# On Exact Methods for Testing Equality of Binomial Proportions

MATSUO, Akihiko [1]

March 23, 2007

[1]Kansai University, Faculty of Economics

# abstract

The purpose of this thesis is to improve conventional exact methods for implementing fixed-level significance test of equality of several binomial proportions. There are two conventional exact methods: the conditional and unconditional tests. Many papers, including Suissa and Shuster (1985) and Mehta and Hilton (1993), have tried power comparisons between these two exact fixed-level significance tests. Suissa and Shuster (1985) discussed the advantage of the unconditional test because of its higher power in testing equality of two binomial proportions. Mehta and Hilton (1993) extended the comparison to three binomial proportions, with the aid of the network algorithm introduced by Mehta and Patel (1983), to reduce the discreteness of a test statistic and observed that the performance of the conditional test was equal to that of the unconditional test when sample sizes were as large as 80 or more. They concluded the advantage of the conditional test on the ground that the computational burden was much less compared with the unconditional test.

The application of these conventional exact methods to fixed-level significance tests may result in far smaller empirical size than the significance level and, as a result, lower power for detecting lack of fit of a model, which has been a strong incentive for adopting asymptotic methods. However, when we look into the behavior of the two exact tests, with the aid of modern computing facilities, we do find some room for improvement on them. We propose two improvements: one being the conditional test using a two-dimensional statistic and the other being the unconditional test using a modified statistic derived from the conditional distribution of a conventional test statistic. By adopting these improvements, we are able to reduce the shortage of the conventional exact tests.

Here, we confine our attention to fixed-level testing, rather than significance testing, i.e. reporting observed level of significance, which is nowadays more popular

among theoretical statisticians. Weerahandi (1995) noted that "In applications such as those in biomedical experiments fixed-level testing are not appealing even when they do exist, because the sufficiency of evidence in favor of or against a hypothesis would depend upon the prevailing circumstances and what being tested, and should be left for the other experts and decision makers to judge". Little (1989) argued that statistical inferences on contingency tables should be conditional and that fixed-level testing should be avoided on a philosophical ground, citing Cox (1984), Yates (1984) and so on, to reinforce his argument. However, fixed-level testing remains valid when rigorous "accept/reject" decision is required, such as when testing the effect of a new medicine, and therefore the effort to raise the power of fixed-level significance tests would be worth making.

We do not consider asymptotic approximationsin this thesis, although there still exist data sets in which exact calculation is infeasible despite of the development of computing facilities. For such cases, we have no choice but resorting to asymptotic approximations. Even asymptotic approximations in conditional inferences are discussed by, for example, Davison(1988). However, we are concerned with such data sets, where each data point is so expensive and vital, like in biomedical research, and where the asymptotic properties are not guaranteed. For such data sets, we should not rely on incorrect asymptotic approximations.

Recent development of computing environment, especially in the last two decades, has made various impacts on the statistical methodology: Multivariate methods, which had been computationally infeasible, has become popular, a considerable volume of statistical tables has been replaced with a set of directives (functions) on computer soft-wares and the demand for approximation methods that save computational burden has been decreasing. Now, it would be the time to exploit more efficient procedures without thinking much of computational load. The test procedures, we are going to introduce here, are more computer intensive but show higher performance than conventional exact tests. All the calculation in this thesis is accomplished on *Mathematica*, ver.3.

In Chapter 1, we describe preliminary notions of statistical test of hypothesis, well-used test statistics and conventional exact test procedures. In Chapter 2, we carry out size and power comparisons among conventional test statistics in the exact conditional test. In Chapter 3, we carry out size and power comparisons between the

exact conditional and unconditional exact tests employing one of the conventional test statistics in turn. In Chapter 4 and 5, we propose two improved exact test procedures to achieve higher power. In Chapter 4, we propose two-dimensional statistic for use in the exact conditional test. In Chapter 5, we propose modified statistic for use in the exact unconditional test. At last, in Chapter 6, we propose modified two-dimensional statistic for use in the exact unconditional test.

MATSUO, Akihiko

**Author's address**
Faculty of Economics
Kansai University
Suita, Osaka, 564-8680, JAPAN
e-mail: amatsuo@ipcku.kansai-u.ac.jp

# Contents

# Chapter 1

# Introduction

In this chapter, we describe fundamental matters: the hypothesis we are going to test, conventional test statistics and exact conditional and unconditional test procedures. We followed, for the most part, the notation of Mehta and Hilton (1993). Our description has been made utilizing Inagaki (1990), Takeuchi and Fujino (1981) and Yanagawa (1986) as reference books.

## 1.1 Hypothesis and test statistics

Let $Y_1, Y_2, \cdots, Y_k$ be a list of independent random variables, where each $Y_i$ is distributed from the binomial distribution $\mathrm{B}(n_i, \pi_i)$, and let $\boldsymbol{y} = (y_1, \cdots, y_k)$ denote the observation of $\boldsymbol{Y} = (Y_1, \cdots, Y_k)$. Throughout this thesis, bold face scripts, like $\boldsymbol{Y}$ and $\boldsymbol{y}$, refer to vectors and plain scripts refer to scalars. Now, the test of equality of binomial proportions is written as,

$$\begin{cases} \mathrm{H}_o : \pi_1 = \pi_2 = \cdots = \pi_k \equiv \pi_o \\ \mathrm{H}_A : \pi_i \neq \pi_j, \text{ for some } i \neq j \end{cases},$$

where $\pi_o$ is the unknown nuisance parameter. This test is also called as goodness-of-fit test for independence in a $2 \times k$ contingency table. We carry out the test with fixed significance level $\alpha$. We fix our attention at $\alpha = 0.05$ for our computation throughout this thesis.

We note that the existence of the sufficient statistic, $S = \sum_{i=1}^{k} Y_i$ of $\pi_o$ under $H_o$, allows the Neyman-Pearson approach to the test. That is, we have a choice of two alternative approaches, conditional and unconditional test procedures.

Before explaining exact conditional and unconditional test procedures, we would like to describe three well-used goodness-of-fit statistics, the *Pearson's* $X^2$, the *deviance*, and the *power divergence* in the specific form for binomial data. For more general form of these statistics, we refer the readers to Read and Cressie (1988).

**Pearson's $X^2$**

This statistic has long been the most frequently used statistic among the three for discrete data. The *Pearson's* $X^2$ is written, in the context of binomial trial, as

$$X^2(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = \sum_{i=1}^{k} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)},$$

where $\hat{\pi}_i$ is an estimate of $\pi_i$. The maximum likelihood estimate ( *m.l.e.* ) is commonly used. Under $H_o$, specifically,

$$X^2(\boldsymbol{y}) = \sum_{i=1}^{k} \frac{(y_i - n_i\hat{\pi})^2}{n_i\hat{\pi}(1-\hat{\pi})}, \tag{1.1}$$

where $\hat{\pi} = s/N$; $s = \sum_{i=1}^{k} y_i$ and $N = \sum_{i=1}^{k} n_i$.

**Deviance**

The *deviance* is actually the likelihood ratio statistic of postulated and saturated models and therefore available as long as the sampling distribution is explicitly specified. We describe the *deviance* in the context of a binomial trial, following the notation of Collett (1991), as follows.

$$D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = -2[\log \hat{L}_p - \log \hat{L}_s],$$

where $\hat{L}_p$ denotes the maximum likelihood of a postulated model,

$$\log \hat{L}_p = \sum_{i=1}^{k} [\ y_i \log \hat{\pi}_i + (n_i - y_i)\log(1-\hat{\pi}_i) + \log C(n_i, y_i)\ ],$$

and $\hat{L}_s$ denotes the of the saturated model,

$$\log \hat{L}_s = \sum_{i=1}^{k} [\ y_i \log y_i + (n_i - y_i)\log(n_i - y_i) - n_i \log n_i + \log C(n_i, y_i)\ ],$$

where $\hat{\pi}_i$ is the *m.l.e.* of $\pi_i$ under the postulated model and $C(n, y)$ denotes the number of distinct combinations when taking $y$ objects from $n$ distinct objects. Under $H_o$, $D$ becomes,

$$\begin{aligned} D(\boldsymbol{y}) &= 2\sum_{i=1}^{k}[y_i \log y_i + (n_i - y_i)\log(n_i - y_i)] \\ &\quad -2\sum_{i=1}^{k} n_i \log n_i - 2\{\ s\log\frac{s}{N} + (N-s)\log\frac{N-s}{N}\ \}. \end{aligned} \tag{1.2}$$

**Power divergence**

This statistic is a member of the *family of power divergence statistics*, introduced by Cressie and Read (1984), for testing goodness-of-fit for discrete multivariate data. This family, denoted by $\{PD^{\lambda}; \lambda \in \Re\}$, is written in the context of binomial trial as,

$$PD^{\lambda}(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k}\{y_i[(\frac{y_i}{n_i\hat{\pi}_i})^{\lambda} - 1] + (n_i - y_i)[(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)})^{\lambda} - 1]\}. \tag{1.3}$$

As this family of statistics includes the *Pearson's* $X^2(\lambda = 1)$ and the *deviance* $(\lambda = 0\ )$ as special cases, we are able to relate these two statistics, and moreover construct a new statistic that might have some optimum properties. We follow the recommendation of Cressie and Read (1984) of adopting $\lambda = 2/3$ and call it the *power divergence*, $PD$,

$$PD(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = \frac{9}{5}\sum_{i=1}^{k}\{y_i[(\frac{y_i}{n_i\hat{\pi}_i})^{2/3} - 1] + (n_i - y_i)[(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)})^{2/3} - 1]\}.$$

Under $H_o$, $\hat{p} = s/N$ and $PD$ becomes

$$PD(\boldsymbol{y}) = \frac{9}{5}\sum_{i=1}^{k}\{y_i[(\frac{y_i}{n_i\hat{\pi}})^{2/3} - 1] + (n_i - y_i)[(\frac{n_i - y_i}{n_i(1-\hat{\pi})})^{2/3} - 1]\}. \tag{1.4}$$

It is a well-known fact that these three statistics are asymptotically equivalent and the limiting distribution is the $\chi^2$ distribution with $n-1$ degree of freedom, see for example Read and Cressie (1988). Their performances in small or moderate sample sizes, however, are considered to be different. We will look into the size and power difference of these statistics, when sample sizes are small or moderate, in Chapter 2 and 3.

## 1.2 Exact conditional test

This procedure is based on the Neyman-Pearson approach to hypothesis testing. That is, we carry out the test conditional on the sufficient statistic, $S = \sum_{i=1}^{k} Y_i = \sum_{i=1}^{k} y_i$, of $\pi_o$ to eliminate the dependence of the distribution of $\boldsymbol{Y}$ on $\pi_o$ under $H_o$. Note that the randomized version of the exact conditional test is not adopted, since decisions should not be based on irrelevant events in practice, which is an usual principle in the literature.

Let denote the sample space of $\boldsymbol{Y}$ by $\Gamma$ and the conditional reference set by $\Gamma_s$, *i.e.* the subset of $\Gamma$ sharing $s$ as the sufficient statistic value,

$$\Gamma_s = \{\boldsymbol{y} \mid \boldsymbol{y} \in \Gamma, \ \sum_{i=1}^{k} y_i = s\}.$$

The conditional distribution of $\boldsymbol{Y}$ on $\Gamma_s$ under $H_A$ is given by,

$$\Pr_{H_A}(\boldsymbol{Y} = \boldsymbol{y} \mid \Gamma_s, \boldsymbol{\pi}) = \frac{\prod_{i=1}^{k} C(n_i, y_i)\ \pi_i^{y_i}\ (1-\pi_i)^{(n_i - y_i)}}{\sum_{\boldsymbol{z} \in \Gamma_s} \prod_{i=1}^{k} C(n_i, z_i)\ \pi_i^{z_i}\ (1-\pi_i)^{(n_i - z_i)}},$$

and under $H_o$ by,

$$\begin{aligned} \Pr_{H_o}(\boldsymbol{Y} = \boldsymbol{y} \mid \Gamma_s, \boldsymbol{\pi}) &= \frac{\prod_{i=1}^{k} C(n_i, y_i)\ \pi_o^{s}\ (1-\pi_o)^{(N-s)}}{\sum_{\boldsymbol{z} \in \Gamma_s} \prod_{i=1}^{k} C(n_i, z_i)\ \pi_o^{s}\ (1-\pi_o)^{(N-s)}} \\ &= \frac{\prod_{i=1}^{k} C(n_i, y_i)}{C(N, s)}, \end{aligned} \tag{1.5}$$

which is independent of the nuisance parameter $\pi_o$.

Let $T = T(\boldsymbol{Y})$ be a goodness-of-fit statistic, then the critical value of the test, $t_s(\alpha)$, is derived within each conditional reference set separately. The critical value for each conditional reference set $\Gamma_s$, $t_s(\alpha)$, is derived as follows,

$$t_\alpha(s) = \min\{t \in \tau_s \mid \Pr_{H_o}\{T(\boldsymbol{Y}) \geq t \mid s, \pi_o\} \leq \alpha\},$$

where $\tau_s = \{\ T(\boldsymbol{y}) \mid \boldsymbol{y} \in \Gamma_s\ \}$, and sample points whose statistic value is greater or equal to $t_s(\alpha)$ form the rejection set in $\Gamma_s$. When we denote the conditional rejection set in $\Gamma_s$ by $W_c(s)$, it is written as,

$$W_c(s) = \{\boldsymbol{y} \mid \boldsymbol{y} \in \Gamma_s, T(\boldsymbol{y}) \geq t_s(\alpha)\}, \tag{1.6}$$

and the overall rejection set of the conditional test, $W_c$, is written by,

$$W_c = \bigcup_{s=0}^{N} W_c(s). \tag{1.7}$$

The conditional size of the test, $\alpha_c(s)$, is written as,

$$\begin{aligned} \alpha_c(s) &= \Pr_{H_o}\{T(\boldsymbol{Y}) \geq t_s(\alpha) \mid \Gamma_s, \boldsymbol{\pi}\} \\ &= \sum_{\{\ \boldsymbol{y}\in\Gamma_s\ \mid\ T(\boldsymbol{y})\geq t_s(\alpha)\}} \frac{\prod_{i=1}^{k} C(n_i, y_i)}{C(N, s)}. \end{aligned}$$

As we noted, at the beginning of this section, that we did not adopt any auxiliary randomization, the conditional sizes are always no more than the nominal significance level, $\alpha$, and feared to be far short of $\alpha$ when $n_1, n_2, \cdots, n_k$ are small. Because the unconditional size function, $\alpha_c(\pi)$ is written as a weighted average of $\alpha_c(s)$,

$$\alpha_c(\pi) = \sum_{s=0}^{N} \alpha_c(s)\ \Pr_{H_o}\{S = s \mid \boldsymbol{\pi}\ \}|_{\pi_o=\pi}, \tag{1.8}$$

it is guaranteed to be no more than $\alpha$. The (unconditional) size of the conditional test is now obtained by maximizing the (unconditional) size function,

$$\alpha_c = \sup_{0\leq\pi\leq 1} \alpha_c(\pi),$$

which is also guaranteed to be no more than $\alpha$.

Thus, the size of the exact conditional test is guaranteed to be no larger than $\alpha$. This procedure is less computer intensive than the unconditional test, because we do not have to treat the test statistic distribution on the whole sample space but on each conditional reference set. However, this procedure can be conservative for the same reason just stated.

The power of the conditional test, denoted by $\beta_c(\boldsymbol{\pi})$ , is written as,

$$\beta_c(\boldsymbol{\pi}) = \Pr_{H_A}\{T(\boldsymbol{Y}) \geq t_\alpha(s) \mid \Gamma_s, \boldsymbol{\pi}\} \cdot \Pr_{H_A}\{S = s \mid \boldsymbol{\pi}\}, \tag{1.9}$$

following the notation by Mehta and Hilton (1993), although $\beta$ usually represents the probability of Type II error.

## 1.3 Exact unconditional test

This test is computationally more intensive, compared with the conditional one, since we must consider the test statistic distribution on the whole sample space when constructing a rejection set. The maximization approach is adopted to eliminate the dependence of the test on the nuisance parameter, following Suissa and Shuster (1985) and Mehta and Hilton (1993). The unconditional critical cutoff value of the exact unconditional test when $\pi_o = \pi$ is,

$$t_\alpha(\pi) = \min\{\ t \in \tau_\pi\ :\ \Pr_{H_o}\{\ T \geq t \mid \boldsymbol{\pi}\}|_{\pi_o=\pi} \leq \alpha\}, \tag{1.10}$$

where $\tau_\pi$ is the support of the unconditional distribution of $T$ for a specific value of $\pi$. We are going to eliminate the dependence of the test on $\pi_o$ by maximizing $t_\alpha(\pi)$ with respect to $\pi$,

$$t_\alpha = \sup_{0\leq\pi\leq 1}\{\ t_\alpha(\pi)\ \}. \tag{1.11}$$

Then, the rejection set of the unconditional exact test, denoted by $W_u$, is the set of sample points whose $T$ values are no less than $t_\alpha$,

$$W_u \ = \ \{\ \boldsymbol{y} \mid \boldsymbol{y} \in \Gamma, T(\boldsymbol{y}) \geq t_\alpha\ \}. \tag{1.12}$$

$W_u$ is also written as,

$$W_u \ = \ \bigcup_{s=0}^{N} W_u(s), \tag{1.13}$$

where

$$W_u(s) = \{\ \boldsymbol{y} \mid \boldsymbol{y} \in \Gamma_s, T(\boldsymbol{y}) \geq t_\alpha\ \}. \tag{1.14}$$

The size function of the unconditional test is,

$$\alpha_u(\pi) = \Pr_{H_o}\{\ T \geq t_\alpha \mid \boldsymbol{\pi}\}|_{\pi_o=\pi}, \tag{1.15}$$

which is guaranteed to be no larger than $\alpha$, owing to the maximization in (1.11). And the size of unconditional test is written as,

$$\alpha_u = \sup_{0\leq\pi\leq 1}\ \alpha_u(\pi), \tag{1.16}$$

which is again no larger than $\alpha$. For the purpose of comparing with the conditional test, we also define the conditional size of the unconditional test, denoted by $\alpha_u(s)$, as

$$\alpha_u(s) = \Pr_{H_o}\{T(\boldsymbol{y}) \geq t_\alpha \mid \Gamma_s,\ \boldsymbol{\pi}\ \}, \tag{1.17}$$

which is independent of the nuisance parameter $\pi_o$. Using this definition, the size function is rewritten as,

$$\alpha_u(\pi) = \sum_{s=0}^{N} \alpha_u(s) \cdot \Pr_{H_o}\{\ S = s\ \}|_{\pi_o=\pi}. \tag{1.18}$$

The power of the unconditional test is,

$$\beta_u(\boldsymbol{\pi}) = \Pr_{H_A}\{\ T(\boldsymbol{Y}) \geq t_\alpha \mid \boldsymbol{\pi}\}. \tag{1.19}$$

In the next two chapters, we will make some comparison among test statistics as well as test procedures we have explained in this chapter.

# Chapter 2

# Comparison in the conditional test

In this chapter, we compare the behavior of the three conventional test statistics, described in Section 1.1, in the exact conditional test, from the viewpoint of size and power of the test. We will consider the situation of three samples hereafter, as in Mehta and Hilton (1993), but we do not confine ourselves to equal sample sizes. Although their extension to three sample sizes was to reduce the discreteness of a test statistic on conditional reference sets, their intention was not fully accomplished because of the equality of sample sizes. From now on, we will remove the restriction of equal sample sizes, which seems to be more natural in biomedical research. Because it is unusual to have a data of equal sample sizes in practice. This relaxation reduces the discreteness of the distribution of a test statistic and, at the same time, increases the amount of computation, as we shall see in this chapter.

We note that, although our numerical illustrations are confined only to three sample case, our discussion hereafter is of course applicable to four or more sample cases. Before implementing size and power calculation in section 2.3, we shall look into the functional form of the three statistics as well as the relation among the values of them in section 2.1, and in section 2.2 we present some useful tips for saving the amount of computation.

The results presented in this chapter are based on Matsuo (1999).

## 2.1 Preliminary survey

At first, we observe the functional form of the three test statistics, as functions of sample sizes.

When we expand the numerator of the *Pearson's* $X^2$ given by (1.1), we have

$$X^2(\boldsymbol{y}) = \frac{1}{\hat{\pi}(1-\hat{\pi})} \sum_{i=1}^{k} \frac{y_i^2}{n_i} + \frac{N\hat{\pi}}{1-\hat{\pi}}, \tag{2.1}$$

where $\sum_{i=1}^{k} y_i^2 \ / \ n_i$ is the only element that may vary on $\Gamma_s$, which means that this statistic is symmetric with respect to $y_i$ and $y_j$ if $n_i = n_j$. In the extreme case that sample sizes are identical, the only element which may vary on $\Gamma_s$ becomes $\sum_{i=1}^{k} y_i^2$, which is symmetric with respect to $\boldsymbol{y}$.

In the case of the *deviance*, the only part of (1.2) that varies on $\Gamma_s$ is

$$\sum_{i=1}^{k} \{y_i \log y_i + (n_i - y_i) \log(n_i - y_i)\}, \tag{2.2}$$

and if sample sizes are identical, this part becomes

$$\sum_{i=1}^{k} \{y_i \log y_i + (n - y_i) \log(n - y_i)\}, \tag{2.3}$$

which is also symmetric with respect to $\boldsymbol{y}$.

The *power divergence* is the most complicated of the three. The entire expression, not a part of it,

$$PD(\boldsymbol{y}) = \frac{9}{5} \sum_{i=1}^{k} \{y_i[(\frac{y_i}{n_i\hat{\pi}})^{2/3} - 1] + (n_i - y_i)[(\frac{n_i - y_i}{n_i(1-\hat{\pi})})^{2/3} - 1]\}, \tag{2.4}$$

varies on $\Gamma_s$, and when sample sizes are equal, this becomes,

$$PD(\boldsymbol{y}) = \frac{9}{5} \sum_{i=1}^{k} \{y_i[(\frac{y_i}{n\hat{\pi}})^{2/3} - 1] + (n - y_i)[(\frac{n - y_i}{n(1-\hat{\pi})})^{2/3} - 1]\}, \tag{2.5}$$

which again is symmetric with respect to $\boldsymbol{y}$.

That sample sizes are equal intrinsically gives tied statistic values among, at least, the observations whose combination is identical and therefore makes the statistic distribution discrete. Another less relevant factor that gives rise to tied statistic values

is the simplicity of a test statistic. It can be said from the functional forms that the *Pearson's* $X^2$ is the simplest and the *power divergence* is the most complex, which coincides our computational experiences that the *Pearson's* $X^2$ tends to give tied value than the *deviance*, and that the *deviance* slightly than the *power divergence*.

Next, we observe how each test statistic has its values on a conditional reference set. For this purpose, we give scatter-plot diagrams with a test statistic value being assigned as x-axis and another as y-value. The following Figure 2.1 show the typical pattern of the relation among three test statistics when sample sizes are identical. The readers could observe the overall relationship by looking at the left column of the figure: the *deviance* (*Dev* in short) tends to have larger values compared to both the *Pearson's* $X^2$ (*PX* in short ) and the *power divergence* (*PD* in short ), and *PD* is slightly larger than *PX*. We note that, in exact tests, the order of a statistic value is relevant, unlike in approximation tests, where the value itself is relevant. The right column of the figure is presented for looking into the points, on which each statistic values are less than 10, a range anticipated to include the 5% critical value of the test.

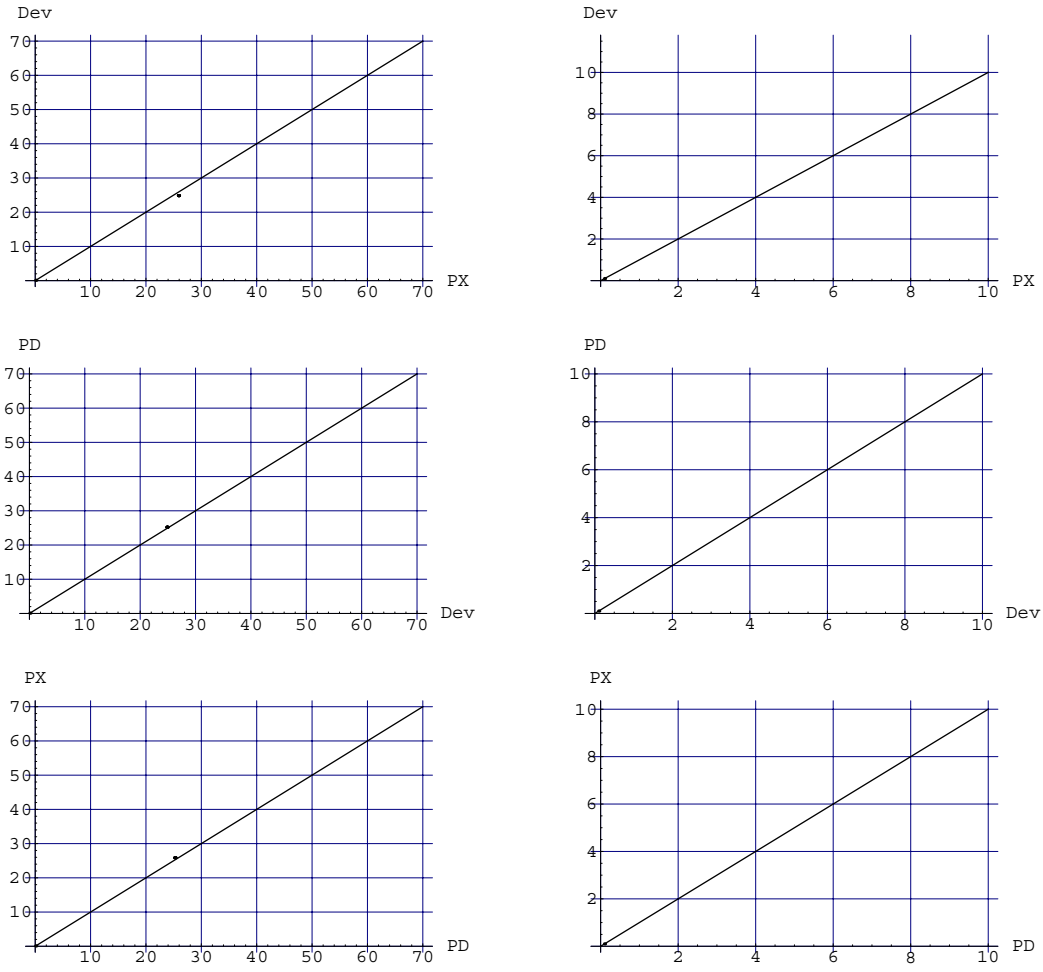



Figure 2.1: Scatter-plots illustrating the relation among the three test statistics, when $\boldsymbol{n} = \{40, 40, 40\}$ and $s = 25$.

We could observe that $PX$ and $PD$ are similar, that is, the points of the third row plots of Figure 2.1 are almost on the line $y = x$, while the *deviance* is different from the other two statistics. When sample sizes are identical, $PD$ distinguishes observations that $PX$ doesn't, which makes $PD$ less discrete than $PX$. When sample sizes are distinct, $PX$, $Dev$ and $PD$ show almost the same degree of discreteness and $PD$ is expected to show intermediate performance because of its definition. Figure 2.2 is presented to show the tendency of the three statistics values when sample sizes are distinct, that is, the discreteness of each statistic distribution is considerably relaxed.

We could also use minus the log conditional probability, $-\log \Pr\{\boldsymbol{Y} = \boldsymbol{y} \mid \Gamma_s\}$, as a test statistic, following Freeman and Halton (1951). However we have not adopted this statistic for the two reasons described below, although it is intuitively preferable to set up an acceptance/rejection set in descending/ascending probability order, so that we can anticipate the conditional size to be nearer to a fixed significance level compared to the three statistics. The first reason is that the value of $-\log \Pr\{\boldsymbol{Y} = \boldsymbol{y} \mid \Gamma_s\}$ depends on the number of elements in $\Gamma_s$, $\#\Gamma_s$, and so this statistic cannot be used in the context of unconditional test because $\#\Gamma_s$ varies with $s$. The second reason is that, from our computational experiences, the descending order of $-\log \Pr\{\boldsymbol{Y} = \boldsymbol{y} \mid \Gamma_s\}$ is quite similar to that of the *deviance* as far as constructing acceptance set, and therefore rejection set, which means tests using the *deviance* and $-\log \Pr\{\boldsymbol{Y} = \boldsymbol{y} \mid \Gamma_s\}$ are similar. Figure 2.3 is presented to show this consistent phenomenon. The plots of the left column display the overall relation of the three statistics values and $-\log \Pr\{\boldsymbol{Y} = \boldsymbol{y} \mid \Gamma_s\}$ values. The plots of the right column are presented for looking into the points, on which both statistics values are less than 10, that is, the set of points including the acceptance set of the test. We can easily observe from the right plot of the second row of Figure 2.3 that the *deviance* almost preserves the ascending order given by $-\log \Pr\{\boldsymbol{Y} = \boldsymbol{y} \mid \Gamma_s\}$ in the acceptance set.

As sample sizes grow larger and $s$ become closer to $N/2$, the values of the three statistics should become close together because the three statistics are asymptotically equivalent. Figure 2.4 is presented to illustrate this fact. The left column of the figure shows the overall relation among the three statistics values. We can easily observe from the right column of Figure 2.4 that all the points, on which statistics

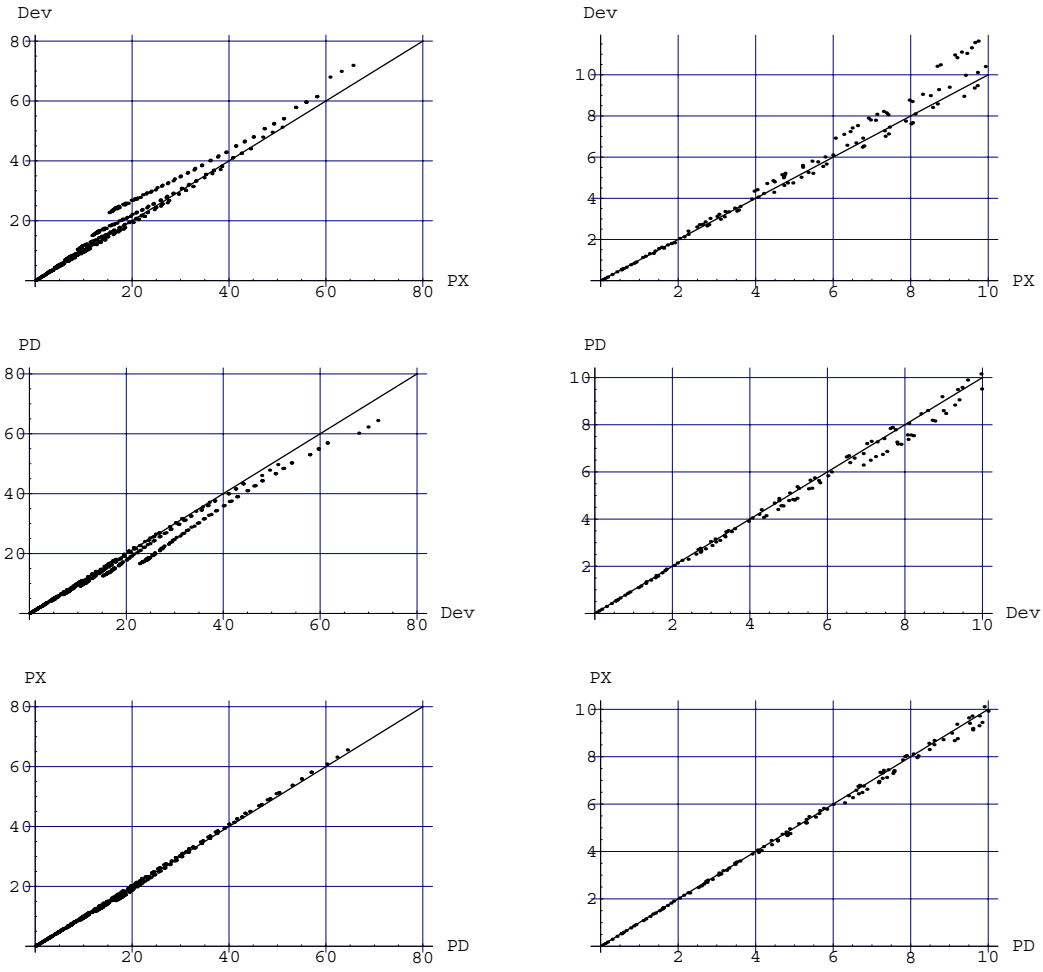



Figure 2.2: Scatter-plots illustrating the relation among the three test statistics, when $\boldsymbol{n} = \{39, 40, 41\}$ and $s = 25$.

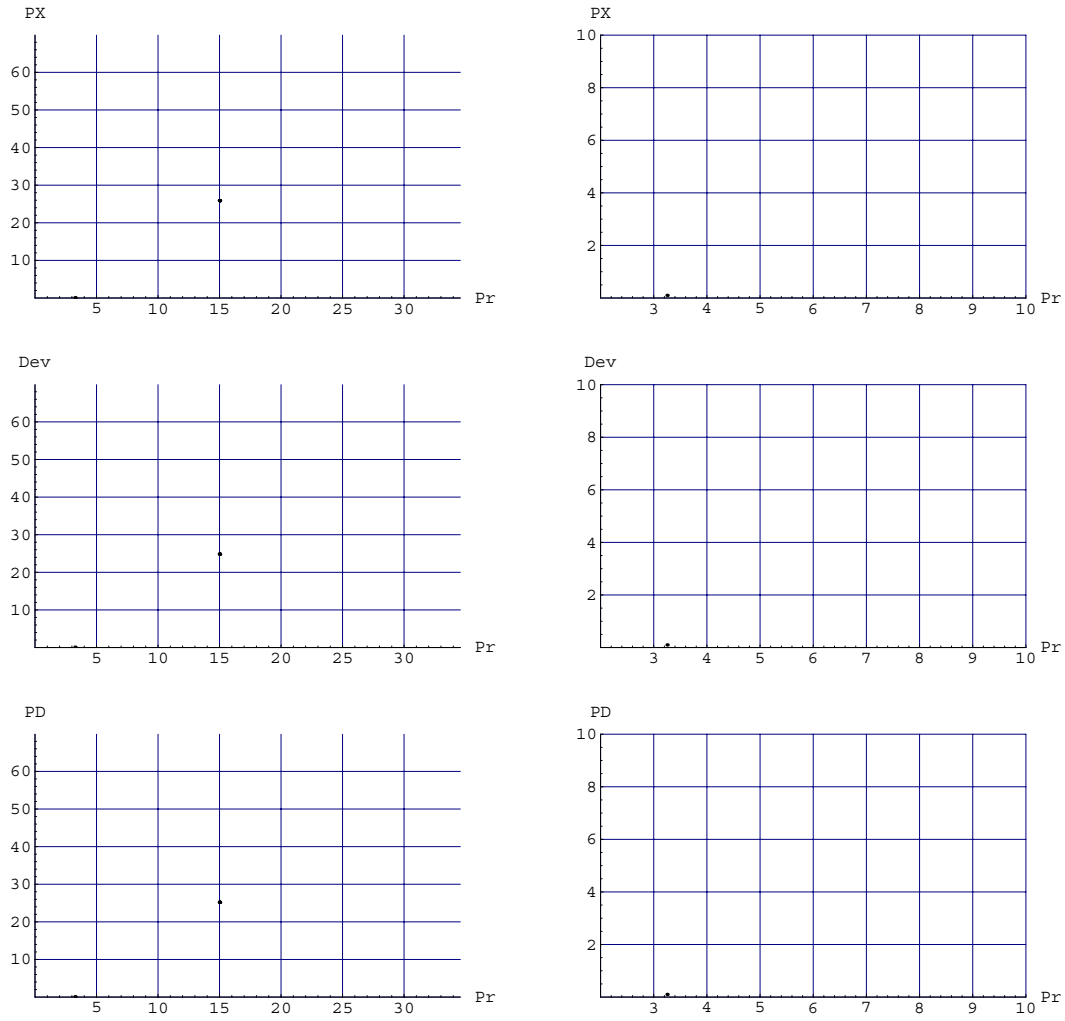



Figure 2.3: Scatter-plots illustrating the relation among the three test statistics and minus the log conditional probability, when $\boldsymbol{n} = \{40, 40, 40\}$ and $s = 25$.

values are less than 10, lie actually on the line $y = x$, which indicate the fact that the distribution functions of the three statistics are very difficult to distinguish up to 10, and therefore over 10. In other words, the three test statistics yield almost the same result in the exact conditional test.

Apparent differences in performance, by changing test statistics, could be observed only when sample sizes are relatively small, which is usual in biomedical research.

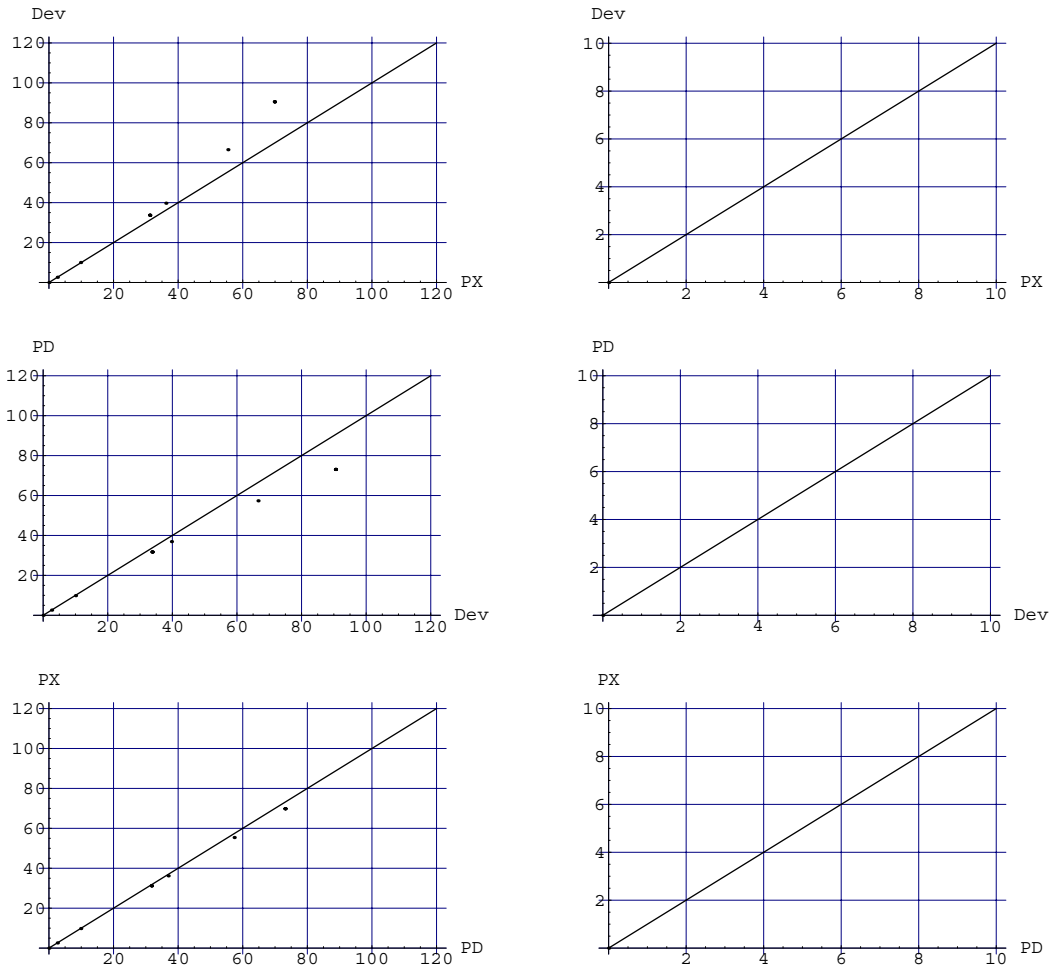



Figure 2.4: Scatter-plots illustrating the relation among the three test statistics, when $\boldsymbol{n} = \{40, 40, 40\}$ and $s = 60$.

## 2.2 Some useful tips in computation

As seen in the previous section, the discreteness of a test statistics varies depending on the set of $n_1, \cdots, n_k$ values. If all sample sizes are equal, any statistic will surely suffer from discreteness unless the size is large enough, say 80 or more. On the other hand, if all of them are distinct, much less discreteness is expected. We will fix $k = 3$ and focus only on the cases, where $n_1, n_2, n_3$ are similar varying from $n_i = 10$ to 50 with the interval of 10.

On each setting of $\boldsymbol{n}$, we calculate size functions for the three statistics and carry out power computations for a selection of alternatives, to display them in graphs. Before listing the alternatives, in the following section, we will give some useful tips which would save the amount of computation.

First of all, what we construct is not *rejection* sets but *acceptance* sets, to save the amount of computation. The number of elements in an *acceptance* set is much smaller than the corresponding *rejection* set. Note that there is a natural one-to-one correspondence between $\Gamma_s$ and $\Gamma_{N-s}$, that is, between $\boldsymbol{y} \in \Gamma_s$ and $\boldsymbol{n} - \boldsymbol{y} \in \Gamma_{N-s}$. Moreover, from the definition (1.3),

$$PD^\lambda(\boldsymbol{y}) \equiv PD^\lambda(\boldsymbol{n} - \boldsymbol{y})$$

holds in the *family of power divergence statistics*. Together with the equation,

$$\Pr(\boldsymbol{y}|\Gamma_s) = \frac{\prod_{i=1}^k C(n_i, y_i)}{C(N, s)} = \frac{\prod_{i=1}^k C(n_i, n_i - y_i)}{C(N, N-s)} = \Pr(\boldsymbol{n} - \boldsymbol{y}|\Gamma_{N-s}),$$

we can conclude that, for any power divergence statistic, the distributions on $\Gamma_s$ and $\Gamma_{N-s}$ are identical. We can, also, derive some properties listed below, which would help us reduce the amount of calculation.

**(p1)** Let $W_c^\lambda(s)$ be the conditional rejection set in $\Gamma_s$ using $PD^\lambda$ as a test statistic and $W^\lambda = \cup_{s=1}^{N-1} W_u^\lambda(s)$ be the over-all rejection set, then $W_c^\lambda \ni \boldsymbol{y} \Longrightarrow W_c^\lambda \ni \boldsymbol{n} - \boldsymbol{y}$ for any $\lambda \in \Re$. Together with the equation,

$$\Pr(\boldsymbol{y}|\boldsymbol{\pi}) \;=\; \Pr(\boldsymbol{n} - \boldsymbol{y}|\boldsymbol{1} - \boldsymbol{\pi}),$$

we have

$$\Pr(\boldsymbol{y}|\boldsymbol{\pi}) + \Pr(\boldsymbol{n} - \boldsymbol{y}|\boldsymbol{\pi}) \;=\; \Pr(\boldsymbol{y}|\boldsymbol{1} - \boldsymbol{\pi}) + \Pr(\boldsymbol{n} - \boldsymbol{y}|\boldsymbol{1} - \boldsymbol{\pi}).$$

So, we can conclude that the power at a simple alternative, $\boldsymbol{\pi}$, is identical with that at $\mathbf{1} - \boldsymbol{\pi}$.

**(p2)** When $n_1 = \cdots = n_k$, $PD^\lambda$ is constant over the permutations of $\boldsymbol{y}$, which means any permutation of $\boldsymbol{y}$ is included in $W_c^\lambda$, if $\boldsymbol{y} \in W_c^\lambda$. Therefore, it is enough to carry out power calculations only for the alternatives satisfying $\pi_1 \leq \pi_2 \leq \cdots \leq \pi_k$. When $n_1, \cdots, n_k$ are not equal but slightly different, we will treat the power difference arises from permuting $\pi_1, \pi_2, \cdots, \pi_k$ as trivial and will report the average power over all the permutations.

**(p3)** Let $\alpha_c^\lambda(s)$ be the conditional size of the conditional test when $S = s$ using $PD^\lambda$, then $\alpha_c^\lambda(s) = \alpha_c^\lambda(N - s)$ holds for any $\lambda \in \Re$ and the unconditional size (null power) function,

$$\alpha_c^\lambda(\pi) = \sum_{s=0}^{N} \alpha_c^\lambda(s) C(N, s) \pi^s (1 - \pi)^{N-s},$$

is symmetric at $\pi = 0.5$ for any $\lambda \in \Re$.

The number of elements in $\Gamma_s$ ( for $s \leq [(N+1)/2]$, where $[x]$ denotes the largest integer that is no larger than $x$ ) , $\sharp\Gamma_s$, is calculated as follows,

$$\begin{array}{rcl} \sharp\Gamma_s & = & C(s+k-1, k-1) \\ & & -\sum_{i=1}^{k} C(s-(n_i+1)+k-1, k-1) \\ & & +\sum_{i<j}^{k} C(s-(n_i+1)-(n_j+1)+k-1, k-1) \\ & & -\sum_{i<j<l}^{k} C(s-(n_i+1)-(n_j+1)-(n_l+1)+k-1, k-1) \\ & & +\cdots \end{array},$$

with the convention that $C(m, l) = 0$ for non-positive $m$. For $k = 3$ and when $n_1, n_2, n_3$ are similar, especially,

$$\sharp\Gamma_s = C(s+k-1, k-1) - \sum_{i=1}^{3} C(s-(n_i+1)+k-1, k-1).$$

We can expect the number of distinct values of the *power divergence* on $\Gamma_s$ to be $\sharp\Gamma_s$, by setting $n_1, \cdots, n_k$ to be distinct. On the other extreme, we would expect only about one sixth of $\sharp\Gamma_s$ when we set $n_1, \cdots, n_k$ to be equal and select the *Pearson's* $X^2$. On the latter setting, Mehta and Hilton (1993) tried to compare the conditional and

unconditional tests. The result is that the sizes of the conditional tests fall short of the fixed significance level and, as a consequence, the conditional test has less power to detect lack-of-fit than the unconditional test, when sample sizes are small.

## 2.3 Numerical results

We will discuss the following three cases separately: the first case being all sample sizes are equal, $\boldsymbol{n} = (10, 10, 10)$, $(20, 20, 20)$, $(30, 30, 30)$, $(40, 40, 40)$ and $(50, 50, 50)$ , the second case being two of three sample sizes are equal, $\boldsymbol{n} = (10, 10, 11)$, $(20, 20, 21)$, $(30, 30, 31)$, $(40, 40, 41)$ and $(50, 50, 51)$, and the third case being all of them are distinct, $\boldsymbol{n} = (9, 10, 11)$, $(19, 20, 21)$, $(29, 30, 31)$, $(39, 40, 41)$ and $(49, 50, 51)$. In the context of the test of equal binomial proportions, they usually treat only the first case, that is, $n_1 = \cdots = n_k$. So, our setting might be slightly more extensive than usual ones, and practical as well. The power calculation is carried out at the collection of simple alternatives, listed below, in the prescribed order.

$$
\begin{array}{rrrr}
1 : (0.1, 0.2, 0.2), & 2 : (0.1, 0.3, 0.3), & 3 : (0.1, 0.4, 0.4), & 4 : (0.1, 0.5, 0.5), \\
5 : (0.2, 0.3, 0.3), & 6 : (0.2, 0.4, 0.4), & 7 : (0.2, 0.5, 0.5), & 8 : (0.2, 0.6, 0.6), \\
9 : (0.3, 0.4, 0.4), & 10 : (0.3, 0.5, 0.5), & 11 : (0.3, 0.6, 0.6), & 12 : (0.4, 0.5, 0.5), \\
13 : (0.1, 0.3, 0.4), & 14 : (0.1, 0.4, 0.5), & 15 : (0.2, 0.5, 0.6), & 16 : (0.2, 0.4, 0.5), \\
17 : (0.3, 0.5, 0.6), & 18 : (0.1, 0.2, 0.3), & 19 : (0.2, 0.3, 0.4), & 20 : (0.3, 0.4, 0.5), \\
21 : (0.4, 0.5, 0.6), & 22 : (0.1, 0.3, 0.5), & 23 : (0.2, 0.4, 0.6), & 24 : (0.3, 0.5, 0.7), \\
25 : (0.3, 0.4, 0.7), & 26 : (0.4, 0.4, 0.6), & 27 : (0.3, 0.3, 0.7), & 28 : (0.3, 0.4, 0.6), \\
29 : (0.2, 0.3, 0.5), & 30 : (0.2, 0.3, 0.6), & 31 : (0.1, 0.2, 0.5), & 32 : (0.1, 0.2, 0.4), \\
33 : (0.4, 0.4, 0.5), & 34 : (0.3, 0.3, 0.6), & 35 : (0.3, 0.3, 0.5), & 36 : (0.3, 0.3, 0.4), \\
37 : (0.2, 0.2, 0.6), & 38 : (0.2, 0.2, 0.5), & 39 : (0.2, 0.2, 0.4), & 40 : (0.2, 0.2, 0.3), \\
41 : (0.1, 0.1, 0.5), & 42 : (0.1, 0.1, 0.4), & 43 : (0.1, 0.1, 0.3), & 44 : (0.1, 0.1, 0.2).
\end{array}
$$

As noted in **(p1)** and **(p2)** in the previous section, the power calculated at $\boldsymbol{\pi} = (0.1, 0.2, 0.3)$ is identical to that given at each permutation of $(0.1, 0.2, 0.3)$ and $(0.7, 0.8, 0.9)$. We also omit all the alternatives where

$$\max(\pi_1, \pi_2, \pi_3) - \min(\pi_1, \pi_2, \pi_3) \geq 0.5,$$

because the power differences at such alternatives are found to be trivial.

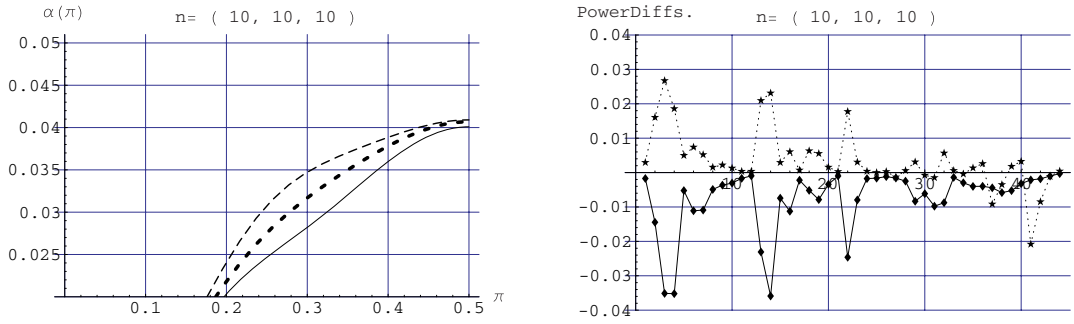
The alternatives listed above, which are arranged after observing numerical results, can be classified into 6 groups. The first group consists of 12 alternatives, from

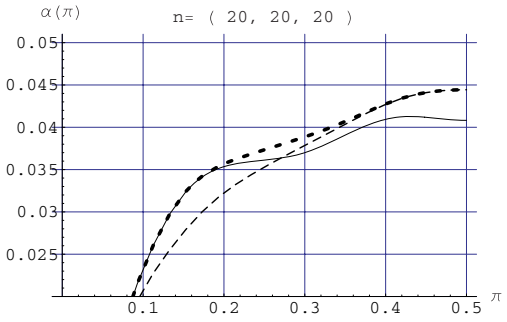
$1 : (0.1, 0.2, 0.2)$ to $12 : (0.4, 0.5, 0.5)$ where $\pi_1 < \pi_2 = \pi_3$. The sixth group consists of 12 alternatives, from $33 : (0.4, 0.4, 0.5)$ to $44 : (0.1, 0.1, 0.2)$ where $\pi_1 = \pi_2 < \pi_3$. There is a $1 - 1$ correspondence, *e.g.*, $1 : (0.1, 0.2, 0.2)$ and $44 : (0.1, 0.1, 0.2)$, between the first and sixth groups. The second group consists of 5 alternatives, from $13 : (0.1, 0.3, 0.4)$ to $17 : (0.3, 0.5, 0.6)$, where $\pi_3 - \pi_2 < \pi_2 - \pi_1$. The fifth group consists of 5 alternatives, from $28 : (0.3, 0.4, 0.6)$ to $32 : (0.1, 0.2, 0.4)$, where $\pi_3 - \pi_2 > \pi_2 - \pi_1$. There is a $1 - 1$ correspondence, *e.g.* $13 : (0.1, 0.3, 0.4)$ and $32 : (0.1, 0.2, 0.4)$, between the second and fifth groups. The third group consists of 7 alternatives, from $18 : (0.1, 0.2, 0.3)$ to $24 : (0.3, 0.5, 0.7)$, where $\pi_3 - \pi_2 = \pi_2 - \pi_1$. The fourth group consists three alternatives, from $25 : (0.3, 0.4, 0.7)$ to $27 : (0.3, 0.3, 0.7)$, each of which could be classified into the fifth or sixth group but has no counter part in the correspondent group.

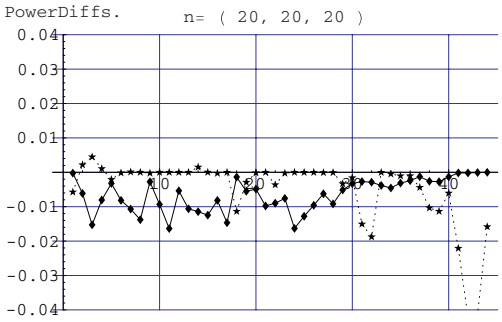
Now we consider the first case, $n_1 = n_2 = n_3$, where the discreteness of a statistic is most feared. Figure 2.5 consists of five portions, $\boldsymbol{n} = (10, 10, 10)$, $\boldsymbol{n} = (20, 20, 20)$, $\boldsymbol{n} = (30, 30, 30)$, $\boldsymbol{n} = (40, 40, 40)$ and $\boldsymbol{n} = (50, 50, 50)$, each portion has two graphs (one graph, located left, is for representing the unconditional size functions of the three statistics, with solid line representing $PX$, broken line $Dev$ and bold dotted line $PD$. And the other graphs, located right, is for representing the power differences,

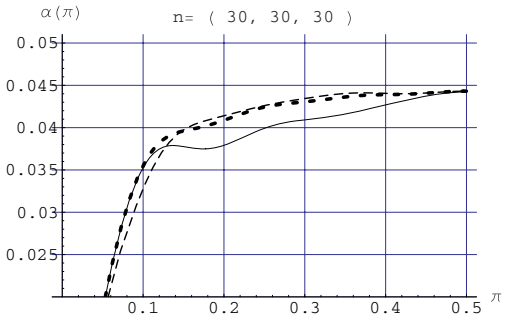
$$\beta_c^{PX}(\boldsymbol{\pi}) - \beta_c^{PD}(\boldsymbol{\pi}) \quad \text{and} \quad \beta_c^{Dev}(\boldsymbol{\pi}) - \beta_c^{PD}(\boldsymbol{\pi}),$$

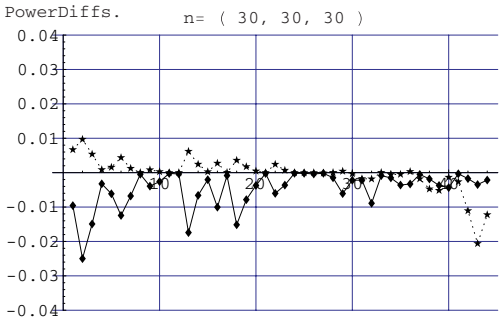
calculated at the simple alternatives in the same order as listed at the beginning of this section, with ⧫ representing $\beta_c^{PX}(\boldsymbol{\pi}) - \beta_c^{PD}(\boldsymbol{\pi})$ and $\star$ representing $\beta_c^{Dev}(\boldsymbol{\pi}) - \beta_c^{PD}(\boldsymbol{\pi})$ .

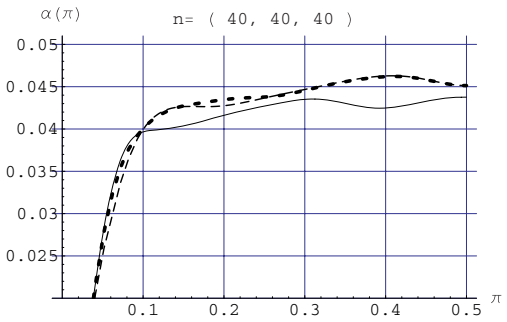


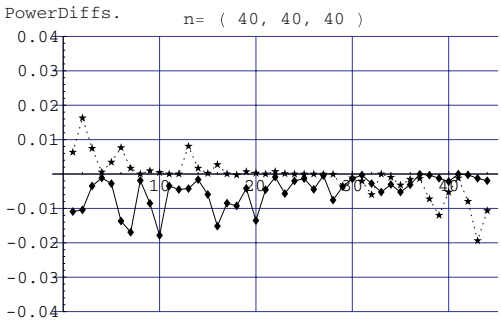

α(π)
n= ( 20, 20, 20 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 20, 20, 20 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40


α(π)
n= ( 30, 30, 30 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 30, 30, 30 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40


α(π)
n= ( 40, 40, 40 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 40, 40, 40 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40

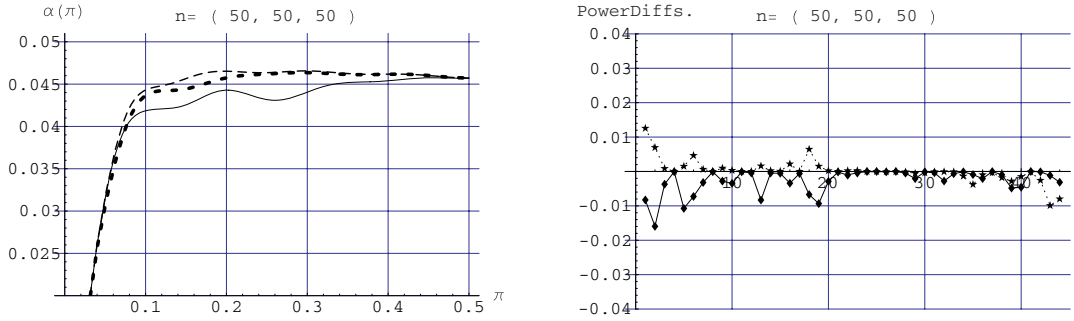
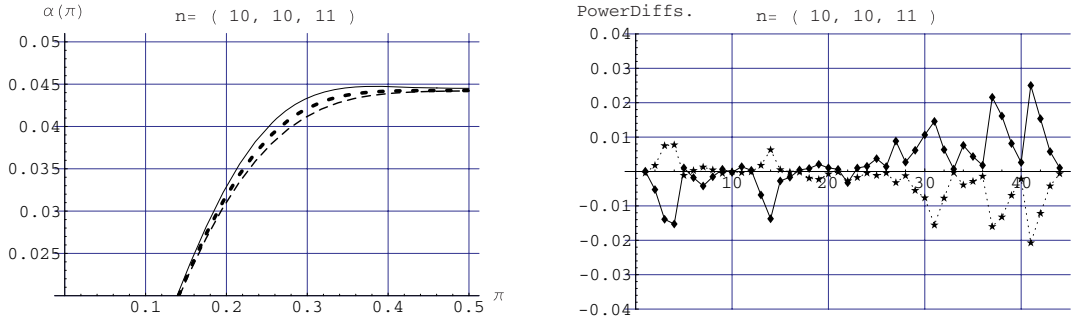
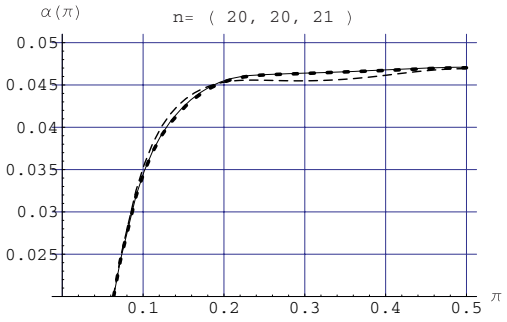


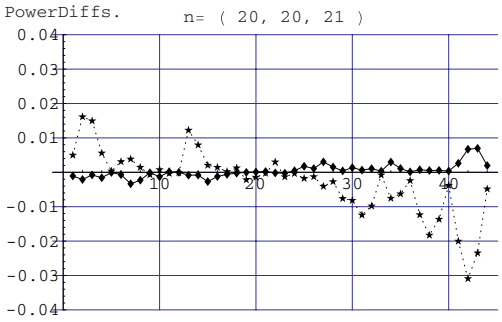

Figure 2.5: Size functions and power differences when $n_1 = n_2 = n_3$.

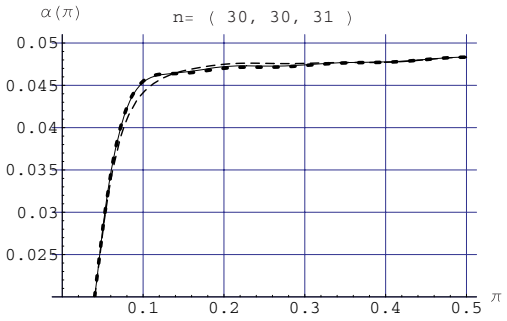
We can easily observe that $PD$ is always more powerful than $PX$. This superiority of $PD$ over $PX$ is consistently observed from $n_1 = n_2 = n_3 = 7$ to 55 with an increment of 2. On the other hand, $Dev$ has a tendency that it is more powerful than $PD$ at the alternatives in groups 1 and 2 and less powerful in groups 5 and 6. This tendency is consistently observed around $n_1 = n_2 = n_3 = 25$ or larger.

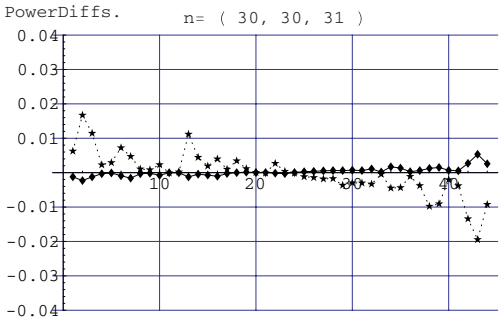
Next, we consider the second case, $n_1 = n_2 = n_3 - 1$, where the discreteness of a statistic is less expected than the first case, $n_1 = n_2 = n_3$. Figure 2.6, just like Figure 2.5, consists of five portions, $n = (10, 10, 11)$ , $n = (20, 20, 21)$, $n = (30, 30, 31)$, $n = (40, 40, 41)$ and $n = (50, 50, 51)$.

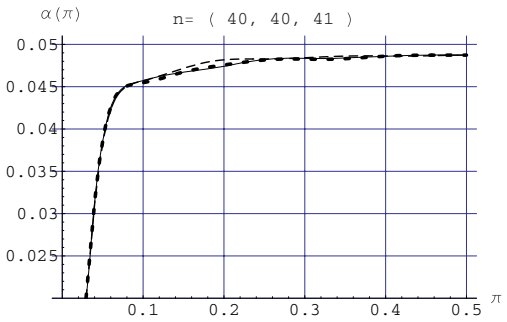


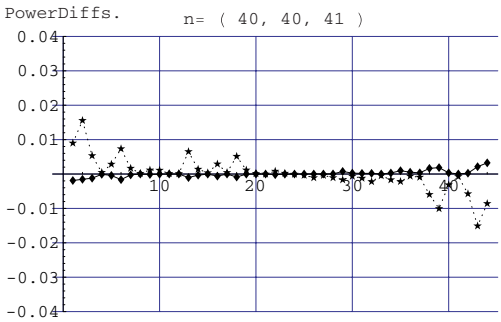

α(π)
n= ( 20, 20, 21 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 20, 20, 21 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40


α(π)
n= ( 30, 30, 31 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 30, 30, 31 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40


α(π)
n= ( 40, 40, 41 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 40, 40, 41 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40

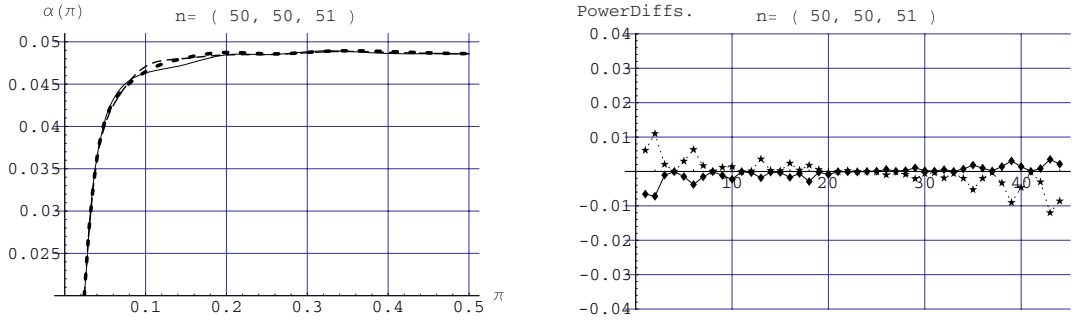
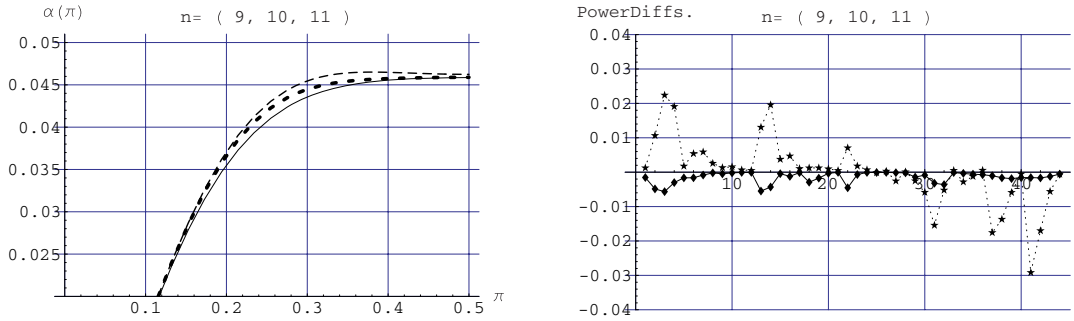
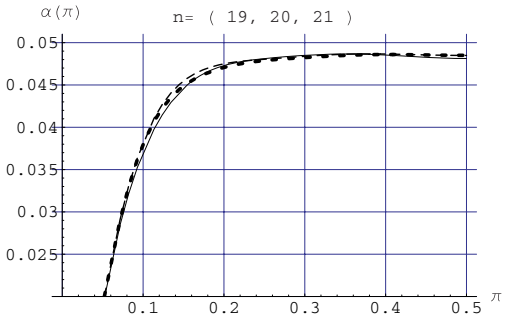


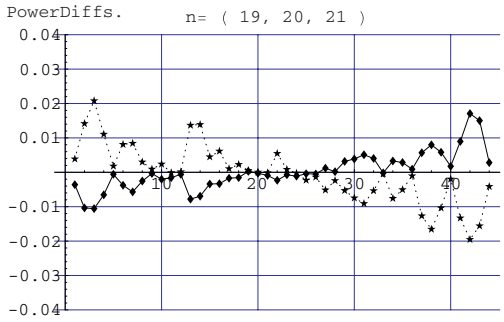

Figure 2.6: Size functions and power differences when $n_1 = n_2 = n_3 - 1$.

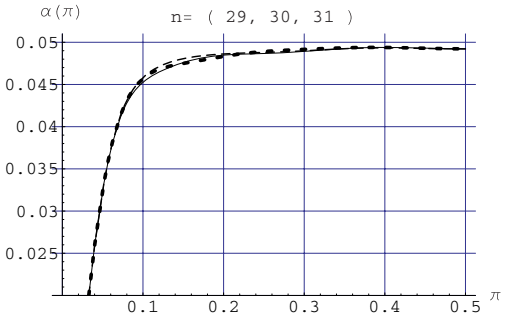
As sample sizes become larger, discreteness of the three statistics becomes less serious and size functions of the conditional test grow faster toward the nominal significance level, $\alpha = 0.05$, compared to the previous case. $PX$ shows a tendency of being most powerful in groups 5 and 6, while least powerful in groups 1 and 2. $Dev$ shows an opposite tendency to that of $PX$. $PD$ usually shows intermediate performance over all the alternatives. These tendencies become stable around $n_i \doteq 25$ or larger. The power differences become smaller as sample sizes grow larger.

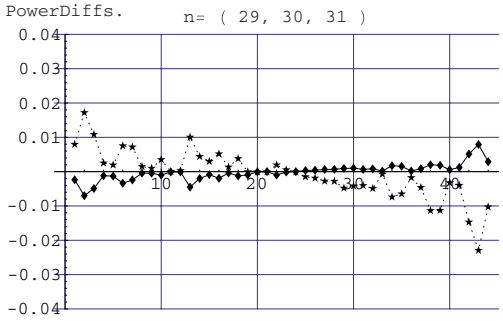
At last, we consider the third case, $n_1 + 1 = n_2 = n_3 - 1$, where the discreteness of the statistics is least serious. Figure 2.7, just as in Figures 2.5 and 2.6, consists of five portions, $n = (9, 10, 11)$, $n = (19, 20, 21)$, $n = (29, 30, 31)$, $n = (39, 40, 41)$ and $n = (49, 50, 51)$.

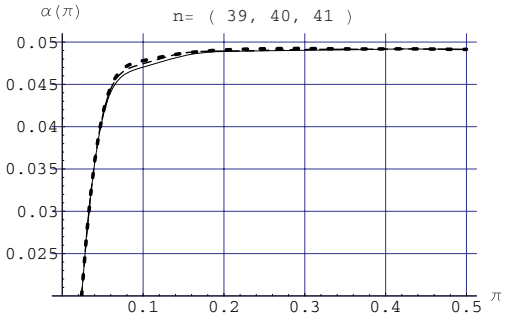


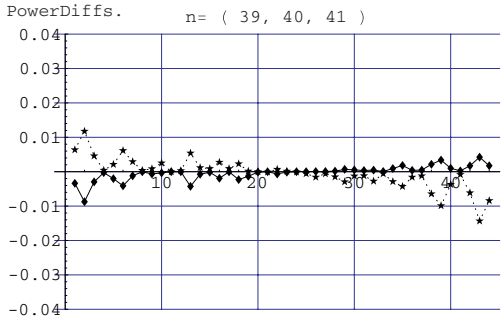

α(π)
n= ( 19, 20, 21 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 19, 20, 21 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40


α(π)
n= ( 29, 30, 31 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 29, 30, 31 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40


α(π)
n= ( 39, 40, 41 )
0.05
0.045
0.04
0.035
0.03
0.025
0.1
0.2
0.3
0.4
0.5
π


PowerDiffs.
n= ( 39, 40, 41 )
0.04
0.03
0.02
0.01
-0.01
-0.02
-0.03
-0.04
10
20
30
40

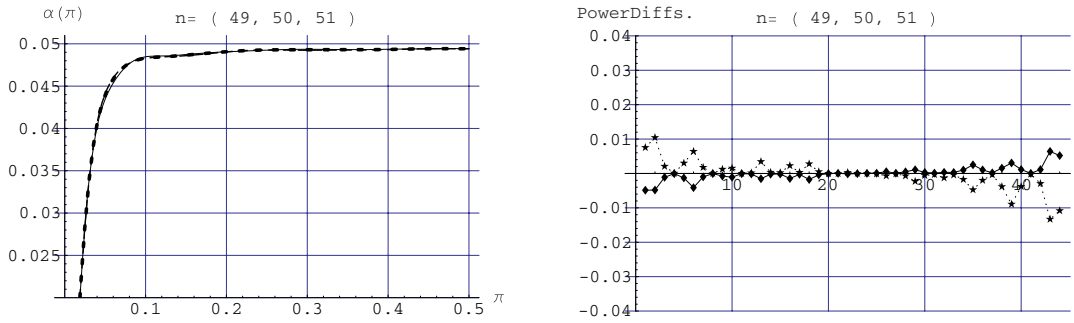



Figure 2.7: Size functions and power differences when $n_1 + 1 = n_2 = n_3 - 1$.

The tendencies observed in the previous two cases are more clearly seen; size functions grow faster toward $\alpha = 0.05$ than the previous two cases, $Dev$ is most powerful in groups 1 and 2, and $PX$ is most powerful in groups 5 and 6. We can also observe that the power differences among the three statistics become smaller as sample sizes become larger. We have carried out additional calculation for the cases $0 \leq n_2 - n_1 \leq 5$ and $0 \leq n_3 - n_2 \leq 5$, varying from $n_2 = 10$ to 50, and observed similar tendencies among the three statistics.

We anticipated that $PD$ would be the most powerful of the three, because of its less discreteness owing to its complex functional form. It is true that $PD$ is uniformly more powerful when $n_1, n_2, n_3$ are equal and alternatives are in groups 5 and 6, but $PD$ failed to attain uniform superiority over $Dev$, which has a close relationship with conditional probability. In the context of a conditional test, where only the order of the statistic values in a conditional reference set is relevant, we did not expect any power tendency like that found by Taneichi and Sekiya (1995) in the context of an unconditional test. But after carrying out intensive computation, we found that there exists a power tendency as below. $Dev$ is most powerful of the three test statistics against the alternatives in groups 1 and 2; especially when $\boldsymbol{\pi} = (0.1, 0.2, 0.2)$, $(0.1, 0.3, 0.3)$, $(0.1, 0.4, 0.4)$, $(0.2, 0.3, 0.3)$, $(0.2, 0.4, 0.4)$ and $(0.2, 0.5, 0.5)$. $PX$ is most powerful in groups 5 and 6; especially when $\boldsymbol{\pi} = (0.1, 0.1, 0.2)$, $(0.1, 0.1, 0.3)$, $(0.1, 0.1, 0.4)$, $(0.2, 0.2, 0.4)$ and $(0.2, 0.2, 0.5)$, except for the cases where serious discreteness is expected. It is therefore recommended to use $Dev$ or $PX$ depending on the assumed alternatives. Here we note again that the tendency seen at $\boldsymbol{\pi} = (0.1, 0.1, 0.2)$(*e.g.*) is also seen at all the permutations of $\boldsymbol{\pi} = (0.1, 0.1, 0.2)$ and

$\mathbf{1} - \boldsymbol{\pi} = (0.9, 0.9, 0.8)$, as explained in **(p1)** and **(p2)** in section 2.2. The performance of $PD$ is, usually, intermediate between the other two, which seems to be a natural consequence of its functional form. If there is no priority in the choice of alternative, it is recommended to choose the *power divergence* as a goodness-of-fit statistic.

# Chapter 3

# Conditional versus unconditional test

In this chapter, we carry out size and power comparisons between the exact conditional and unconditional tests, employing one of the three test statistics in turn. Suissa and Shuster (1985) carried out a comparison between exact conditional and unconditional procedures for testing equality of two binomial proportions. Mehta and Hilton (1993) showed that, by extending the comparison to three binomial proportions with all sample sizes are equal, the power of the conditional test almost equals to that of the unconditional test when sample sizes are 80 or more. They concluded that the conditional test was advantageous because of its far lighter computational burden. As in the previous chapter, we remove the restriction of equal sample sizes, which is more natural in practice. This relaxation reduces the discreteness of the distribution of a test statistic and, at the same time, increases the amount of computation, as we have observed in the previous chapter.

As discussed in Hilton and Mehta (1993), "The primary factor responsible for the conservativeness of the conditional test is the discreteness of the conditional distribution of a test statistic, a factor tending to increase the conditional critical value, $t_\alpha(s)$. The primary factor responsible for the conservativeness of the unconditional test is the need to eliminate the nuisance parameter by considering the worst-case scenario for $\pi$, a factor tending to increase the unconditional critical value, $t_\alpha$." To verify this statement, we calculate the conditional sizes, size functions and powers of the conditional and unconditional tests, using each statistic in turn. The settings of the previous chapter, sample sizes and alternative hypotheses, are again employed

in this chapter.

The exact performance of conventional statistics for small or medium sample sizes has not been fully investigated, because it had been almost impossible to carry out the computation and had to depend on asymptotic properties of these statistics. However, exact inference is becoming much more feasible than it was a decade ago, owing to enormous improvement achieved recently both in algorithms and in computer power. The computational results we are going to display are based on Matsuo (2000a). These results would be informative to practitioners for deciding which statistic should, or should not, be used in practice.

We can summarize the results as follows: $Dev$ consistently performs poorly, to our great surprise, in the unconditional test. On the other hand, $PX$ shows stable performance in the unconditional test and the performance in the unconditional test dominates the conditional test even when sample sizes are as large as 50. $PD$ is expected to show intermediate properties between the above two, as expected from the definition in the previous chapter. In general, the behavior of $PD$ and $PX$ are similar, but the behavior of $Dev$ is different from the other two. We note that $Dev$ should not be used in the unconditional test.

## 3.1 Pearson's $X^2$

We would like to observe the relative performances of the conditional and unconditional tests, when the *Pearson's* $X^2$ is employed as a test statistic. It is of great interest to carry out the comparison, because the *Pearson's* $X^2$ has been the best used goodness-of-fit statistic for analysing discrete data. Sample size settings and alternative hypotheses are just the same as that presented in section 2.3. The results of the size and power calculations are displayed in a graphical form, for each sample sizes four ( $2 \times 2$ ) graphs are presented: upper-left graphs display the conditional sizes of the conditional test, upper-right graphs display the conditional sizes of the unconditional test, lower-left graphs display the size functions of the conditional and unconditional tests, where thick line represents the unconditional test and thin line represents the conditional test, and lower-right graphs display the power plots, where $x-$axis represents the conditional test and $y-$axis represents the unconditional test.

The following Figures 3.1$\sim$3.4 are presented to display the results of the case,

$n_1 = n_2 = n_3$, where the discreteness of a test statistic is most feared. Just as our anticipation, conditional sizes fluctuate heavily under the $\alpha = 0.05$ line in the conditional test, and up and down the line in the unconditional test. That every conditional size should be no more than the significance level, $\alpha$, in the conditional test, which guarantee the size function to be always no more than $\alpha$, makes the test excessively conservative especially in this case. The size functions of the unconditional test are usually above those of the conditional test, which makes us expect the unconditional test more powerful. This expectation is verified by the power plots, in which almost all points are located above the equal power line, $y = x$.

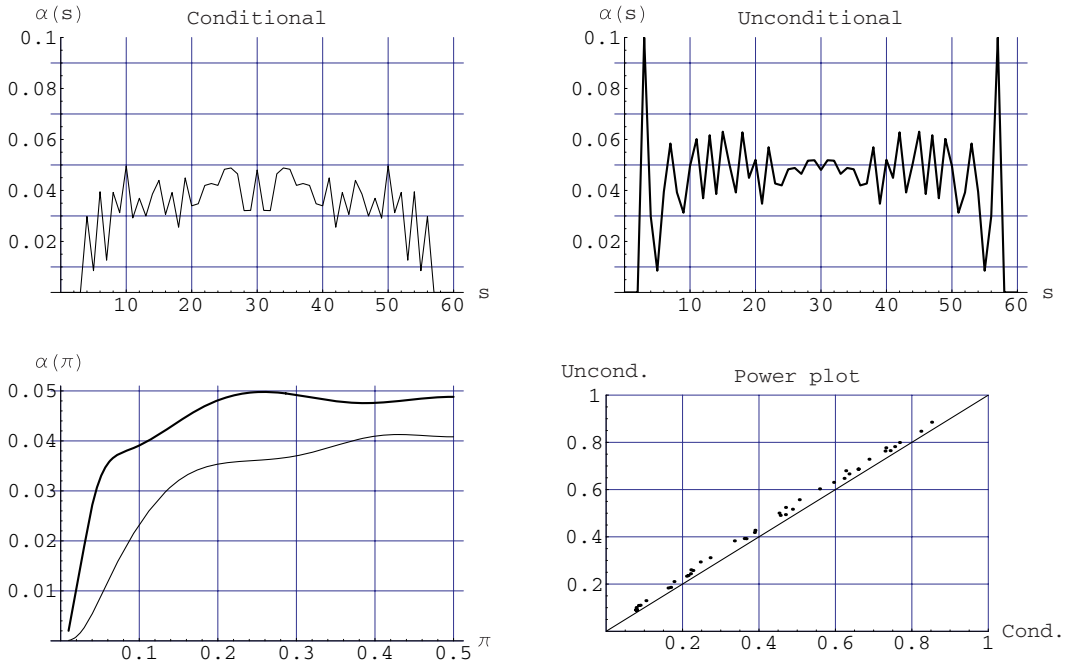



Figure 3.1: $\boldsymbol{n} = (20, 20, 20)$.

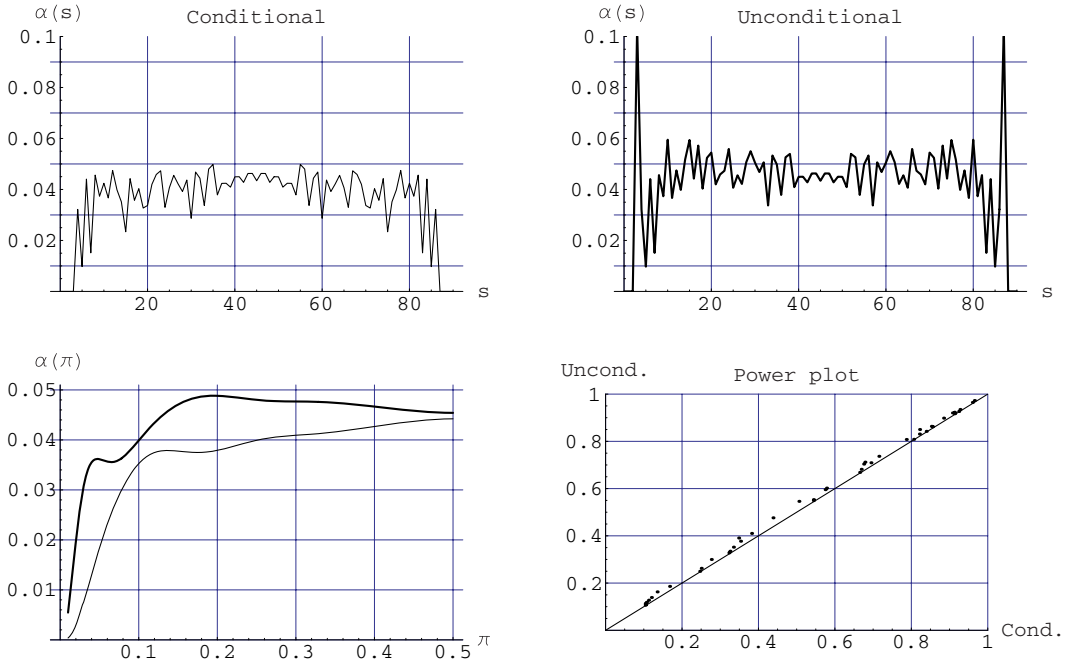



Figure 3.2: $\boldsymbol{n} = (30, 30, 30)$.

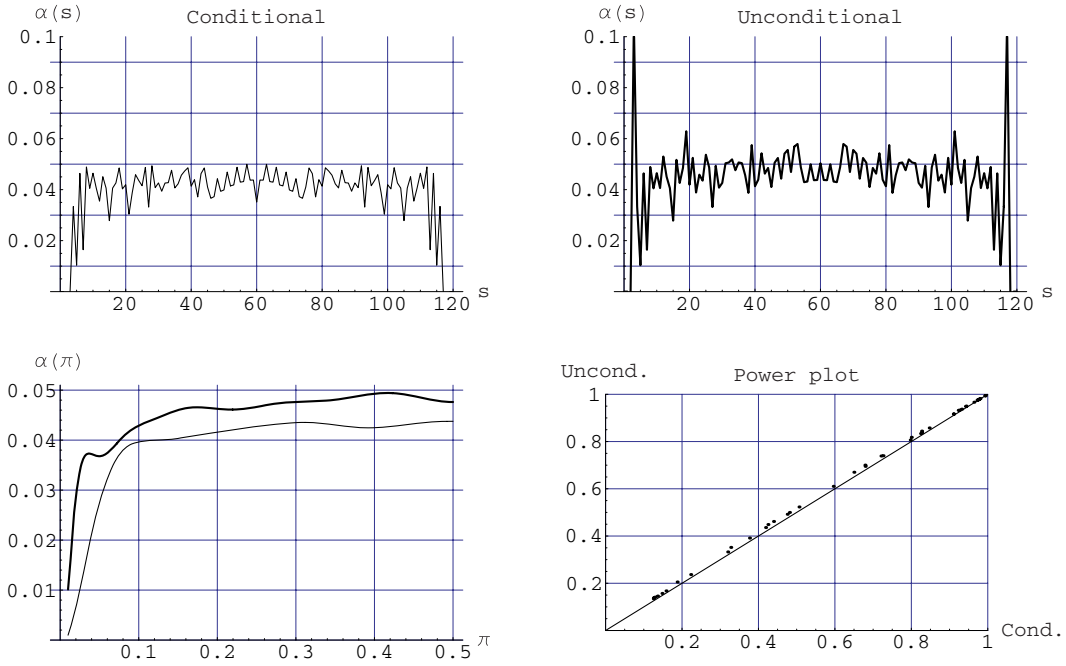



Figure 3.3: $\boldsymbol{n} = (40, 40, 40)$.

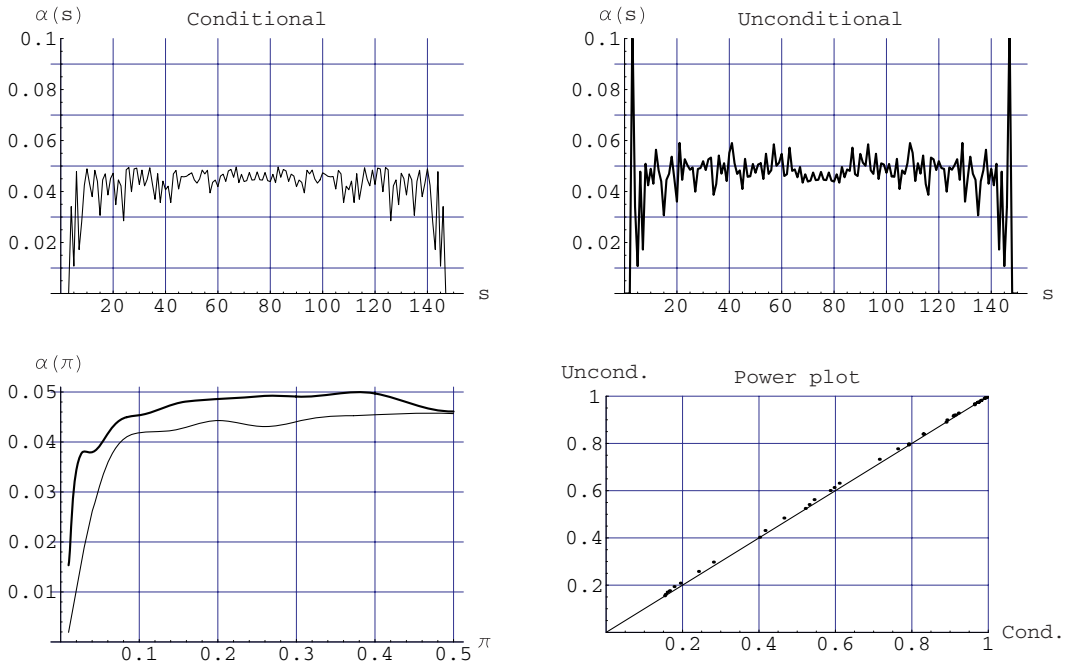



Figure 3.4: $\boldsymbol{n} = (50, 50, 50)$.

Next, we consider the case that two of three sample sizes are equal and the remaining size is larger by 1, $n_1 = n_2 = n_3 - 1$, where the discreteness of the distribution of a test statistic is considerably less marked than the previous case. The following Figures 3.5∼3.8 are presented to display the results. We can observe that the degree of fluctuation of conditional sizes is far smaller compared to the previous case, $n_1 = n_2 = n_3$, and is smaller in the conditional test compared to the unconditional one. The size functions of the unconditional test are not always located above the conditional ones. These observations support the statement, in Mehta and Hilton (1993), that the performance of the conditional test nearly equals that of the unconditional test, as the discreteness of a test statistics disappears. The power plots barely show the power advantage of the unconditional test over the conditional one.

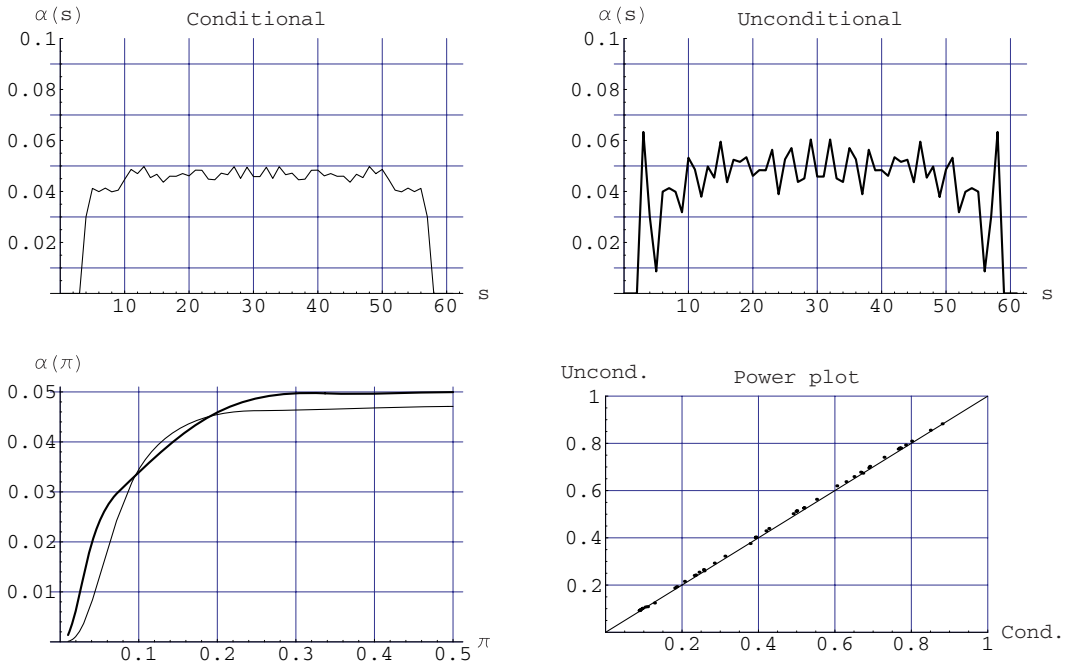



Figure 3.5: $\boldsymbol{n} = (20, 20, 21)$.

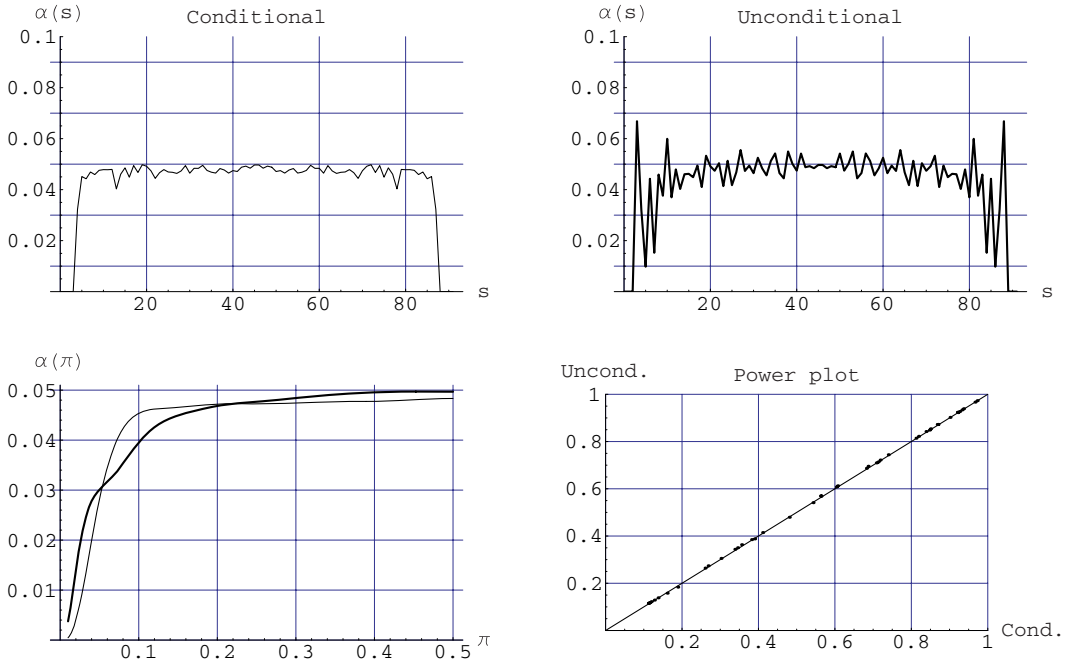



Figure 3.6: $\boldsymbol{n} = (30, 30, 31)$.

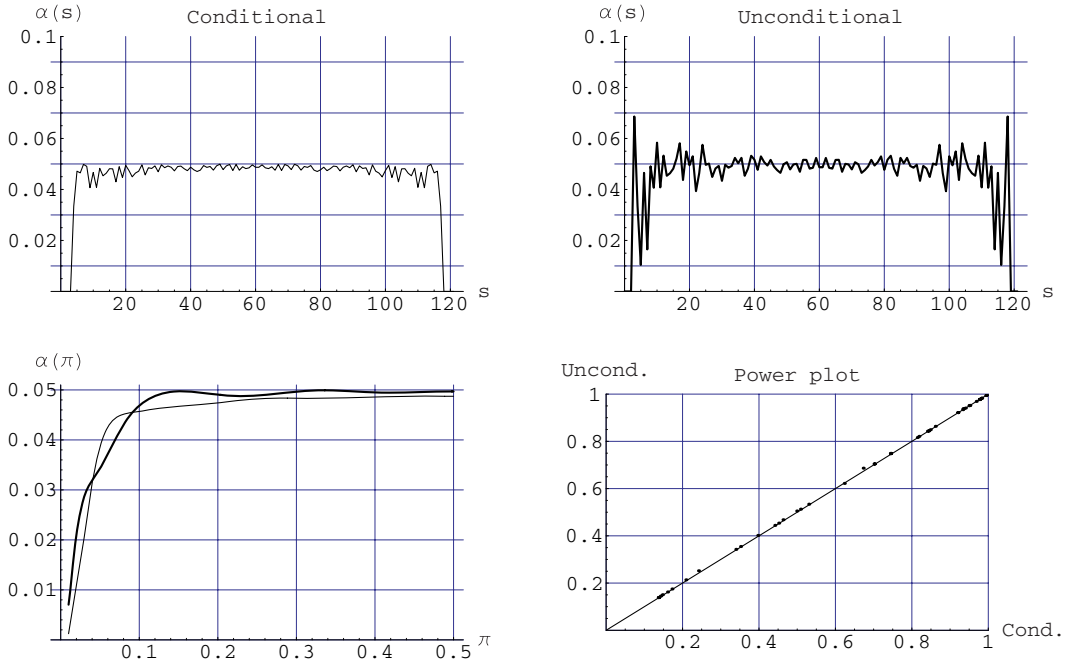



Figure 3.7: $\boldsymbol{n} = (40, 40, 41)$.

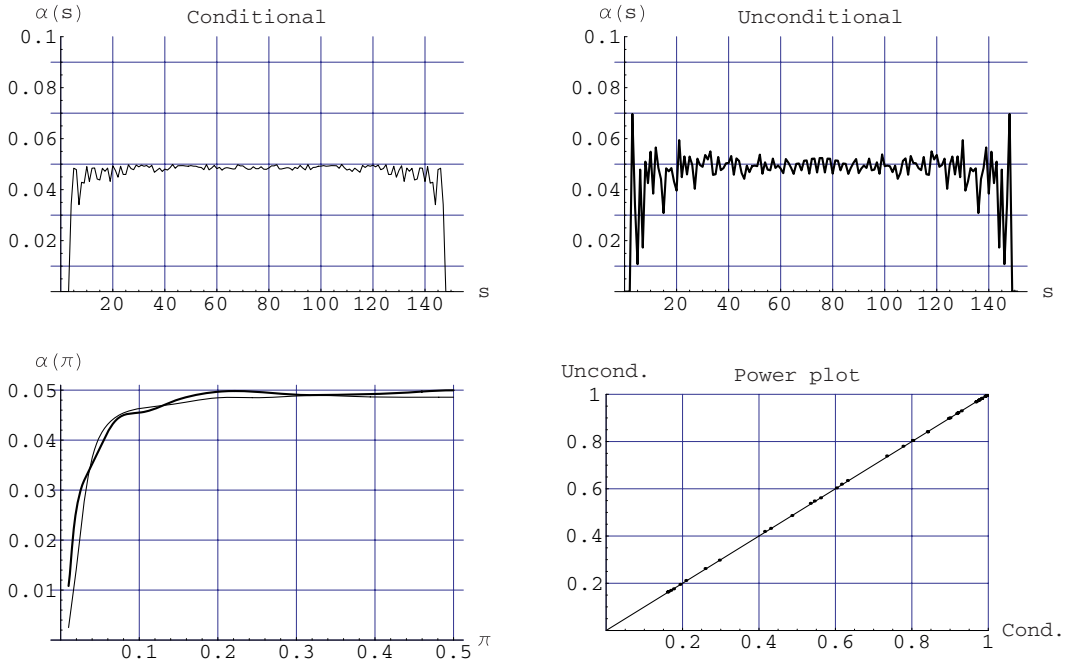



Figure 3.8: $\boldsymbol{n} = (50, 50, 51)$.

At last, we consider the case that all sample sizes are distinct, $n_1+1 = n_2 = n_3-1$, where the discreteness of a test statistic is least expected. The following Figures 3.9∼3.12 are presented to display the results. The change, we have observed from the case, $n_1 = n_2 = n_3$, to the case, $n_1 = n_2 = n_3 - 1$, is again observed and the difference in performance between the conditional and unconditional tests becomes quite trivial, which indicates the disadvantage of the unconditional test because of its heavy computational burden.

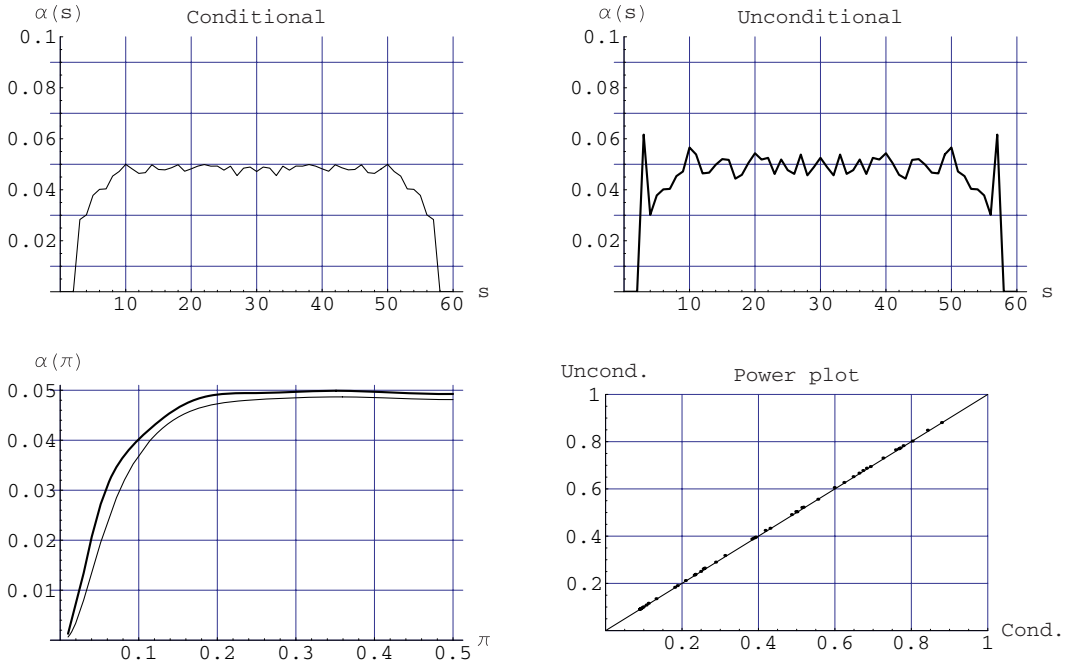



Figure 3.9: $\boldsymbol{n} = (19, 20, 21)$.

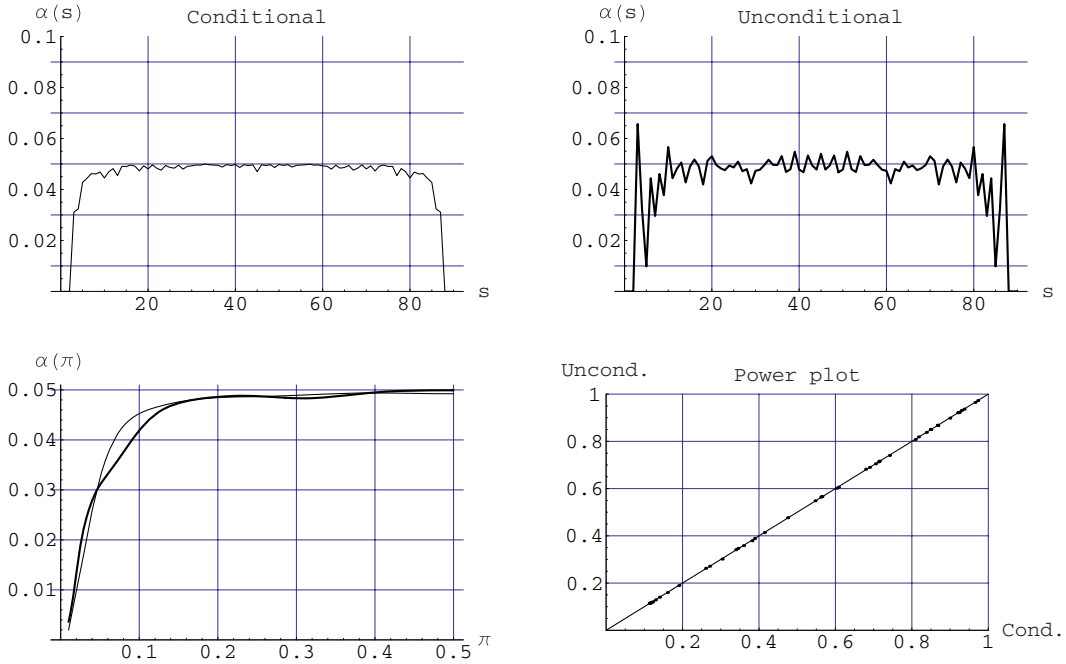



Figure 3.10: $\boldsymbol{n} = (29, 30, 31)$.

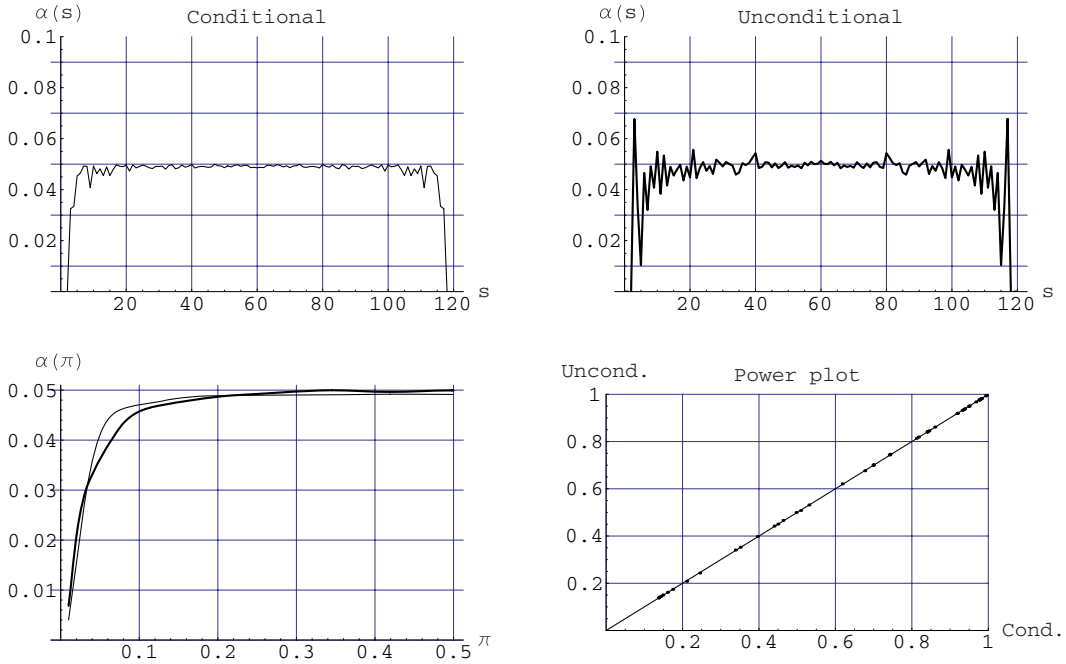



Figure 3.11: $\boldsymbol{n} = (39, 40, 41)$.

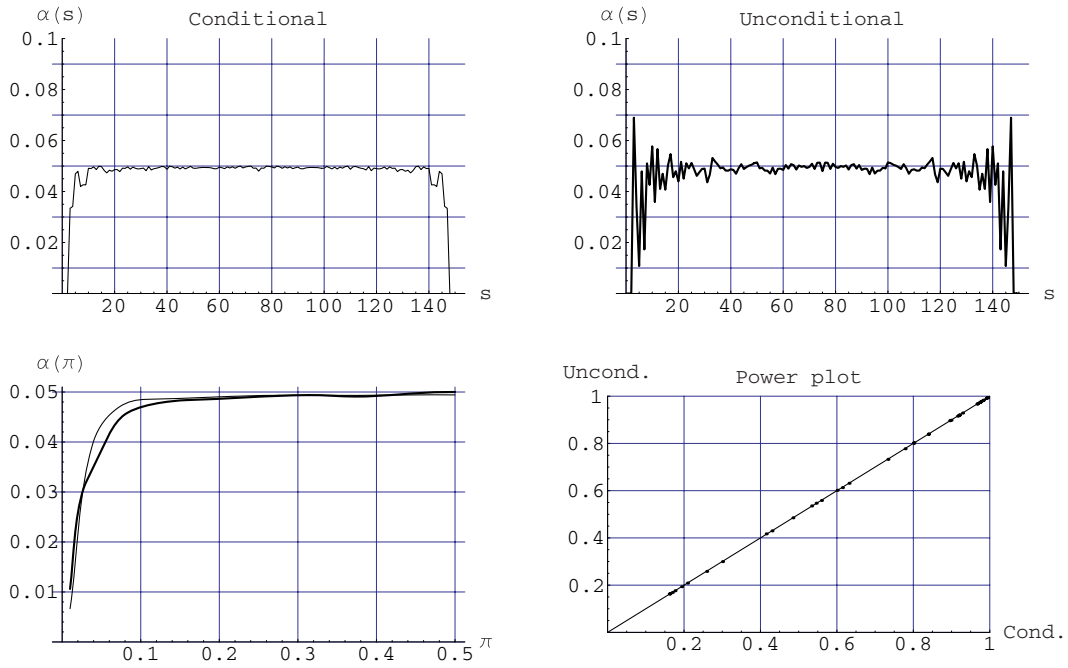



Figure 3.12: $\boldsymbol{n} = (49, 50, 51)$.

## 3.2 Deviance

In this section, we observe the relative performance of the conditional and unconditional tests, when the *deviance*, $Dev$, is employed as a goodness-of-fit statistic. Because $Dev$ is based on the maximum likelihood principle, it is natural to use $Dev$ when maximum likelihood estimate for parameters is used. In fact, in the framework of the *generalized linear models*, the use of $Dev$ is recommended. The results we are going to present would be of great interest for statisticians who are not certain about the exact performance of $Dev$.

Calculations are carried out at the same settings as the *Pearson's* $X^2$, $PX$. The performance of the conditional test using $Dev$ is almost parallel to that of the conditional test using $PX$. The performance of the unconditional test using $Dev$, however, is extremely poor compared to the conditional one. To be more specific, $Dev$ tend to take relatively small values on the consecutive conditional reference sets, $\Gamma_7$, $\Gamma_8$ and $\Gamma_9$, which requires smaller critical values in order to make the size function to be no more than the significance level, $\alpha = 0.05$, around $\pi \doteq 8/N$. As a

consequence, the size functions of the unconditional test, displayed with thick curve in the lower left graph of each Figure 3.13∼3.24, have a common shape that the curves rise rapidly and attain their maximum about $\pi = 8/N$, and then decrease rather slowly towards $\alpha(\pi) = 0.03$. This tendency is consistently observed over the other combinations of sample sizes. It is not too much to say that $Dev$ should not be employed in the unconditional test.

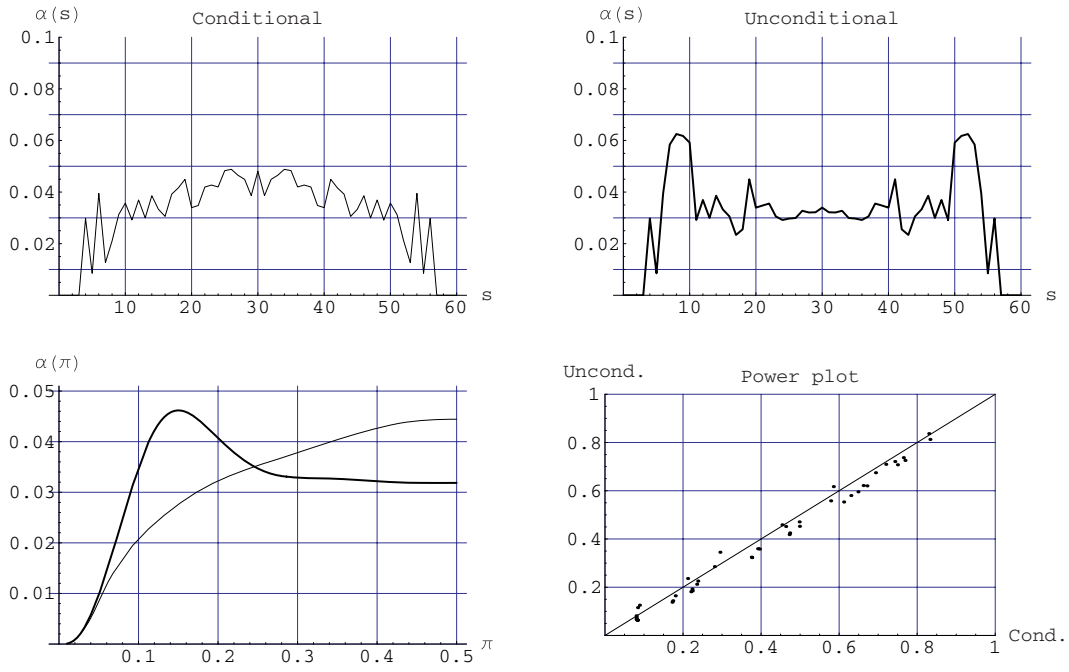



Figure 3.13: $\boldsymbol{n} = (20, 20, 20)$.

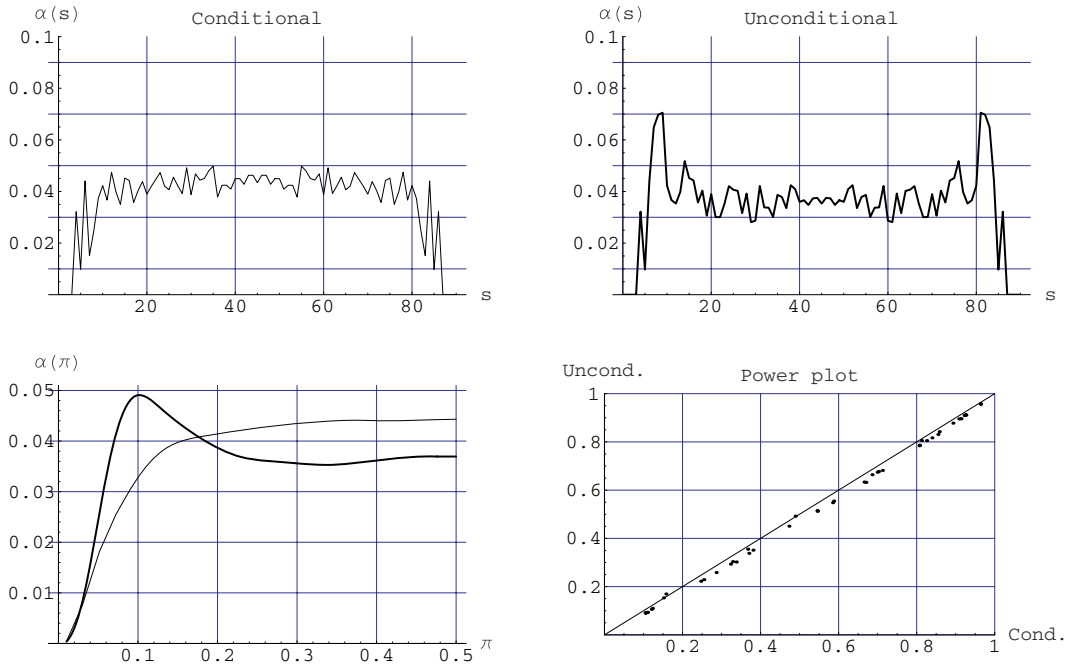



Figure 3.14: $\boldsymbol{n} = (30, 30, 30)$.

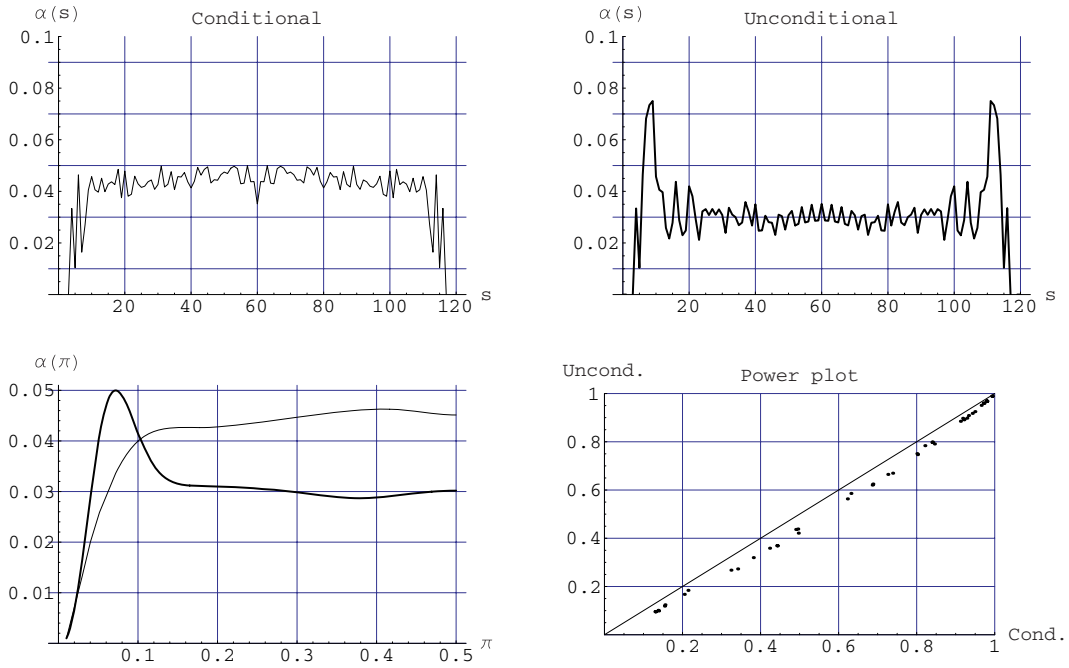



Figure 3.15: $\boldsymbol{n} = (40, 40, 40)$.

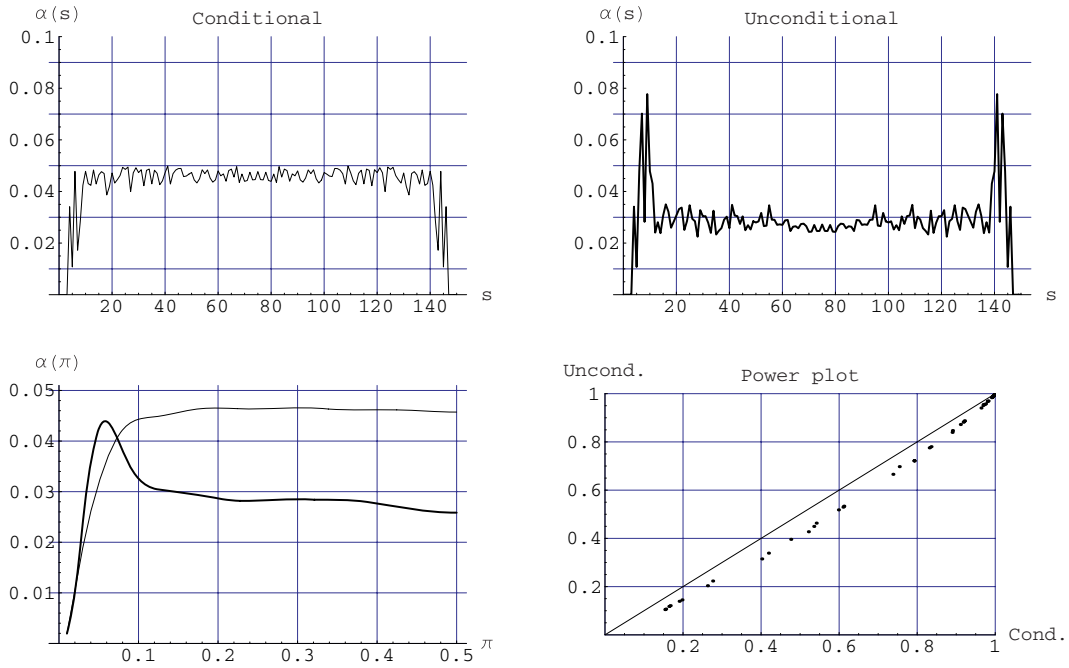



Figure 3.16: $\boldsymbol{n} = (50, 50, 50)$.

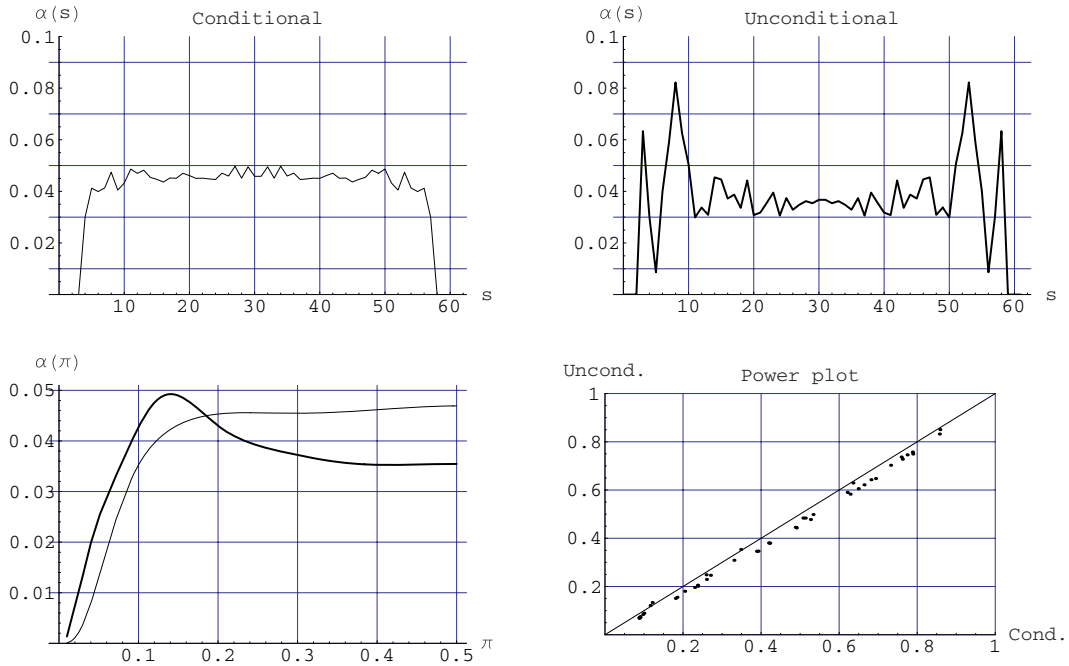



Figure 3.17: $\boldsymbol{n} = (20, 20, 21)$.

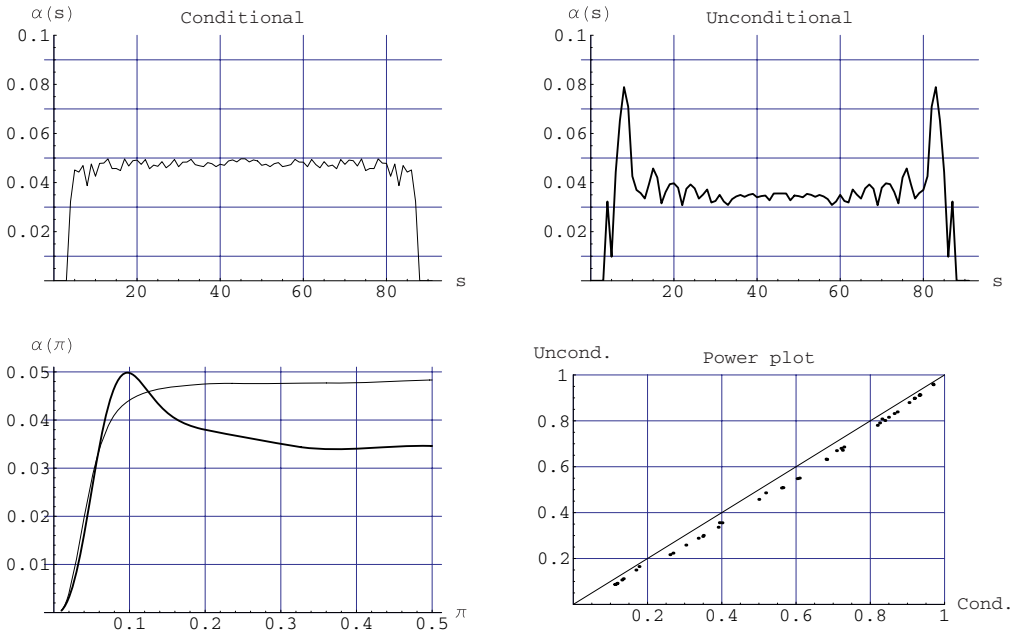



Figure 3.18: $\boldsymbol{n} = (30, 30, 31)$.

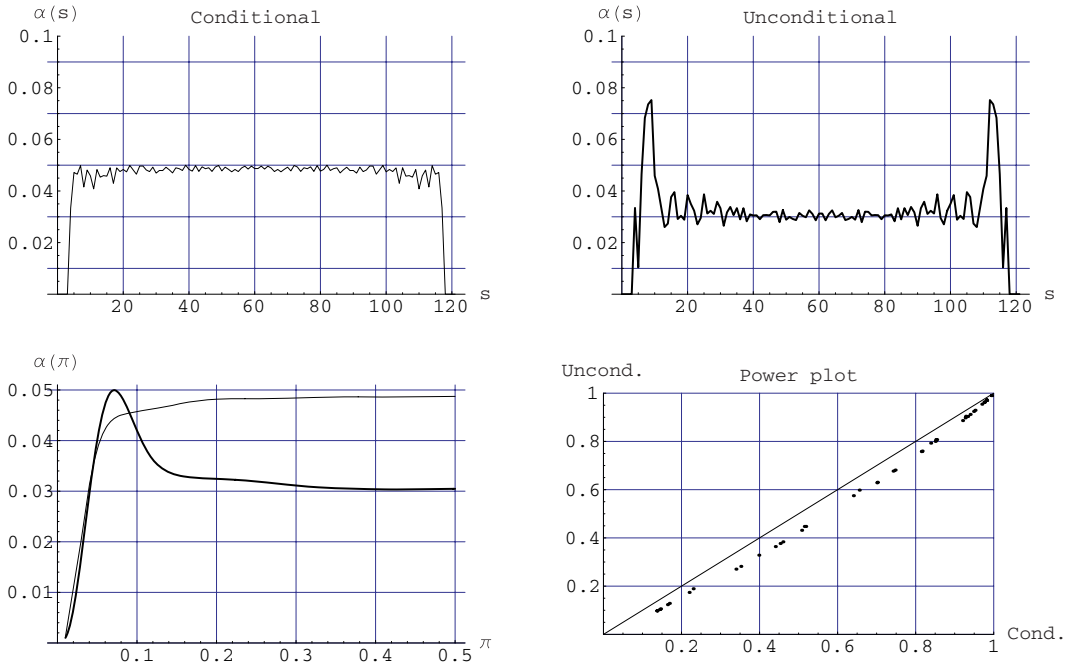



Figure 3.19: $\boldsymbol{n} = (40, 40, 41)$.

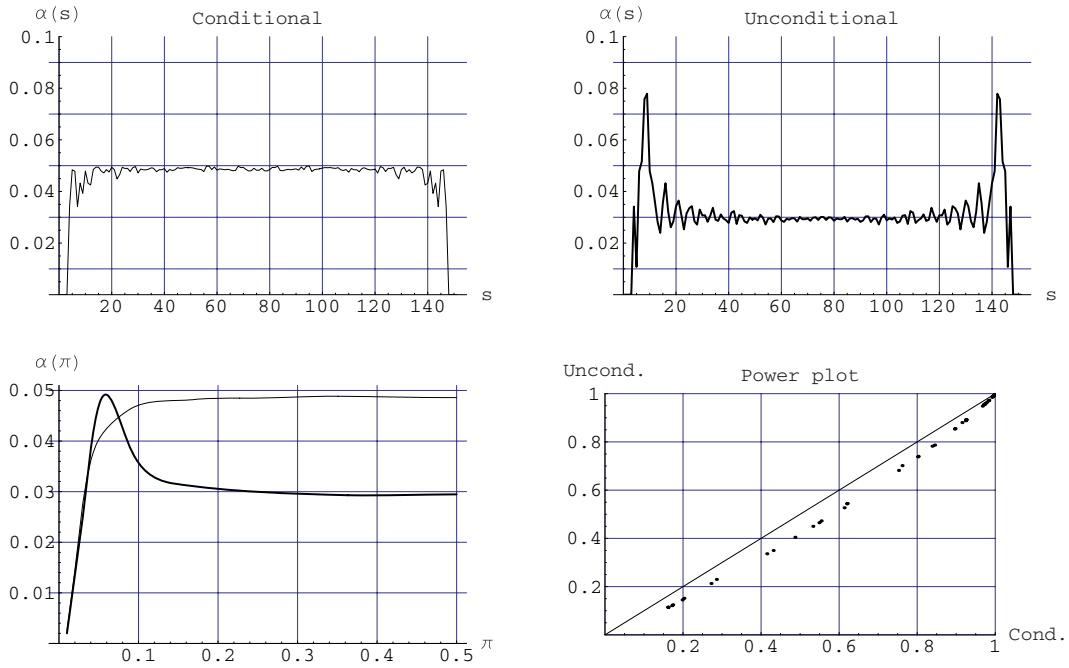



Figure 3.20: $\boldsymbol{n} = (50, 50, 51)$.

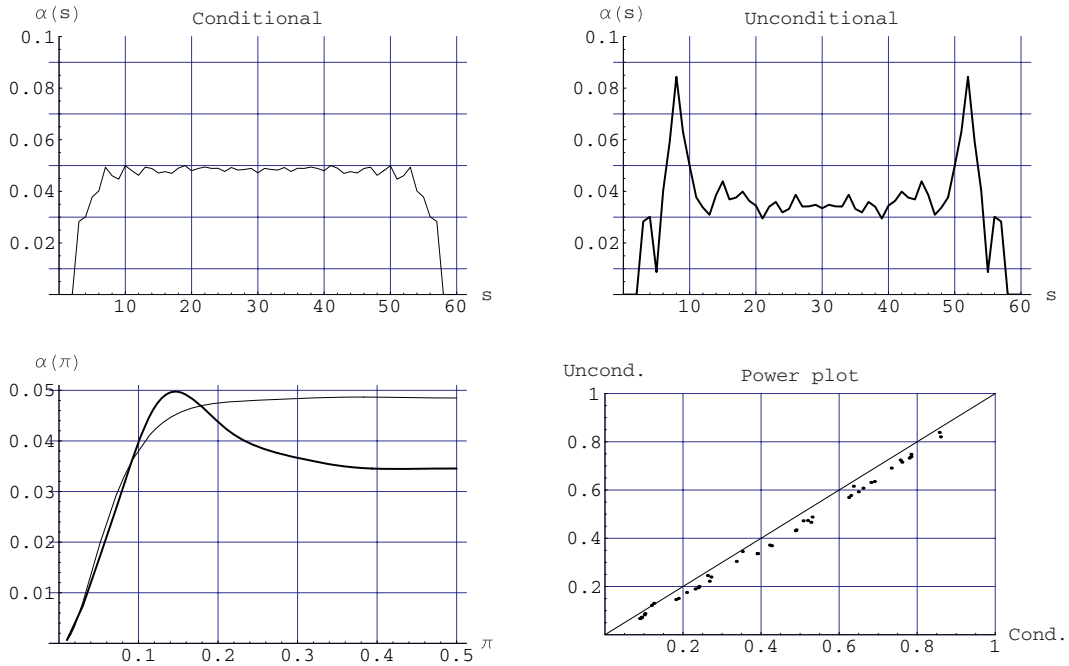



Figure 3.21: $\boldsymbol{n} = (19, 20, 21)$.

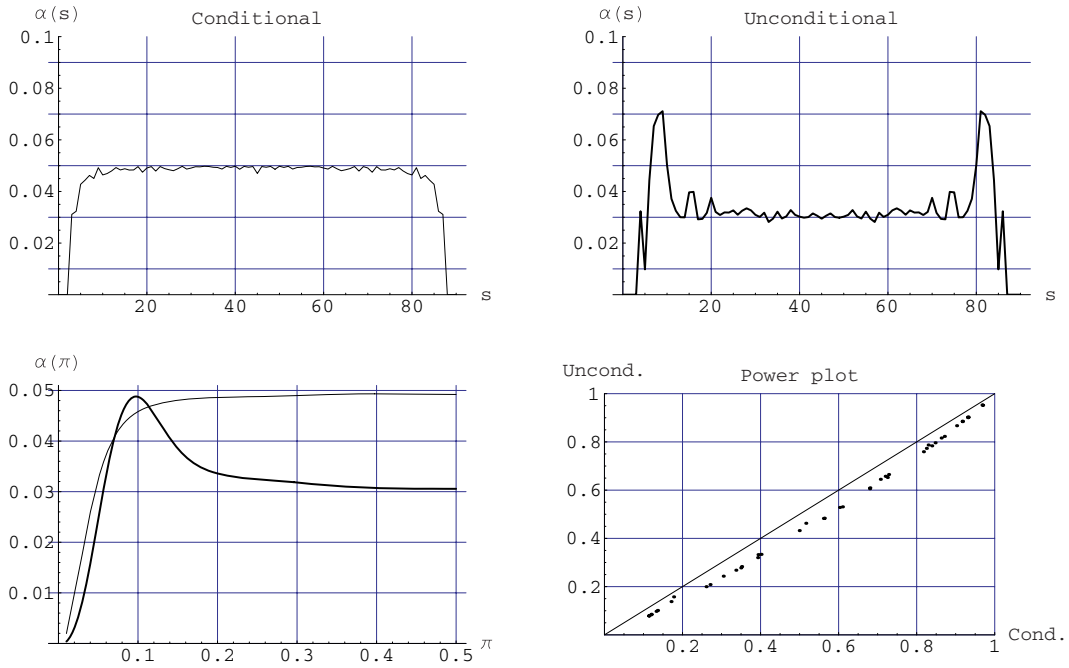



Figure 3.22: $\boldsymbol{n} = (29, 30, 31)$.

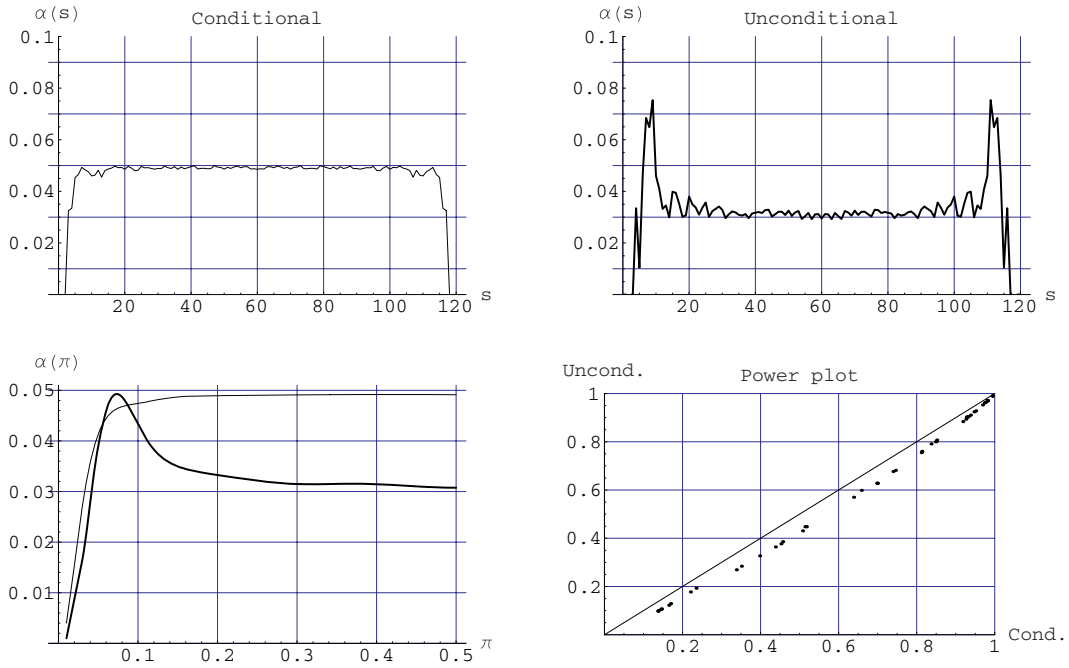



Figure 3.23: $\boldsymbol{n} = (39, 40, 41)$.

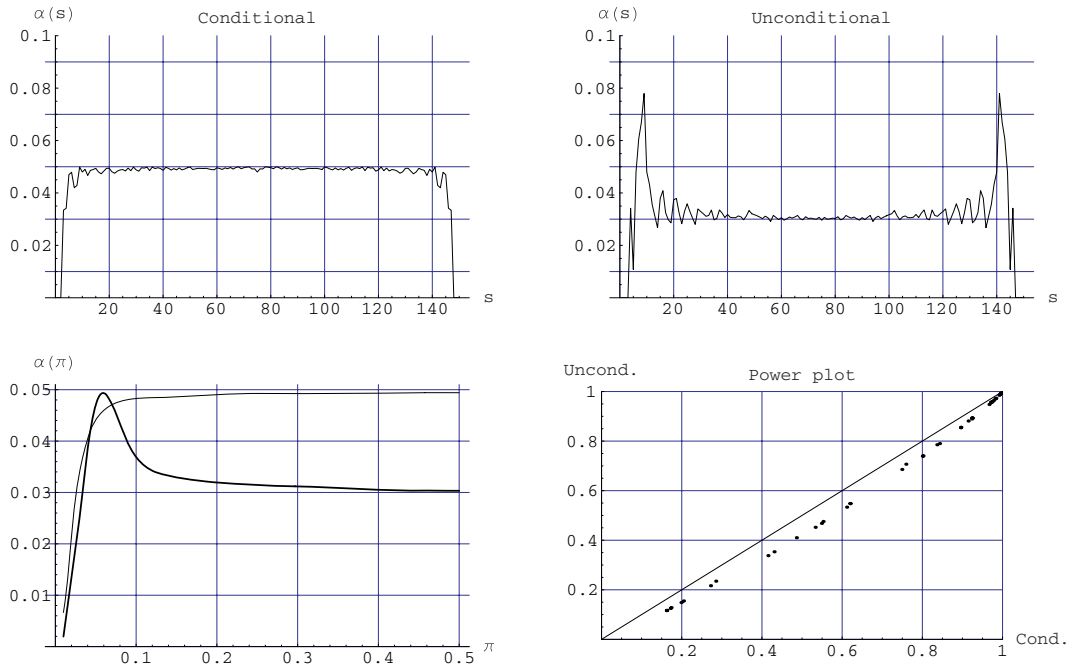



Figure 3.24: $\boldsymbol{n} = (49, 50, 51)$.

## 3.3 Power divergence

The *power divergence*, $PD$, is advocated by Cressie and Read (1984), out of the *family of power divergence statistics*, on asymptotic grounds. From the functional form given in (1.4) and from the Figures 2.1, 2.2 and 2.4, we can expect $PD$ has intermediate properties between $PX$ and $Dev$, more like $PX$ rather than $Dev$.

Calculations are carried out at the same settings as $PX$ and $Dev$. The following Figures 3.25$\sim$ 3.36 are presented to display the results.

The performances of the conditional and unconditional tests using $PD$ are hardly different from those using $PX$. As mentioned in the section 2.1, the complex form of $PD$ prevent from taking tied values, compared to $Dev$ and $PX$, which may be an advantage of employing $PD$.

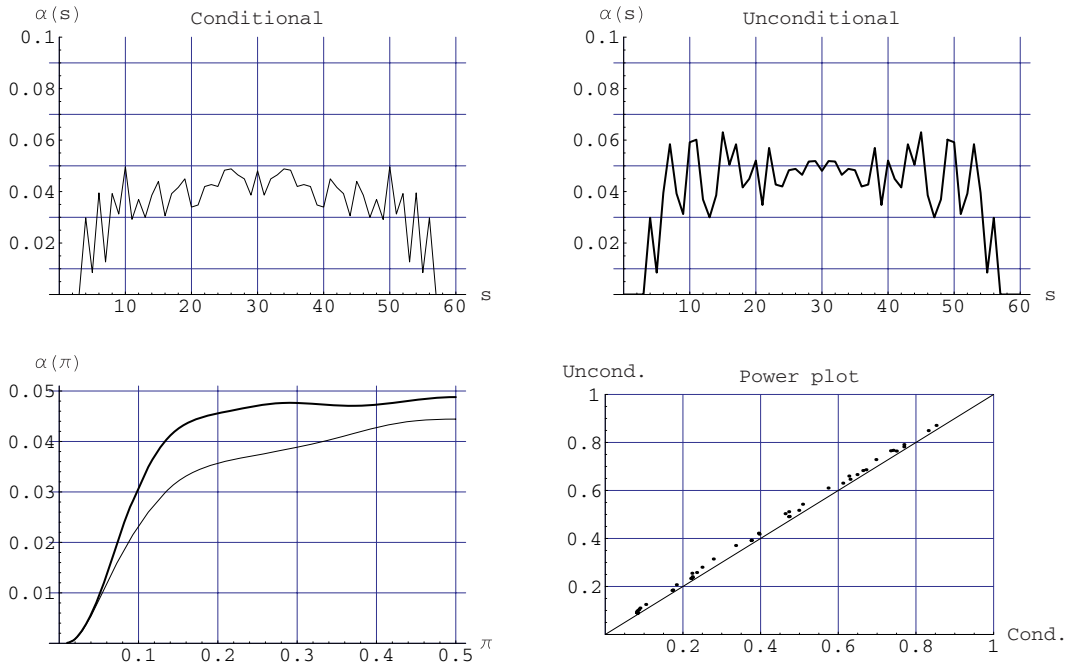



Figure 3.25: $\boldsymbol{n} = (20, 20, 20)$.

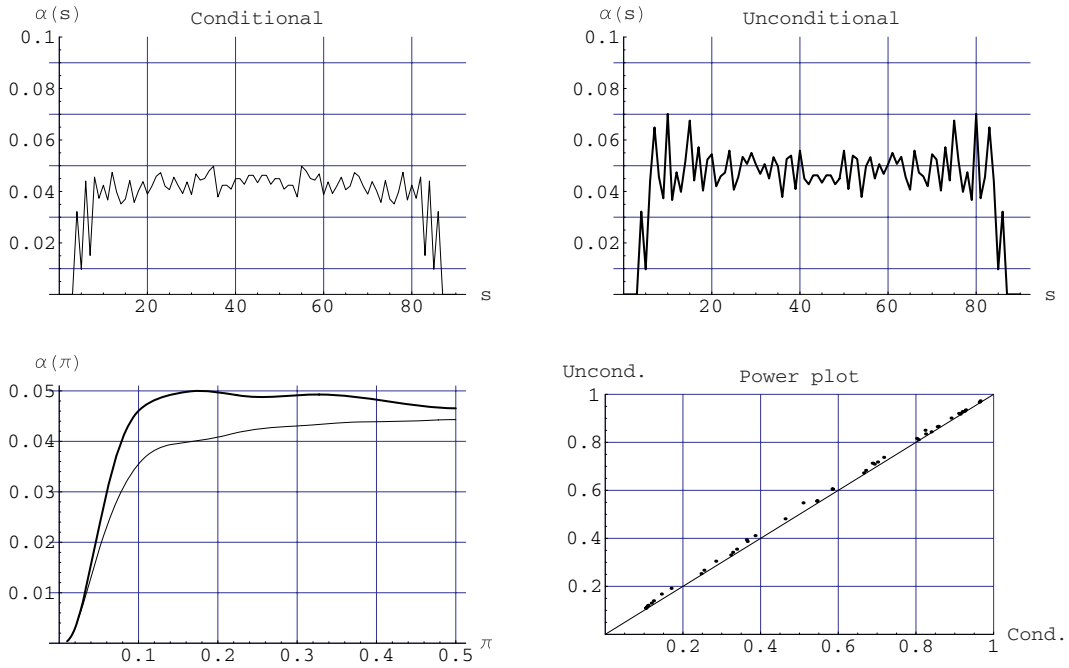



Figure 3.26: $\boldsymbol{n} = (30, 30, 30)$.

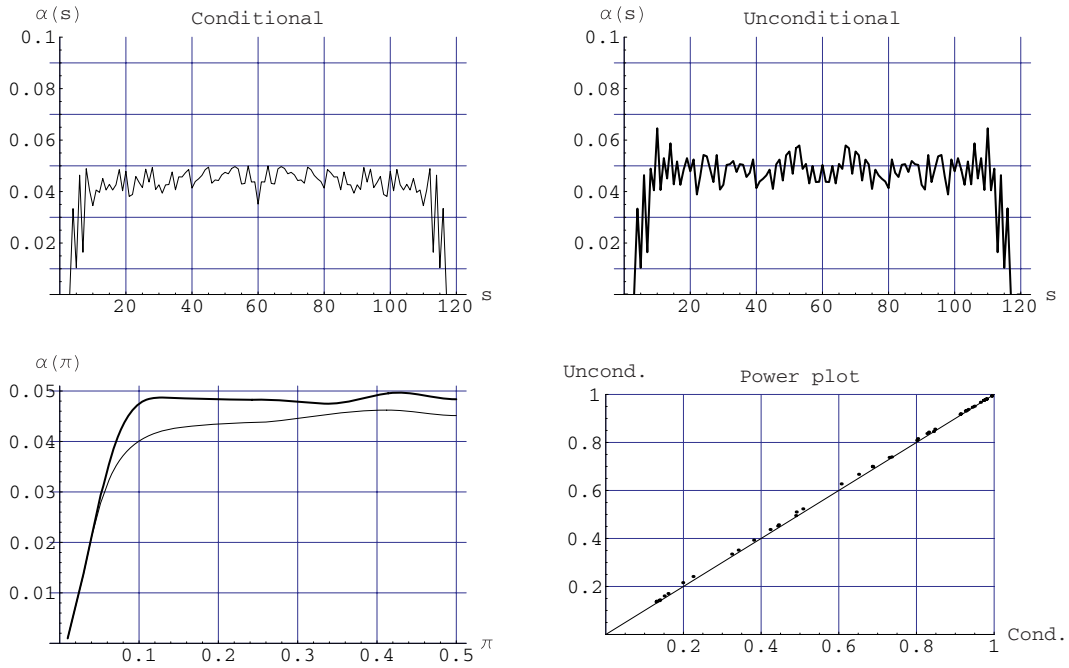



Figure 3.27: $\boldsymbol{n} = (40, 40, 40)$.

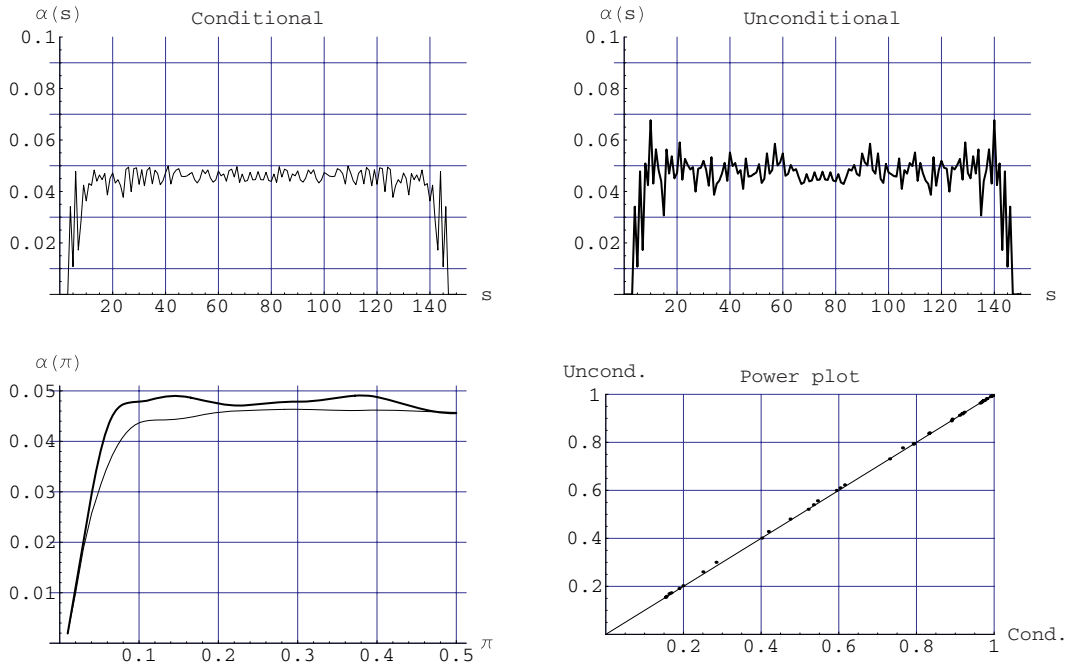



Figure 3.28: $\boldsymbol{n} = (50, 50, 50)$.

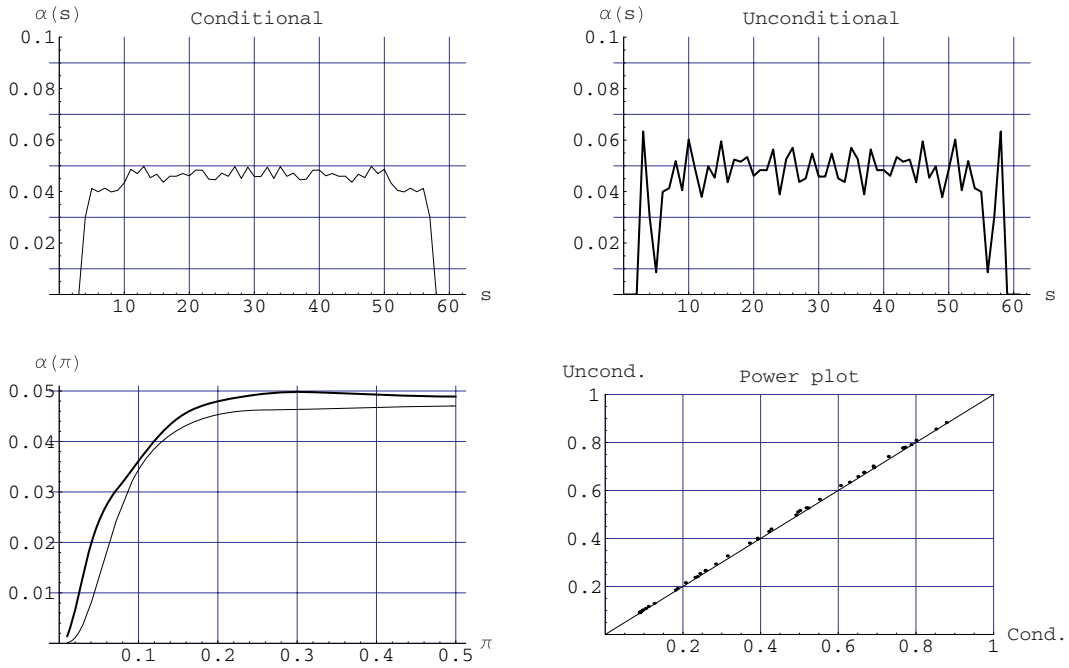



Figure 3.29: $\boldsymbol{n} = (20, 20, 21)$.

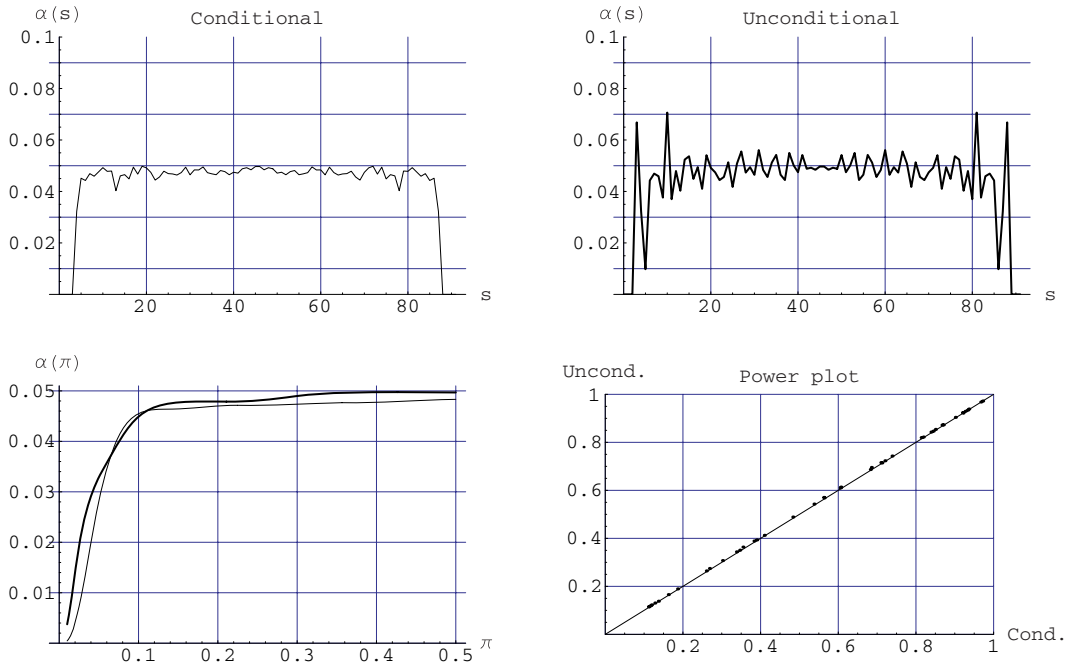



Figure 3.30: $\boldsymbol{n} = (30, 30, 31)$.

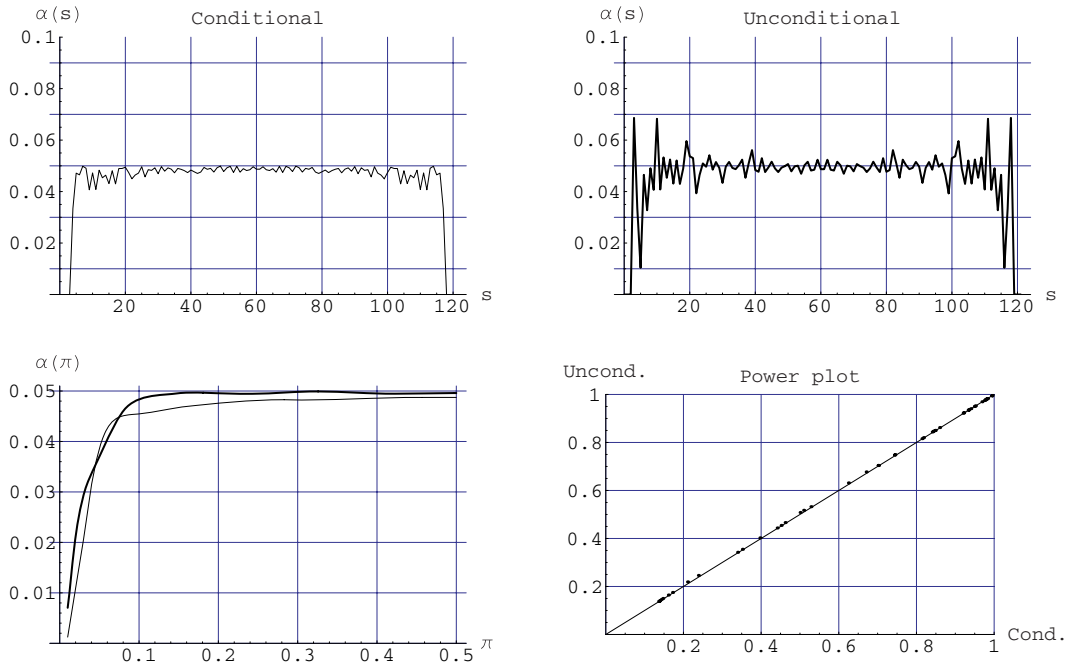



Figure 3.31: $\boldsymbol{n} = (40, 40, 41)$.

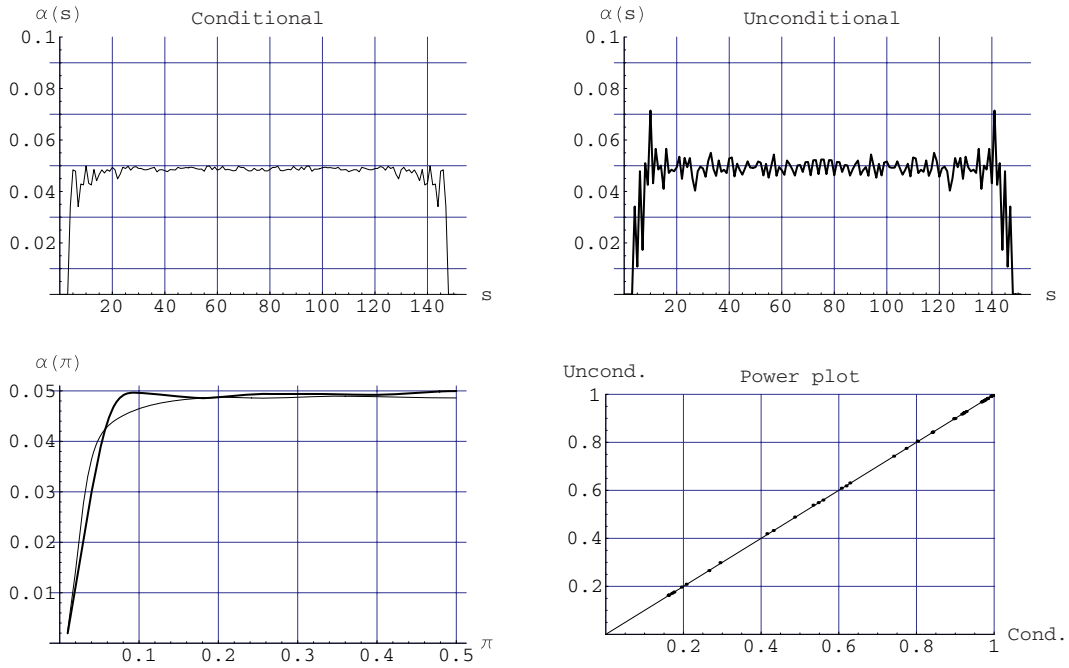



Figure 3.32: $\boldsymbol{n} = (50, 50, 51)$.

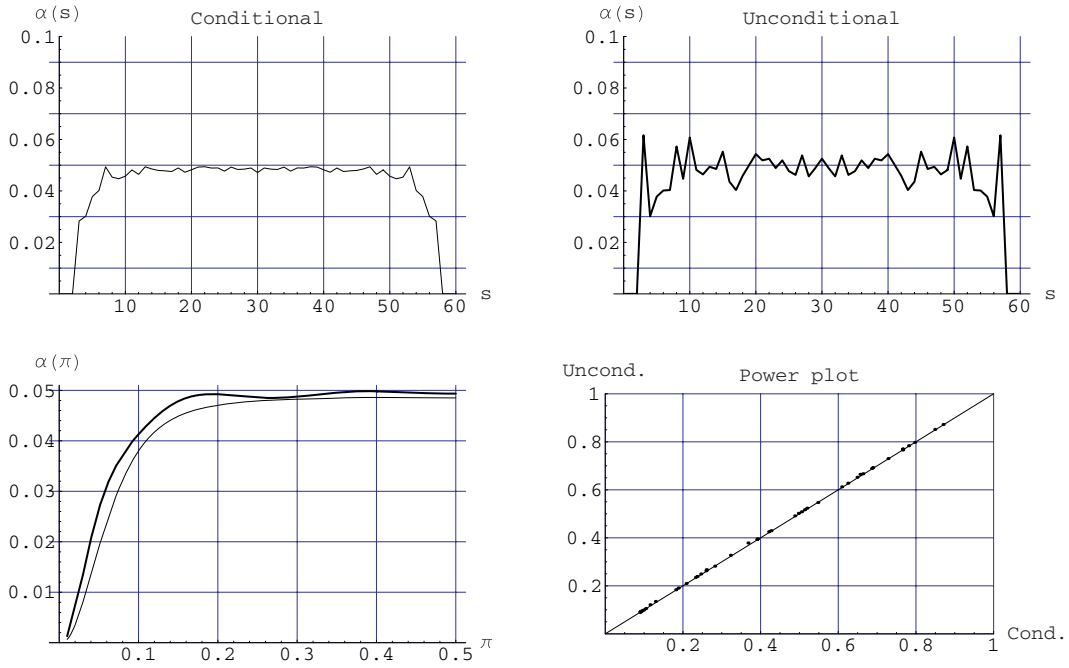



Figure 3.33: $\boldsymbol{n} = (19, 20, 21)$.

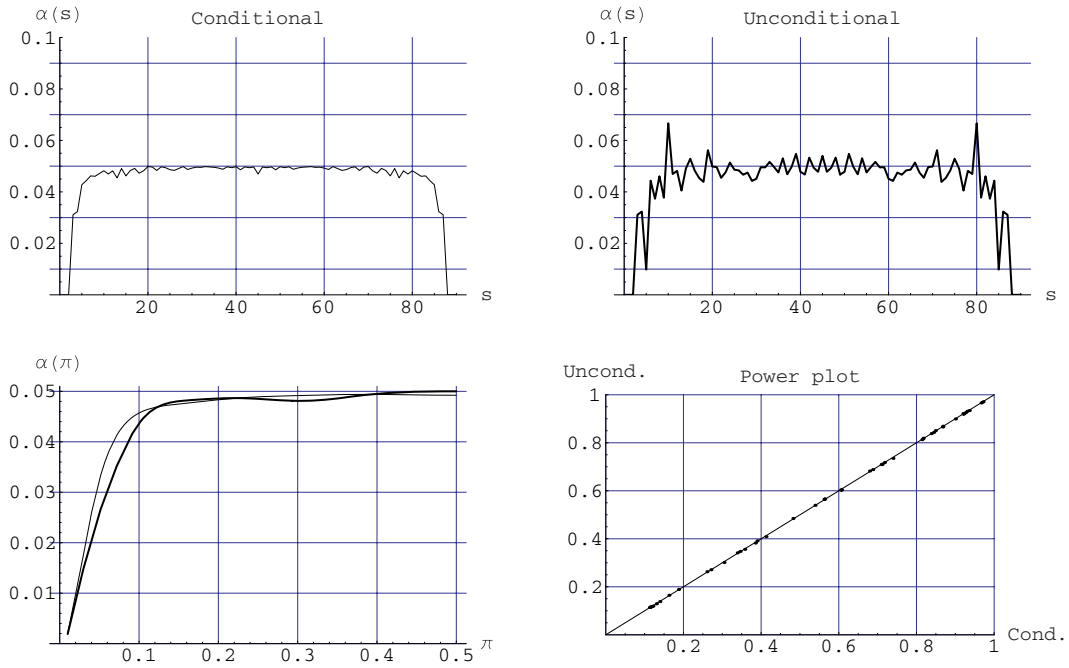



Figure 3.34: $\boldsymbol{n} = (29, 30, 31)$.

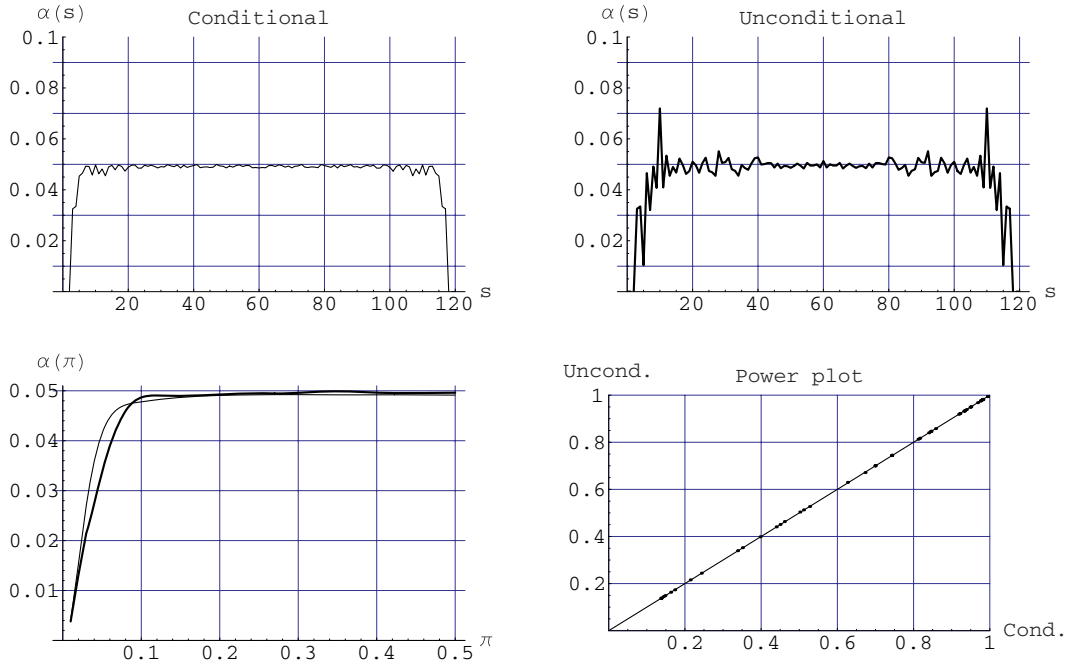



Figure 3.35: $\boldsymbol{n} = (39, 40, 41)$.

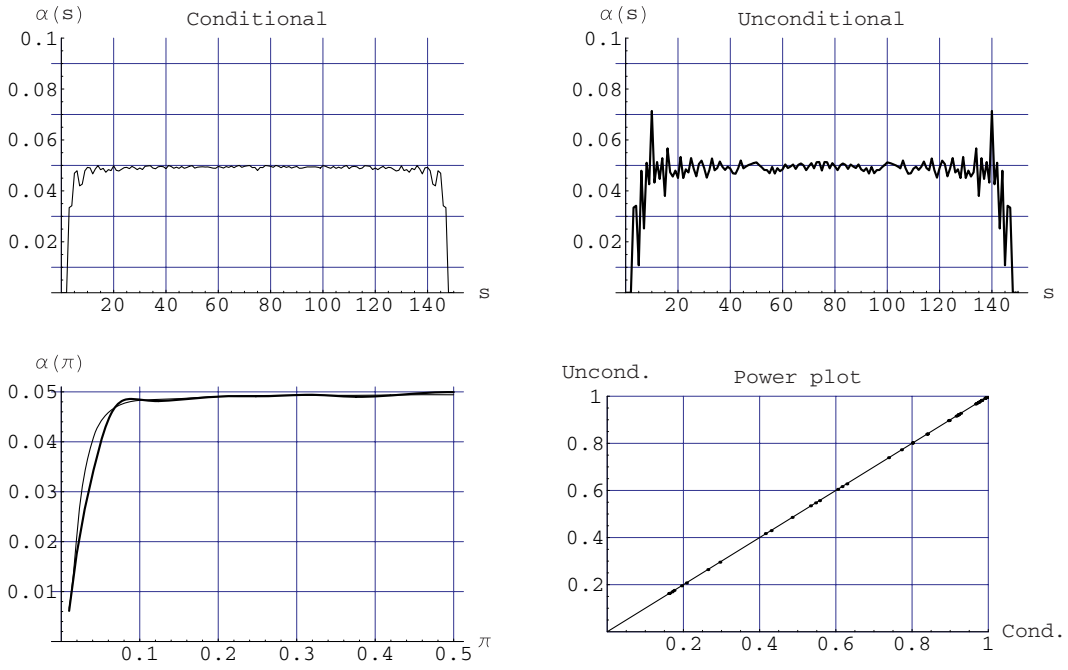



Figure 3.36: $\boldsymbol{n} = (49, 50, 51)$.

# Chapter 4

# Conditional two-dimensional test

In Chapter 3, we compared the conditional and unconditional tests, employing each of the three statistics in turn. Each test has its own advantages as well as disadvantages. Whether the conditional/unconditional test is more powerful than the unconditional/conditional test depends on sample sizes as well as alternative hypotheses. Anyway, the most serious problem is that the both tests have far smaller sizes than a given fixed significance level when sample sizes are small and identical, as observed in Chapter 3.

From now on, we would like to stand upon a practical ground and would not treat the conditional and unconditional tests as competing ones. The purpose of this thesis is to device test procedures which are more powerful than the conventional conditional and unconditional tests, while keeping the rule of fixed-level significance test, that is, the size function of a test should always be no more than the significance level, $\alpha$. We would like to emphasize that the computational burden needed to carry out exact methods is becoming lighter, thanks to the modern development of computational circumstances. Therefore, additional computational burden, caused by a refinement in statistical inferences, would be acceptable as a cost of the higher performance.

We are going to explore test procedures for higher power even when sample sizes are small and identical, hereafter. We note that only $PX$ is considered as a test statistic for the rest of this thesis, because $PD$ performs just like $PX$ and $Dev$ performs poorly in the unconditional test.

The content of this chapter is based on Matsuo (2000a).

## 4.1 Derivation

The procedure we are going to introduce in this chapter is based on one of the observations in Chapter 2 and 3 that, when sample sizes are of the same scale, we expect the discreteness of goodness-of-fit statistic distribution to be least serious if all sample sizes are distinct and most serious if identical. When sample sizes are identical, a conventional test statistic has a tied value over the permutations of an observation, as seen in section 2.1. Whether sample sizes are equal or not seems to be uncritical, it does, however, influence the performance of a test.

To ease the discreteness that arises from the equality of sample sizes, a device, which gives an order among the observations sharing a tied statistic value, might be worth considering. The idea is to use a covariate accompanied with an observation. This covariate needs not necessarily to be of interval scale, but may be of ordered nominal scale such as the order of anticipated binomial proportions. Once a covariate $x_i$ $(i = 1, \cdots, k)$ is specified, we are able to calculate $\sum x_i y_i$ and use it to order the observations sharing a tied statistic value. Since the statistic $\sum x_i y_i$ is the sufficient statistic of the scale parameter $\beta$ when we assume the following logistic regression model,

$$\pi_i = \frac{\exp\{\alpha + \beta x_i\}}{1 + \exp\{\alpha + \beta x_i\}},$$

it is natural to consider that larger $\sum x_i y_i$ value means larger deviation from the null hypothesis, provided $\beta$ is assumed to be positive. The procedure described above is considered as a test using the *two-dimensional statistic*, $( T(\boldsymbol{Y}), \sum x_i Y_i )$. We would like to refer this test as the *conditional two-dimensional test*, hereafter.

## 4.2 Numerical result

Now, we are going to illustrate the relative performance of the conditional two-dimensional test, to the conventional unconditional test, over the settings of sample sizes considered in Chapter 2, except for the case that all sample sizes are distinct. When sample sizes are distinct, there is little effect of introducing the two-dimensional statistic and therefore we do not consider here. To carry out the conditional two-dimensional test, we have to specify the values of the covariate $x_i$. We specify $x_i = i$, which corresponds to a dummy variable. This specification is

expected to yield tied $\sum x_i Y_i$ values and, as a result, give rise to worse performance than the case that $x_i$'s are continuous and not equally spaced. That is, our specification is not fully favorable to the conditional two-dimensional test and better results could be expected than those we are going to display, if a continuous covariate were available.

We consider two cases, one being three sample sizes are equal and another being two of three sample sizes are equal. For each case, we consider four sample size settings and, for each setting, we present four graphs. Upper-left graphs are presented for comparing conditional sizes of the tests, where horizontal axis being $s = \sum y_i$, thick line represents the conditional two-dimensional test and thin line represents the ordinary unconditional test. Upper-right graphs are presented for displaying the size functions, where horizontal axis being $\pi$, thick curve represents the conditional two-dimensional test and thin curve represents the ordinary unconditional test. Lower-left graphs are the power plots of the two tests over the 44 simple alternatives, listed in section 2.3. We note here that, even when all sample sizes are equal, the power of the conditional two-dimensional test varies over the permutations of an alternative, because of the introduction of the two-dimensional test statistic that assigns different values among the permutations of an observation. The powers plotted on these graphs are the average powers over the permutations of each alternative. Lower-right graphs are the power plots of the 44 alternatives themselves, that is, we do not consider any permutations at all. What we want to show, with these graphs, is the performance of the conditional two-dimensional test when the guess, that the observations are listed in ascending expected binomial proportion order, is true.

Let us observe the first case that three sample sizes are equal. We consider the four settings; common sample sizes are 20, 30, 40 and 50. Looking at Figure 3.1 $\sim$ 3.4, it is clearly observed that the conditional sizes of the conditional two-dimensional test (abbreviated to "C2D") consistently come up nearer to $\alpha = 0.05$ line and show much less fluctuation than both the conventional conditional and unconditional tests, which explains why size functions of the conditional two-dimensional test show relatively monotonous behavior. Concerning average powers over the permutations of simple alternatives, the two tests show quite similar results, that is, points in the power plots are almost on the diagonal line, $y = x$, apart

from the setting $n = (20, 20, 20)$, where points are located a little above the diagonal line. The conventional unconditional test sometimes attain power advantage over the two-dimensional conditional test when sample sizes are small, in this case. When it comes to the power comparison over the simple alternatives themselves, the advantage of the conditional two-dimensional test is clearly observed without exception.

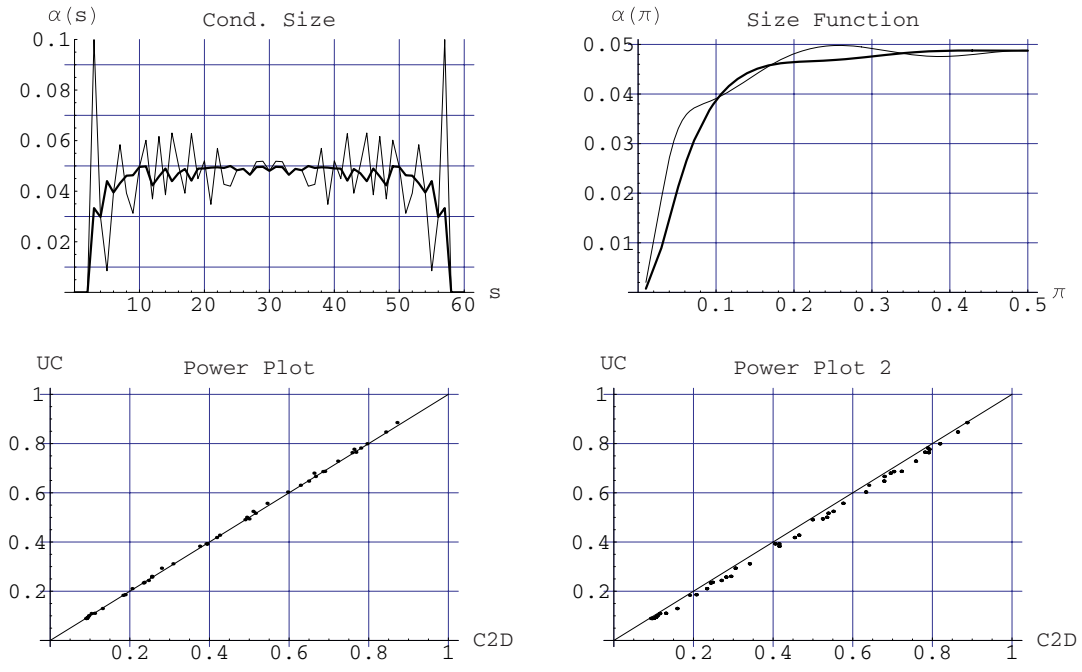



Figure 4.1: $\boldsymbol{n} = (20, 20, 20)$.

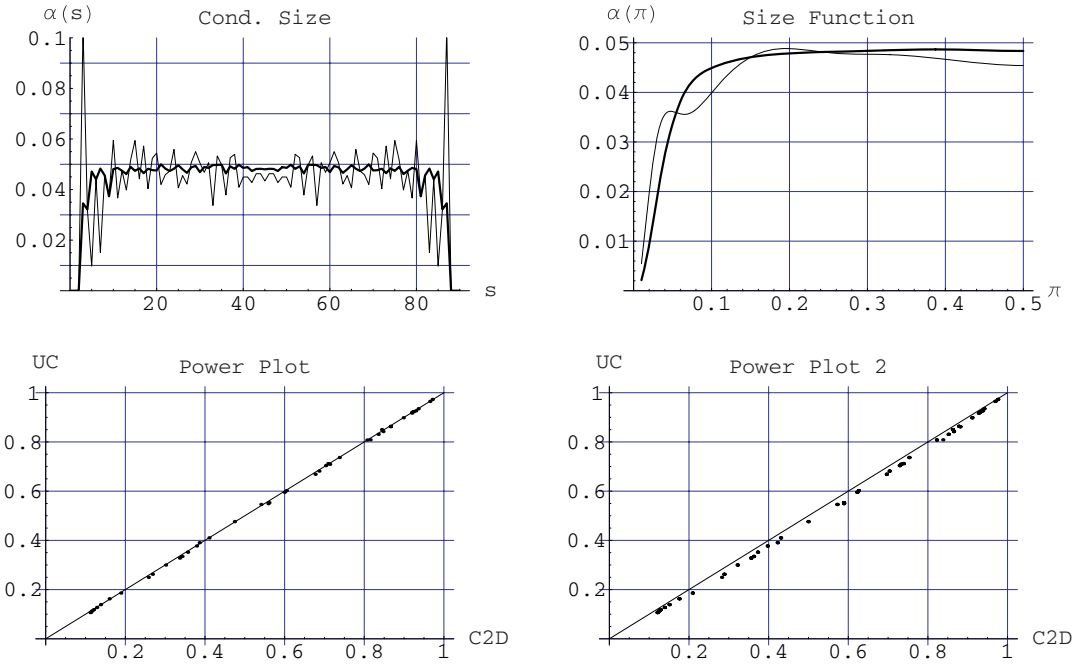



Figure 4.2: $\boldsymbol{n} = (30, 30, 30)$.

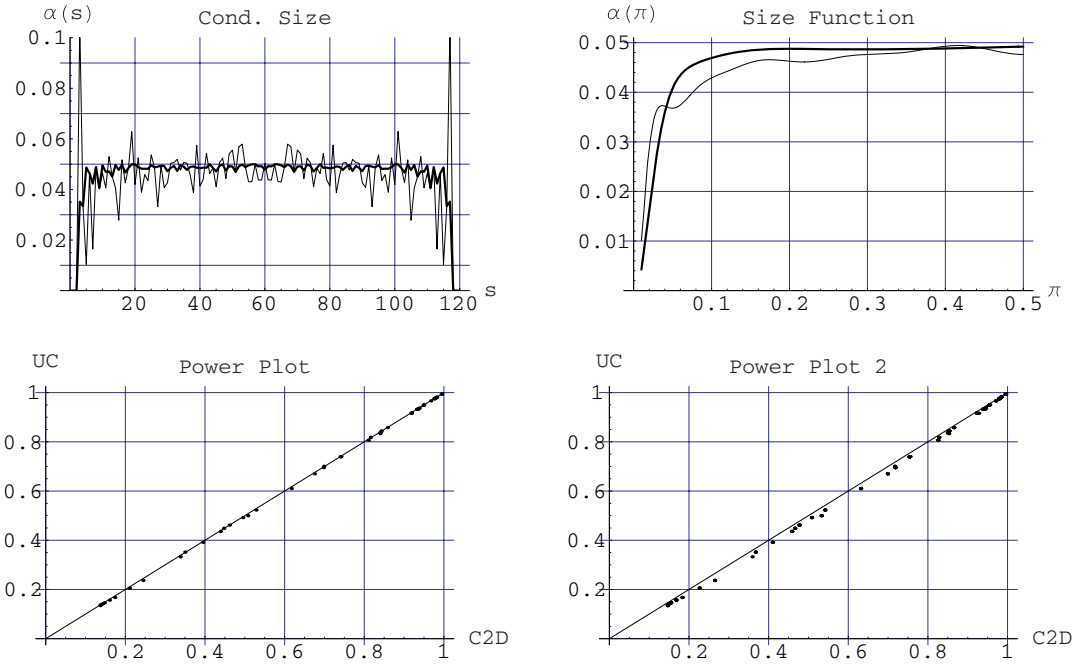



Figure 4.3: $\boldsymbol{n} = (40, 40, 40)$.

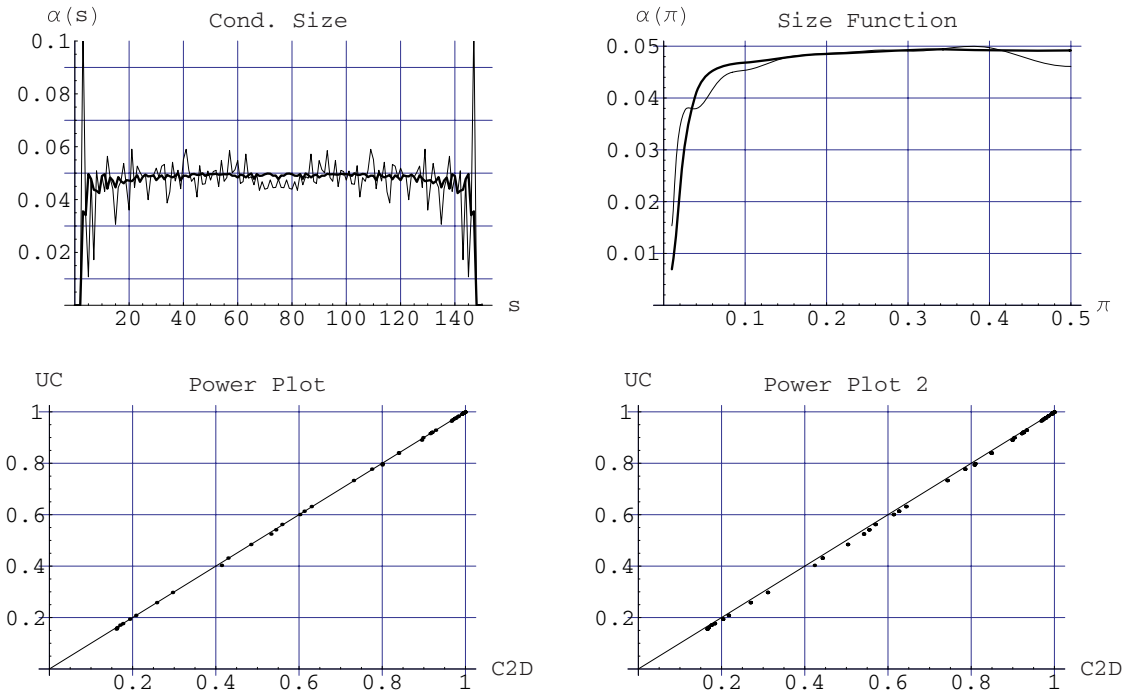



Figure 4.4: $\boldsymbol{n} = (50, 50, 50)$.

Next, we consider the second case that two of three sample sizes are equal and the other sample size is larger than the other two sizes by 1. That one sample size differs from the other two has an effect of reducing the discreteness of a test statistic. We observe from Figure 4.5 $\sim$ 4.8 that the change, we observed when sample sizes grow larger in the previous case, go on and the two tests show similar performances as sample sizes grow larger.

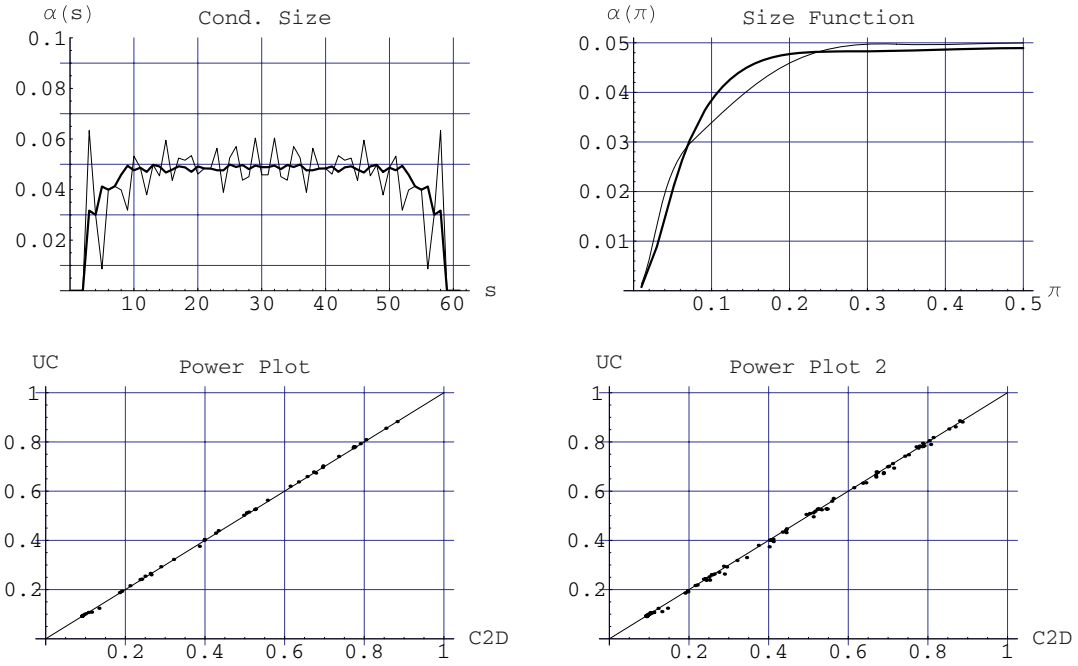



Figure 4.5: $\boldsymbol{n} = (20, 20, 21)$.

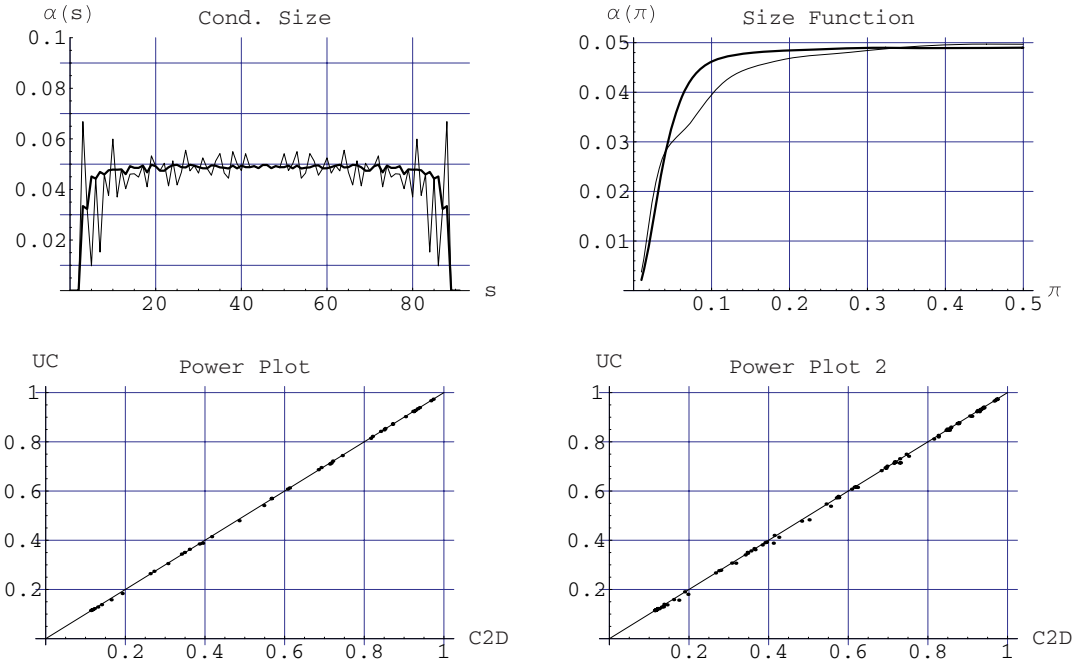



Figure 4.6: $\boldsymbol{n} = (30, 30, 31)$.

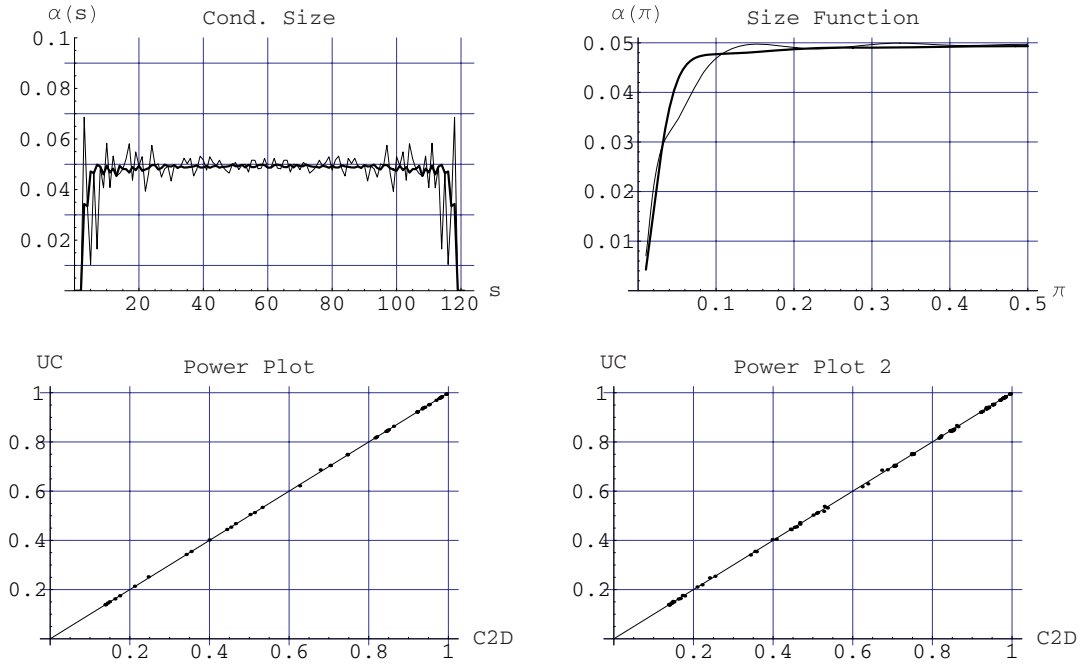



Figure 4.7: $\boldsymbol{n} = (40, 40, 41)$.

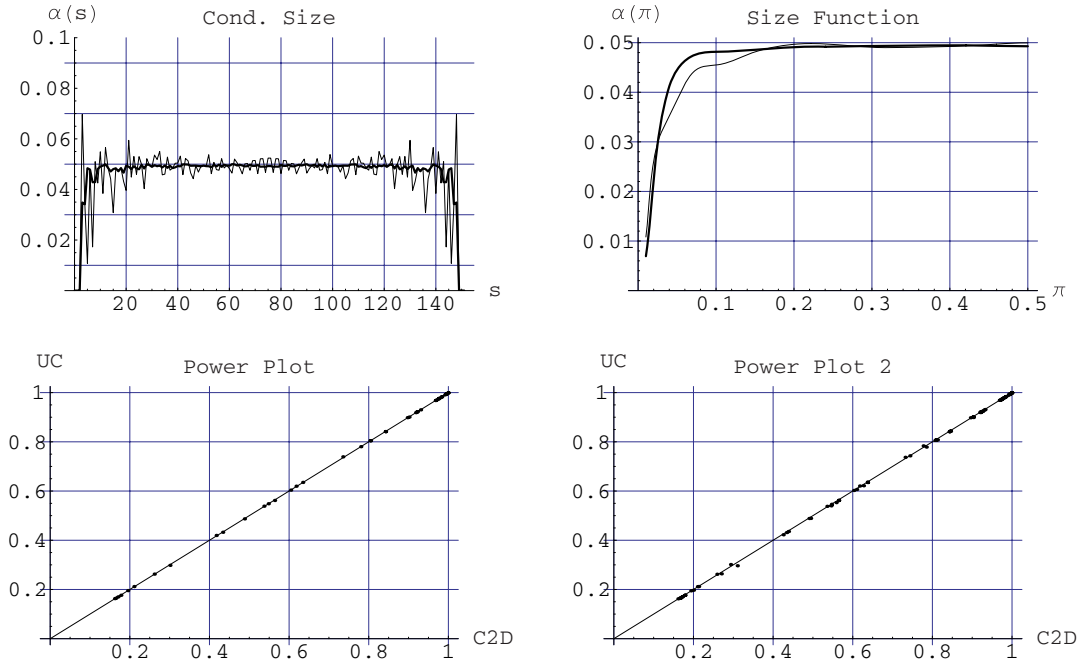



Figure 4.8: $\boldsymbol{n} = (50, 50, 51)$.

Observing over the whole settings comparing the two tests, the restriction of conditional tests that each conditional size should be no more than a nominal significance level is really disadvantageous, because the stability of conditional sizes of conditional tests can not overcome, in some settings, the allowance of conditional sizes of the unconditional test to be larger than the nominal significance level, as long as the size function does not exceed the nominal significance level.

# Chapter 5

# Unconditional modified test

As we observed in the previous chapter, that conditional sizes of a conditional test should be no more than a nominal significance level is really restrictive. In this chapter, we explore another exact test procedure having the desired properties of both the conditional and unconditional tests.

This chapter is based on Matsuo (2000b).

## 5.1 Derivation

The test procedure we are going to introduce is to carry out an unconditional test using a test statistic whose distribution on the whole sample space, $\Gamma$, is stable against the change in $\pi$ value.

The reason, why a conditional test is stable against the change in $\pi$ value, is that it is based on the distribution on the conditional reference set to which an observation belongs. To be more specific, whether an observation is rejected or not depends, not on its statistic value itself, but on the order of the statistic value in the corresponding conditional reference set. Hence, in a conditional test, even if we replace a conventional test statistic $T(\boldsymbol{Y})$ with the modified statistic $T^*(\boldsymbol{Y})$,

$$T^*(\boldsymbol{y}) = \Pr_{H_o}\{\ T(\boldsymbol{Y}) < T(\boldsymbol{y}) \mid \Gamma_s, \pi_o\ \} \quad \text{for } \boldsymbol{y} \in \Gamma_s, \tag{5.1}$$

we will surely have the same result as that given when the original statistic, $T(\boldsymbol{Y})$, is used.

The critical value of the conditional test using the modified statistic is $1-\alpha$, which is constant across the conditional reference sets, and observations whose $T^*$

values are no less than $1-\alpha$ are being rejected.

The advantage of introducing the modified statistic is that it can now be used in the context of unconditional test. The critical value of the unconditional test using the modified statistic is always no more than $1-\alpha$. When the critical value is strictly less than $1-\alpha$, we are able to attain uniformly higher power than the conditional test using the original statistic. We can also expect this test procedure would dominate the unconditional test using the original statistic, except for the odd case that the fluctuation of the conditional sizes of the ordinary unconditional test accidentally give rise to higher power, as observed in Chapter 3. We would like to call this procedure as the *unconditional modified test*, hereafter.

We note that we could use

$$1 - T^*(\boldsymbol{y}) = \Pr_{H_o}\{\ T(\boldsymbol{Y}) \geq T(\boldsymbol{y}) \mid \Gamma_s, \pi_o\ \} \quad \text{for } \boldsymbol{y},$$

instead of $T^*(\boldsymbol{Y})$ and reject $H_o$ if it is no more than $\alpha$. We, however, prefer a statistic that shows larger discrepancy as its value gets larger. So, we do not adopt $1 - T^*(\boldsymbol{y})$.

## 5.2 Numerical result

Now, we are going to display the relative performance of the unconditional modified test to the ordinary unconditional test, over all the settings of sample sizes considered in Chapter 2. As noted in Chapter 4, we employ only the *Pearson's* $X^2$ for the second half of this thesis. So, we construct the modified statistic from the *Pearson's* $X^2$. We consider three cases, the first case being three sample sizes are equal, the second case being two of three sample sizes are equal and the third case being three sample sizes are distinct. For each case, we consider four settings and, for each setting, we present four graphs. Upper-left graphs are presented for comparing conditional sizes of the two tests, where horizontal axis represents $s = \sum y_i$, thick line represents the unconditional modified test and thin line represents the ordinary unconditional test. Upper-right graphs display size functions, where horizontal axis represents $\pi$, thick curve represents the unconditional modified test, and thin curve represents the ordinary unconditional test. Lower-left graphs are the power plots of the two tests at the 44 simple alternatives, listed in section 2.3. We note here that the powers of

these graphs are the average powers over the permutations of each alternative, as in the previous section. Lower-right graphs are average power-difference plots, which are presented to show closely the power differences of the lower-left graph, plotted in the same order as the list in section 2.3.

At first, we consider the case that all three sample sizes are equal, where the discreteness of a test statistic is most feared. Figures 5.1 $\sim$ 5.4 are presented to display our calculation. Although, the ordinary unconditional test accidentally shows superior performance in small sample sizes, we could state that the performance of the unconditional modified test (abbreviated to "UCM") becomes better than the ordinary one, as sample sizes grow larger. From our computational experience, when sample sizes are small, say less than 30, the relative performance of a collection of test procedures depends on the specific sample sizes. It is, therefore, very hard to find a test which is always more powerful over the others with such sample sizes.

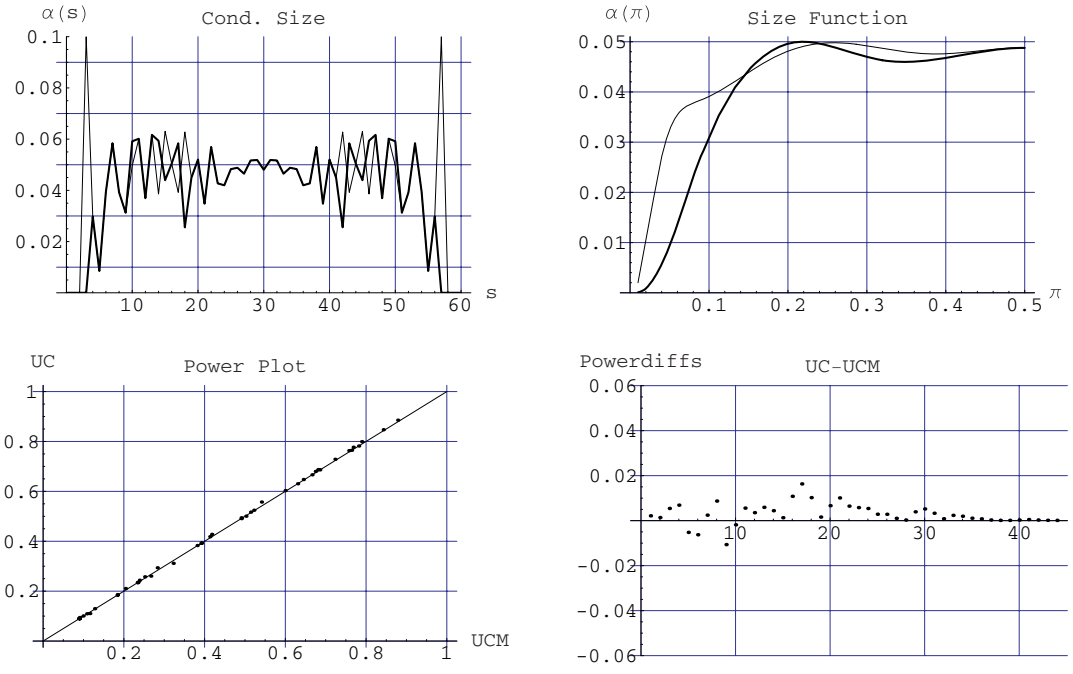



Figure 5.1: $\boldsymbol{n} = (20, 20, 20)$.

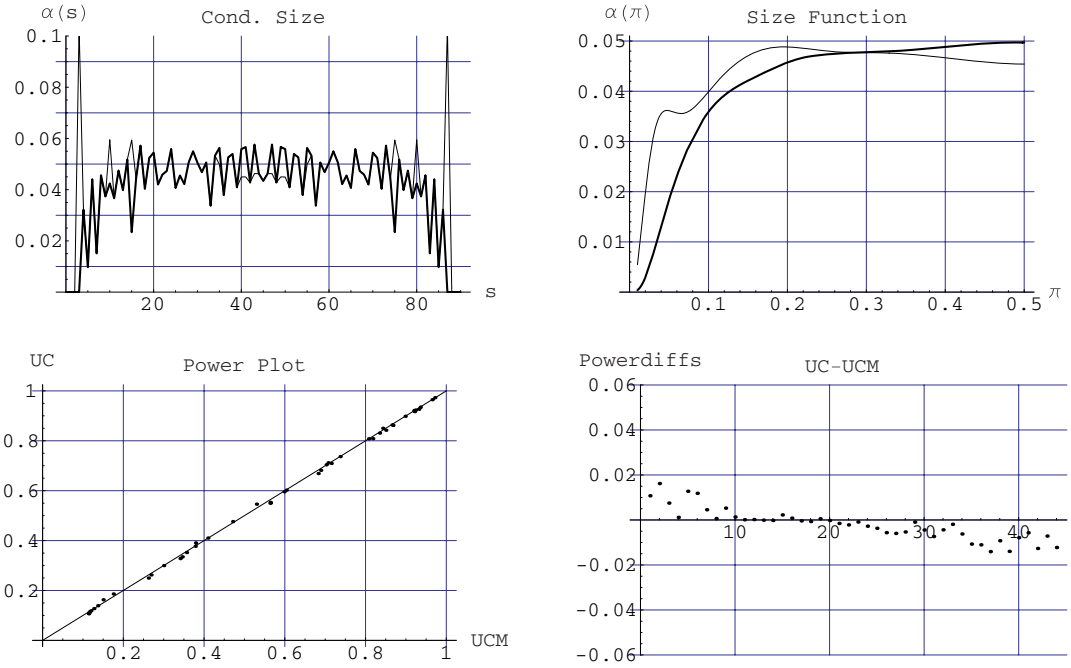



Figure 5.2: $\boldsymbol{n} = (30, 30, 30)$.

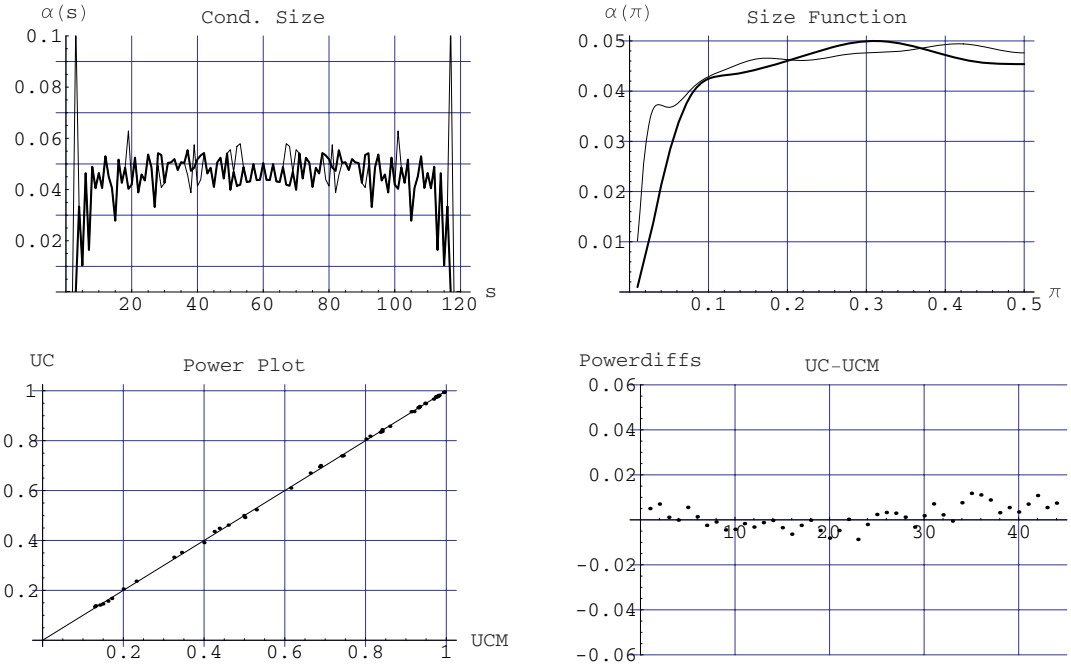



Figure 5.3: $\boldsymbol{n} = (40, 40, 40)$.

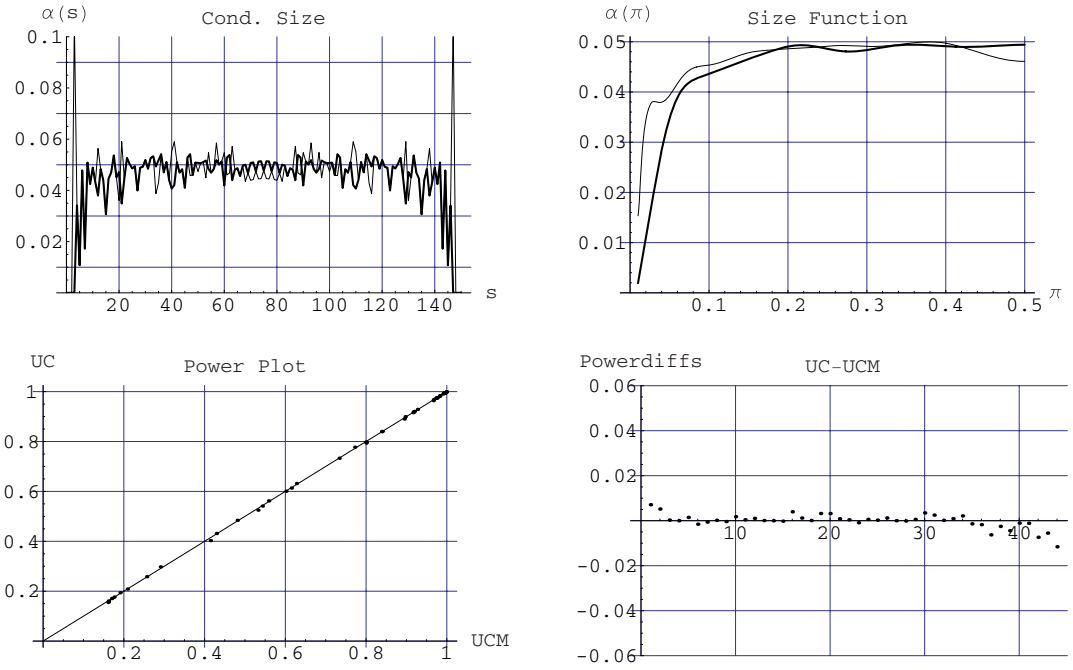



Figure 5.4: $\boldsymbol{n} = (50, 50, 50)$.

Next, we consider the second case that two of three sample sizes are equal and the other is larger by 1. In this case, the discreteness of a test statistic is expected to reduce and the tendency we have observed in the previous case continues to develop. Observing Figure 5.5 $\sim$ 5.8, we could state that our the unconditional modified test showed stable size performance and higher power as a whole, which is just what we have expected.

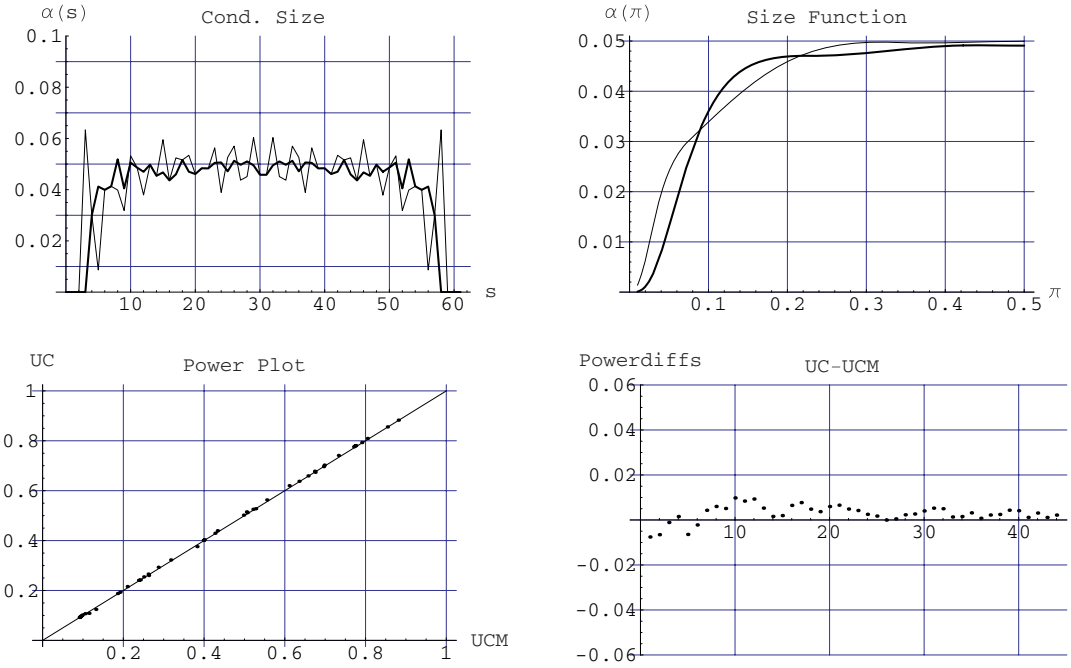



Figure 5.5: $\boldsymbol{n} = (20, 20, 21)$.

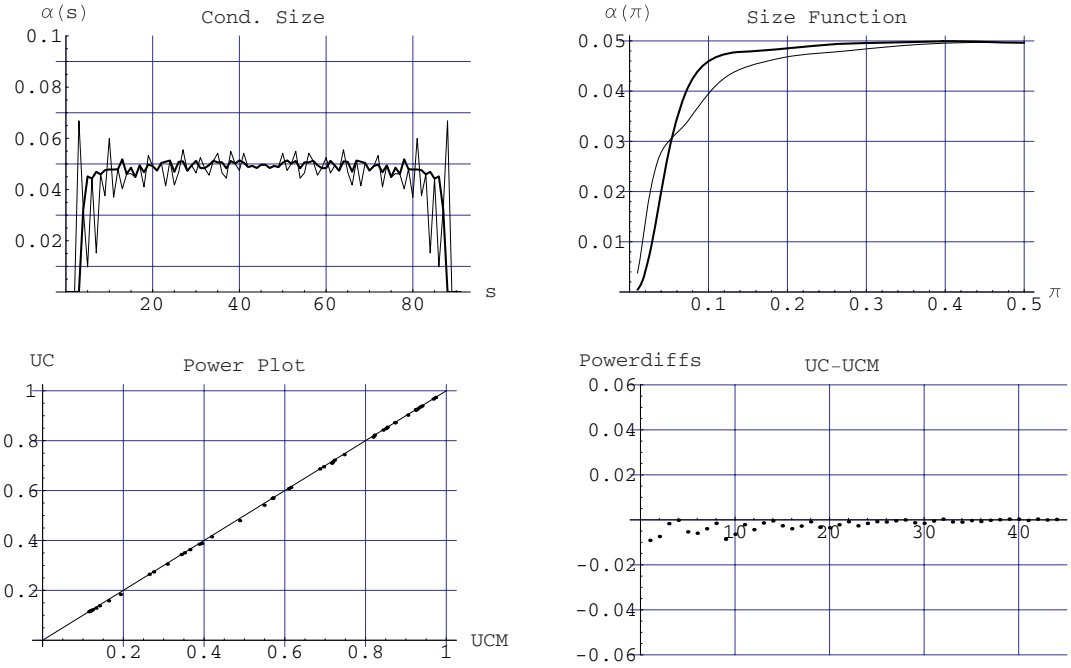



Figure 5.6: $\boldsymbol{n} = (30, 30, 31)$.

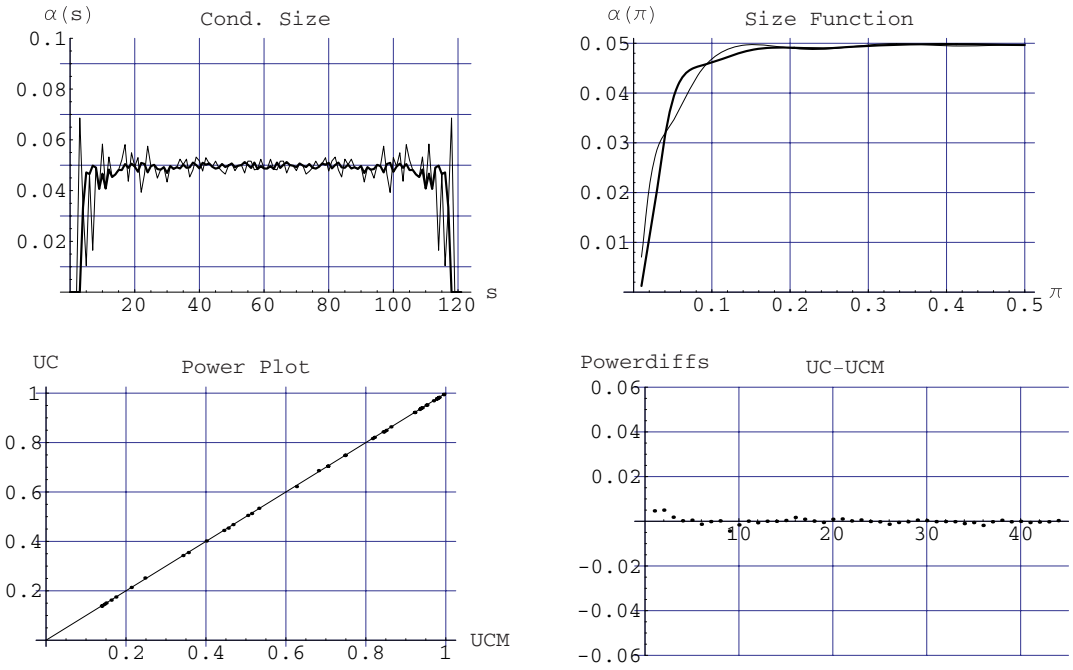



Figure 5.7: $\boldsymbol{n} = (40, 40, 41)$.

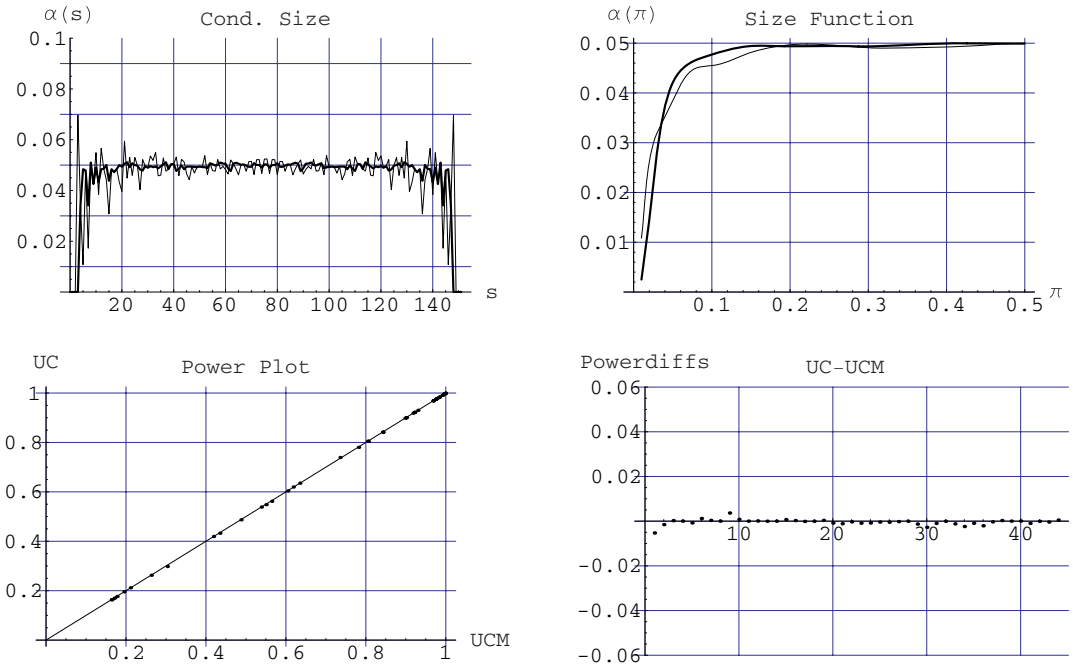



Figure 5.8: $\boldsymbol{n} = (50, 50, 51)$.

At last, we consider the third case that three sample sizes are distinct. We expect the tendency we have observed in the previous two cases would be seen most clearly. The results are displayed in Fugure 5.9 $\sim$ 5.12. The advantage of the unconditional modified test is clearly seen except for the setting $\boldsymbol{n} = (19, 20, 21)$, where sample sizes are small and any general rule is likely to hold. We have carried out more calculation than displayed in this section and observed consistent superiority of the unconditional modified test over the ordinary unconditional test, when sample sizes are more than 30 and distinct.

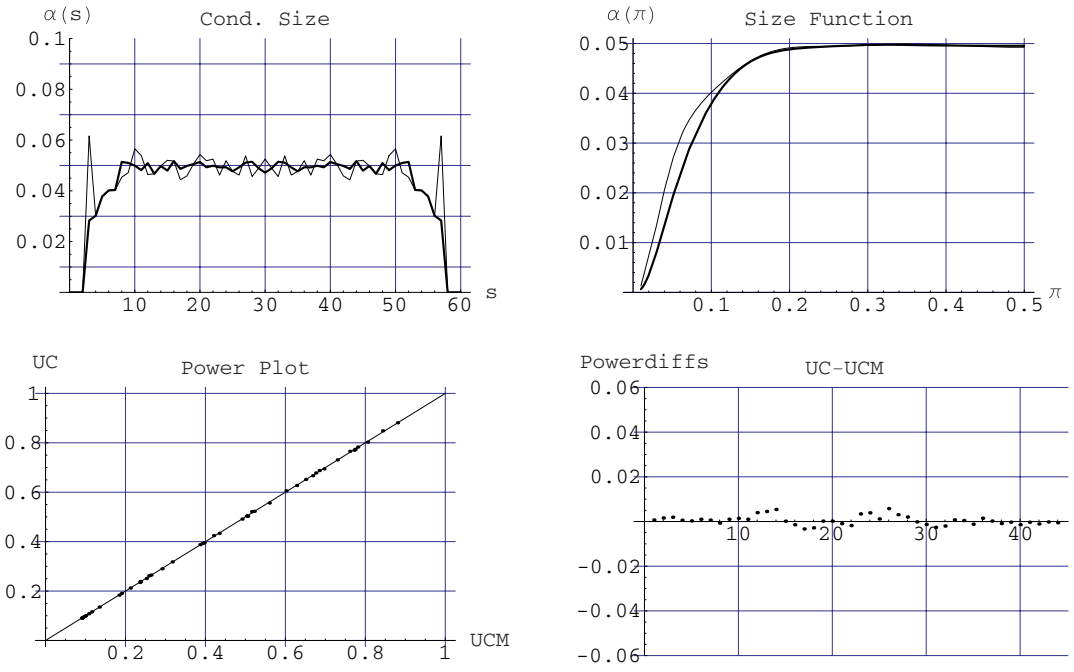



Figure 5.9: $\boldsymbol{n} = (19, 20, 21)$.

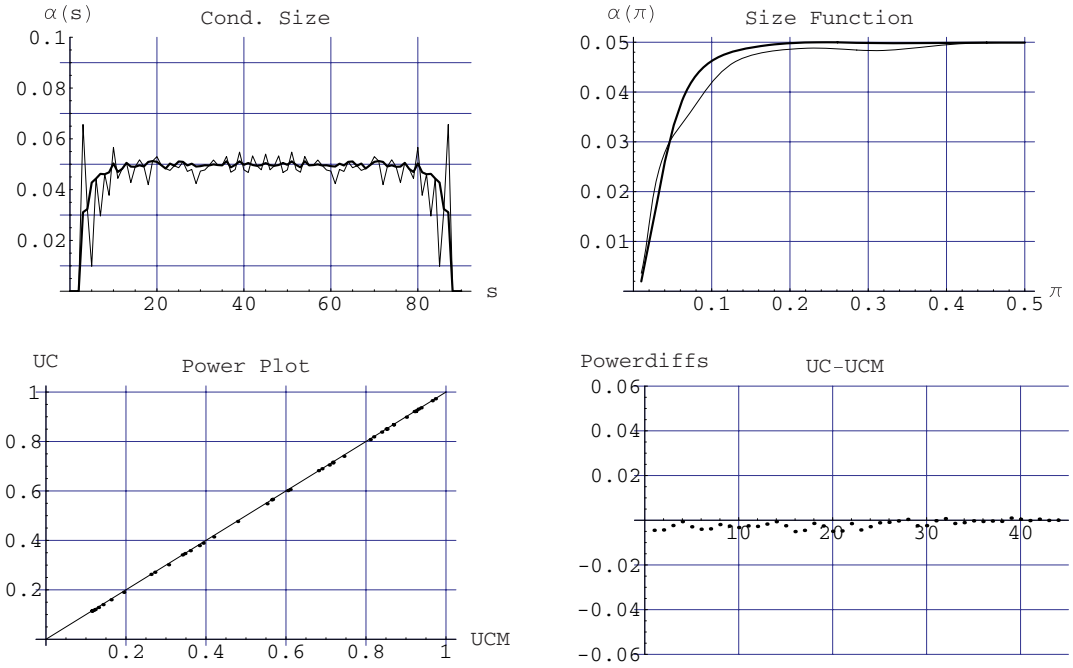



Figure 5.10: $\boldsymbol{n} = (29, 30, 31)$.

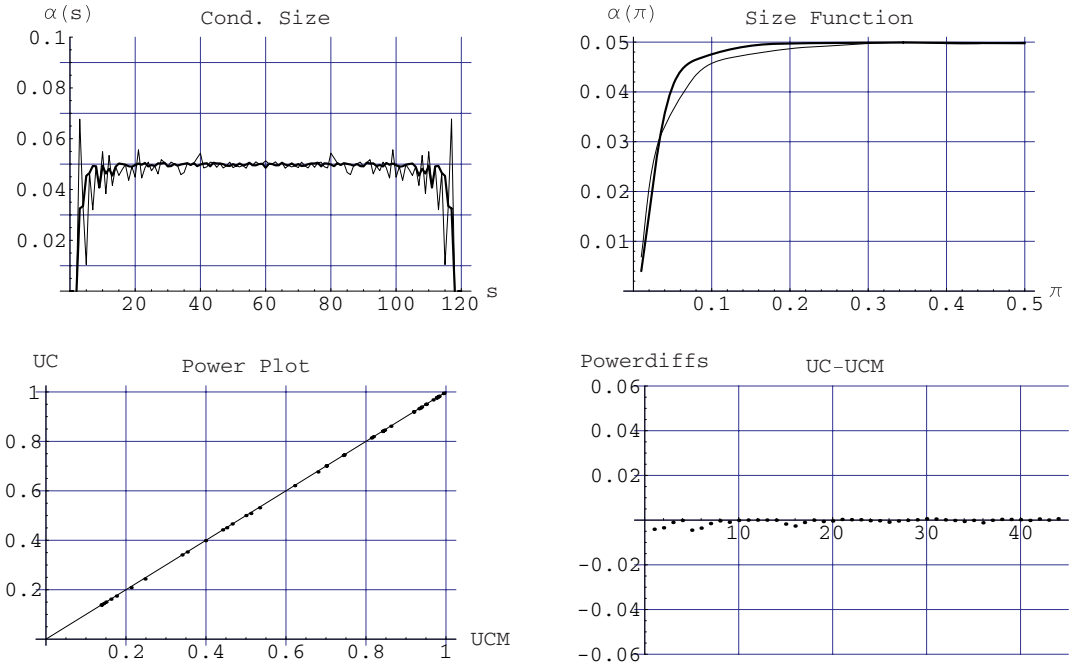



Figure 5.11: $\boldsymbol{n} = (39, 40, 41)$.

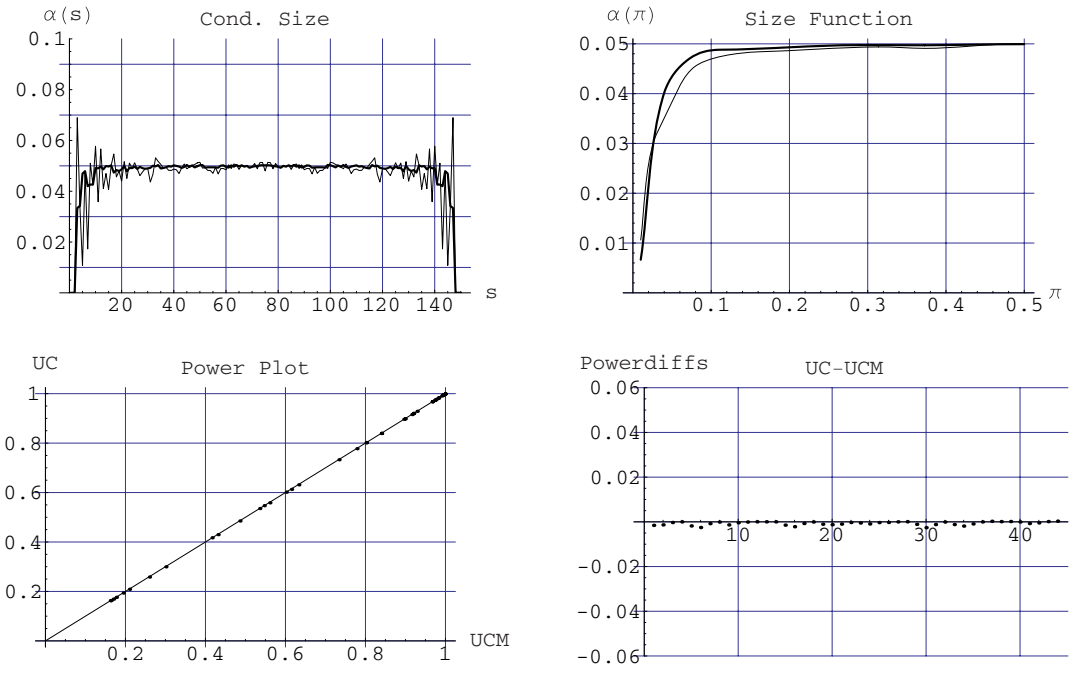



Figure 5.12: $\boldsymbol{n} = (49, 50, 51)$.

# Chapter 6

# Unconditional modified two-dimensional test

The purpose of this thesis is to contrive test procedures that are less conservative than the conventional exact tests, in the sense that the empirical size of the test being virtually a given fixed significance level and having, as a result, higher power. Even if these contrived procedures require a certain amount of additional calculation, this is not a critical issue nowadays considering the dramatic development of modern computing circumstances. Although we have introduced two such procedures in Chapter 4 and 5 and attained certain improvement over ordinary exact tests, the ordinary unconditional test accidentally shows higher powers, on a few sample size settings in the cases , $n_1 = n_2 = n_3$ and $n_1 = n_2 = n_3 - 1$. In this chapter, we are going to introduce a test procedure that is most computer intensive but expected to perform better in the two cases, $n_1 = n_2 = n_3$ and $n_1 = n_2 = n_3 - 1$, than the ordinary unconditional test.

This chapter is based on Matsuo (2001).

## 6.1 Derivation

We have observed in Chapter 4 and 5 that the performance of the conventional unconditional test was pretty good, and that we failed to find a procedure that is uniformly more powerful than that. However, we found that the two procedures introduced in the previous two chapters were not exclusive but can be used at the same time. We, at the end of this thesis, would like to propose a test procedure

that is a combination of two previously introduced procedures; that is, the unconditional test using a modified two-dimensional statistic. We call this procedure as the *unconditional modified two-dimensional test.* This new procedure is guaranteed to have consistently higher power than the conditional two-dimensional test, because the critical value of the unconditional modified two-dimensional test is allowed to be under $1-\alpha$, where $\alpha$ represents a nominal significance level, which is a consequence of introducing the modified statistic (5.1). And, although we can not expect consistent superiority of this procedure over the unconditional modified test and the ordinary unconditional test, we would be able to expect better performance of this procedure than the above two procedures, as a whole.

## 6.2 Numerical result

Here, we illustrate the relative performance of the unconditional modified two-dimensional test, to the ordinary unconditional test, over the settings of sample sizes considered in Chapter 4. As noted in Chapter 4, we employ only the *Pearson's* $X^2$ for the second half of this thesis. We construct the modified statistic from the two-dimensional statistic based on the *Pearson's* $X^2$. When sample sizes are distinct, there is no effect of introducing two-dimensional statistic in the unconditional modified test. Therefore, in this section, we do not treat the case of equal sample sizes. To carry out the unconditional modified two-dimensional test, we have to specify the values of the covariate $x_i$, $(i = 1, \cdots, k)$, as in Chapter 4. We specify $x_i = i$, corresponding to a dummy variable, the specification not fully advantageous to this new procedure, in the same sense explained in Chapter 4.

As in Chapter 4, we are going to consider two cases, first case being three sample sizes are equal, $n_1 = n_2 = n_3$, and second case being two of three sample sizes are equal, $n_1 = n_2 = n_3 - 1$. For each case, we consider four settings and, for each setting, we present four graphs. Upper-left graphs are presented for comparing conditional sizes of the tests, where horizontal axis is $s = \sum y_i$, thick line represents the unconditional modified two-dimensional test and thin line represents the ordinary unconditional test. Upper-right graphs are presented for displaying size functions, where horizontal axis is $\pi$, thick curve represents the unconditional modified two-dimensional test and thin curve represents the ordinary unconditional test.

Lower-left graphs are the power plots of the two test procedures at the 44 simple alternatives listed in section 2.3. As we noted in section 4.2, even when all sample sizes are equal, the power of the unconditional modified two-dimensional test varies over the permutations of an alternative, because of the introduction of the two-dimensional statistic. The powers of these graphs are the average powers over the permutations of each alternative. Lower-right graphs are the power plots of the 44 alternatives themselves, that is, we do not consider any permutations at all. What we want to show with these graphs is the performance of the unconditional modified two-dimensional test when the guess that the observations are listed in ascending expected-probability order is true.

Let us observe the first case that three sample sizes are equal and common sample sizes are 20, 30, 40 and 50. Comparing with Figure 4.1 $\sim$ 4.4 and Figure 5.1 $\sim$ 5.4, it is obviously observed that the conditional sizes of the unconditional modified two-dimensional test ( abbreviated to "UCM2D") consistently come up nearer toward $\alpha = 0.05$ line than the other two procedures introduced in the last two chapters, and that, comparing with the ordinary unconditional test, this new procedure show higher power at almost all alternatives on almost all sample size settings.

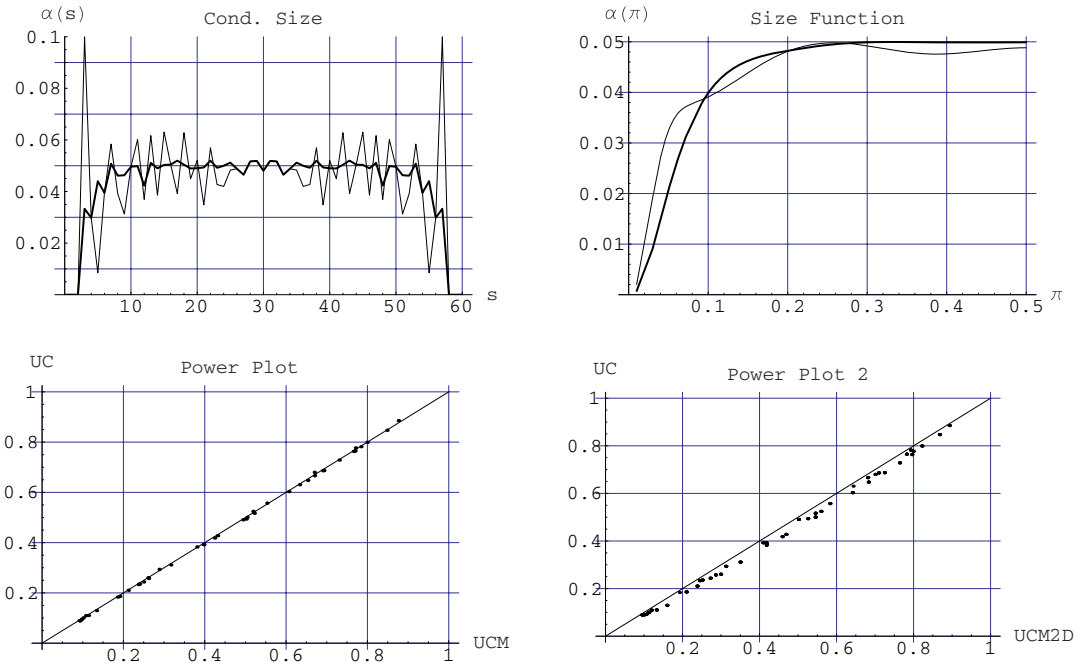



Figure 6.1: $\boldsymbol{n} = (20, 20, 20)$.

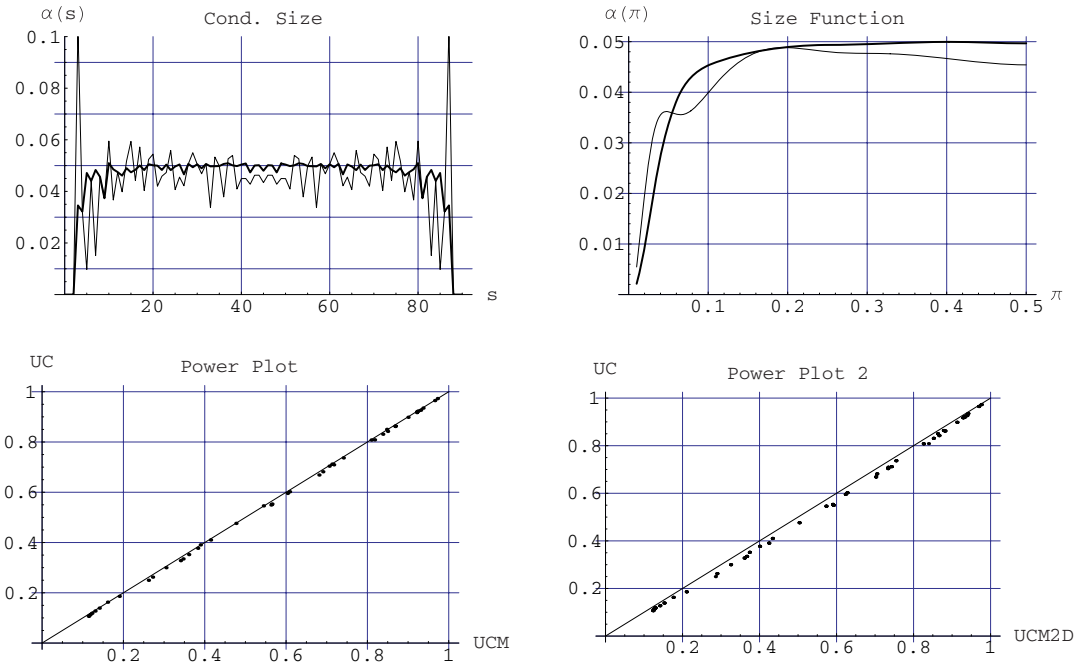



Figure 6.2: $\boldsymbol{n} = (30, 30, 30)$.

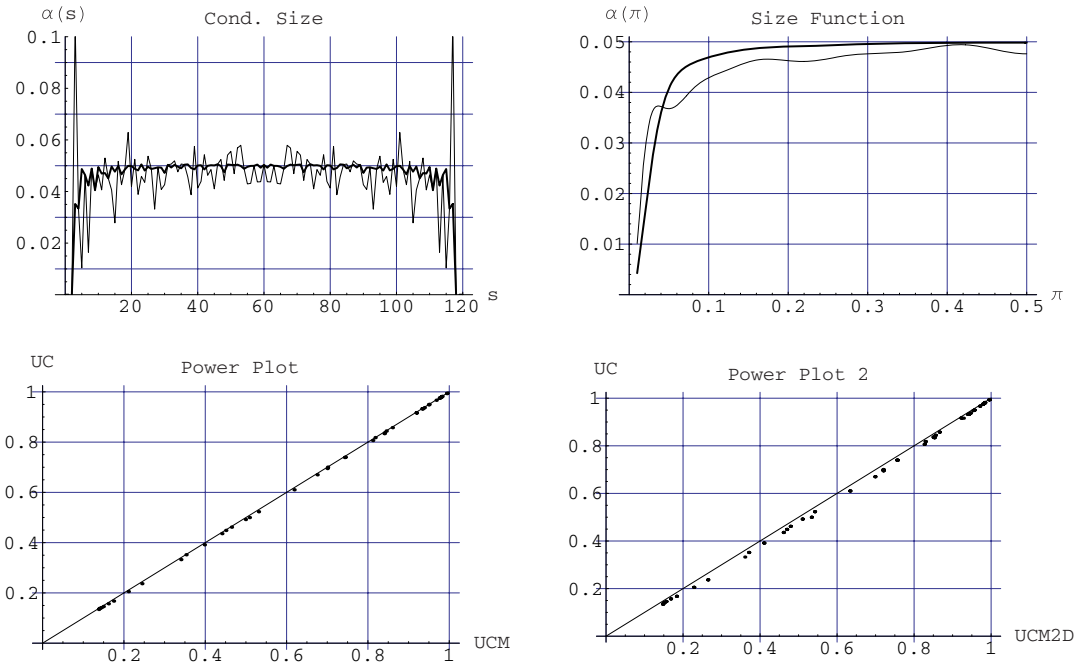



Figure 6.3: $\boldsymbol{n} = (40, 40, 40)$.

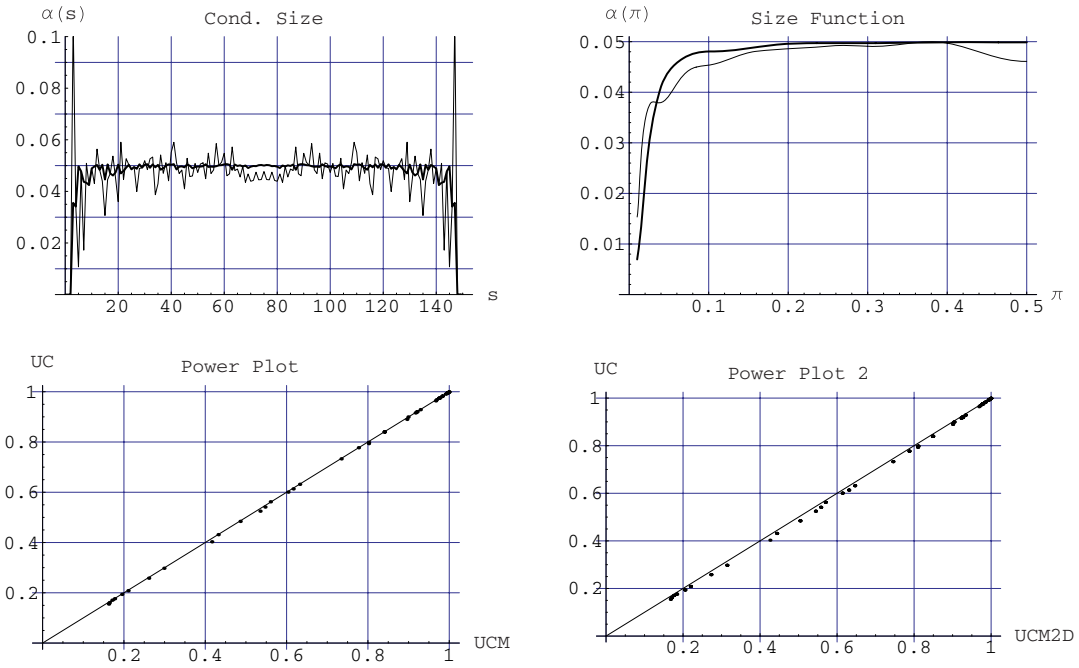



Figure 6.4: $\boldsymbol{n} = (50, 50, 50)$.

Next, we consider the second case that two of three sample sizes are equal and the other sample size is larger than the two by 1. That one sample size differs from the other sample sizes has an effect of reducing the discreteness of a test statistic. We observe from Figure 6.5 $\sim$ 6.8 that the change, we have observed as sample sizes grow larger in the previous case, goes on and the two tests show similar performances as sample sizes grow larger.

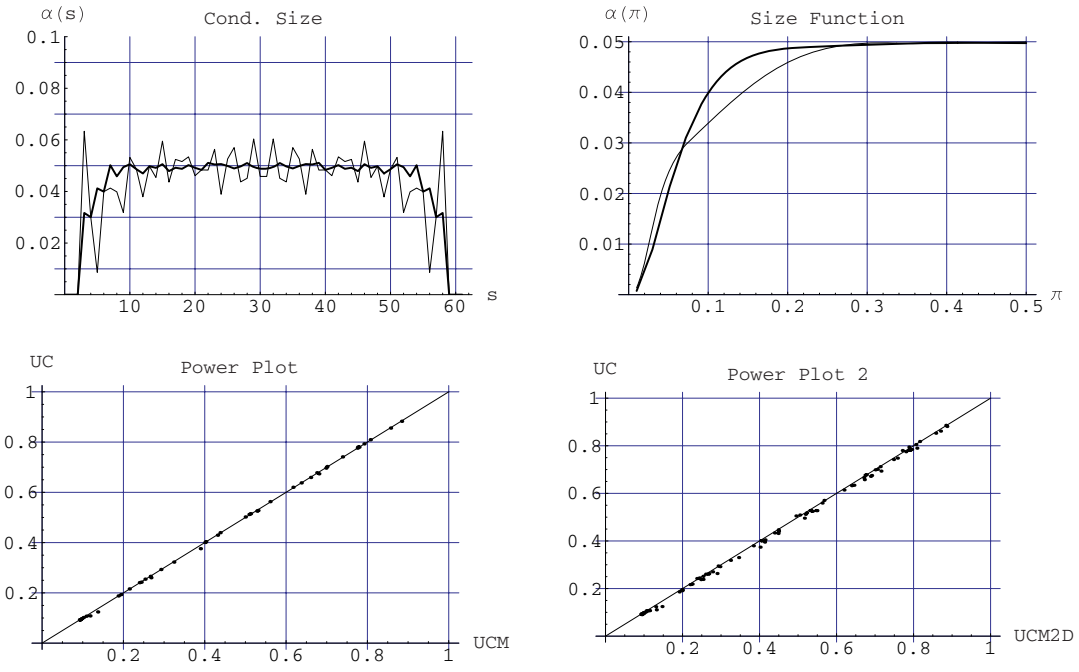



Figure 6.5: $\boldsymbol{n} = (20, 20, 21)$.

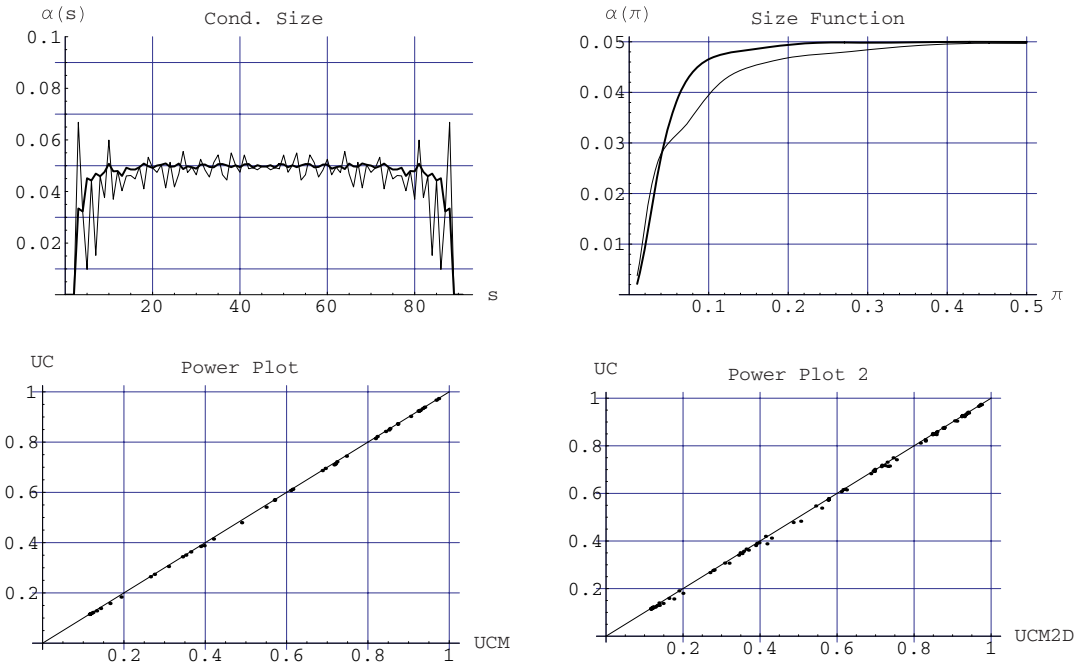



Figure 6.6: $\boldsymbol{n} = (30, 30, 31)$.

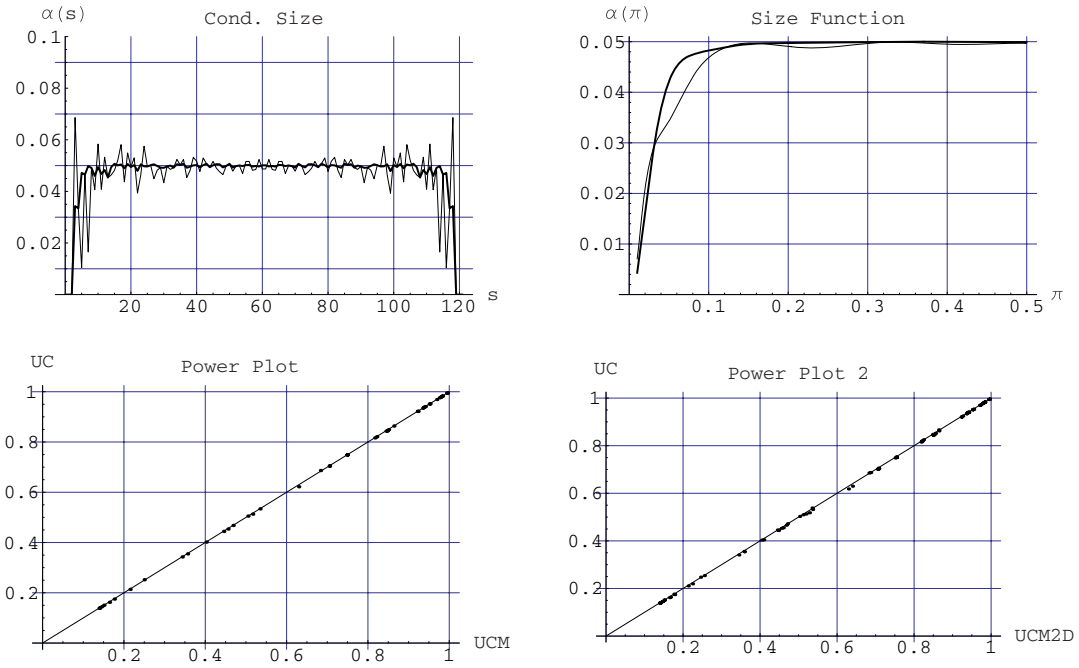



Figure 6.7: $\boldsymbol{n} = (40, 40, 41)$.

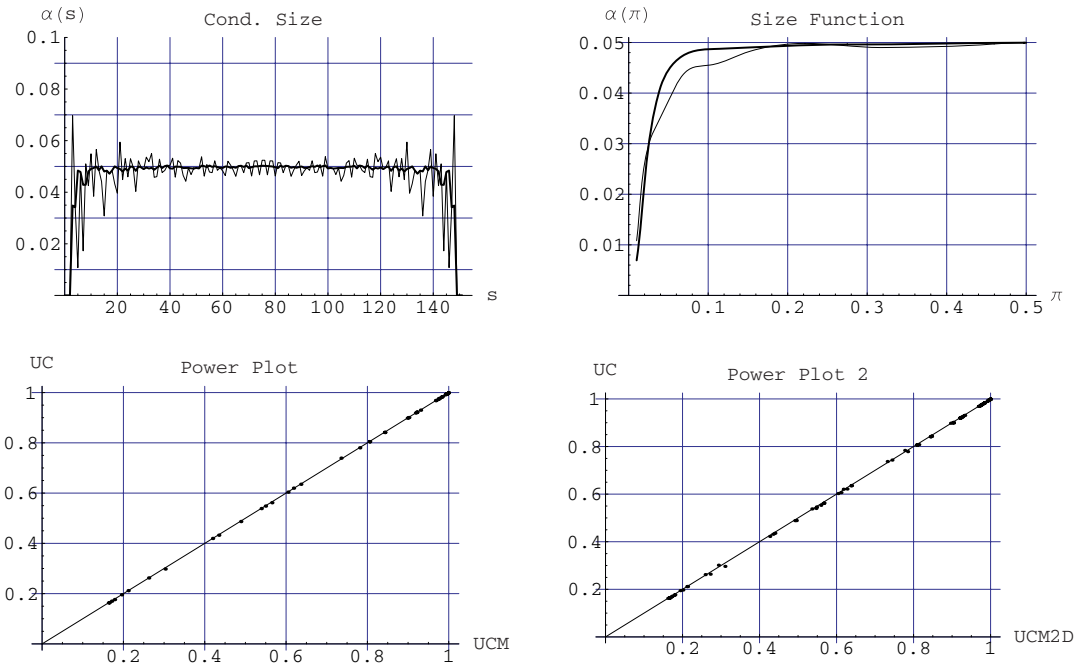



Figure 6.8: $\boldsymbol{n} = (50, 50, 51)$.

Observing the whole figures in this section and Figure 5.9 $\sim$ 5.12, we could state that our new test procedure would yield higher power than the ordinary unconditional test except for irregular settings, where the unstable performance of the ordinary unconditional test accidentally attains higher power.

# Chapter 7

# Conclusion

In the first half of this thesis, we have investigated the exact behavior of the exact conditional and unconditional tests, owing to the modern computational circumstances. The first discovery is that the *deviance* performs very poorly in the unconditional test. On the other hand, the performance of the *Pearson's* $X^2$ and the *power divergence* in the unconditional test is usually better than that in the conditional test. We have also discovered that, when comparing the exact conditional and unconditional tests, there was some room for improving the conventional exact tests.

In the second half of this thesis, following the discovery just above, we have pursued the possibility of improving the conventional exact conditional and unconditional tests. We introduced the conditional two-dimensional test, for the case that some of sample sizes are equal, to order the observations sharing a tied statistic value resulting from the equality. We also introduced the unconditional modified test to implement a test that has both the advantages of the conditional and unconditional tests. We would like to propose to carry out the unconditional modified two-dimensional test, which is the unconditional modified test employing the two-dimensional statistic. We note that, in the case of distinct sample sizes, the unconditional modified two-dimensional test coincides with the unconditional modified test. By adopting our proposal, we are able to implement a test that is always more powerful than the conventional unconditional test, except for some irregular settings of sample sizes and alternative hypotheses. The stable performance of the unconditional modified two-dimensional test can reasonably be expected even with

some settings of sample sizes, where the ordinary unconditional test might show poor performance like that we have observed with the *deviance* (although, we have not encountered such a phenomenon with the *Pearson's* $X^2$ ).

It is true that the amount of calculation, needed to investigate the performance of an unconditional modified two-dimensional test, is enormous. However, the amount of calculation, needed to implement an unconditional modified two-dimensional test, is far trivial than that. To implement an unconditional modified two-dimensional test, we at first calculate the value of the modified two-dimensional statistic referring to the conditional reference set, to which an observation belong. If the value is no less than $1 - \alpha$, we conclude that the evidence against the null hypothesis is significant. Else if the value is slightly less than $1 - \alpha$, we further calculate the minimum value of the modified two-dimensional statistic that is no less than the observed modified two-dimensional statistic value in each conditional reference set, and then draw the graph of the weighted average of the values as a function of $\pi$, just like the size function (1.18). If the function is uniformly no less than $1 - \alpha$, we conclude significant, otherwise not significant. Extra calculation is needed only when the observation has a subtle evidence against the null hypothesis.

Although we have confined our discussion in fixed-level testing throughout the thesis, the test procedures introduced in Chapter 4, 5 and 6 can also be utilized to report the observed significance level ($p-$value). These $p-$values are considered to be more elaborate than ordinary mid $p-$value. At the end of this thesis, we would like to emphasize again that our discussion is applicable to four or more sample cases, in spite of our illustrations being confined to three sample case.

# Acknowledgements

I would like to express my gratitude to Professor N. Inagaki of Osaka University for his patient guidance and valuable advises, which made up my foundation as a reseacher. I am grateful to Professor S. Shirahata of Osaka University for his helpful suggestions and constant encouragement, which enabled me to accomplish this thesis. I would like to thank Professor M. Goto of Osaka University and Professor T. Isogai of Kobe University of Mercantile Marine for their valuable advices.

# Bibliography


[1] Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science,* **7,** 131-177.

[2] Cochran, W.G. (1952). The $\chi^2$ test of goodness-of-fit. *Annals of Mathematical Statistics,* **23,** 315-345.

[3] Collett, D. (1991). *Modelling Binary Data.* London: Chapman and Hall.

[4] Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B,* **46,** 440-464.

[5] Cox, D. R. (1984). Discussion of "Tests of Significance for $2 \times 2$ contingency tables." by F. Yates. *J.Roy.Statist.Soc. Ser. A,* **147**, 451.

[6] Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data, 2nd Ed.* Chapman and Hall, London.

[7] Davison, A. C. (1988). Approximate conditional Inference in generalized linear models. *Journal of the Royal Statistical Society, Series B,* **50,** 445-461.

[8] Freeman, G.H. and Halton, J.H. (1951). Note on an exact treatment of contingency tables, goodness-of-fit and other problems of significance. *Biometrika* **38,** 141-149.

[9] Inagaki, N (1990). (in Japanese) *Statistical Mathematics*, Shokabo, Tokyo.

[10] Kawai, N. (1997). Process for analyzing dichotomous response data. Master thesis, Department of , Faculty of , Osaka University.

[11] Lehmann, E. L. (1986). *Testing Statistical Hypotheses, 2nd Ed.* Wiley.

[12] Little, R. J. A. (1989). Testing the equality of two Independent binomial proportions. *The American Statistician,* **43**, 283-287.

[13] LogXact-Turbo (1993). *Logistic Regression Software Featuring Exact Methods, User Manual*, Cytel Software Corporation.

[14] *The Mathematica book*, 3rd ed. (1996) Cambridge University Press.

[15] Matsuo, A (1999). Exact unconditional power comparison of three statistics in testing the equality of three binomial proportions. *J. Jpn. Soc. Comp. Statist.,* **12,** 1-14.

[16] Matsuo, A (2000a). (in Japanese) On the use of two-dimensional statistics for testing linear logistic model conditionally and exactly. *Japanese J. of Applied Statistics,* **29,** 1-25.

[17] Matsuo, A (2000b). (in Japanese) On exact unconditional tests using modified statistics. *Bulletin of the Comp. Statist. of Japan,* **13,** 41-57.

[18] Matsuo, A (2001). (in Japanese) On the performance of the unconditional test using modified two-dimensional statistic. *The Economic Review of Kansai University,* **51,** 163-177.

[19] McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association,* **81,** 104-107.

[20] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Ed.* London: Chapman and Hall.

[21] Mehta, C. R. and Hilton, J. F. (1993). Exact power of conditional and unconditional tests: going beyond the $2 \times 2$ contingency table. *The American Statistician,* **47,** 91-98.

[22] Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables, *J. Amer. Soc. Assoc.*, **78**, 427-434.

[23] Read, T. and Cressie, N. (1988). *Goodness-of-fit statistics for Discrete Multivariate Data.* Springer Verlag, New York.

[24] Suissa, S. and Shuster, J.J. (1985). Exact unconditional sample sizes for the $2 \times 2$ binomial trial. *Journal of the Royal Statistical Society, Series A* **148,** 317-327.

[25] Takeuchi, K and Fujino, K. (1981). (in Japanese)Binomial distribution and Poisson Distribution, Tokyo University Press, Tokyo.

[26] Taneichi, N. and Sekiya, Y. (1995). Power approximations of the test of homogeneity for multinomial populations. *Journal of the Japan Statistical Society,* **25,** 97-109.

[27] Weerahandi, S. (1995). *Exact Statistical Methods for data analysis.* Springer Verlag, Now York.

[28] Yanagawa, T (1986). (in Japanese) *Analysis of Discrete Multivariate Data*, Kyoritu-shuppan, Tokyo.

[29] Yates, F. (1984). Tests of Significance for $2 \times 2$ contingency tables (with discussion). *J.Roy.Statist.Soc. Ser. A,* **147**, 426-463.