# IE6200 FINAL PROJECT REPORT

## COMPARISON OF CROSS-SECTIONAL MRI DATA IN YOUNG, MIDDLE AGED AND OLDER ADULTS

Vijayakumar Venkataraman
Venkataraman.v@husky.neu.edu

# TABLE OF CONTENTS

## LIST OF SYMBOLS AND ABBREVIATIONS

| | |
|---|---|
| OASIS | Open Access Series of Imaging Studies |
| eTIV | Estimated Total Intracranial Volume in $cm^3$ |
| MRI | Magnetic Resonance Imaging |
| nWBV | Normalized Whole Brain Volume |
| TR | Repetition Time |
| TE | Time to Echo |
| WBV | Whole Brain Volume in $cm^3$ |

# CHAPTER 1

## Introduction

Medical statistics is a subdiscipline of statistics. It has a major role in medical investigation. The identification of statistical patterns can help healthcare providers to track outbreaks. Health statistics can also assist in the allotment of research grants thereby determine the focus of our research efforts. One of the most crucial part of medical/health statistics is predicting a condition that do not have a cure, such as Alzheimer's disease and Dementia. The data from the MRI of the human brain is analyzed to detect patterns that could potentially be indicators of the progression of the disease.[1] Though it must be said that some confounding factors such as genetic disease, lifestyle are hard to quantify in the context of the human brain. In large parts, however, the analysis of the MRI of the human brain gives more information that can be used by researchers. It can also help in treatment of future disorder.[4]

### 1.1 QUESTIONS OF INTEREST

The cross-sectional MRI data of young, middle-aged and older adults was taken from the OASIS study. The OASIS is a series of MRI datasets that have been made public for study and analysis.[2] Every MRI contains two important quantitative variables, Intracranial Volume and Whole Brain Volume. This is an important indicator for brain activity in human beings. Statistical Analysis of these two variable answers various questions. Such as the average intracranial volume of human beings. It is important to note that the whole brain volume is a

portion of the intracranial volume, which also consists of spinal fluid volume. Therefore, the whole brain volume is the most representative of brain activity. However, measurement of the intracranial volume is the most accurate. Therefore, it can be surmised that the analysis of human brain MRI is a key part of medical statistics.

## 1.2 SAMPLING STRATEGY & ITS LIMITATION

The OASIS study selected 416 subjects aged 18 to 96 years. The subjects include both female and male, all of whom are predominantly right-handed. Each subjects participated in a single session of three or four individual T1 weighted MRI scanning.[2] T1- weighted images are produced using short TR and TE times.[3] The study took place in Washington University, Missouri. An additional 20 observations were made on 20 of the 416 subjects to ascertain reliability Thus totaling the observations to 436. For the purpose of this report, the observations for reliability and some outliers are removed prior to the analysis. For the purpose of this statistical report, numerous variables from the dataset of the OASIS study has not been included. The variables include Education Level, Socio-Economic status, Mini Mental State Examination, Clinical Dementia Rating, Atlas Scaling Factor. Many of these variables are subjective and provide no statistical value.

The spread of the subjects in terms of ethnicity is not documented. This is cause for concern because factors such ethnicity, gene identity could have an association with the goal of the study. Therefore, the strength of the analysis and its conclusion could be restricted.

### 1.3 VARIABLE DESCRIPTION

| Variable Name | Type of Variable |
|---|---|
| ID | - |
| **M.F** (Gender) | Categorical Variable |
| Dominant Hand | Categorical Variable |
| **Age** | Categorical Variable with 3 categories |
| Education Level | Quantitative Variable |
| Socioeconomic Status | Quantitative Variable |
| Mini Mental State Examination | Quantitative Variable |
| Clinical Dementia Rating | Quantitative Variable |
| **Estimated Total Intracranial Volume (eTIV)** | Quantitative Variable |
| **Normalize Whole Brain Volume (nWBV)** | Quantitative Variable |
| Atlas Scaling Factor | Quantitative Variable |

**Note:** For the purpose of statistical analysis, only the variables in bold are used.

Due to missing data and outliers, the 436 dataset is reduced to 407.

| Variable Name | Description |
|---|---|
| ID | Identification of the Subjects |
| **M.F** (Gender) | Identifies the Gender of the Subjects |
| Dominant Hand | Identifies the subject's dominant hand |
| **Age** | Classified as Young, Middle or Older |
| Education Level | Indicates the highest level of education completed |
| Socioeconomic Status | Indicates the subject's socio-economic status |
| Mini Mental State Examination | Indicates the result of a mental state examination |
| Clinical Dementia Rating | Indicates the result of clinical test for dementia |
| **eTIV** | eTIV of the subjectmeasured in $cm^3$ |
| **nWBV** | nWBV of the subject measured and normalized. |
| Atlas Scaling Factor | The scaling factor used to normalize the WBV. |

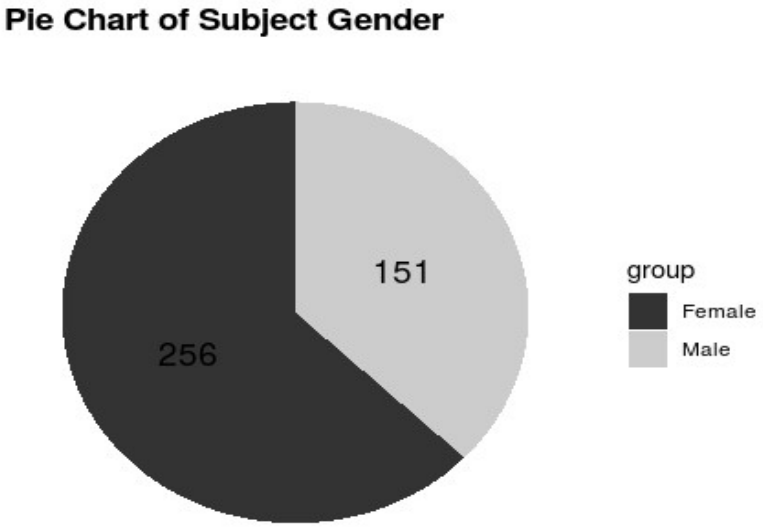
| Variable Name | Scale |
|---|---|
| ID | - |
| **M.F** (Gender) | "M" for MALE and "F" for FEMALE |
| Dominant Hand | "R" for Right-handed and "L" for Left-handed |
| **Age** | Young: Age < 45 Middle: 44< Age < 66 Older: Age >65 |
| Education Level | From 1 to 5 |
| Socioeconomic Status | From 1 to 5 |
| Mini Mental State Examination | From 0 to 30 |
| Clinical Dementia Rating | Between 0 and 2 |
| **eTIV** | Between 0 and 2000 in $cm^3$ |
| **nWBV** | Between 0 and 1 as WBV is normalized |
| Atlas Scaling Factor | Between 0 and 2 |

# CHAPTER 2

## Exploratory Analysis and Visualization

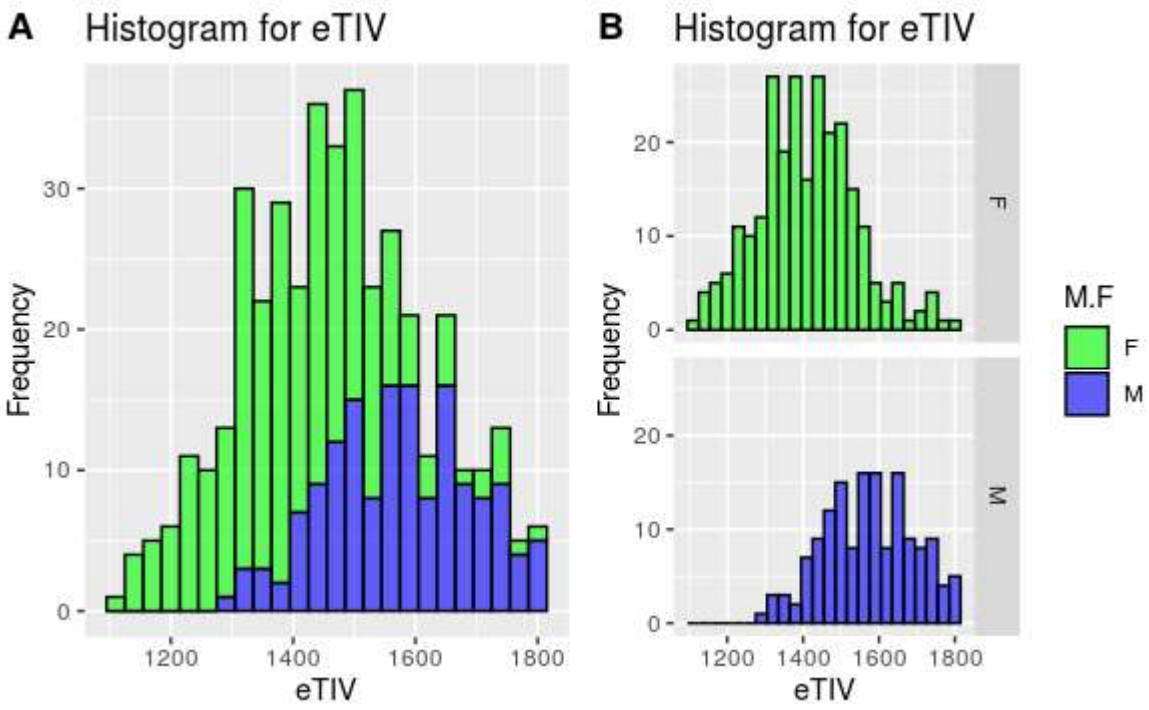
### 2.1 SUMMARY STATISTICS

The sample eTIV has a mean of **1472 cm³,** median of **1473 cm³** standard deviation of **147.8577.** For the purpose of this report, sample nWBV is used after de-normalizing the values by multiplying with the Estimated Intracranial Volume. Other variables such MMSE, CDR, ASF, SES, Edu are not used for the statistical analysis. These variables can be taken as confounding variables and can be used for a broader analysis work.




The pie chart above shows the split of the two genders in the experiment subjects. It is

normally assumed that for a statistical parity, each group is taken in equal proportions. However, this not the case for the OASIS study done.

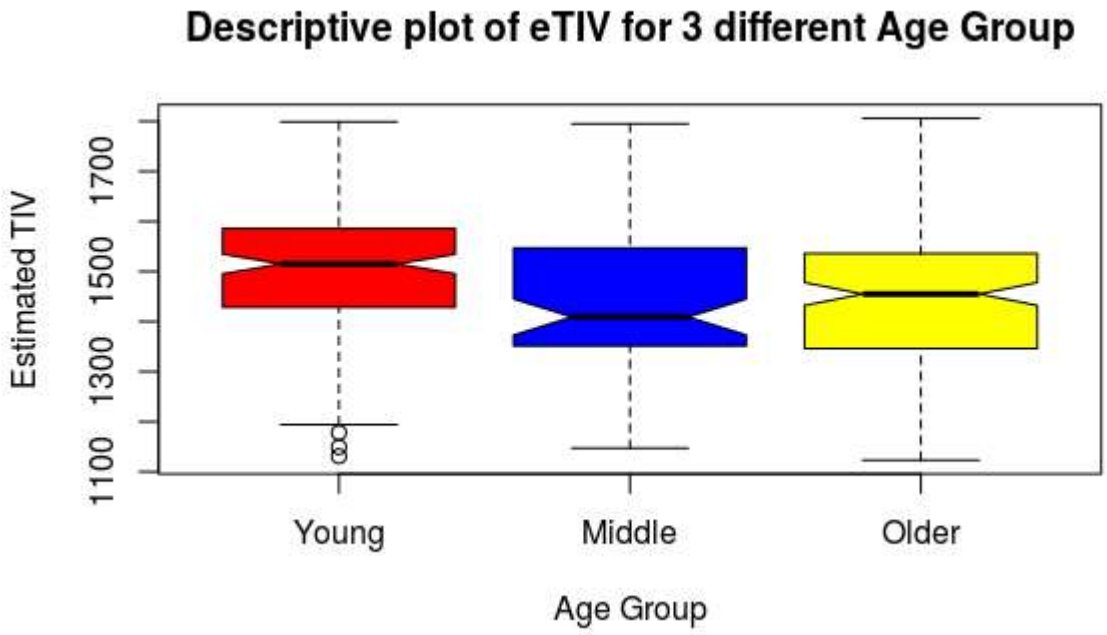



The above histogram gives us a composite picture of the two samples. We can observe that the spread of eTIV is high for females. While that of males, the spread is not as much.

Another interesting observation is that the mean of **eTIV** for **males** is **1574 cm$^3$** and that of **females** is **1410 cm$^3$**. This can be interpreted as "males generally have more ICV than females". However, this does not imply that males have more brain activity than females.

## 2.2 DESCRIPTIVE STATISTICS

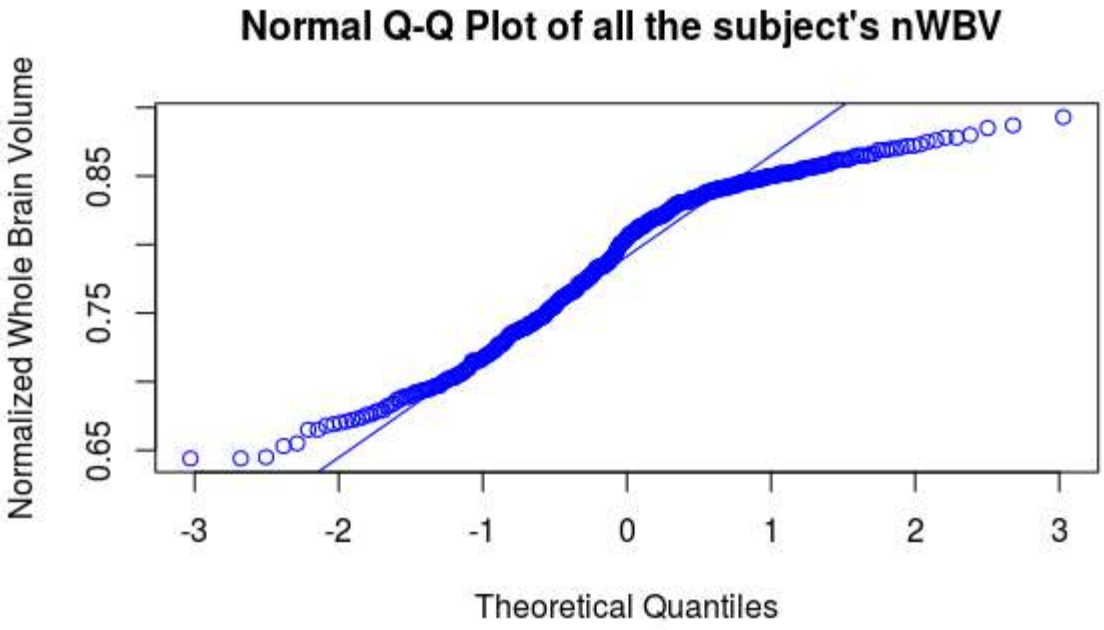
Boxplot is a method by which the data can be described in terms of its statistic. The boxplot below describes the three age groups (Young, Middle and Older) in terms of their respective eTIV.




It can be observed that the sample minimum, sample maximum, sample median and each of the sample quartiles are slightly higher for Young Adults. Though this is not a proof, the general perception that the brain deteriorates as age increases, is supported. Outliers are also observed for the young age group. These outliers are not actual outliers when it relates to the experiment as whole. The outliers in this case are observed only for the respective age group, therefore they will not be removed during preprocessing for the analysis.

## 2.3 NORMALIZED VARIABLE

We can notice from the dataset straightaway that the Whole Brain Volume has been normalized with an Atlas Scaling Factor. However, it is considered that the WBV is a part of the TIV.[5] Normalization is usually done to compare data from different sources more easily by eliminating the units of measurement ($cm^3$ in this case).




The above graph shows that the nWBV of all subjects in the OASIS study is almost symmetric in the distribution. We will see more Normal Q-Q plots in chapter 3 of this report. Many statistical tests require that the data be distributed normally i.e symmetric.

# CHAPTER 3

## Statistical Analysis

The exploratory analysis done in Chapter 2, only has a basic statistical interpretation. In this chapter, we will ask and answer five statistically significant questions pertaining to the MRI dataset.

## 3.1 AVERAGE INTRACRANIAL VOLUME OF HUMANS

Our sample consists of demented and non-demented subjects in it. Therefore, the average eTIV from the sample can be a realistic representation of the true population average intracranial volume of all human beings. By performing both traditional statistical and bootstrap approach we can have more certainty about the results of our test. In general, intracranial volume varies for different people. However, it is generally considered that the average eTIV is around **1500 $cm^3$.**

### 3.1.1 Statistical Test

As we aim to hypothesize a true population average of eTIV which involves one sample and there is no statistical data about the population, we perform a **one sample t-test.**

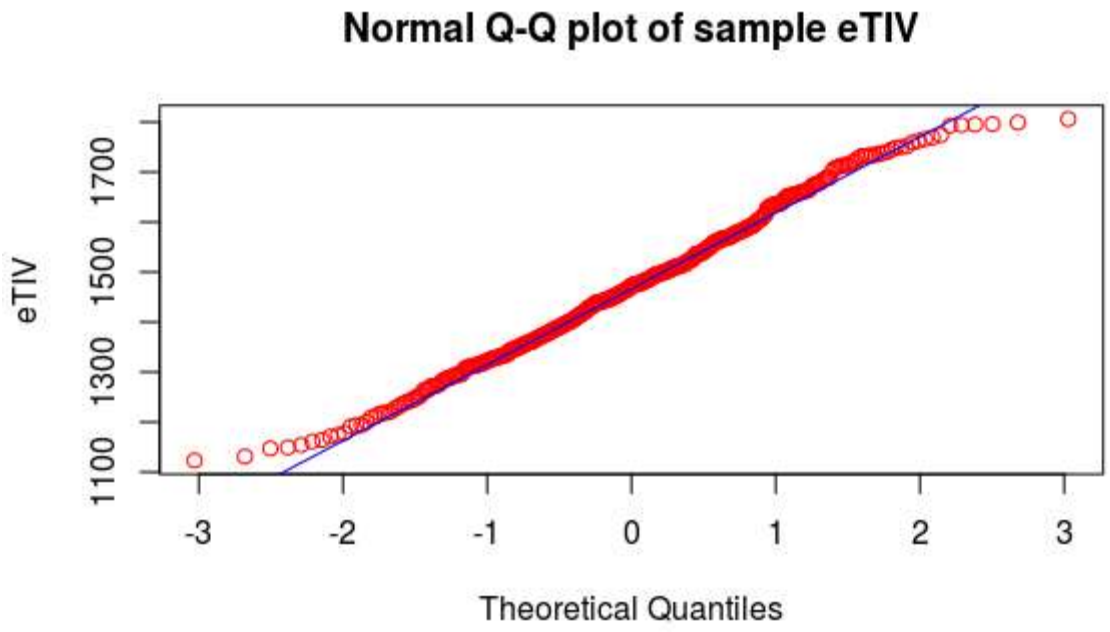
### 3.1.2 Population Parameter

We are interested in the true population, average Estimated Total Intracranial Volume.

$$\mu_{eTIV}$$

### 3.1.3 Validating Selected Statistical Method

To confirm that our dataset and the selected test statistic are compatibility, there are certain requirements that need to be satisfied.




As we can observe from the above graph, the data is symmetric and the datapoints are more dense towards the center of the graph. Thus we can conclude that the sample data is normally distributed.

### 3.1.4 Other requirements for Validation

We have seen that the data is normally distributed. We can also confirm some of the other requires that are necessary for the one-sample t-test.

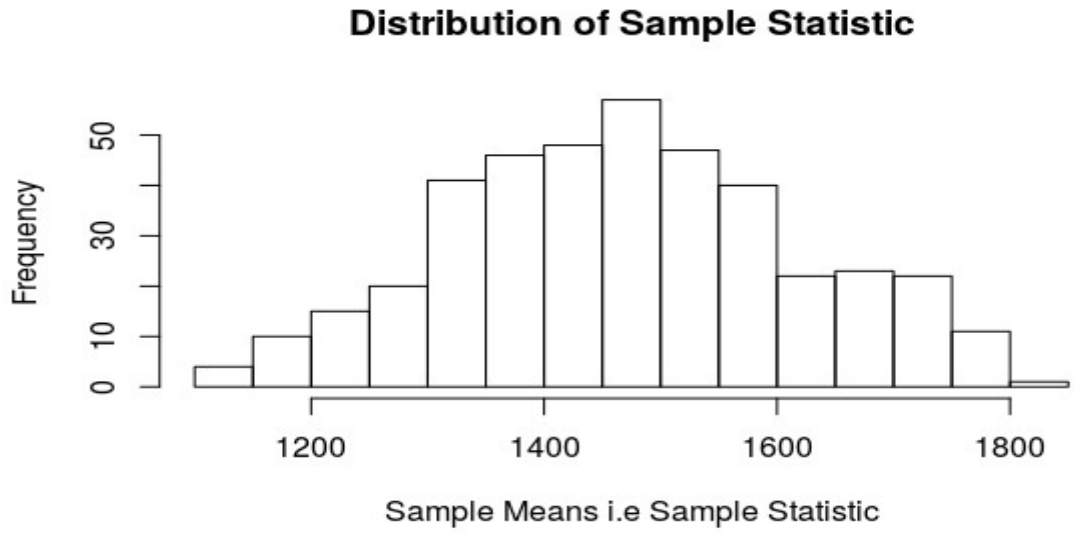
- We know that the sample subjects are from a single population of humans.
- We do not know the means and the standard deviation of population.

- We have only one quantitative variable of interest.

### 3.1.5 SAMPLE STATISTIC & TEST STATISTIC

$$\overline{x}_{eTIV}$$

The above mathematical expression is the sample statistic. The distribution graph below shows that the Sample Statistic is almost symmetric with the mean eTIV near the center.




$$t_{(n-1)} = \frac{(\overline{x}) \ - \ (\mu_0)}{\frac{s}{\sqrt{n}}}$$

The above mathematical expression is the test statistic.

### 3.1.6 Null Hypothesis

We hypothesized that the true population mean of eTIV is equal to **1500 cm³**. This is our Null Hypothesis.

$$H_0: \mu_{eTIV} = 1500\ cm^3$$

### 3.1.7 Alternate Hypothesis

We hypothesized that the mean of eTIV is not equal to **1500 cm3.** This is our Alternate Hypothesis.

$$H_A: \mu_{eTIV} \neq 1500\ cm^3$$

### 3.1.8 TEST RESULTS

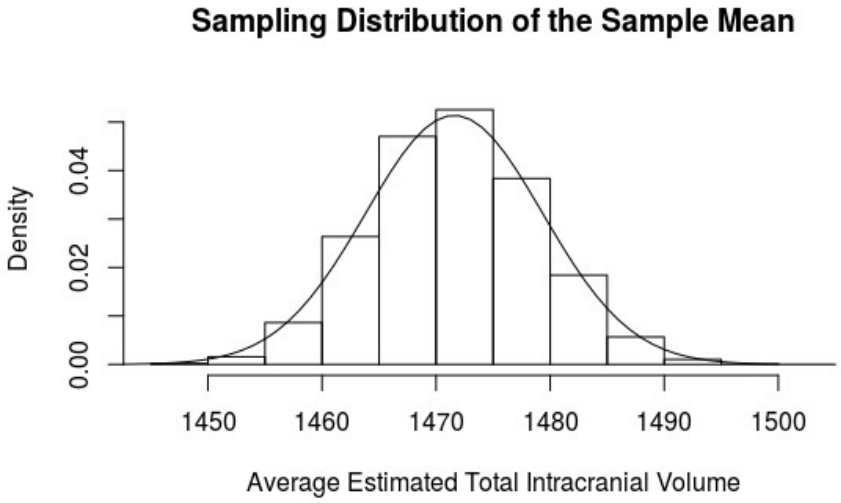
Upon computing the test statistic in R, we obtain test statistic value as **-3.877**. We have assumed a confidence level of 95%, to obtain the p-value. For the null hypothesis, the test statistic has a t-distribution with degree of freedom of 406. To complete the test, p-value was obtained by considering a two-sided test as our alternate hypothesis considers the inverse of the null hypothesis. Hence, both the area under the left and right tails of the t-distribution are calculated. The calculated **p-value** is **0.0001229**.

#### 3.1.8.1 Computation of Confidence Interval

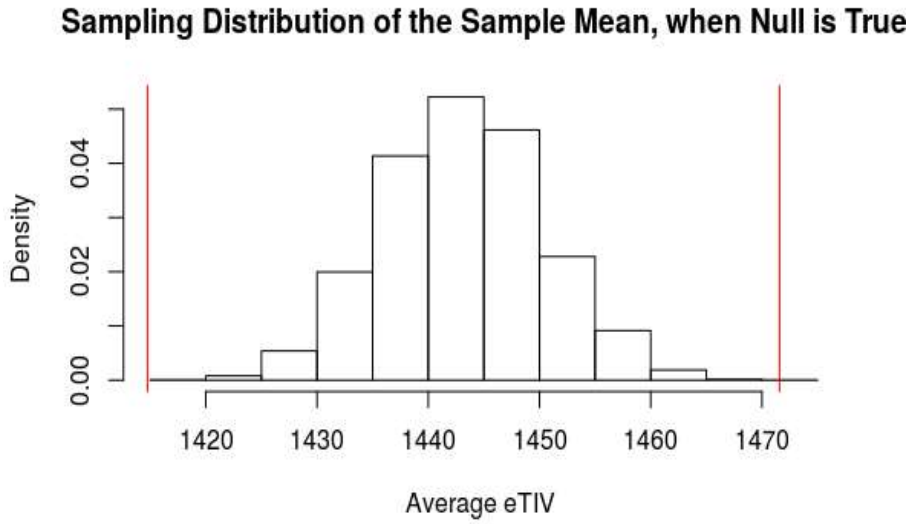
Critical Values are needed to obtain our confidence interval. The critical values obtained are **1.965824** and **-1.965824**. The resulting confidence interval is between **1457.172** and **1485.987.**

### 3.1.9 Bootstrap Approach

For bootstrap testing, we create random samples of our sample by performing ten thousand simulations with replacement. Upon computing the bootstrap test statistic in R, we obtain p-value as **0.0001**. We have assumed a confidence level of 95%, to obtain the p-value.




The above graph represents the sample distribution of the sample means from the bootstrap test. We assume that the null hypothesis is true and proceed with the bootstrapping to find the p-value. The resulting bootstrap confidence interval is between **1457.115** and **1486.045**




### 3.1.10 Interpretation and Test Conclusions

The interval using the traditional method is wider which agrees with our conservative p-value (**p-value** is **0.0001229).** In this case the results of the traditional test and the bootstrap test is very similar, only with minor decimal level changes in confidence intervals.

## 3.2 EQUAL SPLIT OF GENDER IN SUBJECTS SELECTED

Our sample consists of male and female subjects in it. It is usually considered that for any experiment or study, the test subjects must be proportionally split based on factors such as Age Group, Race, Gender, etc. However, for the purpose of this report, let us consider only gender as the factor. By performing both traditional statistical and bootstrap approach we can have more certainty about the results of our test. In general (population) though, there are more females than males.

### 3.2.1 Statistical Test

As we aim to hypothesize the proportion of females and males, we perform a **one sample test of proportions.**

### 3.2.2 Population Parameter

We are interested in the proportion of females in the OASIS study. The inverse of proportion of females gives the proportion of males.

$$\mathrm{p}_f$$

### 3.2.3 Validating Selected Statistical Method

To confirm that our dataset and the selected test statistic are compatibility, there are certain requirements that need to be satisfied.

- We have only one categorical variable of interest, with two categories females and males.
- We take females as success trails and males as failure trails, for the Exact Binomial Test.
- $np \geq 10 \ \ and \ \ n(1-p) \geq 10.$ : This condition is satisfied as we have large sample size 'n'.

### 3.2.4 SAMPLE STATISTIC & TEST STATISTIC

$$\mathrm{p}_f = \frac{Number\ of\ Females\ in\ the\ study}{Total\ Number\ of\ Subjects\ in\ the\ study}$$

The above mathematical expression is the sample statistic.

$$\mathrm{z} = \frac{(p\hat{}_f - p_f)}{\sqrt{\frac{p_f(1-p_f)}{n}}} \sim N(0,1)$$

The above mathematical expression is the test statistic.

### 3.2.6 Null Hypothesis

We hypothesized that the true proportion of females is equal to **0.5**. This is our Null Hypothesis.

$$H_0: p_f = 0.5$$

### 3.2.7 Alternate Hypothesis

We hypothesized the true proportion of females is not equal to **0.5.**. This is our Alternate Hypothesis.

$$H_A: p_f < 0.5$$

### 3.2.8 TEST RESULTS

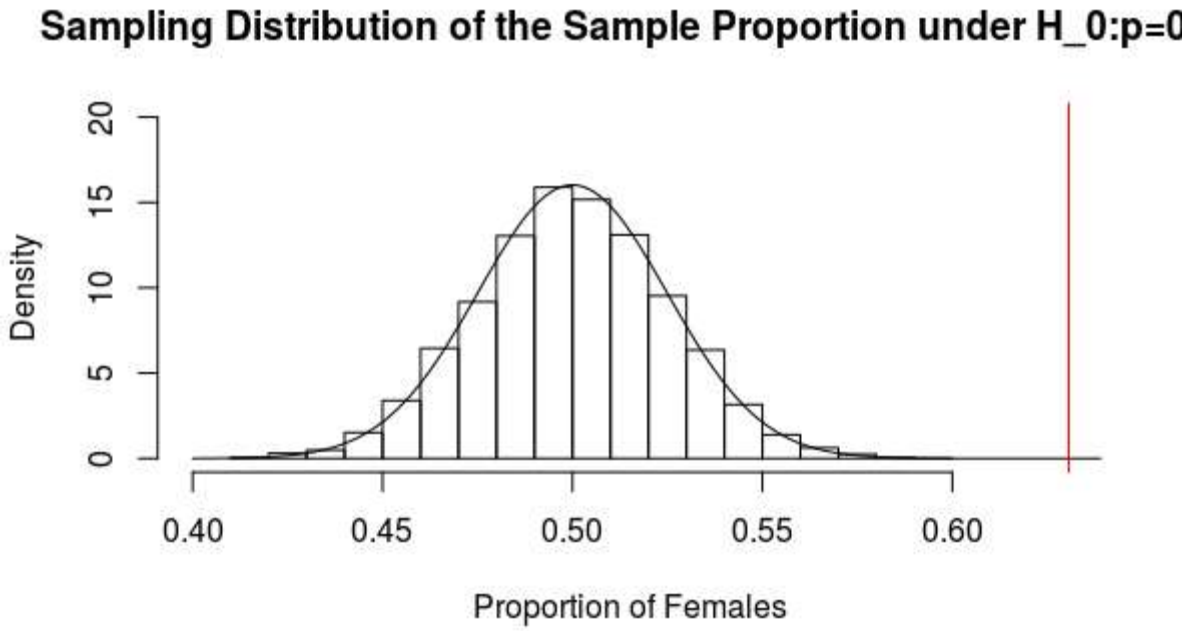
Upon computing the test statistic in R, we obtain test statistic value as **5.271661**. We have assumed a confidence level of 95%, to obtain the p-value. For the null hypothesis, the test statistic has a t-distribution with degree of freedom of 406. To complete the test, p-value was obtained by considering a one-sided test as our alternate hypothesis considers the inverse of the null hypothesis. Hence, both the area under the left and right tails of the t-distribution are calculated. The calculated **p-value** is **0.99999**. Therefore, we fail to reject the null hypothesis.

#### 3.2.8.1 Computation of Confidence Interval

The resulting 95% confidence interval is calculated to be between **0.622679** and **0.622679.**

### 3.2.9 Bootstrap Approach

For bootstrap testing, we create random samples of our sample by performing ten thousand simulations with replacement. Upon computing the bootstrap test statistic in R, we obtain p-value as **1**. We have assumed a confidence level of 95%, to obtain the p-value.




The above graph represents the sample distribution of the sample means from the bootstrap test. We assume that the null hypothesis is true and proceed with, $p_f = 0.5$ to find the p-value. The resulting confidence interval is between **0.6683047**and **0.6683047.**

### 3.2.10 Interpretation and Test Conclusions

The p-value using the traditional method is significant which agrees with our bootstrap p-value (**p-value** is **1).** In this case the results of the traditional test and the bootstrap test is very similar, only with minor decimal level changes to the p-value. The confidence interval obtained at $\alpha = 0.05$ *level* is **0.6226729** with traditional approach. The confidence interval obtained at $\alpha = 0.05$ *level* is between **0.528255** and **0.6683047** with bootstrap approach. The Exact Binomial Test also produced similar results.

## 3.3 COMPARISON OF ETIV OF MALES AND FEMALES

Our sample consists of male and female subjects in it. Therefore, the difference in mean eTIV from the sample can be a realistic representation of the true population difference in mean intracranial volume of all human beings. By performing both traditional statistical and bootstrap approach we can have more certainty about the results of our test. In general, intracranial volume varies for different people.

### 3.3.1 Statistical Test

As we aim to compare the eTIV of males and females, the comparison is made between two samples of two difference populations and there is no statistical data about the population, we therefore perform a **two sample t-test for difference in means.**

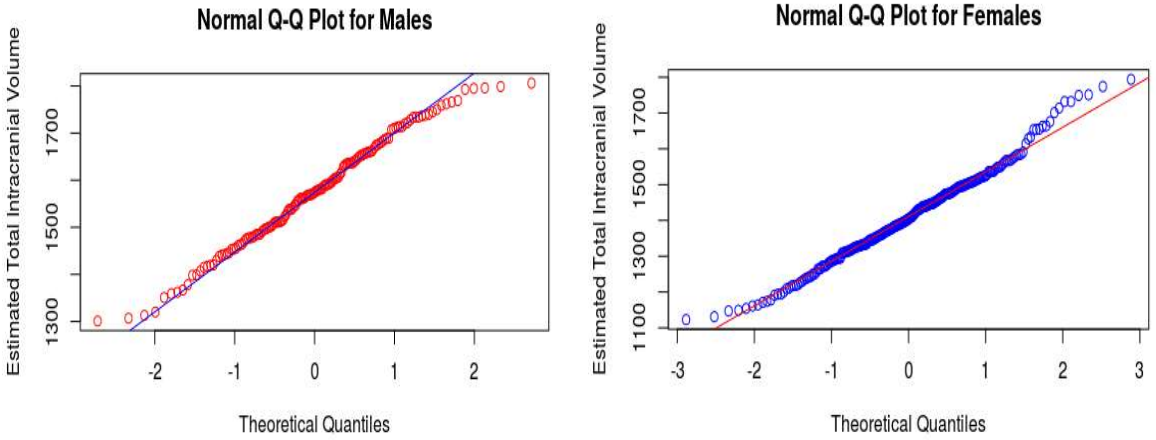
### 3.3.2 Population Parameter

We are interested in the true population mean difference in average Estimated Total Intracranial Volume between males and females.

$$\mu_{male\_eTIV} - \mu_{female\_eTIV}$$

### 3.3.3 Validating Selected Statistical Method

To confirm that our dataset and the selected test statistic are compatibility, there are certain requirements that need to be satisfied.




As we can observe from the above graph, the data is symmetric and the datapoints are more dense towards the center of the graph. Thus we can conclude that the sample data is normally distributed.

### 3.3.4 Other requirements for Validation

We have seen that the data is normally distributed. We can also confirm some of the other requires that are necessary for the two-sample t-test.

- We know the number of categories is two.
- We do not know the population means and the population standard deviation of two categories

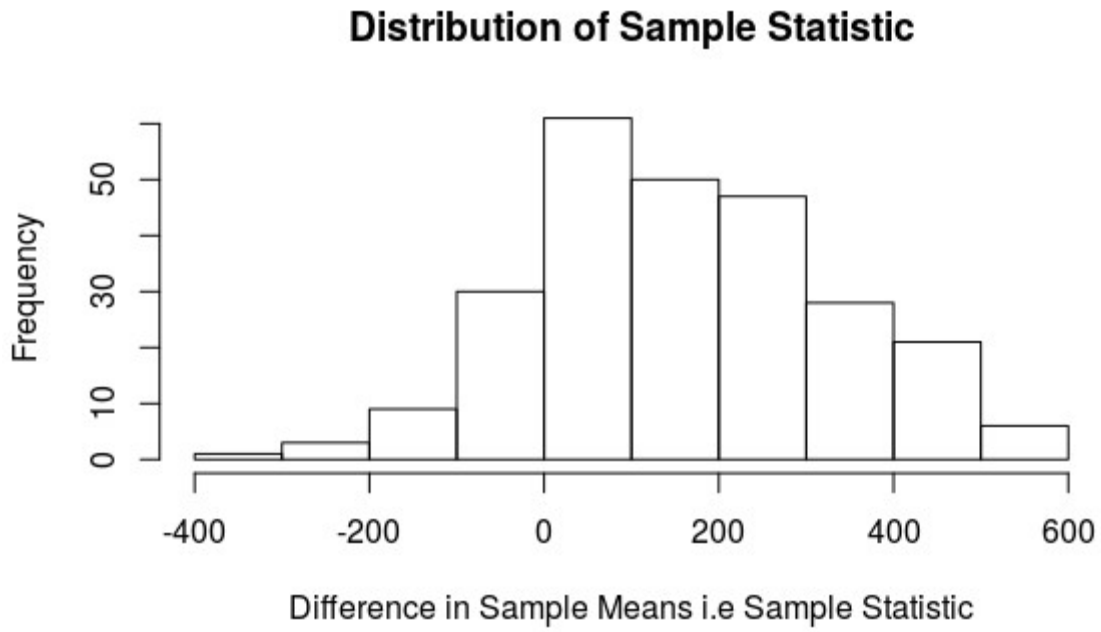
### 3.3.5 SAMPLE STATISTIC & TEST STATISTIC

$$\overline{x}_{male\_eTIV} - \overline{x}_{female\_eTIV}$$

The above mathematical expression is the sample statistic

$$t_{min(n_n-1,n_x-1)} = \frac{(\overline{x}_n - \overline{x}_s) - (\mu_n - \mu_s)}{\sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}}}$$

The above mathematical expression is the test statistic. The histogram below represents the sample distribution. The sample is distributed normally. The histogram below consists of the distribution of difference in sample means i.e. the sample statistic




### 3.3.6 Null Hypothesis

We hypothesized that the true population mean of male eTIV is equal to the true population mean of female eTIV. This is our Null Hypothesis.

$$H_0: \mu_{male\_eTIV} - \mu_{female\_eTIV} = 0$$

### 3.3.7 Alternate Hypothesis

We hypothesized that the true population mean of male eTIV is not equal to the true population mean of female eTIV. This is our Alternate Hypothesis.

$$H_A: \mu_{male\_eTIV} - \mu_{female\_eTIV} \neq 0$$

### 3.3.8 TEST RESULTS

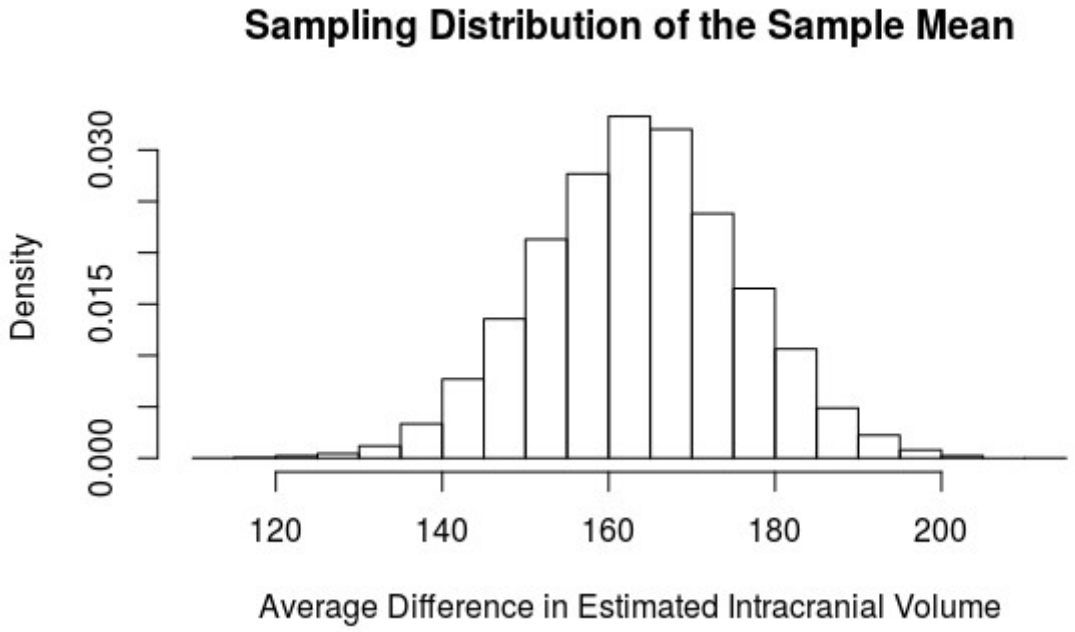
Upon computing the test statistic in R, we obtain test statistic value as **13.0774.** We have assumed a confidence level of 95%, to obtain the p-value. For the null hypothesis, the test statistic has a t-distribution with degree of freedom of 406. To complete the test, p-value was obtained by considering a two-sided test as our alternate hypothesis considers the inverse of the null hypothesis. Hence, both the area under the left and right tails of the t-distribution are calculated. The calculated **p-value** is **$1.4606e^{-26}$**.

#### 3.3.8.1 Computation of Confidence Interval

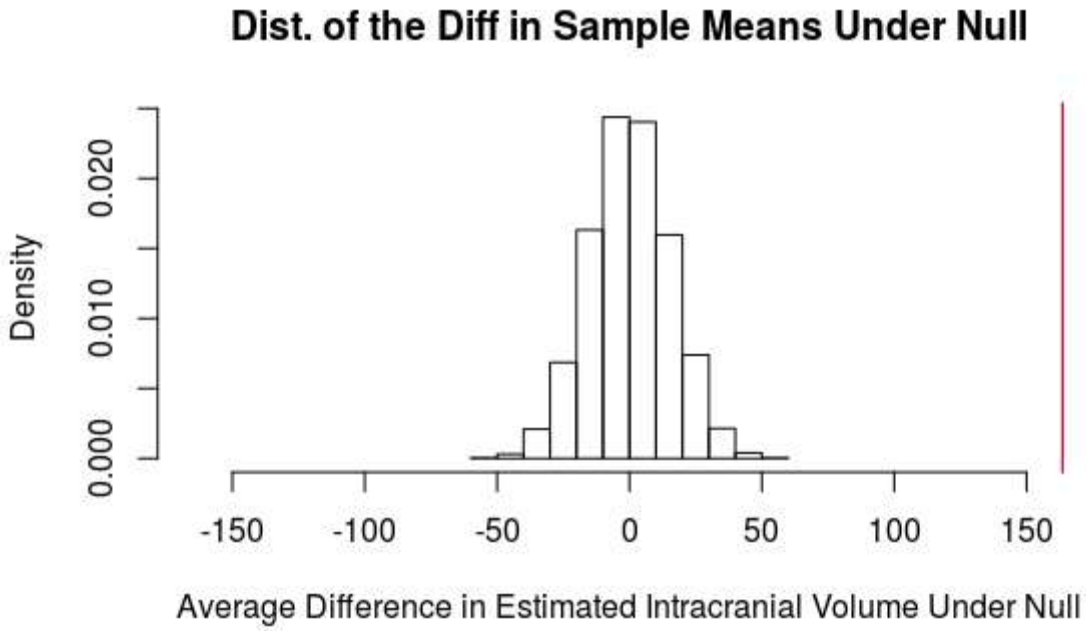
Critical Values are needed to obtain our confidence interval. The critical values obtained are **-1.975905** and **1.975905**. The resulting confidence interval is between **138.8817** and **188.3196.**

### 3.3.9 Bootstrap Approach

For bootstrap testing, we create random samples of our sample by performing ten thousand simulations with replacement. Upon computing the bootstrap test statistic in R, we obtain p-value as **0**. We have assumed a confidence level of 95%, to obtain the p-value.




The above graph represents the sample distribution of the sample means from the bootstrap test. We assume that the null hypothesis is true and proceed with the bootstrapping to find the p-value.




The one-sided Confidence interval under the Null is **187.6268.**

### 3.3.10 Interpretation and Test Conclusions

The interval using the traditional method is wider which agrees with our conservative p-value (**p-value** is **0).** We reject the null hypothesis that there is no difference between eTIV of males and the eTIV of female at the $\alpha = 0.05$ $level$. In this case, the results of the traditional test and the bootstrap test is very similar, only with minor decimal level changes in confidence intervals. This means that the true difference in the mean eTIV of males and eTIV of females is between **139.6880** and **187.6268.** The null hypothesized difference between the mean eTIV is zero, which is not in the 95% confidence interval. This is inconsistent with our hypothesis test. The results suggests that on average the female eTIV is higher than male eTIV.

## 3.4 COMPARISON OF WBV AND ETIV OF MALES AND FEMALES

Our sample consists of male and female subjects in it. The dataset has only a Normalized version of the WBV. The WBV is a proportion of the eTIV. We therefore perform a multiplication of the eTIV and nWBV to obtain WBV. A proportion of WBV to eTIV is calculated from the obtained result. Difference of this proportion is our question of interest. By performing both traditional statistical and bootstrap approach we can have more certainty about the results of our test.

### 3.4.1 Statistical Test

As we aim to compare the proportion of eTIV and WBV of males and females, the comparison is made between two samples of two difference populations and there is no statistical data about the population, we therefore perform a **two sample t-test for difference in proportions.**

### 3.4.2 Population Parameter

We are interested in the true population mean difference in average Estimated Total Intracranial Volume between males and females.

$$p_{male\_(WBV/eTIV)} - p_{female\_(WBV/eTIV)}$$

### 3.4.3 Validating Selected Statistical Method

To confirm that our dataset and the selected test statistic are compatibility, there are certain requirements that need to be satisfied.

- We have two categorical variables of interest, i.e., two independent samples from two populations.
- We take females as success trails and males as failure trails, for the Exact Binomial Test.

- $np \geq 10 \ \ and \ \ n(1-p) \geq 10.$ : This condition is satisfied as we have large sample size 'n'.

### 3.4.4 SAMPLE STATISTIC & TEST STATISTIC

$$\rho_{male_{(WBV/eTIV)}} \ - \ \rho_{female_{(WBV/eTIV)}}$$

The above mathematical expression is the sample statistic

$$\mathrm{z} = \frac{(\rho_{male_{(WBV/eTIV)}} \ - \ \rho_{fema \ \ _{(WBV/eTIV)}}) \ - \ p_{male\_(WBV/eTIV)} \ - \ p_{female\_(WBV/eTIV)}}{\sqrt{\frac{\rho_{male_{(WBV/eTIV)}}(1-\rho_{male_{(WBV/eTIV)}})}{n_{male_{(WBV/eTIV)}}} + \frac{\rho_{femal \ \ _{(WBV/eTIV)}}(1-\rho_{fema \ \ _{(WBV/eTIV)}})}{n_{female_{(WBV/eTIV)}}}}}$$

The above mathematical expression is the test statistic.

### 3.4.5 Null Hypothesis

We hypothesized that the true population proportion of male eTIV and WBV proportion is equal to the true population proportion of female eTIV and WBV. This is our Null Hypothesis.

$$H_0 : p_{male\_(WBV/eTIV)} \ - \ p_{female\_(WBV/eTIV)} = 0$$

### 3.4.6 Alternate Hypothesis

We hypothesized that the true population proportion of male eTIV and WBV proportion is not equal to the true population proportion of female eTIV and WBV.

This is our Alternate Hypothesis.

$$H_A: p_{male\_(WBV/eTIV)} - p_{female\_(WBV/eTIV)} \neq 0$$

### 3.4.7 TEST RESULTS

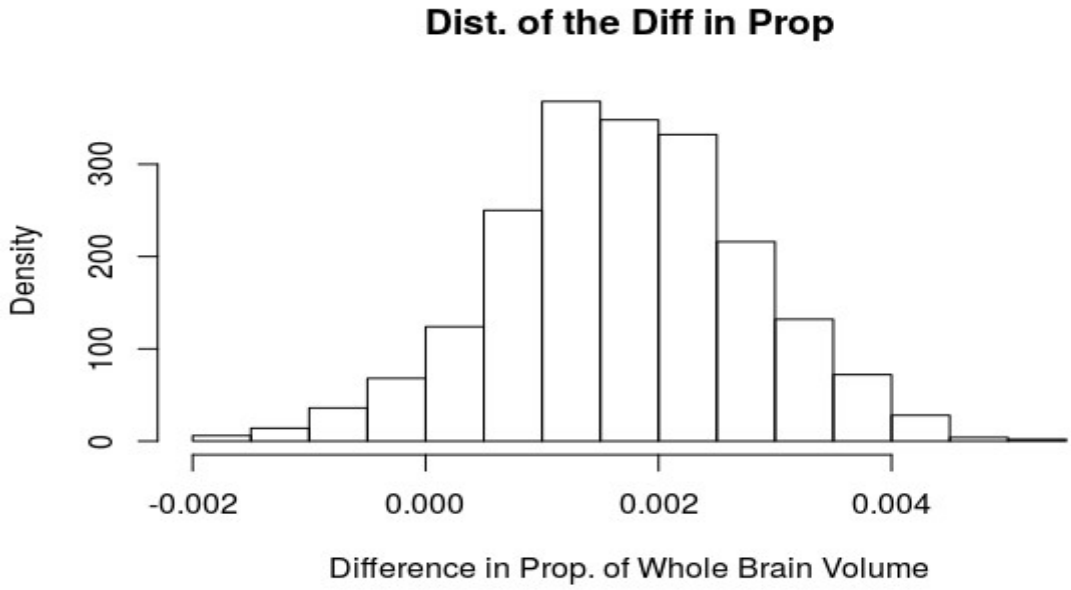
Upon computing the test statistic in R, we obtain test statistic value as **1.560666.** We have assumed a confidence level of 95%, to obtain the p-value. For the null hypothesis, the test statistic has a t-distribution with degree of freedom of 406. To complete the test, p-value was obtained by considering a two-sided test as our alternate hypothesis considers the inverse of the null hypothesis. Hence, both the area under the left and right tails of the t-distribution are calculated. The calculated **p-value** is **0.1186024**. We fail to reject the Null Hypothesis by the traditional method.

#### 3.4.7.1 Computation of Confidence Interval

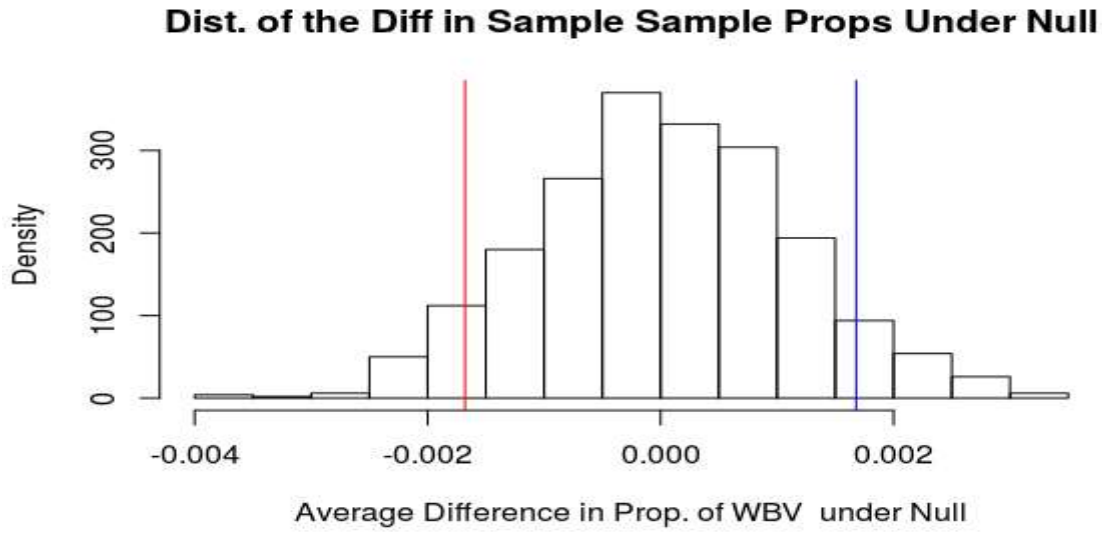
The resulting confidence interval is between **-0.0004295** and **0.003787.**

### 3.4.8 Bootstrap Approach

For bootstrap testing, we create random samples of our sample by performing one thousand simulations with replacement. Upon computing the bootstrap test statistic in R, we obtain p-value as **0.131**. We have assumed a confidence level of 95%, to obtain the p-value.




The above graph represents the distribution of the difference in proportions of the samples from the bootstrap test. We assume that the null hypothesis is true and proceed with the bootstrapping to find the p-value.




The resulting bootstrap confidence interval is between **-0.0005861** and **0.0038703.**

### 3.4.9 Interpretation and Test Conclusions

The interval using the traditional method is wider which agrees with our conservative p-value (**p-value** is **0.1186024).** We failed to reject the null hypothesis that the true population proportion of male eTIV and WBV proportion is equal to the true population proportion of female eTIV and WBV at the $\alpha = 0.05\ level.$ In this case, the results of the traditional test and the bootstrap test is very similar, only with minor decimal level changes in confidence intervals. This means that the true difference in the proportions is between **-0.0004295** and **0.003787.** The null hypothesized difference between the mean eTIV is zero, which is in the 95% confidence interval. This is consistent with our hypothesis test.

## 3.5 EQUAL SPLIT OF AGE GROUPS IN SUBJECTS SELECTED

Our sample consists of various age groups in it. It is usually considered that for any experiment or study, the test subjects must be proportionally split based on factors such as Age Group, Race, Gender, etc. However, for the purpose of this test, let us consider only Age Group as the factor. The dataset has been split into three groups based on the age. Young is classified as Age < 45. Middle is classified as Age between 65 and 45 . Older is classified as Age greater than 65. By performing both traditional statistical and bootstrap approach we can have more certainty about the results of our test. In general (population) though, there are more females than males.

### 3.5.1 Statistical Test

As we aim to hypothesize the proportion of age groups which has three categories, we perform a **chi-square goodness of fit test.**

### 3.5.2 Population Parameter

We are interested in the proportion of females in the OASIS study. The inverse of proportion of females gives the proportion of males.

$$p_{Young}, p_{Middle}, p_{Older}$$

### 3.5.3 SAMPLE STATISTIC & TEST STATISTIC

$$\rho_{Young}, \rho_{Middle}, \rho_{Older}$$

The above mathematical expression is the sample statistic.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E)^2}{E} \sim \chi^2_{k-1}$$

The above mathematical expression is the test statistic.

### 3.5.6 Null Hypothesis

We hypothesized that the true proportion of Young is equal to true proportion of Middle which is equal to true proportion of Older. This is our Null Hypothesis.

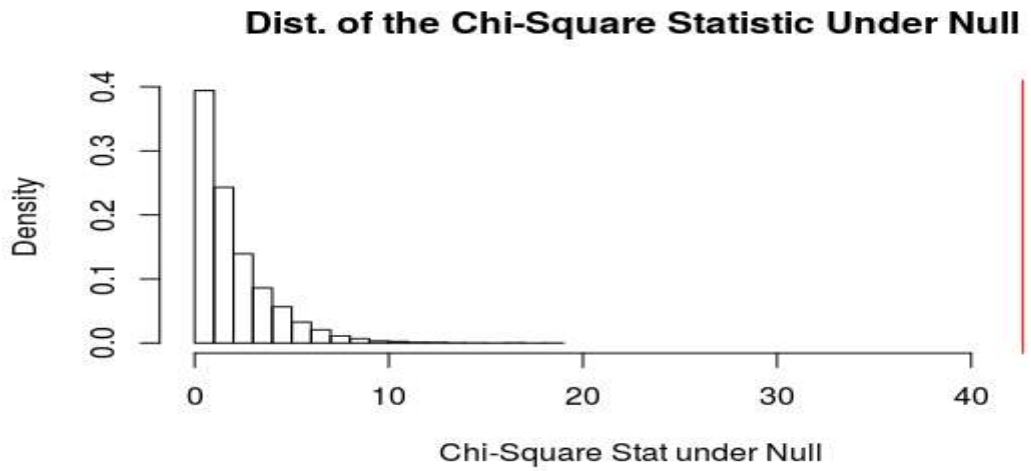
$$H_0: p_{Young} = p_{Middle} = p_{Older}$$

### 3.5.7 Alternate Hypothesis

We hypothesized that the true proportion of Young is not equal to true proportion of Middle which is not equal to true proportion of Older. This is our Alternate Hypothesis.

$$H_A: p_{Young} \neq p_{Middle} \neq p_{Older}$$

### 3.5.8 TEST RESULTS




Upon computing the test statistic in R, we obtain test statistic value as **42.66830**. We have assumed a confidence level of 95%, to obtain the p-value. For the null hypothesis, the test statistic has a t-distribution with degree of freedom of 406. To complete the test, p-value was obtained by considering a upper tail as our alternate hypothesis considers the inverse of the null hypothesis. Hence, both the area under the right tail of the t-distribution are calculated. The calculated **p-value** is **5.42e-10**. Therefore, we reject the null hypothesis.

### 3.5.9 Interpretation and Test Conclusions

There is evidence that any of the proportions of age groups is different to each other. We reject the Null Hypothesis at $\alpha = 0.05$ level. Thus we can conclude the variation in proportion of each age group is high and thereby affect the OASIS study.

# CHAPTER 4

# Conclusion

## 4.1 SUMMARY OF FINDINGS

We conducted 5 statistical tests to answer 5 questions with relation to the OASIS MRI data set. Each of the question and their answers are as follows:

***What is the Average eTIV of human beings?***

We hypothesized that the average eTIV is **1500 cm$^3$**. This null hypothesis was rejected at $\alpha = 0.05\ level$ . We find with 95% confidence that the true average of eTIV is approximately between **1457 cm$^3$** and **1486 cm$^3$**.

***Does the OASIS study have equal proportions of male and female test subjects?***

No, the OASIS study does not have equal proportions of male and female test subjects. Under the null hypothesis, we take the proportion of female is 0.5. Though, we fail to reject the null hypothesis. We can say with 95% confidence that the true proportion of female is between **0.6226729** and **0.668307.** This is supported by our failure to reject the null hypothesis.

***Is there a difference between eTIV of males and eTIV of females?***

Yes, there is a difference between eTIV of males and females. Under the null hypothesis, we consider that there is no difference between eTIV of males and females. This is rejected

as the p-value is 0 for the Bootstrap approach and almost 0 for the traditional approach. We can say with 95% confidence that the true difference in mean of eTIV of males and females is approximately between **139.6 $cm^3$** and **187.6 $cm^3$**.

***Do males and females have the same proportions of WBV to eTIV?***

There is strong evidence to suggest that males and females do have the same proportions of WBV to eTIV. We failed to reject the null hypothesis that the difference in proportion of WBV to eTIV of males and females is zero. We can say with 95% confidence that the true difference in proportions is approximately between **-0.0004295** and **0.003787.**

***Does the OASIS study have equal proportions of the three age groups in its subject pool?***

No, the OASIS study does not have equal proportions of the three age groups (Young, Middle and Older Adults) in it subject pool. We reject the same as our null hypothesis at $\alpha = 0.05$ level. We conclude that the variation in the proportions of the age groups is high. This affects the results of the OASIS study.

All the above tests compare the cross-sectional MRI data of Young, Middle and Older Adults. We find that although there is not an equal representation of females and males, as well as Young, Middle and Older Adults, the average Estimated Total Intracranial Volume and the Whole Brain Volume is a true representation of the general population. Thus, we can summarize that while there is a theoretical expectation that the dataset cannot have a true comparison made, it is not the case. Any comparisons made are generally found to be accurate.

## 4.2 IMPLICATIONS

That result does not mean the sample has no limitations. It implies that the more finetuning is done to the dataset, the more accurate the results get. There are other implications as to why the subjects did not have an equal representation among both genders and the three age groups. This could imply that the generally, there are more females and males. Likewise, there are more Young and Older Adults than Middle aged adults. But there is no implication that any of those two factors namely age group and gender have an affect on the Total Intracranial Volume and Whole Brain Volume.

## 4.3 EXTENSIONS AND LIMITATIONS

**Extensions:** We use statistical methods to study medical data. We can analyze and track many disorders with data such as those taken from an MRI, X-ray, etc... There is numerous study that use data from the medical industry and convert that into actionable use for healthcare providers. This could also help the health insurance industry.

**Limitations:** The questions that we have covered is limited in scope. We do not consider confounding variables and therefore any associations inferred does not hold verifiable. Sample collection method has its limitations. The subjects of the study do not have a great deal of variation about them apart from gender and

age group. This is a real limitation if the objective of the analysis is to find patterns for the general population. In the dataset, there is no information suggesting that the subjects are a true representation of the general population.

## 4.4 FUTURE WORK

Future work on this data-set can include other confounding variable such as genetics, life style, occupation and various other factors that are associated with brain activity. MRI data usually taken for the purpose of identify dementia and Alzheimer's disease generally have all the relevant information regarding the brain. It does not however, contain external variable. For future work, we could build a framework which takes input variables such as TIV, WBV, Age, Quantified Genetic Marker, Quantified Life style parameters. And the framework could return a possible match to disease.

# CHAPTER 5

# Appendix

## 5.1 DATA SOURCE

https://www.kaggle.com/jboysen/mri-and-alzheimers#oasis_cross-sectional.csv

## 5.2 BIBLIOGRAPHY


[1] Barbara A. bartkowiak, and Brian J.Finnegan *"Health Statistics"*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1069092/

[2] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky,
*"Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults ".*
https://www.ncbi.nlm.nih.gov/pubmed/17714011

[3] Magnetic Resonance Imaging (MRI) of the Brain and Spine: Basics
https://casemed.case.edu/clerkships/neurology/Web%20Neurorad/MRI%20Basics.htm.

[4] Richard Nordenskjold, *"Analysis of Human Brain MRI"*,
https://uu.diva-portal.org/smash/get/diva2:713385/FULLTEXT01.pdf

[5] Martin Reite, Erik Reite, Dan Collins, Peter Teale, Donald C. Rojas and Elliot Sandberg
*"Brain size and brain/intracranial volume ration in major mental illness ".*
https://bmcpsychiatry.biomedcentral.com/articles/10.1186/1471-244X-10-79

## 5.3 CODE APPENDIX

```
knitr::opts_chunk$set(echo = FALSE)
### Exploratory Analysis and Data Visualization
## Loading Packages
library(ggplot2)
library(cowplot)
library(scales)

## Reading Dataset
MRI_Data <- read.csv(file="oasis_cross-sectional.csv",header=TRUE)

## Rows 416 to 436 consists of MRI data for the same people already in the test.
##This was taken for testing reliability.

MRI_Data <- MRI_Data[c(1:416),];
MRI_Data <- MRI_Data[MRI_Data$eTIV<1818,]

# eTIV
summary(MRI_Data$eTIV)
mean(MRI_Data$eTIV)
median(MRI_Data$eTIV)
max(MRI_Data$eTIV)
min(MRI_Data$eTIV)
sd(MRI_Data$eTIV)

# nWBV
summary(MRI_Data$nWBV)
mean(MRI_Data$nWBV)
median(MRI_Data$nWBV)
```

```
max(MRI_Data$nWBV)
min(MRI_Data$nWBV)


# Age Classification
MRI_Data_Young  <-  MRI_Data[MRI_Data$Age<45,]
MRI_Data_Middle <- MRI_Data[MRI_Data$Age<66 & MRI_Data$Age>44,]
MRI_Data_Older  <-  MRI_Data[MRI_Data$Age>65,]

# Gender Classification
MRI_Data_M <- subset(MRI_Data,MRI_Data$M.F== "M")
MRI_Data_F <- subset(MRI_Data,MRI_Data$M.F== "F")
n_male=length(MRI_Data_M$ID)
n_female=length(MRI_Data_F$ID)

# eTIV for F
summary(MRI_Data_F$eTIV)
mean(MRI_Data_F$eTIV)
median(MRI_Data_F$eTIV)
max(MRI_Data_F$eTIV)
min(MRI_Data_F$eTIV)
sd(MRI_Data_F$eTIV)
# eTIV for M
summary(MRI_Data_M$eTIV)
mean(MRI_Data_M$eTIV)
median(MRI_Data_M$eTIV)
max(MRI_Data_M$eTIV)
min(MRI_Data_M$eTIV)
sd(MRI_Data_M$eTIV)


## Exploratory Analysis and Data Visualization

# Box Plot of three different age groups in terms of eTIV ----------1
boxplot(MRI_Data_Young$eTIV,MRI_Data_Middle$eTIV,MRI_Data_Older$eTIV,
          main="Descriptive plot of eTIV for 3 different Age Group",
          xlab = "Age Group", names=c("Young","Middle","Older"),
          ylab = "Estimated TIV", font.lab=3,
          notch = TRUE,col=c("red","blue", "yellow"))


# Histogram eTIV for Male and Female ----------2
p1<-ggplot(MRI_Data,aes(eTIV))+
  geom_histogram(binwidth=30,aes(fill=M.F),alpha=0.6,color="black")+
  labs(title="Histogram for eTIV", x="eTIV", y="Frequency")+
  scale_fill_manual(values = c("green","blue"))
p3<-p1+guides(fill=FALSE)
p2<-p1+facet_grid(M.F ~ .)
plot_grid(p3,p2,labels="AUTO")


## A Q-Q plot of nWBV----------3
qqnorm(MRI_Data$nWBV,
```

```
        col="blue",
        ylab="Normalized Whole Brain Volume",
        main="Normal Q-Q Plot of all the subject's nWBV ")
qqline(MRI_Data$nWBV,col="blue")

## Pie Chart of Male and Female Ratio----------4
df <- data.frame(
  group = c("Male", "Female"),
  value = c(n_male,n_female)
)

blank_theme <- theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    plot.title=element_text(size=14, face="bold")
  )

bp<- ggplot(df, aes(x="", y=value, fill=group))+
  geom_bar(width = 1, stat = "identity")
pie <- bp + coord_polar("y", start=0)
pie + scale_fill_grey() +  blank_theme + labs(title="Pie Chart of Subject Gender")+
  theme(axis.text.x=element_blank()) +
  geom_text(aes(y = value/2 + c(0, cumsum(value)[-length(value)]),
                label = (value)), size=5)
### Statistical Analysis
## ONE SAMPLE T- TEST
## Loading Packages
library(ggplot2)
library(cowplot)

## Reading Dataset
MRI_Data <- read.csv(file="oasis_cross-sectional.csv",header=TRUE)

## Rows 416 to 436 consists of MRI data for the same people already in the test.
## This was taken for testing reliability.
MRI_Data <- MRI_Data[c(1:416),];
MRI_Data <- MRI_Data[MRI_Data$eTIV<1818,]

head(MRI_Data)

## Creating a Subset for two Groups "Male" and "Female"
MRI_Data_M <- subset(MRI_Data,MRI_Data$M.F== "M")
MRI_Data_F <- subset(MRI_Data,MRI_Data$M.F== "F")

# Verifying Normality
qqnorm(MRI_Data$eTIV,main="Normal Q-Q plot of sample eTIV",ylab="eTIV",col="red")
qqline(MRI_Data$eTIV,col="blue")

# the parts of the test statistic
```

```r
# sample mean
x_bar <- mean(MRI_Data$eTIV)
# null hypothesized population mean
mu_0 <- 1500
# sample st. dev
s <- sd(MRI_Data$eTIV)
# sample size
n <- length(MRI_Data$eTIV)
# t-test test statistic
t <- (x_bar - mu_0)/(s/sqrt(n))
# two-sided p-value so multiply by 2
two_sided_t_pval <- pt(q = t, df = n-1, lower.tail = TRUE)*2
two_sided_t_pval

## Critical Values
qt(0.025, n-1)
qt(0.975, n-1)

# lower bound
x_bar+(qt(0.025, n-1)*(s/sqrt(n)))
# upper bound
x_bar+(qt(0.975, n-1)*(s/sqrt(n)))

## Verifying with Welch Test - Sanity Check
t.test(MRI_Data$eTIV,
         alternative = "two.sided",
         mu = 1500)

# Histogram of Distribution of Sample Statistic
hist(MRI_Data$eTIV,
      xlab='Sample Means i.e Sample Statistic',
      main="Distribution of Sample Statistic")



## Bootstrap
# Number of Simulations with seed = 0
set.seed(0)
num_sims <- 10000

# A vector to store my results
results <- rep(NA, num_sims)

# A loop for completing the simulation
for(i in 1:num_sims){
  results[i] <- mean(sample(x =MRI_Data$eTIV,
                            size = n, replace
                            = TRUE))
}

# Finally plot the results
hist(results,freq = FALSE, main='Sampling Distribution of the Sample Mean',
xlab = 'Average Estimated Total Intracranial Volume',
```

```
ylab = 'Density', ylim=c(0,0.056),)

# estimate a normal curve over it - this looks pretty good!
lines(x = seq(1400, 1500, 1),
      dnorm(seq(1400, 1500, 1),
      mean = x_bar,
      sd = 158.343744/sqrt(416)))


# Shift the sample so that the null hypothesis is true
etiv_given_H0_true <- MRI_Data$eTIV + mean(MRI_Data$eTIV) - mu_0

# Simulations to get accurate result with seed = 0
set.seed(0)
num_sims <- 10000


# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)

# A loop for completing the simulation
for(i in 1:num_sims){
  results_given_H0_true[i] <- mean(sample(x = etiv_given_H0_true,
                                          size = n, replace
                                          = TRUE))
}

# Finally plot the results
hist(results_given_H0_true, freq = FALSE,
     main='Sampling Distribution of the Sample Mean, when Null is True',
     xlab = 'Average eTIV ', ylab = 'Density')

# add line to show values more extreme on upper end
abline(v=x_bar, col = "red")

# add line to show values more extreme on lower end
low_end_extreme <- mean(results_given_H0_true)+(mean(results_given_H0_true)-x_bar)
abline(v=low_end_extreme, col="red")

# counts of values more extreme than the test statistic in our original sample,given H0is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= low_end_extreme)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= x_bar)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
bootstrap_pvalue

# two sided t p-value
two_sided_t_pval

# need the standard error which is the standard deviation of the results
bootstrap_SE_X_bar <- sd(results)
# an estimate is to use the formula statistic +/- 2*SE
c(x_bar - 2*bootstrap_SE_X_bar, x_bar + 2*bootstrap_SE_X_bar)
```

```r
# you can also use the 5th and 95th quantiles to determine the bounds:
c(quantile(results, c(.025, .975)))

# compare to our t-methods
c(x_bar+(qt(0.025, n-1)*(s/sqrt(n))), x_bar+(qt(0.975, n-1)*(s/sqrt(n))))

## ONE SAMPLE TEST OF PROPORTIONS
## Loading Packages
library(ggplot2)
library(cowplot)

## Reading Dataset
MRI_Data <- read.csv(file="oasis_cross-sectional.csv",header=TRUE)

## Rows 407 to 436 consists of MRI data for the same people already in the test.
## This was taken for testing reliability.
MRI_Data <- MRI_Data[c(1:407),]
MRI_Data <- MRI_Data[MRI_Data$eTIV<1818,]

head(MRI_Data)

## Creating a Subset for two Groups "Male" and "Female"
MRI_Data_M <- subset(MRI_Data,MRI_Data$M.F== "M")
MRI_Data_F <- subset(MRI_Data,MRI_Data$M.F== "F")
length(MRI_Data_M$ID)

n_m <-length(MRI_Data_M$ID)
n_f <-length(MRI_Data_F$ID)

p<- n_f/((n_m)+(n_f))
p

## Test Stat

## p_O = 0.5.. Null Hypothesis that the experiment was done with equal number of male and female
z <- (p - .5) / sqrt((.5*(1-.5)) / 407)
z

# One Sided Lower Tail
binom.test(x=256, n = 407, p=(.5), alternative="less")

pvalue<-pnorm(z, lower.tail = TRUE)
pvalue

binom.test(x=256, n = 461, p=(.5), alternative="less")$conf.int

cat("normal approx")
c(0,p - (qnorm(p))*sqrt(((p)*(1-p))/407))

## Bootstrap

people <- factor(rep(c("Female", "Male"), c(256, 407-256)))
people
```

```r
table(people)

people <- rep(c(1, 0), c(256, 407-256))
people

# Number of Simulations
num_sims <- 10000

# A vector to store my results
results <- rep(NA, num_sims)

# A loop for completing the simulation
for(i in 1:num_sims){
  results[i] <- mean(sample(x = people,
                            size = 407,
                            replace = TRUE))
}

# Finally plot the results
hist(results, freq = FALSE,ylim=c(0,20),
     main='Sampling Distribution of the Sample Proportion',
     xlab = 'Proportion of Femlaes', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(.45, 3.85, .001),
      dnorm(seq(.45, 3.85, .001),
      mean = mean(results), sd = sd(results)))

cat("Bootstrap Confidence Interval")
c(quantile(results, c(0,.95)))
cat("exact binomial test")
binom.test(x=256, n = 461, p=(.5), alternative="less")$conf.int
cat("normal approx")
c(0,p - (qnorm(p))*sqrt(((p)*(1-p))/407))


# Under the assumption that the null hypothesis is true, we have 50% females
people <- rep(c(1, 0), c(407/2, 407/2))
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
  results[i] <- mean(sample(x = people,
                            size = 407,
                            replace = TRUE))
}

# Finally plot the results
hist(results, freq = FALSE,ylim=c(0,20),xlim=c(0.4,0.63),
main='Sampling Distribution of the Sample Proportion under H_0:p=0.5',
xlab = 'Proportion of Females', ylab = 'Density')

# estimate a normal curve over it - this looks pretty good!
```

```
lines(x = seq(.40, 9.85, .001),
      dnorm(seq(.40,9.85, .001),
      mean = mean(results), sd = sd(results)))
abline(v=p, col="red")

## Bootstrap p-value
count_of_more_extreme_lower_tail <- sum(results <= p)
bootstrap_pvalue <- count_of_more_extreme_lower_tail/num_sims
bootstrap_pvalue
binom.test(x=256, n = 407, p=(.5), alternative="less")$p.value
pnorm(z, lower.tail = TRUE)
## TWO SAMPLE T- TEST FOR DIFFERENCE IN MEANS
## Loading Packages
library(ggplot2)
library(cowplot)

## Reading Dataset
MRI_Data <- read.csv(file="oasis_cross-sectional.csv",header=TRUE)

## Rows 416 to 436 consists of MRI data for the same people already in the test.
## This was taken for testing reliability.
MRI_Data <- MRI_Data[c(1:416),]
MRI_Data <- MRI_Data[MRI_Data$eTIV<1818,]
head(MRI_Data)
boxplot(MRI_Data$eTIV)
## Creating a Subset for two Groups "Male" and "Female"
MRI_Data_M <- subset(MRI_Data,MRI_Data$M.F== "M")
MRI_Data_F <- subset(MRI_Data,MRI_Data$M.F== "F")

# Statistical Analysis

# Checking Normality
qqnorm(MRI_Data$eTIV,
       col="green",
       ylab="Estimated Total Intracranial Volume",
       main="Normal Q-Q Plot from the subjects")
qqline(MRI_Data$eTIV,col="blue")
qqnorm(MRI_Data_M$eTIV,
       col="red",
       ylab="Estimated Total Intracranial Volume",
       main="Normal Q-Q Plot for Males")
qqline(MRI_Data_M$eTIV,col="blue")
qqnorm(MRI_Data_F$eTIV,
       col="blue",
       ylab="Estimated Total Intracranial Volume",
       main="Normal Q-Q Plot for Females")
qqline(MRI_Data_F$eTIV,col="red")

# Sample Means of the two Categories
x_bar_n <- mean(MRI_Data_M$eTIV)
x_bar_s <- mean(MRI_Data_F$eTIV)
# Null Hypothesized population mean difference
mu_0 <- 0
```

```r
# Sample Variances
s_n_sq <- sd(MRI_Data_M$eTIV)**2
s_s_sq <- sd(MRI_Data_F$eTIV)**2
# Sample Size
n_n <-length(MRI_Data_M$eTIV)
n_s <- length(MRI_Data_F$eTIV)


# T-Test Statistic
t <- (x_bar_n - x_bar_s - mu_0)/sqrt((s_n_sq/n_n) + (s_s_sq/n_s))
t
# Degrees of Freedom
df = min(n_n, n_s)-1
df
# Two Sided p-value
two_sided_diff_t_pval <- pt(q = t, df = min(n_n, n_s)-1, lower.tail =
                            FALSE)*2
two_sided_diff_t_pval
## Confidence Interval
# Critical Values
qt(0.975, min(n_n, n_s)-1)
qt(0.025, min(n_n, n_s)-1)
# Upper Bound
(x_bar_n-x_bar_s)+(qt(0.975, min(n_n, n_s)-1)*sqrt((s_n_sq/n_n) +
                                                   (s_s_sq/n_s)))
# Lower Bound
(x_bar_n-x_bar_s)+(qt(0.025, min(n_n, n_s)-1)*sqrt((s_n_sq/n_n) +
                                                   (s_s_sq/n_s)))

# Verifying results with Welch Test
t.test(MRI_Data_M$eTIV,MRI_Data_F$eTIV)

# Histogram of Distribution of Sample Statistic
hist(MRI_Data_M$eTIV-MRI_Data_F$eTIV,
     xlab='Difference in Sample Means i.e Sample Statistic',
     main="Distribution of Sample Statistic")



# Bootstrap Approach

num_sims <- 10000
results <- rep(NA, num_sims)
for(i in 1:num_sims){
  mean_Male <- mean(sample(x = MRI_Data_M$eTIV,
                           size = 160,
                           replace = TRUE))
  mean_Female <- mean(sample(x = MRI_Data_F$eTIV,
                             size = 256,
                             replace = TRUE))
  results[i] <- mean_Male - mean_Female
}
# Finally plot the results
```

```
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean',
xlab = 'Average Difference in Estimated Intracranial Volume', ylab = 'Density')

# Bootstrap one-sided CI
c(quantile(results, c(.025, .975)))
# compare to our t-methods
t.test(MRI_Data_M$eTIV, MRI_Data_F$eTIV)


# Check out the transform function used to shuffle
transform(MRI_Data, M.F=sample(M.F))
num_sims <- 10000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
  # idea here is if there is no relationshipm we should be able to shuffle the groups
  shuffled_gender <- transform(MRI_Data, M.F=sample(M.F))
  mean_Male <- mean(shuffled_gender$eTIV[shuffled_gender$M.F=="M"])
  mean_Female <- mean(shuffled_gender$eTIV[shuffled_gender$M.F=="F"])
  results_given_H0_true[i] <- mean_Male - mean_Female
}

# Finally plot the results
hist(results_given_H0_true, freq = FALSE, xlim=c(-165,165),
     main='Dist. of the Diff in Sample Means Under Null',
     xlab = 'Average Difference in Estimated Intracranial Volume Under Null',
     ylab = 'Density')
diff_in_sample_means <- mean(MRI_Data_M$eTIV)- mean(MRI_Data_F$eTIV)
abline(v=diff_in_sample_means, col = "blue")
abline(v=abs(diff_in_sample_means), col = "red")


# counts of values more extreme than the test statistic in our original sample, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= diff_in_sample_means)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= abs(diff_in_sample_means))
bootstrap_pvalue <- ( count_of_more_extreme_upper_tail)/num_sims
cat("Bootstrap p-value")
bootstrap_pvalue
## TWO SAMPLE TEST FOR DIFFERENCE IN PROPORTIONS
## Loading Packages
library(ggplot2)
library(cowplot)

## Reading Dataset
MRI_Data <- read.csv(file="oasis_cross-sectional.csv",header=TRUE)

## Rows 416 to 436 consists of MRI data for the same people already in the test.
## This was taken for testing reliability.
MRI_Data <- MRI_Data[c(1:416),]
MRI_Data <- MRI_Data[MRI_Data$eTIV<1818,]
head(MRI_Data)
```

```r
## Creating a Subset for two Groups "Male" and "Female"
MRI_Data_M <- subset(MRI_Data,MRI_Data$M.F== "M")
MRI_Data_F <- subset(MRI_Data,MRI_Data$M.F== "F")


MRI_Data_Male_WBV <- MRI_Data_M[, c('nWBV')]
MRI_Data_Female_WBV <- MRI_Data_F[, c('nWBV')]


MRI_Data_Male_eTIV <- MRI_Data_M[, c('eTIV')]
MRI_Data_Female_eTIV <- MRI_Data_F[, c('eTIV')]

MRI_Data_Male_WBV <-MRI_Data_Male_WBV * MRI_Data_Male_eTIV
MRI_Data_Female_WBV <-MRI_Data_Female_WBV * MRI_Data_Female_eTIV


# the parts of the test statistic
# sample props
p_hat_m <- sum(MRI_Data_Male_WBV )/sum(MRI_Data_Male_eTIV)
p_hat_f <- sum(MRI_Data_Female_WBV)/sum(MRI_Data_Female_eTIV)
# null hypothesized population prop difference between the two groups
p_0 <- 0
# sample size
n_m <- sum(MRI_Data_Male_eTIV)
n_f <- sum(MRI_Data_Female_eTIV)
# sample variances
den_p_m <- (p_hat_m*(1-p_hat_m))/n_m
den_p_f <- (p_hat_f*(1-p_hat_f))/n_f
# z-test test statistic
z <- (p_hat_m - p_hat_f - p_0)/sqrt(den_p_m + den_p_f)
# two sided p-value
two_sided_diff_prop_pval <- pnorm(q = z, lower.tail = FALSE)*2
two_sided_diff_prop_pval

# lower bound
(p_hat_m - p_hat_f)+(qnorm(0.025)*sqrt(den_p_m + den_p_f))
# upper bound
(p_hat_m - p_hat_f)+(qnorm(0.975)*sqrt(den_p_m + den_p_f))

## Bootstrap Approach

# Make the data
male <- rep(c(1, 0), c(sum(MRI_Data_Male_WBV), n_m -sum(MRI_Data_Male_WBV)))
female <- rep(c(1,0), c(sum(MRI_Data_Female_WBV),n_f - sum(MRI_Data_Female_WBV)))
num_sims <- 1000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
  prop_male<- mean(sample(male,
                          size = n_m,
                          replace = TRUE))
  prop_female <- mean(sample(x = female,
```

```
                                size = n_f,
                                replace = TRUE))
  results[i] <- prop_male - prop_female
}
# Finally plot the results
hist(results, freq = FALSE, main='Dist. of the Diff in Prop',
xlab = 'Difference in Prop. of Whole Brain Volume', ylab = 'Density')

cat("Bootstrap")
c(quantile(results, c(.025, .975)))
cat("Normal Approximation")
c((p_hat_m - p_hat_f)+(qnorm(0.025)*sqrt(den_p_m + den_p_f))
  ,(p_hat_m - p_hat_f)+(qnorm(0.975)*sqrt(den_p_m + den_p_f)))

str(male)
str(female)
# Make the data
df_combined <- data.frame("WBV" = c(male,female),
                          "Gender" = rep(c("male", "female"),
                                         c(n_m-1, n_f-1)))
# Sanity checks
summary(df_combined$Gender)

num_sims <- 1000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
  # idea here is if there is no relationshipm we should be able to shuffle the groups
  shuffled_groups <- transform(df_combined, Gender=sample(Gender))
  prop_male <- mean(shuffled_groups$WBV[shuffled_groups$Gender=="male"])
  prop_female <- mean(shuffled_groups$WBV[shuffled_groups$Gender=="female"])
  results_given_H0_true[i] <- prop_male - prop_female
}

# Finally plot the results
hist(results_given_H0_true, freq = FALSE,
     main='Dist. of the Diff in Sample Sample Props Under Null',
     xlab = 'Average Difference in Prop. of WBV  under Null',
     ylab = 'Density')
diff_in_sample_props <- p_hat_m - p_hat_f
abline(v=diff_in_sample_props, col = "blue")
abline(v=-diff_in_sample_props, col = "red")


# counts of values more extreme than the test statistic in our original sample, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= -diff_in_sample_props)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= diff_in_sample_props)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
cat("Bootstrap p-value")
bootstrap_pvalue
## CHI-SQUARE GOODNESS OF FIT TEST
```

```
## Loading Packages
library(ggplot2)
library(cowplot)

## Reading Dataset
MRI_Data <- read.csv(file="oasis_cross-sectional.csv",header=TRUE)

## Rows 416 to 436 consists of MRI data for the same people already in the test.
##This was taken for testing reliability.
MRI_Data <- MRI_Data[c(1:416),]
MRI_Data <- MRI_Data[MRI_Data$eTIV<1818,]
MRI_Data_Young <- MRI_Data[MRI_Data$Age<45,]
MRI_Data_Middle <- MRI_Data[MRI_Data$Age<66 & MRI_Data$Age>44 ,]
MRI_Data_Older <- MRI_Data[MRI_Data$Age>65,]
Young<-length(MRI_Data_Young$ID)
Middle<-length(MRI_Data_Middle$ID)
Older<-length(MRI_Data_Older$ID)
AgeTab<-rep(c("Young","Middle", "Older"),c(Young,Middle,Older))
table(AgeTab)
prop.table(table(AgeTab))


chisq<-sum(((table(AgeTab) - (407/3))^2)/(407/3))
pchisq(chisq, df = 3-1, lower.tail = FALSE)

# Create our data under the assumption that H_0 is true
solutions_under_H_0 <- rep(c("Young", "Middle", "Older"), (407/3))
# Sanity Check
table(solutions_under_H_0)

num_sims <- 10000
# A vector to store my results
chisq_stats_under_H0 <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
  new_samp <- sample(solutions_under_H_0, 407, replace = T)
  chisq_stats_under_H0[i] <- sum(((table(new_samp) - (407/3))^2)/(407/3))
}
hist(chisq_stats_under_H0, freq = FALSE,xlim=c(0,45),
     main='Dist. of the Chi-Square Statistic Under Null',
     xlab = 'Chi-Square Stat under Null',
     ylab = 'Density')
abline(v=sum(((table(AgeTab) - (407/3))^2)/(407/3)), col="red")

## Randomization p-value
sum(chisq_stats_under_H0 >= sum(((table(AgeTab) - (407/3))^2)/(407/3)))/num_sims
cat("\\newpage")
cat("\\newpage")
```