Nonparametric Techniques for One Sample

This month's publication introduces nonparametric techniques for a single sample. Over the years, we have produced several publications involving analyzing sample results. For example, you might want to determine if the mean of a process is a certain value. To do that, you take samples from the process and then compare the results using either a t-test or a z-test. Many statistical techniques, like the t-test and z-test for a mean, are based on the assumption that your data are normally distributed.

The assumption of normality is often simply ignored. But there are times when this assumption is not valid. For example, lifetime data (such as product survival times) are not normally distributed. Neither are data involving call center waiting times, bacterial growth, or the number of injuries in a plant. What do you do when the assumption of normality is not valid?

There are techniques called nonparametric statistical methods that can be used when the data are not normal. These techniques are distribution-free; they make no assumptions about the distribution from which you take the sample.

In this issue:

Introduction to Nonparametric Techniques

Nonparametric techniques are statistical methods that are distribution-free. You don't have the assumption that the data are normally distributed. One major difference between nonparametric techniques and those requiring normally distributed data is the use of the median instead of the average. The nonparametric techniques will make use of the median, which will be denoted by $\tilde{\mu}$. The median gives a better estimate of the center than the average for non-normal distributions.

We will cover two nonparametric techniques below. These deal with a single sample and discovering something about the population median being sampled. The example data and the mathematical equations to do the analysis come from the book "Statistics and Data Analysis: From Elementary to Intermediate" by Ajit Tamhane and Dorothy Dunlop.

Sign Test for a Single Sample

In this test, a random sample is taken from a population. The results are then used to determine if the population median is equal to some value or different from some value. For example, a sample of ten thermostats are taken at random from a production lot. The design setting for these thermostats is 200. We want to know if this is true for the production lot. So, each thermostat is tested. The results are given below.

Table 1: Thermostat Setting Data

202.2	198.0
203.4	203.7
200.5	200.8
202.5	201.3
206.3	199.0

The sign test for a single sample is used below to see if the population median, based on this sample, is 200. Using the statistics hypothesis route, we are testing the following hypotheses:

$$H_0: \tilde{\mu} = \tilde{\mu}_0 = 200$$

$$H_1: \tilde{\mu} <> \tilde{\mu}_0 = 200$$

where H_0 is the null hypothesis and H_1 is the alternate hypothesis. Note that if the null hypothesis is true, then the probability of a sample being larger or smaller that $\tilde{\mu}_0$ is ½ or 0.5. The sign test methodology is straight-forward. There are essentially three steps:

1. Count the number of individual results (x_i) that are larger than $\tilde{\mu}_0$. This is the number of plus signs and is denoted by s+.
2. Count the number of individual results (x_i) that are smaller than $\tilde{\mu}_0$. This is the number of minus signs and is denoted by s-.
3. Reject H_O if s+ is large or if s- is small

The first steps are easy to do. In this example, s+ is 8, while s- is 2. There are 8 values greater than 200 and 2 values less than 200. Step 3 is the one where you make your decision though. Like many statistical tests, you must select the probability of making a mistake. This usually focuses on the alpha value (α). It is the probability of rejecting the null hypothesis when it is actually true. Typical values of α include 0.05 and 0.01. You decide that you want α to be 0.05. This means that there is only a 5% chance of rejecting the null hypothesis when it is true.

How do you decide to accept or reject the null hypothesis? One way to do this is to assume that the null hypothesis is true and then determine the probability (p value) of getting the sample result. If the p value is large, it means that there is a large probability of getting the sample result when the null hypothesis is true, and you will accept that the null hypothesis is probably true. But if the probability of getting the sample result is small, you will assume that the null hypothesis is probably not true and reject it in favor of the alternative hypothesis. The small is what α controls.

You can calculate the p value for the sign test by using the binomial distribution. With this distribution, there are only two possible outcomes. In our example, it is either larger than or less than 200.

The p-value is given by the following equation:

$$\text{p value} = 2\sum_{i=s_{max}}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^{n} = 2\sum_{i=0}^{s_{min}} \binom{n}{i} \left(\frac{1}{2}\right)^{n}$$

where n = sample size, s_{max} = max(s+, s-) and s_{min} = min(s+,s-). In Excel, you don't have to perform the calculation shown in the equation above. You can use the BINOMDIST or BINOM.DIST functions with the equation above with the $s_{min.}$

$$\text{p value} = 2^* \text{ BINOMDIST}(s_{min}, n, p, \text{TRUE}) = 2^*\text{BINOMDIST}(2,10, 0.5, \text{TRUE}) = 0.110.$$

The p value for the data is 0.110. This is larger than 0.05, the value of α we selected. The conclusion is that the thermostat design setting is not different from 200. We accept the null hypothesis.

You can also construct a confidence interval to see if the design setting of 200 lies in the confidence interval. The confidence intervals are a little different with this type of test than with, for example, the t-test. Since this is binomial data, you can't have an exactly 95% confidence interval (based on $1 - \alpha$). However, you can use the cumulative binomial probabilities to determine the confidence interval. It has the following form:

$$X_{(b+1)} \leq \tilde{\mu} \leq X_{(n-b)}$$

where b is the lower α /2 critical point of the binomial distribution.

The first step in finding the confidence interval is to sort the data in ascending order. This is shown in Table 2.

Table 2: Sorted Thermostat Setting Data

Number (b)	Thermostat Setting
1	198.0
2	199.0
3	200.5
4	200.8
5	201.3
6	202.2
7	202.5
8	203.4
9	203.7
10	206.3

To find the confidence interval, start with the first thermostat setting and calculate the following:

$$1 - \alpha = 1 - 2 * \text{BINOMDIST}(b, 10, 0.5, \text{True}) = 0.9785 \text{ or } 97.85\%$$

where b= 1. The 97.85% confidence interval is then given by:

$$X_{(b+1)} \leq \tilde{\mu} \leq X_{(n-b)}$$

$$X_2 \leq \tilde{\mu} \leq X_9$$

$$199 \leq \tilde{\mu} \leq 203.7$$

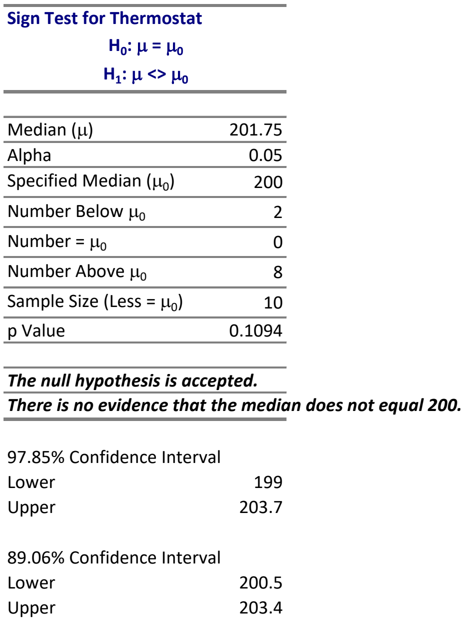
Now go to the second point and do the following calculation:

$$1 – \alpha = 1 - 2 * \text{BINOMDIST}(2, 10, 0.5, \text{True}) = 0.8906 \text{ or } 89.06\%$$

So, 89.06% confidence interval is given by the third and eight results in the table: 200.5 to 203.4.

The output from the SPC for Excel program for this data is shown below.

Figure 1: Sign Test Output



Sign Test for Thermostat

H_0: $\mu = \mu_0$

H_1: $\mu <> \mu_0$

Median (μ)	201.75
Alpha	0.05
Specified Median (μ_0)	200
Number Below μ_0	2
Number = μ_0	0
Number Above μ_0	8
Sample Size (Less = μ_0)	10
p Value	0.1094

The null hypothesis is accepted.

There is no evidence that the median does not equal 200.

97.85% Confidence Interval	
Lower	199
Upper	203.7

89.06% Confidence Interval	
Lower	200.5
Upper	203.4

What happens if the sample result is equal to the design setting ($\tilde{\mu}_0$)? The process above assumes that this does not happen. But, of course, it can happen. The easiest thing to do is to ignore ties and just use the rest of the data. This does impact the sample size of course, but it is rare that there will be many samples that equal $\tilde{\mu}_0$. If there are, then the null hypothesis is probably true – or your measurement system needs some work because it can't tell the difference between samples.

Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test is another parametric method to analyze sample results taken from a non-normal distribution. In general, the steps are:

1. Calculate the absolute value of each sample result from $\tilde{\mu}_0$: $d_i = |x_i - \tilde{\mu}_0|$
2. Rank order the differences with r_i = the rank of d_i
3. Calculate w+ which is the sum of the ranks of the positive differences
4. Calculate w- which is the sum of the ranks of the negative differences
5. Reject H_O if w+ is large or if w- is small

Once again, you have to calculate the p value to determine if w+ is considered large or if w- is considered small. This involves the use of the null distribution. We will continue to use the thermostat data from Table 1.

Table 3 shows the thermostat data with the differences and the ranks.

Table 3: Wilcoxon Signed Rank Test Rankings

Thermostat Setting	Difference from 200	|Difference|	Rank
200.5	0.5	0.5	1
200.8	0.8	0.8	2
199.0	-1.0	1	3
201.3	1.3	1.3	4
198.0	-2.0	2	5
202.2	2.2	2.2	6
202.5	2.5	2.5	7
203.4	3.4	3.4	8
203.7	3.7	3.7	9
206.3	6.3	6.3	10

You can now calculate w+ and w-. w- is the sum of the ranks for those differences that are negative. There are only two differences that are negative. The sum of the ranks is 3 + 5 = 8. So, w- is 8. To find w+, you sum the ranks of the positive differences. The result is w+ = 47

To calculate the p value, you use the null distribution to determine the p value:

$$p \text{ value} = 2^* P\{W \geq w+)$$

You have to look up this probability from a table of the upper probabilities of the null distribution of the Wilcoxon Signed Range statistic. You can download this table at this link. This table handles samples up to 20. The probability from the table is 0.024. Thus,

$$p \text{ value} = 2(0.024) = 0.048.$$

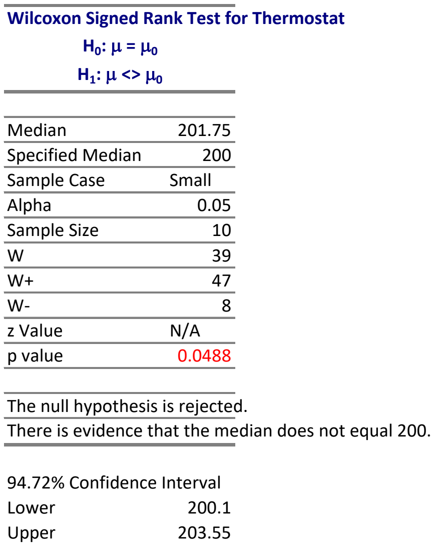
Note that p value calculated for the Wilcoxon Ranked Sign test is less than $\alpha = 0.05$ – so we conclude that the population median is different than 200. The Sign Test did not find a difference.

The Wilcoxon Signed Rank Test has two types of ties. One is when the sample result equals $\tilde{\mu}_0$. Like the sign test, these are ignored. The other tie is when several $|d_i|$ values have the same rank. In this case, you assign an average rank to them. For example, suppose the first two $|d_i|$ values are the same and ranked 1 and 2. Then the average range for both is 1.5.

You can calculate a confidence interval as well, but it involves looking at all pairwise averages. We will not do that here.

Figure 2 shows the output for this test using the SPC for Excel software.

Figure 2: Wilcoxon Signed Rank Test Output



Wilcoxon Signed Rank Test for Thermostat

H_0: $\mu = \mu_0$

H_1: $\mu <> \mu_0$

Median	201.75
Specified Median	200
Sample Case	Small
Alpha	0.05
Sample Size	10
W	39
W+	47
W-	8
z Value	N/A
p value	0.0488

The null hypothesis is rejected.
There is evidence that the median does not equal 200.

94.72% Confidence Interval	
Lower	200.1
Upper	203.55

Summary

This publication examined two methods for analyzing single samples taken from non-normal distributions. One method is the Sign Test. This method involves looking at the number of sample results above $\tilde{\mu}_0$ and the number of sample results below $\tilde{\mu}_0$. The other method is the Wilcoxon Signed Rank Test. This method involves examining the distances the sample results are from $\tilde{\mu}_0$. Both tests focus on the median, not the average.

Quick Links

Visit our home page

SPC for Excel Software

SPC Training

SPC Consulting

SPC Knowledge Base

Ordering Information

Thanks so much for reading our publication. We hope you find it informative and useful. Happy charting and may the data always support your position.

Sincerely,

Dr. Bill McNeese
BPI Consulting, LLC