STAT 40001/STAT 50001 Statistical Computing

Lecture 13

Department of Mathematics and Statistics

Nonparametric Statistics

Nonparametric statistics provides an alternative series of statistical methods that require no or very limited assumptions to be made about the data. There are many methods that can be used in different circumstances.
Unlike their parametric counterparts, non-parametric methods make no specific assumptions about the distribution of the data, nor do they necessarily rely on estimates of population parameters (such as the mean) in order to describe a variable's distribution. For this reason, nonparametric statistics are also often referred to as distribution-free statistics.
A statistical method is nonparametric if it satisfies one of the following criteria(W.J. Conover, Practical Nonparmetric Statistics)

- The method may be used on data with a nominal scale of measurement.
- the method may be used on data with an ordinal scale of measurement.
- The method may be used on data with an interval or ratio scale of measurement, where the distribution function of the random variable is either unspecified or specified except for an infinite number of unknown parameters.

Advantages and disadvantages of nonparametric methods

Given that NP methods make less stringent demands on the data, one might wonder why they are not used more often. There are a few reasons:

- Tradition: Most traditional statistical analyses are parametric.
- Loss of information: Many nonparametric procedures discard information. For example, if we convert severely skewed interval data into ranks, we are discarding the actual values and only retaining their "order".
- Sample Size: Many nonparametric analysis require larger sample size
- Limitations. There are certain types of information that only parametric statistical tests can provide.
- Inefficient: They are usually not as "efficient" as their parametric counterparts when the assumptions can be met.

Advantages of Nonparametric statistics:

- They can be used with non-normally distributed data.
- They can be used with discrete data (nominal or ordinal measurement level), whereas many parametric methods require continuous data (interval, ratio measurement level)
- Methods can be highly customizable.

Nonparametric Methods

Non-parametric methods have long been used to establish whether or not two or more groups are the same. The tests work by first ranking the data and then using the rank values, not the actual data values. For this reason the tests are sometimes referred to as **robust tests:** the tests are robust to the influence of outliers. The situation is analogous to the comparison of the mean and the median. To find the median we rank the data from highest to lowest and look at the middle value. The median is therefore referred to as a measure that is robust to the influence of outliers. This is in contrast to the mean, which can be influenced by outliers.

Rather than use the raw data values, non-parametric methods convert the raw data into ranks first and then use the rank values in the test.

Raw Data:	17	20	12	4	25	27	42	34
Rank Data:	3	4	2	1	5	6	8	7

Testing for Normality

Graphical Methods:

- QQ plots: The quantilie-quantile plot compares ordered values of a variable with quantiles of a specific theoretical distribution
- PP plots: The probability-probability plot (P-P plot or percent plot) compares an empirical cumulative distribution function of a variable with a specific theoretical cumulative distribution function

Numerical(Statistical) Methods

- Shapiro-Wilk test
- Kolmogorov-Smirnov test
- Anderson-Darling test
- Cramer-von Mises test

Note that The Kolmogorov-Smirnov statistic, the Anderson-Darling statistic, and the Cramer-von Mises statistic are based on the empirical distribution function (EDF).

Statistical Test for Normality

Statistical tests for normality are more precise since actual probabilities are calculated. The Kolmogorov-Smirnov and Shapiro-Wilks tests for normality calculate the probability that the sample was drawn from a normal population.
The hypotheses used are:

H_0 : The sample data are not significantly different than a normal population

H_a : The sample data are significantly different than a normal population

It will be a good idea to perform the exploratory data analysis using *EDA()* function available in the `PASWR` package.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (also known as KS) statistic is based on the largest vertical difference between the hypothesized and empirical distribution. Assume $x_1, x_2, \cdots, x_n$ is an iid sample from a continuous distribution with CDF $F(x)$. Let $F_n(x)$ be the empirical cdf . A significance test of $H_0 : F(x) = F_0(x)$ Vs. $H_a : F(x) \neq F_0(x)$ can be constructed with the test statistic

$$D = sup_x |F_n(x) - F(x)|$$

Large value of D support the alternative hypothesis. In R, we can use `ks.test()` to perform this test
Example:

```
> x=rnorm(100, mean=5, sd=2)
> ks.test(x, "pnorm", mean=5, sd=2)
        One-sample Kolmogorov-Smirnov test
data:  x
D = 0.0749, p-value = 0.6294
alternative hypothesis: two-sided
```

KS Test-Example

At 0.05 level of significance, test whether the observations 5,6,7,8 and 9 are from a normal distribution with mean (μ)=6.5 and standard deviation $\sigma = \sqrt{2}$

```
> x<-c(5,6,7,8,9)
> mu<-6.5
> sigma=sqrt(2)
> ks.test(x,y="pnorm",mean=mu, sd=sigma)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.2556, p-value = 0.8269
alternative hypothesis: two-sided
```

Decision: Since $p > 0.05$ there is no evidence to reject the null hypothesis that $F_0(x) \sim N(6.5, 2)$.

Shapiro-Wilk Test

The Kolmogorov-Smirnov test for univariate data set works when the distribution in the null hypothesis is fully specified prior to looking at the data. So we can't use Kolmogrove-Smirnov test to test for normality of the data unless we know the parameters of the underlying distribution. The Shapiro-Wilk test is a powerful normality test appropriate for small samples and allows to test for normality without specifying the parameters. Let $x_1, x_2, \cdots, x_n$ is an iid sample from a continuous distribution. A significance test of

$$\begin{aligned} H_0 &: \text{parent distribution is normal} \\ H_a &: \text{the parent distribution is not normal} \end{aligned}$$

The test statistic measures how closely the empirical quantiles follow the corresponding theoretical quantiles of a normal distribution. In R, we can use `shapiro.test()` to perform this test

Shapiro-Wilk Test-Example

Example 1: Test whether
$47, 50, 57, 54, 52, 54, 53, 65, 62, 67, 69, 74, 51, 57, 57, 59$ is a normal random sample.

```
> x <- c(47,50,57,54,52,54,53,65,62,67,69,74,51,57,57,59)
> shapiro.test(x)
        Shapiro-Wilk normality test

data:  x
W = 0.9471, p-value = 0.4445
```

Decision: Fail to reject the null hypothesis of normality.

Example 2: `stud.recs` data set are available in `UsingR` package. It has 160 observations 6 variables of students' records including the SAT math and SAT verbal scores. Use Shapiro- Wilk test to check whether the SAT math scores and SAT verbal scores are normally distributed.

```
> library(UsingR)
> data(stud.recs)
> attach(stud.recs)
> shapiro.test(sat.m)
```

Additional Tests for Normality

Several other methods have been proposed and studied which are modifiaction of KS test or SW test.
Lilliefors test is a modification of Kolmogorov-Smirnov test appropriate for testing normality when parameters are not known.
`lillie.test` function is located in `nortest` package.

Anderson-Darling test is one of the most powerful normality tests as it will detect the most of departures from normality.
`ad.test` function in `nortest` package.

Significance test for the median

The significance test for the mean relies on large sample , or on an assumption that the parent distribution is normally(or nearly normally) distributed. In order to test for median we need to use nonparametric methods

The Sign Test

The sign test is a simple test for the median of a distribution that has no assumption on the parent distribution except that it is continuous with positive density. Assume $X_1, X_2, \cdots, X_n$ are from a continuous distribution . We want to test

- H_0 : Median $= m$
- H_a: One of the following
  - Median $\neq m$;
  - Median $< m$;
  - Median $> m$

The test statistic is given by

$$T = \text{the number of } X_i \text{ with } X_i > m.$$

If there are data value equal to m simply delete those values.
Note that under the null hypothesis T has Binomial$(n, 1/2)$

Sign Test-Example

The function `SIGN.test()` provided in the `PASWR` package perform the sign test.

It will be a good idea to perform the exploratory data analysis using EDA() function provided in the PASWR package.

Suppose a cell phone bill contains the data below for the number of minutes per call

$$2, 1, 3, 3, 3, 3, 1, 3, 16, 2, 2, 12, 20, 3, 1$$

Test the hypothesis that median length per call is less than 5 minutes.

```
> x<-c(2,1,3,3,3,3,1,3,16,2,2,12,20,3,1)
> SIGN.test(x, md=5, alt="less")
        One-sample Sign-Test
data:  x
s = 3, p-value = 0.01758
alternative hypothesis: true median is less than 5
95 percent confidence interval:
 -Inf    3
sample estimates: median of x
                  3
```

The signed Rank Test

The signed-rank test is an improvement to the signed test when the population is symmetric, but not close enough to normal or t- test. Like the Sign Test, it is based on difference scores, but in addition to analyzing the signs of the differences, it also takes into account the magnitude of the observed differences. Let $X_1, X_2, \cdots, X_n$ is an iid sample from a continuous distribution. We want to test

- H_0 : Median $= m$
- H_a: One of the following
  - Median $\neq m$;
  - Median $< m$;
  - Median $> m$

Test Statistics:

$$T = \sum \text{rank}(|X_i - m|)$$

Note that if there are ties , we need to use midranks. The midrank is defined as the average rank of the tied observations. Calculate $T+$, the sum of the ranks with positive Ds and $T-$, the sum of the ranks with negative Ds. The test statistics is the smaller of $T+$ and $T-$. provided the null hypothesis is true $E(T+) = E(T-)$. When $T+$ is either sufficiently large of sufficiently small the null hypothesis will be rejected.

We can use `wilcox.test(x, mu=, alternative=.)` to perform the test.

Example

The waiting time(in minutes) for a bus is recorded by a passage for 15 days and provided below:

$$8, 2.1, 3.8, 8.6, 7.3, 6.1, 1.4, 2.9, 5.5, 2.7, 4.8, 4.6, 1, 8.7, 0.8$$

Test to see whether the median waiting time is less than 6 minutes.

```
>x<-c(8,2.1,3.8,8.6,7.3,6.1,1.4,2.9,5.5,2.7,4.8,4.6,1,8.7,0.8)
> wilcox.test(x,mu=6, alternative="less")
        Wilcoxon signed rank test
data:  x
V = 28, p-value = 0.0365
alternative hypothesis: true location is less than 6

> wilcoxE.test(x,mu=6, alternative="less")
        Wilcoxon Signed Rank Test
data:  x
t+ = 28, p-value = 0.0365
alternative hypothesis: true median is less than 6
95.26978 percent confidence interval:
 -Inf  5.8
sample estimates:
(pseudo)median
         4.625
```

Wilcoxon Rank sum Test

The Wilcoxon Rank Sum test is used to test for a difference between two samples. It is the nonparametric counterpart to the two-sample Z or t test. Instead of comparing two population means, we compare two population medians. It is also known as Mann-Whitney U-test.
The Wilcoxon test has the following assumptions:

- Both sets of samples (the X and Y observations) are drawn randomly form each population.
- There is an expected mutual independence between the X and Y values as well.
- The variables are at least ordinal.
- There is no requirement that the two populations have normal or any other distribution but have same shape but perhaps different centers

Wilcoxon rank sum Test

We want to test

- H_0: The two sample come from same distribution
- H_a: The two samples come from different distributions.

The test statistic for this analysis is the sum of the ranks of the X variables:

$$W = \sum_{i=1}^{n} R[X_i]$$

If the observations in X and Y are drawn from a single population, as stated in the null hypothesis, then the sum of the ranks of X should be just as large as expected for the sum of the ranks for Y.

Example

A grocery store's management wishes to compare checkout personnel to see if there is a difference in their checkout times. A small sample of data comparing two checkers' times(in minutes) is given below. Compare the mean checkout times.

Checker A	5.8	1.0	1.1	2.1	2.5	1.1	1.0	1.2	3.2	2.7
Checker B	1.5	2.7	6.6	4.6	1.1	1.2	5.7	3.2	1.2	1.3

```
> A=c(5.8, 1.0, 1.1, 2.1, 2.5, 1.1, 1.0, 1.2, 3.2, 2.7)
> B=c(1.5, 2.7, 6.6, 4.6, 1.1, 1.2, 5.7, 3.2, 1.2, 1.3)
> wilcox.test(A,B)
        Wilcoxon rank sum test with continuity correction
data:  A and B
W = 34, p-value = 0.2394
alternative hypothesis: true location shift is not equal to 0
Warning message:
In wilcox.test.default(A, B) : cannot compute exact p-value with
```

Decision: $p > \alpha$ so not enough evidence to reject the null hypothesis.

Example

The amount of rainfall in Boston, MA everyday of the year are discussed in Mario Triola and are provided in the Brightspace. We want to test whether the rainfall amounts for Wednesdays and Saturdays come from the same distribution.

```
> wilcox.test(S,W)
        Wilcoxon rank sum test with continuity correction
data:  S and W
W = 1548, p-value = 0.1992
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(S,W, correct=FALSE)
        Wilcoxon rank sum test
data:  S and W
W = 1548, p-value = 0.1979
alternative hypothesis: true location shift is not equal to 0
```

Example

In a genetic inheritance study discussed by Margolin (1988), samples of individuals from several ethnic groups were taken. Blood samples were collected from each individual and several variables measured. We shall compare the groups labeled "Native American" and "Caucasian" with respect to the variable MSCE (mean sister chromatid exchange). The data is as follows:

```
Native American: 8.50 9.48 8.65 8.16 8.83 7.76 8.63
Caucasian:       8.27 8.20 8.25 8.14 9.00 8.10 7.20 8.32 7.70
```

Test whether the MSCE are the same for Native American and Caucasian.

```
> wilcox.test(N,C)
        Wilcoxon rank sum test
data:  N and C
W = 47, p-value = 0.1142
alternative hypothesis: true location shift is not equal to 0
```

Decision: No evidence against H0 which suggests that median MSCE measurements are the same for both group.

Example

The intelligence test of children was studied by Johnson and Courtesy (Child Development, Vol. 3) by measuring the time taken to complete blocks. Each child was given two trials. Data below gives the time taken by the children in first and second trial.

Child	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
First Trial	30	19	19	23	29	178	42	20	12	39	14	81	17	31	52
Second Trial	30	6	14	8	14	52	14	22	17	8	11	30	14	17	15

Test the hypothesis whether there is a difference between the times of the first and second trial.

```
> wilcox.test(x,y, paired=TRUE)
        Wilcoxon signed rank test with continuity correction
data:  x and y
V = 99.5, p-value = 0.003486
alternative hypothesis: true location shift is not equal to 0
```

Decision: Reject the null hypothesis concluding that there is a difference between times for the first trial and time for the second trial.