



Northeastern

Data Analysis Report

Date: 12/06/2019

BY
BALAJI KOTHANDARAMAN KALANIDHI

Table of Contents

ANALYSIS OF GALLUS BREED OF CHICKEN	2
Abstract.....	2
Objective.....	2
Exploratory Data Analysis.....	4
One Sample test.....	11
Bootstrapping method.....	15
One Sample Proportion test.....	16
Bootstrapping method.....	18
Two sample t-test for difference in means	19
Bootstrapping method.....	22
Two sample difference in proportion test.....	24
Bootstrapping method.....	25
Chi-square test for goodness of fit.....	28
Discussion.....	29
Appendix.....	30
DATA.....	42

ANALYSIS OF GALLUS BREED OF CHICKEN

Northeastern University, IE 6200 - Engineering Probability and Statistics, November 2019

ABSTRACT:

BACKGROUND:

The Gallus Breed is a tropical bird in the family Phasianidae. They are the primary progenitor of the domestic chicken. The red Gallus Chicken had first domesticated at least 5,000 years ago in India. Since then, their domestic form has spread around the world and is kept globally as a very productive food source of both meat and eggs. The range of the Gallus stretches from India, eastwards across Indochina and southern China and into Malaysia, Singapore, the Philippines, and Indonesia. Gallus Breed was one of three main animals carried by early Austronesian peoples from Island Southeast Asia in their voyages to the islands of Oceania in prehistory, starting at around 5,000 BP. Today their ancient descendants are found throughout Micronesia, Melanesia, and Polynesia.

OBJECTIVE:

The objective of this study is to analyze how the Gallus chicken breed differs from two major categories of their breed (Marans and Ameraucana). Various Analysis is conducted on these breeds such as how weights of these chicken breeds differ from one and another, how color egg differs from each breed, how Gallus Egg weight differs, number of hours Gallus chicken is exposed to natural sunlight.

SIGNIFICANCE OF STUDY:

The Study on Gallus breed helps different groups of people. The Primary group is the people in poultry, from this study they can understand how much food is needed for the chicken to be healthy and produce enough eggs per day. It also helps to identify negative effects on chicken and helps them to take the necessary steps to overcome it. The secondary group that is benefited from the study is the owner of the poultry which helps them to take necessary steps to get maximum profit from it.

SCOPE OF THE STUDY:

The scope of the study is limited to the Gallus breed of chicken in Africa. The data is consisting of two main breeds of Gallus chicken (Marans and Ameraucana) and the analysis is limited to these two breeds of the Gallus chickens. For this study various variable such as Gallus breed, Gallus Egg Weight, Gallus Weight, Amount of Feed, Eggs Per Day, Sun Light Exposure were considered.

METHOD OF STUDY:

SOURCE OF DATA:

The dataset consists of data created artificially based partially on subject matter knowledge of Poultry and references derived from various websites. It lists observations made on certain days relating to physical chicken attributes and information relating to the number of eggs laid, the amount of food consumed, and the amount of time exposed to natural light. The dataset name “GallusGallusDomesticus “is the name by which chickens are scientifically known.

SAMPLE SELECTION:

The respondent of Gallus chicken was from Africa. Simple random sampling was followed where each sample was taken for various days to collect data. Controls were used to remove duplicates. By taking a sample on different days it was made sure that each sample was independent of each other.

STATISTICAL METHODS:

A traditional statistical and bootstrapping methods were used to tabulate the study. In the first on sample t-test is used to find the average weight of Gallus chicken, for next question the one-sample proportion test were used to find egg color in Gallus chicken, In the two-sample t-test we are comparing the mean Amount of feed Gallus breed which two categories Marans and Ameraucana, In two-sample proportion test, what are the common beak color of Gallus chicken is analyzed and each of this method was followed by Bootstrapping statistical method. The chi-square method was used to find whether the proportion Gallus Leg Shanks Color is equal or not.

LIMITATION OF THE STUDY:

The study might be limited using a questionnaire as a data collection instrument. It might be a limited place where the data is collected and might provide significant results in that area. Care should be taken before generalizing it to the entire population.

INTRODUCTION TO VARIABLES:

A various variable was used for analyzing the study, the type variable used, and their measures are described below.

Gallus ID- Categorical variable- Nominal (It is used to verify that the samples are independent).

Gallus Breed- Categorical variable- Ordinal (It classifies into two main categories such as Marans and Ameraucana).

Day- Quantitative variable- Discrete (It gives the measures of individual measurement made in a day).

Age- Quantitative variable- Discrete (It gives the age of chicken in days).

Gallus Weight- Quantitative variable- Continuous (It gives the weight of chicken in grams).

Gallus Egg Color- Categorical variable- Nominal (It has four different colors of eggs).

Gallus Egg Weight- Quantitative variable- Continuous (It gives the weight of Gallus chicken in grams).

Amount of Feed- Quantitative variable- Discrete (It gives grams of food given to each Gallus).

Eggs Per Day- Quantitative variable- Discrete (It gives the number of eggs laid by chicken per day).

Sun Light Exposure – Quantitative variable -Continuous (It gives the number of hours Gallus is exposed to sunlight).

Gallus Leg Shanks Color- Categorical variable -Nominal (It has four different colors of legs).

Gallus Beak Color- Categorical variable - Ordinal (It has two different colors of the beak).

Gallus Plumage – Categorical variable – Nominal (It ten different colors of feathers).

EXPLORATORY DATA ANALYSIS:

This section focuses on trends observed using descriptive statistics. The summary of the findings is summarized into each graph and they are listed below.

Figure-1 represents how weight (grams) of the Gallus breed is distributed by the age of the Gallus in days. The histogram shows that the weight of chicken is not dependent on weight, so it is distributed randomly. The Y-axis consists of Gallus weight in grams and X-axis consists of Gallus age in days. The histogram is plotted for overall data points of 950. It is seen clearly from the histogram that the maximum weight of Gallus are around 16000 grams and the lowest being 1200 grams.

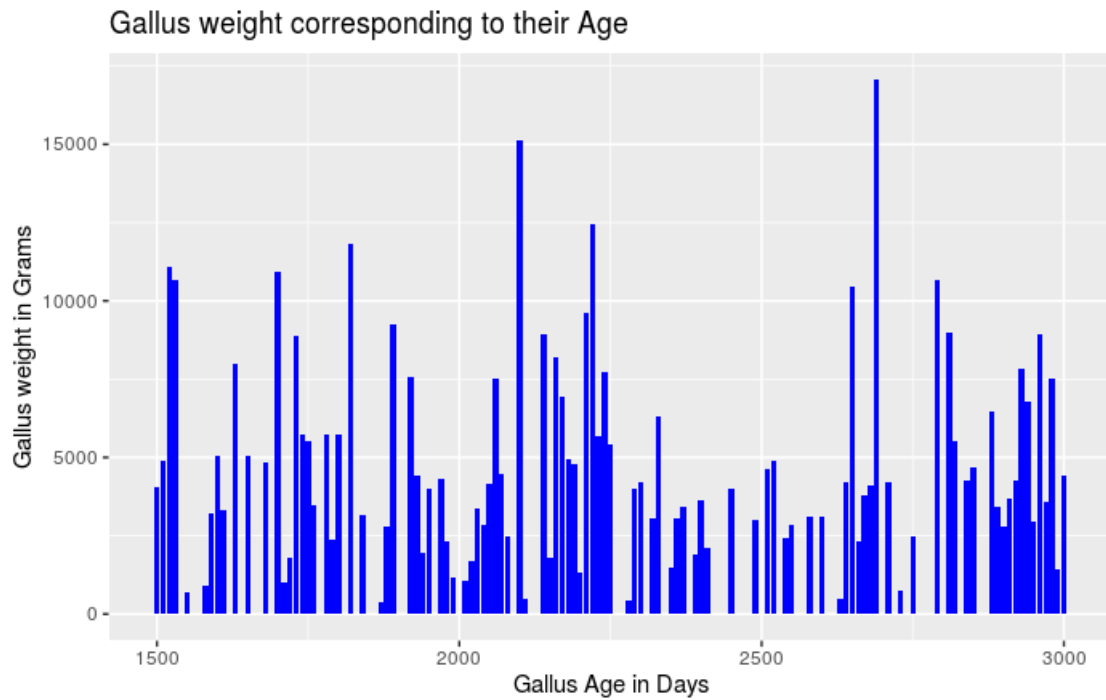
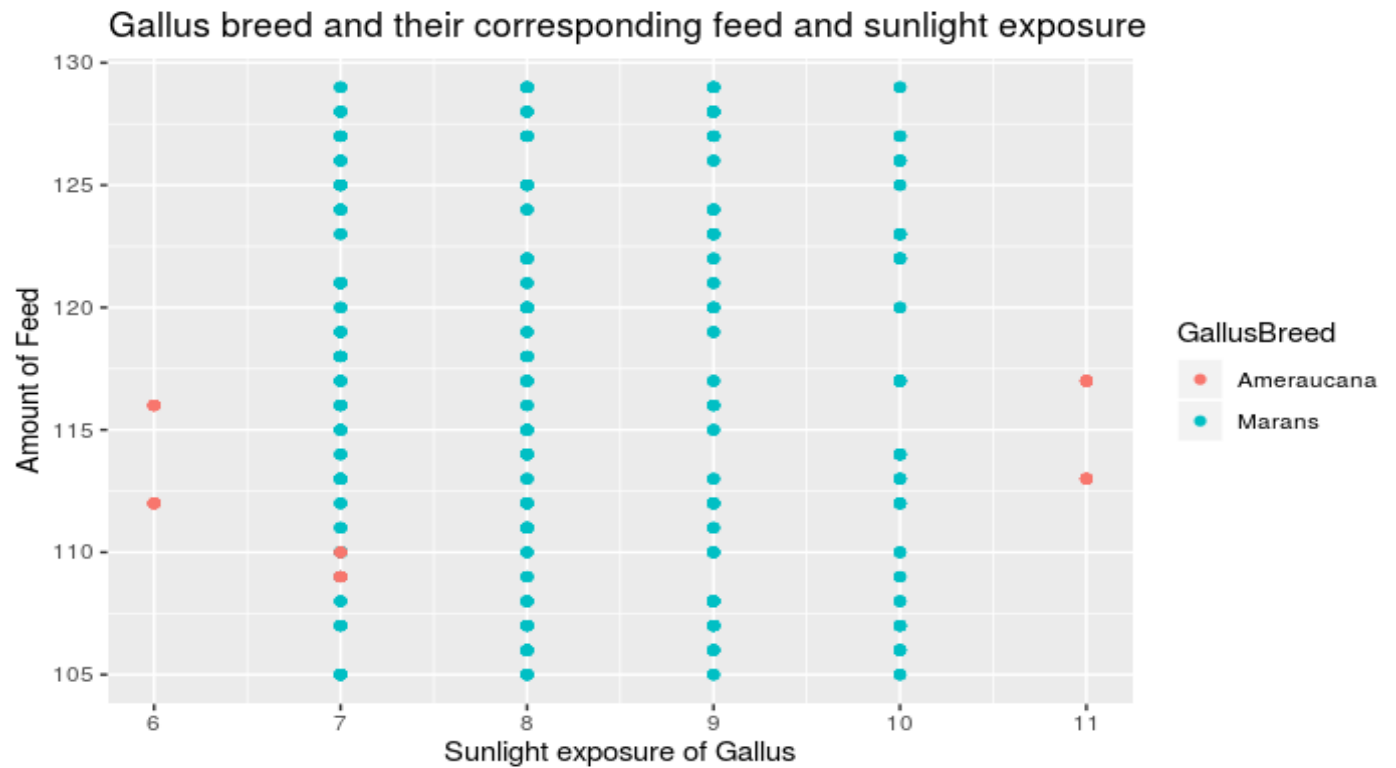
FIGURE - 1**FIGURE - 2**

Figure -2 shows the two main breeds of Gallus (Ameraucana and Marans) and to how much sunlight they are exposed, and it is measured in hours, while the y-axis shows how the amount of food is feed to the Gallus. From the scatter plot you can interpret the Marans are fed higher and exposed to sunlight in higher ratio when it is compared to Ameraucana.

FIGURE -3

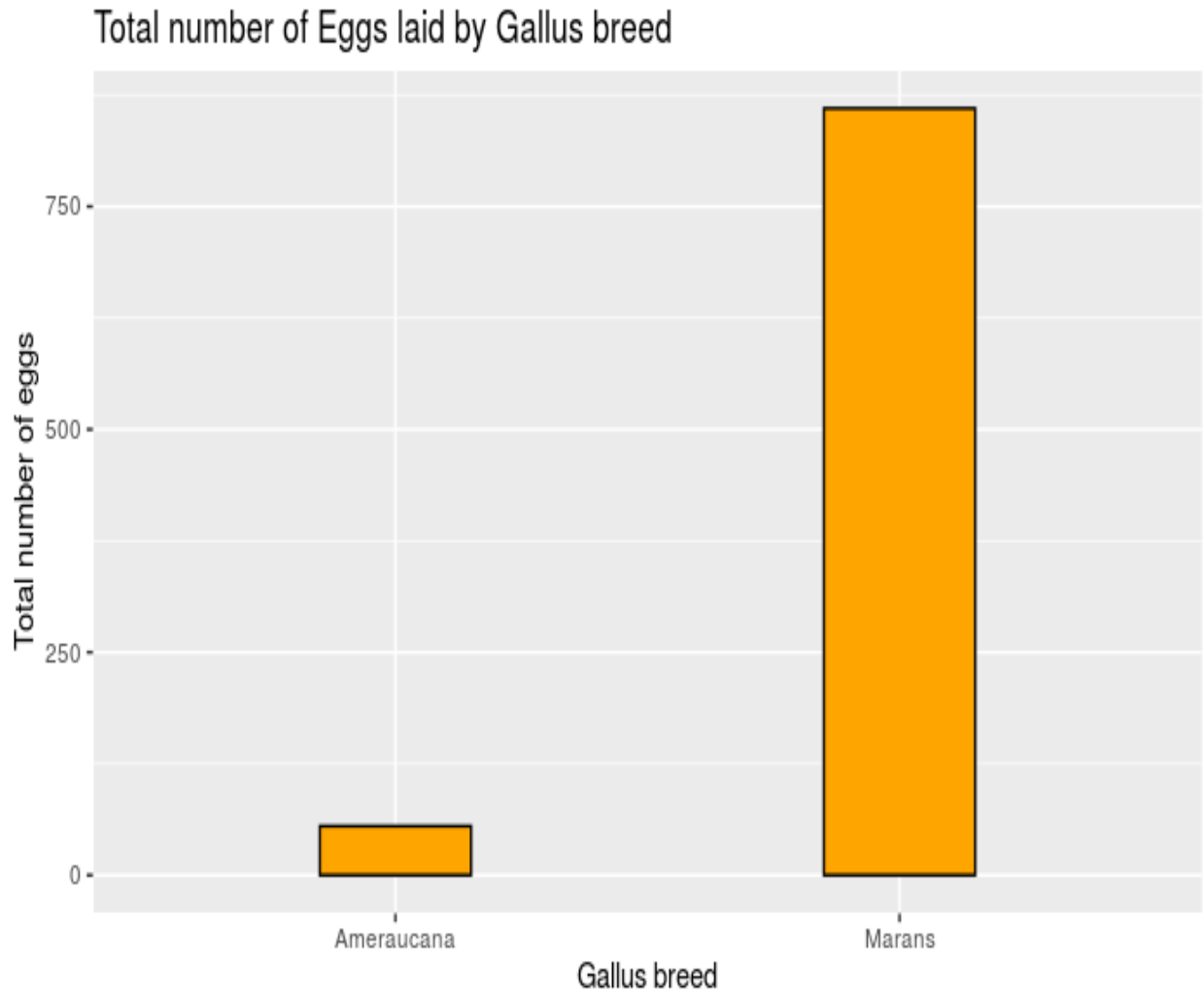


Figure -3 shows the total number of eggs laid by the Gallus chicken. Ameraucana the breed of Gallus chicken has 50 eggs for the sample size of 55, while the Marans breed has a higher number of eggs laid, which is around 800 for the sample size of 895.

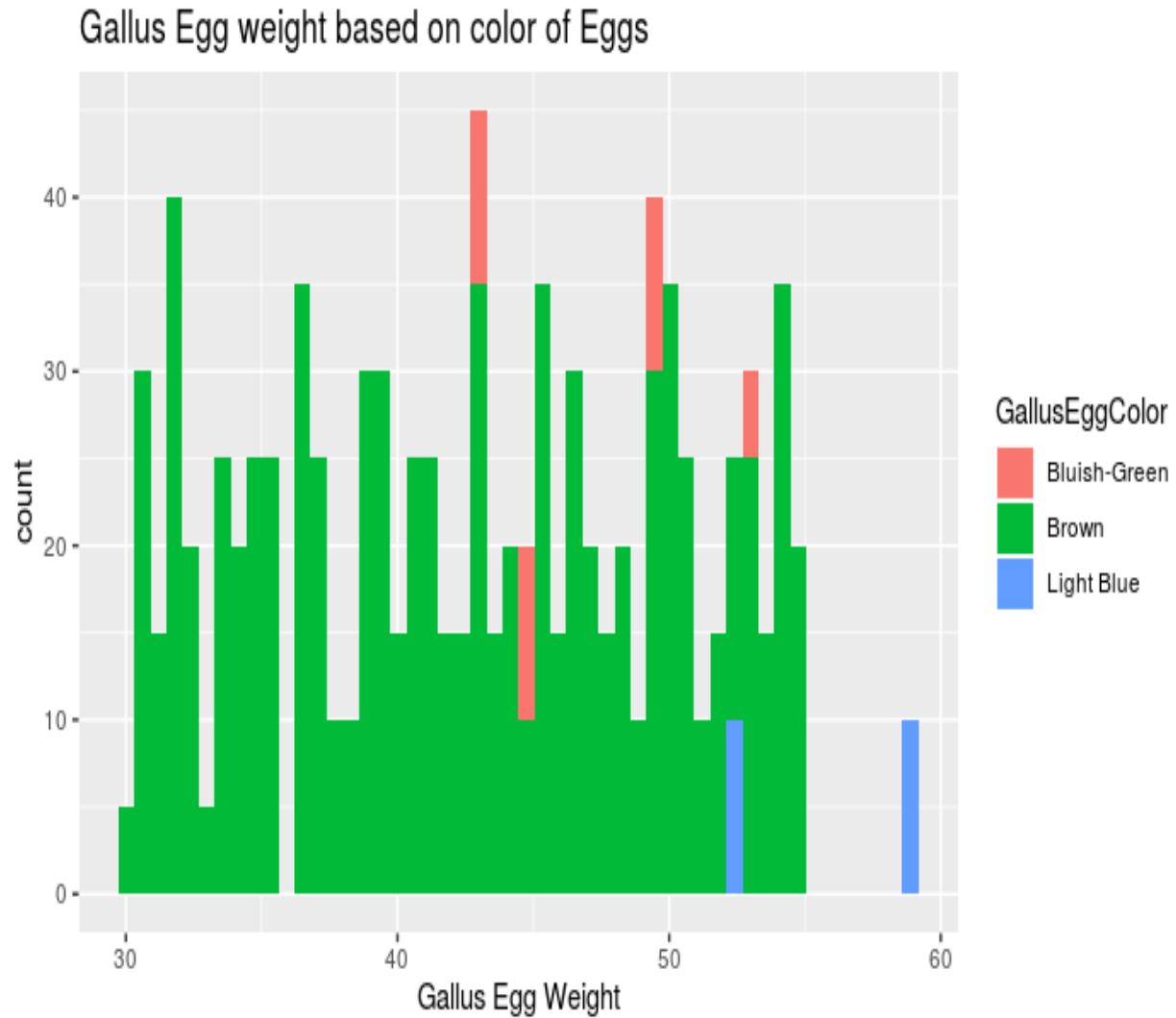
FIGURE - 4

Figure – 4 the Gallus egg color is classified into three types Bluish-green, Brown, Light Blue. These of from Marans and Ameraucana and each egg weight is measured in grams. The x-axis gives the weight of each egg in grams while the y-axis gives the number of eggs based on weight, a legend in the graph helps us to identify the various colors of eggs. From the histogram, the majority of the eggs are brown in color.

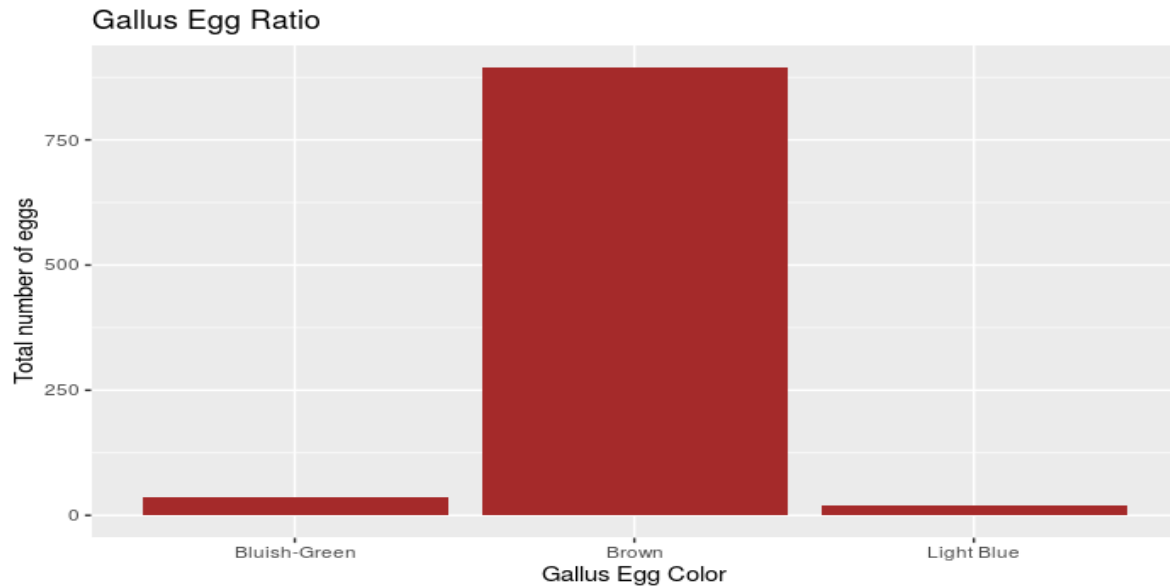
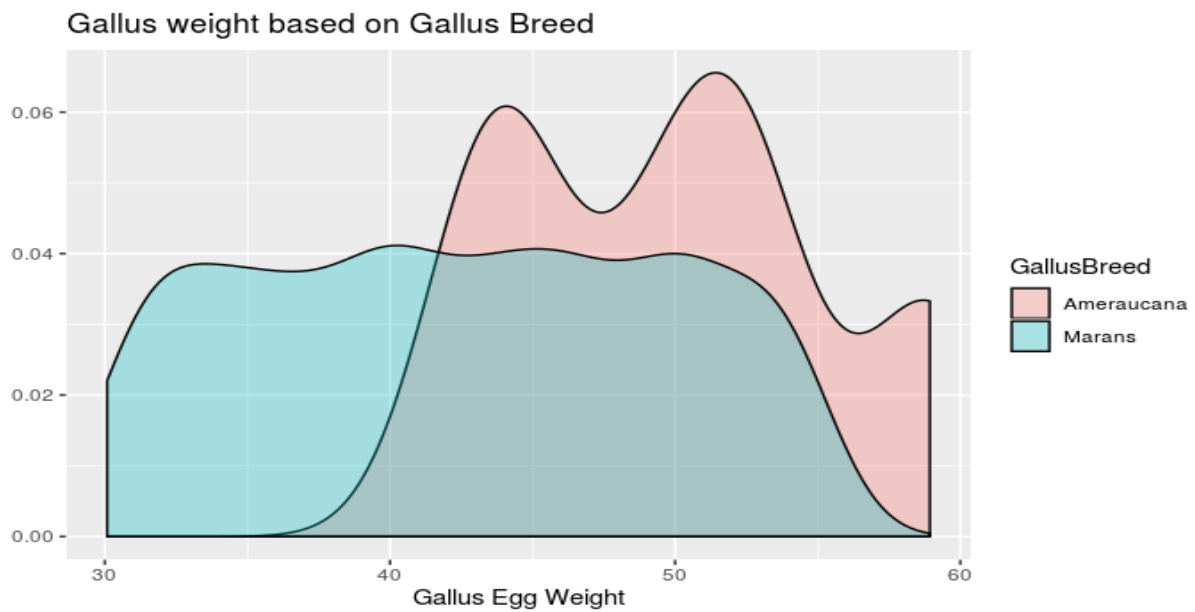
FIGURE - 5

Figure -5 gives a clearer picture of the ratio of eggs distributed among Marans and Ameraucana Breed.

FIGURE -6

The density plot (Figure -6) shows how the egg weight is distributed over two major breeds of Marans and Ameraucana. The legend helps to identify the two different breeds.

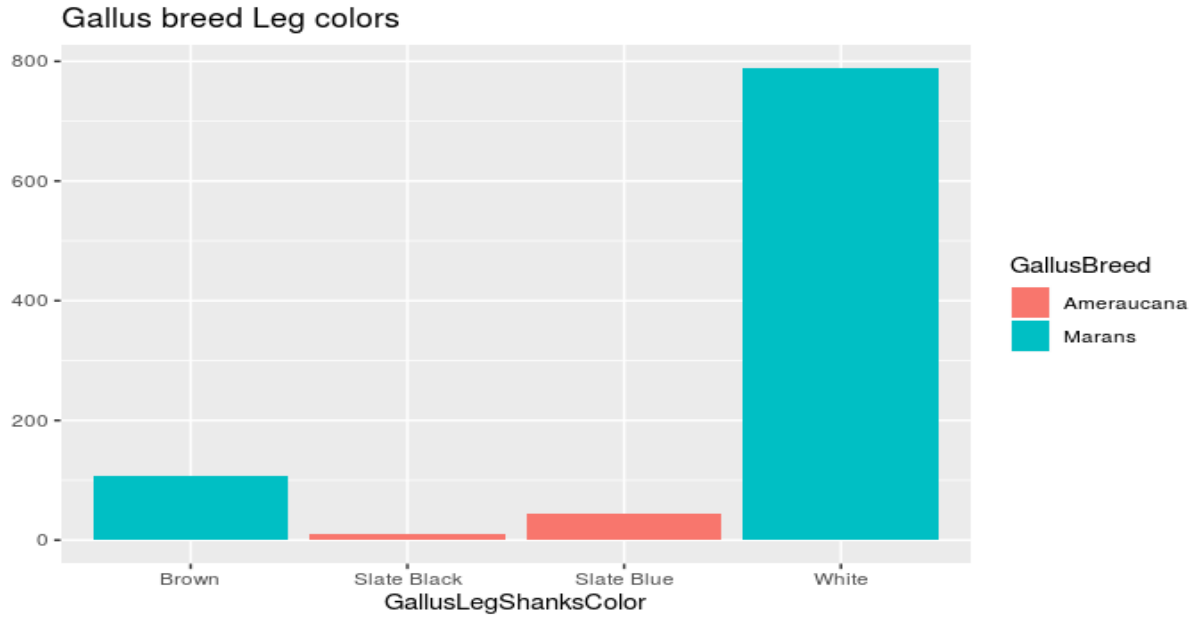


Figure- 7

This bar graph helps us to understand how the leg color of the Gallus breed is distributed. It has a major category of Marans with white color legs.

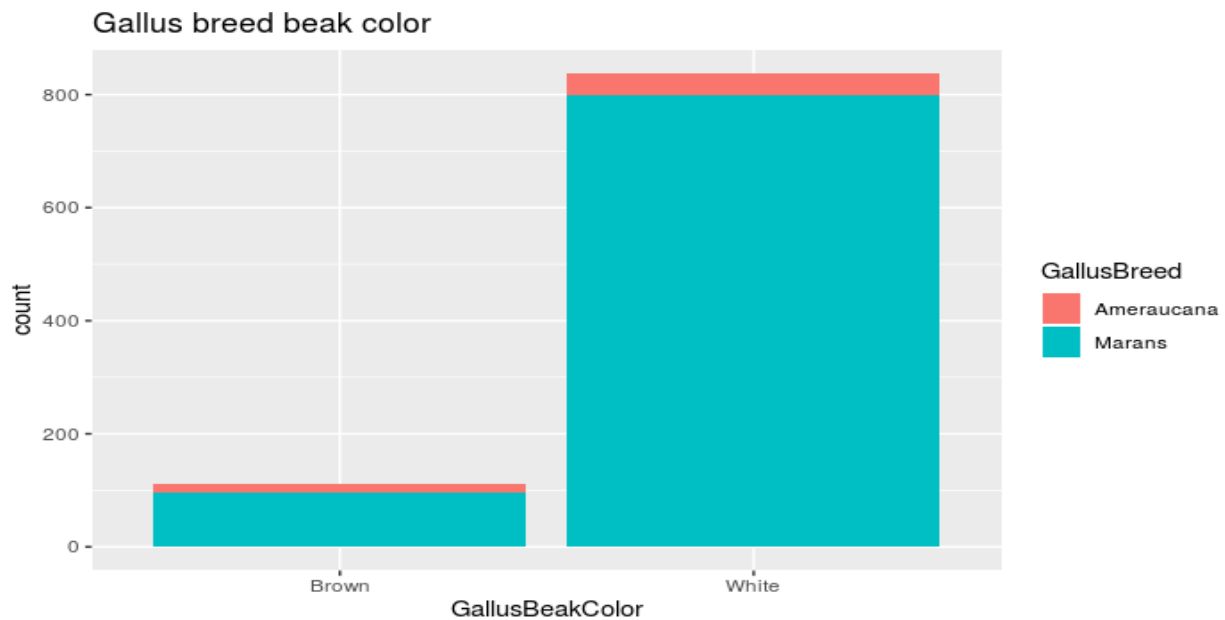


Figure-8

This bar graph helps us to understand how the Gallus beak color is distributed over Marans and Ameraucana.

FIGURE -9

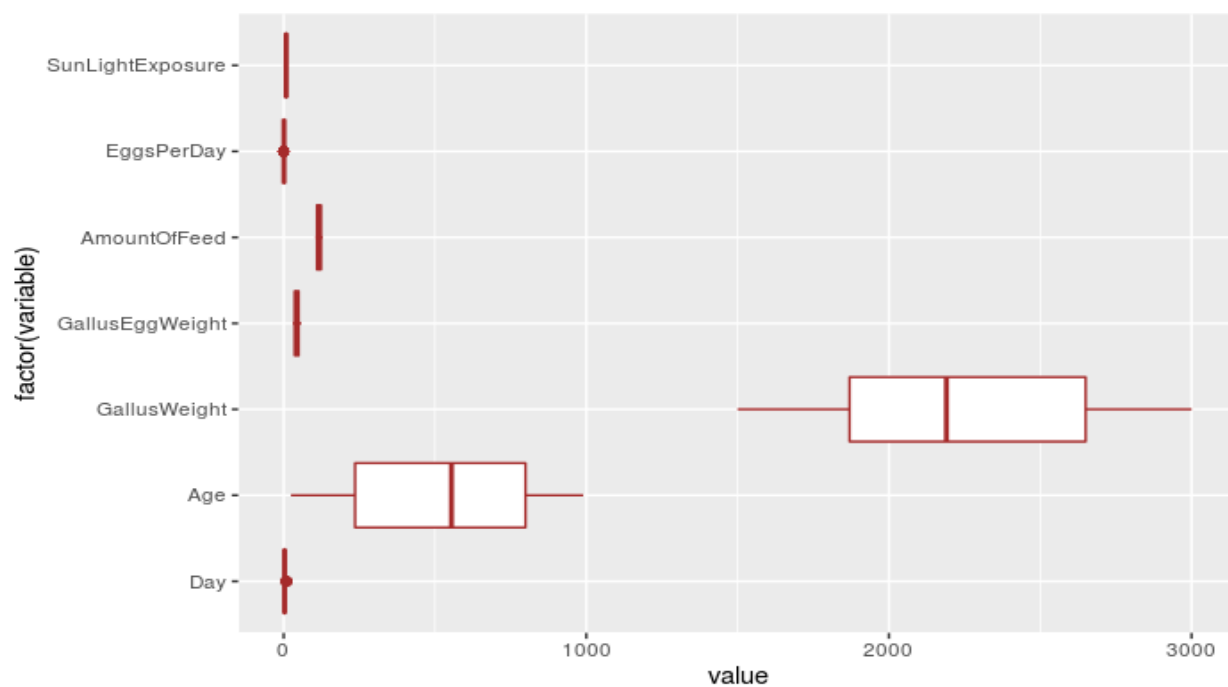
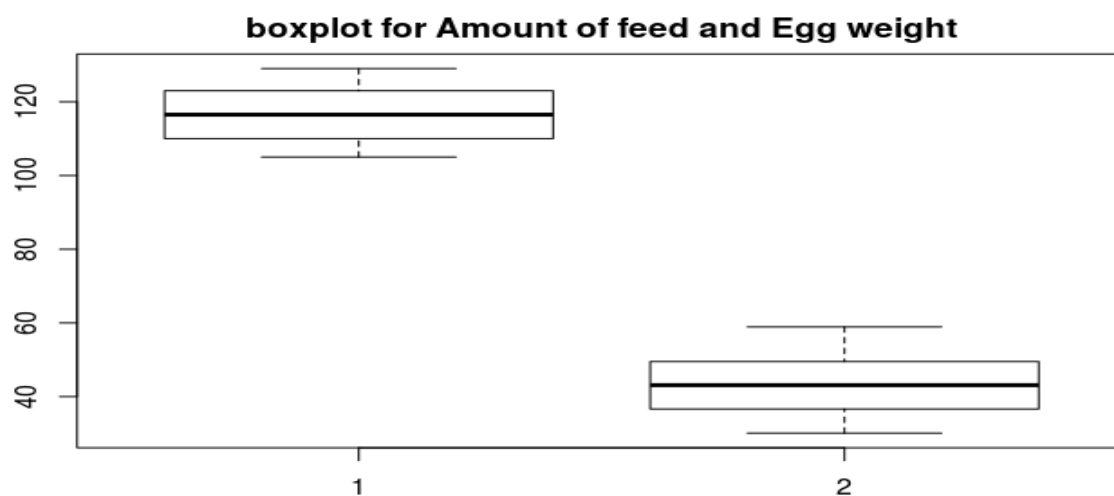


FIGURE- 10



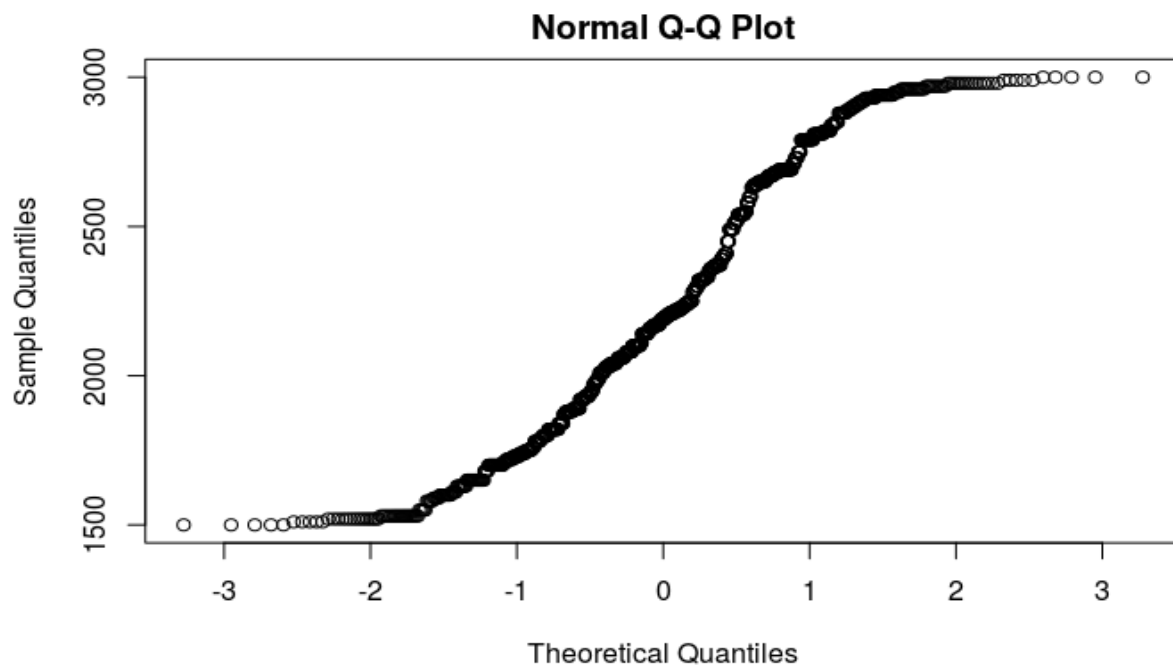
The Box plot (Figure -7 and Figure 8) helps us to identify the outliers across the quantitative variables in the dataset. From the plot we can interpret there are no outliers in each variable, this will help us to make analysis and interpret in a significant manner.

STATISTICAL ANALYSIS:

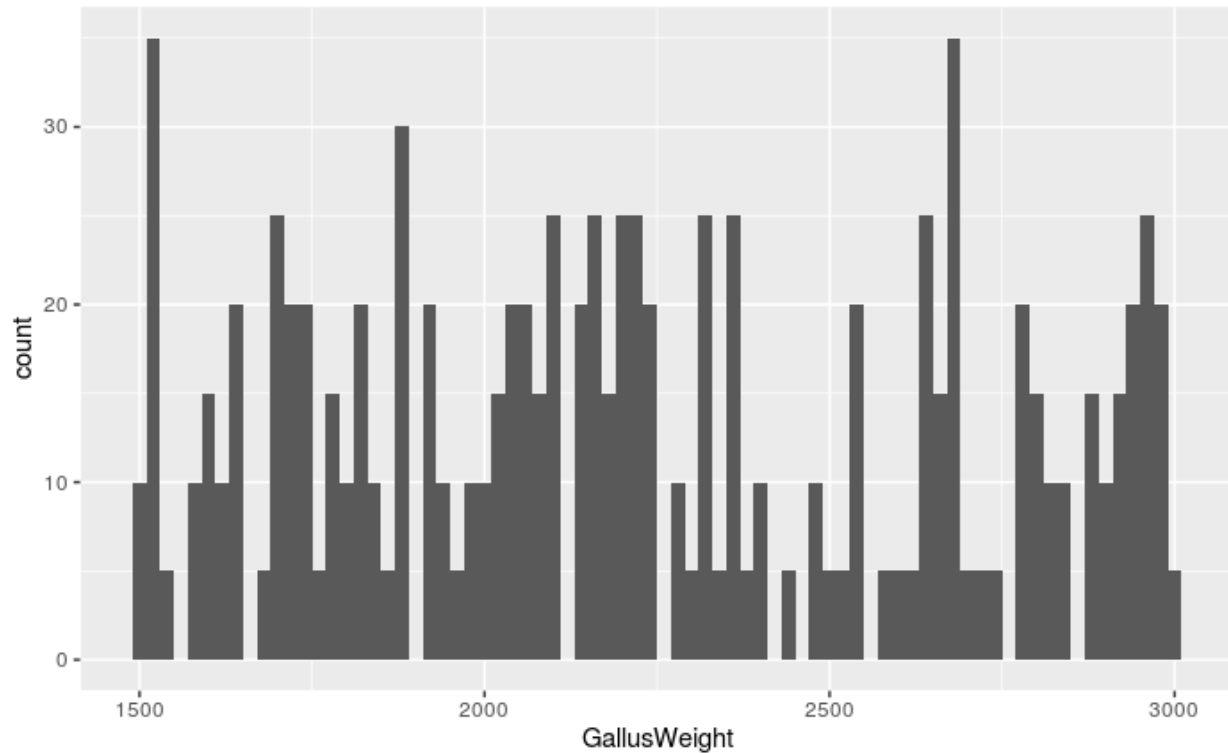
One Sample T-test:

A study proposed that the mean Gallus weight is 2200 grams, we are conducting an one-sample t-test to check whether the true mean weight is 2200.

- To conduct a single sample t-test, We have to satisfy the following conditions.
- The sample data should from a single population- Gallus Breed(single population).
- One quantitative variable of interest- Gallus Weight(Quantitative and Continous).
- We want to make inference about population- We are making inference about the population of Gallus Breed.
- The sample size should be greater than 30- sample size(950).
- The data must be normally distributed- (Should be checked using qqnorm)



The qqnorm plots help us to find out whether the data is normally distributed.



The histogram helps us to identify how the Gallus Weight data is distributed

The histogram proves that the data is not normally distributed. The points are spreading all over the axis. The normality assumption is not met and we should be concerned about p-value and confidence interval.

Population Parameter:

The population parameter we want to make inference is μ .

Hypothesis Testing:

We are going to use Two sided T-test.

The True mean weight of Gallus Chicken is 2200.

$$H_0 : \mu_0 = 2200$$

The True mean weight of Gallus Chicken is different from 2200.

$$H_0 : \mu_A \neq 2200$$

Sample Statistic:

The sample statistic is the sample mean \bar{x}

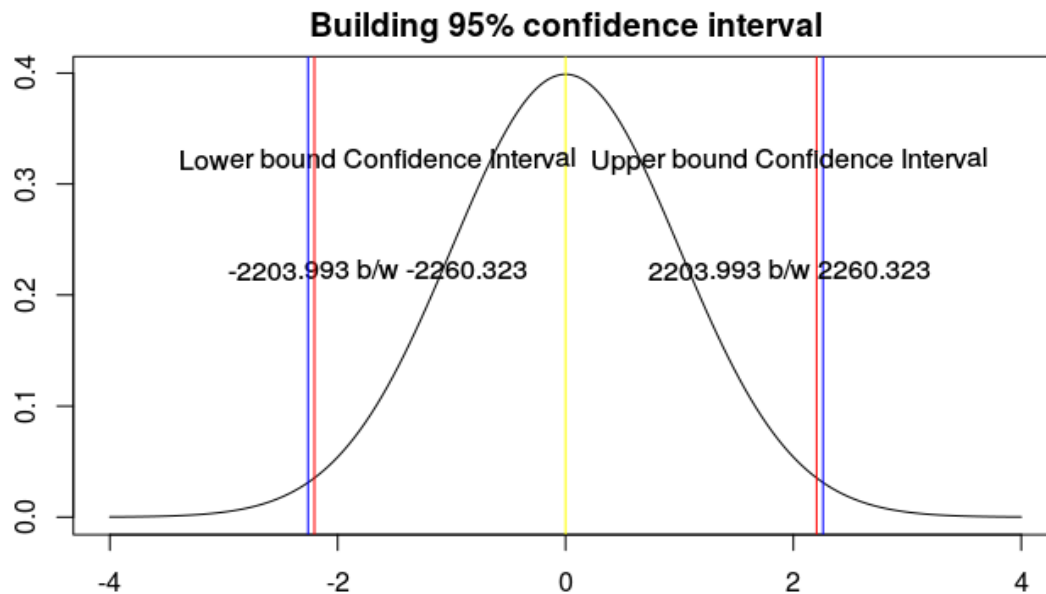
Test Statistic:

We don't know the population variance of the Gallus breed population, so we use sample statistics to estimate it and hence, the sample follows T-Distribution. As the t- distribution reaches infinity it becomes normally distributed.

Built-in T-Test:

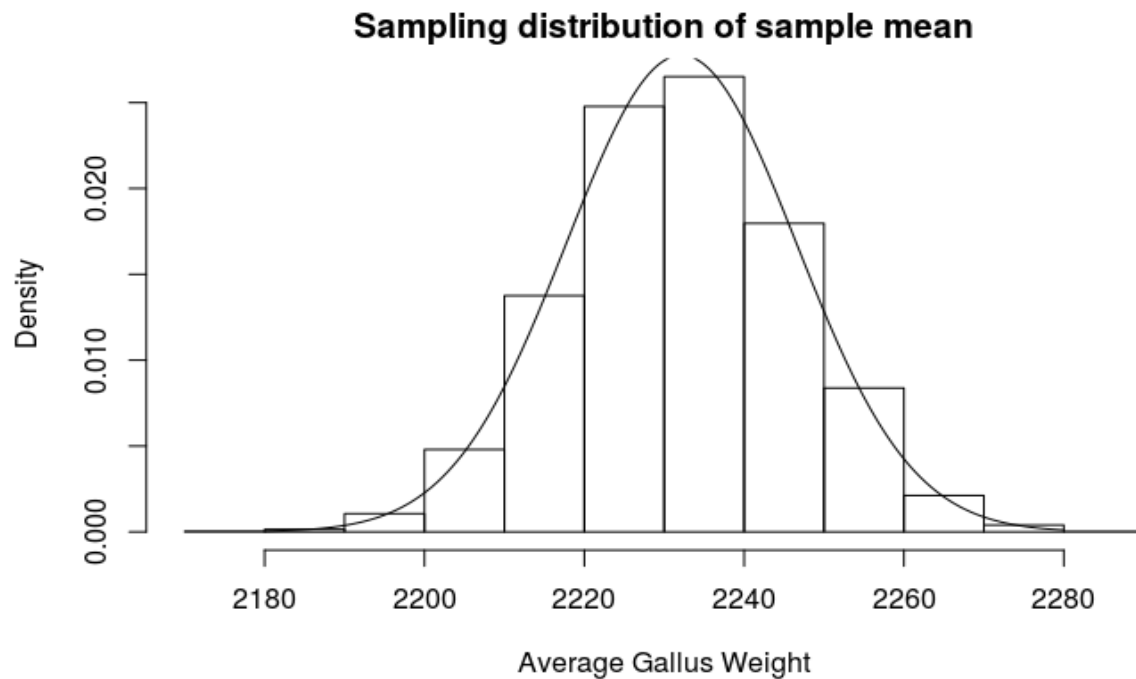
```
One Sample t-test
data: Gallus_weight
t = 2.2407, df = 949, p-value = 0.02528
alternative hypothesis: true mean is not equal to 2200
95 percent confidence interval:
 2203.993 2260.323
sample estimates:
mean of x
 2232.158
```

Confidence Interval for One-Sample T-Test:

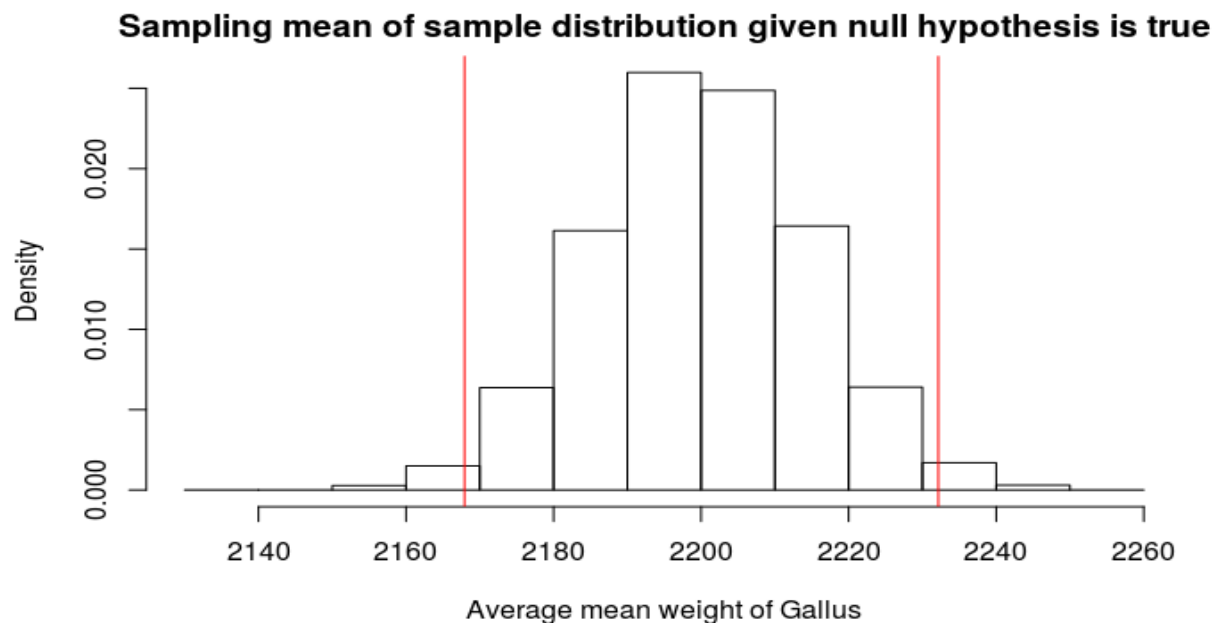


Interpretation:

The One-Sample Confidence Interval shows that we 95% confident the mean weight of Gallus weight is between 2203.993 and 2260.323.

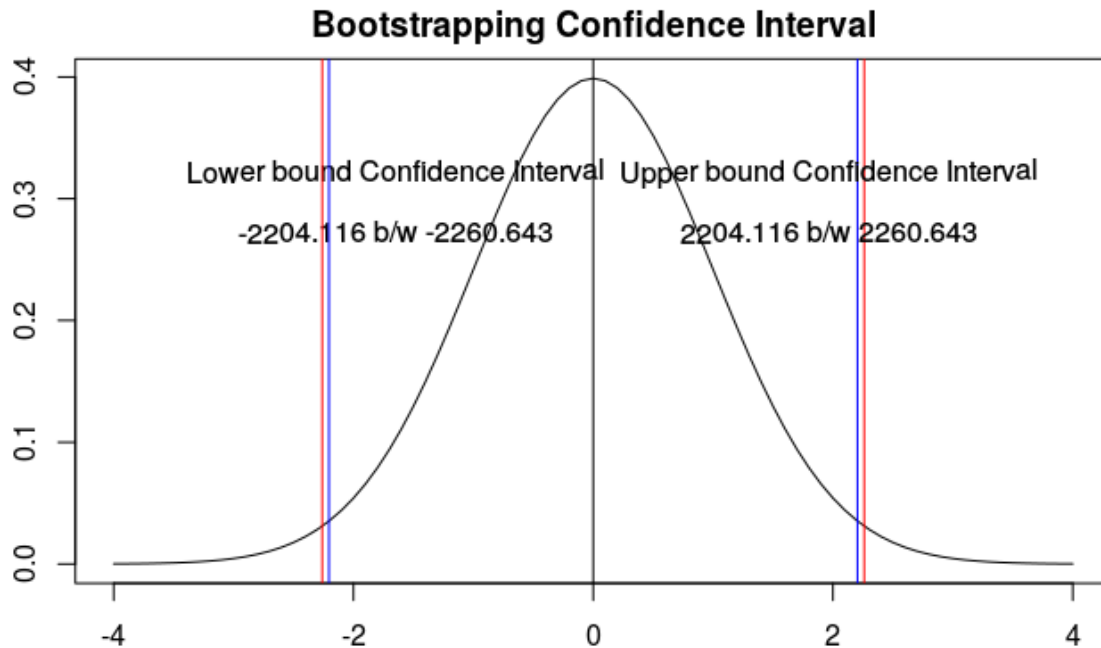
Bootstrapping Method:

The distribution shows how data are sampled using a bootstrapped approach to get normality on the sampled data. The data is sampled for mean Gallus weight in grams.



The data is sampled 10000 using a bootstrapped approach where the null hypothesis is true. The null hypothesis states the mean Gallus weight is 2200 grams.

Bootstrapping Confidence interval:



Interpretation:

There is strong evidence ($p\text{-value} = 0.02527703$) to suggest that the true mean weight of Gallus Chicken is different from 2200 grams. We reject the null hypothesis that the true mean weight of Gallus Chicken is 2200 at $\alpha = 0.05$ level. With 95 %confidence, the true mean for the weight of Gallus Chicken is between 2204.116 and 2260.643 grams, which suggests that the true mean is greater than 2200 grams. The interval using the empirical methods, especially the quantile method is much compact which agrees with our $p\text{-value}$ being a bit conservative and is a bit significant when compared to traditional $t\text{-tools}$.

ONE SAMPLE PROPORTION TEST:

The egg color of Gallus chicken is Brown, Light-Blue, and Bluish-green, I want to find the proportion of Brown colored egg compared to other colors of eggs.

Conditions for doing One-Sample Proportion test:

- The categorical variable of interest with two categories:
- We have categories of Brown, Light-Blue, and Bluish-green so we recode it into Light Blue or other.
- A sample comes from a single population:
- The sample comes from a single population of Gallus breed chicken.

Conditions to Use:

- Exact Binomial test- no requirements
- Normal approximation: $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$

Population Parameter:

The population parameter we want to make an inference is the True population proportion of brown colored eggs.

Hypothesis:

For hypothesis let me consider that brown colored eggs are 95% of true population.

Condition Check:

$$\begin{aligned}
 n\hat{p} &\geq 10 \\
 950(895/950) &= 895 \\
 n(1 - \hat{p}) &\geq 10 \\
 950(55/950) &= 55
 \end{aligned}$$

Hypothesis testing:

We are going to use one sided proportion-test.

Null hypothesis

The True Proportion of Brown egg color of Gallus Chicken is 95 percent.

$$H_0 : \mu_0 = 0.95$$

Alternate hypothesis

The True Proportion of Brown egg color of Gallus Chicken is greater than 95 percent.

$$H_0 : \mu_A > 0.95$$

Sample Statistic:

$$\text{The Sample Statistic is } \hat{p} = \frac{895}{950} = 0.9421$$

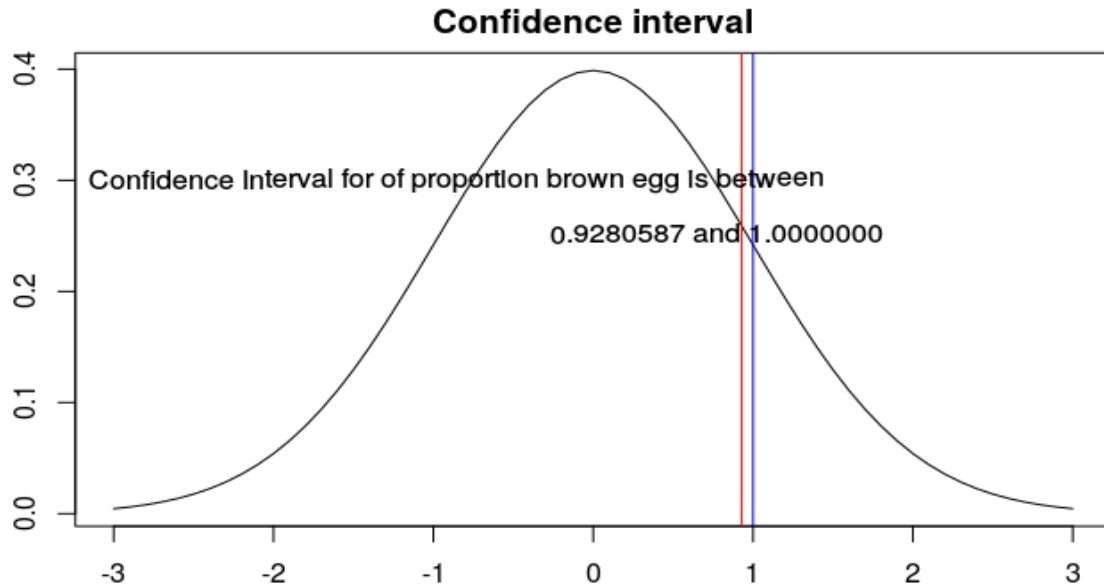
Built-in Proportion Test:

```

data: 895 out of 950
number of successes = 895, number of trials = 950, p-value = 0.8816
alternative hypothesis: true probability of success is greater than 0.95
95 percent confidence interval:
 0.9280587 1.0000000
sample estimates:
probability of success
 0.9421053

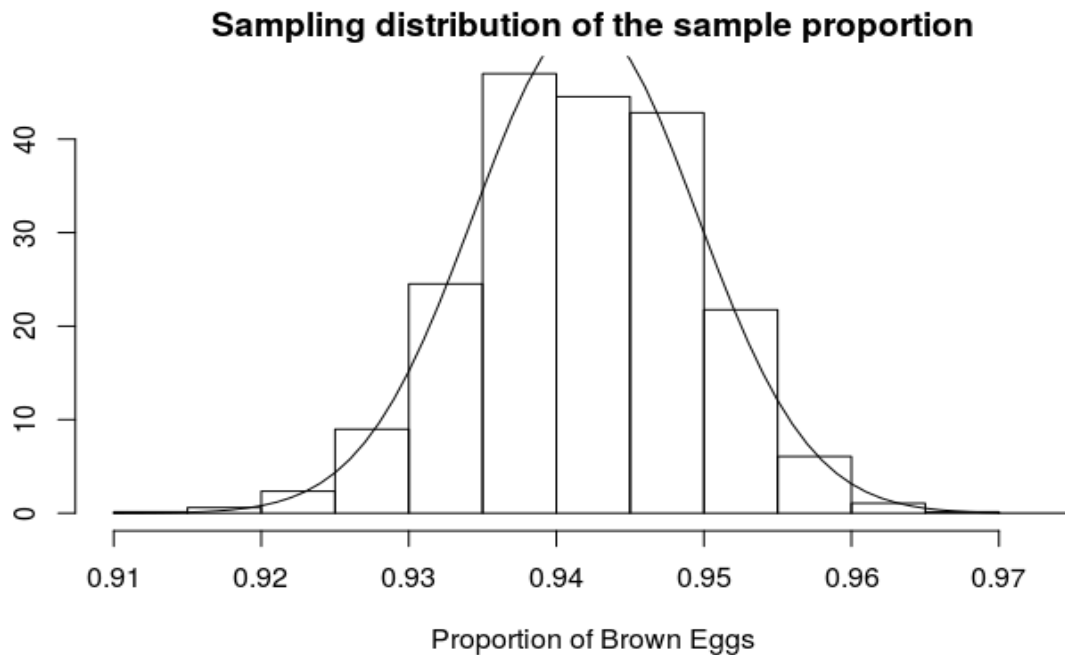
```

From the built-in test, we can interpret that the p-value for the proportion test is 0.8816 and we are 95% confident that the proportion of Brown color egg is between 0.9280587 and 1.

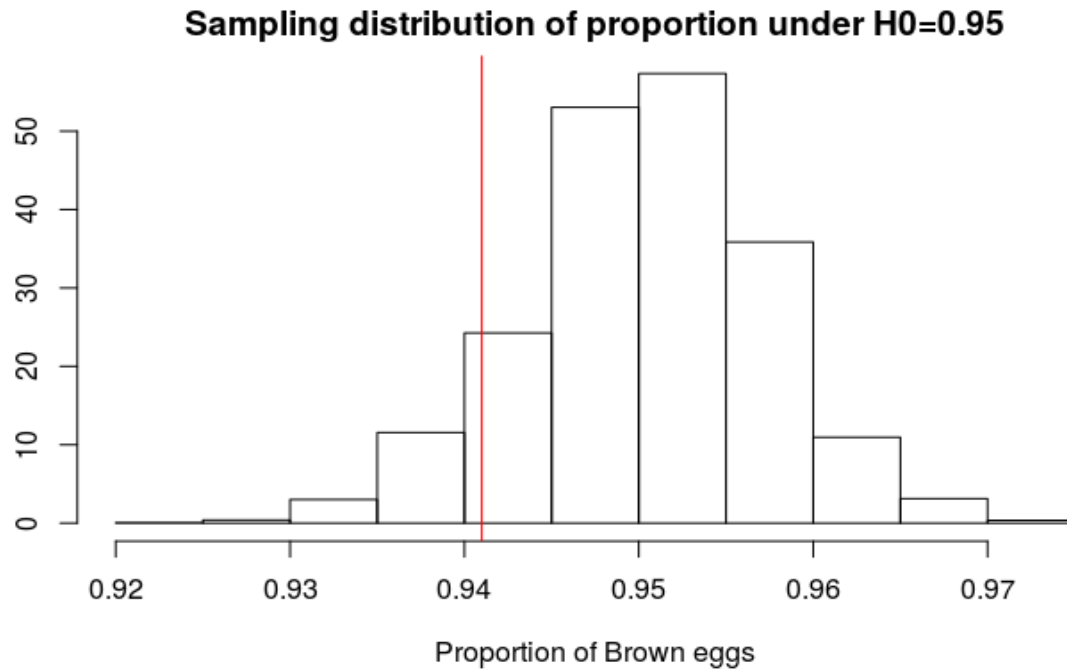


A 95% Confidence interval is drawn for the proportion of brown colored eggs in the Gallus breed, which lies between 0.9280587 and 1.

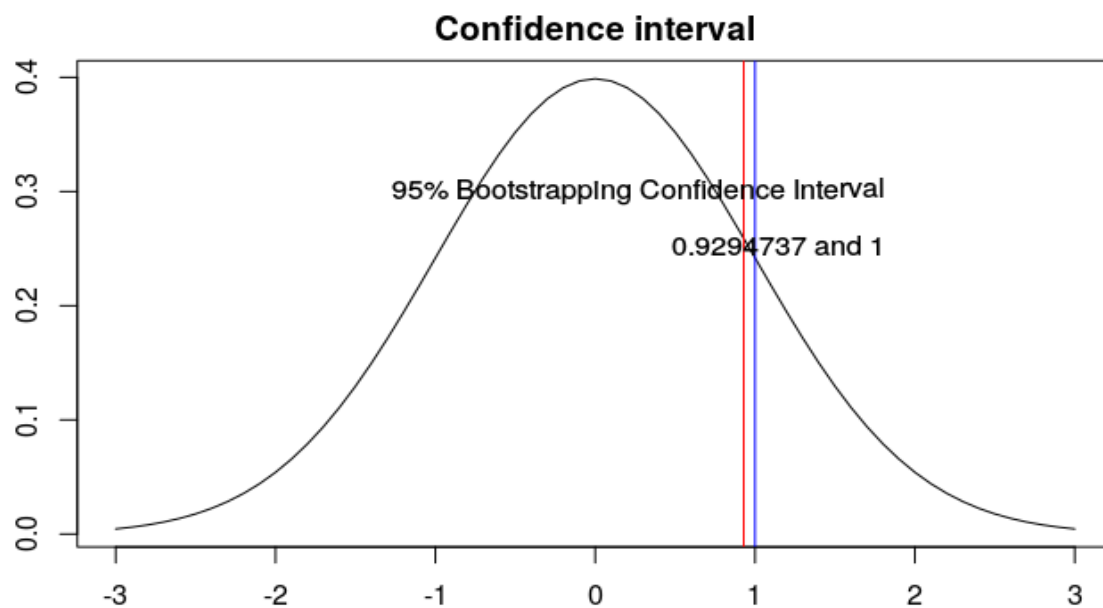
Bootstrapping Method for One-Proportion test:



The distribution shows how data are sampled using a bootstrapped approach to get normality on the sampled data. The data is sampled for the proportion of Gallus brown color egg.



The data is sampled 10000 using a bootstrapped approach where the null hypothesis is true. The null hypothesis states the proportion of brown color Gallus egg is 95 percent.



The Bootstrap Confidence interval shows the proportion of brown eggs in the Gallus breed is between 0.9294737 and 1.

Interpretation:

There is strong evidence (p-value approx. equal to 0.8678926) to suggest the true proportion of brown eggs is not greater than 95%. We cannot reject the null hypothesis that the true proportion of brown egg in the Gallus breed is equal to 95% at the $\alpha = 0.05$. The true proportion of brown eggs in the Gallus breed is between 0.9280587 and 1 at a 95% confidence interval. The interval using the empirical methods, especially the quantile method is much compact which agrees with our p-value being a bit accurate and is bit significant when compared to traditional t-tools.

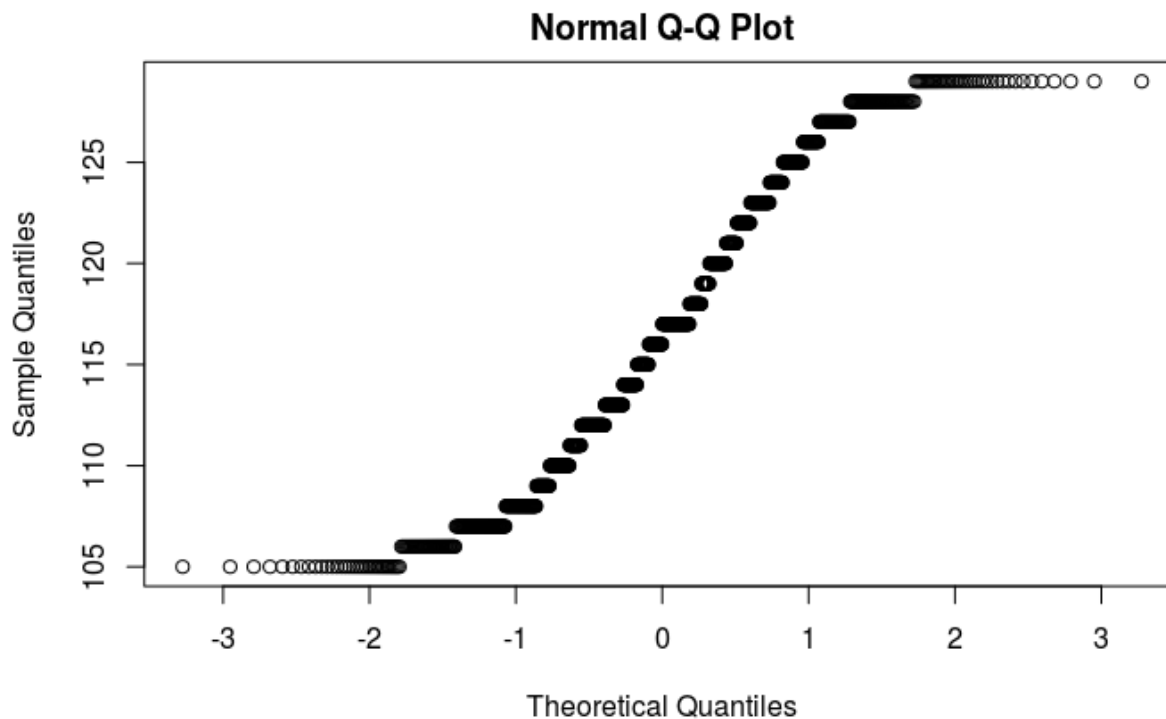
TWO SAMPLE TESTS FOR DIFFERENCE IN MEANS.

A two-sample T-test is used because data were collected from two different populations. Two Sample T-test has been used to find the difference in mean between two Population. We use a t-test because we don't know the Variance in the population.

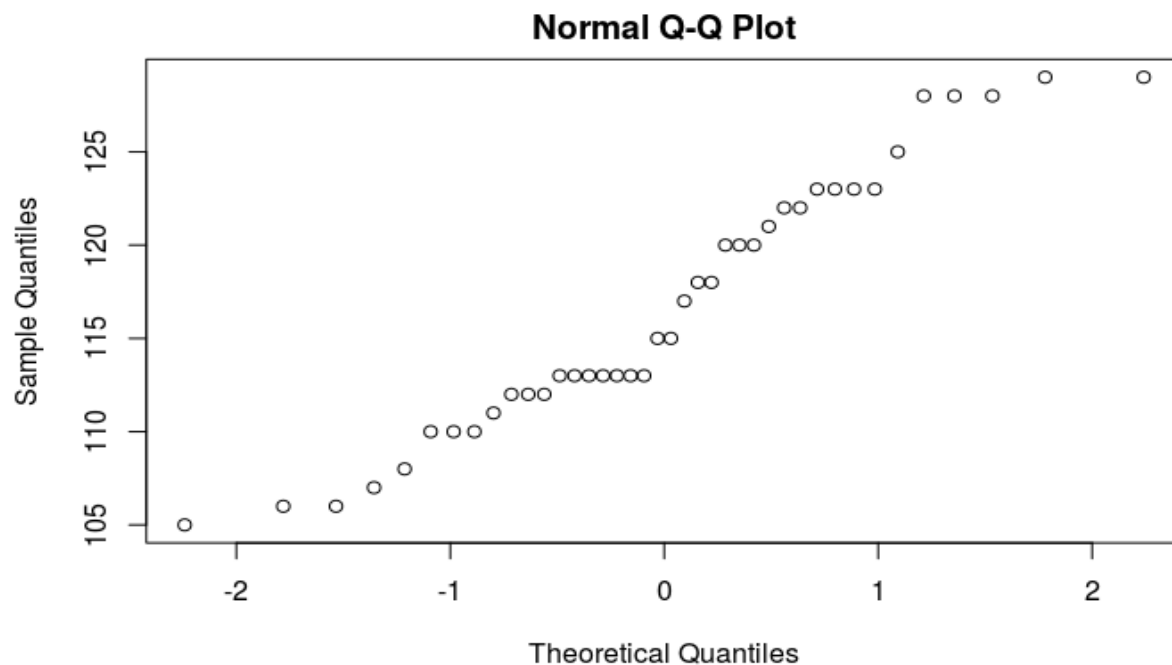
Conditions to do a two-sample t-test.

- The question of interest must be the difference in population.
- 2 independent samples and should belong to a different population.
- The population must be normally distributed.
- $N < 30$, it should greater than 30.

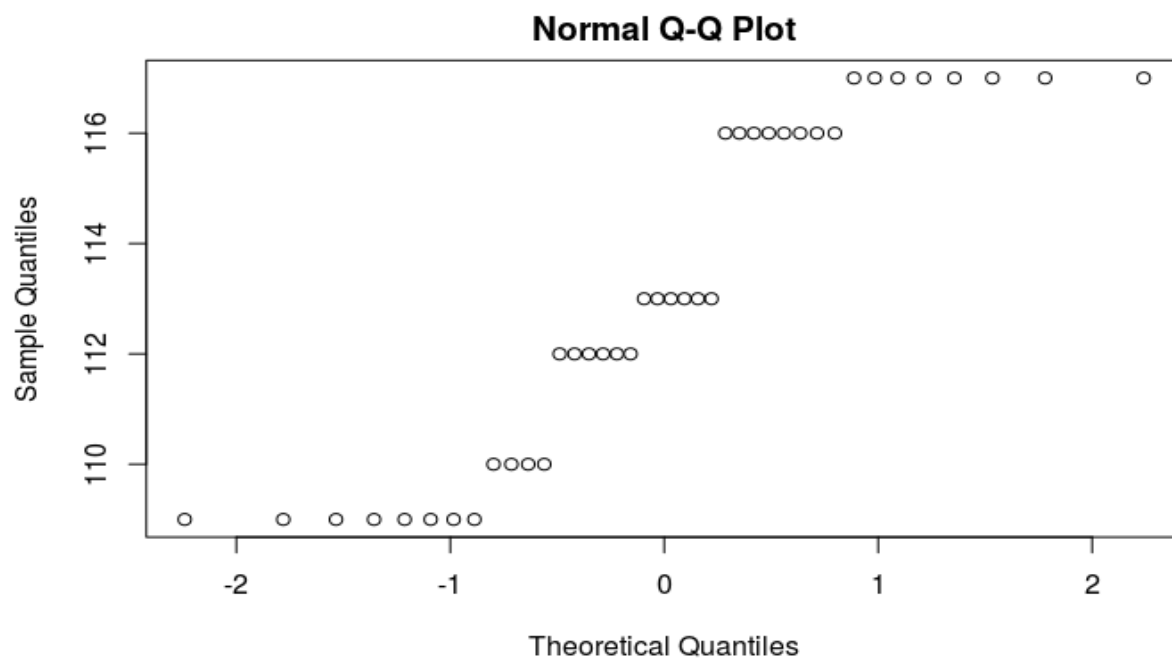
In the two-sample t-test, we are comparing the mean Amount of feed of the Gallus breed which has two categories Marans and Ameraucana.



The qqnorm plots help us to understand how normal data is distributed. This plot gives the total amount of feed given to the Gallus Chicken.



The qqnorm plots help us to understand how normal data is distributed. This plot gives the total amount of feed given to Marans Breed of the Gallus Chicken.



The qqnorm plots help us to understand how normal data is distributed. This plot gives the total amount of feed given to Ameraucana Breed of the Gallus Chicken.

Population Parameter:

*We want to make to make inference about population mean difference
between Marans and Ameraucana μ_{M-A}*

Hypothesis Test:

$$H_0 : \mu_{M-A} = 0$$

The population mean difference between Marans and Ameraucana is zero

$$H_A : \mu_{M-A} \neq 0$$

The population mean difference between Marans and Ameraucana is not equal to 0

Sample Statistics:

The sample mean difference between Marans and Ameraucana.

$$\bar{x}_M - \bar{x}_A$$

Test Statistic for Two sample T-test following T-distribution:

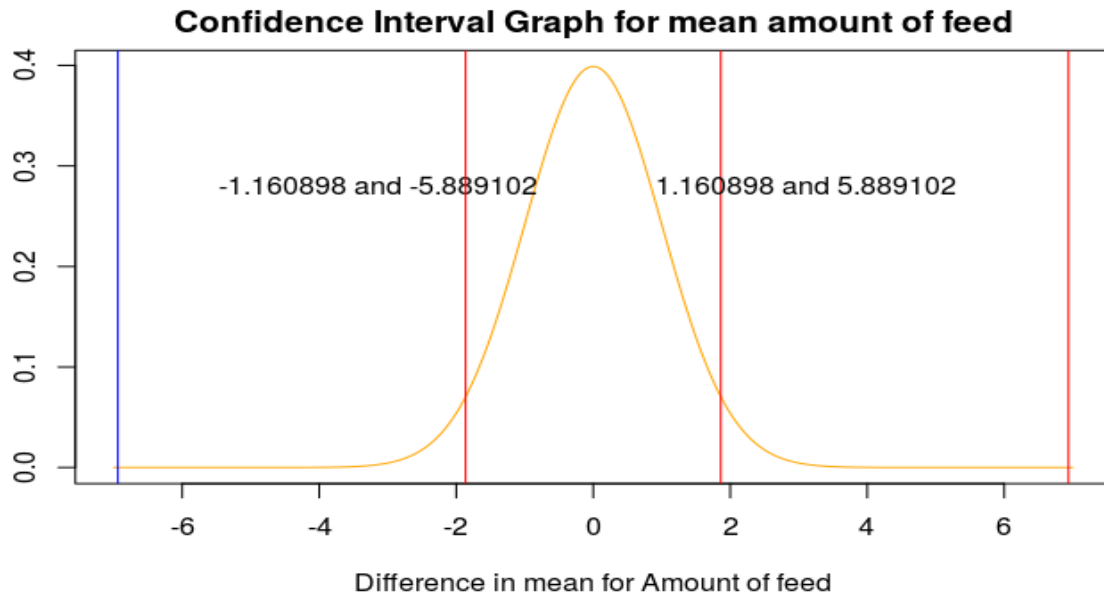
$$t_{\min(n_s - 1, n_b - 1)} = \frac{(x_M - x_A) - (\mu_M - \mu_A)}{\sqrt{\frac{s_x^2}{n_s} + \frac{s_b^2}{n_b}}}$$

Built-in Two sample proportion test:

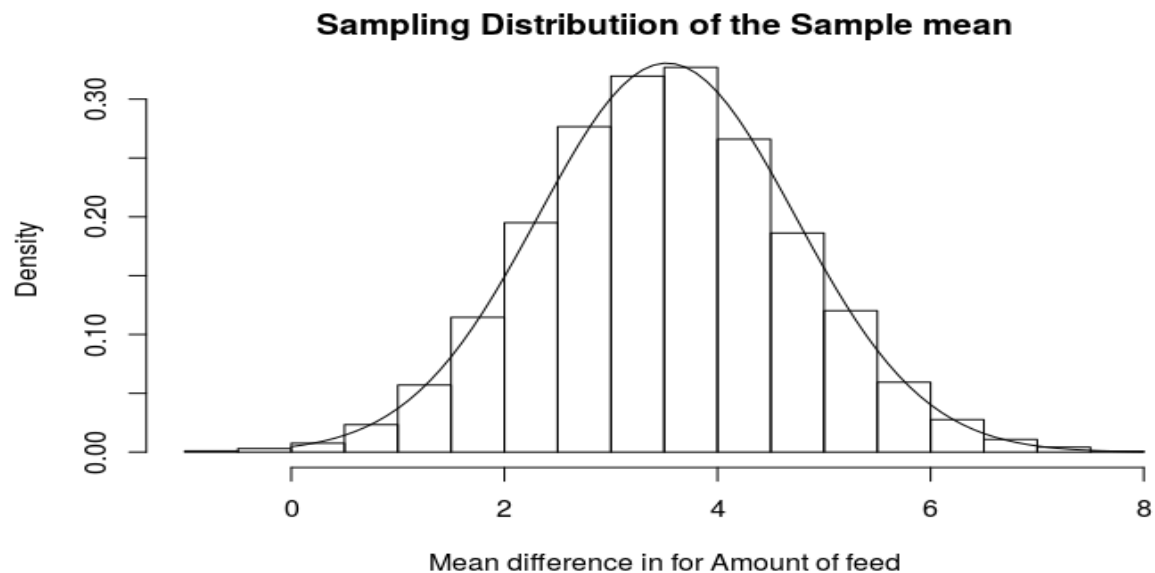
```
Welch Two Sample t-test

data: Marans and Ameraucana
t = 2.9261, df = 53.639, p-value = 0.005023
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.109433 5.940567
sample estimates:
mean of x mean of y
 116.675  113.150
```

From the built-in test, we can find the p-value for two-sample t-tests is 0.005023 and 95 % confident interval is between 1.109433 and 5.940567.

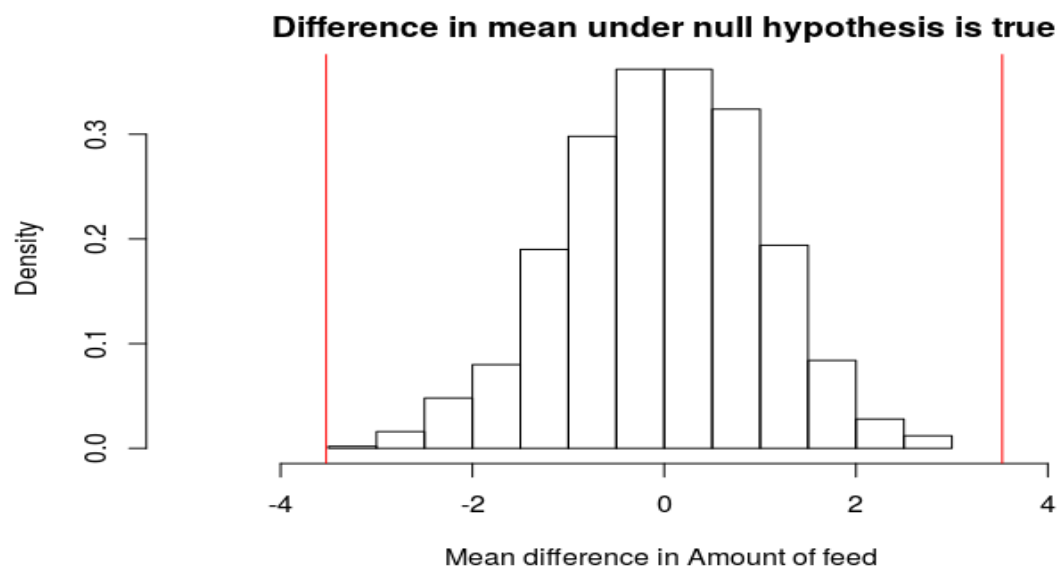
Building 95% confidence interval:

The 95% confidence interval shows the true mean for feed for Marans and Ameraucana lies between 1.160898 and 5.889102.

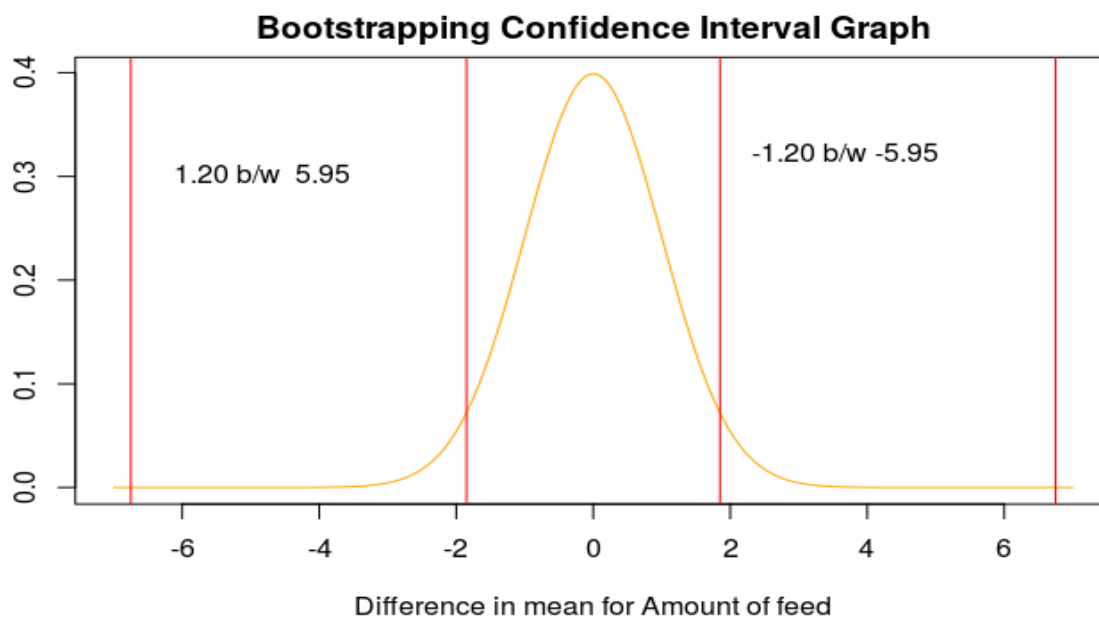
Bootstrapping Method:

The distribution shows how data are sampled using a bootstrapped approach to get normality on the sampled data. The data is sampled for the difference in mean of amount of feed for Marans and Ameraucana.

Null hypothesis using Bootstrap approach:



The data is sampled 10000 using a bootstrapped approach where the null hypothesis is true. The null hypothesis states the difference in mean of Amount of feed between Marans and Ameraucana is zero.



The Bootstrap Confidence interval shows the difference in mean for the Amount of feed between Marans and Ameraucana lies between 1.20 and 5.95.

Interpretation:

There is strong evidence that (p-value=0.005697042) to suggest the true mean difference in the Amount of feed given to Gallus chicken is greater than zero. We reject the null hypothesis that the true mean difference is equal to zero at $\alpha = 0.05$ level. Using Confidence interval created by bootstrap we are 95% confident that the true mean difference is between 1.20 and 5.95. The interval using the empirical methods, especially the quantile method is much compact which agrees with our p-value being a bit conservative and is bit significant when compared to traditional t-tools.

Two Sample Test for difference in Proportion:

What is the proportion of the white beak color of Gallus Chicken with Marans and Ameraucana? Conditions for doing a Two-sample test for proportions.

- The sample needs to be representative of the population-Gallus Breed.
- Categorical response variable with 2 categories- Brown and White.
- 2 independent samples from 2 populations- Marans and Ameraucana.
- $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ for both population

Population parameter:

We are interested in difference in population proportion of white beak color across the Gallus breed of Marans and Ameraucana.

Sample Statistic:

$$\hat{p}_w - \hat{p}_o$$

Test Statistic:

$$z = \frac{(\hat{p}_w - \hat{p}_o) - (p_w - p_o)}{\sqrt{\frac{\hat{p}_w(1-\hat{p}_w)}{n_w} + \frac{\hat{p}_o(1-\hat{p}_o)}{n_o}}} \sim N(0, 1)$$

Hypothesis Testing:

Null Hypothesis

The True Proportion between Marans and Ameraucana white beak color of Gallus Chicken is 0.

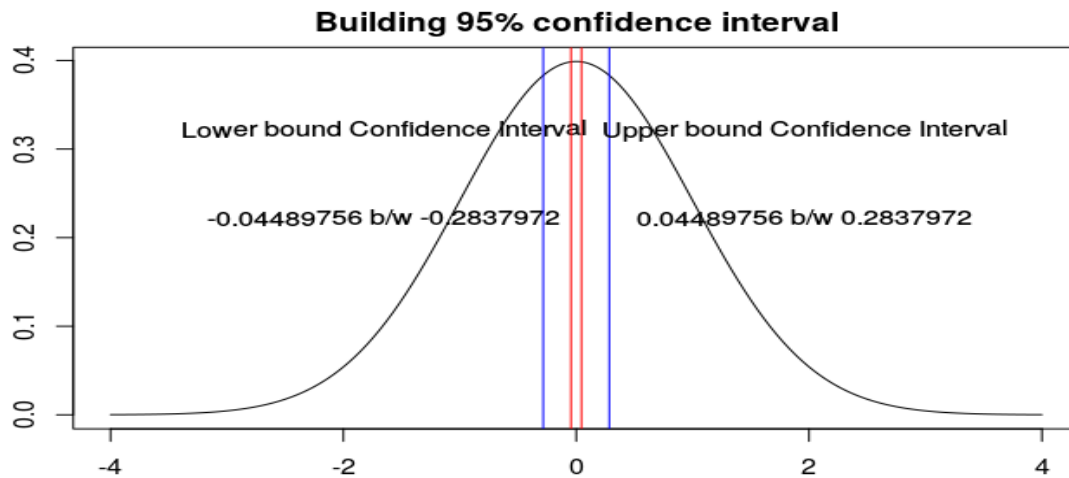
$$H_0 : P_0 - P_w = 0$$

Alternative Hypothesis

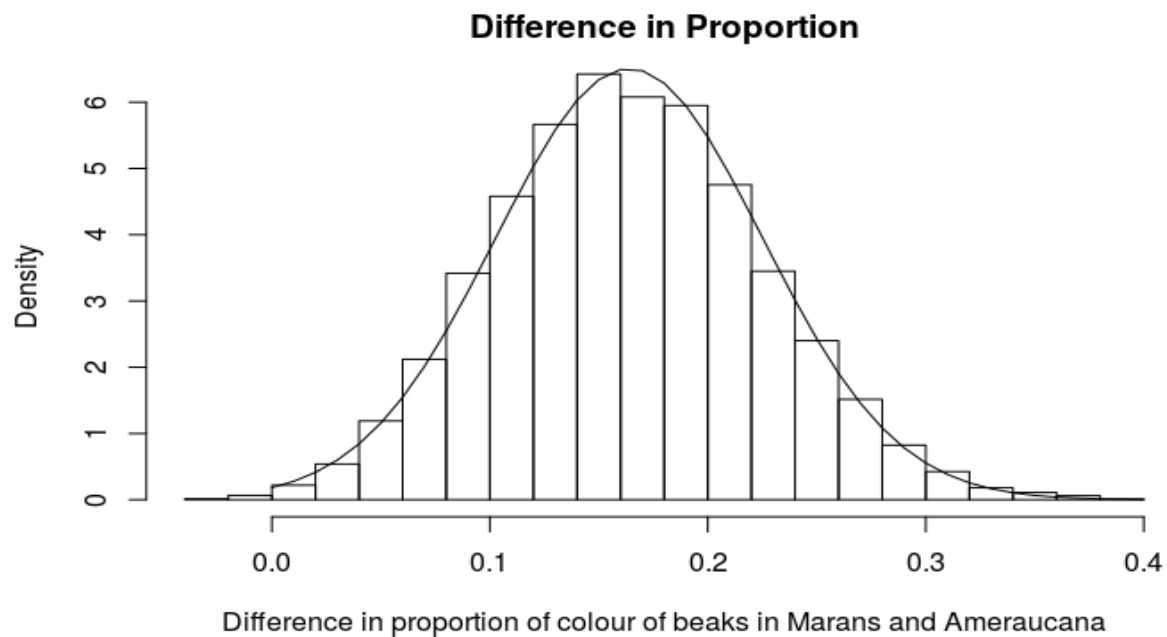
The True Proportion between Marans and Ameraucana white beak color of Gallus Chicken is not zero.

$$H_0 : P_0 - P_w \neq 0$$

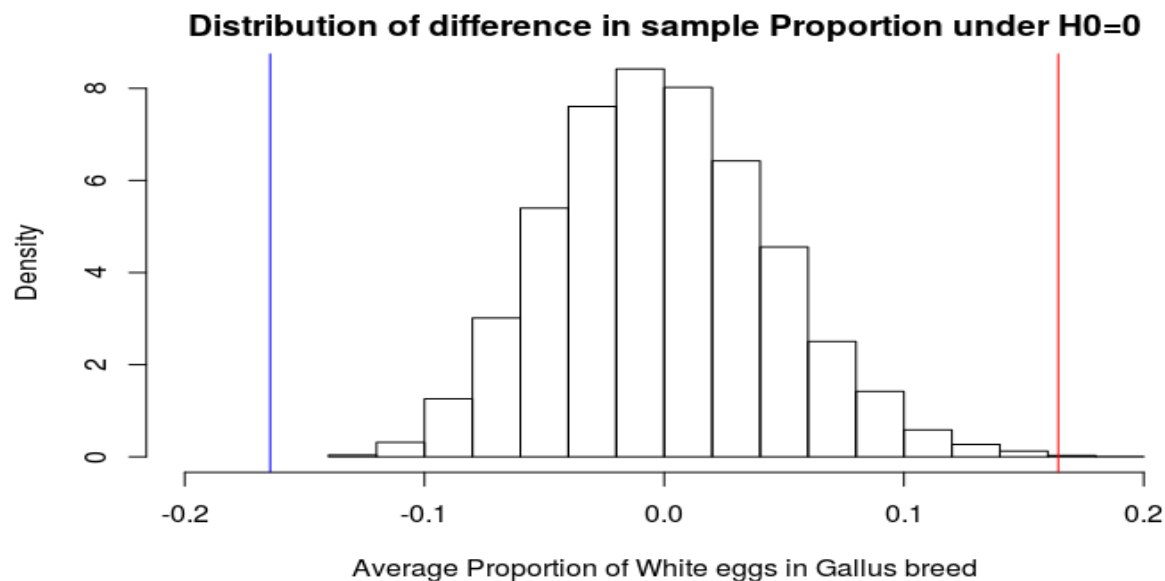
Building 95% confidence interval:



A 95% Confidence interval is drawn for the proportion of white beak color between Marans and Ameraucana and it lies between 0.04489756 and 0.2837972.

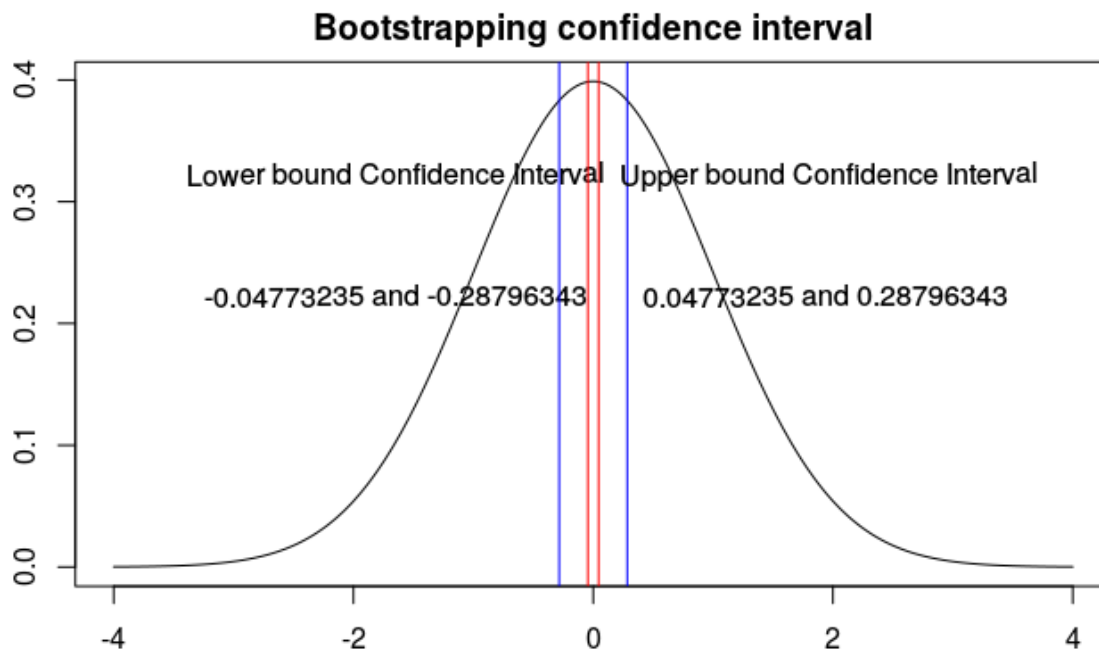


The distribution shows how data are sampled using a bootstrapped approach to get normality on the sampled data. The data is sampled for the difference in proportion between Marans and Ameraucana for white beak color.



The data is sampled 10000 using a bootstrapped approach where the null hypothesis is true. The null hypothesis states the difference in proportion for white beak color between Marans and Ameraucana is zero.

Bootstrapping Confidence interval:



The Bootstrap Confidence interval shows the difference in the proportion of white color beaks between Marans and Ameraucana lies between 0.04773235 and 0.28796343.

Interpretation:

The P-value($8e-04$) suggests that there is strong evidence that the Average difference proportions of White-colored beaks in Marans and Ameraucana are not the same. We reject the null hypothesis at the $\alpha = 0.05$ level that both proportions are not equal. With a 95% confidence interval, we say that the true proportion of White eggs lies between 0.04773235 and 0.28796343. The interval using the empirical methods, especially the quantile method is much compact which agrees with our p-value being a bit accurate and is bit significant when compared to traditional t-tools.

Chi-Square Goodness of fit:**Conditions:**

Single category with more than three categorical variables.
At least five datapoints in each category.

Parameter of interest

To find the proportion of leg color in Gallus Breed.
To find True Value of PA, PB, PC, PD.

Hypothesis Testing:**Null Hypothesis:**

$H_0: PA = PB = PC = PD = 0.4$

The proportion of each Gallus leg color is equal to 0.4

Alternative Hypothesis:

$H_A: \text{Some } p_i \neq .4$

The proportion of at least one Gallus leg color is not equal to 0.4

Under Null hypothesis the expected count is $(950)(0.4) = 380$

P-value:

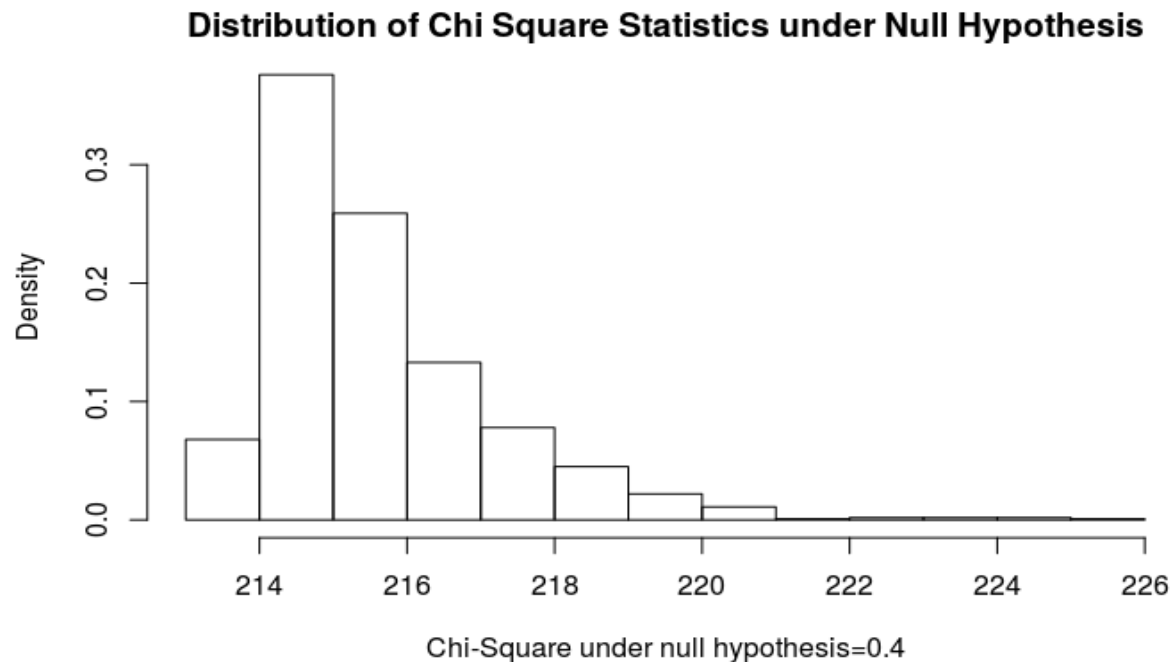
The p-value of chi-square test is $5.099636e-185$.

Test Statistic and Distribution:

The test statistic for chi- square distribution is given by

$$\chi^2 = \sum_{i=1}^k \left(\frac{(O_i - E)^2}{E} \right) \sim \chi_{k-1}^2$$

Chi-square distribution by introducing randomness:



The chi-square distribution of null hypothesis is true is plotted. The null hypothesis states that overall proportion of legs is 0.4 percent.

Interpretation:

The data provides that there is strong evidence that the proportion is not equal to 0.4. We reject the null hypothesis at the $\alpha = 0.05$ level.

Discussion:

Summary of the findings:

In one sample t-test we were able to find true mean weight of Gallus chicken, In one sample proportion test we founded the proportion of brown eggs in Gallus breed, In Two sample difference in mean test we were able to find true mean difference between amount of feed given to Marans and Ameraucana, Two-sample test for proportions we were able to find the proportions of white beak in Marans and Ameraucana. The chi-square method was used to find the proportion color of Gallus leg.

Implication of finding:

Based on the test were able to find various analysis and were not sure whether the result is significant for the overall population of Gallus breed. Since our study is limited to Africa.

Limitation:

This data is collected from an online source and whether the data points are correct and independent is not known.

Resources:

The data were taken from an online platform and the link is given below.

<https://www.kaggle.com/phuzoman/egg-producing-chickens>

Appendix:

```
knitr::opts_chunk$set(echo = TRUE)

install.packages('RCurl')
install.packages('mosaic')
install.packages('tidyverse')
install.packages("rlang")
install.packages('naniar')
install.packages('dplyr')

library(ggplot2)
library(dplyr)
library(reshape2)
library(tidyverse)
library(naniar)
library(mosaic)

df<-read.csv('dataset.csv')
head(df)
summary(df)
glimpse(df)
miss_var_summary(df)
vis_miss(df,cluster = TRUE)
gg_miss_var(df)
gg_miss_upset(df)
```

```

df_new<-drop_na(df)
df_new<-as.data.frame.matrix(df_new)
df_new
summary(df_new)

glimpse(df_new)

ggplot(df_new,aes(GallusWeight,Age))+geom_bar(stat = 'identity',fill='blue')+
xlab('Gallus Age in Days')+ylab('Gallus weight in Grams')+ggtitle('Gallus wei
ght corresponding to their Age')

ggplot(df_new,aes(SunLightExposure,AmountOfFeed,color=GallusBreed))+geom_poin
t()+xlab('Sunlight exposure of Gallus')+ylab('Amount of Feed')+ggtitle('Gallu
s breed and their corresponding feed and sunlight exposure')

breed<- df_new %>% group_by(GallusBreed)%>%summarize(Overall=sum(EggsPerDay))
ggplot(breed,aes(GallusBreed,Overall,width=0.30))+geom_bar(stat = 'identity',
col='black',fill='orange')+xlab('Gallus breed')+ylab('Total number of eggs')+
ggtitle('Total number of Eggs laid by Gallus breed')

ggplot(df_new,aes(x=GallusEggWeight,fill=GallusEggColor))+geom_histogram(bins
=50)+xlab('Gallus Egg Weight')+ggtitle('Gallus Egg weight based on color of E
ggs')

ggplot(df_new,aes(GallusEggColor))+geom_bar(fill='brown')+xlab('Gallus Egg Co
lor')+ylab('Total number of eggs')+ ggtitle('Gallus Egg Ratio')

ggplot(df_new,aes(GallusEggWeight,fill=GallusBreed))+geom_density(alpha=0.3)+
xlab('Gallus Egg Weight')+ylab('')+ggtitle('Gallus weight based on Gallus Br
eed')

melt<-melt(df_new)

ggplot(melt,aes(factor(variable),value))+geom_boxplot(color='brown')+coord_fl
ip()

boxplot(df_new$AmountOfFeed,df_new$GallusEggWeight)

title('boxplot for Amount of feed and Egg weight')

ggplot(df_new,aes(GallusLegShanksColor,fill=GallusBreed))+geom_bar()+ylab('')
+ggtitle('Gallus breed Leg colors')

ggplot(df_new,aes(GallusBeakColor,fill=GallusBreed))+geom_bar()+ggtitle('Gall
us breed beak color')

Gallus_weight<-df_new$GallusWeight
head(Gallus_weight)

```

```

qqnorm(Gallus_weight)

ggplot(df_new, aes(GallusWeight))+geom_histogram(binwidth =20)

mean_gallus<- mean(Gallus_weight)

mu_o<-2200

n<-length(Gallus_weight)

sd_gallus<- sd(Gallus_weight)
sd_gallus

t_value<- (mean_gallus-mu_o)/(sd_gallus/sqrt(n))
t_value

p_value<- pt(q=t_value, df=n-1, lower.tail = FALSE)*2
p_value
t.test(Gallus_weight,alternative = 'two.sided', mu=2200)
lower<-mean_gallus-(qt(0.975,n-1)*(sd_gallus/sqrt(n)))
lower

upper<-mean_gallus+(qt(0.975,n-1)*(sd_gallus/sqrt(n)))
upper
g<-seq(-5,5,0.01)
df_new<-dt(seq(-5,5,0.01),n-1)
plot(g,df_new,type='l',xlab='',ylab='',main=' Plotting t-distribution with n-
1 df_new')
abline(v=c(-t_value,t_value))
text(-t_value,0.025,'t=- 2.240687',srt=0.2,pos=4)
text(t_value,0.025,'t= 2.240687',srt=0.2,pos=2)
abline(v=c(qt(0.025,n-1),qt(0.975,n-1)))
text(qt(0.025,n-1),0.07,"t=-2.042272",srt=0.2,pos=2)

```



```

text(qt(0.975,n-1),0.07,"t=2.042272",srt=0.2,pos=4)
plot(seq(-4,4,0.01),dt(seq(-4,4,0.01),n-1),type='l',xlab = '',ylab = '',main=
'Building 95% confidence interval')
abline(v=0,col=c('yellow'))
abline(v = c( upper*10^-3,lower*10^-3),col=c('blue','red'))
text(qt(0.05,n-1),0.3,'Lower bound Confidence Interval',srt=0.3,pos=3)
text(qt(0.05,n-1),0.2,'-2203.993 b/w -2260.323',srt=0.3,pos=3)
abline(v = c(-upper*10^-3,-lower*10^-3),col=c('blue','red'))
text(qt(0.975,n-1),0.3,'Upper bound Confidence Interval',srt=0.3,pos=3)
text(qt(0.975,n-1),0.2,'2203.993 b/w 2260.323',srt=0.3,pos=3)
num<-10000

results<- rep(NA,num)

for( i in 1: num){
  results[i]<-mean(sample(x=Gallus_weight,size=n,replace=TRUE))
}

hist(results,freq=FALSE,main='Sampling distribution of sample mean',xlab='Ave
rage Gallus Weight')
lines(x=seq(2180,2280,0.1),dnorm(seq(2180,2280,0.1),mean=mean(results),sd= sd
(results)))
Gallus_weight
H0<- Gallus_weight-mean(Gallus_weight)+mu_o

nums<-10000

results_h0<-rep(NA,nums)

for (i in 1 : nums) {
  results_h0[i]<-mean(sample(x=H0,size=n,replace=TRUE))
}

mean(H0)
mean_gallus

```

```

hist(results_h0,freq = FALSE,main= 'Sampling mean of sample distribution give
n null hypothesis is true',xlab='Average mean weight of Gallus')
m_gallus<-Gallus_weight-mean(Gallus_weight)+mu_o
low_and_extreme<-mean(results_h0)+(mean(results_h0)-mean_gallus)
abline(v=c(low_and_extreme),col='red')
high_and_extreme<-mean_gallus
abline(v=c(high_and_extreme),col='red')
sum_of_less_extreme_tail<-sum(results_h0 <= low_and_extreme)
sum_of_less_extreme_tail
sum_of_more_extreme_tail<-sum(results_h0 >= mean_gallus)
sum_of_more_extreme_tail
boot_strap_p_value<- (sum_of_less_extreme_tail+sum_of_more_extreme_tail)/nums
boot_strap_p_value
p_value
bootstrap_sd<- sd(results)

c(quantile(results,c(0.025,0.975)))

plot(seq(-4,4,0.1),dt(seq(-4,4,0.1),n-1),type='l',xlab='',ylab='',main='Boots
trapping Confidence Interval')
abline(v=c(0))
abline(v=c(2204.116*10^-3,2260.643 *10^-3),col=c('blue','red'))
text(qt(0.05,n-1),0.3,'Lower bound Confidence Interval',srt=0.3,pos=3)
text(qt(0.05,n-1),0.25,'-2204.116 b/w -2260.643',srt=0.3,pos=3)
abline(v=c(-2204.116*10^-3,-2260.643 *10^-3),col=c('blue','red'))
text(qt(0.975,n-1),0.3,'Upper bound Confidence Interval',srt=0.3,pos=3)
text(qt(0.975,n-1),0.25,'2204.116 b/w 2260.643',srt=0.3,pos=3)

df_new<-drop_na(df)
df_new<-as.data.frame.matrix(df_new)
df_new%>%group_by(c(GallusBreed))%>%count(c(GallusEggColor))
p<-895/950
z<-(p-0.95)/sqrt((0.95*(1-0.95))/950)
z

```

```

pnorm( z,lower.tail =FALSE)
binom.test(x=895,n=950,p=(0.95),alternative='greater')
cat('exact binomial test')
binom.test(x=895,n=950,p=(0.95),alternative='greater')$conf.int

plot(seq(-3,3,0.1),dt(seq(-3,3,0.1),n-1),type='l',xlab = ' ',ylab='',main='Co
nfidence interval')
abline(v=c(1),col='blue')
abline(v=c( 0.9280587),col='red')
text(qt(0.95,n-1),0.3,'95% Confidence Interval for of proportion brown egg i
s between ',srt=0.2,pos=2.5)
text(qt(0.975,n-1),0.25,'0.9280587 and 1.0000000 ',srt=0.2,pos=2)


BROWN<-factor(rep(c('BROWN','other'),c(895,950-895)))
table(BROWN)


BROWN<-rep(c(1,0),c(895,950-895))
BROWN
nums<-10000

results<- rep(NA,nums)

for (i in 1:nums){

  results[i]<-mean(sample(x=BROWN,size=950,replace=TRUE))

}

hist(results,freq=FALSE,xlab='Proportion of Brown Eggs', main='Sampling distr
ibution of the sample proportion',ylab='')

lines(x=seq(0.91,0.97,0.001),dnorm(seq(0.91,0.97,0.001),mean=mean(results),sd
=sd(results)))

cat('Bootstrap confidence Interval')
c(quantile(results,0.05,1))

```

```

plot(seq(-3,3,0.1),dt(seq(-3,3,0.1),n-1),type='l',xlab = ' ',ylab='',main='Confidence interval')
abline(v=c(1),col='blue')
abline(v=c( 0.9294737),col='red')
text(qt(0.975,n-1),0.3,'95% Bootstrapping Confidence Interval ',srt=0.2,pos=2.5)
text(qt(0.975,n-1),0.25,'0.9294737 and 1 ',srt=0.2,pos=2)

Null_hypo<- rep(c(1,0),c(903,950-903))

nums<-10000

results<-rep(NA,nums)

for (i in 1:nums){
  results[i]<-mean(sample(x=Null_hypo,size=950,replace=TRUE))
}

hist(results,freq=FALSE,xlab='Proportion of Brown eggs',ylab='',main='Sampling distribution of proportion under H0=0.95')

abline(v=c(0.941),col='red')
u<-sum(results >=0.95)
bootstrap_pvalue<- u/nums
cat('Bootstrap P-Value')
bootstrap_pvalue
pnorm(z,lower.tail = FALSE)
binom.test(895,950,p=0.95,alternative = 'greater')

qqnorm(df_new$AmountOfFeed)

mydata<-filter(df_new,df_new$GallusBreed=='Ameraucana')
sample<-mydata[sample(nrow(mydata),40),]

myda<-filter(df_new,df_new$GallusBreed=='Marans')
sam<-myda[sample(nrow(myda),40),]

```

```

df_new.n<-rbind(sample,sam)
df_new.n

Marans<-(df_new.n$AmountOfFeed[df_new.n$GallusBreed=='Marans'])
qqnorm(Marans)

Ameraucana<-(df_new.n$AmountOfFeed[df_new.n$GallusBreed=='Ameraucana'])
qqnorm(Ameraucana)

mean_m<-mean(Marans)
mean_m

mean_a<-mean(Ameraucana)
mean_a

mu_0<-0

n_m<-length(Marans)
n_m

n_a<-length(Ameraucana)
n_a

s_m<-sd(Marans)**2
s_m

s_a<-sd(Ameraucana)**2
s_a

t_value<- (mean_m -mean_a-mu_0)/sqrt((s_m/n_m)+(s_a/n_a))
t_value

p_value<- pt(q=t_value,df=min(n_m,n_a)-1,lower.tail = FALSE)*2
p_value

```

```

x_bar<-mean_m-mean_a
x_bar
t.test(Marans,Ameraucana)
upper<- (mean_m-mean_a)+qt(0.025,n-1)*sqrt((s_m)/(n_m)+(s_a)/(n_a))
upper

lower<-(mean_m-mean_a)+qt(0.975,n-1)*sqrt((s_m)/(n_m)+(s_a)/(n_a))
lower

plot(seq(-7,7,0.01),dt(seq(-7,7,0.01),n-1),type='l',col='orange',xlab='Difference in mean for Amount of feed',ylab='',main='Confidence Interval Graph for mean amount of feed')
abline(v = c (upper,lower), col = c("red"))
text(qt(0.001,n-1),0.3,'-1.160898 and -5.889102 ',pos=1)
abline(v = c(-upper,-lower), col = c("red", "blue"))
text(qt(0.999,n-1),0.3,'1.160898 and 5.889102',pos=1)

nums<-10000

results<-rep(NA,nums)

for (i in 1:nums){

  mean_marans<-mean(sample(Marans,size=40,replace = TRUE))
  mean_Ameraucana<-mean(sample(Ameraucana,size=40,replace = TRUE))

  results[i]<- mean_marans-mean_Ameraucana
}

hist(results,freq=FALSE,xlab='Mean difference in for Amount of feed', main='Sampling Distribution of the Sample mean')

lines(x=seq(0,8,0.1),dnorm(seq(0,8,0.1),mean=mean(results),sd=sd(results)))
c(quantile(results,c(0.025,0.975)))

```

```

plot(seq(-7,7,0.01),dt(seq(-7,7,0.01),n-1),type='l',col='orange',xlab = 'Difference in mean for Amount of feed',ylab = '',main='Bootstrapping Confidence Interval Graph')
abline(v = c( 1.20,5.95 ), col = c("red"))
text(qt(0.001,n-1),0.3,' 1.20 b/w 5.95 ',pos=2)
abline(v = c( -1.29,-5.95 ), col = c("red"))
text(qt(0.9999,n-1),0.3,' -1.20 b/w -5.95 ',pos=3)

nums<-1000

results_h0<-rep(NA,nums)

for (i in 1:nums){

  s<-transform(df_new, GallusBreed=sample(GallusBreed))
  m<-mean(s$AmountOfFeed[s$GallusBreed=='Marans'])
  a<-mean(s$AmountOfFeed[s$GallusBreed=='Ameraucana'])
  results_h0[i]<- m - a
}

hist(results_h0,freq=FALSE,xlab='Mean difference in Amount of feed ',main='Difference in mean under null hypothesis is true',xlim=c(-5,5))

abline(v=x_bar,col='red')
abline(v=-x_bar,col='red')

upper_val<-sum(results_h0 >=x_bar)
lower_val<-sum(results_h0 <=-x_bar)

p_val<- (upper_val+lower_val)/nums

p_val

p_value

```

```

ma<-df_new%>%group_by(GallusBreed)%>%count(GallusBeakColor)
ma
Pw_hat<- 798/895
Pw_hat
Po_hat<-40/55
Po_hat
P_0<-0

n_w<-895
n_o<-55

p_w<-(Pw_hat*(1-Pw_hat))/n_w
p_o<-(Po_hat*(1-Po_hat))/n_o

z<-(Po_hat-Pw_hat-P_0)/sqrt(p_w+p_o)
z

two_sided<-pnorm(q=z,lower.tail = TRUE)*2
two_sided

upper<-(Pw_hat-Po_hat)+(qnorm(0.975)*sqrt(p_w+p_o))
upper

lower<-(Pw_hat-Po_hat)+(qnorm(0.025)*sqrt(p_w+p_o))
lower

plot(seq(-4,4,0.01),dt(seq(-4,4,0.01),n-1),type='l',xlab = '',ylab = '',main=
'Building 95% confidence interval')
abline(v = c( upper,lower),col=c('blue','red'))
text(qt(0.05,n-1),0.3,'Lower bound Confidence Interval',srt=0.3,pos=3)
text(qt(0.05,n-1),0.2,'-0.04489756 b/w -0.2837972',srt=0.3,pos=3)
abline(v = c(-upper,-lower),col=c('blue','red'))
text(qt(0.975,n-1),0.3,'Upper bound Confidence Interval',srt=0.3,pos=3)
text(qt(0.975,n-1),0.2,' 0.04489756 b/w 0.2837972 ',srt=0.3,pos=3)

```



```

M<-rep(c(1,0),c(798,895-798))
A<-rep(c(1,0),c(40,55-40))

nums<-10000

results<-rep(NA,nums)

for(i in 1:nums){

  m<-mean(sample(x=M,size=n_w,replace=TRUE))
  a<-mean(sample(x=A,size=n_o,replace=TRUE))
  results[i]<-m-a

}

hist(results,freq=FALSE,xlab='Difference in proportion of colour of beaks in
Marans and Ameraucana',main='Difference in Proportion')

lines(x=seq(0,0.4,0.01),dnorm(seq(0,0.4,0.01),mean=mean(results),sd=sd(results)))

c(quantile(results,c(0.025,0.975)))

plot(seq(-4,4,0.01),dt(seq(-4,4,0.01),n-1),type='l',xlab = '',ylab = '',main=
'Bootstrapping confidence interval')
abline(v = c( upper,lower),col=c('blue','red'))
text(qt(0.05,n-1),0.3,'Lower bound Confidence Interval',srt=0.3,pos=3)
text(qt(0.05,n-1),0.2,'-0.04773235 and -0.28796343',srt=0.3,pos=3)
abline(v = c(-upper,-lower),col=c('blue','red'))
text(qt(0.975,n-1),0.3,'Upper bound Confidence Interval',srt=0.3,pos=3)
text(qt(0.975,n-1),0.2,'0.04773235 and 0.28796343 ',srt=0.3,pos=3)


new_l<-data.frame("White"=c(M,A),"GallusBreed"=rep(c("Marans","Ameraucana"),c
(n_w,n_o)))
summary(new_l$GallusBreed)

```

```

m<-mean(new_1$White[new_1$GallusBreed=='Marans'])
a<-mean(new_1$White[new_1$GallusBreed=='Ameraucana'])

m==Pw_hat
a==Po_hat
nums<-10000

H0<- rep(NA, nums)

for (i in 1:nums)
{
  s<-transform(new_1,GallusBreed=sample(GallusBreed))
  prop_m<-mean(s$White[s$GallusBreed=='Marans'])
  prop_a<-mean(s$White[s$GallusBreed=='Ameraucana'])
  H0[i]<-prop_m-prop_a
}

hist(H0,freq=FALSE,main='Distribution of difference in sample Proportion unde
r H0=0',xlab='Average Proportion of White eggs in Gallus breed',xlim = c(-0.2
,0.2))

diff_prop<-Pw_hat-Po_hat
abline(v=c(diff_prop),col='red')
abline(v=c(-diff_prop),col='blue')
upper<-sum(H0<=-diff_prop)
lower<-sum(H0>=diff_prop)

p_va<- (upper+lower)/nums
p_va
da<-df_new$GallusLegShanksColor
head(da)
table(da)
prop.table(table(da))
p<-c(106,10,45,789)*0.4
p

```

```

chi<-sum(((table(da)-p)^2)/p)
chi

pchisq(chi,4-1,lower.tail = FALSE)

n_h0<-rep(c('BROWN','Slate Black','Slate Blue','White'),380)
table(n_h0)

nums<-1000

h0<-rep(NA,nums)

for (i in 1: nums)
{

  m<-sample(n_h0,950,replace=TRUE)
  h0[i]<-sum(((table(m)-380)^2)/380)
}

hist(h0,freq=FALSE,main='Distribution of Chi Square Statistics under Null Hypothesis',xlab='Chi-Square under null hypothesis=0.4')

p_<-sum(h0>=sum(((table(da)-p)^2)/p))/nums
p_

```