Chapter 6

Data Analytics 4: Nonparametric Statistics

Prepared by Dr Yuslina Zakaria

Nonparametric tests serve as an alternative to parametric tests when all parametric assumptions are violated. However, these nonparametric tests are less powerful compared to their parametric counterparts. The tests are more appropriate if not in ratio or interval scale, and analysis is carried out using median.

R Packages:

For this lesson, the following packages are used:

- `dplyr`
- `DescTools`

All of these packages has been used in previous sessions.

Datasets

- `subs_heart.csv`
- `subs_anorexia.csv`
- `treatment.csv`

Retrieve the datasets from http://tiny.cc/phc410_da4 and place the files in the `input` folder, in your R project workspace.

Instructions:

-  explains the steps for activity you need to follow.

Load libraries

Before we start, let us load all the required libraries:

```
library(dplyr)
library(DescTools)
```

Convert Scientific Notation into Float

To make it easier for you to evaluate the output from statistical test, convert all scientific notation into float numbers using the following R command.

```
#do option scipen to convert scientific notation into float
options(scipen=999)
```

6.1 Comparing A Single Sample

When the random sample is not from a normally distributed data, we can use Sign test or Wilcoxon signed-rank test.

Question

Suppose we wanted to test whether the average age of female patient with heart disease in this sample is not equal to the average population age of 74.

Assume that the data is not normally distributed.

Hypothesis

Therefore, our null and alternative hypotheses are:

H_0: The median age of female patients with heart disease is equal to 74 ($m = 74$)
H_a: The median age of female patients with heart disease is not equal to 74 ($m \neq 74$)

6.1.1 Sign test

For sign test, you can use the `SignTest()` function from `DescTools` package in R as follows:

```
# one sample Sign test
SignTest(y, mu = median)
```

Preparing the data

In this example, we are using the heart disease dataset `subs_heart.csv` which has been used in the parametric test analysis before.


Import the heart dataset `subs_heart.csv` from `input` folder as `subs_heart` using `read.csv()`. `stringAsFactors` is set to `TRUE` to perform force conversion to all character variables into factors.

```
subs_heart <- read.csv("input/subs_heart.csv",stringsAsFactors = T)
```


Create an object `female_hd` to contain only female patients with heart disease. Observe the structure of `female_hd`.

```
female_hd <- subs_heart %>% filter(target=="presence" & sex=="female")
str(female_hd)
## 'data.frame':    72 obs. of  9 variables:
##  $ age     : int  41 57 56 48 58 50 58 66 69 71 ...
```

```
##  $ sex     : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ cp      : Factor w/ 4 levels "asymptomatic",..: 2 4 2 3 1 3 3 1 1 2 ...
##  $ trestbps: int  130 120 140 130 150 120 120 150 140 160 ...
##  $ chol    : int  204 354 294 275 283 219 340 226 239 302 ...
##  $ fbs     : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
##  $ thalach : int  172 163 153 139 162 158 172 114 151 162 ...
##  $ exang   : Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 1 1 1 ...
##  $ target  : Factor w/ 2 levels "absence","presence": 2 2 2 2 2 2 2 2 2 2 ...
```

Performing the test

Now, let us proceed to the Sign test.

Perform Sign test using the median score of 74.

```
#using SignTest() function from DescTools package
SignTest(female_hd$age,mu = 74)
##
##  One-sample Sign-Test
##
## data:  female_hd$age
## S = 1, number of differences = 71, p-value < 0.00000000000000022
## alternative hypothesis: true median is not equal to 74
## 95.6 percent confidence interval:
##  51 58
## sample estimates:
## median of the differences
##                        54
```

Conclusion and Interpretation

The p-value (0.000) is less than our significance level of 0.05. Therefore, we reject the null hypothesis. There is enough evidence in the data to suggest that the median score of the sample is not equal to 74.

6.1.2 Wilcoxon Signed-rank Test

The one-sample Wilcoxon signed-rank test is another non-parametric alternative to one-sample t-test when the normality assumption is violated. The test is more powerful than Sign test and is usually used for a bigger sample size.

In R, `wilcox.test()` function is used to perform the Wilcoxon test. The followings are the functions that we can use for one sample Wilcoxon test:

```
# one sample Wilcoxon test (one-tailed test)
wilcox.test(y, mu = median, alternative = "greater") # upper-tailed
wilcox.test(y, mu = median, alternative = "less") # lower-tailed

# one sample Wilcoxon test (two-tailed test)
wilcox.test(y, mu = median, alternative = "two.sided")
```

Wilcoxon test is a rank-based test. Therefore, `exact=FALSE` needs to be added in the function in cases where the data contain ties.

`DescTools` package provides `AllDuplicated()` function to identify if the observed values contain ties. In this example, we will check if the variable `age` in `subs_heart` contains ties before proceeding to the Wilcoxon Signed Rank test.

Identify ties in `age` in `female_hd`

```
#TRUE means duplicated values (ties)
dup_age <- AllDuplicated(female_hd$age)
female_hd$age[dup_age]
##  [1] 41 57 58 50 58 66 71 65 41 46 54 65 65 51 53 53 53 51 44 63 57 71 45 62 55
## [26] 60 42 67 54 58 54 45 62 63 45 50 50 64 64 46 46 64 41 54 39 67 54 49 41 49
## [51] 60 51 42 67 44 60 71 66 39 58 55
```

The result shows that there are 61 duplicate values in `age`, and thus we can confirm that the variable `age` contains ties. We should use `exact=FALSE` in the `wilcox.test()` function.

Perform Wilcoxon signed-rank test using the median score of 74.

```
#set ties to TRUE (exact=FALSE)
wilcox.test(female_hd$age,mu=74,exact=FALSE)
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  female_hd$age
## V = 1, p-value = 0.0000000000002565
## alternative hypothesis: true location is not equal to 74
```

Conclusion and Interpretation

The p-value (0.000) retrieved from the Wilcoxon test is less than our significance level and therefore we reject the null hypothesis. There is enough evidence in the data to suggest that the median score of the sample is not equal to 74.

6.2 Comparing Two Independent Samples

In the following sections, we will estimate the difference in the average of two independent groups and the average of paired samples.

6.2.1 Mann-Whitney U test / Wilcoxon Rank Sum test

The Mann-Whitney U test or also known as Wilcoxon Rank Sum test is a non-parametric alternative to the independent samples t-test, which can be used to compare two independent groups of samples. The test is used when your data are not normally distributed. `wilcox.test()` function is also used to perform Wilcoxon Rank Sum test in R, as follows:

```
# Wilcoxon Rank Sum test (without ties)
wilcox.test(y~x) # y is numerical variable and x is categorical variable

# Wilcoxon Rank Sum test (with ties)
wilcox.test(y~x, exact = FALSE) # y is numerical variable and x is categorical variable
```

Question

Suppose we wanted to test whether the median age of female patients differs from the median age of male patients. Assume that the data is not normally distributed.

Hypothesis

H_0: The median age of female patients is equal to the median age of male patients.
H_a: The median age of female patients is not equal to the median age of male patients.

Performing the test

Perform Wilcoxon Rank Sum test between `age` for female and male patients.

```
#between age for female and male patients
wilcox.test(subs_heart$age ~ subs_heart$sex, exact=FALSE)
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  subs_heart$age by subs_heart$sex
## W = 11158, p-value = 0.08507
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion and Interpretation

The p-value $0.09 > 0.05$ - thus we fail to reject H_0. Therefore, the median age of female patients is equal to the median age of male patients. In other way, there is no significant difference in the median age between male and female patients.

6.3 Comparing Two Paired Samples

The Wilcoxon signed-rank test is used to compare dependent or paired data when the data are not normally distributed. Similar to the one sample test, the Sign test can also be used to compare paired data. However, the Sign test only takes consideration of sign of the differences without taking observation rank into account. Therefore, the Sign test is considered as weaker than the Wilcoxon signed-rank test.

In this section, we will only focus on the use of Wilcoxon signed-rank test to compare the differences between `before` and `after` values.

To perform a non parametric test for paired samples:

```
# Wilcoxon Signed Rank test - y1 & y2 are numerical variables
wilcox.test(y1,y2,paired=TRUE) #
```

We will be using the dataset on weight change of 72 female anorexic patients in `subs_anorexia.csv` in this section.

Question

Suppose we are interested in finding out whether there are effects for patients who underwent a family therapy (FT) to gain some weight in the study.

Assume that the data is not normally distributed.

Hypothesis

H_0: There is no difference in median weight before and after the treatment for those in the FT group.
H_a: There is a difference in median weight before and after the treatment for those in the FT group.

Preparing the data

Assign the data set from `subs_anorexia.csv` with an object `subs_anrx`

```
subs_anrx <- read.csv("input/subs_anorexia.csv",stringsAsFactors = T)
```

Get the subset of `subs_anrx` which only has `FT` for treatment.

Filter to get observations of FT subject (`FT`) with `ft_subs`

```
ft_subs <- subs_anrx %>% filter(Treat=="FT")
```

Performing the test

Perform the Sign test and the Wilcoxon signed-rank test.

```
# Wilcoxon Signed Rank test - contains tie
wilcox.test(ft_subs$Postwt,ft_subs$Prewt,paired=TRUE,exact = FALSE)
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  ft_subs$Postwt and ft_subs$Prewt
## V = 142, p-value = 0.002091
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion and Interpretation

Based on the results, the p-value 0.002 is less than the significance value of 0.05. Therefore, we reject the null hypothesis. There is a significant difference in median weight before and after the treatment for the patients who underwent family therapy (FT).

6.4 Comparing Three Groups and More (3+)

Kruskal-Wallis test is a non-parametric alternative to one-way ANOVA test. This test is used if there are more than two groups in comparison when the assumptions of one-way ANOVA test are not met.

To perform the Kruskal-Wallis test:

```
# Kruskal Wallis test - y is numerical variable and x is categorical variable
kruskal.test(y~x) #
```

Question

Suppose we wanted to compare whether the weight after therapy of anorexic patients differs between three different types of therapy.

Hypothesis

H_0: The median of weight for the three different types of therapy are the same.
H_a: At least one sample median of weight is not equal to the others.

Performing the test

Perform the Kruskal-Wallis test using weight after therapy (`Postwt`) for different types of therapy (`Treat`).

```
#Kruskal Wallis test (Post-weight VS Therapy)
kruskal.test(subs_anrx$Postwt ~ subs_anrx$Treat)
##
##  Kruskal-Wallis rank sum test
##
## data:  subs_anrx$Postwt by subs_anrx$Treat
## Kruskal-Wallis chi-squared = 12.881, df = 2, p-value = 0.001596
```

Conclusion and Interpretation

From the output, we know that the there is a significant difference in between therapy groups. We will perform post-hoc pairwise comparisons to know which pairs of therapy that are significantly different from each other.

6.4.1 Post-hoc Multiple Pairwise-comparisons

The function `pairwise.wilcox.test()` to determine which levels of the independent variable differ from each other level.

To specify the p-value adjustment method, you can set `p.adjust.method` to the following options:

- the Bonferroni correction ("bonferroni")
- Holm (1979) ("holm")
- Hochberg (1988) ("hochberg")
- Hommel (1988) ("hommel")
- Benjamini & Hochberg (1995) ("BH" or its alias "fdr")
- Benjamini & Yekutieli (2001) ("BY")
- pass-through option ("none")

Perform the post-hoc multiple pairwise-comparisons using Bonferroni for p-value adjustment

```
#Post-hoc comparison using Wilcoxon (Post-weight VS Therapy)
pairwise.wilcox.test(subs_anrx$Postwt, subs_anrx$Treat,
                     p.adjust.method = "bonferroni", exact=F)
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  subs_anrx$Postwt and subs_anrx$Treat
##
##      CBT    Cont
## Cont 0.0732 -
## FT   0.2392 0.0021
##
## P value adjustment method: bonferroni
```

Based on the pairwise comparison, only FT and Cont are significantly different ($p<0.05$).

6.5 Comparing Proportions of Categorical Variables

6.6 Chi-square Tests

There are two main types of chi-square test. The chi-square test for goodness of fit applies to the analysis of a single categorical variable, the chi-square test for independence or relatedness applies to the analysis of the relationship between categorical variables.

6.7 Chi-square Test for Goodness of Fit

The chi-square goodness of fit test is used to compare the observed distribution to an expected distribution, in a situation where we have two or more categories in a discrete data. In other words, it compares multiple observed proportions to expected probabilities. Thus, determine if there is any significant difference between the observed proportions and the expected proportions.

To perform the Chi-Square test for goodness of fit:

```
#chi-square test for goodness of fit
chisq.test(x, p) #x is = numeric vector, p =  vector of probabilities (sum of p equals 1)
```

Question

Suppose we are studying a cohort of cancer patients to determine if cancer was more likely to be diagnosed in patients who are in a low-income category, based on socio-economic status (SES) quartiles.

Category of SES	Frequency of Cancer Patients
Highest SES	165
Moderate SES	283
Lower SES	622
Very Low SES	980

The expected distribution of cancer patients within the community is equally distributed across the four income categories so that there are 25% in all categories (probability of 0.25).

Hypothesis

H_0: : There is no significant difference between the observed and the expected value.
H_a: : There is a significant difference between the observed and the expected value.

Prepare the data

Create a dataframe `ses_freq` based on the given distribution.

```
#Create a vector of SES and frequency
ses <- c("Highest SES","Moderate SES","Lower SES","Very Low SES")
freq <- c(165,283,622,980)
ses_freq <- data.frame(ses,freq)
names(ses_freq) <- c("seq","freq")
ses_freq
```

```
##            seq freq
## 1  Highest SES  165
## 2 Moderate SES  283
## 3    Lower SES  622
## 4 Very Low SES  980
```

Performing the test

Perform the chi-square test for the goodness of fit for `ses_freq`

```
#Chi-square for goodness of fit - for equal probability of 0.25 (25%)
chisq.test(ses_freq$freq,p=c(0.25,0.25,0.25,0.25))
##
##  Chi-squared test for given probabilities
##
## data:  ses_freq$freq
## X-squared = 788.24, df = 3, p-value < 0.00000000000000022
res <- chisq.test(ses_freq$freq,p=c(0.25,0.25,0.25,0.25))
```

Conclusion and Interpretation

From the output, we can see that the p-value of the test is 0.0000 ($< \alpha = 0.05$). We can conclude that the observed proportions are significantly different from the expected proportions.

6.8 Chi-square Test of Independence

The Chi-square test of independence tests whether there is a relationship between two categorical variables. The null and alternative hypotheses of this test are:

H_0: There is no relationship between the two categorical variables.
H_a: There is a relationship between the two categorical variables.

For the null hypothesis, it simply means by knowing the value of one variable, it does not help to predict the value of the other variable. In contrast, for the alternative hypothesis, knowing the value of one variable helps to predict the value of the other variable.

Question

Suppose we have 105 patients to test the effectiveness of drug for a certain medical condition. 50 patients were treated with a drug and the remaining 55 patients were kept as control. After the treatment, the health condition of the patients was checked to see if it has improved.

Hypothesis

H_0: There is no relationship between the treatment and the improvement in health condition.
H_a: There is a relationship between the treatment and the improvement in health condition

Prepare the data

 Import the treatment dataset `treatment.csv` from `input` folder as `patient` using `read.csv()`. `stringAsFactors` is set to `TRUE` to perform force conversion to all character variables into factors.

```
patient <- read.csv("input/treatment.csv",stringsAsFactors = T)
str(patient)
## 'data.frame':    105 obs. of  3 variables:
##  $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ treatment  : Factor w/ 2 levels "not-treated",..: 2 2 1 2 2 2 1 2 1 2 ...
##  $ improvement: Factor w/ 2 levels "improved","not-improved": 1 1 1 1 2 2 2 2 1 1 ...
```

The variables under study are `treatment` and `improvement`.

Performing the test  Prepare a contingency table `cont_tbl` of the two variables i.e. `treatment` and `improvement`.

```
cont_tbl <- table(patient$treatment,patient$improvement)
cont_tbl
##
##               improved not-improved
##   not-treated       26           29
##   treated           35           15
```

 Perform Chi-Square test of independence. Turn off Yates' continuity correction by setting up `correct=FALSE` in the function.

```
#Yates' continuity correction is used if n < 40
chisq.test(cont_tbl,correct = FALSE)
##
##  Pearson's Chi-squared test
##
## data:  cont_tbl
## X-squared = 5.5569, df = 1, p-value = 0.01841
```

Conclusion and Interpretation

The chi-squared value is 5.5569 and the p-value (0.01841) is less than 0.05. Thus, we reject the null hypothesis and conclude that there is a relationship between the treatment and the improvement in health condition of patients.