National College of Ireland

Higher Diploma in Science in Data Analytics

Advanced Business Data Analysis

Continuous Assessment 1 – Non-Parametrical Statistical Tests

Investigating Melanoma Thickness

Lucas Morato

Table of Contents

Introduction

The goal of this project is to use a dataset that fit the necessary conditions to perform non-parametrical statistical tests, using computational tools to perform the calculations and also delivering personal observations and questions in order to practice common challenges a data analyst will face on her/his career.

The documentation of this project includes:

- A word document;
- A pdf version of this report;
- The csv file of the data;
- The R code used, including all the comments.

All documents are available here, at my GitHub.

Dataset

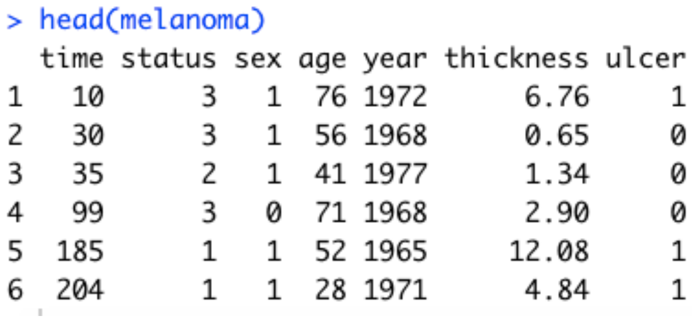
The dataset is available on the following website:

http://vincentarelbundock.github.io/Rdatasets/doc/boot/melanoma.html

It has data about melanoma, a common type of skin cancer, including several variables, that are described below (description found at the data website):

Name	Description
time	Survival time in days since the operation, possibly censored.
status	The patient's status at the end of the study. 1 indicates that they had died from melanoma, 2 indicates that they were still alive and 3 indicates that they had died from causes unrelated to their melanoma.
sex	The patient's sex; 1=male, 0=female.
age	Age in years at the time of the operation.
Year	Year of operation.
Thickness	Tumor thickness in mm.
ulcer	Indicator of ulceration; 1=present, 0=absent.

The dataset has 7 variables and 205 rows, each row representing one individual and unique patient. Let's have a look on the first columns:



```
> head(melanoma)
  time status sex age year thickness ulcer
1   10      3   1  76 1972      6.76     1
2   30      3   1  56 1968      0.65     0
3   35      2   1  41 1977      1.34     0
4   99      3   0  71 1968      2.90     0
5  185      1   1  52 1965     12.08     1
6  204      1   1  28 1971      4.84     1
```

Wilcoxon's Independent Test

Question

Investigate if there’s difference between thickness of melanoma for patients who died from melanoma, based on gender.

Why is this question relevant?

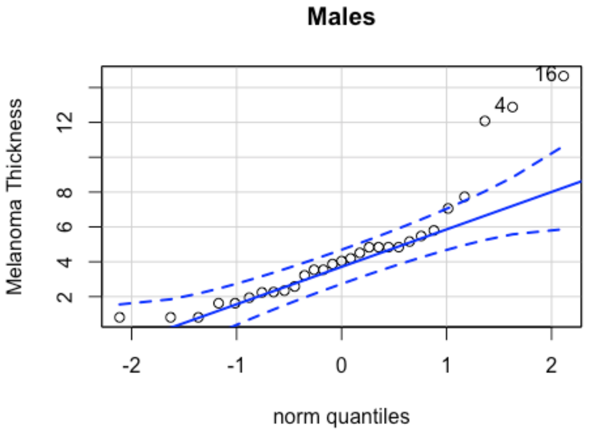
If it’s found that the thickness of the melanoma, that means, how deep it is inside skin tissue, is different between males and females who died from this type of cancer, it could be used on new diagnosis techniques, like a new hint for doctors about how aggressive the cancer is. Example, melanoma of the same thickness could represent different degrees of cancer aggressivity for men and women, and this kind of knowledge could be part of new strategy to treat with correct doses of medicine patients based on their gender.

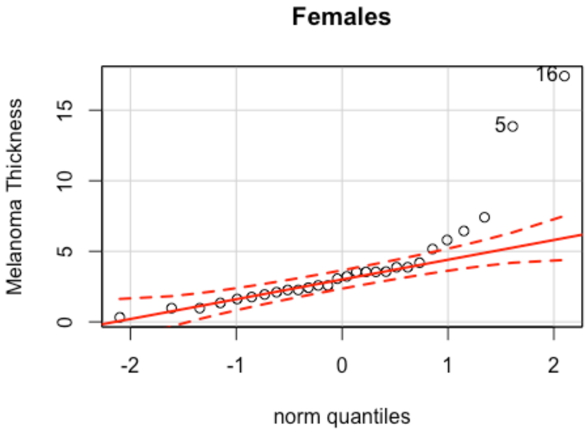
Null Hypothesis: the median difference between pairs of melanoma thickness for males and females is zero.

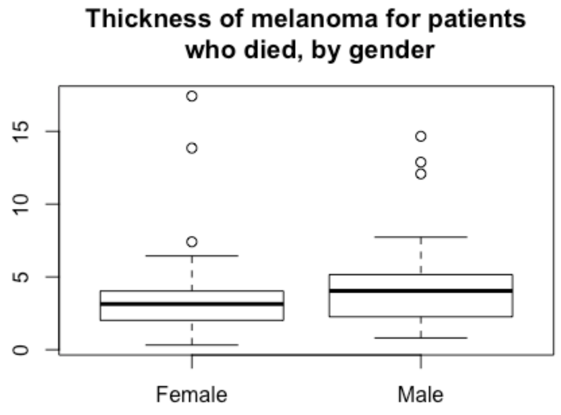
Alternative Hypothesis: there is a median difference between the pairs of melanoma thickness between males and females.

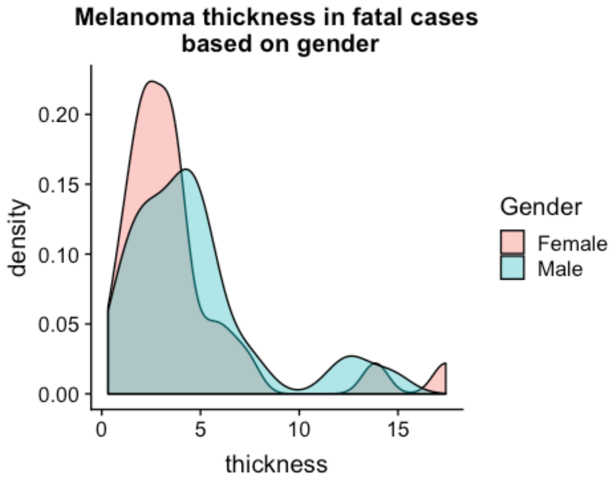
Auxiliary Stats

We are going to split the data first taking just patients who died from melanoma, then creating two groups based on gender. After, we are going to investigate how ‘thickness’ is distributed based on ‘sex’ and check the normality of this distribution. To do that, we’re going to use qqplot, boxplot, density plot and Shapiro-Wilk Test.

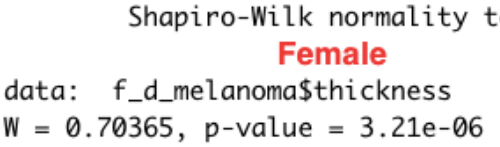




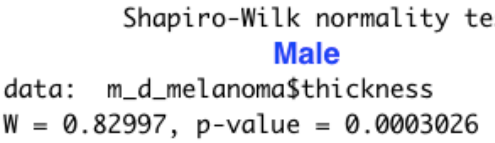




The qqplots showed that we have outliers both for males and females, what was confirmed on the boxplot. The density plot does not indicate normality neither. Let’s double check that with a Shapiro-Wilk test:



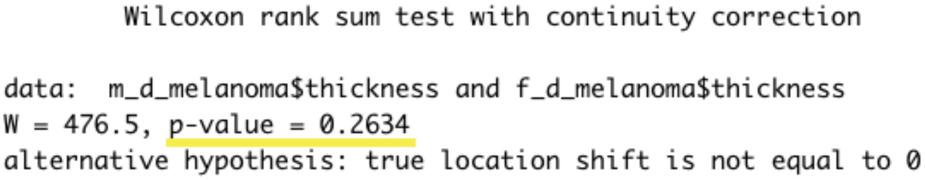
```
        Shapiro-Wilk normality test
                   Female
data:  f_d_melanoma$thickness
W = 0.70365, p-value = 3.21e-06
```



```
        Shapiro-Wilk normality test
                   Male
data:  m_d_melanoma$thickness
W = 0.82997, p-value = 0.0003026
```

For both groups we have found p values lower than our alpha of 0.05, what rejects the null hypothesis that the group's data are normally distributed. The recommend test on this case is Wilcoxon's, and it will be an independent test as we are not testing the same group twice.

Test output



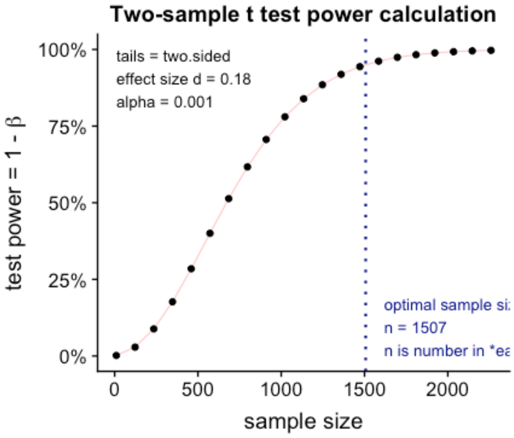
```
        Wilcoxon rank sum test with continuity correction

data:  m_d_melanoma$thickness and f_d_melanoma$thickness
W = 476.5, p-value = 0.2634
alternative hypothesis: true location shift is not equal to 0
```

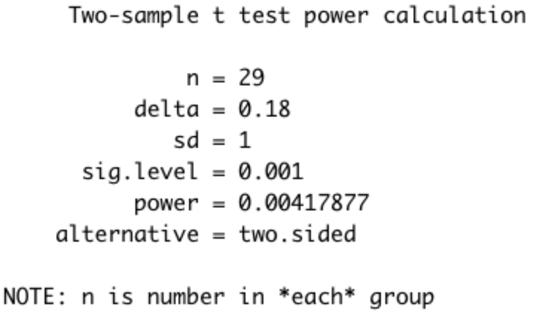
Formal Report

A Wilcoxon Signed-ranks test indicated that melanoma thickness does not have significant difference between males (*Mdn* = 4.04) and females (*Mdn* = 3.14) who died from this type of cancer, Z = 1.12, p = 0.26.

Power Test

We are going to do a Cohen test to determine the d value, then use that information into a power test to determine the sample size need for a power of 95% and significance level of 0.001. We are also going to perform a power test with the same parameters using n = 28 (the smallest group size we have) and check the power.





```
     Two-sample t test power calculation

              n = 29
          delta = 0.18
             sd = 1
      sig.level = 0.001
          power = 0.00417877
    alternative = two.sided

NOTE: n is number in *each* group
```

The conclusion is that we need 1 507 people per group. As we are testing two groups, that makes the total of 3 014 people needed for this test using real-world parameters.
Using the same significance level, we discover that our current power is less than 1%, when the recommend power is at least of 80%.

Interpretation of the Result

The original question was if there's difference on melanoma thickness between males and females who died from this cancer. With the result that there's no significant difference, it is possible to conclude:

- We can investigate more about how the measurement of thickness can help on the diagnosis, or even use it to create a rank about the aggressivity of the cancer based on that factor combined with other variables, like ulcer or how long the person lived after removing it before dying from the disease. One of the possible hypotheses could be that thickness is bigger for the ones who died less than a year after removing it, and so on.

- Now that we discovered that gender of people who died from skin cancer doesn't seem to change the thickness of the melanoma, another research could be conducted with the groups of people to survived and the ones who survived but died from unrelated reasons.

- The power test showed that our groups are too small to allow us to have a strong conclusion, and that our test doesn't have an strong power, so we should re-do that using larger groups to have reliable results.

Who care about the results?

Doctors and researchers from different areas could benefit from that result, as it discards a possible misconception based simply on the gender of the patient suffering from skin cancer, and open doors for more investigations. The patients of course benefit from the results, as it adds more knowledge about their disease.

Kruskal-Wallis test

Question

We are keep investigating melanoma thickness. This time we are going to split our data in three groups:

- Group 1: Patients who died from melanoma;
- Group 2: Patients who survived the melanoma and are still alive;
- Group 3: Patients who survived the melanoma but died for unrelated reason.

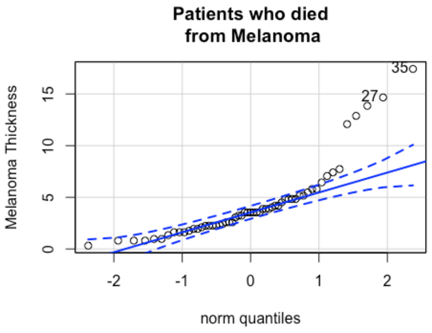
We want now to discover if there's any difference between the thickness depending on the status of the patient.

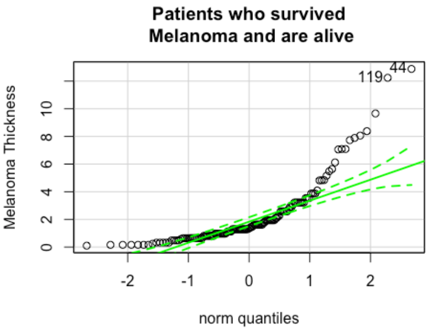
Why is this question relevant?

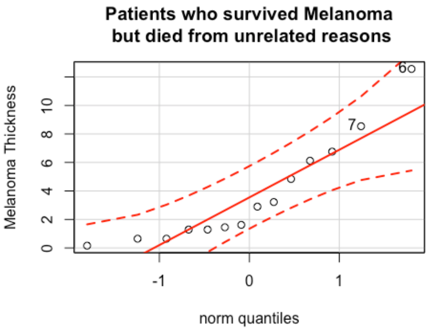
Like on the previous case, this question can open new ways to understand the effect of thickness of the melanoma as a factor of gravity based on the historical data. With this initial step, we can investigate more to try to find if deeper melanomas are an indicative of mortality and use this information to try to help patients according.

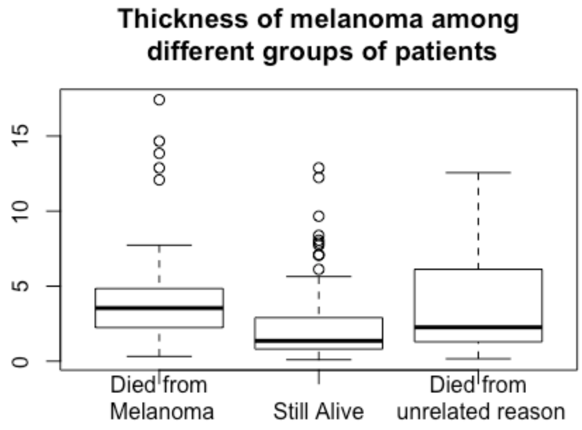
Auxiliary Stats

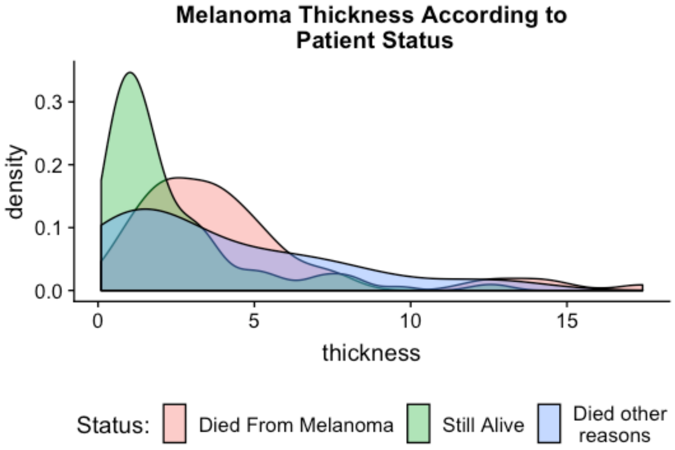
As before, we're going to use qqplot, boxplot, density plot and Shapiro test to determine normal distribuition:



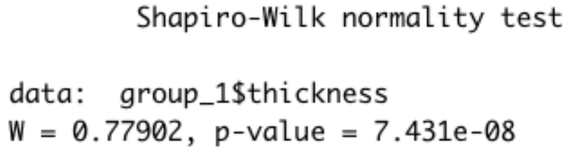




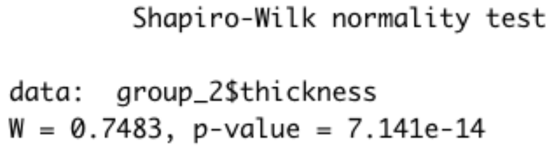




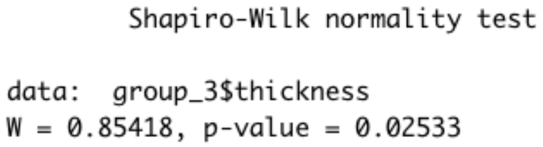
The plots showed non-parametrical data, with visible outliers on the qqplot and boxplot, besides the group of people who died from other reasons (group 3). The density plot also indicates that we need a non-parametrical test. We will double check with Shapiro test for the three groups.



```
        Shapiro-Wilk normality test

data:  group_1$thickness
W = 0.77902, p-value = 7.431e-08
```



```
        Shapiro-Wilk normality test

data:  group_2$thickness
W = 0.7483, p-value = 7.141e-14
```

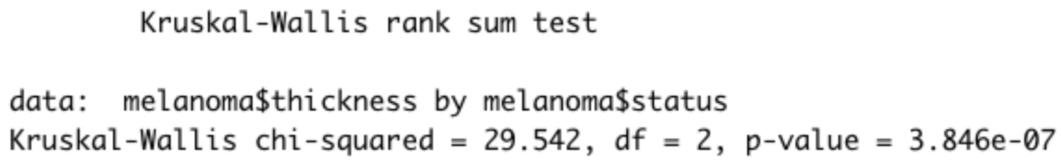


```
        Shapiro-Wilk normality test

data:  group_3$thickness
W = 0.85418, p-value = 0.02533
```

For the groups 1 and 2 we had very small p-numbers, validating what it was suggested before. For the third groups, the p-value was 0.02, closer to alpha (0.05), so it was also on the rejection zone and we can assume the three groups are not normally distributed. With that information, we will choose to perform a Kruskal-Wallis test.

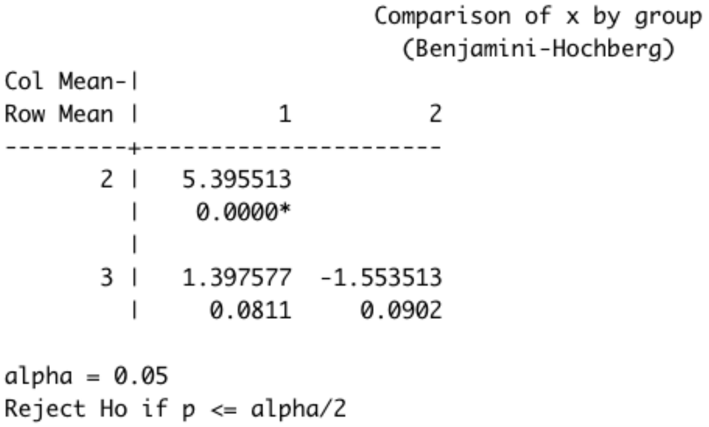
Tests output

Kruskal-Wallis test:



```
        Kruskal-Wallis rank sum test

data:  melanoma$thickness by melanoma$status
Kruskal-Wallis chi-squared = 29.542, df = 2, p-value = 3.846e-07
```

Dunn test (Post-Hoc):



```
                           Comparison of x by group
                             (Benjamini-Hochberg)
Col Mean-|
Row Mean |          1          2
---------+----------------------
       2 |   5.395513
         |    0.0000*
         |
       3 |   1.397577  -1.553513
         |     0.0811     0.0902

alpha = 0.05
Reject Ho if p <= alpha/2
```

Formal Report

A Kruskal-Wallis test indicated a statistically significant difference between the thickness of melanoma between patients with different status (Chi-Squared = 29.54, p < 0.001, df = 2). Dunn's pairwise post-hoc test showed a significant melanoma thickness difference between patients who died from melanoma (group 1) and patients who survived the disease and are still alive (group 2).

Interpretation of the Result

On this test we have found some important information regarding melanoma thickness:

- There is a difference between people who died and survived the disease. It could be a sign that a correlation exists between thickness and mortality, which could be explored in further research;

- There is not a significant difference between who survived and who died from unrelated reasons. It could be interpreted like the melanoma didn't affect the post-life of the patients who were cured.

Who cares about those results?

This result, if conclusive, basically affects the life of all people enrolled into cancer research. As on the previous test, it could be used as a tool to identify or classify the gravity of a cancer based on the thickness of the melanoma lesion, permitting a more assertive treatment prescription and maybe saving more people.

Multi-linear Regression

Question

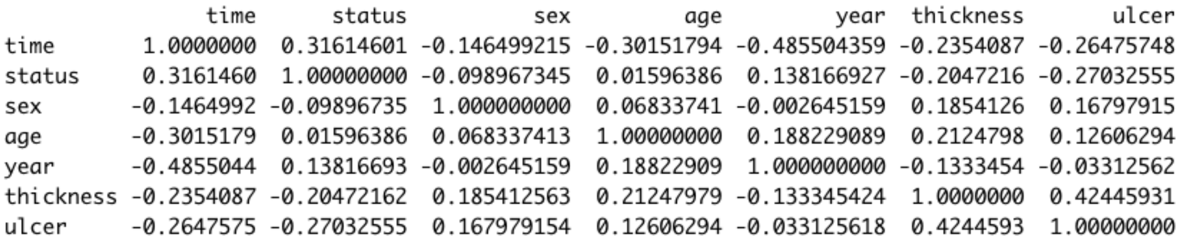
Develop a model to predict melanoma thickness based on its correlation within other variables.

Why it is relevant?

If other statistical tool can help to understand the effect of melanoma thickness, the multi linear regression can help to predict it based in the other variables available. The prediction can help to diagnose more aggressive cancer depending on the outcome, and also it can be changed over time to study accuracy and investigation about different correlation within the variables.

Auxiliary Stats

Let's have a look on the correlation between variables:

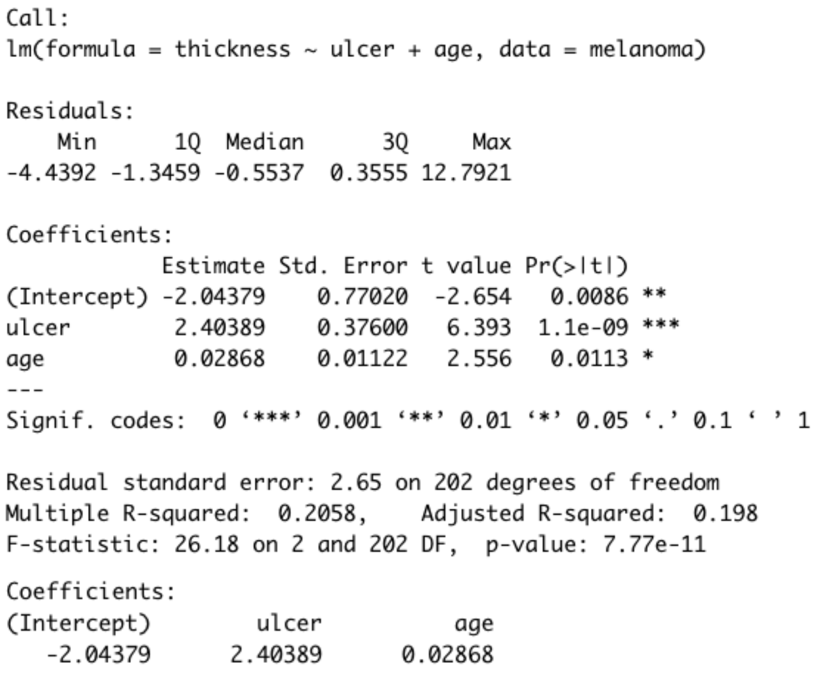


```
                time      status          sex         age          year  thickness       ulcer
time       1.0000000  0.31614601 -0.146499215 -0.30151794 -0.485504359 -0.2354087 -0.26475748
status     0.3161460  1.00000000 -0.098967345  0.01596386  0.138166927 -0.2047216 -0.27032555
sex       -0.1464992 -0.09896735  1.000000000  0.06833741 -0.002645159  0.1854126  0.16797915
age       -0.3015179  0.01596386  0.068337413  1.00000000  0.188229089  0.2124798  0.12606294
year      -0.4855044  0.13816693 -0.002645159  0.18822909  1.000000000 -0.1333454 -0.03312562
thickness -0.2354087 -0.20472162  0.185412563  0.21247979 -0.133345424  1.0000000  0.42445931
ulcer     -0.2647575 -0.27032555  0.167979154  0.12606294 -0.033125618  0.4244593  1.00000000
```

We have found a small negative correlation between thickness/time and thickness/year, and a small positive correlation between thickness/age. Normally we should use strong correlations to build a linear regression model, with coefficients higher than 0.8. However, for the purpose of this project we are going to use the available data.

I will build my model to predict thickness based on ulcer and age.

Test Output



```
Call:
lm(formula = thickness ~ ulcer + age, data = melanoma)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4392 -1.3459 -0.5537  0.3555 12.7921 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.04379    0.77020  -2.654   0.0086 ** 
ulcer        2.40389    0.37600   6.393  1.1e-09 ***
age          0.02868    0.01122   2.556   0.0113 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 202 degrees of freedom
Multiple R-squared:  0.2058,	Adjusted R-squared:  0.198 
F-statistic: 26.18 on 2 and 202 DF,  p-value: 7.77e-11

Coefficients:
(Intercept)        ulcer          age  
   -2.04379      2.40389      0.02868  
```

Formal Report

A multiple linear regression was calculated to predict thickness based on the presence of ulcer in the melanoma and age of patient. A significant regression equation was found (F(2, 202) = 26.18, $p < 0.001$), with R^2 of 0.20 . Patient’s predict tumor thickness is equal - 2.04 + 2.40(ulcer) + 0.03(age), where ulcer is coded as 1= present, 0= not present, and age is measured in years. Patients’ tumor thickness decreased 2.04 millimeters for each year added to age of patient and increased 2.4 millimeters when ulcer is present on the tumor. Both ulcer and age were significant predictors of thickness.

Interpretation of the Results

The R squared of our model is way too low to be reliable, $R^2 = 0.2$, but we have found that a correlation exists between the variables.

This model could be perfectioned to be more efficient. On that case, the results could have multiple uses:

- Fill gaps in large datasets with reliable values, so data can be used in more machine learning models.
- Predict the thickness of a melanoma tumor, combined with previous studies about the effect of thickness on patient’s health, could help on better diagnosis and give to doctors more treatment options.

This model could be also reshaped with other variables, combined with the variable with higher correlation on our case, the presence of ulcer (0.42). Also, as we didn’t find any relevant correlation between our variables, we could try to gather more data, like ethnicity, use of sunscreen before the tumor, geographical area, etc., so this study would have a more reliable model that could be use in real world.

Two Way Anova

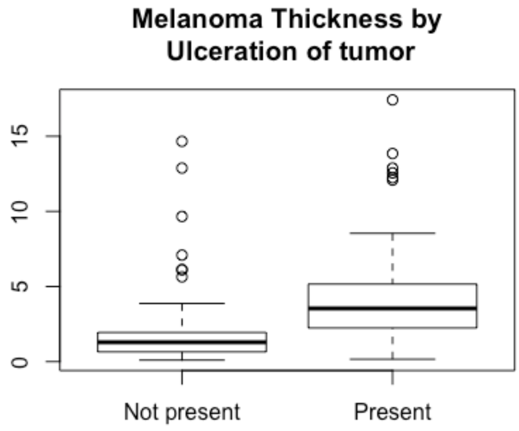
Question

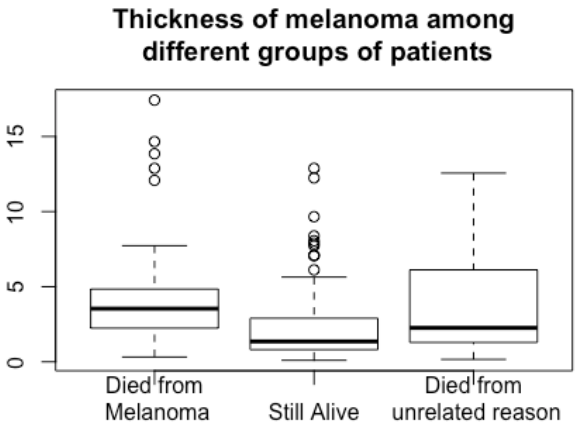
Find the effect of ulcer and patient status, separated and combined, on the thickness of melanoma tumor.

Why is this question relevant?

Unlike the first two test we performed on this report, this test offers a result that shows the effect of two variables at the same time. The result would be important to determine the interaction between different variables over melanoma thickness.

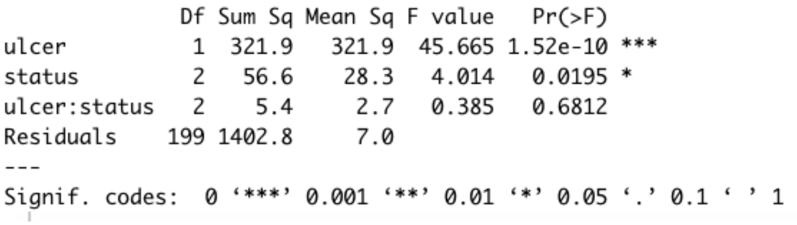
Auxiliary Stats





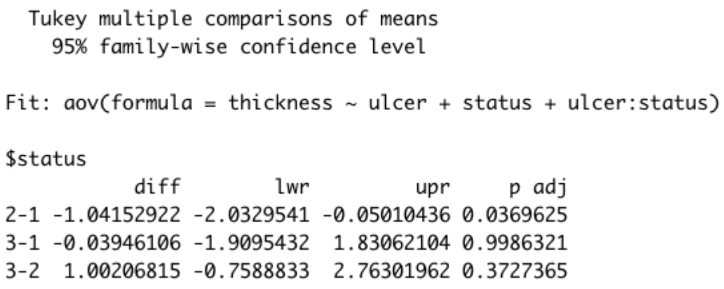
Besides the outliers indicating non-parametrical data, we can assume the tendency of normality, according to the tendency of values to cluster around it’s mean/median/mode. That’s the reason we are going to perform a two-way ANOVA.

Test Output

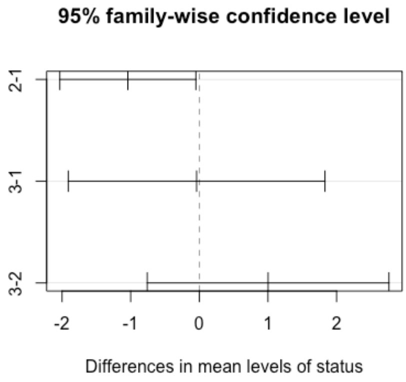


```
              Df Sum Sq Mean Sq F value   Pr(>F)    
ulcer          1  321.9   321.9  45.665 1.52e-10 ***
status         2   56.6    28.3   4.014   0.0195 *  
ulcer:status   2    5.4     2.7   0.385   0.6812    
Residuals    199 1402.8     7.0                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Post Hoc



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = thickness ~ ulcer + status + ulcer:status)

$status
           diff        lwr         upr     p adj
2-1 -1.04152922 -2.0329541 -0.05010436 0.0369625
3-1 -0.03946106 -1.9095432  1.83062104 0.9986321
3-2  1.00206815 -0.7588833  2.76301962 0.3727365
```



Formal Report

A factorial ANOVA was conducted to compare the main effects of two independent variables (presence of ulcer on melanoma tumor and the status of the patient), and the interaction effect between them on the thickness of tumors. Presence of ulcer included two levels (0= no ulcer, 1= ulcer on tumor), and status consisted of three levels (1= died of melanoma, 2= survived and alive, 3= survived and died from unrelated reasons). The independent variables were statistically significant at the 0.05 significance level. The interaction effect was not significant, F(2,199) = 0,38, p = 0,68. A Tukey post hoc test showed that the status 1 and 2 differed significantly at $p < 0.05$.

Interpretation of the Results

The result that the interaction between ulcer and status is not significant, indicates that the relationship between status and melanoma thickness doesn't depend on the ulcer status.

This result is important to avoid decisions based on empirical observations from doctors, preventing incorrect assumptions based on the presence of ulcer. Tests like that one have great value to validated or not medical diagnosis, and could be used as an important tool for medical research, and the interaction between independent variables could lead to more robust conclusions on the depth, or thickness, of melanoma tumor, as part of research to prevent and treat this disease.

References

McDonald, J.H. (2014). *Handbook of Biological Statistics.* 3rd edn. Available at http://www.biostathandbook.com/ [Accessed 25 June 2019].