Nottingham Trent University

Statistical Data Analysis and Visualisation

Name: YASH UDAY SULE

NTU ID: N1040923

Module Code: MATH40031

18th March, 2022

Table of Contents

Table of Figures

Table of Tables

Part 1

a) How many variables are in the dataset?

Answer: In all, there are 6 variables with 90 observations namely Subject, ColourShapeCorrect, GreyShapeCorrect, Gender, Age, Alzheimers Disease.

b) What types of variables are they?

Answer:

Variables	Types	Output obtained in R (using typeof())
Subject	Ordinal	integer
ColourShapeCorrect	Numerical(Quantitative) – Discrete	integer
GreyShapeCorrect	Numerical(Quantitative) – Discrete	integer
Gender	Categorical(Qualitative) – Nominal	character
Age	Numerical(Quantitative) – Discrete (Since we are only considering completed years)	integer
Alzheimers Disease	Categorical(Qualitative) – Nominal	logical

c) Report the summary statistics of the variables.

Answer:
The table shows the summary statistics of each variable from the Alzheimer's Dataset.

Summary Statistics / Variables	Minimum Value	1st Quartile	Median	Mean	3rd Quartile	Maximum Value	Standard Deviation	Variance
Subject	1.00	23.25	45.50	45.50	67.75	90.00	26.12	682.5
ColourShapeCorrect	9.00	31.00	45.00	45.73	61.00	75.00	16.62	276.13
GreyShapeCorrect	50.00	65.25	75.00	73.68	81.75	97.00	10.58	111.99
Age	46.00	59.00	66.00	67.24	74.00	93.00	9.70	94.19

	Female	Male	Total
Gender	48	42	90

	FALSE	TRUE	Total
AlzheimersDisease	45	45	90

d) What is the range of ages in the Alzheimer's Disease patients and healthy controls?

Answer:

	Range of Ages
Healthy Controls:	46 – 72
Alzheimer's Disease patients:	55 – 93

e) Is the dataset balanced? Explain.

Answer: Yes, the dataset provided for the analysis is a balanced dataset. It can be confirmed by counting the number of Alzheimer's Disease patients and Healthy Controls, since, the dataset contains a column of a variable which states whether a particular patient has an Alzheimer's Disease or not having a logical output of TRUE / FALSE. Using some inbuilt functions of R it can be seen that there are 45 patients in each case as shown in Figure – 1 below.

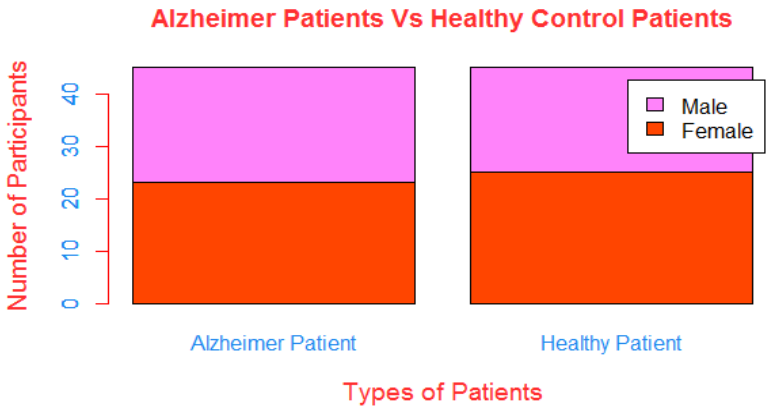


Figure – 1: Barplot of Types of Patients

f) Compute the summary statistics of the variables for Alzheimer's Disease and healthy controls separately. Discuss.

Answer:

The table shows the summary statistics of each variable from the Alzheimer's Dataset for Healthy Controls Patients.

Summary Statistics / Variables	Minimum Value	1st Quartile	Median	Mean	3rd Quartile	Maximum Value	Standard Deviation	Variance
Subject	46	57	68	68	79	90	13.13	172.5
ColourShapeCorrect	40.00	54.00	61.00	60.09	66.00	75.00	8.03	64.40
GreyShapeCorrect	60.00	77.25	81.00	80.24	85.00	97.00	8.11	65.82
Age	46.00	56.00	59.00	60.64	65.00	72.00	5.99	35.92

	Female	Male	Total
Gender	23	22	45

The table shows the summary statistics of each variable from the Alzheimer's Dataset for Alzheimer's Disease Patients.

Summary Statistics / Variables	Minimum Value	1st Quartile	Median	Mean	3rd Quartile	Maximum Value	Standard Deviation	Variance
Subject	1	12	23	23	34	45	13.13	172.5
ColourShapeCorrect	9.00	27.00	31.00	31.38	36.00	48.00	8.52	72.56
GreyShapeCorrect	50.00	60.00	67.00	67.11	72.00	87.00	8.51	72.51
Age	55.00	68.00	74.00	73.84	80.00	93.00	8.09	65.49

	Female	Male	Total
Gender	25	20	45

Part 2

Introduction

In this statistical report, we want to analyse the relationship between cognitive tasks and Alzheimer's disease patients. Our hypothesis is that Alzheimer's disease patients irrespective of sex are deficient at cognitive tasks as they tend to grow older.

The data consists of 90 subjects, out of these 42 were male and 48 were female. Also, among 90 participants 45 participants were having Alzheimer's disease and other 45 people were belonging to the healthy control group. All of them went through 100 trials of two cognitive tasks – one was related to identifying geometrical representations having different colours and another possess only grey coloured shapes. Out of which only the number of correct trials were reported for both the tasks and will be considered for our further analysis.

In order to justify our hypothesis the analysis aspects are divided into four main sections:
Section – 1: Analysing whether Age or Gender is a crucial factor through statistical tests.
Section – 2: Identifying important correlations.
Section – 3: Establishing linear relationships between some variables.
Section – 4: Performing some more statistical tests to support our hypothesis.

Section – 1: Analysing whether Age or Gender is a crucial factor through statistical tests

In order to begin with the analysis part in this section will bifurcate it into two main segments. They are as follows:

Segment – 1: Checking differences in Age between the Alzheimer's disease group and the healthy control group.
Segment – 2: Checking for association of sex between the above mentioned two groups.

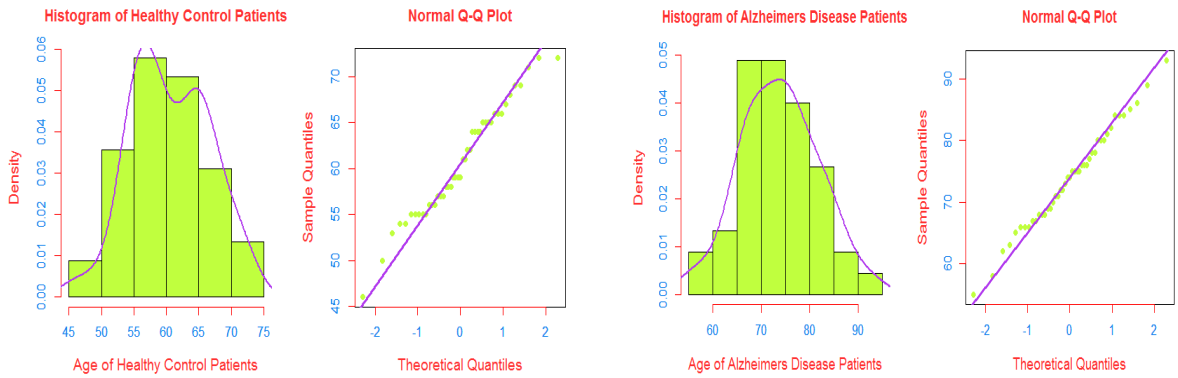
Segment – 1: To carrying out this, will first test the normality of ages for the two groups – Alzheimer's disease group and Healthy Control group separately using the Kolmogorov – Smirnov tests. In this case, normality of two groups was tested against the normal distributions with same means and standard deviations of the data. The null hypothesis and alternative hypothesis was:

	Mathematically
H_0: Both the groups tend to follow normal distribution. H_a: Neither of the groups is normally distributed.	$H_0: G_{ages\_Healthy} \sim N(\mu_1, \sigma_1)$, $G_{ages\_Alzheimer} \sim N(\mu_2, \sigma_2)$ $H_a: G_{ages\_Healthy} \nsim N(\mu_1, \sigma_1), G_{ages\_Alzheimer} \nsim N(\mu_2, \sigma_2)$ where, $\mu_1, \mu_2, \sigma_1, \sigma_2$ are the means and standard deviations of both the groups ($G_{ages\_Healthy}, G_{ages\_Alzheimer}$).

The level of significance was $\alpha = 0.05$. After performing the Kolmogorov – Smirnov tests, the p – value for Healthy control group was 0.5445 while for the Alzheimer's disease group was 0.9955 respectively. Since the p – values are much higher than 0.05, we cannot reject the null hypothesis and thus we can conclude that both the groups have distributions which are approximately normally distributed which suggest us that will have to use the parametric tests for further analysis. Figure – 2 shows the histograms and QQ – plots for both the groups.

We then examine statistically whether there is any difference in the age between the two groups. In this case, will use the Welch t – test (also referred as Two – sample t – test) since the participants in both the groups are different which suggests that the data is an unpaired data. The null and the alternative hypothesis were:

	Mathematically:
H_0: There is no difference in ages of both the groups. H_a: The ages among the two groups are not same.	$H_0: \mu_{ages\_Healthy} = \mu_{ages\_Alzheimer}$ $H_a: \mu_{ages\_Healthy} \neq \mu_{ages\_Alzheimer}$



(a) Healthy control patients (b) Alzheimer's disease patients

Figure – 2: Histograms and QQ – plots for two groups

The p – value from the t – test was 2.021×10^{-13} with 0.05 level of significance. Thus, we reject the null hypothesis and based on this evidence we can conclude that there is a significant difference between the ages of Healthy and Alzheimer's disease patients. Hence we should investigate both the groups separately for analysis. Figure 3 shows the boxplot of the age differences between the two groups.

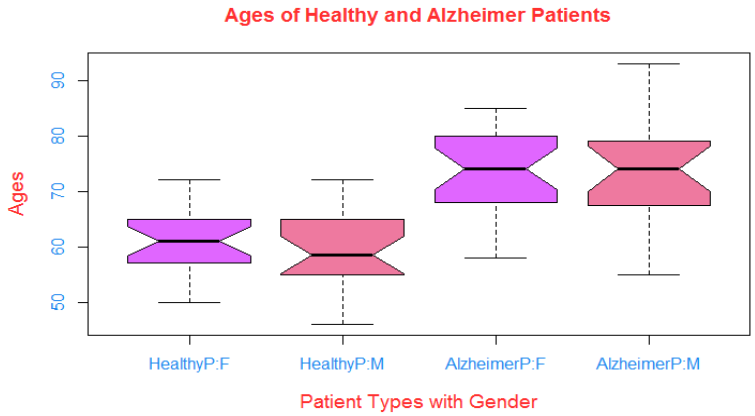


Figure – 3: Classification of Ages for Healthy control patients and Alzheimer's disease patients among male and female.

Segment – 2: In order to begin with this part of analysis will first have to consider the following Table – 1.

	Male	**Female**	**Total**
Healthy control patients	22	23	45
Alzheimer's disease patients	20	25	45
Total	42	48	90

Table – 1: Number of Male and Female among the two groups

Since it is a categorical data will use the Chi – Square distribution test for checking the association of sex among the two groups. In this case the level of significance was $\alpha = 0.05$ and our null hypothesis and alternative hypothesis was:

H_0: There is an association of gender with the types of patients.
H_a: There is no association of gender with the types of patients.

The test result gives the p – value = 0.8327 which states that we do not have much evidence to reject the null hypothesis and we can conclude that there is some association of gender with the two groups.

Section – 2: Identifying important correlations

In this section we will examine whether there are correlations among the data with some specific variables. For this, we had segmented it into two different segments of analysis. They are as follows:
Segment – 1: Determining whether there exist correlation between age and number of correct trials recorded in cognitive tasks.
Segment – 2: Determining whether there exist correlation between gender and number of correct trials recorded in cognitive tasks.

Segment – 1: In this particular case, we had assumed the significance level as $\alpha = 0.05$ and our null hypothesis and alternative hypothesis was as follows:

	Mathematically
H_0: There is no correlation between age and number of correct trials reported in two cognitive tasks – colour shapes and grey shapes.	$H_0: \rho = 0$
H_a: There is a correlation among age and trials reported in the cognitive tasks.	$H_a: \rho \neq 0$

The following results were obtained after applying the tests related to correlation and are summarized in the form of a table in Table – 2:

	Value of ρ	p - value
Case – 1: Correlation between age and colour shape correct variables	-0.8885357	2.2×10^{-16}
Case – 2: Correlation between age and grey shape correct variables	-0.8929053	2.2×10^{-16}

Table – 2: Summarized Table of correlations values and p – values related to age

As it can been seen from Table – 2 that the p – value is less than 0.05 will reject the null hypothesis stating that there is no correlation and also it can be observed that the value of ρ is negative in both the cases. Since it is very nearer to -1 we suggest it is strongly negatively correlated which means that as the age increases the identification of correct number of shapes decreases whether it is grey shaped or colour shaped. Figure – 4 shows the graphical perspective of correlations between the variables.

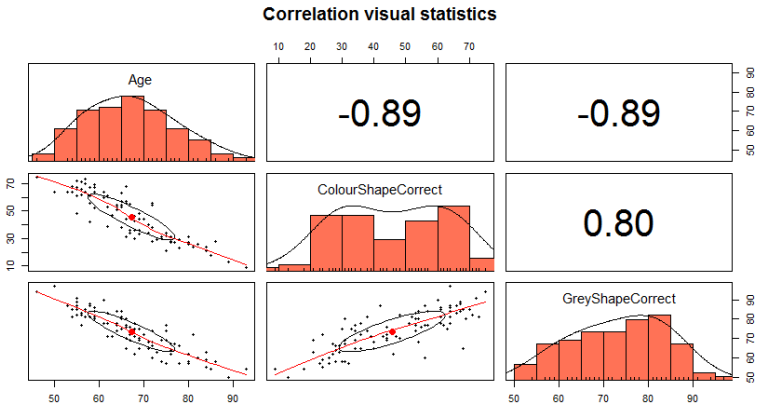


Figure – 4: Pictorial representation of correlations among the variables

Segment – 2: For this, we will use biserial correlation because gender is a dichotomous variable and correct trials are quantitative variables. After applying the biserial correlation function we got the following results which are summarized in Table – 3.

	Value of ρ
Case – 1: Correlation between gender and colour shape correct variables	0.04339901
Case – 2: Correlation between gender and grey shape correct variables	-0.01156917

Table – 3: Summarized Table of correlations values related to gender

Therefore, it can be concluded that there is a definite negative correlation between age and correct number of trials but since the value of $\rho = 0$ in both the cases of gender (as observed in Table – 3) there is no association of gender with correct trials.

Section – 3: Establishing linear relationships between some variables

In this section, will look out to establish a linear relationship between age and number of trials for colour shapes and grey shapes since from the previous section we could find out that there is no correlation among gender and trials so we discard this particular linear relationship. For this purpose will create two linear models and then will check the normality of the residuals using Kolmogorov – Smirnov tests to assure that the obtained variances follow a normal distribution which inline assures that linear equation follows the principle of least square.

In this case, our null hypothesis and alternative hypothesis for both the linear models with significance level $\alpha = 0.05$ was:

	Mathematically
H_0: The value of the slope in a linear regression line is equal to zero. H_a : The value of the slope is significantly different from zero.	If $Y = bX + a$ is the line of linear regression where slope = b and intercept = a then the null and alternative hypothesis will be, $H_0: b = 0.$ $H_a: b \neq 0.$

The following results were obtained after performing the tests which are encapsulated in Table – 4.

	Linear Models	
	Model – 1	**Model – 2**
X	Age	Age
Y	ColourShapeCorrect	GreyShapeCorrect
b (estimated slope)	-1.52138	-0.97367
a (estimated intercept)	148.03752	139.15170
Equation of Linear regression line	$Y = -1.52138 * X + 148.03752$	$Y = -0.97367 * X + 139.15170$
p – value	2.2×10^{-16}	2.2×10^{-16}
	Residuals	
p – value	0.8107	0.9885

Table – 4: Encapsulated results of tests for two linear models

Thus, it can be observed from the Table – 4 that the p – value for both the models is very less than 0.05 therefore we can reject the null hypothesis and conclude that the slope is different from zero. This confirms that the relationship is linear. Moreover, looking to the p – value of the residuals obtained from the Kolmogorov – Smirnov tests are much higher than 0.05 which in-turn tells us that the residuals are approximately normally distributed. This proves that the equation of linear relationship obtained in Table – 4 is a valid relation. Figure 5 shows the graphs of the line best fit to the data for both the models.

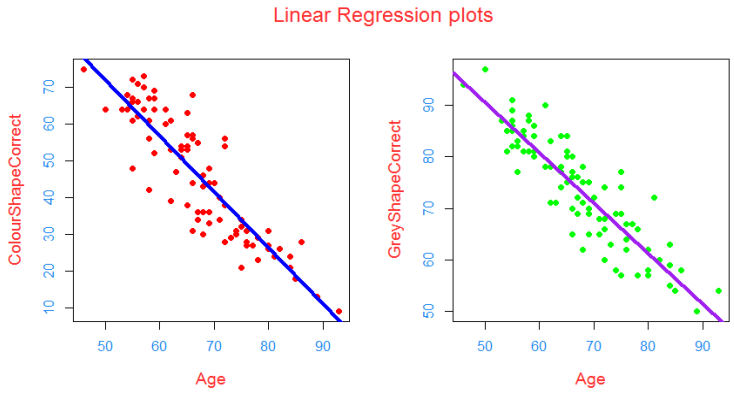


Figure – 5: Graphs of Two Linear Regression Models

Moreover, since from Figure – 5 and Table – 4 it can be seen that value of the slope is negative which concludes that age is negatively correlated with colour and grey shapes trials and nature of the straight line states that there is a linear relationship between age and cognitive tasks.

Section – 4: Performing some more statistical tests to support our hypothesis

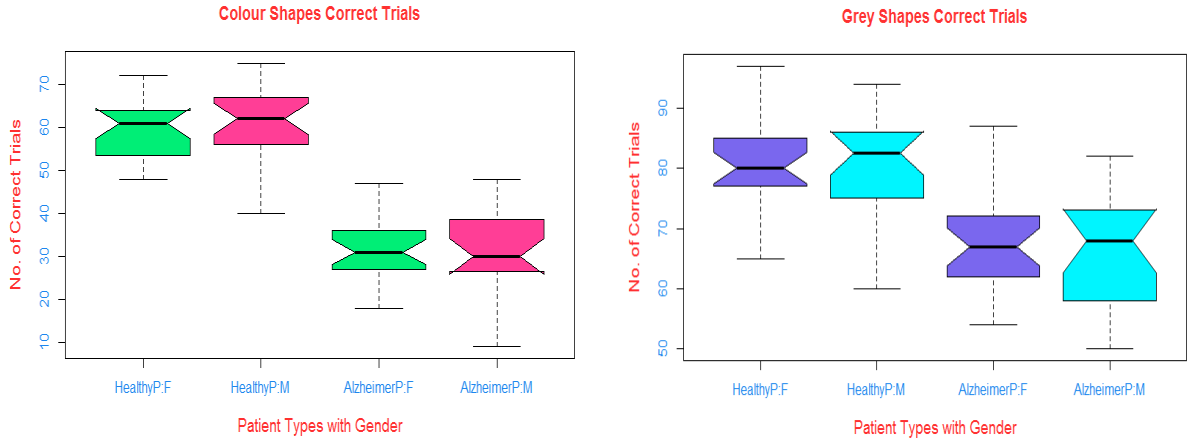
Before going for further analysis in this section the following are the highlights of conclusions from the previous sections:

1. Analyse the two groups – Alzheimer's disease and Healthy control patients groups separately and found that there are differences in ages between the two groups.
2. Age is a factor which is negatively correlated with correct number of trials – identifying colour shapes and grey shapes in cognitive tasks while sex is not correlated with any of the trials.
3. There is a strong linear relationship between age and number of trials.

The above analysis points out that now we check whether there are any differences between the correct number of trials for colour and grey shapes among the two groups separately. In this case, will use the Wilcoxon rank sum test (also referred as Mann – Whitney U–test) since that data is unpaired, the null and alternative hypothesis was:

H_0: There is no difference in the number of correct trials between the Healthy control patients and Alzheimer's disease patients.
H_a: There is a statistical difference in the number of trials between the two groups.
Mathematically it can be said that, H_0: Medians are same between the groups and H_a: Statistically there is a variation in the medians of both the groups.
The level of significance was $\alpha = 0.05$.

After performing the tests, the p – value for both the cognitive tasks – colour shapes and grey shapes are 7.525×10^{-16} and 2.357×10^{-16} respectively which is much less than 0.05 and thus we reject the null hypothesis and conclude that there is a statistical significant difference in number of trials among both the groups of patients. Figure 6 shows the boxplots for the Colour shapes correct trials and Grey shapes correct trials respectively where it can be observed that medians are different for both groups.



(a) Colour shapes correct trials (b) Grey shapes correct trials

Figure – 6: Boxplots of correct shapes trials

From this it can be concluded that Healthy patients can identify more shapes and performed well in both the cognitive tasks compared to Alzheimer's disease patients irrespective of gender.

Conclusion

Overall, after analyzing each aspect among the different sections, it can be concluded that Alzheimer's disease patients performed badly at the cognitive engagements and no matter what type of sex it is as the age increases the ability to recognize colour shapes and grey shapes decreases.

References

[1] Alboukadel Kassambara (2018). *Machine learning essentials : practical guide in R*. Scotts Valley, Kalifornien] Createspace Independent Publishing Platform.

[2] Corder, G.W. and Foreman, D.I. (2009). *Nonparametric statistics for non-statisticians : a step-by-step approach*. Hoboken, N.J.: Wiley, pp.136–139.

[3] Revelle, W. (2020). *How To: Install R and the psych package*.

[4] Rizopoulos, D. (2022). *Package "ltm" Title Latent Trait Models under IRT*. [online] Available at: https://github.com/drizopoulos/ltm [Accessed 18 Mar. 2022].

[5] Robert Russell Johnson (1996). *Elementary statistics*. Seventh ed. Belmont Etc.: Duxbury Press, pp.378–384.

[6] Wei, Y. (2021). *Colors in R*. [online] Available at: https://tz33cu.blogspot.com/2006/09/have-you-seen-so-manycolors-in-r.html [Accessed 18 Mar. 2022].