

**National College of Ireland**

**Project Submission Sheet – 2017/2018**

**School of Computing**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student Name:** | JORDI BATLLE | | |
| **Student ID:** | 17133246 | | |
| **Programme:** | HDSDAICTSEP3 | **Year:** | 2017/2018 |
| **Module:** | ADVANCED BUSINESS DATA ANALYSIS | | |
| **Lecturer:** | GIOVANI ESTRADA | | |
| **Due Date:** | 12/03/2018 | | |
| **Project Title:** | CA1: PARAMETRIC STATISTICAL TESTS | | |
|  |  | | |

**I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.**

**ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.**

|  |  |
| --- | --- |
| **Signature:** | ……………………………………………………………………………………………………………… |
| **Date:** | ……………………………………………………………………………………………………………… |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).

2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

|  |  |
| --- | --- |
| **Office Use Only** | |
| Signature: |  |
| Date: |  |
| Penalty Applied (if applicable): |  |

PARAMETRIC STATISTICAL TESTS

Table of Contents

[1. T-TEST 3](#_Toc508732091)

[1.1. Sample details 3](#_Toc508732092)

[1.2. Checking the data 4](#_Toc508732093)

[1.3. Running the statistical test 6](#_Toc508732094)

[2. ONE-WAY ANOVA 9](#_Toc508732095)

[2.1. Sample details 9](#_Toc508732096)

[2.2. Checking the data 10](#_Toc508732097)

[2.3. Running the statistical test 15](#_Toc508732098)

[3. TWO-WAY ANOVA 20](#_Toc508732099)

[3.1. Sample details 20](#_Toc508732100)

[3.2. Checking the data 21](#_Toc508732101)

[3.3. Running the statistical test 24](#_Toc508732102)

[References 29](#_Toc508732103)

# T-TEST

# Sample details

For this first test, I have downloaded a dataset that shows the salaries of public employees at the White House under the Obama administration and the Trump administration. For the Obama administration, the file shows the data for several years but I will only use the salaries for his last year 2016. For Trump administration, the file shows the salaries for his first year as President, 2017.

The file shows salaries for both employees and detailees; for the analysis, I will only consider the employees’ salaries.

The dataset can be found at: <https://www.kaggle.com/adamschroeder/white-house-salaries/data>

And the subset used can be found in the project submission attached files.

Below a brief description of the data I will review:

|  |
| --- |
| > # Structure of the file  > str(salary)  'data.frame': 796 obs. of 2 variables:  $ ADMIN : Factor w/ 2 levels "OBAMA","TRUMP": 2 2 2 2 2 2 2 2 2 2 ...  $ SALARY: int 30000 30000 40000 40000 40000 40000 40000 40000 40000 40000 ... |

The objective is to compare whether there is a significant difference between the salaries under Obama administration and Trump administration.

# Checking the data

Before to start, in order to determine what test best suits, I check whether the following assumptions are met:

* Assumption #1:

The dependent variable, in this case the employees’ salaries, is continuous.

* Assumption #2:

The independent variable, administration, consists of two categorical, independent groups. In this case, every observations falls in one and only one of the categories: either the employee worked for Obama administration; either they work for Trump.

* Assumption #3:

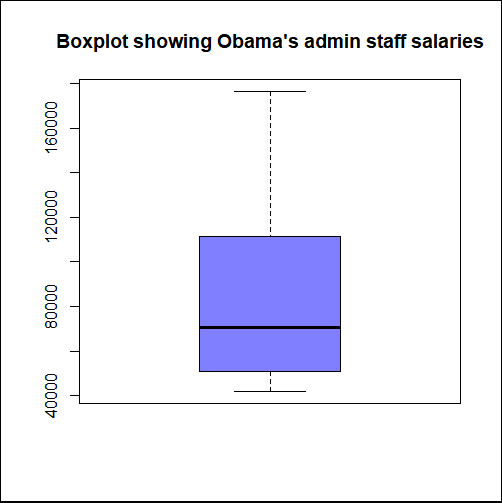
The observations are independent. There is no relationship between the observations in each category or between the categories themselves. In this case, the salaries of every employee are unique and therefore independent from each other.

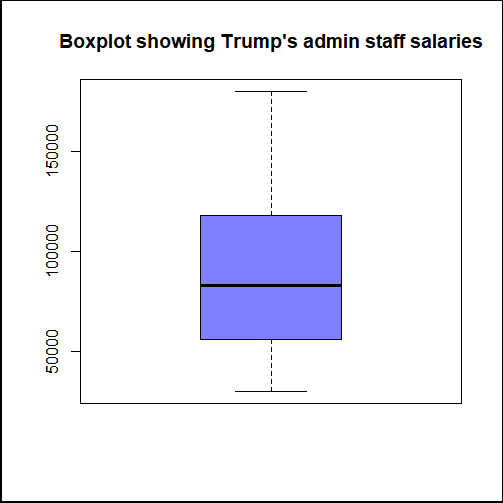
* Assumption #4:

Checking for the outliers.

From the original file, I removed three rows for employees for both Obama and Trump administrations were no salary was stated.

Below boxplots showing the salaries under Trump and Obama administrations:





No values are outside the boxplots therefore we can conclude there are no outliers.

* Assumption #5:

In order to decide what test is the most appropriate, I checked whether the dependent variables, salaries, are approximately normally distributed for each category of the independent variable, Obama and Trump administration.

Below, the results of the Shapiro-Wilk test for normality ran in R:

> # Shapiro test for normality

> shapiro.test(obama$SALARY)

Shapiro-Wilk normality test

data: obama$SALARY

W = 0.87747, p-value < 2.2e-16

> shapiro.test(trump$SALARY)

Shapiro-Wilk normality test

data: trump$SALARY

W = 0.91801, p-value = 5.038e-13

I set the hypotheses (valid for both samples):

H0: The sample data are normally distributed

H1: The sample data are not normally distributed

As we can see, for both samples, the result (p-value) is nearly 0.

If we set the level of significance α at 5% (the standard value), the p-value is smaller than α and, therefore, we fail to accept the null hypothesis; we cannot confirm the samples are normally distributed.

* Assumption #6:

I check whether the variance of the samples is homogenic.

Below, the results of Levene test for homogeneity of variances ran in R:

> # Levene test for homogeneity of variances

> library(car)

> leveneTest(SALARY~ADMIN, data=salary)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 1 5.7073 0.01713 \*

794

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

I set the hypotheses:

H0: The variance of the sample data is homogenic.

H1: The variance of the sample data is not homogenic.

As we can see, the result (p-value) is 0.017.

If we set the level of significance α at 5% (the standard value), the p-value is smaller than α and, therefore, we fail to accept the null hypothesis; we cannot confirm the variance of the samples is homogenic.

# Running the statistical test

As stated previously the aim of the study is to compare the salaries of the two administrations. For this, I will run a t-test based on the assumptions checked (even though not all of them are met).

In statistics, the t test is considered robust to violations of the checked assumptions therefore, I will go ahead with the t test. Although, afterwards, I will conduct the non-parametric equivalent Mann Whitney U Test to perform a comparison between the results.

1. **State the Null and Alternate hypotheses**

If we state the following:

O: Obama administration salaries

T: Trump administration salaries

H0 : µO = µT

there is no difference between the means of Obama and Trump administration salaries

H1 : µO ≠ µT

there is a difference between the means of Obama and Trump administration salaries

1. **I set the level of significance α at 0.05 as the standard practice.**

The probability of rejecting the null hypothesis when it is true (type I error) is 5%.

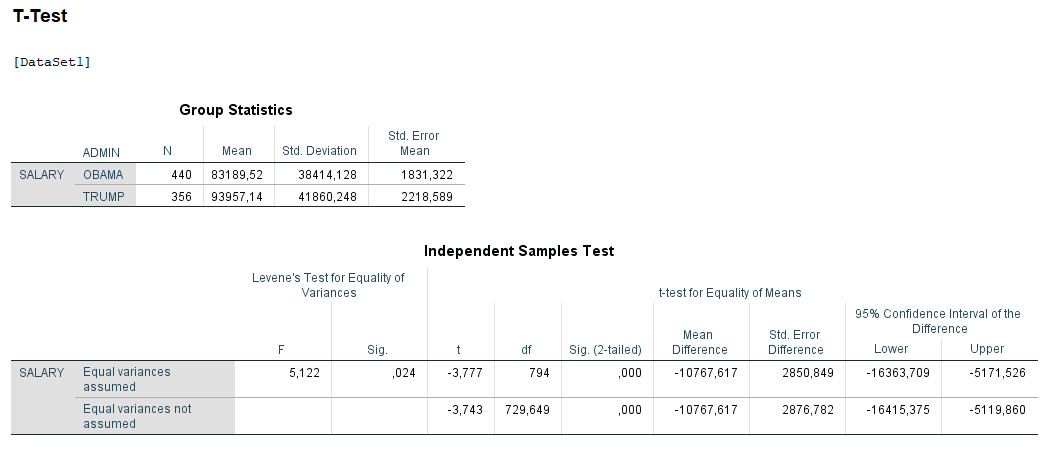
1. **As per the above-mentioned assumptions, I will run a t-test for independent variables.**
2. **Compute the test statistic value using appropriate formulas**

I ran the t-test using Excel Data Analysis toolkit, SPSS and R. Results follow:

1. **Excel results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | **t-Test: Two-Sample Assuming Unequal Variances** |  |  |  |
|  |  |  |  |  |
|  |  | *Variable 1* | *Variable 2* |  |
|  | Mean | 83189.52045 | 93957.13764 |  |
|  | Variance | 1475645260 | 1752280399 |  |
|  | Observations | 440 | 356 |  |
|  | Hypothesized Mean Difference | 0 |  |  |
|  | df | 730 |  |  |
|  | t Stat | -3.742938066 |  |  |
|  | P(T<=t) one-tail | 9.80878E-05 |  |  |
|  | t Critical one-tail | 1.64694365 |  |  |
|  | P(T<=t) two-tail | 0.000196176 |  |  |
|  | t Critical two-tail | 1.963218974 |  |  |
|  |  |  |  |  |

1. **SPSS results**



1. **R results**

> # T test to compare populations

> t.test(obama$SALARY, trump$SALARY, alternative = "two.sided", paired = FALSE)

Welch Two Sample t-test

data: obama$SALARY and trump$SALARY

t = -3.7429, df = 729.65, p-value = 0.0001962

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-16415.37 -5119.86

sample estimates:

mean of x mean of y

83189.52 93957.14

As we can see, the results for the tests performed with different tools are very similar:

|  |  |  |  |
| --- | --- | --- | --- |
|  | EXCEL | SPSS | R |
| df | 730 | 729.649 | 729.65 |
| t\_stat | -3.742938066 | -3.743 | -3.7429 |
| t\_crit | 1.963218974 | - | - |
| p-value | 0.000196176 | .000 | 0.0001962 |

1. **Conclusion**

We see that: **p < α**

We fail to accept H0. At a 95% level of confidence, we can conclude that there is a significant difference in the population means and therefore, the salaries for Obama and Trump administrations are different.

Although, as stated previously, the samples do not meet the assumptions for a t-test. I will then perform further analysis on the samples.

**Mann-Whitney U test**

I ran the Mann-Whitney U test that allows me to compare 2 populations when the samples are not normally distributed.

Below the result of the test ran in R:

> # Mann-Whitney U test to compare non normal populations

> wilcox.test(obama$SALARY, trump$SALARY)

Wilcoxon rank sum test with continuity correction

data: obama$SALARY and trump$SALARY

W = 65883, p-value = 0.0001149

alternative hypothesis: true location shift is not equal to 0

We observe again that the p-value is smaller than α, therefore, with this test, we reach the same conclusion than previously: the salaries for both administrations are different.

# ONE-WAY ANOVA

# 2.1. Sample details

For the second test, I have downloaded a dataset that shows Major League Baseball Players characteristics. Among the information included in the file there is their position in the pitch and their weight in pounds.

The dataset can be found at: https://people.sc.fsu.edu/~jburkardt/data/csv/mlb\_players.csv

And the subset used can be found in the project submission attached files.

Below a brief description of the data I will review:

> # Structure of the file

> str(players)

'data.frame': 1034 obs. of 5 variables:

$ Name : Factor w/ 1032 levels "A.J. Burnett",..: 13 795 818 629 200 135 732 718 67 23 ...

$ Position: Factor w/ 10 levels "","Catcher","Designated Hitter",..: 2 2 2 4 4 7 8 10 10 5 ...

$ Height : int 74 74 72 72 73 69 69 71 76 71 ...

$ Weight : int 180 215 210 210 188 176 209 200 231 180 ...

$ Age : num 23 34.7 30.8 35.4 35.7 ...

The idea is to study whether the morphology (weight) of the players will determine their position in the pitch.

# 2.2. Checking the data

Before to start, in order to determine what test best suits, I check whether the following assumptions are met:

* Assumption #1:

The dependent variable, in this case the players’ weights, is continuous.

* Assumption #2:

The independent variable, position, consists of 9 categorical, independent groups. In this case, every observation falls in one and only one of the categories: each player plays in one single position.

* Assumption #3:

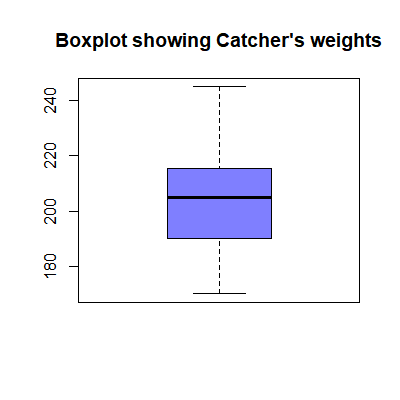
The observations are independent. There is no relationship between the observations in each category or between the categories themselves. In this case, the weights of every single player are unique and therefore independent from each other.

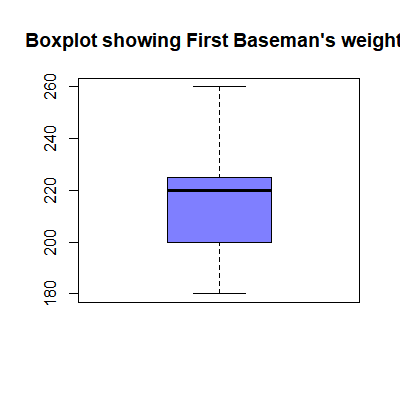
* Assumption #4:

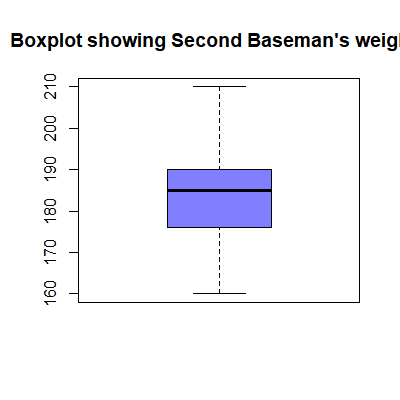
Checking for the outliers.

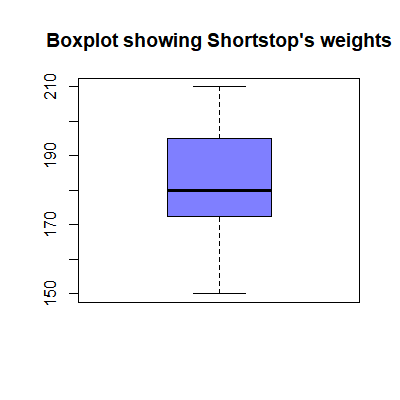
From the original file, I removed one row for a player where no weight was stated.

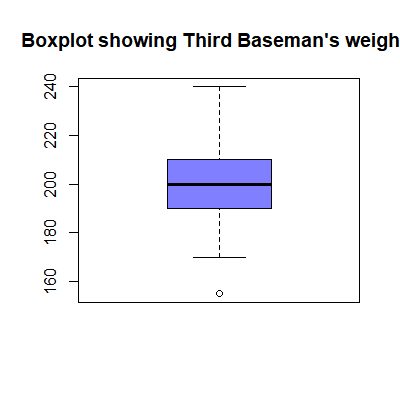
Below, I show the boxplots showing the weights for the positions that I will analyse because, as we will see, there are 2 categories I will discard in my analysis as they are not normally distributed:

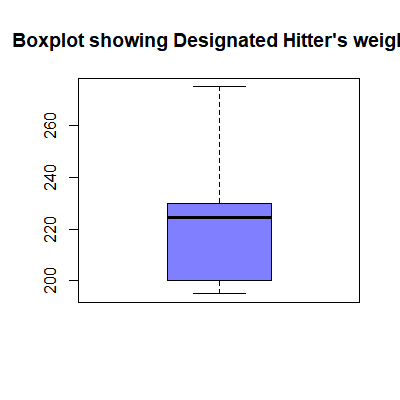


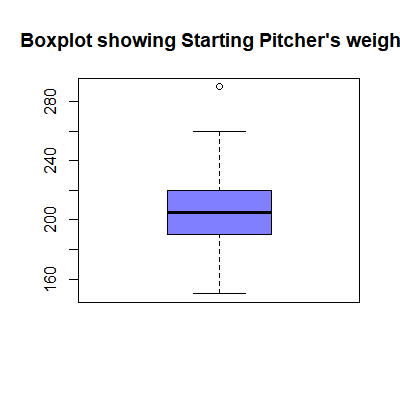












From the boxplots, we can see there could be outliers for the Starting pitcher and Third basement positions; once the values checked, we can say the values are realistic as we can see in the other positions similar values. Therefore, I do not take any action on them.

* Assumption #5:

In order to decide what test is the most appropriate, I checked whether the dependent variable, weight, is approximately normally distributed for each one of the 9 categories of the independent variable.

Next, the results of the Shapiro-Wilk test of normality ran in R:

> # Shapiro test for normality

> shapiro.test(Catcher$Weight)

Shapiro-Wilk normality test

data: Catcher$Weight

W = 0.97875, p-value = 0.2318

> shapiro.test(FirstBaseman$Weight)

Shapiro-Wilk normality test

data: FirstBaseman$Weight

W = 0.96622, p-value = 0.1242

> shapiro.test(SecondBaseman$Weight)

Shapiro-Wilk normality test

data: SecondBaseman$Weight

W = 0.98289, p-value = 0.5847

> shapiro.test(Shortstop$Weight)

Shapiro-Wilk normality test

data: Shortstop$Weight

W = 0.97988, p-value = 0.5205

> shapiro.test(ThirdBaseman$Weight)

Shapiro-Wilk normality test

data: ThirdBaseman$Weight

W = 0.98059, p-value = 0.6437

> shapiro.test(Outfielder$Weight)

Shapiro-Wilk normality test

data: Outfielder$Weight

W = 0.98575, p-value = 0.04727

> shapiro.test(DesignatedHitter$Weight)

Shapiro-Wilk normality test

data: DesignatedHitter$Weight

W = 0.91735, p-value = 0.116

> shapiro.test(StartingPitcher$Weight)

Shapiro-Wilk normality test

data: StartingPitcher$Weight

W = 0.98746, p-value = 0.05047

> shapiro.test(ReliefPitcher$Weight)

Shapiro-Wilk normality test

data: ReliefPitcher$Weight

W = 0.98823, p-value = 0.01177

I set the hypotheses (valid for both samples):

H0: The sample data are normally distributed

H1: The sample data are not normally distributed

If we set the level of significance α at 5% (the standard value), we can see that for Outfielder and Relief pitcher the p-value is smaller than α and therefore we fail to accept the null hypothesis; we cannot confirm they are normally distributed.

I will then discard these two categories to perform the analysis.

For the other categories, we see a p-value bigger than α and therefore we can accept the null hypotheses and conclude they are normally distributed.

* Assumption #6:

I check whether the variance of the samples is homogenic.

Below, the results of Levene for homogeneity of variances ran in R:

> # Levene test for homogeneity of variances

> library(car)

> players <- subset(players, Position == 'Catcher' | Position == 'First Baseman' | Position == 'Second Baseman' | Position == 'Shortstop' | Position == 'Third Baseman' | Position == 'Designated Hitter' | Position == 'Starting Pitcher')

> leveneTest(Weight~Position, data=players)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 6 5.972 4.758e-06 \*\*\*

517

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

I set the hypotheses:

H0: The variance of the sample data is homogenic.

H1: The variance of the sample data is not homogenic.

As we can see, the result (p-value) is near 0.

If we set the level of significance α at 5% (the standard value), the p-value is smaller than α and, therefore, we fail to accept the null hypothesis; we cannot confirm the variance of the samples is homogenic.

# 2.3. Running the statistical test

I will run a test to compare whether there is a difference in the players weights with relation to the position they play. For this, I will run a one way Anova test based on the assumptions checked previously.

1. **State the Null and Alternate hypotheses**

If we state the following:

1:Catcher

2:First Baseman

3: Second Baseman

4: Shortstop

5: Third Baseman

6: Designated Hitter

7: Starting Pitcher

H0 : µ1 = µ2 = µ3 = µ4 = µ5 = µ6 = µ7

there is no difference between the means of the different categories

H1 : µ1 ≠ µ2 ≠ µ3 ≠ µ4 ≠ µ5 ≠ µ6 ≠ µ7

there is a difference between the means of the different categories

1. **I set the level of significance α at 0.05 as the standard practice.**

The probability of rejecting the null hypothesis when it is true (type I error) is 5%.

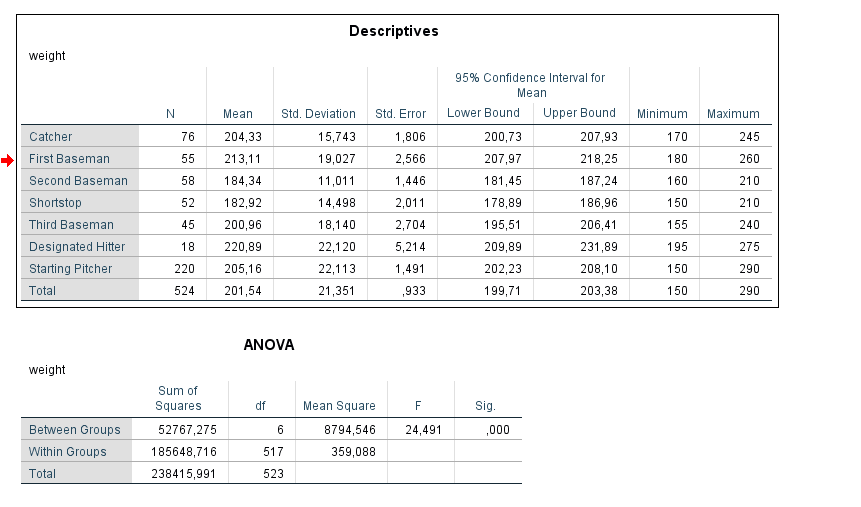
1. **As per the above-mentioned assumptions, I will run a one way Anova test.**
2. **Compute the test statistic value using appropriate formulas**

I ran the one way Anova test using Excel Data Analysis toolkit, SPSS and R. Results follow:

1. **Excel results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| Catcher | 76 | 15529 | 204.3289474 | 247.8503509 |  |  |
| First Baseman | 55 | 11721 | 213.1090909 | 362.0249158 |  |  |
| Second Baseman | 58 | 10692 | 184.3448276 | 121.2474289 |  |  |
| Shortstop | 52 | 9512 | 182.9230769 | 210.1900452 |  |  |
| Third Baseman | 45 | 9043 | 200.9555556 | 329.0434343 |  |  |
| Designated Hitter | 18 | 3976 | 220.8888889 | 489.2810458 |  |  |
| Starting Pitcher | 220 | 45136 | 205.1636364 | 488.9685347 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 52767.27495 | 6 | 8794.545825 | 24.49130973 | 1.40E-25 | 2.116103613 |
| Within Groups | 185648.7155 | 517 | 359.0884246 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 238415.9905 | 523 |  |  |  |  |

1. **SPSS results**



1. **R results**

> # display ANOVA table

> summary(results)

Df Sum Sq Mean Sq F value Pr(>F)

Position 6 52767 8795 24.49 <2e-16 \*\*\*

Residuals 517 185649 359

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

As we can see, the results for the tests performed with different tools are very similar:.

|  |  |  |  |
| --- | --- | --- | --- |
|  | EXCEL | SPSS | R |
| df | 523 | 523 | 523 |
| F\_stat | 24.49130973 | 24.491 | 24.49 |
| F\_crit | 2.116103613 | - | - |
| p-value | 1.40E-25 | .000 | <2e-16 |

**5.     Conclusion**

We see that: **p < α**

We fail to accept H0. At a 95% level of confidence, we can conclude that there is a significant difference between at least 2 categories.

Although, as we saw previously, the sample did not meet all the assumptions, no homogeneity of the variances. Therefore, further tests should be run as for example the Kruskal-Wallis H Test.

**TukeyHSD test**

As we found a difference for at least 2 of the categories, we run now the Tukey HSD test to check which of the categories are different.

Below, the results in R:

> # Tukey test to compare the groups

> TukeyHSD(results)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Weight ~ Position, data = players)

$Position

diff lwr upr p adj

Designated Hitter-Catcher 16.559942 1.856728 31.2631552 0.0159152

First Baseman-Catcher 8.780144 -1.149620 18.7099069 0.1228612

Second Baseman-Catcher -19.984120 -29.763763 -10.2044769 0.0000001

Shortstop-Catcher -21.405870 -31.500442 -11.3112985 0.0000000

Starting Pitcher-Catcher 0.834689 -6.628408 8.2977864 0.9998942

Third Baseman-Catcher -3.373392 -13.923843 7.1770592 0.9647127

First Baseman-Designated Hitter -7.779798 -23.011037 7.4514407 0.7376716

Second Baseman-Designated Hitter -36.544061 -51.677859 -21.4102631 0.0000000

Shortstop-Designated Hitter -37.965812 -53.305004 -22.6266198 0.0000000

Starting Pitcher-Designated Hitter -15.725253 -29.476189 -1.9743160 0.0134352

Third Baseman-Designated Hitter -19.933333 -35.576301 -4.2903653 0.0034045

Second Baseman-First Baseman -28.764263 -39.321136 -18.2073909 0.0000000

Shortstop-First Baseman -30.186014 -41.035278 -19.3367503 0.0000000

Starting Pitcher-First Baseman -7.945455 -16.401456 0.5105470 0.0814165

Third Baseman-First Baseman -12.153535 -23.428204 -0.8788666 0.0251479

Shortstop-Second Baseman -1.421751 -12.133788 9.2902864 0.9997128

Starting Pitcher-Second Baseman 20.818809 12.539607 29.0980103 0.0000000

Third Baseman-Second Baseman 16.610728 5.468046 27.7534103 0.0002500

Starting Pitcher-Shortstop 22.240559 13.591620 30.8894984 0.0000000

Third Baseman-Shortstop 18.032479 6.612394 29.4525635 0.0000770

Third Baseman-Starting Pitcher -4.208081 -13.384996 4.9688347 0.8241057

As per the results, p smaller than α for the following combinations:

|  |  |
| --- | --- |
| GROUPS | P-VALUE |
| Designated Hitter-Catcher | 0.0159152 |
| Second Baseman-Catcher | 0.0000001 |
| Shortstop-Catcher | 0 |
| Second Baseman-Designated Hitter | 0 |
| Shortstop-Designated Hitter | 0 |
| Starting Pitcher-Designated Hitter | 0.0134352 |
| Third Baseman-Designated Hitter | 0.0034045 |
| Second Baseman-First Baseman | 0 |
| Shortstop-First Baseman | 0 |
| Third Baseman-First Baseman | 0.0251479 |
| Starting Pitcher-Second Baseman | 0 |
| Third Baseman-Second Baseman | 0.00025 |
| Starting Pitcher-Shortstop | 0 |
| Third Baseman-Shortstop | 0.000077 |

For these combinations we can conclude there is a significant difference between these categories (that will explain the result of the Anova test).

As per the results, p bigger than α for the following combinations:

|  |  |
| --- | --- |
| First Baseman-Catcher | 0.1228612 |
| Starting Pitcher-Catcher | 0.9998942 |
| Third Baseman-Catcher | 0.9647127 |
| First Baseman-Designated Hitter | 0.7376716 |
| Starting Pitcher-First Baseman | 0.0814165 |
| Shortstop-Second Baseman | 0.9997128 |
| Third Baseman-Starting Pitcher | 0.8241057 |

We can then conclude that between these groups there is no significant difference.

# TWO-WAY ANOVA

# 3.1. Sample details

For the third test, I have downloaded from Github a dataset that shows a sample of people and some of their characteristics. For the analysis, I will only use the attributes sex (Female or Male) and race (White, Black, Hispanic or Other). Although, as the Other and Hispanic races have very few observations, I grouped them in the same category that I called Other as well.

Finally we have then a 3 x 2 factorial analysis:

* sex: female, male
* race: black, other, white

As well, as in Excel it is not possible to run a Two way Anova with different group sizes, I limited each group combination to 40 observations.

As other variables are not normally distributed with relation to sex and race, I randomly set a variable IQ.

IQ setting:

* + Firstly, I considered IQ range scale: it can take values from 0 to 200. (IQ test, 2015)
  + Then in excel I generated a random number using the below formula that returns normally distributed values:

=NORMINV(RAND(); 10; 30)

* + The range of values found was between -84 and 98
  + I added to the range 100
  + The result is a range of values between 16 and 198

Therefore, the results range is in the range of IQ standard values accepted.

The dataset can be found at: https://github.com/hadley/r4ds/blob/master/data/heights.csv

And the subset used can be found in the project submission attached files.

Below a brief description of the data I will review:

> str(people)

'data.frame': 240 obs. of 3 variables:

$ sex : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...

$ race: Factor w/ 3 levels "black","other",..: 1 1 1 1 1 1 1 1 1 1 ...

$ IQ : int 112 100 97 124 146 61 125 65 157 105 ...

The objective is to examine the influence of sex and race on people’s IQ.

# 3.2. Checking the data

Before to start, in order to determine what test best suits, I check whether the following assumptions are met:

* Assumption #1:

The dependent variable, in this case IQ, is continuous.

* Assumption #2:

The independent variables, sex and race, consists of categorical, independent groups. In this case, every observation falls in one and only one of the categories: each person is grouped in one single sex and race.

* Assumption #3:

The observations are independent. There is no relationship between the observations in each category or between the categories themselves. In this case, the IQ scores of every person are unique and therefore independent from each other.

* Assumption #4:

Checking for the outliers.

There are no outliers as I set the independent variable initially and I already checked the results fall within the range of IQ values, between 0 and 200.

* Assumption #5:

In order to decide what test is the most appropriate, I checked whether the dependent variable, IQ, is approximately normally distributed for each combination of the groups of the two independent variables, sex and race.

Below, the results of the Shapiro-Wilk test of normality ran in R:

> # Shapiro test for normality

> shapiro.test(Wblack$IQ)

Shapiro-Wilk normality test

data: Wblack$IQ

W = 0.97442, p-value = 0.4907

> shapiro.test(Wother$IQ)

Shapiro-Wilk normality test

data: Wother$IQ

W = 0.98278, p-value = 0.7908

> shapiro.test(Wwhite$IQ)

Shapiro-Wilk normality test

data: Wwhite$IQ

W = 0.98361, p-value = 0.8195

> shapiro.test(Mblack$IQ)

Shapiro-Wilk normality test

data: Mblack$IQ

W = 0.98481, p-value = 0.8583

> shapiro.test(Mother$IQ)

Shapiro-Wilk normality test

data: Mother$IQ

W = 0.98984, p-value = 0.9725

> shapiro.test(Mwhite$IQ)

Shapiro-Wilk normality test

data: Mwhite$IQ

W = 0.9666, p-value = 0.2794

I set the hypotheses (valid for every group):

H0: The sample data are normally distributed

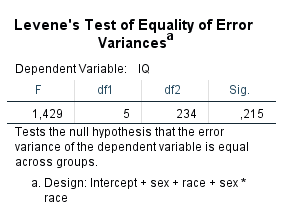
H1: The sample data are not normally distributed

If we set the level of significance α at 5% (the standard value), we can see that for all the categories the p-value is bigger than α and therefore we fail to reject the null hypothesis; we can consider they are all normally distributed.

* Assumption #6:

I check whether the variance of the samples is homogenic.

Below, the results of Levene for homogeneity of variances ran in SPSS:



I set the hypotheses:

H0: The variance of the sample data is homogenic.

H1: The variance of the sample data is not homogenic.

As we can see, the result (p-value) is 0.215

If we set the level of significance α at 5% (the standard value), the p-value is higher than α and, therefore, we fail to reject the null hypothesis; we confirm the variance of the groups is homogenic.

# 3.3. Running the statistical test

I will run a test to try to establish if there is a relation between IQ score and the sex and race of the individuals. For this, I will run a two way Anova test based on the assumptions checked previously.

1. **State the Null hypotheses**

H1: The means of the observations grouped by sex are equal

H2: The means of the observations grouped by race are equal

H3: There is no interaction between sex and race

1. **I set the level of significance α at 0.05 as the standard practice.**

The probability of rejecting the null hypothesis when it is true (type I error) is 5%.

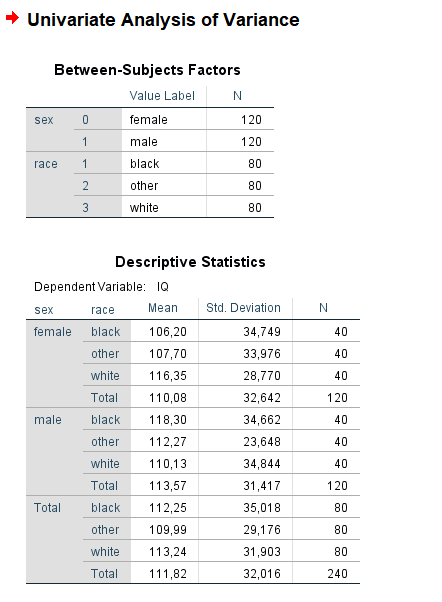
1. **As per the previously confirmed assumptions, I will run a two way Anova test.**
2. **Compute the test statistic value using appropriate formulas**

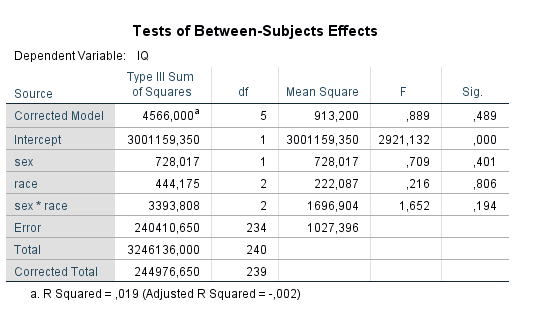
I ran the one way Anova test using Excel Data Analysis toolkit, SPSS and R. Results follow:

1. **Excel results**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Anova: Two-Factor With Replication | | |  |  |  |  |  | |
|  |  |  |  |  |  |  |  | |
| SUMMARY | black | other | white | Total |  |  |  | |
| *female* |  |  |  |  |  |  |  | |
| Count | 40 | 40 | 40 | 120 |  |  |  | |
| Sum | 4248 | 4308 | 4654 | 13210 |  |  |  | |
| Average | 106.2 | 107.7 | 116.35 | 110.0833333 |  |  |  | |
| Variance | 1207.497436 | 1154.369231 | 827.7205128 | 1065.505602 |  |  |  | |
|  |  |  |  |  |  |  |  | |
| *male* |  |  |  |  |  |  |  | |
| Count | 40 | 40 | 40 | 120 |  |  |  | |
| Sum | 4732 | 4491 | 4405 | 13628 |  |  |  | |
| Average | 118.3 | 112.275 | 110.125 | 113.5666667 |  |  |  | |
| Variance | 1201.446154 | 559.2301282 | 1214.112179 | 987.0039216 |  |  |  | |
|  |  |  |  |  |  |  |  | |
| *Total* |  |  |  |  |  |  |  | |
| Count | 80 | 80 | 80 |  |  |  |  | |
| Sum | 8980 | 8799 | 9059 |  |  |  |  | |
| Average | 112.25 | 109.9875 | 113.2375 |  |  |  |  | |
| Variance | 1226.291139 | 851.2530063 | 1017.803639 |  |  |  |  | |
|  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  | |
| ANOVA |  |  |  |  |  |  |  | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |  | |
| Sample | 728.0166667 | 1 | 728.0166667 | 0.708603799 | 0.400767 | 3.881505 |  | |
| Columns | 444.175 | 2 | 222.0875 | 0.216165444 | 0.805763 | 3.034414 | |  |
| Interaction | 3393.808333 | 2 | 1696.904167 | 1.651655511 | 0.193959 | 3.034414 |  | |
| Within | 240410.65 | 234 | 1027.39594 |  |  |  |  | |
|  |  |  |  |  |  |  |  | |
| Total | 244976.65 | 239 |  |  |  |  |  | |
|  |  |  |  |  |  |  |  | |

1. **SPSS results**





1. **R results**

> # Run the anova function

> result <- aov( IQ ~ sex \* race, data=people)

> summary(result)

Df Sum Sq Mean Sq F value Pr(>F)

sex 1 728 728.0 0.709 0.401

race 2 444 222.1 0.216 0.806

sex:race 2 3394 1696.9 1.652 0.194

Residuals 234 240411 1027.4

As we can see, the results for the tests performed with different tools are very similar:

|  |  |  |  |
| --- | --- | --- | --- |
|  | EXCEL | SPSS | R |
| df | 239 | 239 | 239 |
| F\_stat | 1.651655511 | 1.652 | 1.652 |
| F\_crit | 3.034414 | - | - |
| p-value | 0.193959 | 0.194 | 0.194 |

**5.     Conclusion**

|  |  |  |  |
| --- | --- | --- | --- |
| P-VALUES RESULTS | | | |
|  | EXCEL | SPSS | R |
| SEX | 0.400767 | 0.401 | 0.401 |
| RACE | 0.805763 | 0.806 | 0.806 |
| SEX \* RACE | 0.193959 | 0.194 | 0.194 |

We see that: **p > α** (for the 3 hypotheses)

We fail to reject the null hypotheses. At a 95% level of confidence, we can conclude that there is no significant difference between the means of the 2 factors and that there is no interaction between them.

# References

IQ test, 2015. *IQ scale.* [Online]   
Available at: https://www.iqtestforfree.net/iq-scale.html  
[Accessed March 2018].

Laerd Statistics, 2013. *Independent t-test using SPSS Statistics.* [Online]   
Available at: https://statistics.laerd.com/spss-tutorials/independent-t-test-using-spss-statistics.php  
[Accessed March 2018].

Laerd Statistics, 2013. *One-way ANOVA in SPSS Statistics.* [Online]   
Available at: https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php  
[Accessed March 2018].

Laerd Statistics, 2013. *Two-way ANOVA in SPSS Statistics.* [Online]   
Available at: https://statistics.laerd.com/spss-tutorials/two-way-anova-using-spss-statistics.php  
[Accessed March 2018].

O'Loughlin, D. E., 2016. *How To... Perform a Two-Way ANOVA in Excel 2013.* [Online]   
Available at: https://www.youtube.com/watch?v=hPG8IvKMXoc  
[Accessed March 2018].

R EXPLORATIONS, 2015. *SIMPLE OUTLIER DETECTION IN R.* [Online]   
Available at: https://rexplorations.wordpress.com/2015/09/05/simple-outlier-detection-in-r/  
[Accessed March 2018].

Salkind, N. J., 2017. *Statistics for People Who (Think They) Hate Statistics: Microsoft Excel 2016.* 4th ed. Los Angeles: Sage Publications.