MATH1324 Applied Analytics Assignment 2

Indian Agriculture Analysis- Two Sample t-test

Shinoj Philip John [s4022577], Sebastian George [s3997919], Mimamsa Singh[s4036609]

Last updated: 14 October, 2023

RPubs link information

- **Rpubs link:**
  http://rpubs.com/shinojphilipjohn/1098610

Introduction

- *Agriculture* sector is key to Indian economy contributing around 20% to the country's GDP.
- We wanted to analyse the cashew nut market particularly as it is one of the most exotic nuts and India being the top three producer of cashew nuts, made it an interesting topic to investigate on.
- We will be investigating, if the production of cashew nuts are the same in the two seasons, 'Kharif' and 'Rabi'. Kharif in which crops are sown at the beginning of the monsoon rains and Rabi in which crops are sown at the end of the monsoon rains.
- To see the effect of seasons on the crop production.

Problem Statement

- We used a Two-Sample T-Test and investigated whether the mean production of cashew nuts in Kharif and Rabi season are same. If they are different, we will reject the null hypothesis.
- The 'Production' variable was used for descriptive statistics across the two seasons Kharif and Rabi.
- QQ plots was used to explore the normality of the analysis

Data

- We are using Indian Agriculture data set which is a csv data set, here we have used this data set as it has a wide variety of crops across different seasons and time spans and in particularly has cashew nut production data from 2008 to 2020.
- Kaggle was used to find our data set and it is open sourced with Creative Commons Licence. Oleg Pyatakov is the author of the data set.
- **Data Description**

Here we have a data set which consists of 10 attributes and 345407 rows. It has 7 categorical attribute and 3 numeric data.

- **Important Variable Description**
- Crop - Name of the crop
- Year - Year of production
- Season - Divisions of the year characterised by weather patterns
- Production - Measure of total production

Data Cont.

- ## Data Loading

```
agriculture <- read_csv("India Agriculture Crop Production.csv")
dim(agriculture)
```

```
## [1] 345407     10
```

```
str(agriculture)
```

```
## spc_tbl_ [345,407 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ State           : chr [1:345407] "Andaman and Nicobar Islands" "Andaman and Nicobar
Islands" "Andaman and Nicobar Islands" "Andaman and Nicobar Islands" ...
##  $ District        : chr [1:345407] "NICOBARS" "NICOBARS" "NICOBARS" "NORTH AND MIDDLE
ANDAMAN" ...
##  $ Crop            : chr [1:345407] "Arecanut" "Arecanut" "Arecanut" "Arecanut" ...
##  $ Year            : chr [1:345407] "2001-02" "2002-03" "2003-04" "2001-02" ...
##  $ Season          : chr [1:345407] "Kharif" "Whole Year" "Whole Year" "Kharif" ...
##  $ Area            : num [1:345407] 1254 1258 1261 3100 3105 ...
##  $ Area Units      : chr [1:345407] "Hectare" "Hectare" "Hectare" "Hectare" ...
##  $ Production      : num [1:345407] 2061 2083 1525 5239 5267 ...
##  $ Production Units: chr [1:345407] "Tonnes" "Tonnes" "Tonnes" "Tonnes" ...
##  $ Yield           : num [1:345407] 1.64 1.66 1.21 1.69 1.7 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   State = col_character(),
##   ..   District = col_character(),
##   ..   Crop = col_character(),
##   ..   Year = col_character(),
##   ..   Season = col_character(),
##   ..   Area = col_double(),
##   ..   `Area Units` = col_character(),
##   ..   Production = col_double(),
##   ..   `Production Units` = col_character(),
##   ..   Yield = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Pre Processing

```
#Filtering the data set
agriculture_filtered <- agriculture %>% filter(Crop=="Cashewnut" &(Season=="Kharif" | Season=="Rabi"))

#Selecting wanted columns
agriculture_filtered <- agriculture_filtered %>% select(Crop,Year,Season,Production)

#Data type conversion
agriculture_filtered$Year <- as.factor(agriculture_filtered$Year)
agriculture_filtered$Season <- as.factor(agriculture_filtered$Season)

#Checking the  Levels of the factor variables

levels(agriculture_filtered$Year)
```

```
##  [1] "2008-09" "2009-10" "2010-11" "2011-12" "2012-13" "2013-14" "2014-15"
##  [8] "2015-16" "2016-17" "2017-18" "2018-19" "2019-20"
```

```
levels(agriculture_filtered$Season)
```

```
## [1] "Kharif" "Rabi"
```

Pre Processing Cont.

```
#Checking for missing values and special values
colSums(is.na(agriculture))
```

```
##             State          District              Crop              Year 
##                 0                 0                32                 0 
##            Season              Area        Area Units        Production 
##                 0                33                 0              4993 
## Production Units             Yield 
##                 0                33
```

```
is.sp_val <-function(x){
if(is.numeric(x)) (is.infinite(x) | is.nan(x)) }

sapply(agriculture,function(x) sum(is.sp_val(agriculture)))
```

```
##             State          District              Crop              Year 
##                 0                 0                 0                 0 
##            Season              Area        Area Units        Production 
##                 0                 0                 0                 0 
## Production Units             Yield 
##                 0                 0
```

Decsriptive Statistics

```
agriculture_filtered %>% group_by(Season) %>% summarise(Min = min(agriculture_filtered$Production,na.rm = TRUE),
                                        Q1 = quantile(agriculture_filtered$Production,probs = .25,na.rm = TRUE),
                                        Median = median(agriculture_filtered$Production, na.rm = TRUE),
                                        Q3 = quantile(agriculture_filtered$Production,probs = .75,na.rm = TRUE),
                                        Max = max(agriculture_filtered$Production,na.rm = TRUE),
                                        Mean = mean(agriculture_filtered$Production, na.rm = TRUE),
                                        SD = sd(agriculture_filtered$Production, na.rm = TRUE),
                                        n = n(),
                                        Missing = sum(is.na(agriculture_filtered$Production))) -> table1
knitr::kable(table1)
```

Season	Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
Kharif	0.44	15	73.5	3217.5	34663	2941.847	6191.521	151	4
Rabi	0.44	15	73.5	3217.5	34663	2941.847	6191.521	55	4

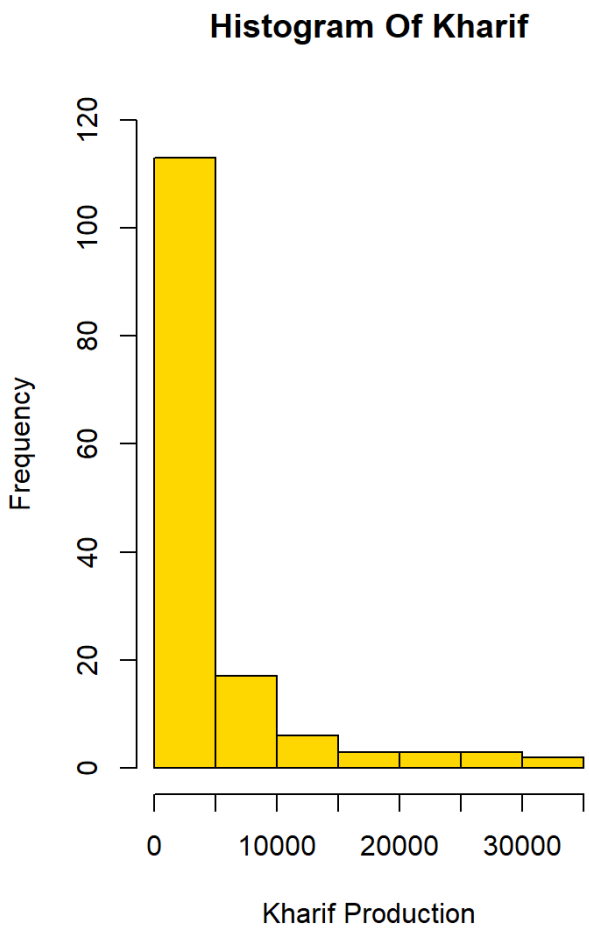
```
# Removing NA values
agriculture_filtered <- na.omit(agriculture_filtered)
```

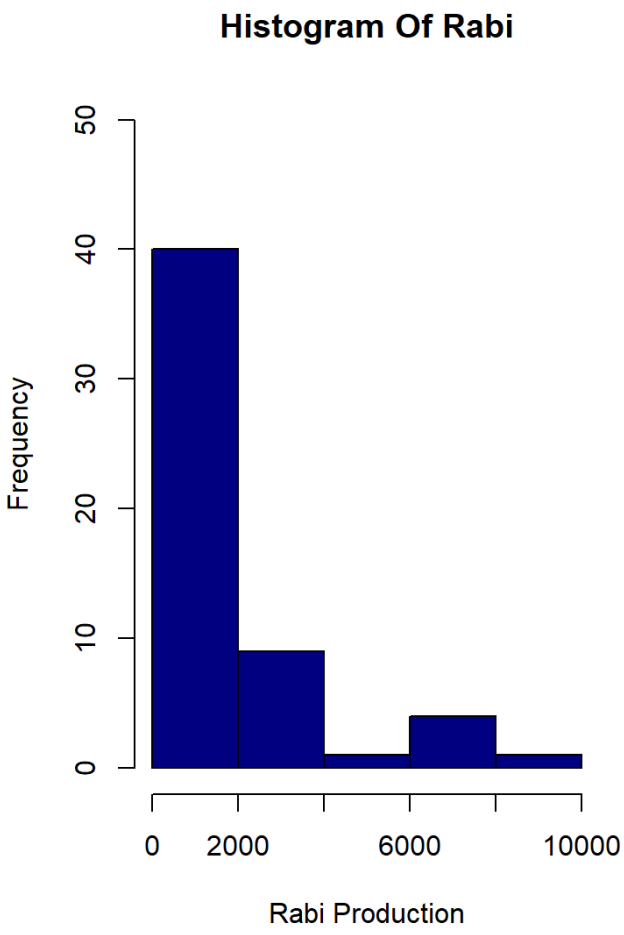
- Mean of Production is equal in both the seasons.
- Mean is higher than the Median in both the Seasons, thus causing it to have positive skewness.
- Missing value were removed from the data sets as it was just 4 observations in each of the data set.

Descriptive Statistics and Visualisation

```
#Segregating Seasons
agriculture_kharif <- agriculture_filtered %>% filter(Season=="Kharif")
agriculture_rabi <- agriculture_filtered %>% filter(Season=="Rabi")

#Histogram Plot
par(mfrow = c(1,2))
hist(agriculture_kharif$Production, xlab ="Kharif Production", ylim=c(0,120),main="Histogram Of Kharif",col="gold")
hist(agriculture_rabi$Production,  xlab ="Rabi Production",ylim=c(0,50),main="Histogram Of Rabi",col="navy")
```

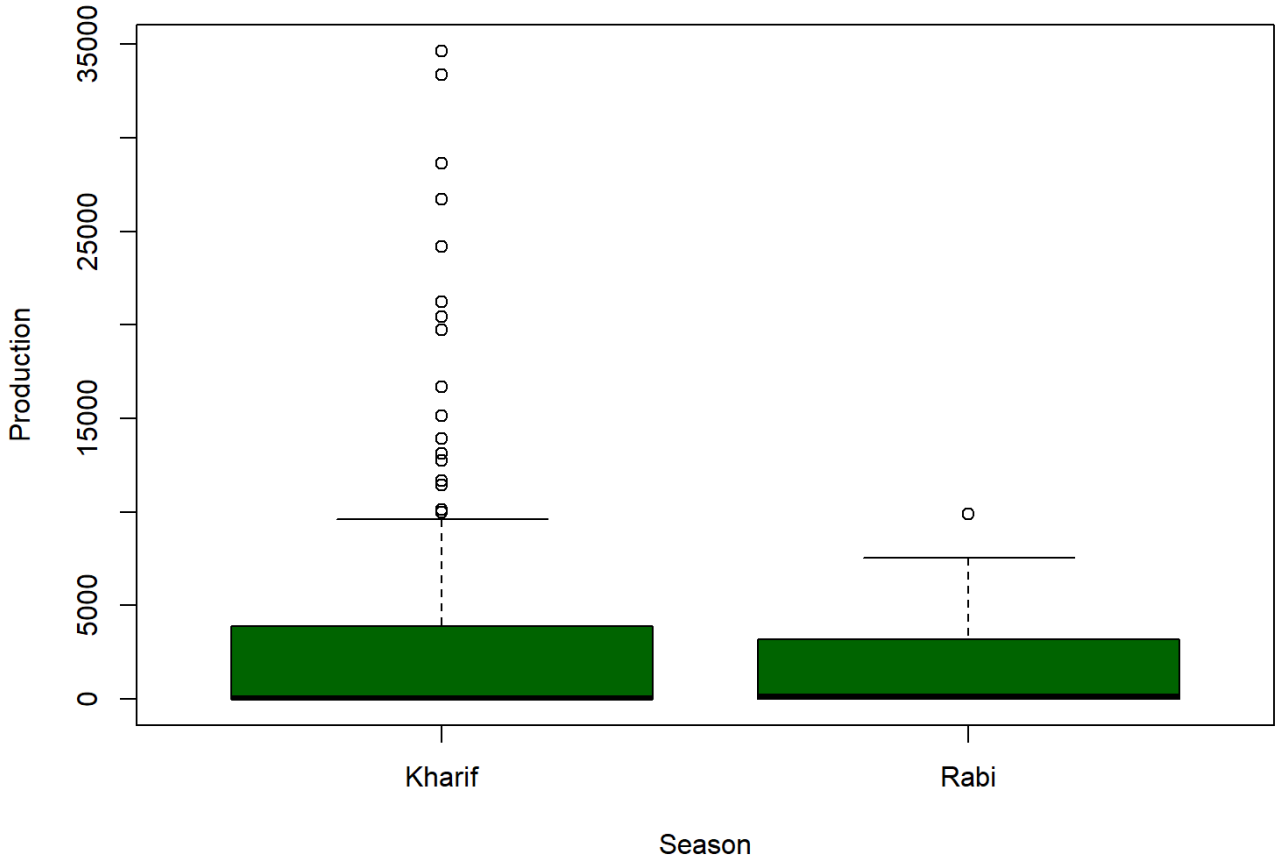




Outliers - Visual Representation

```r
# Quartiles and IQR
Q1 <- quantile(agriculture_filtered$Population, 0.25)
Q3 <- quantile(agriculture_filtered$Population, 0.75)
IQR <- Q3 - Q1
#Lower and upper bounds for outliers
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

#Boxplot
agriculture_filtered %>% boxplot(Production~Season, data=.,col = "darkgreen")
```



- It looks like Kharif Crops have a higher Production than Rabi, but we can conduct two sample t-test to get a conclusion.
- Both the Seasons have outliers, But Kharif has more outliers than Rabi.

Outliers Cont.

```r
# Capping outliers
cap <- function(x){
    quantiles <- quantile( x, c(.05, 0.25, 0.75, .95 ) )
    x[ x < quantiles[2] - 1.5*IQR(x) ] <- quantiles[1]
    x[ x > quantiles[3] + 1.5*IQR(x) ] <- quantiles[4]
    x}
agriculture_filtered$Production <- agriculture_filtered$Production %>% cap()

# Data in the first six rows
head(agriculture_filtered)
```

Crop	Year	Season	Production
<chr>	<fct>	<fct>	<dbl>
Cashewnut	2008-09	Kharif	5
Cashewnut	2009-10	Kharif	6
Cashewnut	2008-09	Kharif	4
Cashewnut	2009-10	Kharif	3
Cashewnut	2008-09	Kharif	164
Cashewnut	2009-10	Kharif	137

6 rows

- Outliers were capped in the Production variable.

Hypothesis Testing

H0-Cashew nut Crop produced in Kharif Season and Rabi Season have the same average Production in Tonnes.

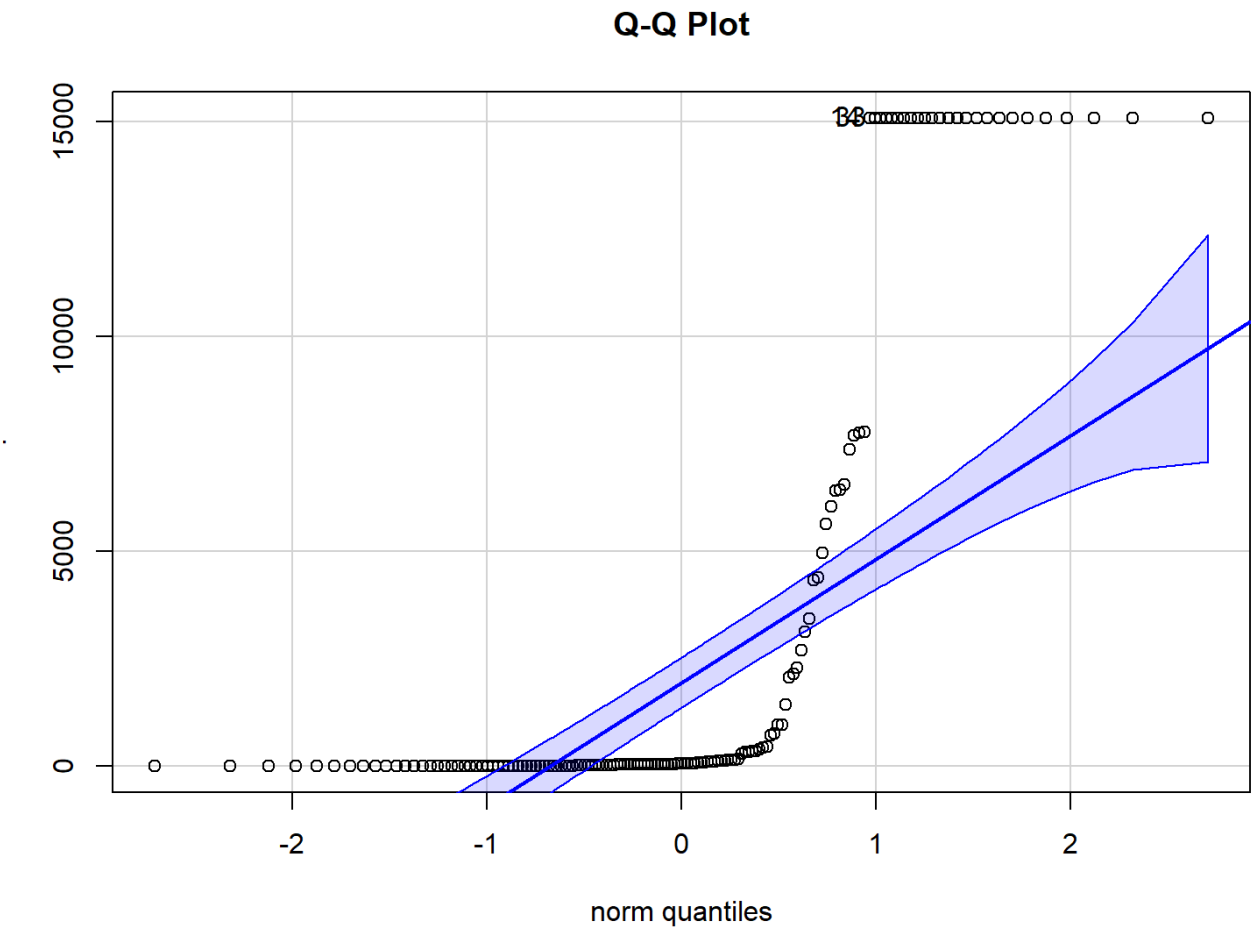
$$H_0 : \mu_1 - \mu_2 = 0$$

HA-Cashew nut Crop produced in Kharif Season and Rabi Season have different average Production in Tonnes.

$$H_A : \mu_1 - \mu_2 \neq 0$$

Hypothesis Testing- Q-Q Plots

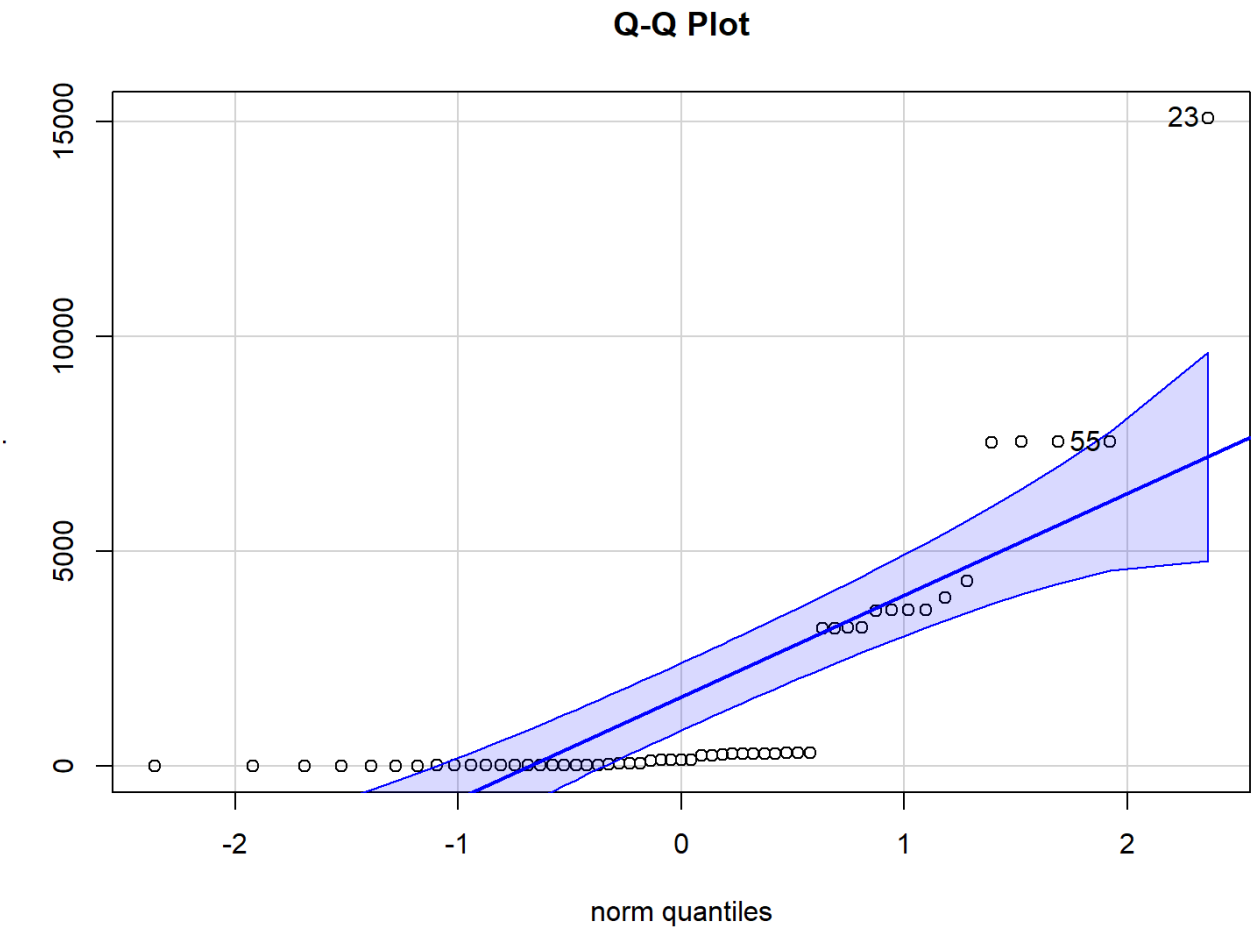
```
agriculture_kharif <- agriculture_filtered %>% filter(Season=="Kharif")
agriculture_kharif$Production %>% qqPlot(dist="norm",main="Q-Q Plot")
```




```
## [1] 14 33
```

Hypothesis Testing- Q-Q Plots Cont.

```
agriculture_rabi <- agriculture_filtered %>% filter(Season=="Rabi")
agriculture_rabi$Production %>% qqPlot(dist="norm",main="Q-Q Plot")
```




```
## [1] 23 55
```

Hypothesis Testing Cont.

Normality Inference

- Both Normal Q-Q Plots does not follow normal distribution .
- But it is not a concern as sample size in both the dataframes are Kharif=147 and Rabi=55 which is n>30 so according to *Central Limit Theorem*, the mean should be approximately normally distributed thus we can continue with our analysis.

Homogeneity of Variance- Levenes Test

```
leveneTest(Production~Season,data=agriculture_filtered) %>% as.data.frame()
```

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	1	4.727676	0.03085486
	200	*NA*	*NA*

2 rows

- The p value for production of cashew nuts is found to be p=0.03, which is p<0.05, therefore we reject H0. Hence we assume unequal variance. The Levenes Test is statistically significant.

Hypothesis Testing- Two Sample T-Test

```
t.test(
  Production ~ Season,
  data = agriculture_filtered,
  var.equal = FALSE,
  alternative = "two.sided"
  )
```

```
##
##  Welch Two Sample t-test
##
## data:  Production by Season
## t = 2.8143, df = 181.48, p-value = 0.005428
## alternative hypothesis: true difference in means between group Kharif and group Rabi is
not equal to 0
## 95 percent confidence interval:
##   511.9009 2913.4970
## sample estimates:
## mean in group Kharif   mean in group Rabi
##             3261.415             1548.716
```

```
# Difference between two means given through the t-test
3261.415-1548.716
```

```
## [1] 1712.699
```

Hypothesis Testing- Two Sample T-Test Cont.

Critical Value Approach

```
qt(p=0.025,df=147+55-2)
```

```
## [1] -1.971896
```

The test statistic from the two sample t-test assuming unequal variance is t=2.81 which is a more extreme value than the t-critical value of -1.97, thus we reject H0. According to the critical value approach, there is a statistically significant difference in the means of production of cashew nuts in the Kharif and Rabi seasons.

p-value approach As p=0.005 which is p<0.05,we reject H0. There is a statistical evidence to prove HA, that is there is a difference in the means of the production of cashew nuts in both the seasons.

Confidence Interval Approach The 95% CI difference between the means is 1712.699. The 95% CI are 511.9 and 2913.5. The interval does not capture the H0 value thus we will reject it. Thus there is a difference in the means of the production.

Discussion

- The hypothesis testing was done to check, if the production of cashew nuts are the same in the two seasons.
- Plotted Q-Q plot to check the normality and it showed that the data set does not follow normal distribution.
- But due to higher number of observations in both the data sets that is n>30, CLT was applied and thus analysis proceeded to the test.
- Levene's Test of Homogeneity of Variance showed unequal variance.
- Performed two sample t-test assuming unequal variance, to check if the hypothesis held true but we rejected H0 through the three approaches.
- Critical value, p-value and confidence interval approaches all rejected H0.
- Thus the results suggested that the production of cashew nuts in both the Kharif and Rabi seasons are different. Therefore we reject the null hypothesis.
- **Strengths**- Data had distinct values for the crops and thus easy to filter and select,
- **Limitations**- Number of the observations were limited for cashew nuts in the data set.
- **Future Investigation Direction**- To see what affects the difference in production in both the seasons and conduct the test on attributes like rainfall, temperature,etc.

References

- Pyatakov O (2023) India Agriculture Crop Production kaggle website, accessed 11 October 2023. https://www.kaggle.com/datasets/pyatakov/india-agriculture-crop-production/
- Baglin J (2016a) 'Module 5 Sampling: Randomly Representative' [Module 5 Notes, MATH1324], RMIT University, Melbourne.
- Baglin J (2016b) 'Module 6 Estimating Uncertainty Confidently' [Module 6 Notes, MATH1324], RMIT University, Melbourne.
- Baglin J (2016c) 'Module 7 Testing the Null: Data on Trial' [Module 7 Notes, MATH1324], RMIT University, Melbourne.
- contributors WC (2023) File:Cashew apples.jpg, Wikimedia Commons website, accessed 14 October 2023. https://commons.wikimedia.org/w/index.php?title=File:Cashew_apples.jpg&oldid=763295964
- contributors WC (2021) File:Cashew Flower.JPG, Wikimedia Commons website, accessed 14 October 2023. https://commons.wikimedia.org/w/index.php?title=File:Cashew_Flower.JPG&oldid=587582998