ISSN No:-2456-2165

Exploratory Data Analysis of Migraine Data

P Chandra
Department of Research - Ph.D. Computer science
Tirupur Kumaran College for Women
Tirupur, India

Dr. V Arulmozhi
Associate Professor,
Department of Research - Ph.D. Computer science
Tirupur Kumaran College for Women
Tirupur, India

Abstract:- This paper aims to provide a methodology for the quantitative analysis of migraine data. The main objective is to facilitate the health practitioners who are fascinated by data study and have it in mind to offer a concise steer that may prove useful across a wide range of medical applications. To illustrate the proposed study, a typical migraine dataset is used to demonstrate how these steps are useful in practice. However, nowadays migraine becoming a common problem in almost all kinds of people. Due to stress full working environment, the impact of parent's heredity in children, lifestyle change, irregular food habits, weather conditions, excess consumption of caffeine, medication overuse, menstrual time headache, and menopause stress in women and tension are the reasons for a migraine attack. The data set Kostecki Dillon downloaded from the UCI repository with 4152 observations on 133 subjects for 9 variables is considered for the learning and from this, separation of records on migraine handling collected by Tammy Kostecki-Dillon consists of headache entries set aside in a treatment program. The study will discuss some standard ideas of correlations and p – values to quantify "importance" (or more mathematically accurately statistical significance). Also, it discusses some standard statistical analysis and hypothesis testing which offers an improved understanding.

Keywords:- *Migraine; Statistical analysis; Hypothesis testing; Correlation)*

I. INTRODUCTION

Migraine is one of the most common happening complaints disturbing the nervous organism of humans. There are various forms of migraines, influence persons depend on the surroundings, age, gender, and other aspects. This paper endeavors to execute some statistical tests on Migraine data and to present some implications concerning the patient's age, gender, and headache varieties. In turn, these realities will facilitate the patients and other individuals to be conscious of the happening of Migraine. It would be reasonable to say that statistical techniques are essential to effectively work in the course of machine learning schemes. Statistical methods can be used to clean and organize the data geared up for modeling. Statistical hypothesis tests can support in model assortment and in presenting the expertise and predictions from concluding models. This study is concerned with presenting the associations between the variables, performing some standard statistical and hypothesis tests on the variables, and gives the conclusions about the data. The paper continues with a literature review, data set description, a methodology of performing the statistical analysis, conclusion, and Further enhancements and ends with references.

II. LITERATURE REVIEW

Among all the medical reasons, the most dominating feature is the number of times of the occurrence of the headache. Mostly women with above 40 years of age are affected with migraine extensively. The throbbing and pain disturb daily routine functioning. The most useful factors for governing headaches in women are age, education, and frequency of headaches. On account of the headache rate of recurrence shows potential as the most influencing factor, it is of the utmost importance to enlighten patients of the value of taking prophylactic measures. Also, it is important to find the cause of migraine occurrence. This approach will help the person in conducting themselves either with the help of medicines or other approaches [1].

The main reason for the occurrence of Migraine is the emotional strain. Migraine fatalities are found while the emotional and stress full events take place. If a person gets emotional, certain chemicals in the brain are unconfined and fight the state of affairs. The discharge of these chemicals causes migraines. Stress is an important aspect of the occurrence of a Tension headache. Tension headaches can both be sporadic or continual. Episodic tension headache is triggered by a demanding circumstance or a build-up of stress. It can be treated by over-the-counter painkillers. Daily strain, such as high-anxiety jobs, will direct to a chronic tension headache. Treatment for chronic tension headaches usually involves stress managing, therapy, biofeedback, and probably the use of antidepressant or anxiety-sinking medicines [2].

Environmental reasons such as transformation in the atmosphere or weather, a change in altitude or barometric pressure, high winds, traveling, or a change in habit are factors which generate migraine. Other ecological triggers take account of a bright or gleaming brightness (sunlight reflections, glower, luminous lighting, television, or movies), boundaries of heat and resonance, and vigorous smells or fumes. Any change in the environment of the headache patient will aggravate the headache suffering. Change in job and school and a transaction if adaptation will affect the migraine patients. Travel, change in diet, change in the ecological and atmospheric circumstances may raise the headache. Some physical factors also can trigger migraine headaches; including overexertion such as bending, straining, or lifting; toothache; or contained head or neck pains [3].

Seng, et al. of Albert Einstein College of Medicine, presented some conclusions from a study on the impact of a parent's migraine problem on their kids. In a month the parents generally had 6.8 headaches as an average. Both parents and kids reported that the occurrence of headache gives a sensible force on their daily routine. It affects parent-child relations, followed by the hinder of providing everyday help and a significant crash. The major population of the kids wants more assistance in helping their parents with migraines [4].

Studies conducted by Lipton et al, women with migraines were separated into groups of Pure, Menstrual Related, and Non-Menstrual Related Migraine. Researchers signifying the results that women with Non-menstrual related were more possible to experience high pain concentration, more overall pain disturbance, and practical demolition within an hour of headache commencing when compared to women with other groups [5].

Yu-Chen Cheng, et al. of Massachusetts General Hospital in Boston offered conclusions that 60% of the female with migraine history at the menopausal time developed migraine; the change in the occurrence of the migraine is at the time of pre-menopausal or post-menopausal. Based on MRI scans, the researchers concluded that pituitary abnormalities were more persistent in patients with new-onset migraine [6] [7].

III. DATASET DESCRIPTION

The data set is downloaded from the UCI repository with 4152 observations on 133 subjects for 9 variables is taken for the study and from this a detachment of data on migraine treatments composed by Tammy Kostecki-Dillon consist of headache entries kept in a treatment program. Patients entered the program at an altered period over a stage of about 3 years. Table 1 will show the description of the dataset.

Table 1 Data set description

Variable Name	Description
id	Patient id.
time	Time in days comparative to the start of treatment, which comes at time 0.
dos	A period in days from the beginning of the study.
hatype	A reason with levels Aura Mixed No Aura, the type of migraine practiced by a focus.
age	At the onset of treatment, in years.
airq	A measure of air quality.
medication	An issue with levels none concentrated systematic, demonstrating subjects who discontinued their medication, who sustained but at a reduced measure, or who persistent at the previous dose.

IV. METHODOLOGY

The type of the variable in a dataset may be continuous or categorical. The data for a continuous variable will be continuous. The data entries for a categorical variable will be of categories like yes or no and specific classes. These data entries are cleaned and if necessary it will be converted to a required data type. If the categories are to be converted to numeric, it can be done and prepared for testing. This study will present all the data preparation and testing methods by using R Language.

Now, the variable "hatype" can be viewed by using Table 2 for finding whether it is a continuous or categorical variable and the number of occurrences of it.

Table 2 Type and occurrence of variable hatype

Aura	Mixed	No Aura
1710	457	1985

From Table 2 it is observed that there are three categories of data for the variable "hatype" with the types "Aura", "Mixed" and "No Aura". The observation respectively for Aura is 1710 entries, for a mixed type of head ache is 457 and for No Aura is 1985.

Fig 1 will show the bar chart for the "hatype".

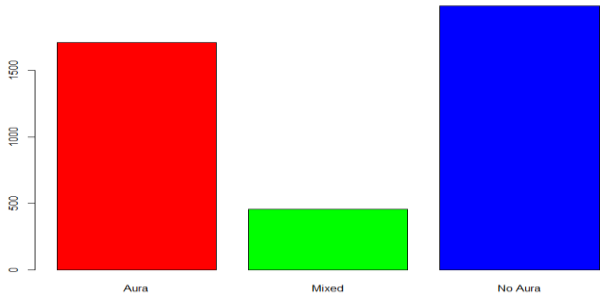

Fig 1 Bar plot for hatype variable.

The continuous variable age will be displayed by using a bar plot in Fig 2.

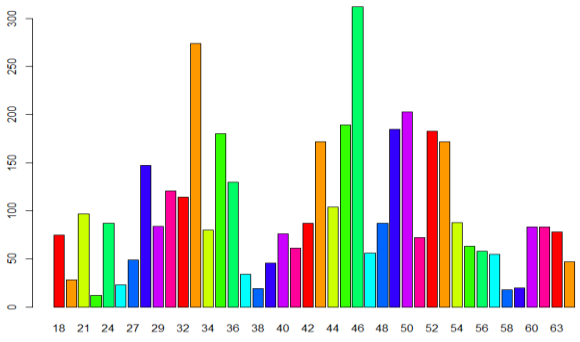

Fig 2 Bar plot for age variable.

From Fig 2 it is observed that the continuous value for age varies from 18 to approximately 68. So, the assumption can be made that, the marine attack will take place approximately from age 18 up to 70.

Fig 3 & 4 displays the age variable with a Histogram and bar plot.

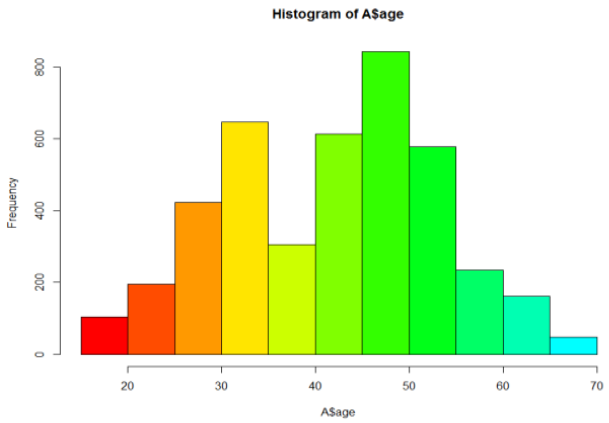

Fig 3 Histogram for the age variable

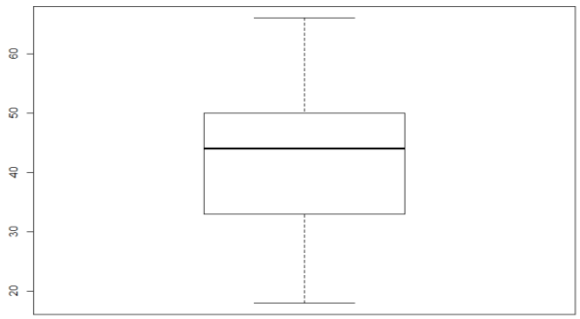

Fig 4 Box plot of variable age

A. Correlation of variables

In Fig 5 the correlation of variables of the entire data set is displayed.

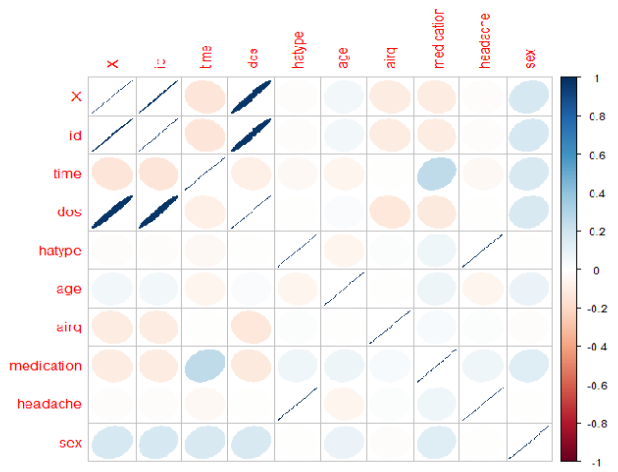

Fig 5 Correlation plot for migraine dataset

B. P-Value

P-Value is the indicator of the test results that explains a significant difference between variables. Many statistical tests provide both numeric results and a p-value. The P-value is a value that ranges between 0 and 1. The P-value is the probability of obtaining the observed results of a hypothesis test, assuming that the null hypothesis (here in this scenario the comparing two groups are the same) is correct. The low P-value means that the smallest level of significance shows the assuming groups are not the same thereby the null hypothesis H_0 would be rejected. Therefore it can be called the difference between the two groups statistically significant. The high P-value (Alternative hypothesis) means that the two groups were not that much different. A P-value 1 indicates that there is no disparity at all between the two groups. Generally, if the p-value is less than 0.05, the difference observed is considered as statistically significant.

V. STATISTICAL TESTS

A. T-Test

The T-Test used to calculate the mean of two groups of samples is called 2 sample T-test. The test is to evaluate the means of the two sets of data are statistically significantly vary from each other. Here in this study, the unpaired two-sample t-test is used to compare the means of two independent samples “age” and “hatype”.

The method for T-test if the means are different is [8]

t-value = (mean 1 – mean 2) / sp
where,

sp = √ (var1 / n1) + (var2 / n2)

By using the RStudioo, the t value is calculated, and the evaluated variables are “age” and “hatype” by using the following algorithm.

The algorithm for T-Test is as follows:

```
rm (list = ls())

a<- read.csv("KosteckiDillon.csv")
# Check the mean age vary Significantly with headache type or not
boxplot(a$age~a$hatype)
M <- tapply(a$age,a$hatype,mean)
v <- tapply(a$age,a$hatype,var)
n <- table(a$hatype)
tdata = (M[1] - M[2])/ sqrt((v[1]/n[1])+(v[2]/n[2]))
# probability of error
pvalue <- 2*pt(tdata,df = min(n[1] - 1,n[2] - 1), lower.tail = FALSE)
```

To check the mean age varies significantly with headache type or not, the P-value is calculated by using the t – value. To reject the null hypothesis P - value should be < 0.025

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
a$hatype	2	13181	6590	53.86	<2e-16 ***
Residuals	4149	507637	122		

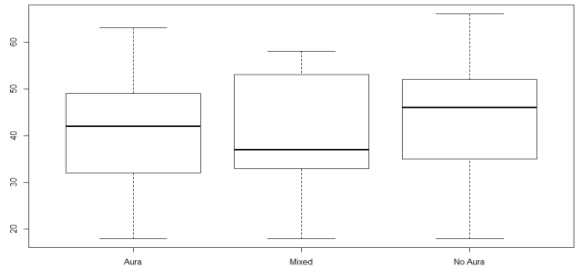

Fig 6 Box plot for comparison of means (age and hatype)

Figure 6 illustrates the means of the variables "age" and "hatype".

The Results for t score and p-value while running the algorithm in RStudio is as follows:

P-value
Aura
0.1650248

Tdata
Aura
1.390601

The inference from the above said T-test value will be The higher P-value indicates the t-test fails to reject the null hypothesis (h0). There is no significant dissimilarity among "age" concerning "hatype".

B. Z – Test

A Z-test is a type of hypothesis test that will tell the results from the test are valid or repeatable. A hypothesis test will tell if it is probably true, or not. A Z – test is used when the data is roughly normally distributed.

A two proportion z-test compares two extents.

- The null hypothesis (H_0) is to test the proportions are the same.
- The alternate hypothesis (H_1) is to test the proportions are not the same [9].

1) Pooled sample proportion:

The null hypothesis states that $P_1=P_2$, the pooled sample proportion (p) is to calculate the standard error of the sample distribution.

$p = (p_1 * n_1 + p_2 * n_2) / (n_1 + n_2)$

Where p_1 and p2 are the sample proportions from populations 1 & 2, n_1 and n_2 are the size of samples 1 & 2.

2) Standard error:

This gives the standard error (SE) of the model distribution dissimilarity between two proportions.

$SE = sqrt \{p * (1 - p) * [(1/n_1) + (1/n_2)] \}$

Where p is the proportion of pooled sample, n_1 and n_2 are size of samples 1 & 2.

3) Test statistic:

The formula for z-score (z) is defined by

$z = (p_1 - p_2) / SE$

Where p_1 is the proportion from sample 1, p_2 is the proportion from sample 2, and SE is the standard error of the sampling distribution [10].

The algorithm for Z – Test is as follows:

```
rm (list = ls())
#a<- read.csv("KosteckiDillon.csv",header = TRUE,sep = "\t",row.names = 1)
# check the proportion of headachetype with respect to sex vary significantly or not
a<- read.csv("KosteckiDillon.csv")
n <- table(a$sex)
f <- table(a$hatype,a$sex)
p1 <- f[2,1]/(f[1,1] + f[2,1])
p2<- f[2,2] / (f[1,2] + f[2,2])
pp <- (f[2,1] + f[2,2]) / (f[1,1] + f[2,1] + f[1,2] + f[2,2])
zdata <- (p1 - p2)/ sqrt(pp * (1 - pp) * (1/n[1] + 1/n[2]))
pvalue <- 2 * pnorm(abs(zdata),lower.tail = FALSE)
```

To check the proportion of headache type concerning sex varies significantly or not the formula for z score has been used and the results as follows:

p-value
female
1.769618e-128

From the result, due to the high p-value, the Z test fails to reject the null hypothesis. Hence there is no significant headache type in proportion concerning gender.

C. Analysis of Variance (ANOVA - F test):

Data do not always fit into discrete categories, but numeric data can also be of interest in a field of investigation. The comparison of one continuous and one categorical variable will be compared by using this test.

To calculate the Mean Square Error (MSE) and Mean Square Treatment, first calculate the Error Sum of Squares (SSE), Treatment Sum of Squares (SSTR), and total Sum of Squares (SST) [11].

The formula for finding the Error Mean Square is

$MSE = SSE / N – t$

N – Total number of observations
t – Total Number of treatments.

The formula for Treatment Mean Square is:

$MSTR = SSTR / t – 1$

The formula for Test Statistic (F – Statistic) is:

F = MSTR / MSE

The algorithm for ANOVA Test is as follows:

```
rm (list = ls())
a<- read.csv("KosteckiDillon.csv")
#boxplot(a$age~a$hatype,col=rainbow(3))
boxplot(a$age~a$hatype,col=rainbow(3))
M <- tapply(a$age,a$hatype,mean)
res <- aov(a$age~a$hatype)
summary(res)
```

The builtin function aov(a$age~a$hatype) has been used to calculate the Analysis of variance.

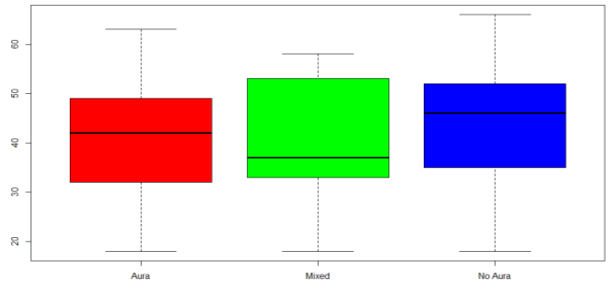

Fig 7 Box plot comparing "age" and "hatype"

The Result for the above-mentioned statement will be
Table 3 Result of ANOVA test

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

From the results shown in Table 3, the p-value is high and the ANOVA test fails to reject the Null hypothesis (H0). Hence there is no significant difference between the samples "age" and "hatype".

D. Chi-Square Test

There are two types of chi-square tests

1. A chi-square test for goodness of fit - determines if the sample data matches a population in other words goodness of fit is used to test if sample data fits a distribution from a certain population.
2. A chi-square test for independence – tests to see whether distributions of categorical variables differ from each other.

A very small chi-square test statistic means that the observed data fit the expected data extremely well. There is a relationship that exists between the two variables. A very larger chi-square value means that the data does not fit very well and there is no relationship exists between the variables tested.

A Chi-Square statistic is a dimension of how opportunities contrast to results. The data used in calculating a chi-square must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large sample [12].

The formula for calculating Chi-Square test is:

$$\chi^2_c = \Sigma ((o_i - E_i)^2 / E_i)$$

The subscript "c" is degrees of freedom. "o" and "E" observed and Expected values.

The algorithm to calculate Chi-Square value implemented in R is as follows:

```
rm (list = ls())
a<- read.csv("KosteckiDillon.csv")
#ChiSquare Test
res <- chisq.test(a$hatype,a$sex)
```

To calculate the Chi-Square value the function chisq.test(a$hatype,a$sex) has been called and the result will be as follows:

Pearson's Chi-squared test

data: a$hatype and a$sex

X-squared = 259.95, df = 2, p-value < 2.2e-16

From the result above it is clear that due to high P-Value the Chi-square test fails to reject the Null hypothesis (H_0). Hence the proportion of "hatype" is not the same with gender and there is no relationship between headache type and gender.

VI. CONCLUSION AND FURTHER ENHANCEMENTS

A headache becomes a common disease due to some viral infections, present working environment, and other lifestyle reasons, it is worth to explore about it by using the existing data and to know the actual relations of the age, gender, and the type of headache. The main objective is to help the clinical practitioners who are interested in data analysis and intends to offer a succinct guide that may prove useful across a wide range of medical applications. It is possible to find the inferences related to migraines by using the statistical tests. Further, the predictions and the classifications can be made by using the statistical models available.

REFERENCES

[1]. Dorota Talarska,1 MaBgorzata Zgorzalewicz-Stachowiak,2 MichaB Michalak,3 Agrypina Czajkowska,1 and Karolina HudaV 2," Functioning of Women with Migraine Headaches", Hindawi Publishing Corporation, Scientific World Journal Volume 2014, Article ID 492350, 8 pages http://dx.doi.org/10.1155/2014/492350.
[2]. https://my.clevelandclinic.org/health/articles/9646-stress-and-headaches
[3]. https://headaches.org/2007/10/25/environmental-physical-factors/.

[4]. Seng EK, et al. "When mom has migraine: An observational study of the burden of parental migraine on children". AHS 2018; Abstract OR-07.
[5]. Lipton RB, et al." Assessing Unmet Treatment Needs and Associated Disability in Persons with Migraine: Results from Migraine in America Symptoms and Treatment (MAST) Study". AHS 2018; Abstract OR-02.
[6]. Cheng YC, et al. "Migraine Pattern Changes in Women During the Menopause Transition". AHS 2018; Abstract OR-16.
[7]. https://www.practicalpainmanagement.com/meeting-summary/impact-migraine-women-health
[8]. http://www.sthda.com/english/wiki/t-test-formula
[9]. https://www.statisticshowto.datasciencecentral.com/z-test/
[10]. https://stattrek.com/hypothesis-test/difference-in-proportions.aspx
[11]. https://www.dummies.com/education/math/business-statistics/how-to- find-the-test-statistic-for-anova-using-the-error-mean-square-and-the-treatment-mean-square/
[12]. https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/