Mathematical Statistics 2B

Practical 4: Parametric vs. Non-Parametric Tests

Department of Statistics - University of Johannesburg

1. Introduction

Parametric statistics are based on assumptions about the distribution of the population from which the sample was taken. For parametric tests, the data must meet the following requirements:

- Continuous variable of interest (interval or ratio)
- A random sample of data from the population
- Independent observations
- Normal distribution (approximately) of the sample and population on the variable of interest
- Homogeneity of variances (for two or more independent groups)
- No outliers

If the parametric assumptions are not met, we can use the corresponding nonparametric tests that do not rely on assumptions about the shape or parameters of the underlying population distribution. Nonparametric tests are typically used when the sample is not normally distributed, or the sample size is small, or the variables are measured on nominal or ordinal scales. It can sometimes be difficult to decide whether to use a parametric or nonparametric procedure. Nonparametric procedures generally have less power for the same sample size than the corresponding parametric procedure if the data truly are normal. Interpretation of nonparametric procedures can also be more difficult than for parametric procedures.

For this module we will look at how to perform and interpret parametric vs. non-parametric tests in R for one sample, paired samples, two independent samples, multiple independent samples, and correlation. The one-sample and two-sample hypothesis formulation in the notes are all given for two-sided tests. The form of the alternative hypothesis can be adjusted in the built-in functions to reflect a left-sided or a right-sided test, depending on the problem statement.

2. Datasets For Examples

The following datasets are used in this practical.

Create the following vectors/data frames in R

```
data1=c(84,86,90,97,98,98,99,102,102,103,108,111)
data2=c(91,95,96,96,96,97,98,101,103,105,106,123,130,143,145)
test_data=cbind(1:8,c(59,68,60,60,59,62,58,76),c(61,78,65,56,55,62,60,68))
colnames(test_data)=c("Student","Test1","Test2")
```

Create the *growth* dataset

A researcher wants to know if three different fertilizers lead to different levels of plant growth. A sample of 30 plants are randomly divided into three groups of 10 and a different fertilizer applied to each group. The height of each plant was measured at the end of one month. Use the following code to construct the dataset in R:

```
growth=data.frame(group=rep(c('A', 'B', 'C'), each=10),
      height=c(7, 14, 14, 13, 12, 9, 6, 14, 12, 8, 15, 17, 13, 15, 15, 13, 9, 12, 10, 8, 6, 8, 8, 9, 5, 14, 13, 8, 10, 9))
```

The *cars* dataset is available in R

The data give the speed of 50 cars and the distances taken to stop.

speed	Speed (mph)	dist	Stopping distance (ft)

The *mtcars* dataset is available in R

The data give fuel consumption and 10 aspects of automobile design and performance for 32 cars.

mpg	Miles/(US) gallon	qsec	1/4 mile time
cyl	Number of cylinders	vs	Engine (0 = V-shaped, 1 = straight)
disp	Displacement (cu.in.)	am	Transmission (0 = automatic, 1 = manual)
hp	Gross horsepower	gear	Number of forward gears
drat	Rear axle ratio	carb	Number of carburetors
wt	Weight (1000 lbs)		

3. One-Sample Tests

The one-sample *t*-test examines whether the mean of a population is statistically different from a known or hypothesized value. It is a parametric test, commonly used to test the statistical difference between a mean and a known or hypothesized value of the mean in the population, or the statistical difference between a change (difference) score and zero. It tests the hypothesis $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. The *t.test()* function in R is used to test if the mean is different from a specific value.

The non-parametric equivalent is called the one-sample Wilcoxon signed rank test, which tests hypotheses about the median of the population instead of the mean (as in the parametric test), namely $H_0 : m = m_0$ vs. $H_1 : m \neq m_0$. The *wilcox.test()* function in R is used to test if the median is different from a specific value.

For both the parametric and non-parametric test functions we will use the arguments *x* (variable of interest), *y* (second variable for paired data), *alternative* (to define the form), *mu* (if this is a number different from 0) and *paired* (to indicate a paired test). By default the *wilcox.test()* calculates an exact *p*-value if the samples contain less than 50 finite values and there are no ties. If there are ties in the data, you will see the warning message "cannot compute exact *p*-value with ties" as R adjusts the *p*-value by the continuity correction. In such a case we can use the option *correct*=FALSE to turn off the continuity correction, or the option *exact*=FALSE.

Example 1: One-sample (test_data)

The lecturer wants to test if the average/median mark for test 1 is significantly better than the average/median mark of last year, namely: mean = 56 and median = 55.

```
#check assumptions
> shapiro.test(test1)
        p-value = 0.009245
```

```
#One-sample t-test
> t.test(test1,mu=56,alternative = "greater")
p-value = 0.008946
```

```
# Wilcoxon signed rank test
>wilcox.test(test1,mu=55,alternative = "greater", exact=FALSE)
p-value = 0.007015
```

Example: Paired sample (test_data)

The lecturer wants to know if there is a significant difference in students' marks for test 1 compared to test 2.

```
#Dependent samples t-test (test1 mark minus test2 mark)
> t.test(test1,test2,paired = T)
p-value = 0.8578
mean difference: -0.375
```

```
# Wilcoxon signed rank test
> wilcox.test(test1,test2,paired = T,exact=FALSE)
p-value = 0.9324
```

I will leave it up to you to check the summary statistics and any other assumptions tests not included in the above.

4. Two Independent Sample Tests

The independent samples *t*-test is a parametric test used to test for mean differences between two independent groups. It tests the hypothesis $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$.

The *t.test()* function in R is used to compare the mean of a numerical variable of interest (dependent variable) between two independent groups (independent variable). The first argument of the function identifies the dependent and independent variables in the formula format, namely *dependent ~ independent*. The dataset is then defined using the *data* argument. The test assumes equality of variance and computes a pooled variance. If this assumption is not met, the test statistic calculation is adjusted for unequal variances, and the *var.equal* argument is set as FALSE (default *var.equal* = FALSE), otherwise it is set to TRUE for equal variances assumed.

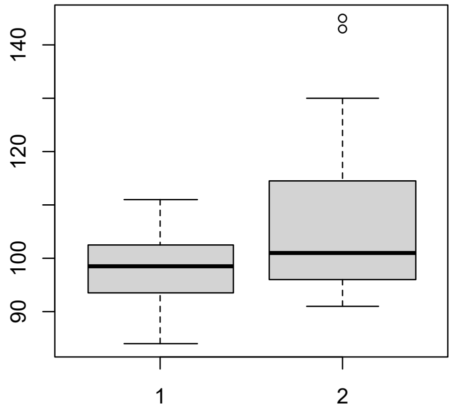
The non-parametric equivalent is called the Mann-Whitney U test, also referred to as the Mann Whitney Wilcoxon test or the Wilcoxon rank sum test, and is used to test whether there is a difference with respect to central tendency, specifically the median, between the two groups, namely $H_0 : m_1 = m_2$ vs. $H_1 : m_1 \neq m_2$. The *wilcox.test()* function in R is used for the Mann-Whitney U test. The measurements for the two groups can be defined in the *x* and *y* arguments, or in formula notation in the *x* argument. In the latter case, we also define the dataset using the *data* argument.

Example: Two independent samples (data1 and data2)

```
#summary statistics
data_summary=rbind(c(mean(data1),mean(data2)),c(median(data1),median(data2)),c(sd(data1),sd(data2)))
colnames(data_summary)=c("Data1","Data2")
row.names(data_summary)=c("Mean","Median","Standard deviation")
> data_summary
                    Data1       Data2
Mean                98.166667   108.3333
Median              98.500000   101.0000
Standard deviation  8.155682    17.9271
```

```
#check assumptions
> shapiro.test(data1)
        p-value = 0.6122
> shapiro.test(data2)
        p-value = 0.002517
> var.test(data1,data2)
        p-value = 0.0125
```

```
#box-and-whisker plot
> boxplot(data1,data2)
```



```
#Independent samples t-test
> t.test(data1,data2,var.equal = F)
p-value = 0.06406
```

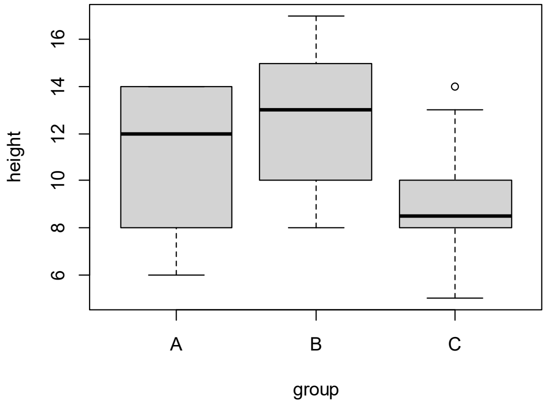
```
#Mann-Whitney U test
> wilcox.test(data1,data2,exact=FALSE )
p-value = 0.3924
```

5. Multiple Independent Sample Tests

The independent one-way analysis of variance (ANOVA) is a parametric test used to test for mean differences for three or more independent groups. It tests the hypothesis H_0: all group means are equal vs. H_1: at least one group has a different mean value. The *aov()* function in R is used to compare the mean of a numerical variable of interest (dependent variable) for three or more independent groups (independent variable). The first argument of the function identifies the dependent and independent variables in the formula format, namely *dependent ~ independent*. The dataset is then defined using the *data* argument. The non-parametric equivalent is called the Kruskal-Wallis test, and is used to test whether there is a difference in medians of the independent groups, namely H_0: all group medians are equal vs. H_1: at least one group has a different median value. The *kruskal.test()* function in R is used for the Kruskal-Wallis test. The dependent and independent variables are identified in formula notation.

Example: Three independent samples (growth data)

```
#box-and-whisker plot
boxplot(height~group,data=growth)
```



```
#check assumptions
> shapiro.test(growth$height[growth$group=="A"])
        p-value = 0.08727
> shapiro.test(growth$height[growth$group=="B"])
        p-value = 0.5756
> shapiro.test(growth$height[growth$group=="C"])
        p-value = 0.4242
> bartlett.test(height~group,data=growth)
        p-value = 0.9512
> fligner.test(height~group,data=growth)
        p-value = 0.838
```

```
#ANOVA
> anova_out=aov(height~group,data=growth)
> summary(anova_out)

        p-value = 0.0317
```

```
#Kruskal-Wallis
> KW_out=kruskal.test(height~group,data=growth)
> KW_out
        p-value = 0.04311
```

6. Correlation

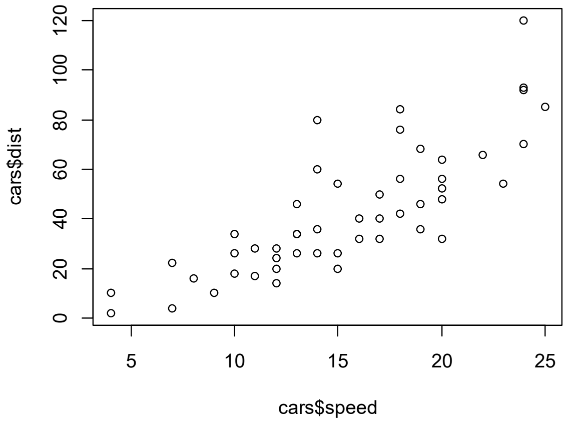
The most common correlation coefficient is Pearson's product moment correlation coefficient (or simply Pearson's correlation), denoted with *r*, which measures the strength of the linear relationship between two numerical variables. For Pearson's correlation, both variables should be normally distributed, have a linear relationship, be equally distributed about the regression line (homoscedasticity) and have no outliers.

Rank correlation methods such as Kendall's Tau and Spearman's Rho are non-parametric alternatives to Pearson's correlation and are based on the ranks of the data, rather than the observed values. While Pearson measures the linear relationship between two variables, Kendall and Spearman both measure the monotonic relationship. In a linear relationship the two variables move together at a constant rate (straight line). A monotonic relationship measures how likely it is for two variables to move in the same direction, but not necessarily at a constant rate. For example, the upward exponential curve $y = x^2, x > 0$ has a strictly positive monotonicity because as x increases y also increases, but the curved line is not linear since the rate changes in y vary at different values of x.

Kendall and Spearman can be used with ordinal data. If the Pearson and Spearman coefficients are not much different, the data tend to not have extreme values (outliers), but if they are very different, it is likely that outliers are present. The Kendall correlation is more robust and shows consistent values whether there are outliers or not. All three correlation measures will return a value between −1 and +1 and are interpreted in the same way. The *cor.test*() function in R is used to calculate Pearson, Kendal and Spearman correlation coefficients. The x and y arguments identify the two variables, and the specific coefficient is selected in the *method* argument (default method = "pearson".

Example: Correlation (cars data)

```
plot(cars$speed,cars$dist)
```



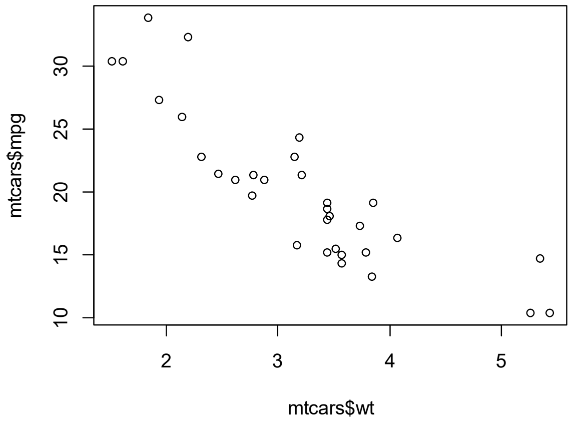
```
> cor(cars$speed,cars$dist)                          #Pearson's correlation
[1] 0.8068949

> cor(cars$speed,cars$dist,method = "spearman")      #Spearmans's Rho
[1] 0.8303568

> cor(cars$speed,cars$dist,method = "kendall")       #Kendall's Tau
[1] 0.6689901
```

Example: Correlation (mtcars data)

```
plot(mtcars$wt, mtcars$mpg)
```



```
> cor(mtcars$wt, mtcars$mpg)                               #Pearson's correlation
[1] -0.8676594

> cor(mtcars$wt, mtcars$mpg,method = "spearman")           #Spearmans's Rho
[1] -0.886422

> cor(mtcars$wt, mtcars$mpg,method = "kendall")            #Kendall's Tau
[1] -0.7278321
```