Nonparametric Hypothesis Testing

Parametric statistical tests are tests that make assumptions about the underlying population distribution (and hence parameters) from which the sample data are drawn, while non-parametric tests are tests that make no such assumptions. Thus, nonparametric tests are often called *distribution free tests*. In general, parametric statistical tests assume that the population data follow the normal distribution. Note that the use of nonparametric tests instead of parametric ones, when the distributional assumptions are satisfied, is not recommended since the nonparametric tests are less powerful in such situation.

Nonparametric Hypothesis Testing about one Population Mean (use median) [for sample data]:
(One-Sample Wilcoxon Signed Rank Test)

The assumptions of this test are:

This is a nonparametric alternative to a one-sample t-test. It's used to examine whether the median of the population is equal to some specified/hypothesized value. This test assumes that the data are distributed symmetrically about the median. Recall that if skewness is 0, then the data are perfectly symmetrical. In real data however, it's almost impossible to obtain exactly a skewness value of zero. Therefore, as a rule of thumb, we use the followings:

- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
- If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

The null and alternative hypotheses, respectively, are:

$$H_0 : \text{Either } \tilde{\mu} \geq \tilde{\mu}_0 \quad \text{or} \quad \tilde{\mu} \leq \tilde{\mu}_0 \quad \text{or} \quad \tilde{\mu} = \tilde{\mu}_0 \ .$$
$$H_1 : \text{Either } \tilde{\mu} < \tilde{\mu}_0 \quad \text{or} \quad \tilde{\mu} > \tilde{\mu}_0 \quad \text{or} \quad \tilde{\mu} \neq \tilde{\mu}_0 \ .$$

REMARK: To implement the One-Sample Wilcoxon Signed Rank Test, In STATA, we use the following syntax:

`signrank Variable=Hypothesized-value`

However, STATA provides only the p-value corresponding to two sided test. To get the p-value for the one-sided test, just divide the provided p-value by 2.

What to do if the data are NOT distributed symmetrically about the median?

Nonparametric Hypothesis Testing about one Population Mean (use median) [for sample data]:
(One-Sample Sign Test)

The assumptions of this test are:

This is a nonparametric alternative to a one-sample t-test. It's used to examine whether the median of the population is equal to some specified/hypothesized value. This test makes no assumption about the shape of the population distribution, therefore this test can handle non-symmetric data.

REMARK: To implement the One-Sample Sign Test, In STATA, we use the following syntax:

`signtest Variable=Hypothesized-value`

(1) **problem 1:** Consider the Diabetes and obesity, cardiovascular risk factors data set we have used in Lab 1 (link: http://www.mathalpha.com/lab1/diabetes.dta) to test whether the true population mean height of African Americans in Virginia is less than 66.5"? Please indicate the name of the appropriate statistical test to use in here.

(2) **problem 2:** Consider the Diabetes and obesity, cardiovascular risk factors data set we have used in Lab 1 (link: http://www.mathalpha.com/lab1/diabetes.dta) to test whether the true population mean age of African American adults in Virginia is more than 44.7 years old? Please indicate the name of the appropriate statistical test to use in here.

(3) **problem 3:** Consider the Diabetes and obesity, cardiovascular risk factors data set we have used in Lab 1 (link: http://www.mathalpha.com/lab1/diabetes.dta) to test whether the true population mean Cholesterol/HDL Ratio among African Americans in Virginia is more than 4? Please indicate the name of the appropriate statistical test to use in here.

Nonparametric Hypothesis Testing about two Population means (use median) [for sample data]:
(Wilcoxon Rank Sum Test for two Independent Samples [sometimes it's called the Mann-Whitney U-test])

The assumptions of this test are:

   This is a nonparametric alternative to a Two Independent Samples T-test. It's used in order to overcome the underlying assumption of normality in the parametric Two Independent Samples T-test. This test does not require that the underlying populations be normally distributed or that their variances be equal. It does, however, assumes that the samples are random, and that observations drawn for each sample are independent from each other. This test examines whether the two distributions are identical or whether the medians of the two populations are equal:

The null and alternative hypotheses, respectively, are:

$$H_0 : \text{The distributions of the two populations are identical .}$$
$$H_1 : \text{The distributions of the two populations are not identical.}$$

   OR

$$H_0 : \text{Either } \tilde{\mu}_1 \geq \tilde{\mu}_2 \quad \text{or} \quad \tilde{\mu}_1 \leq \tilde{\mu}_2 \quad \text{or} \quad \tilde{\mu}_1 = \tilde{\mu}_2 .$$
$$H_1 : \text{Either } \tilde{\mu}_1 < \tilde{\mu}_2 \quad \text{or} \quad \tilde{\mu}_1 > \tilde{\mu}_2 \quad \text{or} \quad \tilde{\mu}_1 \neq \tilde{\mu}_2 .$$

REMARK 1: To implement the Wilcoxon Rank Sum Test, In STATA, we use the following syntax:

```
ranksum variable, by(group)
```

REMARK 2: An alternative nonparametric tests for the Wilcoxon Rank Sum Test are the median test and kolmogorov-smirnov test. In STATA, we implement the median and kolmogorov-smirnov tests by using the following syntax respectively:

```
median variable, by(group)
ksmirnov variable, by(group)
```

REMARK 3: The kolmogorov-smirnov (KS) test is a test for equality of two distributions and thus the KS test is sensitive to any differences in the two distributions as opposed to the Wilcoxon Rank Sum (Mann-Whitney) Test. The median test only considers the position of each observation relative to the overall median, whereas the Wilcoxon Rank Sum (Mann-Whitney) Test takes the ranks of each observation into account. Thus the Wilcoxon Rank Sum (Mann-Whitney) Test is usually more powerful than the median test.

(4) **problem 4:** Consider the Diabetes and obesity, cardiovascular risk factors data set we have used in Lab 1 (link: http://www.mathalpha.com/lab1/diabetes.dta) to test whether the true population HDL mean among African Americans females in Virginia is more than that of males? Please indicate the name of the appropriate statistical test to use in here.

Nonparametric Hypothesis Testing for paired data (use median) [for sample data]:
(Wilcoxon signed rank test)

The assumptions of this test are:
The Wilcoxon signed rank test is the non-parametric version of a paired samples t-test. It's used in order to overcome the underlying assumption of normality in the parametric paired samples t-test. This test does not require that the underlying population be normally distributed. It does, however, assumes that
- The sample is random.
- The paired observations are independent from each other.
- The distribution of the differences between the two related variables are symmetrical in shape.

This test examines the following hypotheses:

$$H_0 : \text{Either } \tilde{\mu}_1 \geq \tilde{\mu}_2 \quad \text{or} \quad \tilde{\mu}_1 \leq \tilde{\mu}_2 \quad \text{or} \quad \tilde{\mu}_1 = \tilde{\mu}_2 \ .$$
$$H_1 : \text{Either } \tilde{\mu}_1 < \tilde{\mu}_2 \quad \text{or} \quad \tilde{\mu}_1 > \tilde{\mu}_2 \quad \text{or} \quad \tilde{\mu}_1 \neq \tilde{\mu}_2 \ .$$

OR

$$H_0 : \text{Either } \tilde{\mu}_{diff} \geq 0 \quad \text{or} \quad \tilde{\mu}_{diff} \leq 0 \quad \text{or} \quad \tilde{\mu}_{diff} = 0 \ .$$
$$H_1 : \text{Either } \tilde{\mu}_{diff} < 0 \quad \text{or} \quad \tilde{\mu}_{diff} > 0 \quad \text{or} \quad \tilde{\mu}_{diff} \neq 0 \ .$$

REMARK: To implement the Wilcoxon Signed Rank Test for paired data, In STATA, we use the following syntax:

```
signrank Variable1=Variable2
```

What to do if the difference data are NOT distributed symmetrically about the median?

Nonparametric Hypothesis Testing for paired data (use median) [for sample data]:
(The Sign Test for Paired data)

The assumptions of this test are:
This is a nonparametric alternative to a paired samples t-test. This test makes no assumption about the shape of the population distribution of the difference, therefore this test can handle non-symmetric data.

REMARK: To implement the Sign Test for Paired data, In STATA, we use the following syntax:

```
signtest Variable1=Variable2
```

8

(5) **problem 5:** In a pilot study of 44 adults, measurements of systolic blood pressure were taken on each subject at two different times. The first measurement was taken one hour before lunch and the second measurement was taken one hour after lunch. Use the following sample data to test whether the mean systolic blood pressure after lunch is significantly different than the systolic blood pressure before lunch? Please indicate the name of the appropriate statistical test to use in here.

```
clear all
input sbefore safter sdiff
190 185 -5
161 161 0
160 128 -32
130 130 0
130 120 -10
148 148 0
137 149 12
158 160 2
100 110 10
180 170 -10
146 139 -7
170 168 -2
176 180 4
158 160 2
149 136 -13
150 142 -8
158 140 -18
158 149 -9
165 153 -12
159 160 1
158 148 -10
150 145 -5
160 160 0
138 140 2
142 130 -12
156 144 -12
178 182 4
151 148 -3
125 117 -8
142 156 14
160 150 -10
160 158 -2
136 130 -6
136 126 -10
150 156 6
170 170 0
158 158 0
160 162 2
159 150 -9
150 150 0
150 150 0
```

```
160 158 -2
180 210 30
146 168 22
end
```