Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6010 - Probability Theory and Introductory Statistics**

**Assignment:**

MODULE 4 PRACTICE ASSIGNMENT 4

**Submitted on:**

December 4, 2022

**Submitted to:**  **Submitted by:**

Professor: BEHZAD AHMADI NIKSHITA RANGANATHAN

# **INTRODUCTION**

Data analytics is the process of understanding raw information with the objective of drawing inferences about that data. It involves using various techniques, such as statistical analysis, machine learning, and data mining, to uncover patterns and trends in the data. Analytics is utilized in a range of fields, from the sciences to business, to help organizations and companies make better business decisions. By analyzing data, organizations can discover trends, spot opportunities, and make better predictions. Data analytics also helps organizations identify potential risks and optimize operations. Through data analytics, organizations can gain insights into areas such as customer experience, marketing, product development, and financial performance.

To test a hypothesis regarding a population parameter, a statistical approach known as hypothesis testing is performed. It is used to make decisions about a population or process on the basis of information from a sample. They are used to test relationships between variables, assess the impact of a change or intervention, and compare the performance of different groups.

Hypothesis testing is used to answer questions in many fields, including medical research, economics, engineering, and social sciences. Hypothesis testing plays a significant role in ensuring that outcomes are reliable.

**About the dataset:**

We extracted the data set from the “MASS” package consisting of data about cats. There are 144 cats listed here, along with information on their sex (M/F), body weight (kg), and heart weight (g).

The Data set of cats with the dictionary is as follows: -

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Feature** | **Dictionary** |
| 1. | Sex | The cats' gender (Male or Female) |
| 2. | Bwt | The cats' body weight (kilograms) |
| 3. | Hwt | Heart weights of the cats (grams) |

*Table 1: Dictionary*

The goal of this analysis is to use 2 types of hypothesis testing in this assignment – Two sample t-test and Paired t-test

**Performing Hypothesis testing:**

1. When doing a hypothesis test, you must first identify both the null and alternative hypotheses.

2. Determine the significance level for the test, usually 0.05 or 0.01.

3. Calculate the test statistic to determine the likelihood of getting the observed results, given the null hypothesis is true.

4. Determine whether the test statistic supports or contradicts the hypothesis.

5. Interpret the results in terms of the problem and explain the implications of the decision.

**EXPLORATORY DATA ANALYSIS**

* Data Cleaning
  + Checking for <NA> values

The entire data frame is visualized using the vis\_dat() function, it also displays the datatypes involved in the dataset and information about missing data.

The cats dataset does not contain NA values.

Chart, waterfall chart

Description automatically generatedChart, waterfall chart

Description automatically generated****

Logo

Description automatically generated with low confidence

**Figure 1-<NA> values**

* Understanding the cats dataset
* The variables in the dataset include the Sex of the cats, their body, and heart weights. We can see datatypes factor and number here.
* There are 47 observations in the Female category and 97 observations under the Male category.

**A picture containing text

Description automatically generated**

**Text

Description automatically generated**

**Figure 2-str() and summary()**

* There are 144 observations and 3 features in this dataset.

A picture containing icon

Description automatically generated

**Figure 3-dim()**

* We can see the first and last 4 observations using headTail() function.

**A picture containing table

Description automatically generated**

**\**

**Figure 4-headTail()**

* An alternative to summary() that quickly gives a general overview of a data frame is skim().

A screenshot of a computer

Description automatically generated with medium confidence

**Figure 5-skim()**

* To obtain the descriptive statistics of the data set's features, we utilized the describe() function from the package "PSYCH."
* The mean of Bwt (Body weight) is approximately 2.72, with a standard deviation of 0.49. The minimum and maximum values are 2.0kgs and 3.9kgs respectively.
* The mean of Hwt (Heart weight) is approximately 10.63, with a standard deviation of 0.49. The minimum and maximum values are 6.3g and 20.5g respectively.
* The skew of both Body and heart weight columns is between -0.5 to 0.5 which represents symmetric distribution.
* The features Bwt and Hwt have kurtosis values below 3, which indicates a distribution that is flat and has fewer peaks (Platykurtic).

Text

Description automatically generated with medium confidence

**Figure 6-describe()**

* describeBy() returns the statistical summary of grouped data. In this case, it is by Sex (Female and Male)

**A screenshot of a computer

Description automatically generated with medium confidence**

**Figure 7-describeBy()**

**PART-1**

In order to compare the body weights of male and female cats, I made different subgroups, checked the normality of features, and then performed the two-sample t-test.

* Subgroups of male and female cats can be seen below. I also grouped the dataset by separating Female and Male body and heart weights.

Text

Description automatically generated

**Figure 8-subgroups using subset()**

* Normality tests
* Shapiro-Wilk test

The Shapiro-Wilk test, a statistical method, is used to determine if a sample of data has a normal distribution. If the p-value is below 0.05, the normality of the features is rejected. The degree of normality of a data set is determined by how near the W value is to 1.

Male cats body weight has a p-value of 0.119 and a W value (0.97883) denoting a normal distribution. Whereas female cats body weight has a p-value of 0.0003754 and a W value (0.89096) which denotes a non-normal distribution.

**Graphical user interface, text, application

Description automatically generated**

**Figure 9-Shapiro-Wilk tests for both groups**

* Q-Q Plots

A Q-Q plot is a graphical representation used to assess how closely a given set of data follows a normal distribution.

**Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated**

**Figure 10-QQ plots for cats (female and male)**

* Density plots

The density plot will resemble a bell curve if the data is distributed normally. This bell curve will have a peak in the middle and will taper off at the edges.

Both the plots seem normal.

Chart

Description automatically generated

**Figure 11 - Density plot between body weights of Female and male cats**

* Other Visualizations
* Histogram and Boxplot

**Chart, bar chart, histogram

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Figure 12 - Histogram and Boxplots between body weights of Female and male cats**

* Two-Sample T-test

When two independent samples, such as two distinct groups of people or two distinct treatments, are compared, this kind of test is used to assess whether there is a statistically significant difference between the two.

Welch’s t-test is a type of two-sample test used to compare the means of two independent groups. Welch’s t-test does not assume equal variances in the two samples.

**Null Hypothesis H0:** The difference in means between Female and Male cats is equal to 0

**Alternate Hypothesis H1:** The difference in means between Female and Male cats is not equal to 0

**Text

Description automatically generated**

**Figure 13 – Two sample t-test**

**Degree of Freedom: 136.84**

**Confidence Interval: 95 %**

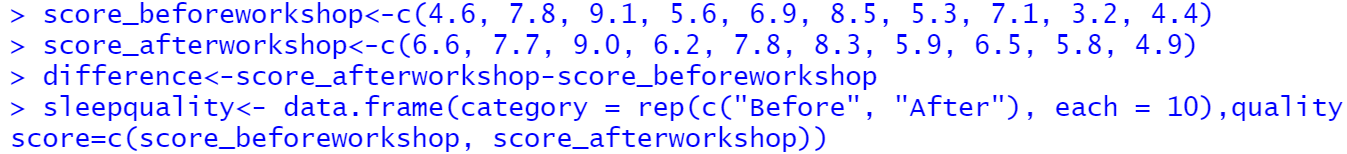
**P value: 8.831034e-15**

**Question - Do male and female cat samples have the same body weight (“Bwt”)?**

We reject the null hypothesis of the test because the p-value is below the level of significance of 0.05. There is sufficient evidence to say that the male and female samples do not have the same body weight.

**PART-2**

Sleep scores of 10 students were measured before and after the meditation workshop and in this project, we have to find out if there is any effect of the workshop on students. Similar to Part 1 of the assignment, first I checked the normality of the variables. To store the scores before and after the session, I used two different vectors. The scores were subtracted to determine the difference in the sleep quality scores for both workshops. Additionally, another data frame sleepquality with quality scores before and after the medication workshop was created.



**Figure 14 – Quality scores**

* Understanding sleepquality dataset

There are 20 rows and 2 features (Category and quality score). category column has a datatype character and qualityscore is numerical.

Text

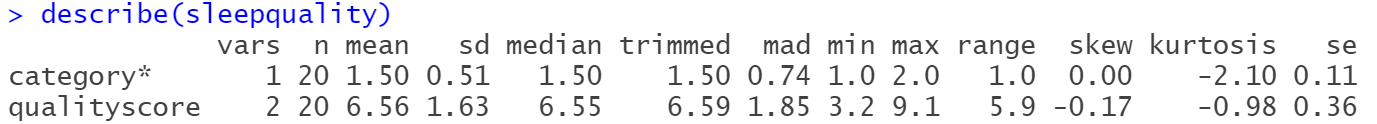
Description automatically generated with medium confidenceTable

Description automatically generated

**Figure 15 – headTail() and summary()**

Descriptive statistics of sleepquality dataset

The quality score has a mean of 6.56 and a standard deviation of 1.63. Skew of -0.17 indicates the normal distribution and kurtosis denotes flat distribution with less peaks.

****

**Calendar

Description automatically generated with low confidence**

**Figure 16 – describe() and describeBy()**

* Normality tests
* Shapiro-Wilk test

The difference vector has a p-value of 0,2177 and a W value (0.89975) which represents a non-normal distribution. On the contrary, when the Shapiro test was conducted by for the groups(before and after), we can consider that they have a normal distribution,

**Text

Description automatically generated with medium confidence**

**Figure 17 – Shapiro tests**

* Q-Q Plots

The observation is that the points are generally linear and follow a straight line suggesting that the data is regularly distributed.

**Chart

Description automatically generated**Chart, line chart

Description automatically generated

**Figure 18 – QQ plots**

* Density Plots

**Chart, histogram

Description automatically generatedChart

Description automatically generated**

**Figure 19 – Density plots A and B**

* Other Visualizations

**Chart, line chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**Figure 20 – Basic boxplot and Paired boxplot**

* Paired t-test

**Question -  Justify the testing procedure you choose, why use the test of your choice rather than other tests?**

The paired t-test is used to check whether the two related samples vary statistically significantly from one another. In this case, before and after measurements are taken from the same group of subjects that is medication workshop. Because of this reason, I chose this test.

**Null Hypothesis H0:** The difference in average sleeping scores before and after the medication workshop is equal to 0

**Alternate Hypothesis H1:** The difference in average sleeping scores before and after the medication workshop is not equal to 0

**Text

Description automatically generated**

**Figure 21 – Paired T-test**

**Degree of Freedom: 9**

**Confidence Interval: 95 %**

**P value: 0.08322052**

We fail to reject the test's null hypothesis since the p-value is more than the significance level(α) of 0.05. There is insufficient evidence to disprove that the sleep quality scores before and after the workshop are equal and medication workshop has an impact on the scores.

**Question - Does your conclusion change if the level of significance changes from 0.05 to 0.1?**

If we take an alpha-value of 0.1 (90%), the p-value would be less than this alpha-value (0.1) and the Null hypothesis would get rejected in this situation. So, if the significance level is 0.1, we will have sufficient evidence to say that workshop has an impact on the sleep quality scores of the students.

**CONCLUSION**

Based on the gender of the cats, the dataset of "cats" from the library "MASS" has revealed numerous insights about the body weights and heart weights of the cats. For the second part of the assignment,10 people's sleeping patterns before and after the meditation workshop were also examined. In this project, I completed exploratory data analysis, made statistical calculations, and plotted data visualization graphs to comprehend the data. I also performed normality tests and various hypothesis tests for both datasets.

The key findings are :

* According to Shapiro tests, the body weights of male cats are normally distributed.
* To evaluate the characteristics' visual normality, density plots can be used. It can be seen that bodyweight feature follow a normal distribution.
* It was possible to learn more about the normality of the attribute using QQ plots, boxplots, and histograms too.
* A two-sample t-test was carried out to ascertain the relationship between two separate samples. It can be concluded that the means of the two categories (male cats and female cats) are not equal.
* The average of the quality scores before and after the workshop was compared using a paired t-test. If we take a significance level of 0.05 we are unable to reject the null hypothesis of the test. We could, however, reject the test's null hypothesis with an alpha value of 0.1.

**REFERENCES**

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Paired Samples T-test in R - Easy Guides - Wiki - STHDA. (n.d.-b). <http://www.sthda.com/english/wiki/paired-samples-t-test-in-r>

Herv, M. (2022, March 18). Shapiro-Wilk test for factor levels. <https://rdrr.io/cran/RVAideMemoire/man/byf.shapiro.html>

Plot Grouped Data: Box plot, Bar Plot and More. (2017, November 17). Articles - STHDA.  http://www.sthda.com/english/articles/32-r-graphics-essentials/132-plot-grouped-data-box-plot-bar-plot-and-more/

**APPENDIX: CODE**

#---------------------- Week\_4\_Module\_4 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Week 4 Assignment - Module 4 R Practice")

print("Course Name - ALY6010: Probability Theory and Introductory Statistics")

# Loading the packages

library(ggthemes)

library(plotly)

library(visdat)

library(ggpubr)

library(dlookr)

library(psych)

library(skimr)

library(MASS)

# Checking NA values

vis\_dat(cats)

vis\_miss(cats)

sum(is.na(cats))

sum(is.null(cats))

#------------------- Exploratory Data Analysis -------------------#

# Analysis

data(cats)

headTail(cats)

str(cats)

summary(cats)

dim(cats)

skim(cats)

describe(cats)

describeBy(cats,group=cats$Sex)

# Creating Subgroups

Femalecats<-subset(cats,cats$Sex=="F")

Malecats<-subset(cats,cats$Sex=="M")

Femalecatsbwt<-subset(cats,cats$Sex=="F",select=c("Bwt"))

Malecatsbwt<-subset(cats,cats$Sex=="M",select=c("Bwt"))

# Shapiro tests

shapiro.test(Malecats$Bwt)

shapiro.test(Femalecats$Bwt)

# Normal QQplot

ggqqplot(Femalecats$Bwt,main="Q-Q Plot (Body weight of Female Cats)")

ggqqplot(Malecats$Bwt,main="Q-Q Plot (Body weight of Male Cats)")

# Density Plots

library(gridExtra)

library(grid)

library(ggplot2)

library(wesanderson)

devtools::install\_github("ropenscilabs/ochRe")

require(ochRe)

library(ochRe)

library(naniar)

ggplot(cats, aes(x = `Bwt`, fill = `Sex`)) +geom\_density(alpha = 0.6)+xlim(0,5)+scale\_fill\_ochre(palette="williams\_pilbara")+theme(text = element\_text(size = 20))+xlab("Body Weight")+ggtitle("Density plots")

# Histograms

ggplot( cats,aes(x=Bwt, fill=Sex)) + geom\_histogram( alpha=0.7, position = 'identity',bins = 30) + scale\_fill\_manual(values= wes\_palette("GrandBudapest2", n = 2)) +ggtitle("Histograms for Body weight")+xlab("Body Weight")+labs(fill="")+theme(text = element\_text(size = 20))

# Boxplots

ggplot(cats, aes(x = Sex, y = Bwt,fill=Sex)) +geom\_boxplot(notch=TRUE,outlier.fill = "red",outlier.shape = 22,outlier.colour = "black")+ggtitle("Boxplot")+scale\_fill\_ochre(palette="mccrea")+theme\_minimal()+geom\_jitter(cex = 0.9)+theme(text = element\_text(size = 20))+ylab("Body Weight")

#------------------- Part 1 -------------------#

# Two sample t tests

ttest1<-t.test(Malecatsbwt,Femalecatsbwt,alternative = "two.sided")

ttest1

ttest1$p.value

#------------------- Part 2 -------------------#

score\_beforeworkshop<-c(4.6, 7.8, 9.1, 5.6, 6.9, 8.5, 5.3, 7.1, 3.2, 4.4)

score\_afterworkshop<-c(6.6, 7.7, 9.0, 6.2, 7.8, 8.3, 5.9, 6.5, 5.8, 4.9)

difference<-score\_afterworkshop-score\_beforeworkshop

sleepquality<- data.frame(category = rep(c("Before", "After"), each = 10),qualityscore=c(score\_beforeworkshop, score\_afterworkshop))

# Analysis

headTail(sleepquality)

summary(sleepquality)

describe(sleepquality)

describeBy(sleepquality,group=sleepquality$category)

# Shapiro tests

shapiro.test(difference)

install.packages("RVAideMemoire")

library(RVAideMemoire)

byf.shapiro(qualityscore~category,data=sleepquality)

# Normal QQplots

library(ggpubr)

ggqqplot(difference,main="Q-Q Plot Difference in Quality scores (After and Before Workshop)")

ggqqplot(sleepquality,main="Q-Q Plot",x = "qualityscore",facet.by = "category",color="category",fill="category")+scale\_color\_ochre(palette="healthy\_reef")+scale\_fill\_ochre(palette="healthy\_reef")

# Density Plots

cols <- c("#79AF97FF","#FFDC91FF")

ggplot(sleepquality, aes(x = `qualityscore`,fill=`category`)) +theme(text = element\_text(size = 20))+geom\_density(alpha=0.6)+xlab("Quality Score")+ scale\_fill\_manual(values=cols)+xlab("Quality scores")+ggtitle("Density plot A")

difference=as.data.frame(difference)

ggplot(difference,aes(x=`difference`))+geom\_density(fill = "#925E9F99")+theme(text = element\_text(size = 20))+xlab("Difference between Quality scores")+ggtitle("Density plot B")

# Boxplot

ggplot(sleepquality, aes(x=category,y=qualityscore, fill=category)) + geom\_boxplot()+scale\_fill\_brewer(palette = "Spectral")+theme(text = element\_text(size = 20))+ylab("Quality scores")+ggtitle("Boxplot")

# Paired Boxplot

ggpaired(sleepquality, x = "category", y = "qualityscore",

color = "category", line.color = "seashell4", line.size = 0.5,

palette = "aaas")+ stat\_compare\_means(paired = TRUE)+theme(text = element\_text(size = 20))

# Paired T test

ttest2<-t.test(data=sleepquality,qualityscore~category,paired=TRUE)

ttest2

ttest2$p.value