

Chapter 5

Data Analytics 3: Parametric Statistics

Prepared by Dr Yuslina Zakaria

In statistical analysis, the parametric statistical techniques require certain assumptions to be met before conducting the analysis. In this practical, you will be learning on how to perform preliminary tests to check the assumptions and hence, perform the parametric statistical tests. The lessons include:

1. Exploring assumptions (normality and homogeneity of variances)
2. Performing one sample t-tests
3. Comparing means between two groups and more than two groups

R Packages:

For this lesson, the following packages are used:

- dplyr
- car

Package Installation:

- `install.packages("car")`

Datasets

Four institutions i.e. Hungarian Institute of Cardiology, Zurich University Hospital, Basel University Hospital and V.A. Medical Center observed 76 features in 303 patients with and without heart disease. Here, we will be using the subset of original dataset consisting of 303 patients with 9 features set.



- `subs_heart.csv` (description: `subs_heart.txt`)

72 young anorexic female patients underwent three different possible therapies and their weights before and after therapies were recorded.

- `subs_anorexia.csv` (description: `subs_anorexia_desc.txt`)

Download the datasets (with their description files) from http://tiny.cc/phc410_da3 and place the files in the `input` folder, in your R project workspace.

Instructions:

-  explains the steps for activity you need to follow.
-  section contains practice questions for you to work on and submit as your lab report (Lab Report 3).

Load libraries

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

```
library(dplyr)
library(car)
```

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)
```

5.1 Exploring Assumptions (Normality)

Based on the central limit theorem, regardless of the data distributions, the sample distribution tends to be normal if the size of sample is more than 30 ($n > 30$).

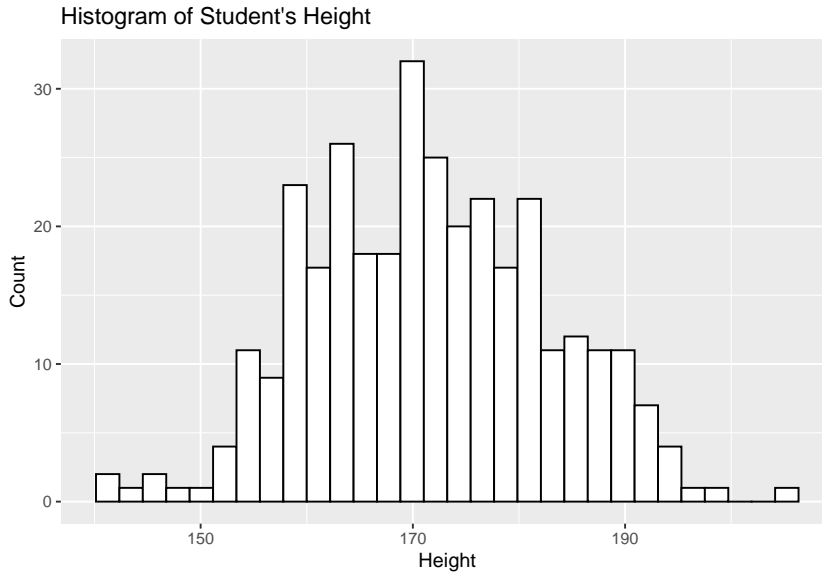
In this practical, the size of datasets that you are going to use is more than 30. Therefore, we can ignore the normality testing of data distribution (assume that it is normally distributed) and hence, use the parametric tests. However, it is beneficial to learn how we can assess the normality via visual inspection and significant tests.

For visual inspections, commonly used plots to observe normality of data distributions are histogram and normal probability quantile-quantile plot, or also known as Q-Q plot.

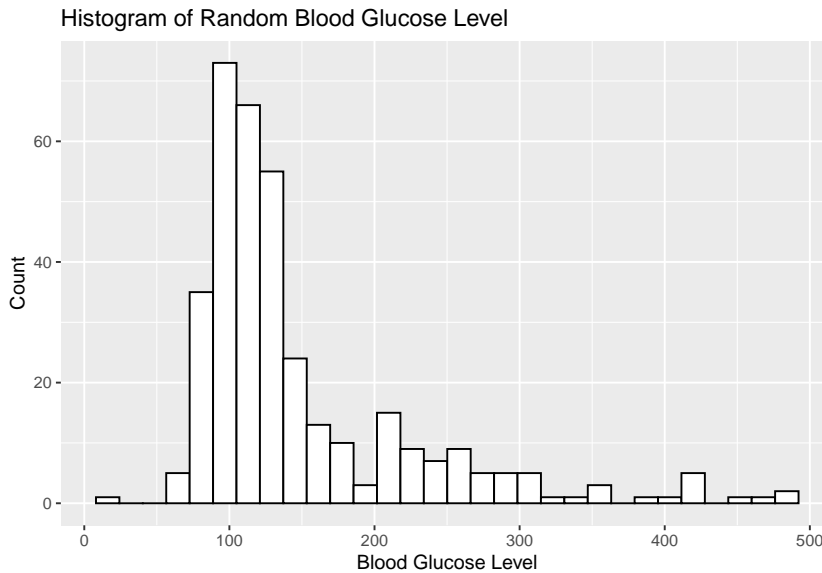
5.1.1 Histogram

We can tell if a distribution is approximately normal by visualising the distribution using a histogram. The graph must form a symmetric distribution or a bell-shaped curve for us to assume normality.

Example: Histogram for Normally Distributed Data (Student's Height)



Example: Histogram for Non-normally Distributed Data (Blood Glucose Level)



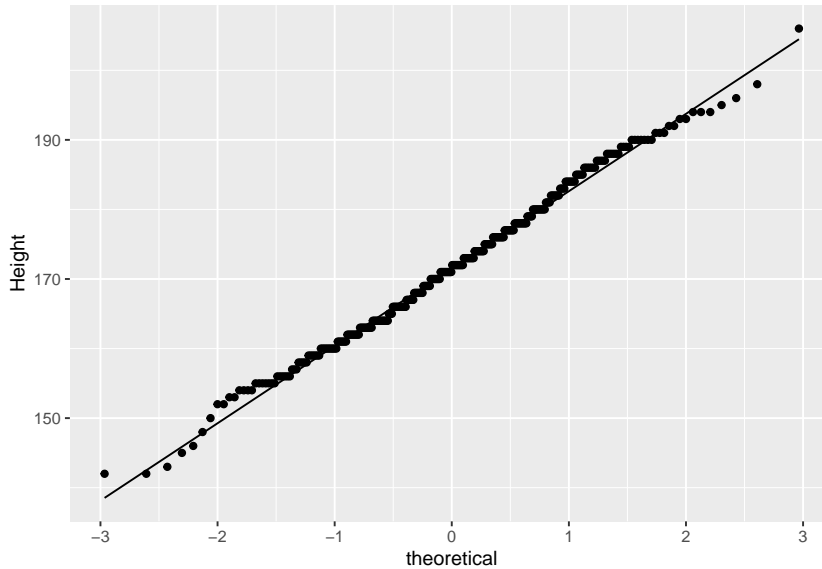
By observing the histogram, we may conclude that the variable is normally or not normally distributed. However, it is usually difficult to observe a bell-shaped distribution for a small size of samples. Thus, normal probability quantile-quantile plots (Q-Q plots) provide a more sensitive graphical technique for assessing normality.

5.1.2 Normal Probability Q-Q plot

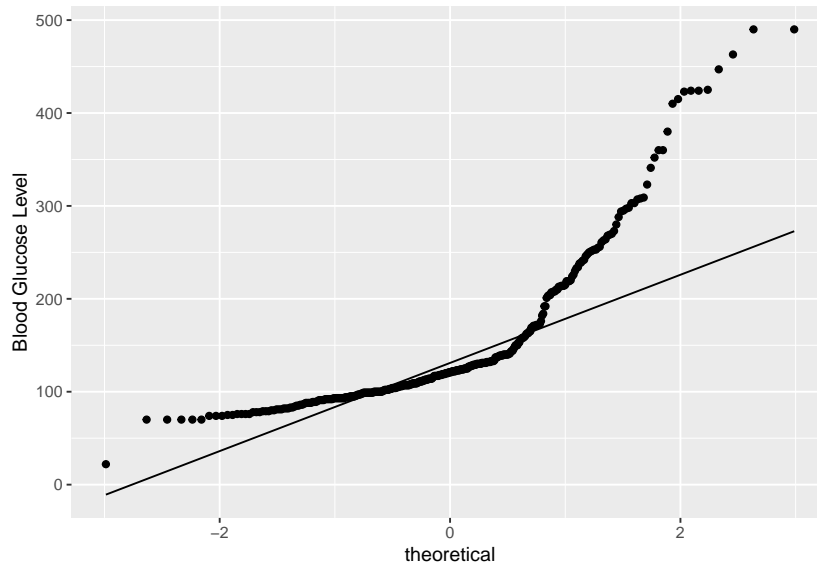
Normal probability Q-Q plot or quantile-quantile plots display the observed values against normally distributed data. If all the points fall approximately along the straight line, we can assume that the sample is from a normal distribution.

In R, we may apply `qqnorm` and `qqline` functions for plotting normal probability Q-Q plots.

Example: Q-Q plot for Normally Distributed Data (Student's Height)



Example: Q-Q plot for Non-normally Distributed Data (Blood Glucose Level)



5.1.3 Shapiro-Wilk

Using visual inspection, as described previously, is typically not very useful for small sample size data. Statistical significance tests that calculate actual probabilities to compare the sample distribution to a normal distribution are more precise.

There are several methods for normality test such as Shapiro-Wilk's (SW) normality test and Kolmogorov-Smirnov (KS) test. Note that, these significant tests can only be applied to continuous

variables. SW is more powerful than KS and it is widely recommended for assessing normality of small sample size (< 50 samples). In R, it can handle up to 5000 sample size. In this practical, only Shapiro-Wilk's test is covered.

The hypotheses used in normality test 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.

The significant p-value obtained from the test will be used to conclude if the null hypothesis is rejected ($p < 0.05$) and assume that the data is not normally distributed. Otherwise, $p > 0.05$ indicates the data is normally distributed.

To perform Shapiro-Wilk's test of normality in R we use:

```
shapiro.test()
```

Example: Output from Shapiro-Wilk's test for Normally Distributed Data (Student's Height)

```
##
##  Shapiro-Wilk normality test
##
## data:  muslim_s$height
## W = 0.99451, p-value = 0.2851
```

From the output, the $p\text{-value} > 0.05$ implies that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

Example: Output from Shapiro-Wilk's test for Non-normally Distributed data (Blood Glucose Level)

```
##
##  Shapiro-Wilk normality test
##
## data:  ckd$bgr
## W = 0.76786, p-value < 0.00000000000000022
```

The $p\text{-value} < 0.05$ implies that the distribution of the data are significantly different from normal distribution and thus, we assume that the data is not normally distributed. Hence, we should proceed with non-parametric statistical tests (will be covered in the next practical session).

Note that, normality test is sensitive to sample size and small samples most often pass normality tests. Therefore, it is important to combine visual inspection and significance test in order to take the right decision.

5.2 Exploring Assumptions (Homogeneity of Variance)

Some statistical tests require you to check the homogeneity of variance of your data distribution as one of the assumptions. Levene's test is commonly used to verify the assumption by measuring the equality of variances across samples.

In this practical, `leveneTest()` function from `car` package is used to assess whether the population variances are relatively equal.

To perform Levene's test in R we use:

```
library(car)
leveneTest()
```

5.3 One sample t-Tests

One sample t-test is used to determine whether the sample mean is statistically different from a known or hypothesized population mean (μ). If the aim is to measure any difference, regardless of the direction, a two-tailed hypothesis is used. If we want to measure both the difference and the direction, either a lower-tailed or upper-tailed hypothesis is used.

The corresponding null and alternative hypotheses of a two-tailed test:

H_0 : The sample mean is equal to the population mean ($\bar{x} = \mu$)

H_a : The sample mean is not equal to the population mean ($\bar{x} \neq \mu$)

For a one-tailed test, the null and alternative hypotheses are as follows:

- Upper-tailed
 H_0 : The sample mean is less than or equal to the population mean ($\bar{x} \leq \mu$)
 H_a : The sample mean is greater than the population mean ($\bar{x} > \mu$)
- Lower-tailed
 H_0 : The sample mean is greater than or equal to the population mean ($\bar{x} \geq \mu$)
 H_a : The sample mean is less than the population mean ($\bar{x} < \mu$)

For one sample t-test in R, the following `t.test()` is used:

```
# one sample t-test (two-tailed test)
t.test(y,mu=mean) # mu is population or hypothesized mean
```

You can use the `alternative="less"` or `alternative="greater"` option to specify a one tailed test.

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

Assumptions

The main three assumptions for one sample t-test are:

1. The variables should be continuous (interval/ratio)
2. The variables are independent of one another (no relationship between the observations).
3. The variables should be normally distributed.

Question

Suppose we wanted to test whether the mean age of male patient with heart disease in this sample is not equal to the population mean of 70.

Hypothesis

Therefore, our null and alternative hypotheses are:

H_0 : The mean age of male patients with heart disease is equal to 70 ($\bar{x} = 70$)

H_a : The mean age of male patients with heart disease is not equal to 70 ($\bar{x} \neq 70$)

Preparing the data

Let's proceed with analyzing the data set. Here, we are using the heart disease data set `subs_heart.csv`.



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

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

Since we are focusing on getting the mean of age for male patients with heart disease, we need to get the required subset of the data first.



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

```
male_hd <- subs_heart %>% filter(target=="presence" & sex=="male")
str(male_hd)
## 'data.frame': 93 obs. of 9 variables:
## $ age      : int  63 37 56 57 44 52 57 54 49 64 ...
## $ sex      : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
## $ cp       : Factor w/ 4 levels "asymptomatic",...: 1 3 2 4 2 3 3 4 2 1 ...
## $ trestbps : int  145 130 120 140 120 172 150 140 130 110 ...
## $ chol     : int  233 250 236 192 263 199 168 239 266 211 ...
## $ fbs      : Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 1 1 1 1 ...
## $ thalach  : int  150 187 178 148 173 162 174 160 171 144 ...
## $ exang    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 2 ...
## $ target   : Factor w/ 2 levels "absence","presence": 2 2 2 2 2 2 2 2 2 2 ...
```

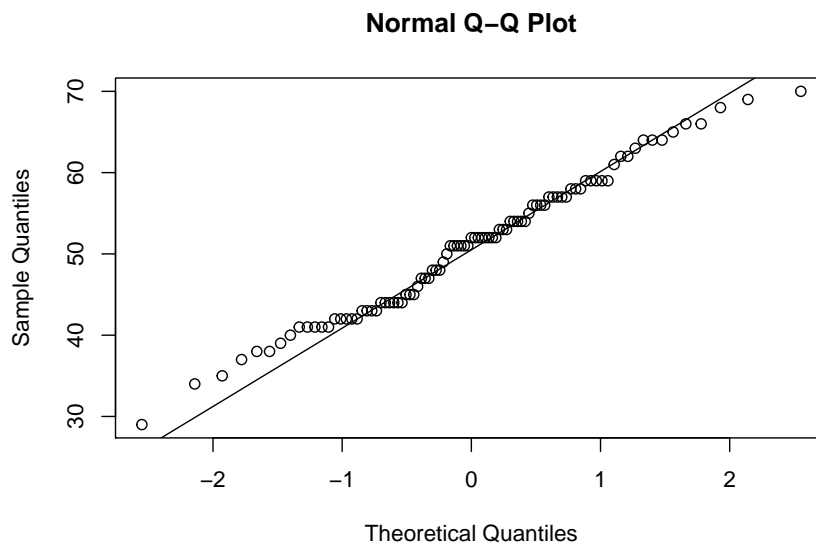
`male_hd` contains 93 male patients with heart disease ($n=93$) with 9 features. Although it is unnecessary to assess the normality of distribution due to the large sample size (big data), let us just take a look at how we can assess normality of the data.

Checking the assumptions



Check the normality using Q-Q plot

```
qqnorm(male_hd$age)
qqline(male_hd$age)
```



From the Q-Q plot, we can see that the points fall approximately lie on the straight line (normality assumed).

Alternatively, we may conduct a formal significant test, the Shapiro-Wilk test, to see whether the `age` of this sample come from a normal distribution.

Check normality using Shapiro-Wilk.

```
shapiro.test(male_hd$age)
##
##  Shapiro-Wilk normality test
##
## data:  male_hd$age
## W = 0.98616, p-value = 0.4359
```

A significant p-value of 0.4359 (> 0.05) means that we failed to reject the null hypothesis and conclude that the data is normally distributed.

Performing the test

Now, let us proceed to the one sample t-test.

Perform one sample t-test using hypothesized population mean of 70.

```
t.test(male_hd$age,mu=70)
##
##  One Sample t-test
##
## data:  male_hd$age
## t = -21.21, df = 92, p-value < 0.0000000000000022
## alternative hypothesis: true mean is not equal to 70
## 95 percent confidence interval:
##  49.11500 52.69145
## sample estimates:
## mean of x
##  50.90323
```

Conclusion and Interpretation

The point estimate of the sample mean is 50.9, and the 95% confidence interval is from 49.12 to 52.69. The hypothesis testing p-value is smaller than 0.05 which means the difference between the means is significant (reject null hypothesis).

The output indicates that the mean age of male patients with heart disease is not equal to 70. In other way, there are a significant difference in age between the sample mean and the population mean. That is, based on the sample mean, the male patients having heart disease in the sample seem to have age younger than those in the population $t(92)=-21.21$, $p<0.05$.

5.4 Comparing Means of Two Groups

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

5.4.1 Independent Samples t-Test

The independent samples t-test is used to compare the means between two independent groups. In R, `t.test()` function is used to compare independent groups. :

```
# independent 2-group t-test (default for unequal variance)
t.test(y~x) # y is numerical variable and x is categorical variable

# independent 2-group t-test with equal variances
t.test(y~x, var.equal=TRUE) # y is numerical variable and x is categorical variable
```

Assumptions

The independent sample t-test has two additional assumptions:

1. The observation or participant should only be measured in only one group and the groups are unrelated.
2. Homogeneity of variance - equal variances for each of the two groups.

`var.equal=T` is used to assume that the data has equal variances. By default, unequal variances assumed (i.e. `var.equal=F`)

Question

Suppose we wanted to test whether the mean age of female patients differs from the mean age of male patients.

Hypothesis

H_0 : The mean age of female patients is equal to the mean age of male patients.

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

Checking the assumptions

To check for equal variances, `leveneTest()` function from `car` package is used.

Check the equality of variances using Levene's test.

```
leveneTest(subs_heart$age ~ subs_heart$sex)
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value Pr(>F)
## group    1    0.363 0.5473
##           301
```

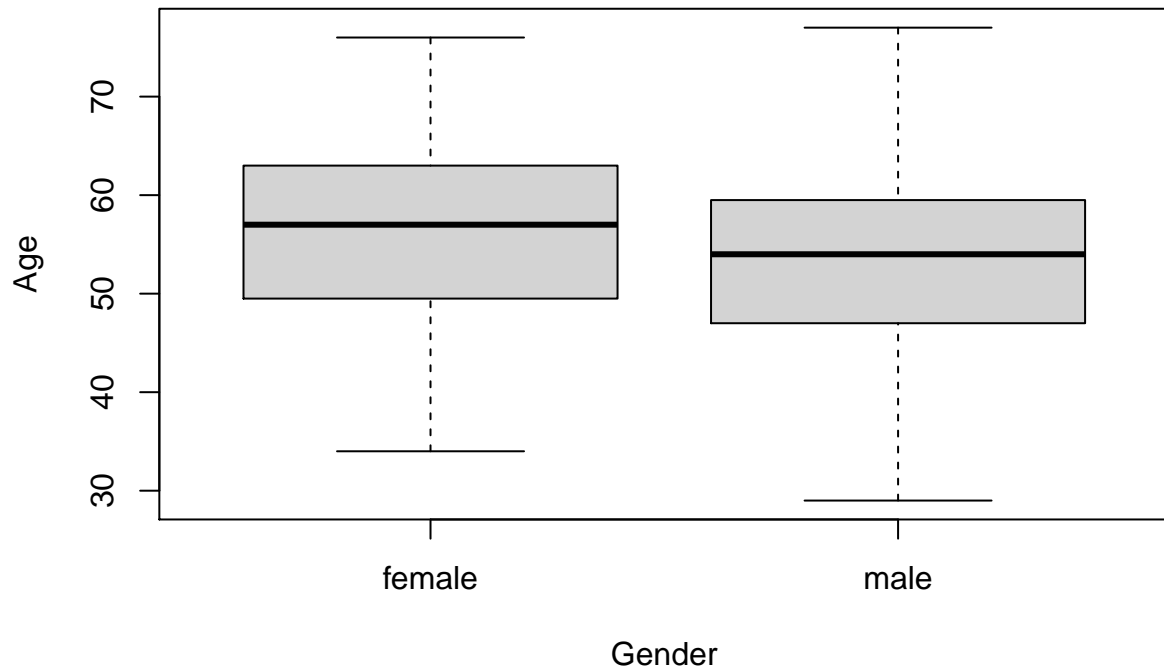
Conclusion and Interpretation

The p-value of the test is 0.5473, which is greater than our significance level of 0.05. Thus, we fail to reject the null hypothesis and conclude that the variances between the groups are equal.

We can also assess the variance equality from the length of the boxplot.

Using boxplot to estimate the variance

```
boxplot(subs_heart$age ~ subs_heart$sex, xlab="Gender", ylab="Age")
```



From the boxplot, we can see the length of box between female and male is almost equal. This confirms that the population has equal variances.

Performing the test

Perform independent t-test with equal variances

```
#between age for female and male patients
t.test(subs_heart$age ~ subs_heart$sex, var.equal=T )
##
## Two Sample t-test
##
## data: subs_heart$age by subs_heart$sex
## t = 1.7163, df = 301, p-value = 0.08713
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.2812059 4.1184643
## sample estimates:
## mean in group female mean in group male
## 55.67708 53.75845
```

Conclusion and Interpretation

The p-value $0.08 > 0.05$ - thus we fail to reject H_0 . Therefore, the mean age of female patients is equal to the mean age of male patients. In other way, there is no significant difference in the mean age between male and female patients, $t(301)=1.7163$, $p>0.05$.

Now let us compare the mean of age between patients with and without heart disease.

Question

Suppose we wanted to test whether the mean age of patients with heart disease differs from the mean age of patients without heart disease.

Hypothesis

H_0 : The mean age of patients with heart disease is equal to the mean age of patients without heart diseases.

H_a : The mean age of patients with heart disease is not equal to the mean age of patients without heart diseases.

Checking the assumptions

Estimate the variance using Levene's test

```
#between age for disease and without disease
leveneTest(subs_heart$age ~ subs_heart$target)
```

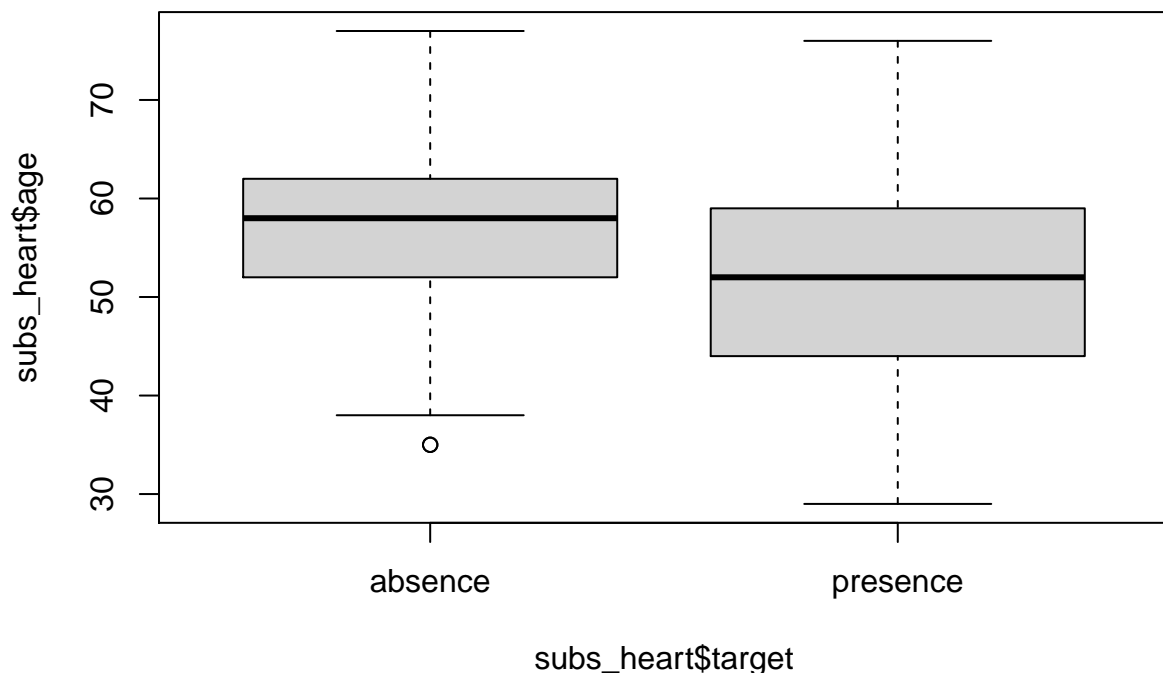
```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value  Pr(>F)
## group 1  7.9854 0.005031 **
##      301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion and Interpretation

The p-value of the test is 0.005031, which is smaller than our significance level of 0.05. Thus, we reject the null hypothesis and conclude that the variances between the groups are not equal (unequal variances assumed).

Estimate the variance using the boxplot

```
boxplot(subs_heart$age ~ subs_heart$target)
```



Obviously, it can be seen the length of boxplot between patients without heart disease (**absence**) and patients with heart disease (**presence**) is different.

Performing the test

Let us proceed with the independent t-test for unequal variances

Independent t-test for unequal variances

```
t.test(subs_heart$age ~ subs_heart$target, var.equal=F)
##
## Welch Two Sample t-test
##
## data: subs_heart$age by subs_heart$target
## t = 4.0797, df = 301, p-value = 0.00005781
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.124635 6.084324
## sample estimates:
## mean in group absence mean in group presence
##           56.60145           52.49697

#or
#t.test(subs_heart$age ~ subs_heart$target)
```

Conclusion and Interpretation

The p-value $0.00005781 < 0.05$ - thus reject H_0 . There is a significant difference in the mean age between patients with and without the heart disease. $t(301)=4.0797$, $p<0.05$.

5.4.2 Paired Samples t-Test / Repeated Measures t-Test

Two data samples are matched if they come from repeated observations of the same subject. The paired t-test has the same assumptions of independence and normality as a one sample t-test.

To perform a paired samples t-test:

```
# paired t-test
t.test(y1,y2,paired=TRUE) # y1 & y2 are numerical variables
```

In this section, let us look at a data set on weight change of 72 female anorexic patients in `subs_anorexia.csv`.

Question

Here we are interested in finding out whether there is a placebo effect for patients who do not get treated (Control) to gain some weight in the study.

Hypothesis

H_0 : There is no difference in mean weight before and after the treatment for those in the Control group.

H_a : There is a difference in mean weight before and after the treatment for those in the Control 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)
```

Since the question asks us to compare the findings for patients who do not get treated, we need to get the subset of `subs_anrx` which only has `Cont` for treatment.

Filter to get observations of Control subject (`Cont`) with `cont_subs`

```
cont_subs <- subs_anrx %>% filter(Treat=="Cont")
```

Performing the test

Assuming that the data is normally distribution, proceed to the paired t-test.

Perform the paired t-test

```
t.test(cont_subs$Postwt, cont_subs$Prewt, paired=TRUE)
##
## Paired t-test
##
## data: cont_subs$Postwt and cont_subs$Prewt
## t = -0.28723, df = 25, p-value = 0.7763
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.676708 2.776708
## sample estimates:
## mean of the differences
## -0.45
```

Conclusion and Interpretation

Based on the results, we fail to reject the null hypothesis ($t = -0.28723$, $df = 25$, $p\text{-value} = 0.7763$) as $p\text{-value} > 0.05$. There is no difference in mean weight before and after the treatment in the Control group, $t(25)=-0.28723$. $p>0.05$.

5.5 Comparing Means of More Than Two Groups

For comparing means between more than two groups, one-way analysis of variance (ANOVA) is used.

Assumptions

The assumptions of for ANOVA are the same as those for the t-test:

1. The variables should be normally distributed.
2. Homogeneity of variance.

Question

We would like to compare whether the maximum heart rate (**thalach**) of patients differs between four different types of chest pain (**cp**).

Hypothesis

H_0 : The means of maximum heart rate of the four different types of chest pain are the same.

H_a : At least one sample mean of maximum heart rate is not equal to the others.

We will be using the same **subs_heart** object which we have used in the previous one sample and independent sample t-tests.

Checking the assumptions

Since $n>30$, let us proceed to the Levene's test to verify homogeneity of variance.

Perform the Levene's test

```
leveneTest(subs_heart$thalach ~ subs_heart$cp)
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 3    2.46 0.06286 .
```

```
##          299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion and Interpretation

The p-value of the test is 0.06286 ($p > 0.05$) suggests that the assumption for homogeneity of variances has not been violated. We can conclude that the variances between the groups are equal and therefore, we can proceed with one-way ANOVA.

Performing the test

Perform one-way ANOVA

```
heart_aov <- aov(subs_heart$thalach ~ subs_heart$cp)
summary(heart_aov)
##          Df Sum Sq Mean Sq F value    Pr(>F)
## subs_heart$cp    3   24030     8010   17.82 0.000000000115 ***
## Residuals      299  134413      450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion and Interpretation

The F-ratio (17.82) with an F-probability value (0.000) of less than 0.05 is significant. Thus, we would reject the null hypothesis. We believe that the type of chest pain does significantly influence the maximum heart rate.

5.6 Practice



Data set:

Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels. This medical records data set consists of 70,000 patients aged from 29 to 65 years old. Each record is provided with 12 features with four objective input features i.e. age, gender, height, weight; four examination features i.e. systolic and diastolic blood pressures, cholesterol and glucose levels; and four subjective features i.e. smoking, alcohol intake, physical activity and whether the patient has CVD. The data is recorded in `cvd_train.csv` and the description of each variable is explained in `cvd_train_desc.txt`

Import the data using the following R command:

```
cvd <- read.csv("input/cvd_train.csv", sep=";")
```

Assume that the data are normally distributed, answer the following questions:

1. Using the description in `cvd_train_desc.txt`, recode `cardio`. Don't forget to load required libraries.
2. Does the systolic blood pressure differ between patients with and without cardiovascular disease?
 1. State the appropriate statistical test.
 2. Perform Levene's test, the suitable parametric test and interpret your results.