**Lead concentration Statistical Analysis**

BMD407-Final Project

|  |  |  |
| --- | --- | --- |
| **Group B** | | |
| Mennatullah Mahmoud Amin |  | 18101713 |
|  |  |  |
| Hesham Atef Abdelfattah |  | 18101682 |
|  |  |  |
| Abdelrahman Mohamede Elgebaly |  | 18102591 |
|  |  |  |
| Amr Mohamed Aboalsaoud |  | 18101688 |
|  |  |  |
| Yousef Emad Gadallah |  | 18100565 |

## Descriptive statistics

After reading the data “RDATA” File using load() and get() functions, multiple functions such as str(), summary(), and summarize\_data() “Implemented from scratch function” are all used for :

* Identifying and exploring the data
* Knowing more information about columns data types
* Summarizing the data: calculating mean, median, minimum, maximum, first and third quartile.

Multiple Factors “factor()” and tables “table()” functions are used in order to deal with categorical data and create frequency tables. “Area”, “Sex”, “Lead type”, and “Exposed” columns are such examples for existing categorical data.

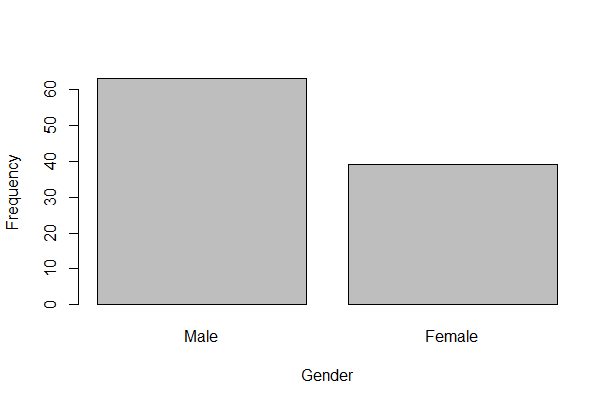
## Calculating correlation coefficient

Calculating correlation coefficient process “for (MAXWT and *Ld72)* and (MAXWT and Ld73)” had been done considering the following information:

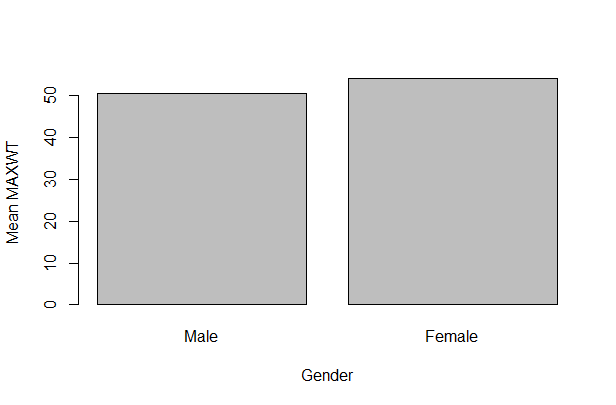
* **For calculating correlation coefficient, there are 3 methods:**
* "Pearson"/Assumptions: Normality/no outliers/linear relationship/homoscedasticity/continuous
* "Spearman"/Assumptions: Monotonic relationship "more robust to outliers"/ordinal
* “Kendall"/Assumptions: Similar to Spearman "More robust"
* Scatter-plots are used to know if the relationships between the specified relationships are linear, monotonic, or non-monotonic
* Shapiro-test used in order to check the normality of the specified columns. The results of the test suggested that the data seems to be not normally distributed, so “Spearman” was selected as an appropriate method in this case.
* Correlation coefficient results came like the following:
* The correlation coefficient (MAXWT and Ld72)
* "Output: negative correlation/ weak correlation/ weak relationship/ less than 0.5 (-0.3984387)"
* "Negative relationship: when x increase, y decrease and vice versa"
* The correlation coefficient (MAXWT and Ld73)
* "Output: negative correlation/ weak correlation/ weak relationship/ less than 0.5 (-0.3182327)"

## Graphics

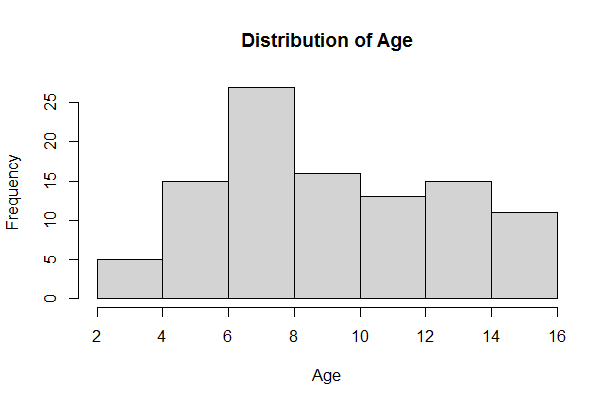
* Bar chart of a categorical variable for the gender (Sex parameter) was plotted in order to know which gender was with the highest frequency “Number of Males (gender 1) >Number of Females (gender 2)”.

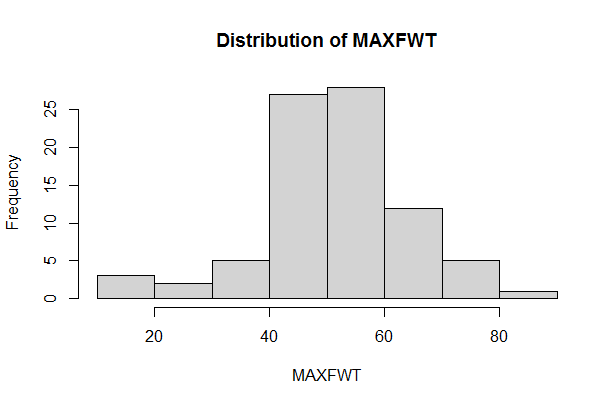


* Bar chart graph with mean MAXWT in males and females was generated in order to visualize and know which gender have the higher mean of finger tapping “The mean MAXFWT of Females was higher than the mean MAXFWT of Men”.

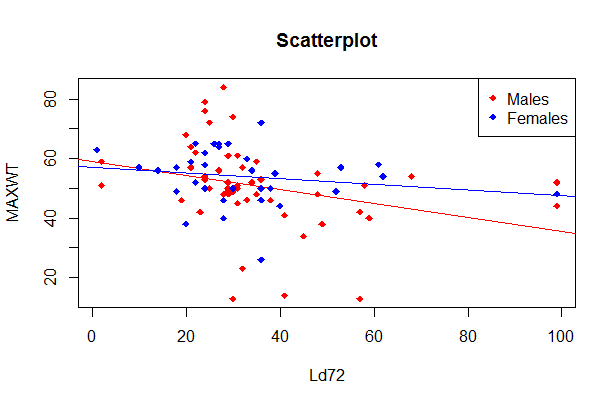


* A histogram of a continuous variable: “*age*” as well as “MAXWT” allowed us to know the spread and the normality of these variables “Both histograms showed that the data seems to be not normally distributed”

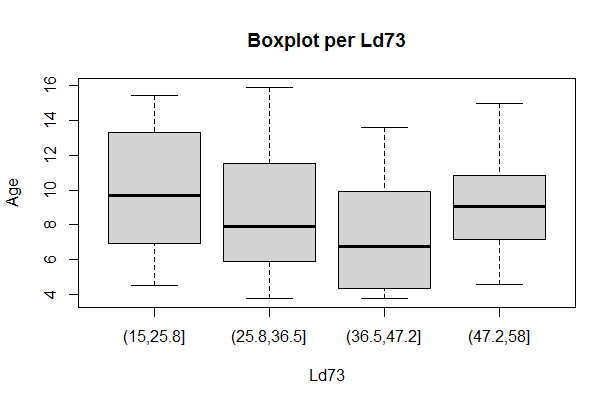


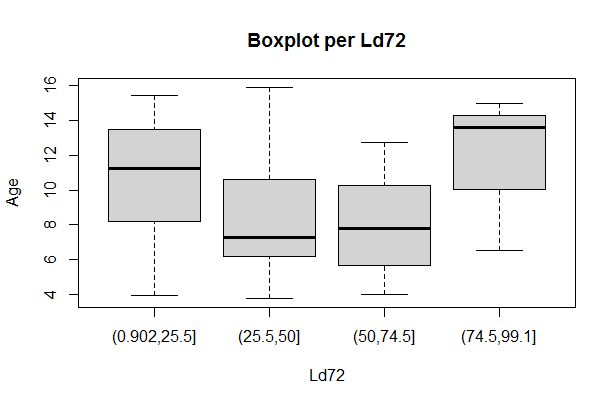


* A scatterplot of 2 continuous variables *Ld72* and MAXWT, and the regression lines for each gender was generated for allowing us to know the relationship between MAXWT “Outcome/dependent” and the Ld72 “Regressor/Independent” which seems to be a negative relationship (Negative Slope) “If Ld72 Increases, MAXWT will decrease”



* Separate box-plots such as “Age per Ld72” and “Age per Ld73” were generated in order to see the variation of the data. “Cut () function for dividing Ld73 data into “4” intervals”

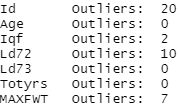


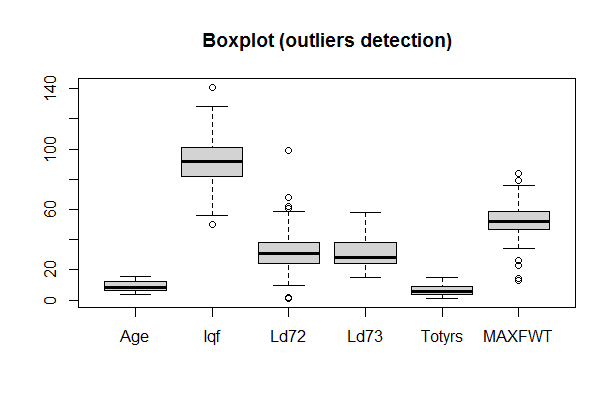


## Outlier detection

## Exploring the data for any existing outliers

For detecting outliers Step, A loop was implemented in order to get all the numeric classes and plotting a box-plot for each column. Box-plots are very helpful in testing homoscedasticity, checking variation, and detecting outliers. In addition, detect outliers () function implemented in order to detect and know the number of outliers in each column in the data frame. The results came as the following: “there are some outliers found in "iqf/Ld72/MAXFWT" columns”





## Testing for normality/ homoscedasticity

First, we started to check the normality by two different methods with is Histogram to visualize the data also by Shapiro test to test the normality statistically.

H0: sample distribution is normal

H1: sample distribution is not normal

* The p value between the gender “Male” and MAXFWT is 0.005127 which is less than the value of the alpha which means we that we have evidence to reject the null hypothesis and assume it is not normally distribution between them.
* The p-value between the gender “Female” and MAXFWT is 0.2299which is more than the value of the alpha which means we that we do not have evidence to reject the null hypothesis and assume it’s normally distributed.
* The p-value of MAXFWT is 0.0005636 which is less than the value of the alpha which means we that we have evidence to reject the null hypothesis and assume it is not normally distribution between them.
* The p-value of the Age is 0.0004677 which is less than the value of the alpha which means we that we have evidence to reject the null hypothesis and assume it is not normally distribution between them.
* The p-value of Ld72 is 3.188e-08 which is less than the value of the alpha which means we that we have evidence to reject the null hypothesis and assume it is not normally distribution between them.

Secondly, the homoscedasticity we started to check it by the boxplot as a visualized method and also used Levene test to check the variance statistically.

H0: all groups variances are equal(homoscedastic)

H1: at least the variances of two groups are different (the variance is not the same for the all the groups) (not homoscedastic)

* The pr(f-value) of MAXFWT is 0.1778 which is more than the value of the alpha which means we that we do not have evidence to reject the null hypothesis and assume it is homoscedastic.
* The pr(f-value) of Sex is 0.389 which is more than the value of the alpha which means we that we do not have evidence to reject the null hypothesis and assume it is homoscedastic.
* The pr(f-value) of Ld72 is 0.8542 which is more than the value of the alpha which means we that we do not have evidence to reject the null hypothesis and assume it is homoscedastic.
* The pr(f-value) of Ld73 is 0.9177 which is more than the value of the alpha which means we that we do not have evidence to reject the null hypothesis and assume it is homoscedastic.

1. **Statistical Inference**

Confidence Intervals

* A confidence interval is an interval that contains the population parameter with probability Text

  Description automatically generated
* Confidence interval takes on the form: Text

  Description automatically generated
* Where   is the value needed to generate an area of α/2 in each tail of a t-distribution with n-1 degrees of freedom and   is the standard error of the mean.

**Calculating a confidence interval**

To calculate a confidence interval, we need the following steps:

1. **Calculate the mean**

* Use the mean() command to calculate it.

1. **Calculate the standard error of the mean**

The formula for the standard error of the mean is

Graphical user interface

Description automatically generated with low confidence

and if we do not know the population standard deviation



* The sd() command can be used to find the standard deviation. The length() command can be use to determine the sample size.

1. **Find the t-score that corresponds to the confidence level**

* We need to have  probability in the lower and upper tails, we divide by two because there are two tails. The qt() command will calculate the t-score,



1. **Calculate the margin of error and construct the confidence interval**

* The margin of error is Text

  Description automatically generated
* The confidence interval is the mean +/- margin of error.

.

**Result**

|  |  |  |
| --- | --- | --- |
| **Confidence** **level** | **Male** | **Female** |
| **90 %** | **47.58566 : 53.77434** | **51.59610 : 56.64632** |
| **95%** | **46.97567 : 54.38433** | **51.08921 : 57.15321** |
| **99 %** | **45.75538 : 55.60462** | **50.06003 : 58.18240** |

* **Male**

**Chart, histogram

Description automatically generated**

**Fig 1 : 90% C.I for males**

**Chart, histogram

Description automatically generated**

**Fig 2 : 95% C.I for males**

**Chart, histogram

Description automatically generated**

**Fig 3 : 99% C.I for males**

* **Females**

**Chart, histogram

Description automatically generated**

**Fig 1 : 90% C.I for females**

**Chart, histogram

Description automatically generated**

**Fig 2 : 95% C.I for females**

**Chart, histogram

Description automatically generated**

**Fig 3 : 99% C.I for females**

* Applying the 95 percent rule, the result show that we can be 95 percent confident that males has range between 46.97567 : 54.38433 and females range between 51.08921 : 57.15321. which indicate that females has *MAXFWT better values rather than males ,because(* The higher the score the better).
* When request higher confidence the range is being larger .

1. **Hypothesis testing**

### We hypothesis that MAXWT is different between male vs female. Assuming normality and homoscedasticity, can you test this hypothesis using statistical hypothesis framework.

1. Firstly, we state the research question:

Does MAXWT is different between male vs female?

1. We convert the research question to a statistical one:

Does the mean of the group male differ from the mean of the group females in MAXWT?

1. We state the null and alternative hypothesis:

* Null hypothesis: there is no difference between the 2 groups
* Alternative hypothesis: there is a difference between the 2 groups

1. we calculate p-value: if normality and homoscedasticity are assumed

two-sample t-test will be used.

1. Conclusion: p-value = 0.2364, which is greater than the alpha (=0.05). our result is not significant, and we do not have enough evidence to reject the null hypothesis. Thus, we do not have evidence to say that MAXFWT is different between male and female

### Assess whether the previous test assumptions have been met for the test.

1. We checked normality of males and females by QQplot, histogram and Shapiro test.

In male, QQplot’s line has deviation from data points, histogram is not normally distributed, Shapiro test p-value equals 0.005127, which is lower than the alpha so we have enough evidence to reject the null hypothesis, thus it is not normal. Opposite for females, QQplot’s line has no deviation from data points, histogram is almost normally distributed, shapiro test result is 0.2299 which is greater than the alphas o we do not have enough evidence to reject the null hypothesis, thus it is normal. In conclusion, we assumed that our sex data is not normally distributed.

1. We checked variance of sex column by levene test because it does not assume normality and more robust. Our p-value was 0.1778, it is greater than the alpha, so we do not have enough evidence to reject the null hypothesis, thus it is homoscedasticity

Therefore, our data is not normal and homoscedasticity so not all assumptions are met. And it is better to use Mann Whitney rank-based test instead of two sample t-test. After using man whitney test, our p-value is 0.1399, which is greater than the alpha (=0.05). our result is not significant, and we do not have enough evidence to reject the null hypothesis. thus, we do not have evidence to say that MAXFWT is different between male and female

### We hypothesis that MAXWT is “lower” in the group receiving *Ld72 > 40* compared to the control *Ld72 =< 40*. Can you test this hypothesis assuming heteroscedasticity?

1. As the previous question, we state the research question: Does MAXWT is "lower" in the group receiving Ld72 > 40 compared to the control Ld72 =< 40?
2. We convert the research question to a statistical one:

Does the mean of the group ld72>40 lower than the mean of the group ld72=<40 in MAXWT?

1. We state the null and alternative hypothesis:

- Null hypothesis: MAXWT is not "lower" in the group receiving Ld72 > 40 compared to the control Ld72 =< 40

- Alternative hypothesis: MAXWT is "lower" in the group receiving Ld72 > 40 compared to the control Ld72 =< 40

1. we will test normality by QQplot, histogram and Shapiro test in the 2 groups which are ld72>40 (receiving) and ld72 =<40 (control)

In receiving group, the QQplot’s line has deviation from data points, histogram is not normally distributed, Shapiro test p-value equals 0.005391 which is lower than the alpha, so we have enough evidence to reject the null hypothesis, thus it is not normal. Same as in control group, the QQplot’s line has deviation from data points, histogram is not normally distributed, Shapiro test p-value equals 009785 which is lower than the alpha, so we have enough evidence to reject the null hypothesis, thus it is not normal.

1. we calculate p-value: if heteroscedasticity are assumed, and being not normal

Mann Whitney rank-based test was used, its p-value is 0.003036, which is smaller than the alpha(=0.05), then we have the evidence to reject the null hypothesis, thus we have enough evidence to say that there the group receiving Ld72 > 40 is lower than the control Ld72 =< 40.

### Assess the previous test assumption

We checked variance of Ld72 by levene test because it does not assume normality and more robust and our F value was 0.7382, it is greater than the alpha, so we do not have enough evidence to reject the null hypothesis, so it is homoscedastic

To conclude, it is not normal and homoscedastic, so the assumptions are not met. But Mann Whitney rank-based test is also the best for this case.

### We hypothesis that MAXWT is different between the different Lead types with the different genders (i.e., 4 groups male\_leadtype1, male\_leadtype2, female\_leadtype1, female\_leadtype2). Can you perform comparison between the different groups, after assessing the assumptions and performing post-hoc testing (assuming normality and homoscedasticity)?

1. We state the research question:

Is MAXWT different between the different Lead types with the different genders

1. We convert the research question to a statistical one:

Does the mean of the group MAXWT differs from the mean of the group Lead types and gender?

1. We state the null and alternative hypothesis:

- Null hypothesis: there is no difference between groups

- Alternative hypothesis: there is a difference between groups

d- we check the normality of sex and lead type

* We checked the four groups (male leadtype 1, male leadtype 2, female leadtype1, female leadtype 2)
* We found that male leadtype 1 and male leadtype 2 are not normally distributed while the others are normally distributed
* Thus, we considered it not normal.

1. We calculate p-value: our data is not normal, so we used kruskal wallis,

* p-value = 0.004795 which is smaller than the alpha(=0.05).Our result is below significant and we have enough evidence to reject the null hypothesis .thus, we have evidence to say that MAXFWT is different between different gender and different lead types

1. Posthoc then will be performed by Tukey test (assuming normality and homoscedasticity):

* We firstly did anova because normality is assumed
* the f-value of the sex is 0.20764, which is greater than the alpha (=0.05), our result is not significant, and we do not have enough evidence to reject the null hypothesis, thus, we do not have evidence to say that MAXFWT is different between different gender
* the f-value of the lead type is 0.00195, which is lower than the alpha (=0.05), our result is significant and we have enough evidence to reject the null hypothesis, thus, we have evidence to say that MAXFWT is different between different lead types
* the f-value of the sex and lead type is 0.11060, which is greater than the alpha (=0.05), our result is not significant, and we do not have enough evidence to reject the null hypothesis, thus, we do not have evidence to say that MAXFWT is different between different gender and different lead type.
* Then we did tukey test
* Tukey results in 6 combinations, each has 4 columns which are 1- difference between the mean of both cases, 2- lower and upper: they are the confidence interval the model is is 95% confident that true population mean of the 2 groups lies in this interval, 3- p-adjust: it is corrected p-value
* p adj of Male:2-Male:1(0.0036208) and Male:2-Female:1(0.0044057) are less than the alpha(=0.05), our result is significant and we have enough evidence to reject the null hypothesis, thus, male lead type 2 with male lead type 1 is different in MAXFWT, and male lead type 2 and female lead type 1 is different in MAXFWT, besides that the confidence interval in these 2 cases doesnt have the 0 value and that also indicates the low p adjusted
* the rest are greater than the alpha (=0.05) so they are not different in MAXFWT and the 0 value lies in the confidence interval and that indicates the high p-adjusted as well.

Chart, box and whisker chart

Description automatically generated

Fig(1) describes each group with his confidence interval

1. **Linear model**

### Fit a linear regression to the data and interpret the regression coefficient (for the one of the hypotheses mentioned above)

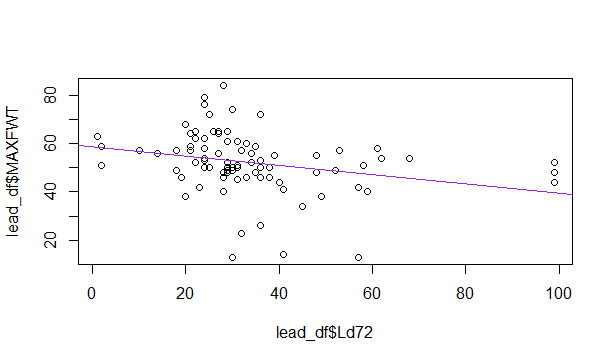
Firstly, we plot the data and fit our linear model to fit the linear regression line on the plot for Ld72 column with MAXFWT.

In our model:

* The residual median is 0.160 which is relatively small, so the model is significant.
* The estimates were 58.4812 for the intercept and -0.1887 for Ld72 (slope), creating our equation as y= 58.4812 -0.1887x.
* The t-test value for the slope is 0.0152 which is smaller than the alpha (=0.05), so our result is significant, and we do have enough evidence to reject the null hypothesis. Thus, there is relationship between our two variables so y can be predicted by x.

(The null hypothesis is that the slope of the variable equal 0 and the alternative differ from 0)

* Multiple R-squared(R^2) equals 0.07053 which is very bad. This value means that LD72 explains 7.053% of the variability in MAXFWT.
* Residual standard error (mean square error) is 12.51



Fig(2) plot of linear regression model

### Calculate and interpret a 95% confidence interval of the regression slope.

we are 95% confident that Ld72 will increase between -0.3400905 and -0.03725285