On Zeroes in Sign and Signed Rank Tests

Rajarshi Dey
University of South Alabma United States
Department of Mathematics and Statistics
rajarshidey@southalabama.edu

Justin Manjourides
Northeastern University United States
Department of Health Sciences
jmanjourides@northeastern.edu

Ronald H Randles
University of Florida United States
Department of Statistics
rrandles@stat.ufl.edu

Abstract

When zeroes (or ties within pairs) occur in data being analyzed with a sign test or a signed rank test, nonparametric methods textbooks and software consistently recommend that the zeroes be deleted and the data analyzed as though zeroes did not exist. This advice is not consistent with the objectives of the majority of applications. In most settings a better approach would be to view the tests as testing hypotheses about a population median. There are relatively simple p-values available that are consistent with this viewpoint of the tests. These methods produce tests with good properties for testing a different (often more appropriate) set of hypotheses than those addressed by tests that delete the zeroes.

Keywords: Sign test; Signed rank test; Paired t-test; Meaningful zeroes; nonparametric

Introduction: The sign test

The sign test has a lengthy history in statistics, including its early application by Arbuthnot (1710) in eighteenth century and its formal description by Dixon and Mood (1946). Throughout, there has been substantial controversy (Randles 2001) about the role and use of zero (neutral) responses.

Example 1: A study was conducted which had as one of its objectives to determine whether taking dichloroacetate (DCA) affects the hearts of patients. Since DCA is typically administered to correct energy metabolism disorders, effects on heart rate, either increases or decreases, could be viewed as a undesirable side-effect. Measurements of heart rate both before and 30 minutes after administration of DCA are displayed in Table 1. We focus our attention on the difference (after-before) column in this table. There are 15 positive differences, 3 negative differences and 2 zeroes. When conducting the sign test on this data, what roles should be paid by the 2 zero observations?

To model problems of this type, let $\{X_1, X_2, \ldots, X_n\}$ be independent observations with

$p_+ = P(X_i > 0)$, $p_0 = P(X_i = 0)$ and $p_- = P(X_i < 0)$, for every i. Let $N_+(N_0$ and $N_-)$ denote the random number of positive (zero and negative) values within a sample of size n, respectively. The vector (N_+, N_0, N_-) thus has a multinomial distribution with parameters (n, p_+, p_0, p_-). The sign test is often described as a test of

$$H_0: p_+ = p_- \text{ vs } H_a: p_+ \neq p_- \qquad (1)$$

or the one-sided alternative versions. The usual sign test, reported in most textbooks recommends deleting zeroes and reporting a

$$p - value = 2P[B \geq max(n_+, n_-) | B \sim bin(n - n_0, 0.5)] \qquad (2)$$

where $n_+(n_0$ and $n_-)$ are the observed values of $N_+(N_0$ and $N_-)$ and $bin(n, p)$ denotes a binomial distribution with parameters n and p. For the data displayed in Table 1, we get the p-value calculated as

$$p - value = 2P\big(B \geq 15 \big| B \sim bin(18{,}0.5)\big) = 0.0075$$

This conditional sign test's p-value is commonly used in practices (deleting zeroes). It reports the same p-value whether there are only a few or many zeroes observed. For example, if this data was obtained from 30 patients instead of 20 but we had observed 12 zeroes, we would report the same p-value. Hollander et al (2014) and Siegel and Castellan (1988) recommend this, among others. This approach is also implemented commonly in statistical software like SAS.

Coakley and Heise (1996) studied a large number of the methods from the literature for handling zeroes in the sign test. When testing the two-sided hypothesis in (1), they recommended use of the

$$p - value = 2\Phi\left(\frac{\min(n_+, n_-) - \max(n_+, n_-)}{n_+ + n_-}\right).$$

Note that, this p-value also only depends on n_+ and n_-, ignoring the number of zeroes.

Statisticians with practical experience often claim that zeroes, which represent "no change in condition", are meaningful and important responses that should not be discarded. They argue that in most, but not all, settings, the zeroes should lend credence to the null hypothesis.

Some authors have used the zeroes to improve the power of the sign test as a test of (1). See for example, Starks (1979), Suissa and Shuster (1991), and Presnell (1996). The tests discussed in these papers have the property that with n_+ and n_- fixed, the p-values generally decrease as n_0 increases. Thus, zeroes add credence to the alternative when using these methods.

The purpose of this article is to recommend that the sign and signed rank tests be viewed as tests about population medians when handling observed zeroes. This would be consistent with the point estimates and confidence intervals that correspond to these tests, since they estimate population medians. It will also ensure that any zeroes would be viewed as meaningful and lending credence to the null hypothesis. In the majority of problem settings, this is the more appropriate viewpoint toward zeroes. Moreover, as this article presents, there are simple, practical ways to find p-values for the tests corresponding to this viewpoint.

The Median Sign Test

Consider the multinomial model as in Figure 1. Let M denote the population median, and consider the one-sided test of the hypothesis

$$H_0\text{: } M = 0 \text{ vs } H_a\text{: } M > 0 \quad . \tag{3}$$

Clearly $p_0$plays a role in defining the population median. If $p_0 = 0$, the hypothesis described above are the same as testing

$$H_0\text{: } p_+ = p_- \text{ vs } H_0\text{: } p_+ > p_-. \tag{4}$$

If, however, $p_0 > 0$, then (3) is testing

$$H_0\text{: } p_+ \leq 0.5 \text{ vs } H_0\text{: } p_+ > 0.5. \tag{5}$$

Therefore, as atest of (3), the p-value is:

$$p - value = P[B \geq n_+ | B \sim bin(n, 0.5)]. \tag{6}$$

Here the zeroes are combined with the negatives and both types are considered “failures” in the binomial setting. Using this p-value has sometimes been described as the ultra-conservative approach to handling zeroes in the sign test. But, it is actually a very appropriate and powerful test of (3), which is a distinctly different objective from (4), the problem addressed by the usual (delete zeroes) sign test. The sign test is often described as a test about the population median. See, for example, Hollander et al (2014), page 90. Yet, when it comes to handling zeroes, this objective is usually abandoned.

The two-sided alternative test

$$H_0\text{: } M = 0 \text{ vs } \; H_a\text{: } M \neq 0 \tag{7}$$

is more of a challenge. It tests

$$H_0\text{: } \max(p_+, p_-) \leq 0.5 \text{ vs } H_a\text{: } \max(p_+, p_-) > 0.5. \tag{8}$$

Fong et al (2003) identified this as an interesting problem. They noted that doubling the smallest tail probability, i.e.,

$$p - value = 2P[B \geq \max(n_+, n_-) | B \sim bin(n, 0.5)] \tag{9}$$

leads to p-values that are much too large and, in fact, may exceed 1. They proposed the following method of finding a p-value:

$$p - value_{FKLL} = \frac{P[B \geq \max(n_+, n_-) | B \sim bin(n, 0.5)]}{P[B \geq [|(\mathrm{n}+1-n_0)/2|] | B \sim bin(n, 0.5)]}, \tag{10}$$

where [|.|] is the greatest integer function. This is very simple and easy to implement. It only requires $Bin(n, 0.5)$ tables. The denominator in (10) is the maximum value possible for the numerator. Thus the p-value in (10) is always less than or equal to one.

We propose that the two-sided test be based on $n_* = \max(n_+, n_-)$, given the value of n_0. If $p_0$was known, we could construct a p-value at the boundary of the null via

$$p - value(p_0) = P[N_* \geq n_* | (N_+, N_0, N_-) \sim Multinomial(n, p_{+*}, p_{0*}, p_{-*})], \tag{11}$$

where $p_{+*} = 0.5$, $p_{0*} = \min(p_0, 0.5)$ and $p_{-*} = 1 - p_{+*} - p_{0*}$. In practice, we would use $p - value(\widehat{p_0})$, where $\widehat{p_0} = \min(0.5, n_0/n)$. This can be viewed as a plug-in bootstrap p-value, where we have estimated the unknown p_0. It is more complex than (10), but is also less conservative. With modern computing packages and languages, the proposed p-value can be found easily via

$$p - value(\widehat{p_0}) = \sum_{k=0}^{n} Q(k, n_*, n, \widehat{p_0}) b(k|n, \widehat{p_0}), \tag{12}$$

with

$$Q(k, n_*, n, \widehat{p_0}) = \begin{cases} 1 & if\; n_* \leq (n-k)/2 \\ 1 - B_*(n_* - 1) + B_*(n - k - n_*) & if\; \frac{n-k}{2} < n_* \leq n - k \\ 0 & otherwise \end{cases}$$

where $b(k|n, p)$ is the binomial probability function and $B_*(t)$ is the distribution function of a binomial $(n - k, p_c)$ random variable with $p_c = \left(2(1 - \widehat{p_0})\right)^{-1}$. The p-value in (12) has some nice properties. It is equal to the usual two-sided binomial p-value when there are no zeroes. It has a natural relationship to the one-sided p-value in (6) and it is relatively easy to compute.

To illustrate the influence of zeroes on the p-vaues, table 2 uses the DCA data with $n_+ = 15$ and $n_- = 3$ fixed, but varying the number of zeroes. The median sign tests are testing hypothesis (7) instead of hypothesis (1) which are tested by usual sign test (delete zeroes). The median sign tests have p-values which increase as the number of zeroes increases. The proposed p-value does not differ substantially from (10), described by Fong, Kwan, Lam and Lam, but the p-values are generally somewhat smaller.

Power Functions

The power of the two-sided tests based on the p-values described earlier: the usual sign test (2), the Fong, Kwan, Lam and Lam (10) and the proposed (12), were compared for a fixed probability $p_0 = 0.1, 0.2, 0.4$; varying $delta = p_+ - p_-$.Note that, for the median sign test, the boundary of the null hypothesis occurs when $delta = p_0$. Graphs of the actual power curves (enumerated, not simulated) are shown in Figure 2 for different sample sizes with $p_0 = 0.2$ and $\alpha = 0.05$. This graph show that the proposed p-value controls the levels and improves the power of the median sign test for smaller sample sizes. The proposed p-values (12) are typically smaller than those in (10). But for a fixed α value the tests may have the same rejection region, because of the discrete nature of the tests. The cases pictured are ones in which the rejection regions differ. They show that the proposed method can improve the power of the test because of the smaller p-values.

Table 1: Heart Rates of Patients Before and After DCA

Patient	**Baseline**	$30\,minutes$	$Difference$
1	73	73	0
2	62	76	14
3	67	74	7

4	100	105	5
5	97	105	8
6	84	100	16
7	51	53	2
8	56	56	0
9	83	87	4
10	72	77	5
11	82	96	14
12	79	82	3
13	74	68	−6
14	73	76	3
15	81	87	6
16	59	72	13
17	104	103	−1
18	96	76	−20
19	62	67	5
20	74	81	7

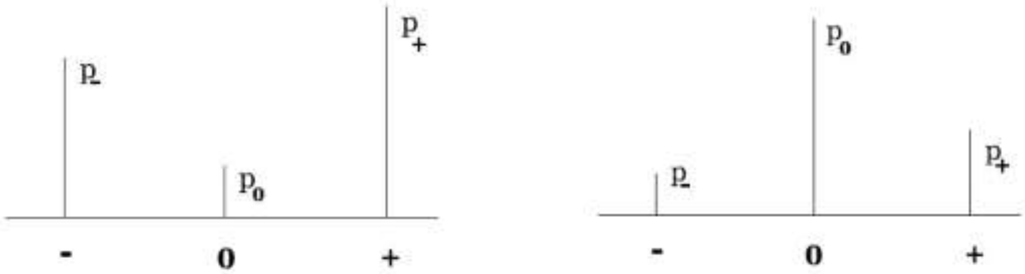

Figure 1: Multinomial models

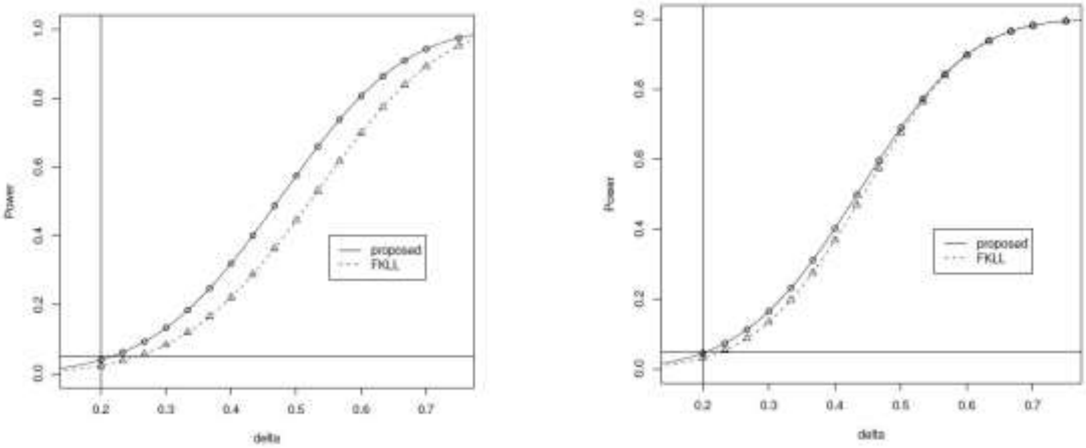

Figure 2: Power functions for $p_0 = 0.2, n = 37$ (left), 51(right)

Signed Rank Tests

In the signed rank test, we assume $\{X_1, X_2, \dots, X_n\}$ are independent and identically distributed with distribution function $F(.)$. Under the null hypothesis, the distribution

$F(.)$ is symmetrically distributed around 0. The signed rank test is related to the parameters:

$$p_{+avg} = P\left[\frac{X_1+X_2}{2} > 0\right] \text{ and } p_{-avg} = P\left[\frac{X_1+X_2}{2} < 0\right].$$

Underthe null hypothesis $\frac{X_1+X_2}{2}$ is symmetrically distributed around 0 and hence $p_{+avg} = p_{-avg}$. Wilcoxon (1945) and Wilcoxon and Wilcox (1964) suggested ranking the absolute values of all non-zero observations from $1,2,\ldots,n-n_0$ and then forming

$$w_+ = (sum\ of\ absolute\ vaue\ ranks\ of\ all\ x_i > 0)$$

and

$$w_- = (sum\ of\ absolute\ vaue\ ranks\ of\ all\ x_i < 0)$$

The test designed to detect H_a: $p_{+avg} > p_{-avg}$, uses

$p - value = P[W_+ \geq w_+ |\ W_+ \text{under} 2^{n-n_0} \text{sign sets assigned to } 1,2,,\ldots,n-n_0]$. (13)

The two-sided tests designed to detect H_a: $p_{+avg} \neq p_{-avg}$, uses $w_* = \max(w_+, w_-)$ and

$p - value = 2P[W_+ \geq w_* |\ W_+ \text{under} 2^{n-n_0} \text{sign sets assigned to } 1,2,,\ldots,n-n_0]$. (14)

The null distribution of the test statistic is determined by independently assigning equally likely signes $(+1\ or -1)$ to each of the ranks $1,2,\ldots,n-n_0$. Wilcoxon's suggestion deletes all zero responses from the analysis regardless of the number or the interpretation of the observed zeroes.

Pratt (1959) noticed a peculiar property resulting from the way Wilcoxon proposed handling zeroes. He showed that in a one-sided test using Wilcoxon's p-value in (13), it is possible to shift the data in a positive direction on the number line and actually get a larger p-value than before they shifted. To avoid this counter-intuitive property, Pratt proposed ranking the absolute values of all non-zero observations from $n_0+1, n_0+2,\ldots,n$ and forming

$$v_+ = (\text{sum of absolute vaue ranks of all} x_i > 0)$$

and

$$v_- = (\text{sum of absolute vaue ranks of all} x_i < 0).$$

For the test designed to detect H_a: $p_{+avg} > p_{-avg}$, Pratt suggested using the

$$p - value = P[V_+ \geq v_+ |\ V_+ \text{under} 2^{n-n_0} \text{sign sets assigned to} n_0+1,\ldots,n]. \quad (15)$$

The two-sided tests designed to detect H_a: $p_{+avg} \neq p_{-avg}$, uses $v_* = \max(v_+, v_-)$ and

$$p - value = 2P[V_+ \geq v_* |\ V_+ \text{under} 2^{n-n_0} \text{sign sets assigned to } n_0+1,\ldots,n]. \quad (16)$$

Pratt's p-values are often close to the same values as Wilcoxon's, but not always. Pratt's method does explicitly use the number of zeroes $n_0$in the assignment of ranks. Derrick and White (2017) explains why this method is more robust for data on an ordinal scale.

Median Signed Rank Test

Dey (2018) argues why the Wilcoxon's and Mann-Whitney-Wilcoxon Tests should be about medians. Let us consider a median signed rank test. Let M_{avg} denote the median of the distribution of $\frac{X_1+X_2}{2}$ when $X_1, X_2$are $\text{i.i.d.} F(.)$. If $F(.)$ is a symmetric distribution, then, $M_{avg} = M$, the population median. Signed rank tests are concerned with the

parameter M_{avg}. In particular, the point estimator and the confidence interval corresponding this test estimate M_{avg}. Clearly the zeroes should play a role in estimating $M_{avg}$and likewise should play a fundamental role in a test about M_{avg}.
To test

$$H_0\text{: } F(.)\text{is symmetric around 0 vs. } H_a\text{: } M_{avg} > 0 \tag{17}$$

use Pratt ranks and form v_+. The test is based on:

$$p - value = P[W_+ \geq v_+ |\ W_+\text{under}2^n\text{sign sets assigned to } 1,2,,\dots,n]. \tag{18}$$

This assigns the same p-value as would be found by the signed rank test, if the observed zeroes were actually negative numbers that were very close to zero. This approach is ths analogous to (6), because the zeroes are counted as evidence against the alternative.
A two-sided test of

$$H_0\text{: } F(.)\text{is symmetric around 0 } vs. H_a\text{: } M_{avg} \neq 0, \tag{19}$$

could be based on $v_* = \max(v_+, v_-)$ and use the

$$p - value = \frac{P[W_+ \geq v_*]}{P\left[W_+ \geq \left\lfloor \left| \frac{n(n+1)}{4} - \frac{n_0(n_0+1)}{4} + 0.5 \right| \right\rfloor \right]}, \tag{20}$$

where again W_+corresponds to use of all 2^nequally likely sign sets attached to $1,2,\dots,n$. This p-value is the analogue to the one proposed by Fong et al (2003) for the sign test. The denominator is the largest possible for the numerator. It is computationally very easy as it only uses tables of distribution of W_+ under null hypothesis.

Table 3 displays the p-values of the signed rank tests for the DCA data, varying the number of zeroes. For this particular data set, there is an initial decrease in the p-value as number of zeroes increase, but it eventually increases sharply as number of zeroes keep increasing. Pratt's p-values decrease as number of zeroes increase for this data.

Power Simulation

The power of the two-sided test p-values were simulated when handling zeroes in the manner suggested by Wilcoxon (14), by Pratt (16) and by the proposed method (20). The model used a fixed value of p_0 on 0 and probability $(1 - p_0)$ spread over a continuous distribution that is symmetric around a location parameter θ. As θ increases from 0, the value of p_{+avg}also increases, eventually exceeding 0.5. The powers were simulated using a normal distribution and Cauchy distribution for the continuous part of the distribution. The normal distribution results are displayed in Figure 3 and the Cauchy distribution results are displayed in Figure 4. Each set useus $p_0 = 0.2$ and 0.4. When $p_0 = 0$, all three methods are equivalent. The sample sizes we considered were $n = 30{,}50{,}70$ with $\alpha = 0.05$ and the data setswere simulated $10{,}000$ times. In figures, we only show the simulations with $n = 70$ as the graphs were all similar. The powers are graphed as a function of p_{+avg}(denoted p+ave). For the median signed rank test, the boundary of the null hypothesis occurs at $p_{+avg} = 0.5$. The Wilcoxon and Pratt methods test the alternative $p_{+avg} = p_{-avg}$, whereas the median signed rank test, uses the alternative $\max(p_{+avg}, p_{-avg}) > 0.5$. The performance of the tests reflects this difference in objectives.

Table 2: Sign test p-values for the DCA data varying $n_0, n_+ = 15, n_- = 3$

n_0	One-tailed sign test	One-tailed median	Two-tailed	Small tail doubled	FKLL modified	Proposed
0	0.0038	0.0038	0.0075	0.0075	0.0064	0.0075
1	0.0038	0.0096	0.0075	0.0192	0.0142	0.0121
2	0.0038	0.0207	0.0075	0.0414	0.0277	0.0223
3	0.0038	0.0392	0.0075	0.0784	0.0485	0.0401
4	0.0038	0.0669	0.0075	0.1338	0.0781	0.0674
7	0.0038	0.2122	0.0075	0.4244	0.2243	0.2122
12	0.0038	0.5722	0.0075	1.1445	0.5769	0.5722
22	0.0038	0.9597	0.0075	1.9193	0.9597	0.9608

Table 3: Signed rank test p-values for the DCA data varying $n_0, n_+ = 15, n_- = 3$

n_0	One-tailed Wilcoxon	One-tailed Pratt SR	One-tailed Median SR	Two-tailed Wilcoxon	Two-tailed Pratt SR	Two-tailed Median
0	0.0053	0.0053	0.0053	0.0107	0.0107	0.0105
1	0.0053	0.0046	0.0049	0.0107	0.0091	0.0096
2	0.0053	0.0039	0.0048	0.0107	0.0079	0.0093
3	0.0053	0.0036	0.0052	0.0107	0.0072	0.0097
4	0.0053	0.0033	0.0059	0.0107	0.0066	0.0105
7	0.0053	0.0027	0.0103	0.0107	0.0055	0.0161
12	0.0053	0.0024	0.0321	0.0107	0.0048	0.0410
22	0.0053	0.0023	0.2034	0.0107	0.0045	0.2130

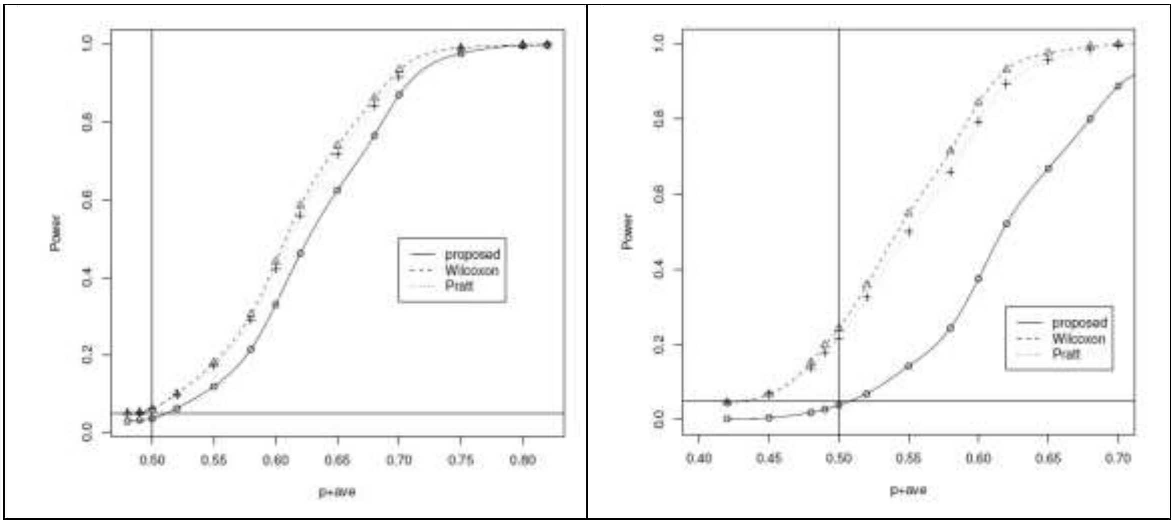

Figure 3: Estimated power for F()=normal, $p_0 = 0.2$(left), 0.4(right),$n = 70$

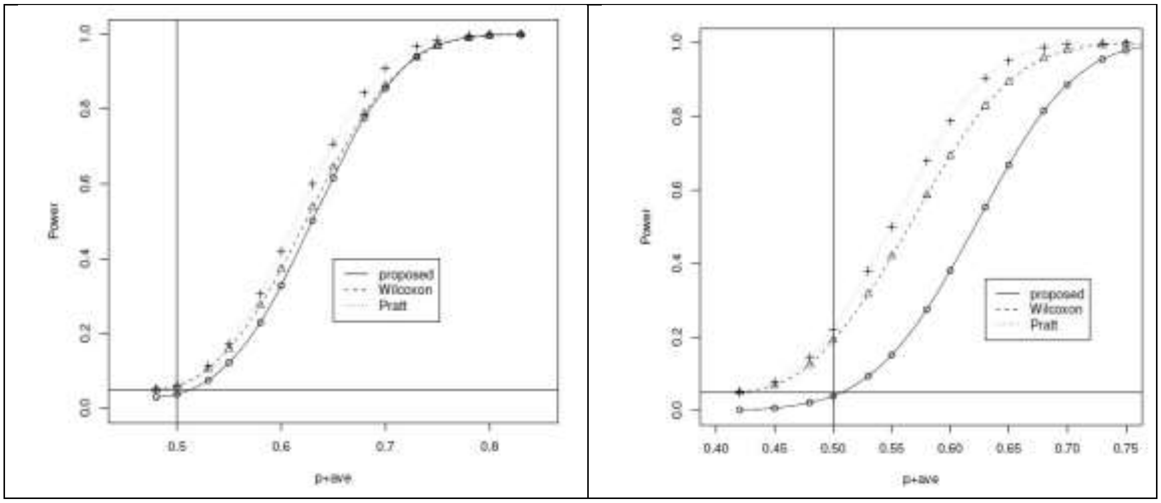

Figure 4: Estimated Power for F()=Cauchy, $p_0 = 0.2$(left), 0.4(right),$n = 70$

Conclusions

In most applications, the zeroes are meaningful and a test about a population median is more appropriate than simply deleting the zeroes. While true in most settings, it is not always the case. Consider the AZT data reported by Makutch and Parks (1988) which displays serum antigen levels for 20 AIDS patients before and after treatment with AZT. Some of the patients had 0 serum antigen levels before treatment so their levels could only go up or stay the same. This data includes several types of zeroes, many with different interpretations. Including the zeroes in the analysis would seem to be problematic because of varying interpretations.

So, how should one decide whether or not to include zeroes in the analysis? The researcher needs to decide whether the focus is on

(a) What is a typical response (a difference with a paired data)

or

(b) Which type of change (increase or decrease) is more prevalent?

If a typical (median) response is the focus, then a median sign or median signed-rank test as proposed in this paper is the proper way to handle the zeroes. This is completely analogous to the paired t-test where typical is interpreted as the average response (difference) and zeroes are always included. If, on the other hand, the focus is on (b), then the zeroes are irrelevant and should be discarded. Asking whether a population like one shown in Figure 1 (right panel) should be detected, may help to elicit the choice between (a) or (b).

References

1. Arbuthnot, J. (1710). An argument for devine providence, taken from the constant regularity observed in the births of both sexes. Philosophical Transactions, 27, 186-190.

2. Coakley, C. W. and Heise, M. A. (1996). Versions of the sign test in the presence of ties. Biometrics, 52, 1242-1251.
3. Dixon, W. J. and Mood, A. M. (1946). The statistical sign test. Journal of the American Statistical Association, 41, 557-566.
4. Derrick, B. and White, P. (2017). Comparing two samples from an individual likert question. International Journal of Mathematics and Statistics, 18(3), 1-13.
5. Dey, R. (2018). Hypotheses tests with precedence probabilities and precedence-type tests. Wires Computational Statistics, 10:e1417.
6. Fong, D. Y. T., Kwan, C. W.,Lam, K. F. and Lam, K. S. L. (2003). Use of the sign test for the median in the presence of ties. The American Statistician, 57, 237-240.
7. Hollander, M.,Wolfe, D. A. and Chicken, E. (2014). Nonparametric Statistical Methods. New York, NY. Wiley.
8. Makuch, R. W. and Parks, W. P. (1988). Response of Serum Antigen level to AZT for the treatment of AIDS. AIDS research and Human Retroviruses, 4, 305-316.
9. Pratt, J. W.(1959). Remarks on zeros and ties in the Wilcoxon signed rank procedures. Journal of the American Statistical Association, 54, 655-667.
10. Presnell, B. (1996). Bootstrap uncoditional p-values for the sign test with ties and the 2X2 matched pairs trial. Journal of Nonparametric Statistics, 7, 47-55.
11. Randles, R. H.(2001). On neutral responses (zeros) in the sign test and ties in Wilcoxon-Mann-Whitney test. The American Statistician, 55(2), 96-101.
12. Siegel, S. and Castellan, N. J. Jr. (1988). Nonparametric Statistics for the Behavioral Sciences. Boston, MA. McGraw Hill.
13. Starks, T. H. (1979). An improved sign test for experiments in which neutral responses are possible. Technometrics, 21, 525-530.
14. Suissa, S. and Shuster, J. J. (1991). The 2X2 matched pairs trial:Exact unconditional design and analysis. Biometrics, 47, 361-372.
15. Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics, 1, 80-83.
16. Wilcoxon, F. and Wilcox, R. A. (1964). Some rapid approximate statistical procedures. Pearl River, NY. American Cyanamid Co., Lederle Laboratories.