

VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING



Assignment for

“Probability and Statistics”

Advisor: Mr. Nguyen Tien Dung
Students: Tran Nguyen Anh Khoa - 1911419
Le Quoc Anh - 1852006
Tran Nguyen Phuoc Nhan - 1952893
Ta Minh Huy - 1952268

HO CHI MINH CITY, MAY 2021



Contents

1	Project I - Topic 3	2
1.1	Problem 1	2
1.1.1	Theoretical basic	2
1.1.2	Implementation	2
1.1.3	Implementation using R-Studio	3
1.2	Problem 2	3
1.2.1	Theoretical basic	4
1.2.2	Implementation	4
1.3	Problem 3	5
1.3.1	Theoretical basic	5
1.3.2	Implementation	5
1.4	Problem 4	6
1.4.1	Theoretical basic	7
1.4.2	Implementation	8
2	Project II - Topic 3	11
2.1	Data description	11
2.2	Importing data	11
2.3	Data cleaning	11
2.4	Data visualization	12
2.4.1	Descriptive statistics for each variable	12
2.4.2	Box plot	13
2.5	t.test: between pre.weight and weight6weeks	15
2.5.1	Theoretical basic	15
2.5.2	Implementation	15
2.6	One way ANOVA: What is the best diet for weight loss?	16
2.6.1	Implementation of One Way ANOVA test in R	16
2.6.2	Conclusion	18
2.7	Two way ANOVA : How do <i>Diet</i> and <i>gender</i> affect <i>weightLOST</i> ?	19

1 Project I - Topic 3

1.1 Problem 1

Question: Trachoma has 4 stages T1, T2, T3 and T4. The results of trachoma examination in 3 provinces A, B, C are given in the following table:

	Stages			
	T1	T2	T3	T4
A	47	189	807	1768
B	53	746	1387	946
C	16	228	438	115

1.1.1 Theoretical basic

Type of problem: testing for goodness of fit using Chi-squared Test.

What is Chi-squared: A chi-square (χ^2) statistic is a test that measures how a model compares to actual observed data. The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample. For example, the results of tossing a fair coin meet these criteria.

Chi-square tests are often used in hypothesis testing. The chi-square statistic compares the size any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship. For these tests, degrees of freedom are utilized to determine if a certain null hypothesis can be rejected based on the total number of variables and samples within the experiment. As with any statistic, the larger the sample size, the more reliable the results.

1.1.2 Implementation

Assumption:

- H_0 : The distribution of the stages of trachoma is the same in the three provinces.
- H_1 : The distribution of the stages of trachoma is different in the three provinces.

Calculating expected values:

	Stages				Total
	T1	T2	T3	T4	
A	48.3792	485.0435	1097.7080	1197.8693	2811
B	53.9039	540.4326	1223.0599	1314.6036	3132
C	13.7169	137.5239	311.2320	334.5272	797
Total	116	1163	2632	2829	6740

Calculating degree of freedom:

$$df = (4 - 1)(3 - 1) = 6$$

Applying formula for to calculate Chi-square statistic:

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Where: χ^2 = Chi-squared, $O_{i,j}$ = Observed value, $E_{i,j}$ = Expected value.

$$\chi^2 = \frac{(47 - 48.3792)^2}{48.3792} + \frac{(189 - 485.0435)^2}{485.0435} + \dots + \frac{(115 - 334.5272)^2}{334.5272} = 1010.0166$$

We have $\chi^2_{0.01,6} = 16.81 < 1010.0166$ out of accepted region so we reject H_0 and accept H_1 .

Conclusion: the distribution of the stages of trachoma is different in the three provinces.

1.1.3 Implementation using R-Studio

Initializing matrix: We give the matrix necessary numbers and change the column's name and row's name of the matrix.

```
1 data = matrix(c(47, 53, 16, 189, 746, 228, 807, 1387, 438, 1768, 946, 115), ncol =  
2 4, nrow = 3)  
3 colnames(data) = c("T1", "T2", "T3", "T4")  
4 rownames(data) = c("A", "B", "C")
```

The matrix after initialization:

```
1   T1  T2  T3  T4  
2 A 47 189 807 1768  
3 B 53 746 1387 946  
4 C 16 228 438 115
```

Using **chisq.test()** function to get result: We use **chisq.test()** function to get χ^2 , degree of freedom, p-value out of the matrix.

```
1 chisq.test(data)
```

The results of the function:

```
1  
2 Pearson's Chi-squared test  
3  
4 data: data  
5 X-squared = 1010, df = 6, p-value < 2.2e-16
```

We have $p\text{-value} < 2.2e-16 < 0.01$ so we reject H_0 and accept H_1 .

Conclusion: the distribution of the stages of trachoma is different in the three provinces.

1.2 Problem 2

Question: The following table shows data on the number of cancer deaths in the United States, Japan and the UK. Cancers are classified by the type of tissues in which the cancer originates. Compare the cancer death rate in the three countries at the significance level of $\alpha = 1\%$

Types of cancer	Countries		
	US	Japan	UK
Colorectal cancer	11	5	5
Breast cancer	15	3	7
Stomach cancer	3	22	3
Other types of	41	30	15

1.2.1 Theoretical basic

Type of problem: *Two-way ANOVA*.

In this problem, we have to determine if two independent variables, which are Countries and Types of cancers have the effect on the dependent variable which is the cancer death rate. Therefore, we will use the **Two-way ANOVA** method to implement the checking step.

1.2.2 Implementation

Assumptions:

- H_0 : The distribution of cancer death rate is the same in three countries.
- H_1 : The distribution of cancer death rate is different in three countries.

Implementation on R-Studio:

```
1 ##generating data
2 numCase <- c(11, 5, 5,
3             15, 3, 7,
4             3, 22, 3,
5             41, 30, 15
6             )
7 countries <- rep(c("US", "Japan", "UK"), 4)
8 typeOfCancer <- c(rep("Colorectal", 3), rep("Breast", 3), rep("Stomach", 3), rep("Other",
9             3))
10 DeathRate <- data.frame(typeOfCancer, countries, numCase)
11
12 ##displaying result
13 result = aov(numCase~typeOfCancer+countries, data = DeathRate)
14 anova(result)
```

Result we get is:

Generating data:

```
1   typeOfCancer countries numCase
2 1 Colorectal      US      11
3 2 Colorectal     Japan      5
4 3 Colorectal      UK      5
5 4 Breast         US     15
6 5 Breast         Japan      3
7 6 Breast         UK       7
8 7 Stomach        US       3
9 8 Stomach        Japan     22
10 9 Stomach        UK       3
11 10 Other         US     41
12 11 Other         Japan     30
13 12 Other         UK      15
```

Summary of ANOVA:

1	Analysis of Variance Table												
2													
3	Response: numCase												
4		Df	Sum Sq	Mean Sq	F value	Pr(>F)							
5	typeOfCancer	3	948.67	316.222	4.0950	0.06705	.						
6	countries	2	216.67	108.333	1.4029	0.31634							
7	Residuals	6	463.33	77.222									
8	---												
9	Signif. codes:	0	***	0.001	**	0.01	*	0.05	.	0.1	1		

Find the F value and compare:

With the significant level of 1% we have to calculate the value of quantile function over F distribution with R-Studio implementation:

```
> qf(0.99, 2, 6)
10.92477
```

Conclusions: With $F = 1.4029 < F_{0.01,2,6} = 10.92477$. In addition, $P = 0.31634 > \alpha = 0.01$. Therefore, we fail to reject H_0 and there is no significant different between mean death rate in the three countries.

1.3 Problem 3

Question: The sales of 4 stores of a company (million/month) are give in the following table:

Month	Stores			
	1	2	3	4
1	12.3	14.2	15.6	17.2
2	12.6	12.4	17.2	15.8
3	11.6	11.5	18.2	12.2
4	15.2	11.6	12.5	
5	18.6		11.8	
6	17.1			

At the significance level of $\alpha = 5\%$, compare the sales of these stores.

1.3.1 Theoretical basic

Type of problem: one-way analysis of variance (ANOVA).

What is One-way ANOVA: One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. One-Way ANOVA is a parametric test.

1.3.2 Implementation

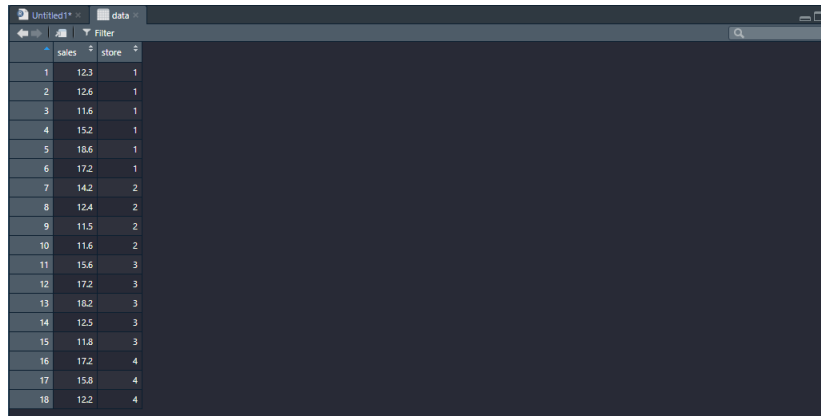
Assumption:

H_0 : The revenue of sale/month is the roughly same for 4 given stores.

H_1 : There are at least 2 stores have a significant difference in sale/month revenue.

Initializing matrix: Import necessary data into a csv file and import it into code.

```
1 library(readxl)
2 data <- read_excel("C:/Users/xps/Desktop/data.xlsx")
3 View(data)
```



	sales	store
1	12.3	1
2	12.6	1
3	11.6	1
4	15.2	1
5	18.6	1
6	17.2	1
7	14.2	2
8	12.4	2
9	11.5	2
10	11.6	2
11	15.6	3
12	17.2	3
13	18.2	3
14	12.5	3
15	11.8	3
16	17.2	4
17	15.8	4
18	12.2	4

Figure 1: The data after being imported

Using `aov()` function to conduct the one-way ANOVA analysis on the given data set.

```
1 data$store = as.factor(data$store)
2 aov = aov(data=data, formula = sales~store)
3 summary(aov)
```

Before run the `aov()` function we have to set the `aov()` variable as a factor by using `aov()` for conversion, because the problem statement wants us to compare the sales according to each store. After running the `aov()` to conduct the analysis, we store our output values in to a variable in R (in this case I'll name it `aov()`), then I'll use the `summary()` function to give us a summary chart of our results as below:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
store	3	19.19	6.397	0.981	0.43
Residuals	14	91.31	6.522		

Finally, we compare the output P-value = 0.43 (the “Pr(>F)” value in the chart). As we compared with the 5% significance given from the statement , P-value > Alpha-value (0.43 > 0.05) we failed to reject H_0 .

Conclusion: There is no significant difference between the sales revenue the stores.

1.4 Problem 4

Question: At the level of significance $\alpha = 5\%$, compare the business performance of some industries in the four urban districts on the basis of the sales of some stores given in the following table.

	Districts			
	1	2	3	4
Refrigeration	2.5, 2.7, 2.0, 3.0	13.1, 3.5, 2.7	2.0, 2.4	5.0, 5.4
Construction Materials	0.6, 10.4	15.0	9.5, 9.3, 9.1	19.5, 17.5
Computer Services	1.2, 1.0, 9.8, 1.8	2.0, 2.2, 1.8	1.2, 1.3, 1.2	5.0, 4.8, 5.2

1.4.1 Theoretical basic

Type of problem: testing the performance of some industries on the basis of sales (with replicates type of two-way ANOVA).

What is two-way ANOVA: this method used for comparing the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

An experiment that utilizes every combination of factor levels as treatments is called a factorial experiment.

In a factorial experiment with factor A at a levels and factor B at b levels, the model for the general layout can be written as:

$$Y_{ij} = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

for $i = 1, 2, 3, \dots, A$
for $j = 1, 2, 3, \dots, B$
 $k = 1, 2, 3, \dots, r$

where μ is the overall mean response, τ_i is the effect due to the i -th level of factor A, β_j is the effect due to the j -th level of factor B and γ_{ij} is the effect due to any interaction between the i -th level of A and the j -th level of B.

When an $a \times b$ factorial experiment is conducted with an equal number of observations per treatment combination, the total (corrected) sum of squares is partitioned as:

$$SS(total) = SS(A) + SS(B) + SS(AB) + SSE$$

where AB represents the interaction between A and B.

For reference, the formulas for the sums of squares are:

$$SS(A) = rb \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SS(B) = ra \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SS(AB) = r \sum_{j=1}^b \sum_{i=1}^a (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$SSE = \sum_{k=1}^r \sum_{j=1}^b \sum_{i=1}^a (y_{ijk} - \bar{y}_{ij.})^2$$

Source	SS	df	Mean Square
Factor A	SS(A)	(a-1)	SS(A)/(a-1)
Factor B	SS(B)	(b-1)	SS(B)/(b-1)
Interaction	SS(AB)	(a-1)(b-1)	SS(AB)/((a-1)(b-1))
Error	SSE	(N-ab)	SSE/(N-ab)
Total	SS(Total)	(N-1)	

$$SS(Total) = \sum_{k=1}^r \sum_{j=1}^b \sum_{i=1}^a (y_{ijk} - \bar{y}_{...})^2$$

The various hypotheses that can be tested using this ANOVA table concern whether the different levels of Factor A, or Factor B, really make a difference in the response, and whether the AB interaction is significant.

1.4.2 Implementation

Assumption:

H_A : The Sales of those Industries are the same.

H_B : The Sales of those Districts are the same.

H_{AB} : The Sales of Industries and Districts have a interaction

Create 3 vector for three different industries, Districts, and Sales by using `gl(n, k, length, labels)` function:

- n is the range number of factor
- k is number of repeat of that factors
- length is the number of factor
- labels is the name of each factor

Create data table by syntax `data.frame`:

```
1 Performance <- data.frame(Industries, Districts, Sales)
```

Using instruction `aov` to analyze variance and then assign it to “Analysis” variable.

```
1 Analysis <- aov(Sales ~ Industries + Districts + Industries*Districts)
```

Step 4:

Using `summary` to perform the ANOVA table **Step 5:**

Comparing with f

We have: $r = 3, c = 4, n = 48, \alpha = 0.05$

$f(r-1, n-cr) = f(2, 36) = 3.259446$

$f(c-1, n-cr) = f(3, 36) = 2.866266$

$f((r-1)(c-1), n-rc) = f(6, 36) = 2.363751$

We have 3 value of f due to instruction $1 - \alpha, k_1, k_2$ The final result is

$$F_A = 2.149 < f(2, 36) = 3.259446 \Rightarrow H_A \text{ is acceptable}$$

	Industries	Districts	Sales
1	Refrigeration	1	2.5
2	Refrigeration	1	2.7
3	Refrigeration	1	2.0
4	Refrigeration	1	3.0
5	Refrigeration	2	13.1
6	Refrigeration	2	3.5
7	Refrigeration	2	2.7
8	Refrigeration	2	0.0
9	Refrigeration	3	2.0
10	Refrigeration	3	2.4
11	Refrigeration	3	0.0
12	Refrigeration	3	0.0
13	Refrigeration	4	5.0
14	Refrigeration	4	5.4
15	Refrigeration	4	0.0
16	Refrigeration	4	0.0
17	Construction Materials	1	0.6
18	Construction Materials	1	10.4
19	Construction Materials	1	0.0
20	Construction Materials	1	0.0
21	Construction Materials	2	15.0
22	Construction Materials	2	0.0
23	Construction Materials	2	0.0
24	Construction Materials	2	0.0

Figure 2: The performance table

```
> summary(Analysis)
              Df Sum Sq Mean Sq F value Pr(>F)
Industries      2  103.1    51.57   2.149   0.131
Districts       3   41.3    13.78   0.574   0.636
Industries:Districts  6  117.6    19.60   0.817   0.564
Residuals      36  864.0    24.00
>
```

Figure 3: The result of ANOVA

```
>
> qf(1-0.05, 2, 36)
[1] 3.259446
> qf(1 - 0.05, 3, 36)
[1] 2.866266
> qf(1 - 0.05, 6, 36)
[1] 2.363751
```



$$F_B = 0.574 < f(3, 36) = 2.866266 \Rightarrow H_B \text{ is acceptable}$$

$$F_{AB} = 0.817 < f(6, 36) = 2.363751 \Rightarrow H_{AB} \text{ is acceptable}$$

Therefore, the business performance between 3 Industries and also Districts are the same. There is no interaction between Industries and Districts.

2 Project II - Topic 3

2.1 Data description

The given data set **Diet.csv** contains information on 78 people using one of three diets (The University of Sheffield). Attribute information:

- *Person* : Participant - number
- *gender* : Gender (1 = male, 0 = female) - Binary
- *Age* : Age (years) - Scale
- *Height* : Height (cm) - Scale
- *preweight* : Weight before the diet (kg) - Scale
- *Diet* : Diet
- *weight6weeks* : Weight after 6 weeks (kg) - Scale
- *weightLOST* : Weight lost after 6 weeks (kg) - Scale

2.2 Importing data

This section includes setting up directory, then we read the dataset using *readr* library.

The implementation in R-Studio:

For setting up directory:

```
1 setwd("directory e.g: D:\\")
2 getwd()
```

For importing dataset:

```
1 library(readr)
2 diet = read.csv("Diet.csv", row.names = 1)
```

Also, we will refactor "gender" and "age" to label "Female", "Male" and label "A", "B", "C" respectively.

2.3 Data cleaning

Due to lacking of information, it is necessary to remove some rows which contain NA sections. At this time we will calculate the $Weight.lost = pre.weight - weight6weeks$ column for further uses.

Implementation on R-Studio to remove "NA" sections:

```
1 ####data cleaning (NA)
2 diet = na.omit(diet)
```

After removing all "NA" sections, the dataset remain 76 rows.

2.4 Data visualization

2.4.1 Descriptive statistics for each variable

To find min, median, first-quantile, third quantile and max respectively, we use `summary()` function to report in the above order. Here is the implementation and result in R-Studio.

```
1 #for descriptive statistic info
2 print(summary(diett))
```

Then we obtain the result:

```
1      gender      Age      Height      pre.weight
2 Female:43   Min.    :16.00   Min.    :141.0   Min.    :58.00
3 Male  :33   1st Qu.:32.50   1st Qu.:163.8   1st Qu.:66.00
4           Median :39.00   Median :169.0   Median :72.00
5           Mean   :39.22   Mean   :170.8   Mean   :72.29
6           3rd Qu.:47.25   3rd Qu.:175.2   3rd Qu.:78.00
7           Max.    :60.00   Max.    :201.0   Max.    :88.00
8
9      Diet      weight6weeks      Weight.lost      diet_type
10 Min.    :1.000   Min.    :53.00   Min.    : -2.100   A:24
11 1st Qu.:1.000   1st Qu.:61.95   1st Qu.:  2.300   B:25
12 Median :2.000   Median :68.95   Median :  3.700   C:27
13 Mean   :2.039   Mean   :68.34   Mean   :  3.946
14 3rd Qu.:3.000   3rd Qu.:73.67   3rd Qu.:  5.650
15 Max.    :3.000   Max.    :84.50   Max.    :  9.200
```

Also, other information was reported, too. Including *Variance*, *Standard Deviation* and *Standard Error Mean*.

Implementation on R-Studio:

```
1 var(diett$weight6weeks)
2 sd(diett$weight6weeks)
3 sd(diett$weight6weeks)/sqrt(length(diett$gender))
4 var(diett$pre.weight)
5 sd(diett$pre.weight)
6 sd(diett$pre.weight)/sqrt(length(diett$gender))
7 var(diett$pre.weight)
8 sd(diett$Weight.lost)
9 sd(diett$Weight.lost)/sqrt(length(diett$gender))
```

Then we obtain the result of:

```
1 > var(diettt$weight6weeks)
2 [1] 64.94649
3 > sd(diettt$weight6weeks)
4 [1] 8.058938
5 > sd(diettt$weight6weeks)/sqrt(length(diettt$gender))
6 [1] 0.9244236
7 > var(diettt$pre.weight)
8 [1] 63.59509
9 > sd(diettt$pre.weight)
10 [1] 7.974653
11 > sd(diettt$pre.weight)/sqrt(length(diettt$gender))
12 [1] 0.9147554
13 > var(diettt$pre.weight)
14 [1] 63.59509
15 > sd(diettt$Weight.lost)
16 [1] 2.505803
17 > sd(diettt$Weight.lost)/sqrt(length(diettt$gender))
18 [1] 0.2874354
```

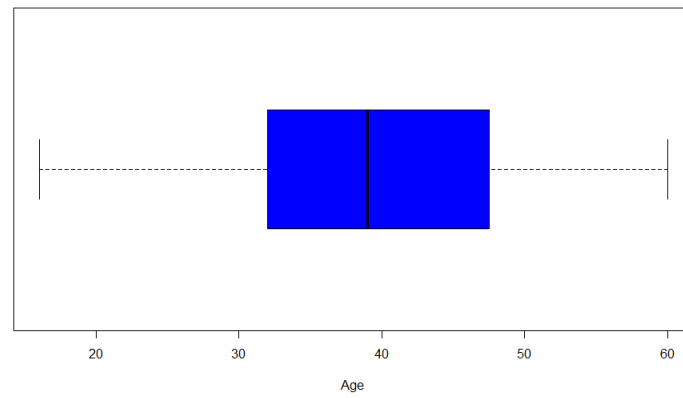
2.4.2 Box plot

Implementation on R-studio for plotting box plots of criteria *Height*, *Age*, *Weight* and *lost.weight*

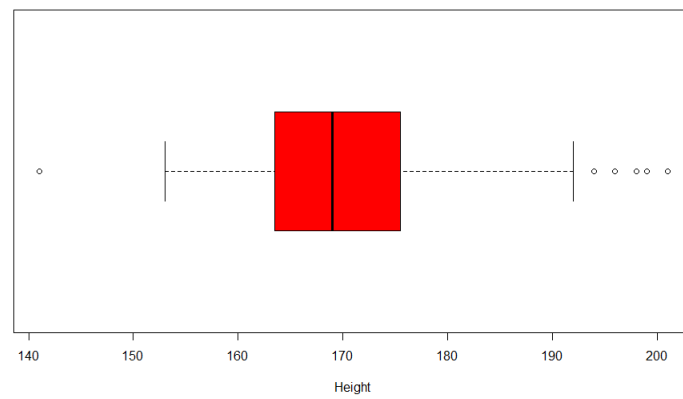
```
1 ###Box plotting
2 ##Age
3 boxplot(diettt$Age,
4         main = "Age box plot",
5         xlab = "Age",
6         col = "blue",
7         horizontal = TRUE
8         )
9 ##Height
10 boxplot(diettt$Height,
11         main = "Height box plot",
12         xlab = "Height",
13         col = "red",
14         horizontal = TRUE
15         )
16
17 ##Weight
18 boxplot(diettt$pre.weight,diettt$weight6weeks,
19         main = "Weight",
20         at = c(1,2),
21         names = c("preweight", "6weeks"),
22         las = 2,
23         col = c("orange","red"),
24         border = "brown",
25         horizontal = TRUE
26         )
27
28 ##lost.weight
29 boxplot(diettt$Weight.lost,
30         main = "Lost weight",
31         xlab = "Weight",
32         col = "orange",
33         horizontal = TRUE
34         )
```

The box plots for each variables:

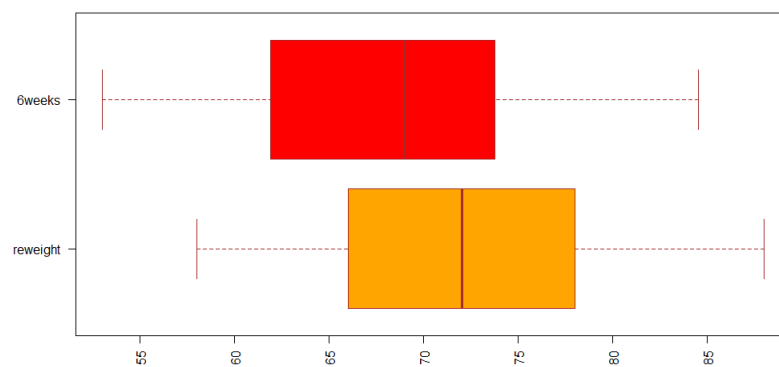
Age box plot

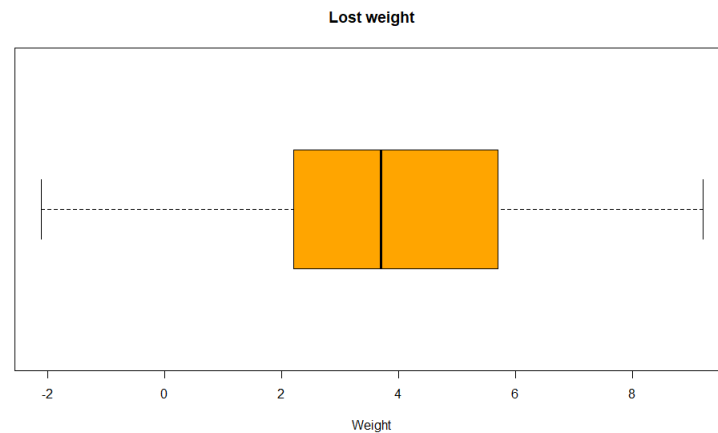


Height box plot



Weight





2.5 t.test: between pre.weight and weight6weeks

2.5.1 Theoretical basic

Type of problem: Tests whether the mean of the differences between dependent or paired observations is equal to a target value using paired t-test. In this problem, we will test whether the difference between the previous weight and the weight after 6 months of diet is significant.

What is t-test: A t-test is a hypothesis test of the mean of one or two normally distributed populations. Several types of t-tests exist for different situations, but they all use a test statistic that follows a t-distribution under the null hypothesis. There is several kinds of t-test, but the one we are using to deal with this problem is paired t-test

2.5.2 Implementation

Assumptions:

H_0 : There is no significant difference between the weight before diet and the weight after 6-week diet.

H_1 : There is a significant difference between the weight before diet and the weight after 6-week diet.

Applying formula for to calculate t-value

$$t = \frac{\frac{\sum D}{N}}{\sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{(N-1)N}}}$$

Where: $\sum D$ = Sum of the differences, $\sum D^2$ = Sum of the squared difference.

Calculating using t.test() function: Implementation on R-Studio:

```
1 t.test(diett$pre.weight, diett$weight6weeks, paired = TRUE)
```

After that, we got the result of:


```
1 Paired t-test
2
3 data: diett$pre.weight and diett$weight6weeks
4 t = 13.728, df = 75, p-value < 2.2e-16
5 alternative hypothesis: true difference in means is not equal to 0
6 95 percent confidence interval:
7  3.373452 4.518653
8 sample estimates:
9 mean of the differences
10      3.946053
```

We can clearly see that the t-value lies out the region of 95% confident interval. So we reject the null hypothesis and conclude that there is a significant difference between the weight before diet and the weight after 6-week diet.

2.6 One way ANOVA: What is the best diet for weight loss?

To solve the problem, we'll use One Way ANOVA to detect whether or not exist a difference between amount weight loss of each diet and compute the significance of that difference. Then, we'll use TukeyHSD as a post hoc test to find out the best diet for weight loss.

According to problem statement, we'll carry out the ANOVA test for the dependent variable *weight.loss* and the independent variable *Diet*, which are the two variables we have viewed above in the data set.

2.6.1 Implementation of One Way ANOVA test in R

Performing One Way ANOVA test using the `aov()` function in R with the corresponding variables of *weight.loss* and *Diet*. Then store the result in to a variable called *aov_1*.

```
1 aov_1 <- aov(formula = weight.loss ~ Diet, data = diet)
```

We can see the result of the test by printing *aov_1*:

```
1 Call:
2   aov(formula = weight.loss ~ Diet, data = diet)
3
4 Terms:
5           Diet Residuals
6 Sum of Squares  60.5270  410.4018
7 Deg. of Freedom    2      73
8
9 Residual standard error: 2.371064
```

Before go to the conclusion on detecting the difference between the diet types by using the ANOVA test, we have to go through some assumptions test for ANOVA to confirm that the given result will be reliable. Firstly, we test the normality assumption by graphing a histogram of the standardised residuals of the ANOVA Test by using the code below:

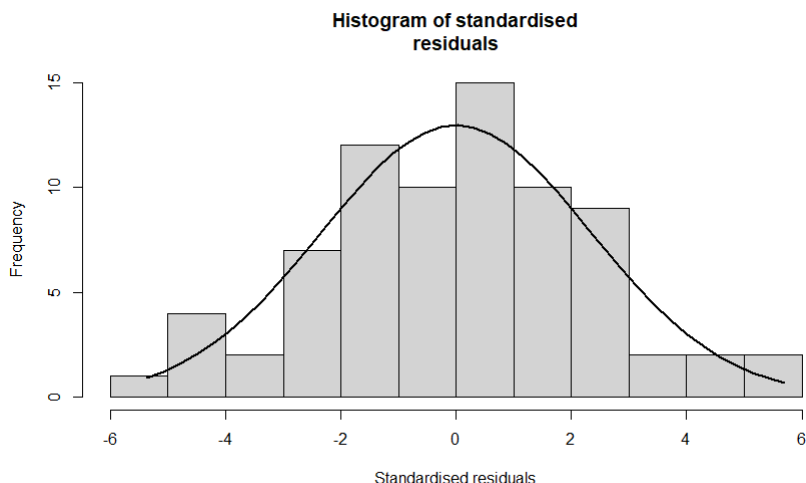
```
1 diet = read_csv("C:/Users/xps/Desktop/stcp-Rdataset-Diet.csv")
2 View(diet)
3 #Clean NA
4 sum(is.na(diet))
5 diet <- na.omit(diet)
6 sum(is.na(diet))
```

```

7 View(diet)
8
9 #Features engineer
10 diet$weight.loss = diet$pre.weight - diet$weight6weeks
11 attach(diet)
12 diet$Diet = factor(diet$Diet, labels = c("A","B","C"))
13 diet$gender = factor(diet$gender, labels = c("Female","Male"))
14 View(diet)
15
16 #Plotting Boxplot
17 boxplot(weight.loss~Diet,data=diet,col="light gray",
18         ylab = "Weight loss (kg)", xlab = "Diet type",na.rm = TRUE)
19 abline(h=0,col="blue")
20
21 #One way ANOVA
22 aov_1 <- aov(formula = weight.loss ~ Diet, data = diet)
23 print(aov_1)
24 summary(aov_1)
25
26 #using TukeyHSD to find the best diet
27 TukeyHSD(aov_1)
28
29 #Levene Test for Homogeneity of Variances assumption
30 library(car)
31 leveneTest(weight.loss~Diet, data = diet)
32
33 #Res Histogram for Normality assumption
34 res <- aov_1$residuals
35 h<-hist(res, main="Histogram of standardised
36 residuals",xlab="Standardised residuals")
37 xfit <- seq(min(res), max(res), length = 40)
38 yfit <- dnorm(xfit, mean = mean(res), sd = sd(res))
39 yfit <- yfit * diff(h$mids[1:2]) * length(res)
40 lines(xfit, yfit, col = "black", lwd = 2)

```

- By running the above code, we will have the histogram as below:
- As we can observe from the histogram, the distribution of the residuals of the ANOVA test is



following the pattern resembles the bell curve hence the residuals is normally distributed \Rightarrow So the normality assumption of ANOVA can be trusted.

- Next, we'll perform the Levene's test for the homogeneity of variances assumption of ANOVA:

```
1 library(car)
2 leveneTest(weight.loss ~ Diet, data = diet)
```

- Which give us the result of:

```
1 Levene's Test for Homogeneity of Variance (center = median)
2      Df F value Pr(>F)
3 group  2  0.4629 0.6313
4      73
```

- As p-value = 0.6313 > 0.05, the homogeneity of variances can be assumed.

- Giving that the normality of data and homogeneity of variances can be assumed, we can use the *summary()* function to state our conclusion for the ANOVA test:

```
1 summary(aov_1)
```

```
1 > summary(aov_1)
2      Df Sum Sq Mean Sq F value Pr(>F)
3 Diet      2    60.5   30.264    5.383 0.0066 **
4 Residuals 73   410.4    5.622
5 ---
6 Signif. codes:  0   ***    0.001   **   0.01   *   0.05   .   0.1   1
```

- From the summary above, we can conclude that there is a $F(2,73) = 5.383$ difference in mean weight lost by the significance $p = 0.0066$ between the diets.

- By the result above, we can use the TukeyHSD post hoc test to compare the differences and significance of each pair of diets, from that we can draw an answer for the problem statement. The TukeyHSD test can be performed by:

```
1 TukeyHSD(aov_1)
```

- The test result is:

```
1 Tukey multiple comparisons of means
2 95% family-wise confidence level
3
4 Fit: aov(formula = weight.loss ~ Diet, data = diet)
5
6 $Diet
7      diff      lwr      upr      p adj
8 B-A -0.032000 -1.6530850 1.589085 0.9987711
9 C-A  1.848148  0.2567422 3.439554 0.0188047
10 C-B  1.880148  0.3056826 3.454614 0.0152020
```

- From the result, we can see that pair-wised, there was a significant difference between diet C and diet A ($p = 0.019$) and diet C lost more than 1.85 (kg) on average of diet A. Also, there was a significant difference between diet C and diet B ($p = 0.015$) and diet C lost more than 1.88 (kg) on average of diet B

2.6.2 Conclusion

A one-way ANOVA was conducted to compare the effectiveness of three diets. Normality checks and Levene's test were carried out and the assumptions met. There was a significant difference

in mean weight lost [$F(2,73) = 5.383$, $p = 0.0066$] between the diets. Post hoc comparisons using the Tukey test were carried out. There was a significant difference between diets A and C ($p = 0.019$) with people on diet C lost on average 1.85 kg more than those on diet A. There was also a significant difference between diets B and C difference ($p = 0.015$) with people on diet C lost on average 1.88 kg more than those on diet B. So in conclusion, we can state that diet C is the best diet for weight loss.

2.7 Two way ANOVA : How do *Diet* and *gender* affect *weightLOST*?

As we previously declared before of a column named: *weight.lost* has values of *pre.weight - weight6weeks*. The aim of the dataset was to define which diet was the best for losing weight but it was also considered as the best diets for males and females so the independent variables are *diet* and *gender*.

There are three hypotheses with a two-way ANOVA. There are tests for main effects(diet & gender) as well as the test for interaction between them.

CHECKING THE ASSUMPTIONS FOR TWO-WAY ANOVA & STEPS ON R

First, we assume that residuals should be normally distributed. Then we perform a check by saving the residuals from *aov()* command output and create a histogram or either a normality test.

For further tests, we use the Homogeneity of Variance(Levene's Test) for test of equality of variances.

Implementation using R-Studio

- Performing Two-way ANOVA test and store to *anova2* variable

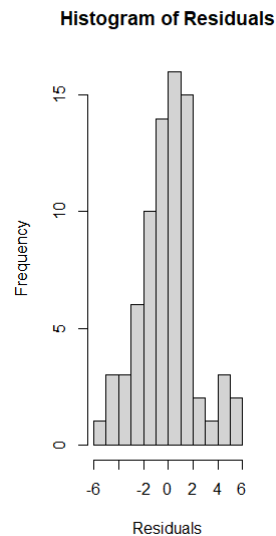
```
1 ##anova2way
2 anova2 = aov(diet$Weight.lost ~ diet$gender * diet$diet_type)
```

The result of *anova2* is:

```
1 Call:
2   aov(formula = diet$Weight.lost ~ diet$gender * diet$diet_type)
3
4 Terms:
5             diet$gender diet$diet_type diet$gender:diet$diet_type
6 Sum of Squares      0.2785         60.4172             33.9041
7 Deg. of Freedom           1              2              2
8
9 Residuals
10 Sum of Squares    376.3290
11 Deg. of Freedom       70
12
13 Residual standard error: 2.318648
```

- Checking the residual for normality, we can either use a histogram or *shapiro.test()* to test for normal distribution.

```
1 result = anova2$residuals
2 hist(result, main = "Histogram of Residuals", xlab = "Residuals")
3 #alternative way to test normality using shapiro test
4 print(shapiro.test(result))
```



The Histogram for Residuals is: Also, the normality test performed by *shapiro.test()* function result is:

```

1 Shapiro-Wilk normality test
2
3 data:  result
4 W = 0.97738, p-value = 0.1923
5
```

Levene's Test for equality variances is implemented by:

```

1 #Levene's Test for equality variances
2 library(car)
3 print(leveneTest(diet$Weight.lost~diet$gender*diet$diet_type))

```

- The result for the test is:

```

1 Levene's Test for Homogeneity of Variance (center = median)
2      Df F value Pr(>F)
3 group  5  0.3867 0.8563
4      70

```

- Printing summary ANOVA Table:

```

1 #view ANOVA Table
2 print(summary(anova2))

```

- The summary of ANOVA Table is:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
diett\$gender	1	0.3	0.278	0.052	0.82062	
diett\$diet_type	2	60.4	30.209	5.619	0.00546	**
diett\$gender:diett\$diet_type	2	33.9	16.952	3.153	0.04884	*
Residuals	70	376.3	5.376			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1						1

- Finally, we perform the post-hoc tests which is produced by using the *TukeyHSD()* function for creating the main effect and interactions. Only interpret post hoc test for the significant factors from the ANOVA. If the interaction is NOT significant, interpret the post hoc tests for significant main effects but if it is significant, only interpret the interactions post hoc tests.

Post hoc tests for main effects of diet and gender:

- Implementation:

```
1 #post-hoc test for finding the best diet
2 print(TukeyHSD(anova2))
```

Result:

```
1 Tukey multiple comparisons of means
2 95% family-wise confidence level
3
4 Fit: aov(formula = diett$Weight.lost ~ diett$gender * diett$diet_type)
5
6 $'diett$gender'
7      diff      lwr      upr      p adj
8 Male-Female 0.1221283 -0.9480861 1.192343 0.8206233
9
10 $'diett$diet_type'
11      diff      lwr      upr      p adj
12 B-A -0.03484966 -1.6215073 1.551808 0.9984761
13 C-A 1.84475570 0.2871469 3.402365 0.0162482
14 C-B 1.87960536 0.3385771 3.420634 0.0128844
15
16 $'diett$gender:diett$diet_type'
17      diff      lwr      upr      p adj
18 Male:A-Female:A 0.6000000 -2.2129628 3.4129628 0.9887997
19 Female:B-Female:A -0.4428571 -3.0107291 2.1250148 0.9958151
20 Male:B-Female:A 1.0590909 -1.6782698 3.7964516 0.8656520
21 Female:C-Female:A 2.8300000 0.3052886 5.3547114 0.0191170
22 Male:C-Female:A 1.1833333 -1.4893925 3.8560592 0.7855223
23 Female:B-Male:A -1.0428571 -3.8558199 1.7701056 0.8852416
24 Male:B-Male:A 0.4590909 -2.5093998 3.4275816 0.9975014
25 Female:C-Male:A 2.2300000 -0.5436187 5.0036187 0.1863470
26 Male:C-Male:A 0.5833333 -2.3256625 3.4923292 0.9915569
27 Male:B-Female:B 1.5019481 -1.2354126 4.2393087 0.5963201
28 Female:C-Female:B 3.2728571 0.7481458 5.7975685 0.0040103
29 Male:C-Female:B 1.6261905 -1.0465354 4.2989163 0.4833188
30 Female:C-Male:B 1.7709091 -0.9260048 4.4678230 0.3965102
31 Male:C-Male:B 0.1242424 -2.7117126 2.9601974 0.9999949
32 Male:C-Female:C -1.6466667 -4.2779524 0.9846191 0.4513580
```

CONCLUSIONS ABOUT TWO-WAY ANOVA ON HOW *Diet & gender* AFFECT *weightLOST*

There are a lot of combinations of diets and genders. There was a statistically significant inter-



action between the effect of *Diet* and *Gender* on weight loss, e.g: $F(2, 70) = 3.153, p = 0.049$. Tukey's HSD post hoc test was carried out. For females, diet 3 was significantly different to diet 1 ($p = 0.0191$) and diet 2 ($p = 0.004$) but there is no evidence to suggest that any diets differed for males. Women on diet 3 lost on average $2.83kg$ more than those on diet 1 and $3.27kg$ more than those on diet 2.

Finally, normality checks and Levene's test were carried out and the assumptions met.