The p-value for the sign test

Ana Marta Lisboa Vila Fernandes
anavilafernandes@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2015

Abstract

The sign test is a non-parametrical statistical procedure used to evaluate hypothesis about the q-th continuous population quantile. Denoted \( \chi_q \), with \( 0 < q < 1 \), the q-th continuous population quantile, on the two-sided sign test the null hypothesis states \( H_0 : \chi_q = \chi_0 \) while in the alternative \( H_1 : \chi_q \neq \chi_0 \). The test procedure is based on the sign differences between the sample observations and \( \chi_0 \). Considering the statistic \( S_n \) that accounts the frequency of the sign (+) on the sample of \( n \) observations, it is clear that \( S_n \sim \text{Bin}(n,p) \) with \( p = 1 - q \). When the observed value of the statistic \( s_n \) is not close to \( np \), the sample favours to the alternative hypothesis. A decision rule can be conducted based on the p-value. Nevertheless, the commonly used p-value formula was originally applied on two-sided tests for statistics with continuous distributions. In the case of discrete or continuous and asymmetric statistics, the usual p-value formula can lead to incorrect values. For the two-sided sign test incoherent p-values can be obtained under particular conditions of the binomial statistic. The main goal of this thesis is to address in which situations the two-sided sign test p-values meaning is not clear. In order to solve this problem, there are some alternatives proposed in the literature and their advantages and disadvantages were analyzed. Also, a new p-value formula was introduced and its efficiency was compared with the alternative methods for the two-sided sign test.

Keywords: Binomial Distribution, P-value methods, Power of the Sign Test, Two-sided Sign Test.

1. Introduction

The great boost for the realization of this thesis was the fact that the common p-value formula applied to two-sided sign tests may return probabilities greater than one. Nevertheless this happening occurs under particular conditions, which depend on the symmetricity or asymmetricity of the binomial statistic distribution and its concretization for the sample. Therefore this fact is often ignored or even not noticed by whom applies the test. For the two-sided tests, and taking the observed value of the statistics as the reference, it is common practice to compute the p-value as twice the probability on the tail with lower probability. However, problems can arise when the test statistic is discrete. In the continuous case, all points of the real line has an associated probability that can be attainable. In the case of discrete the scenario is different. To illustrate this, let us assume two consecutive points \( n \) and \( m = n + 1 \) under the null hypothesis of the test statistics, with probabilities \( p_n \) and \( p_m \), respectively, such that \( p_n \neq p_m \). In this case, since there are no intermediate points between \( n \) and \( m \), the probability \( p \) such that \( p_n < p < p_m \), is never attainable. This can raise problems in hypothesis tests, for example a fixed significance level (e.g. 5 %) may be not attainable. To illustrate the occurrence of p-value greater than one, taking up the specific case of the random variable \( X \sim N(10,2) \) and the sample of 6 observations generated with the statistical package R (R Core Team, 2014):

\[
H_0 : \chi_{\frac{1}{2}} = 10 \; \text{versus} \; H_1 : \chi_{\frac{1}{2}} \neq 10.
\]

The sign test statistic is \( S_6 \sim \text{Bin}(6, \frac{1}{2}) \) and the observed statistic value is \( s_6 = 3 \). Using the usual formula for calculating the p-value, the obtained value is:

\[
p-value = 2 \times \min\{P(S_6 \leq 3), P(S_6 \geq 3)\} = 2 \times 0.6563 = 1.3125 > 1.
\]

The main goal of this thesis is to warn of situations that produce p-values greater than one in the two-sided sign test. Furthermore we also introduce a new alternative to compute the p-value, which led to satisfactory results as compared to existing alternatives in the literature.

2. P-Value

Under the null hypothesis \( H_0 \), the p-value is the probability of observing a result equal to or more
extreme than the observed value of the test statistic. To calculate a p-value it is necessary to know at least three conditions: null hypothesis under test, distribution function of the test statistic and a possible ordering of the data to allow identification of the data greater than \( \chi_0 \). This ordering is conveniently organized in terms of a test statistic. Usually as a p-value is a frequentist concept, the sample space is defined in terms of possible outcomes that can occur from an infinite repetition of the random experience.

2.1. The Ronald Fisher p-value
Fisher (1920) was the work that formally introduced the notion of p-value and its calculation. In the context of classical inference, Fisher’s approach focuses only in one hypothesis and as final conclusion the hypothesis rejection or lack of evidence for that. According to Fisher (1935), “Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis”. The definition of the p-value according to Fisher is similar to the currently used, but the notions of how it should be used and interpreted are a bit different: first of all, it is a measure of evidences in a single experiment, used to reflect the credibility of the null hypothesis in the data. Secondly, as a measure of evidences, the p-value should be combined with other sources of information of the phenomenon under study. In short, in Fisher’s point of view the p-value is an index that calculates the degree of evidence against the null hypothesis. It says that a p-value ≤ 0.05 (0.05 × 100% significance) is a standard level that allows us to conclude that there is evidence against the hypothesis tested, although not as an absolute rule.

2.2. The Neyman-Pearson p-value
Jerzy Neyman and Egon Pearson of Neyman and Pearson (1928) developed an alternative theoretical framework. In 1928, they published a background paper on the theoretical foundation for a process they called hypothesis test. They introduced the idea of alternative hypothesis and type II error associated to it. In this approach, the error rate must be given priority to data collection, defining by \( \alpha \) the error rate of type I and by \( \beta \) the error rate of type II. The fact of conducting the test with emphasis in the Type I error value is related with repeatability concept of the sample. For Neyman and Pearson, the improper rejection of \( H_0 \) can be done in \( \alpha \times 100\% \) of the decisions and therefore they propose to attribute small values to \( \alpha \).

2.3. Comparison between the two approaches
It is interesting that these two approaches are not competitors, as they address the problem in a different perspective. Frequentists analysis usually perform the calculations by adopting the properties of the Neyman-Pearson methodology. The conclusion is given from the Fisher perspective, rejecting the null hypothesis and taking the decision according to the comparison of the pre-specified level of significance and the p-value obtained. It is a messy procedure since these approaches are developed according to different principles and aiming to different responses, and may even result in antagonistic conclusions. The main differences identified in these two approaches are the hypothesis on test and associated error rates. Fisher mainly focuses on the type I error and in the probability of rejecting the null hypothesis when it is in fact true. Neynam and Pearson added to the Fisher’s concern the type II error. For those statisticians, the probability of not rejecting the null hypothesis when it is false should also be controlled.

3. The sign test
The sign test is a nonparametric alternative to evaluate hypotheses about the quantiles, location parameters of distributions. The sign test may be applied to one sample or two paired samples. In this paper only the first situation is addressed, namely the test for the quantile \( \chi_q \) of a distribution \( X \). The stated null hypothesis, \( H_0 \), is that \( q – th \) quantile of \( X \), \( \chi_q \), is equal to \( \chi_0 \). The distribution of \( X \) does not require to be continuous or symmetric, as mentioned Chakraborti and Gibbons (2004, page 62). However, this test is usually applied when the population \( X \) is continuous. For simplicity, in this paper we will deal only with samples from continuous populations. Assuming that the random variable \( X \) as distribution function \( F(x) \), is a continuous and strictly increasing function. This test is based on the fact that under \( H_0 \) approximately \( nq \times 100\% \) of the n sample observations will be below \( \chi_0 \) and \( n(1 - q) \times 100\% \) will be greater than \( \chi_0 \). Taking in to account the differences \( (x_i - \chi_0) \), \( i = 1 \ldots n \), if the number of positive signs is approximately equal to \( n(1 - q) \) then \( H_0 \) should not be rejected.

3.1. Hypothesis testing
Consider the case of a continuous population \( X \) with distribution function \( F(x) \), and a concrete sample \( (x_1, x_2, \ldots, x_n) \) to perform a test on the parameter \( \chi_q \). The two-sided test states the hypotheses:

\[
H_0 : \chi_q = \chi_0 \text{ versus } H_1 : \chi_q \neq \chi_0,
\]

with \( 0 < q < 1 \).

3.2. Test statistic
As previously mentioned under the validity of \( H_0 \) it is expected that in the \( n \) sample values approximately \( nq \) of those values are smaller than \( \chi_0 \) and approximately \( n(1 - q) \) of those values are greater
than \( \chi_0 \). Denote by \( S_n \) the random variable consisting of the number of observations greater than an \( \chi_0 \) in a random sample of size \( n \) from \( X \). Using the indicator function, of follows that \( S_n = \sum_{i=0}^{n} \mathbb{I}(x_i - \chi_0) \) equals the number of positive signs in the sample of differences \( (x_i - \chi_0) \), where:

\[
\mathbb{I}(x_i - \chi_0) = \begin{cases} 
1, & \text{if } x_i > \chi_0 \\
0, & \text{if } x_i \leq \chi_0 .
\end{cases}
\]

Since \( P(\mathbb{I}(X - \chi_0) = 1|H_0) = 1 - q \), it follows that \( S_n|H_0 \sim Bin(n, 1 - q) \). Note that the distribution of \( S_n \) under \( H_0 \) does not dependent on the population distribution, \( F(X) \), from which the sample is originally. To simplify the notation, we will consider \( (1 - q) = p \), so that \( S_n|H_0 \sim Bin(n, p) \).

3.3. The rejection region

For the two-sided test the rejection region results on the union of two sets for the values of the \( S_n \) statistic. Taking a fixed significance level \( \alpha \), the rejection region of the test is \( S_n \leq c_{\alpha/2} \) or \( S_n \geq c_{\alpha/2}^* \). The two constants \( c_{\alpha/2} \) and \( c_{\alpha/2}^* \) are given by: \( c_{\alpha/2} = \frac{\log(\sum_{i=0}^{n} (\chi_i)^p(1-p)^{n-i})}{\log(\alpha/2)} \) and \( c_{\alpha/2}^* = \frac{\log(\sum_{i=0}^{n} (\chi_i)^p(1-p)^{n-i})}{\log(\alpha/2)} \), wherein given the formulas considered by Gibbons and Pratt (1975) and Kulinskaya (2008), which indicate alternative formulas for calculating the p-value. In this section, we presente the alternatives used in Gibbons and Pratt (1975) and Kulinskaya (2008).

4. Alternatives for p-value calculation

The purpose of this section is to present some alternative formulas for calculating the p-value of two-sided sign tests. As previously mentioned, in the two-sided hypothesis testing usually the p-value is defined as twice the value of the smaller p-value associated to the two one-sided tests. This duplication is particularly significant and informative only for continuous statistics. In contrast, when the distribution of the test statistic is discrete, as in the sign test, there is no consensus about the calculation of the p-value for two-sided tests. Although the literature that addresses this issue is not extensive, some authors have demonstrated concern with this issue. That was the case of Gibbons and Pratt (1975) and Kulinskaya (2008), which indicate alternative formulas for calculating the p-value. In this section, we present the alternatives used in Gibbons and Pratt (1975) and Kulinskaya (2008).

4.1. Method of Placing

One method that can be used to obtain the p-value is the method of placing (Plac), wherein given the observed value of the statistic, \( s_n \), selects an equal number of values in the two tails of the distribution. In the case of a binomial statistic, assume first that the observed number of successes, \( s_n \), is a value in the right tail of the distribution. The p-value is defined as the sum of the probabilities in values larger or equal to \( s_n \) or smaller or equal to \( (n - s_n) \), entailing that:

\[
p\text{-value}_{\text{Plac}} = P(S_n \geq s_n) + P(S \leq n - s_n).
\]

Similarly, it \( s_n \) is a value in the left tail of the distribution, then:

\[
p\text{-value}_{\text{Plac}} = P(S_n \leq s_n) + P(S_n \geq n - s_n).
\]

Example: Assuming a binomial statistic with \( n = 10 \) and \( p = 0.6 \), under \( H_0 \), and take up \( s_{10} = 3 \) for the observed value of the statistic. The probability function of \( S_{10}|H_0 \sim Bin(10, 0.6) \) is presented in Figure 1. Note that this is a unimodal and slightly right-skewed distribution. In Table 1 the probability function of \( S_{10} \) is presented. As the value \( s_{10} = 3 \) is the fourth largest value in the left tail of the distribution, the correspondent value in the right tail counting down from 10 is the value 7. Thus the p-value is given by:

\[
p\text{-value}_{\text{Plac}} = P(S_{10} \leq 3) + P(S_{10} \geq 7) = 0.055 + 0.382 = 0.437.
\]
Since \( P(S_{10} \geq 7) \) is about 7 times larger than \( P(S_{10} \leq 3) \), there is an increase of the probability regarding to \( \text{p-value}_{\text{usual}} = 2 \times P(S_{10} \leq 3) = 0.11 \) (approximately 4 times).

| \( s \) | \( P(S_{10} = s|H_0) \) |
|---|---|
| 0 | 0.0001 |
| 1 | 0.0016 |
| 2 | 0.0106 |
| 3 | 0.0425 |
| 4 | 0.1115 |
| 5 | 0.2007 |
| 6 | 0.2508 |
| 7 | 0.2150 |
| 8 | 0.1209 |
| 9 | 0.0403 |
| 10 | 0.0060 |

Table 1: Probability function of the \( \text{Bin}(10, 0.6) \) distribution.

4.2. Method Principle of Minimum Likelihood
With the principle of minimum likelihood method, \( p_{\text{ml}} \), the \( \text{p-value} \) is the sum of the probabilities in all values of \( S_n \) whose probability does not exceeds \( P(S_n = s_n) \). Therefore,

\[
p_{\text{p-value}_{\text{pl}}} = \sum_{i=0}^{n} P(S_n = i|H_0) \times 1_{\{P(S_n = i|H_0) \leq P(S_n = s_n|H_0)\}}.
\]

Example: Considering the previous example with \( S_{10}|H_0 \sim \text{Bin}(10, 0.6) \) and \( s_{10} = 3 \). The probability of the value 3, as shown in Table 1, is \( P(S_{10} = 3|H_0) = 0.0425 \). The set of points in the range of \( S_{10} \), whose probabilities that do not exceed 0.0425 are \( \{0, 1, 2, 3\} \cup \{9, 10\} \). Thus,

\[
\text{p-value}_{\text{pml}} = P(S_{10} \leq 3) + P(S_{10} \geq 9) = 0.055 + 0.046 = 0.101.
\]

4.3. Conditional method
The calculation formula of the conditional \( \text{p-value} \), proposed by Kulinskaya (2008), is given by:

\[
p_{\text{p-value}_{\text{cond}}} = \frac{P(S_n \leq s_n)}{P(S_n \leq \nu)} \times 1_{\{s_n < \nu\}} + 1_{\{s_n = \nu\}} + \frac{P(S_n \geq s_n)}{P(S_n \geq \nu)} \times 1_{\{s_n > \nu\}}. \tag{2}
\]

where \( s_n \) is the observed value of the statistic and \( \nu \) is a location measure, such as the mean, the median or the mode of \( S_n \).

Equation 2 can be rewritten as:

\[
p_{\text{p-value}_{\text{cond}}} = \min\left\{ \frac{P(S_n \leq s_n)}{P(S_n \leq \nu)}, \frac{P(S_n \geq s_n)}{P(S_n \geq \nu)} \right\},
\]

since both produce the same results. Took up for the parameter \( \nu \) the median of \( S_n \), denoted by \( w \), since the median is robust measure of location.

Example: Returning to the example of the statistic \( S_{10}|H_0 \sim \text{Bin}(10, 0.6) \) with median \( w = 6 \), since \( s_{10} = 3 < w \) entails that:

\[
p_{\text{p-value}_{\text{cond}}} = \frac{P(S_{10} \leq 3)}{P(S_{10} \leq 6)} = 0.0887.
\]

The \( \text{p-value} \) produced by this calculation formula is lower than \( \text{p-value}_{\text{pl}} \) and \( \text{p-value}_{\text{pml}} \). Note that considering the significance level \( \alpha \) fixed at 0.1, the \( \text{p-value}_{\text{cond}} \) would lead to the rejection of the hypothesis \( H_0 : \chi_{0.4} = \chi_0 \) contrary to the decision based on the \( \text{p-value}_{\text{pl}} \) and \( \text{p-value}_{\text{pml}} \). To evaluate the performance of this conditional formula, assume that the observed value of the statistic is \( s_{10} = 8 > w \). It follows that:

\[
p_{\text{p-value}_{\text{cond}}} = \frac{P(S_{10} \geq 8)}{P(S_{10} \geq 6)} = 0.2642.
\]

If the observed value of the statistic equals the median of \( S_{10} \), \( s_{10} = 6 = w \), then:

\[
p_{\text{p-value}_{\text{cond}}} = 1.
\]

Few observations can be taken when this method is applied to the sign test. In the first analysis, the mean and median of the binomial distribution can take values that do not belong to the set of natural numbers, \( N \), and in Kulinskaya (2008) there is no reference to this situation. When this occurs, the wisest course is to consider the integer part of the location parameter. Another observation can...
be made about the statistic $S_n|H_0 \sim Bin(n, \frac{1}{2})$ when $n$ is even. In this case since $P(S_n \leq \nu) = P(S_n \geq \nu) = \frac{1}{2}$, the p-value$_{\text{cond}}$ equation equals to p-value$_{\text{usual}}$:
\[
p-value_{\text{cond}} = \begin{cases} \min\{P(S_n \leq s_n), P(S_n \geq s_n)\} & \text{if } s_n \text{ is even} \\ 2 \times \min\{P(S_n \leq s_n), P(S_n \geq s_n)\} & \text{if } s_n \text{ is odd} \end{cases}
\]
However, for $S_n|H_0 \sim Bin(n, \frac{1}{2})$ with $n$ even, both $P(S_n \leq \nu)$ and $P(S_n \geq \nu)$ are not equal to 0.5. This occurs because $S_n$ is a discrete statistic and some probabilities are not attainable. Therefore, when $n$ is even the cond method is not equal to the usual method.

4.4. New Proposal - Distance Method

Bearing in mind the studied methods, we will develop a new method (dist) to compute the p-value that is consistent with the hypothesis $H_0$ and that does not result in p-values greater than one. Consider the test statistic $S_n|H_0 \sim Bin(n, p)$ of the sign test and $w = \chi_4$ the median of $S_n$ as the central measure of location. Suppose that the observed value of statistic in a two-sided test to the $q$-th quantile of $X$ is $s_n$ and $d$ let denote the euclidean distance between $s_n$ and $w$, $d = |w - s_n|$. The p-value$_{\text{dist}}$ is given by the sum of probabilities of the points whose distance to $w$ is larger or equal than $d$. This procedure is especially intuitive when seeking to interpret the p-value as the degree of agreement or disagreement between the observed value of the statistic and the median of $S_n$.

Example: Return to $S_{10}|H_0 \sim Bin(10, 0.6)$ with $w = 6$ and $s_n = 3$, so that $d = |w - s_n| = 3$. In Figure 2, we represent in blue the probability value in points whose distance to the median $w$ is equal or larger than $d$ and in dashed (green) the probability value in the other points. The p-value$_{\text{dist}}$ is given by the sum of the probabilities presented in blue:
\[
\begin{align*}
p-value_{\text{dist}} &= P(S_n \leq 3) + P(S_n \geq 9) \\
&= 0.101.
\end{align*}
\]

The p-value with the method dist can be computed as:
\[
\begin{align*}
p-value_{\text{dist}} &= 1 - \left\{ \sum_{i=s_n+1}^{w+d+1} P(S_n = i) \times I_{\{s_n<w\}} + \sum_{i=w-d+1}^{s_n-1} P(S_n = i) \times I_{\{s_n>w\}} \right\},
\end{align*}
\]

The dist method is very similar to the pml method, therefore sometimes both methods lead to the same results.

4.5. Comparison and analysis of the p-value methods

The aim is compare the p-values obtained by the five methods, usual, plac, pml, cond and dist, when the distribution of the statistic $S_n$ given $H_0$ is symmetric or asymmetric. To carry out this study, we developed computational implementations, with software R, of the p-value methods. The analysis consists in obtaining the p-value with the different methods of the two-sided test: $H_0 : \chi_q = \chi_0$ vs $H_1 : \chi_q \neq \chi_0$, and $q = 0.1, 0.25, 0.5, 0.75$ and 0.9. The goal is to verify which methods, usual or plac, the p-value is greater than one and in which cases the decision of the test can change. In Tables 2 and 3 we report the p-values obtained by the five methods for samples with size $n = 6$ and $n = 7$, respectively. The results of the study can be summarized as follows:

1. plac method
   - When the distribution of $S_n$ is quite asymmetric, the p-value$_{\text{plac}}$ can lead to false decisions if $s_n$ is unlikely to occur.
   - If $S_n$ asymmetric and $s_n = w$ then the p-value is smaller than one.
   - When $n$ is an even number and $s_n = \lfloor \frac{w}{2} \rfloor$, then p-value$_{\text{plac}} > 1$.
   - If $n$ is an odd number and the observed value of the statistic is $s_n = \lfloor \frac{w}{2} \rfloor$ or $s_n = \lceil \frac{w}{2} \rceil$, then the p-value is equal to one.
   - This method produces acceptable results when the statistic $S_n$ has symmetric or slightly asymmetric distribution.

2. pml method
   - In samples with even or odd size and binomial statistics with symmetric distributions p-value$_{\text{pml}} = $ p-value$_{\text{usual}}$. If $n$ is

Figure 2: Probability function of $S_{10} \sim Bin(10, 0.6)$ illustrative of the distance method.
If $S_n | H_0$ has an asymmetric distribution, on the right (left), where every point $i$ that satisfies $P(S_n = i) \leq P(S_n = s_n)$ is located at the right (left) of $s_n$ it is verified that $p_{\text{value}}_{\text{pml}} = \frac{p_{\text{value}}_{\text{usual}}}{2}$.

When $s_n$ is the mode of $S_n$ then $p_{\text{value}}_{\text{pml}} = 1$.

3. Cond method

If $S_n$ is symmetric and $n$ is an even number, the $p$-value returned by the cond method is never equal to $p_{\text{value}}_{\text{usual}}$. This is a consequence of $P(S_n \geq w)$ and $P(S_n \leq w)$ never being equal to 0.5.

In general this method returns coherent $p$-values. However the associated values are relatively lower than $p_{\text{value}}_{\text{usual}}$.

4. Dist method

The dist method sometimes produces similar results to the pml method. However, it is noted that when the test statistic has a symmetric distribution and $n$ is an odd number $p_{\text{value}}_{\text{dist}} \neq p_{\text{value}}_{\text{pml}}$, except when $s_n$ corresponds to the limits of the median class.

If the distribution of $S_n$ is asymmetric with $q < 0.5$ whenever $s_n < w$ and $P(S_n = s_n) < P(S_n = w + d)$ or $s_n > w$ and $P(S_n = s_n) > P(S_n = w - d)$, then $p_{\text{value}}_{\text{dist}} \neq p_{\text{value}}_{\text{pml}}$. The same happens with $q < 0.5$, whenever $s_n > w$ and $P(S_n = s_n) < P(S_n = w - d)$ or $s_n < w$ and $P(S_n = s_n) > P(S_n = w + d)$.

Whenever $p_{\text{value}}_{\text{dist}} \neq p_{\text{value}}_{\text{pml}}$ it also happens that $p_{\text{value}}_{\text{dist}}$ is different from the $p$-value returned by the other methods.

The dist method returns coherent $p$-values that are never greater than one.

5. Simulation Study

We will conduct a simulation study in order to evaluate the performance of the previously presented methods to compute the $p$-value and power calculations of the sign test:

$H_0 : \chi_q = \chi_0 \text{ versus } H_1 : \chi_q \neq \chi_0$

with $q = 0.1, 0.25, 0.5, 0.75$ and 0.9. To perform that, we simulated $N = 10000$ samples with $n$ data from symmetric or asymmetric distributions. The $n$ data sets were generated, in the R program, as originating from reduced normal distribution and Gama(3,0.1). Furthermore, we also used generated

$$\begin{array}{|c|c|c|c|c|c|}
\hline
n & S & CHI2 & CHI2 & \text{USUAL} & \text{USUAL} \\
\hline
1 & 0.930 & 0.538 & 0.538 & 0.538 & 0.538 \\
2 & 0.900 & 0.885 & 0.885 & 0.885 & 0.885 \\
3 & 0.861 & 1.016 & 1.016 & 1.016 & 1.016 \\
4 & 0.822 & 0.954 & 0.954 & 0.954 & 0.954 \\
5 & 0.783 & 0.854 & 0.854 & 0.854 & 0.854 \\
\hline
\end{array}$$

Table 3: P-values of the test statistic $S_n$ with $S_n | H_0 \sim Bin(\chi, p)$, for the test $H_0 : \chi_q = \chi_0 \text{ versus } H_1 : \chi_q \neq \chi_0$ obtained by the usual, plac, pml, cond and dist methods.

Table 2: P-values of the test statistic $S_n$, with $S_n | H_0 \sim Bin(\chi, p)$, for the test $H_0 : \chi_q = \chi_0 \text{ versus } H_1 : \chi_q \neq \chi_0$ obtained by the usual, plac, pml, cond and dist methods.
data from a neighborhood of the $F = N(0,1)$ distribution, with $0 < \varepsilon < 1$, as:
\[
F_\varepsilon(F) = \{ G : G = (1 - \varepsilon)F + \varepsilon H \},
\]
and $H = N(2, 4)$. The sample size $n$ was taken as $n = 6, 7, 10, 11, 20, 21, 30$ and $31$. The $q$-th quantiles, $\chi_q$, were obtained in R and assigned to $\chi_0$.

For each method we recorded the relative frequency of $p$-value $\leq \alpha$, with $\alpha = 0.01, 0.05$ and $0.1$, on the 10000 generated samples. Furthermore, we have recorded also the frequency of $p$-value $> 1$ for the usual and the plac methods.

The aim of this study was check whether there are significant differences among the methods. To assess the power test, another simulation was performed by establishing in $H_0$ contiguous hypotheses to $H_0$. Here we considered the test $H_0 : \chi_q = \chi_0$ versus $H_1 : \chi_q = \chi_0 + \Delta_i$, where $\chi_0$ refers to the $q$-th quantile of $X \sim N(0,1)$. Now, under $H_1$, the samples were generated from $F_1 = N(\Delta, 1)$ when $\Delta > 0$, or from $F_2 = N(\Delta, 1)$ case $\Delta < 0$. For each sample, we evaluated $\gamma(\Delta) = P(\text{reject } H_0 | H_1)$ with the five methods. The study of the sign test power with data sets from asymmetric distributions was not performed because in this case it is not easy to change the $q$-th quantile without changing the distribution variance. The purpose of this simulation was to assess the methods that conduct to small values of the probability of type II error.

### 5.1. Simulation of the p-value

- Generation of $N = 10000$ data sets with dimension $n = 6, 7, 10, 11, 20, 21, 30$ and $31$ from the distribution $F = N(0,1)$.

In tables 4 and 5 we reported the results obtained to the two-sided tests for the median of $X \sim N(0,1)$. For each value of $n$ and each method, the number of samples (from 0 to 10000) that checked the $p$-value conditions listed in the first column of these tables are presented.

### 5.2. Power Simulation

As it is well known, a statistical test with low power to detect that the sample is consistent with $H_1$ causes indecision about the validity of $H_0$. In order to compare the power value of the sign test produced by the methods usual, plac, pm1, cond and dist, we considered the two-sided test:

$H_0 : \chi_q = \chi_0$ vs $H_1 : \chi_q = \chi_0 + \Delta_i,$

where $\chi_0$ indicates the $q$-quantile of $X \sim N(0,1)$ and $\Delta_i = \pm 0.25, \pm 0.5, \pm 1.5, \pm 2.5$ and $\pm 3$. Under $H_1$ we generated:

- $N = 10000$ samples with dimension $n = 6, 7, 10, 11, 20, 21, 30$ and $31$ from $F_1 = N(\Delta_i, 1)$, $\Delta_i > 0$, or from $F_2 = N(\Delta_i, 1)$, $\Delta_i < 0$.

With the five methods, we calculated $\gamma(\Delta_i) = P(\text{reject } H_0 | H_1)$. Under $H_1$, as the the data comes from $F_1$ or $F_2$, it follows that $S_{1n} \sim Bin(n, p_1^*\gamma)$ with $p_1^* = P(X > \chi_0 | F_1^*)$, and $S_{2n} \sim Bin(n, p_2^*)$ with $p_2^* = P(X > \chi_0 | F_2^*)$. Note that the power functions of these tests, except for the median test of $X$, does not exhibit any symmetry around the value $\chi_0$. We evaluated the average powers obtained in the $N = 10000$ simulated samples. For the usual and plac methods the samples which gives rise to $\gamma(\Delta_i) > 1$ were removed, since these two methods can return a probability larger than one. Figures 3 and 4 present the average power obtained for the median test of $F$.

### 6. Interpretation of simulation results

Then, we present the results obtained in the simulation of the p-value and power simulation.
6.1. Summary results for the p-value simulation study

- In general, the sample size \( n \) influences the decision of the sign test. Different decisions are obtained for each value of \( n \).

- The p-value results for the \( q \)-th quantile tests were not influenced by the shape, symmetric or asymmetric, behaviour of the distribution of \( X \). The cardinal samples that verify each of the conditions about the p-value obtained in the \( q \)-th quantile test were similar, which leads us to conclude that the results are independent of the asymmetry of the distribution of \( X \).

- Regarding the conditions, p-value \( \leq 0.01 \), p-value \( \leq 0.05 \) and p-value \( \leq 0.1 \), as larger is the sample size \( n \) the larger is the number of samples that verify such conditions. The opposite happens with the p-value > 1 condition.

- It is notorious that p-value\textsubscript{usual} > 1 occurs more often than p-value\textsubscript{PML} > 1.

- For the median test, when \( n \) is odd, the number of samples that verify the p-value > 1 condition is negligible on the 10000 simulated samples.

- When \( H_0 \) establishes that the quantiles of \( F \) are equal to the quantiles of \( F = N(0, 1) \), it was observed that the results of the test to quantiles \( q = 0.75 \) and \( q = 0.9 \) are obviously the most affected. The contamination of samples with observations generated from \( N(2, 4) \) does not affect the smaller quantiles of \( F \). The introduction of noise in the samples can decrease the sign test performance.

6.2. Summary results of the power simulation study

- In general, the usual method proved to be a poor method, with smaller average power than the other methods.

- As larger is the difference between the expected value of the distributions stipulated under \( H_0 \) and \( H_1 \), the greater will be the power of the test. This increase of \( \gamma \) was observed in all methods, except for the usual method which falls steeply (\( \gamma \approx 0 \)) with the increase of \( |\Delta_i| \), showing the instability of the method. One reason for this to occur is the number of samples that are removed of the power study because verified \( P(\text{reject} H_0 | H_1) > 1 \).

It was found that as larger is the fixed value for \( |\Delta| \), greater is the number of removed samples from the study. The samples are generated according to the distribution of \( H_1 \) and the number of observations in the sample of \( X \) above \( \chi_0 \) increases when \( \Delta > 0 \) and decreases when \( \Delta < 0 \).

As a consequence, the observed value of the test statistic \( s_n \) coincides more times with the median of \( S_{1n}^1 \) or of \( S_{2n}^2 \), respectively, when \( \Delta > 0 \) or \( \Delta < 0 \).

The usual method returns p-values greater than one when \( s_n = w \), therefore the number
of samples that satisfy $\gamma > 1$ increases with increasing $|\Delta|$.

- Good results were recorded with dist method, in general, for small values of $|\Delta|$, proving dist to be the method with the greatest average power. Furthermore, in general, the increase of $\Delta$ lead to greater or equal powers than in the other methods.

- For small values of $|\Delta|$, the average of $\gamma$ belongs to $[0.55, 0.7]$. When $\Delta$ increase, the average of $\gamma$ also increases, getting very close to one, with the exception of the usual method.

- When the $q$–th quantile increases, we observe a slight decrease in the average of $\gamma$ when $\Delta > 0$ and a slight increase where $\Delta < 0$.

7. Conclusions

After all the extensive analysis of the five p-value calculation methods, usual, plac, pml, cond and dist, the main features of each were appointed. The usual and plac methods are the only ones that return p-values greater than one. In the case of the usual method, p-value_{usual} > 1 when $s_n = w$. Whenever $S_n$ has a median class, the p-value_{usual} = 1 when $s_n$ equals one of the limits of this class. In the plac method, the p-value is greater than one when $n$ is even and $s_n = \frac{n}{2}$. Another note about this method is that it only returns acceptable p-values when $S_n$ is a symmetric or a weakly asymmetric distribution. Regarding the methods pml, cond and dist, these never produce p-values greater than one. The p-value_{pml} = p-value_{usual} when the test statistic is symmetric. When $S_n$ is asymmetric, there are situations where p-value_{pml} is equal to one half of p-value_{usual}. When the observed value of the statistic corresponds to the mode of $S_n$, p-value_{pml} = 1. Comparing the methods cond and usual, it can be observed that, in general, p-value_{cond} < p-value_{usual}. When the value of $s_n$ corresponds to the median of $S_n$ then p-value_{cond} = 1. The recently introduced method, dist method, usually produces the same results that are obtained by the pml method. The dist method showed a very positive performance, producing coherent results. Based on the comparison study of the p-value calculation methods and simulation study, setting $\alpha \times 100\%$ in the usual significance levels (1%, 5% and 10%), we observe that:

- When the distribution of $S_n$ is symmetric, the cond and dist methods may return p-values relatively lower than other methods. Thus, for the usual significance levels, the cond and dist methods reject more times the hypothesis $H_0$.

- Case the distribution of $S_n$ is asymmetric, all methods can change the decision of the sign test when selected instead of usual. The usual method rejects the hypothesis $H_0$ less often.

- If the $S_n$ distribution is fairly asymmetric, the application of the usual, dist and pml methods can lead to changes on the decision of the test.

7.1. Achievements

From the alternative methods studied, plac, pml, cond and dist, it is not possible to support in favor of a particular method, since all of them have their on limitations. However, the plac method proved to be an inconsistent method and therefore is not recommended.

The results obtained by the dist method at power simulation of the $q$–th quantile test for the standard normal were quite satisfactory. Typically, this method produces a greater average power than other methods.

It was possible to conclude that, apart from some exceptional cases, low powers are obtained with the usual method, enhancing the inconsistency of the method. In short, the study in this paper gives to evidence the weaknesses of the usual method. Although the usual method is the most used to perform statistical tests, it did not show to be appropriate to determine the p-value in the two-sided test. In contrast, the dist method proved to be a good option for the p-value calculation.

7.2. Future work

As proposal for future work we suggest to address the justification for the choice of the euclidean distance, $d = |s_n - w|$, used to calculate the p-value by the dist method, or an alternative metric that optimizes the method. The realization of an equal study for the Wilcoxon test or another test that has a test statistic with discrete or continuous asymmetric distribution, is another possible topic for future work. This is a challenge, since the test statistic of the Wilcoxon test is not as simple as the binomial distribution of the sign test. Furthermore, the p-value is greater than one when $s_n = w$ and therefore the determination of the median of $S_n$ for the Wilcoxon test is also an hurdle to overcome.

Acknowledgements

I warmly thank to Prof. Isabel Rodrigues for her support, availability and dedication. It was a pleasure working with her. I also want to thank my family and friends for believing in me and giving me strength in every time I did not believe in myself. I got it!

References

Fisher, R. A. (1920). A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error and by the Mean Square Error. Monthly No-


