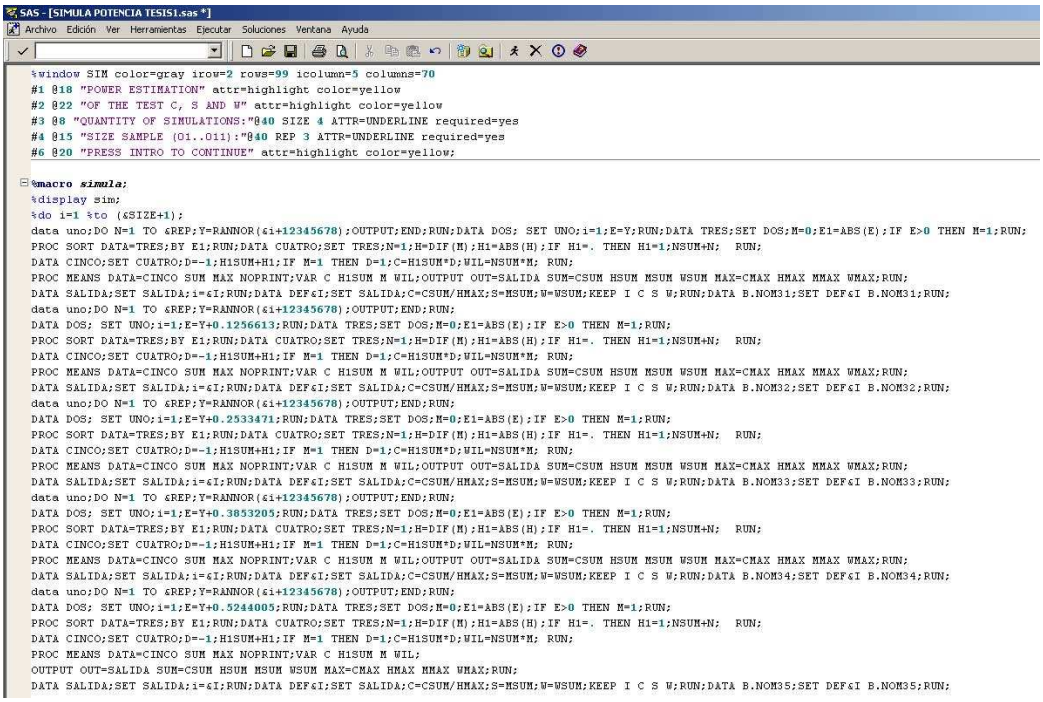

Figure H.2: SAS Macro SIM

```
required=yes
#4 @15 "SAMPLE SIZE:"@40 REP 3 ATTR=UNDERLINE
required=yes
#6 @20 "PRESS INTRO TO CONTINUE" attr=highlight
color=yellow;

%macro simula;
%display sim;
```

```
%do i=1 %to (&SIZE+1);

data uno;DO N=1 TO &REP;Y=RANNOR(&i+12345678);
OUTPUT;END;RUN;
DATA DOS; SET UNO;i=1;E=Y;RUN;
DATA TRES;SET DOS;M=0;E1=ABS(E);IF E>0 THEN M=1;RUN;
PROC SORT DATA=TRES;BY E1;RUN;
DATA CUATRO;SET TRES;N=1;H=DIF(M);H1=ABS(H);IF H1=.
THEN H1=1;NSUM+N;  RUN;
DATA CINCO;SET CUATRO;D=-1;H1SUM+H1;IF M=1 THEN D=1;
C=H1SUM*D;WIL=NSUM*M; RUN;
PROC MEANS DATA=CINCO SUM MAX NOPRINT;VAR C H1SUM M WIL;
OUTPUT OUT=SALIDA SUM=CSUM HSUM MSUM WSUM MAX=CMAX
HMAX MMAX WMAX;RUN;
DATA SALIDA;SET SALIDA;i=&I;RUN;DATA DEF&I;
SET SALIDA;C=CSUM/HMAX;S=MSUM;W=WSUM;KEEP I C S W;RUN;
DATA B.NOM31;SET DEF&I B.NOM31;RUN;

data uno;DO N=1 TO &REP;Y=RANNOR(&i+12345678);
OUTPUT;END;RUN;
DATA DOS; SET UNO;i=1;E=Y+0.1256613;RUN;
DATA TRES;SET DOS;M=0;E1=ABS(E);IF E>0 THEN M=1;RUN;
PROC SORT DATA=TRES;BY E1;RUN;
DATA CUATRO;SET TRES;N=1;H=DIF(M);H1=ABS(H);IF H1=.
THEN H1=1;NSUM+N;  RUN;
DATA CINCO;SET CUATRO;D=-1;H1SUM+H1;IF M=1 THEN D=1;
C=H1SUM*D;WIL=NSUM*M; RUN;
PROC MEANS DATA=CINCO SUM MAX NOPRINT;VAR C H1SUM M WIL;
OUTPUT OUT=SALIDA SUM=CSUM HSUM MSUM WSUM MAX=CMAX
HMAX MMAX WMAX;RUN;
DATA SALIDA;SET SALIDA;i=&I;RUN;DATA DEF&I;
SET SALIDA;C=CSUM/HMAX;S=MSUM;W=WSUM;KEEP I C S W;RUN;
DATA B.NOM32;SET DEF&I B.NOM32;RUN;

data uno;DO N=1 TO &REP;Y=RANNOR(&i+12345678);
OUTPUT;END;RUN;
DATA DOS; SET UNO;i=1;E=Y+0.2533471;RUN;
DATA TRES;SET DOS;M=0;E1=ABS(E);IF E>0 THEN M=1;RUN;
PROC SORT DATA=TRES;BY E1;RUN;
DATA CUATRO;SET TRES;N=1;H=DIF(M);H1=ABS(H);IF H1=.
 THEN H1=1;NSUM+N;  RUN;
```

```
DATA CINCO;SET CUATRO;D=-1;H1SUM+H1;IF M=1 THEN D=1;
C=H1SUM*D;WIL=NSUM*M; RUN;
PROC MEANS DATA=CINCO SUM MAX NOPRINT;VAR C H1SUM M WIL;
OUTPUT OUT=SALIDA SUM=CSUM HSUM MSUM WSUM MAX=CMAX
HMAX MMAX WMAX;RUN;
DATA SALIDA;SET SALIDA;i=&I;RUN;DATA DEF&I;
SET SALIDA;C=CSUM/HMAX;S=MSUM;W=WSUM;KEEP I C S W;RUN;
DATA B.NOM33;SET DEF&I B.NOM33;RUN;

data uno;DO N=1 TO &REP;Y=RANNOR(&i+12345678);
OUTPUT;END;RUN;
DATA DOS; SET UNO;i=1;E=Y+0.3853205;RUN;
DATA TRES;SET DOS;M=0;E1=ABS(E);IF E>0 THEN M=1;RUN;
PROC SORT DATA=TRES;BY E1;RUN;
DATA CUATRO;SET TRES;N=1;H=DIF(M);H1=ABS(H);IF H1=.
THEN H1=1;NSUM+N;  RUN;
DATA CINCO;SET CUATRO;D=-1;H1SUM+H1;IF M=1 THEN D=1;
C=H1SUM*D;WIL=NSUM*M; RUN;
PROC MEANS DATA=CINCO SUM MAX NOPRINT;VAR C H1SUM M WIL;
OUTPUT OUT=SALIDA SUM=CSUM HSUM MSUM WSUM MAX=CMAX
HMAX MMAX WMAX;RUN;
DATA SALIDA;SET SALIDA;i=&I;RUN;DATA DEF&I;
SET SALIDA;C=CSUM/HMAX;S=MSUM;W=WSUM;KEEP I C S W;RUN;
DATA B.NOM34;SET DEF&I B.NOM34;RUN;

data uno;DO N=1 TO &REP;Y=RANNOR(&i+12345678);
OUTPUT;END;RUN;
DATA DOS; SET UNO;i=1;E=Y+0.5244005;RUN;
DATA TRES;SET DOS;M=0;E1=ABS(E);IF E>0 THEN M=1;RUN;
PROC SORT DATA=TRES;BY E1;RUN;
DATA CUATRO;SET TRES;N=1;H=DIF(M);H1=ABS(H);IF H1=.
THEN H1=1;NSUM+N;  RUN;
DATA CINCO;SET CUATRO;D=-1;H1SUM+H1;IF M=1 THEN D=1;
C=H1SUM*D;WIL=NSUM*M; RUN;
PROC MEANS DATA=CINCO SUM MAX NOPRINT;VAR C H1SUM M WIL;
OUTPUT OUT=SALIDA SUM=CSUM HSUM MSUM WSUM MAX=CMAX
HMAX MMAX WMAX;RUN;
DATA SALIDA;SET SALIDA;i=&I;RUN;DATA DEF&I;
SET SALIDA;C=CSUM/HMAX;S=MSUM;W=WSUM;KEEP I C S W;RUN;
DATA B.NOM35;SET DEF&I B.NOM35;RUN;
```

```
%end;
%mend simula;
%simula
```

Bibliography

[1] G. E. Andrews, The Theory of Partitions, *London: Addison-Wesley*, 1976.

[2] G. E. Andrews, K. Eriksson, Integer Partitions, *Cambridge University Press*, 2004.

[3] A. Baklizi, A continuously adaptive rank test for the shift in location, *Australian and New Zealand Journal of Statistics* 47 (2005) 203-209.

[4] N. Balakrishnan, M.V. Koutras, Runs and Scans With Applications, *New York: Wiley*, 2002.

[5] N. Balakrishnan, Statistics for Industry and Technology, *Hardcover*, 1997.

[6] U. Bandyopadhyay, D. Dutta, Adaptive nonparametric tests for a single sample location problem, *Statistical Methodology* 4 Issue 4 (2007) 423-433.

[7] D. E. Barton, F. N. David, Multiple Runs, *Biometrika* 44 (1957) 168-178.

[8] S. Chadjiconstantinidis, M.V. Koutras, Distributions of the numbers of failures an successes in a waiting time problem, *Ann. Inst. Statist. Math* (2001) 53,3, 576-598.

[9] J. Corzo, Verallgemeinerte Runtest für Lage-und Skalen-Alternativen, *Dissertation . Universität de Dortmund. Deutschland* (1989).

[10] J. Corzo, Teoría de rachas, *Revista Colombiana de Estadística Universidad Nacional de Colombia* 19-20 (1989).

[11] J. Corzo, Pruebas insesgadas basadas en rachas para alternativas de localización y escala, *Revista Colombiana de Estadística. Universidad Nacional de Colombia* 20-21 (1990).

[12] R.L. Dobrushin, Limit theorems for a Markov chain of two states, *Izv. Akad. Nauk S.S.S.R. Ser.Mat.* 17 (1953) 291-330.

[13] H. Fernández, J. Ortiz, Pruebas de Hipotesis basadas en Secuencias, *Revista Colombiana de Estadística. Universidad Nacional de Colombia* 12 (1986).

[14] J. Fu, M. Koutras, Distribution theory of runs, Markov chain approach, *Journal of the American Statistical Association* 89 (1994) 1050-1058.

[15] J.C. Fu, W.Y. Lou, Distribution theory of runs and patterns and its applications. A finite markov chain imbedding technique, *World Scientific Pub., USA*, 2003.

[16] J. D. Gibbons, S. Chakraborti, Nonparametric Statistical Inference, *Marcel Dekker, NewYork* , 1992.

[17] W.M.Y. Goh, E. Schmutz, The number of distinct part sizes in a random integer partition, *J. Combinatorial Theor.* Series A 69 (1995) 149-158.

[18] L. Gordon, M.F. Schilling, M.S. Waterman, An extreme value theory for long head runs, *Prob. Theory Rel. Fields* 72 (1986) 279-288.

[19] Q. Han, S. Aki, Joint distributions of runs in a sequence of multi-state trials, *Ann. Inst. Statist. Math.* 51,3 (1999) 419-447.

[20] T. Hettmansperger, Statistical Inference Based On Ranks, *Jhon Wiley & Sons*, 1984.

[21] E. Ising, Adaptive Beiträge zur Theorie des Ferromagnetismus, *Z. Physik* 31 (1925) 235-288.

[22] Y. Kong, A simple method for evaluating partition functions of linear polymers, *Revista Journal of Physical Chemistry* Ser. B, 105 (2001) 10111-10114.

[23] Y. Kong, Distribution of runs and longest runs: a new generating function approach, *Journal of the American Statistical Association* 101, No. 475 (2006) 1253-1263.

[24] P.A. Koopman, Testing symmetry with a procedure, combining the sign test and the signed rank test, *Statistics NeerZandica* 33 (1979) 137-142.

[25] M.V. Koutras, V.A. Alexandrou, Runs, scans and urn model distributions: A unified Markov chain approach, *Ann. Inst. Statist. Math* 47 (1995) 743-766.

[26] H.H. Lemmer, On the distribution of the longest run in number partitions, an adaptive test for the median, *Research Report 85/1, Department of Statistics, Rand Afrikaans University* (1985a).

[27] H.H. Lemmer, On the distribution of the longest run in number partitions, A test for the median, combining the sign and signed rank tests. *Research Report 8512, Department of Statistics, Rand Afrikaans University* (1985b).

[28] H.H. Lemmer, A test for the median, combining the sign and the signed rank test, *Communications in Statistics, Simulation and Computation* 16 (1987) 621-627.

[29] H.H. Lemmer, Adaptive test for the median, *IEEE Transactions on Reliability* 42 (1993) 442-448.

[30] H. Levene, J. Wolfowitz, The covariance matrix of runs up and down, *Annals of Mathematical Statistics* 15 (1944) 58-69.

[31] T.P. McWilliams, A distribution-free test for symmetry based on a runs statistic, *Journal of the American Statistical Association* 85 (1990) 1130-1133.

[32] A.M. Mood, The distribution theory of runs, *Annals of Mathematical Statistics* 11 (1940) 367-392.

[33] A. Mood, On The Asymptotic Efficiency Of Certain Nonparametric Two Sample Tests, *Annals of Mathematical Statistics* 11 (1954) 367-392.

[34] M.S. Nikulin, P.M. Price, D. Turetayev, V.G. Voinov, An application of a multivariate discrete probability distribution for sampling survey analysis of supply chains, *Technical Report N9905, UniversitBe Victor Segalen Bordeaux 2, Bordeaux, France* (1999).

[35] M.S. Nikulin, T.I. Smirnov, V.G. Voinov, Multivariate discrete distributions induced by an urn scheme, linear diophantine equations, unbiased estimating, testing and applications, *Journal of statistical planning and inference*, Vol.101 No 1-2 (2002) 255-266 .

[36] T.W. O'Gorman, An adaptive two sample test based on modified Wilcoxon scores, *Comm. Statist.Simulation Comput.* 25 (1996) 459-479.

[37] J. Ortiz, Pruebas de hipotesis sobre parametros de localizacion basadas en secuencias, *Revista Colombiana de Estadística. Universidad Nacional de Colombia* 8 (1983) 20-33.

[38] J. Ortiz, J. Corzo, Una prueba de dispersión basada en secuencias, *Revista Colombiana de Estadística. Universidad Nacional de Colombia* 8 (1983) 34-48.

[39] A.N. Philippou, F.S. Makri, Longest success runs and Fibonacci type polynomials, *The Fibonacci Quarterly* 23 (1985) 338-346.

[40] A.N. Philippou, F.S. Makri, Successes, runs and longest runs, *Statist. Probab. Lett.* 4 (1986) 101-105.

[41] A.N. Philippou, G.E.Bergum, A.F Horadam, Distributions and Fibonacci polynomials of order k, longest runs, and reliability of consecutive k-out-of-n, *F systems. Fibonacci Numbers and Their Applications, Dordrecht.* (1986) 203-227.

[42] H. Randles, D. Wolfe, Introduction To The Theory Of Nonparametric Statistcs, *Jhon Wiley & Sons* , 1979.

[43] R.H. Randles, M.A. Fligner, G.E. Policello, D.A. Wolfe. An asymptotically distribution-free test for symmetry versus asymmetry, *Journal of the American Statistical Association* 75 (1980) 168-172.

[44] B. Richmond, A. Knopfmacher, Compositions with distinct parts, *Aequationes Mathematicae* 49 (1995) 86-97.

[45] P.W. Schaughnessy, Multiple Run Distributions: Recurrences and Critical Values, *Journal of the American Statistical Association* 76 (1981) 732-736.

[46] M. Schilling, The longest run of heads, *College Mathematics Journal* 21 (1990) 196-207.

[47] E. F. Schuster, X. Gu, One the Conditional and Unconditional Distributions of the Number of Runs in a Sample From a Multiletter Alphabet, *Communications in Statistics, Part B - Simulation and Computation* 26 (1997) 423-442.

[48] S. Schwager, Run probabilities in sequences of Markov dependent trials, *J. Amer. Statist. Assoc.* 78 (1983) 168-175.

[49] V.T. Stefanov, On run statistics for binary trials, *Journal of Statistical Planning and Inference.* 87 (2000) 177-185.

[50] W.L. Stevens, Distribution of groups in a sequence of alternatives, *Ann. Eugen* 9 (1939) 10-17.

[51] G. Szekeres, Some asymptotic formulae in the theory of partitions (I), *Quart. J. Math. Oxford 2.* 2 (1951) 85-108.

[52] G. Szekeres, Some asymptotic formulae in the theory of partitions (II), *Quart. J. Math. Oxford 2,* 4 (1953) 96-111.

[53] V.G. Voinov, M.S. Nikulin, On power series, Bell polynomials, *Hardy-Ramanujan-Rademacher problem and its statistical applications, Kybernetika* 30 (1994) 343-358.

[54] V.G. Voinov, M.S. Nikulin, Generating functions, problems of additive number theory and some statistical applications, *Rev.Roumaine Math. Pures Appl.* 40 (1995) 107-147.

[55] V.G. Voinov, M.S. Nikulin, Unbiased Estimators and Their Applications, Volume 2: Multivariate Case, *Dordrecht: Kluwer Academic Publishers* , 1996.

[56] V.G. Voinov, M.S. Nikulin, On a subset sum algorithm and its probabilistic and other applications. In: Balakrishnan, N. (Ed.), Advances in Combinatorial Methods and Applications to Probability and Statistics. BirkhVauser, *Boston*, 1997 153-163.

[57] S. Wagner, On the distribution of the longest run in number partitions, *The Ramanujan Journal* 20 (2009) 189-206.

[58] A. Wald, J. Wolfowitz, On A Tests Whether Two Samples Are From The Same Population, *Annals of mathematical Statistics* 11 (1940) 147-162.

[59] A. Wald, J. Wolfowitz, Non-Parametric Statistical Inference, *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability" University of california*, (1949) 93-113.

[60] W.A. Whitworth, Choice and Chance 5th Ed, *New York: Hafner* (1959).

[61] J. Wolfowitz, Asymptotic distribution of runs up and down, *Annals of Mathematical Statistics* 15 (1944) 163-172.