Jayalaxmi Agro Tech Assignment

Manish Grewal

24/08/2020

library(readxl)  
library(ggplot2)  
library(data.table)  
library(rstatix)  
library(tidyr)

setwd("C:/Users/manish.grewal/emdp/R/JATAgri")  
  
data <- read\_excel("IMB733-XLS-ENG.xlsx", sheet = "Data Sheet")  
bdata <- read\_excel("IMB733-XLS-ENG.xlsx", sheet = "Belagavi\_weather")  
ddata <- read\_excel("IMB733-XLS-ENG.xlsx", sheet = "Dharwad\_weather")  
  
#Remove the last row with totals  
data <- data[-124,]

## QUESTIONS

## 1. Anand, the cofounder of JAT, claims that disease 6 (leaf curl) information was accessed at least 60 times every month on average since October 2017 due to this disease outbreak. Test this claim at a significance level of 0.05 using an appropriate hypothesis test.

**Null hypothesis:** Information was accessed less than 60 times every month on an average since October 2017

**Alternative hypothesis:** Information was accessed 60 or more times every month on an average since October 2017

data1 <- data[95:123,"D6"]  
  
test <- t.test(data1, mu = 60, alternative = "greater")  
test

##   
## One Sample t-test  
##   
## data: data1  
## t = 2.341, df = 28, p-value = 0.01329  
## alternative hypothesis: true mean is greater than 60  
## 95 percent confidence interval:  
## 62.29976 Inf  
## sample estimates:  
## mean of x   
## 68.41379

**Interpretation:** As the p-value = 0.0133 is less than 0.05 in this case, we reject the null hypothesis and infer that the claim made by Anand that the access of disease 6 in his app is at least 60 times a month on average due to the disease outbreak is correct.

## 2. Among the app users for disease information, at least 15% of them access disease information related to disease 6. Use an appropriate hypothesis test to check this claim at a = 0.05.

**Null hypothesis:** Less than 15% of users accessed disease 6 information

**Alternative hypothesis:** 15% of more users accessed disease 6 information

data2 <- data[,5:15] # disease columns  
sums <- lapply(data2, sum)  
tot\_usage <- Reduce("+", sums)  
  
test <- prop.test(sums$D6, tot\_usage, p = 0.15, alternative = "greater")  
test

##   
## 1-sample proportions test with continuity correction  
##   
## data: sums$D6 out of tot\_usage, null probability 0.15  
## X-squared = 21.311, df = 1, p-value = 1.953e-06  
## alternative hypothesis: true p is greater than 0.15  
## 95 percent confidence interval:  
## 0.1564156 1.0000000  
## sample estimates:  
## p   
## 0.160082

**Interpretation:**: As the p-value = 1.952779710^{-6} is less than 0.05, we reject the null hypothesis. The claim is correct that 15% or more users accessed disease 6 information.

## 3. JAT believes that over the years, the average number of app users have increased significantly. Is there statistical evidence to support that the average number of users in year 2017-2018 is more than average number of users in year 2015-2016 at a=0.05. Support your answer with all necessary tests.

**Null hypothesis:** The average number of users in year 2017-2018 is **same as** average number of users in year 2015-2016

**Alternative hypothesis:** The average number of users in year 2017-2018 is **more than** average number of users in year 2015-2016

data3 <- data  
data3$dummy <- c(rep("2015-16", 73), rep("2017-18", 50)) # flag variable  
  
test <- t.test(data3$"No of users"~data3$dummy, mu = 0, alternative = "greater", var.eq = TRUE)  
test

##   
## Two Sample t-test  
##   
## data: data3$"No of users" by data3$dummy  
## t = -9.2567, df = 121, p-value = 1  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -154.4945 Inf  
## sample estimates:  
## mean in group 2015-16 mean in group 2017-18   
## 50.06849 181.10000

**Interpretation:**: The critical value for a=0.05 with 121 degrees of freedom is 1.65. Since the calculated statistic value is greater than critical value, we reject the null hypothesis and accept the claim that there is statistically significant difference between the users base of the years 2017-2018 and 2015-2016.

## 4. Farmers use apps to access information throughout the month. Using the data, check whether app usage is same or different across the four weeks of a month. Anand claims that app usage picked up after January 2016; so, test this hypothesis using data from January-2016 - May 2018.

#### Shapiro Wilk’s Test for normality

**Null hypothesis:** Data are normally distributed

**Alternative hypothesis:** data are not normally distributed

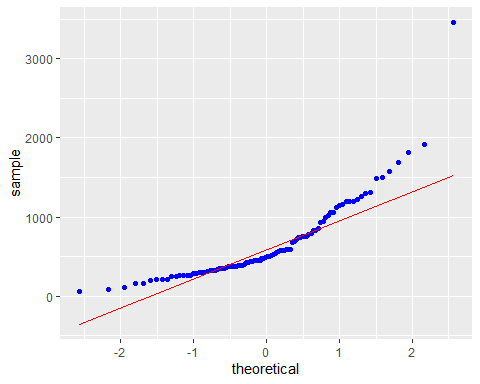
data4 <- data [26:123,]  
  
shapiro.test(data4$Usage)

##   
## Shapiro-Wilk normality test  
##   
## data: data4$Usage  
## W = 0.79955, p-value = 3.178e-10

**Interpretation:** The low p-value makes us reject the null hypothesis and accept the alternative hypothesis that data are not normally distributed.

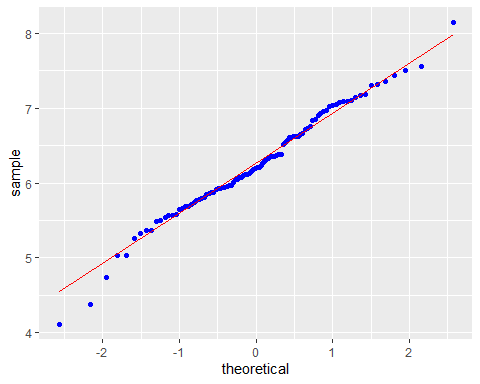
We can confirm the same with a normal probability curve.

ggplot(data4, aes(sample = Usage)) +  
 stat\_qq(col="blue") +  
 stat\_qq\_line(col="red")



We transform the data using log transformation to make it more normally distributed and find that the normal probability curve confirms the same.

## converting the data using log transformation and checking the normality  
  
data4$logusage <- log(data4$Usage)  
  
ggplot(data4, aes(sample = logusage)) +  
 stat\_qq(col="blue") +  
 stat\_qq\_line(col="red")



#### Shapiro Wilk’s Test for normality on log transformation data

**Null hypothesis:** Data are normally distributed

**Alternative hypothesis:** data are not normally distributed

shapiro.test(data4$logusage)

##   
## Shapiro-Wilk normality test  
##   
## data: data4$logusage  
## W = 0.98766, p-value = 0.4984

p-value close to 0.5 makes us accept the null hypothesis that the data are normally distributed.

### One way ANOVA of means of log transformed usage data by week

## using log transformed variables  
anova <- aov (data4$logusage~data4$Week)  
model.tables(anova, type = "means")

## Tables of means  
## Grand mean  
##   
## 6.251738   
##   
## data4$Week   
## Week1 Week2 Week3 Week4  
## 6.227 6.193 6.112 6.466  
## rep 25.000 24.000 24.000 25.000

summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)  
## data4$Week 3 1.72 0.5718 1.166 0.327  
## Residuals 94 46.09 0.4903

**Interpretation:**: One-way ANOVA output shows that p-value = 0.326 > 0.05 and F-stat =1.16, which is less than F critical. Hence, we fail to reject the null hypothesis; and hence we conclude that there is no statistically significant difference in the app usage across various weeks.

## using orginal variable

anova1 <- aov (data4$Usage~data4$Week)  
model.tables(anova1, type = "means")

## Tables of means  
## Grand mean  
##   
## 657.7041   
##   
## data4$Week   
## Week1 Week2 Week3 Week4  
## 622.2 583.8 541.5 875.8  
## rep 25.0 24.0 24.0 25.0

summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)  
## data4$Week 3 1.72 0.5718 1.166 0.327  
## Residuals 94 46.09 0.4903

## 5. Anand claims that number of users have increased over a period of two years. He wants to understand if app usage (number of times his app is accessed in a month by various users) has increased with the increased number of users. Prove this claim statistically. Also suggest a suitable statistical test to prove that the correlation between users and usage is non-zero.

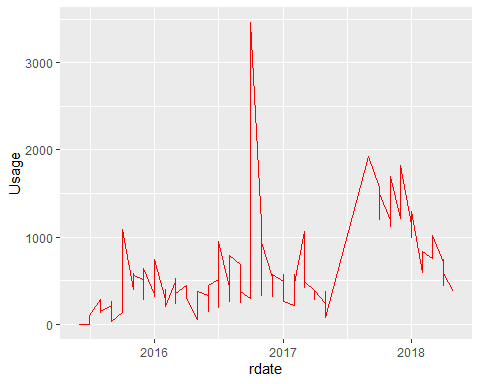
cor <- cor.test(data$"No of users", data$Usage)  
cor

##   
## Pearson's product-moment correlation  
##   
## data: data$"No of users" and data$Usage  
## t = 17.674, df = 121, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7908180 0.8919634  
## sample estimates:  
## cor   
## 0.8489918

**Interpretation:**: This shows a very high correlation between the number of users enrolled and the usage of the app; hence with the increase in the number of users, the usage also increased. Based on the above test, we can prove that users and usage have a correlation.

## 6. A new version of the app was released in August 2016. Anand wants to understand which month in the given time frame after the launch of the new version, the mean usage pattern would start to show a statistically significant shift.

data$rdate <- data$"Month-Year"  
ggplot(data, aes(rdate, Usage)) +  
 geom\_line(group = 1, color = "red")



From the graph, we can observe a spike in the usage from October 2016.

### creating a subset(from October 2016 onwards)

data6 <- data[1:61,]  
data7 <- data[62:123,]  
  
data6$dummy <- "before"  
data7$dummy <- "after"  
newdata1 <- rbind(data6, data7)

**Null hypothesis:** App usage is similar before and after October 2016

**Alternative hypothesis:** App usage is different before and after October 2016

t.test (newdata1$Usage~newdata1$dummy, mu = 0, alternative = "two.sided", conf.level = 0.95, paired = FALSE, var.eq = TRUE)

##   
## Two Sample t-test  
##   
## data: newdata1$Usage by newdata1$dummy  
## t = 6.1002, df = 121, p-value = 1.311e-08  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 315.8774 619.4203  
## sample estimates:  
## mean in group after mean in group before   
## 814.4194 346.7705

**Interpretation:**: The calculated t-statistic is more than the critical value of t; and thus, we reject the null hypothesis and conclude that the app usage is statistically significantly different before and after October 2016. So, after the new release in August 2016, the usage of JAT’s app increased, starting from October 2016.

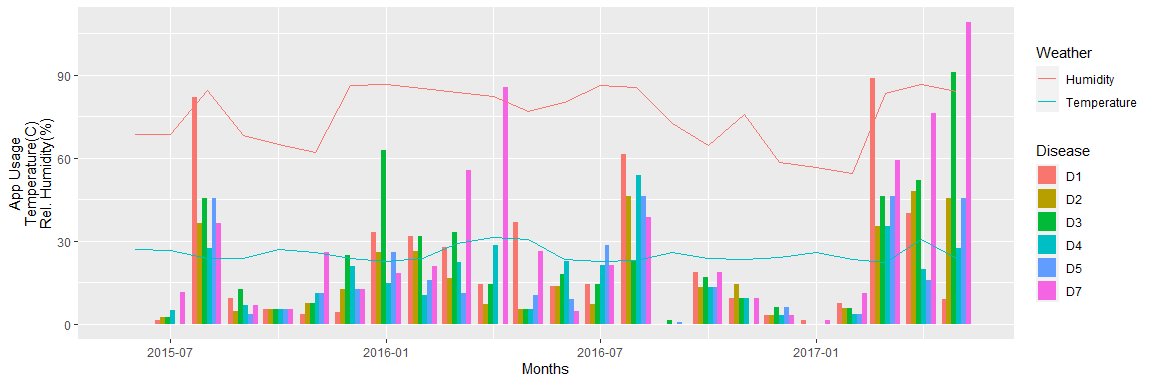
## 7. If a disease is likely to spread in particular weather condition (data given in the disease index sheet), then the access of that disease should be more in the months having suitable weather conditions. Help the analyst in coming up with a statistical test to support the claim for two districts for which the sample of weather and disease access data is provided in the data sheet. Identify the diseases for which you can support this claim. Test this claim both for temperature and relative humidity at 95% confidence.

#options(digits = 2)  
conds <- data.frame(Hmin = c(80, 83, 0, 85, 77, 80),  
 Hmax = c(100, 100, 100, 100, 85, 100),  
 Tmin = c(20, 21.5, 22, 22, 22, 25),  
 Tmax = c(24, 24.5, 24, 26, 24.5, 100),  
 row.names = paste0("D", c(1:5, 7)))  
  
myttest <- function(bdata, temp\_humid) {  
 Hmin <- temp\_humid$Hmin  
 Hmax <- temp\_humid$Hmax  
 Tmin <- temp\_humid$Tmin  
 Tmax <- temp\_humid$Tmax  
   
 bdata1 <- subset(bdata, Temp >= Tmin & Temp <= Tmax & Humid >= Hmin & Humid <= Hmax)  
 bdata2 <- subset(bdata, !(Temp >= Tmin & Temp <= Tmax & Humid >= Hmin & Humid <= Hmax))  
   
 bdata1$weather <- "favourable"  
 bdata2$weather <- "unfavourable"  
 newbdata <- rbind(bdata1, bdata2)  
   
 test1 <- t.test (newbdata$D ~ newbdata$weather, mu = 0, alternative = "two.sided", conf.level = 0.95, paired = FALSE, var.eq = FALSE)  
   
 df <- test1$parameter[[1]]  
 #critical <- abs(qt(0.05, df))  
 critical <- abs(qt(0.05, df))  
 result <- if(test1$p.value < 0.05) "Reject" else "Accept"  
   
 # return the vector of desired values from t-test  
 c(round(c(nrow(bdata1), nrow(bdata2), test1$estimate[[1]], test1$estimate[[2]],   
 sd(bdata1$D), sd(bdata2$D), test1$statistic[[1]], test1$p.value,   
 df, critical), 2), result)  
}  
  
  
resb <- data.frame(matrix(NA, nrow = 11, ncol = 6))  
row.names(resb) <- c('Count of favourable months', 'Count of unfavourable months', 'Mean of favourable months', 'Mean of unfavourable months', 'Std. Dev. of favourable months', 'Std. Dev. of unfavourable months', 't-value', 'p-value', 'Degrees of Freedom', 'Critical Value', "Result (Accept / Reject Null Hypothesis)")  
names(resb) <- paste0("D", c(1:5, 7))  
  
# Belagavi Analysis  
mydata <- data.frame(Temp = bdata$Temperature, Humid = bdata$Humidity)  
  
for (name in names(resb)) {  
 mydata$D <- unlist(bdata[name])  
 resb[, name] <- myttest(mydata, conds[name,])  
}  
  
# Dharwad analysis  
mydata <- data.frame(Temp = ddata$Temperature, Humid = ddata$`Relative Humidity`)  
  
resd <- resb  
for (name in names(resd)) {  
 mydata$D <- unlist(ddata[name])  
 resd[, name] <- myttest(mydata, conds[name,])  
}

myplot <- function(data) {  
  
 data1 <- data[,!(names(data) %in% c("D6", "D8", "D9", "D10", "D11"   
 , "Temperature", "Humidity"  
 ))]  
   
 data2 <- data[, names(data) %in% c("Months", "Temperature", "Humidity")] %>%  
 pivot\_longer(-Months, names\_to = "Weather")  
  
 colors <- c("Temperature" = "Red", "Humidity" = "Green")  
   
 data1 %>%  
 #select(-c(D8, D9, Temperature, Humidity))  
 pivot\_longer(-Months, names\_to = "Disease") %>%  
 ggplot( aes(x = Months)) +  
 geom\_bar(aes(y = value, fill = Disease), position = "dodge", group = 1, stat = "identity", size = 1.2) +  
 geom\_line(data = data2, aes(Months, y = value, group = Weather, color = Weather)) +  
 labs(y = "App Usage\nTemperature(C)\nRel. Humidity(%)")  
}

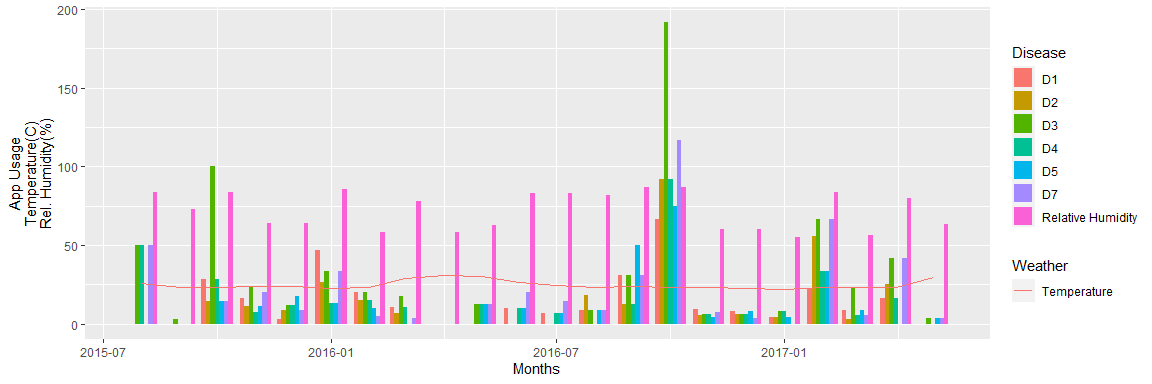
### Belagavi district app usage and weather by month

myplot(bdata)



### Dharwad district app usage and weather by month

myplot(ddata)



The plots show spikes in the access usage for various diseases. Even though weather conditions are shown, the plot does not indicate the favorable weather conditions for various diseases. Need to enhance or supplement the plot. Reading the graph with the Disease\_index data sheet provided is a little tedious.

**Null hypothesis:** Mean of usage is same in favourable and unfavourable weather conditions.

**Alternative hypothesis:** Mean of usage is not same in favourable and unfavourable weather conditions.

The claim that the access in the months having suitable weather conditions is true for the diseases for which we can reject the null hypothesis.

### Belagavi district weather and disease Analysis

knitr::kable(resb, caption = "Belagavi Weather and Disease Analysis")

Belagavi Weather and Disease Analysis

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | D1 | D2 | D3 | D4 | D5 | D7 |
| Count of favourable months | 9 | 8 | 13 | 5 | 4 | 3 |
| Count of unfavourable months | 15 | 16 | 11 | 19 | 20 | 21 |
| Mean of favourable months | 37.59 | 29.38 | 30.96 | 24.29 | 36.57 | 72.42 |
| Mean of unfavourable months | 11.92 | 9.17 | 11.61 | 12.97 | 10.52 | 21.01 |
| Std. Dev. of favourable months | 32.14 | 14.26 | 24.56 | 17.12 | 18.33 | 15.39 |
| Std. Dev. of unfavourable months | 13.21 | 11.63 | 16.41 | 11.29 | 11.91 | 25.02 |
| t-value | 2.28 | 3.47 | 2.3 | 1.4 | 2.73 | 4.93 |
| p-value | 0.05 | 0 | 0.03 | 0.22 | 0.06 | 0.01 |
| Degrees of Freedom | 9.65 | 11.83 | 20.99 | 4.95 | 3.52 | 3.74 |
| Critical Value | 1.82 | 1.78 | 1.72 | 2.02 | 2.22 | 2.17 |
| Result (Accept / Reject Null Hypothesis) | Reject | Reject | Reject | Accept | Accept | Reject |

**Interpretation:** For Belagavi district, we can reject the null hypothesis for **D1, D2, D3 and D7**. For D5, the p-value is 0.06 which makes us accept the null hypothesis, however, the t-value 2.73 is more than the critical value 2.22. Thus, we can accept the claim for **D5** as well but at 94% confidence.

### Dharwad district weather and disease Analysis

knitr::kable(resd, caption = "Dharwad Weather and Disease Analysis")

Dharwad Weather and Disease Analysis

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | D1 | D2 | D3 | D4 | D5 | D7 |
| Count of favourable months | 7 | 5 | 14 | 3 | 4 | 2 |
| Count of unfavourable months | 15 | 17 | 8 | 19 | 18 | 20 |
| Mean of favourable months | 31.59 | 40.13 | 40.27 | 39.17 | 14.18 | 35 |
| Mean of unfavourable months | 6.52 | 6.1 | 11.96 | 12.11 | 13.07 | 19.91 |
| Std. Dev. of favourable months | 19.51 | 33.57 | 51.15 | 45.47 | 14.07 | 21.21 |
| Std. Dev. of unfavourable months | 6.33 | 7.46 | 16.83 | 12.79 | 19.18 | 28.32 |
| t-value | 3.32 | 2.25 | 1.9 | 1.02 | 0.13 | 0.93 |
| p-value | 0.01 | 0.09 | 0.07 | 0.41 | 0.9 | 0.49 |
| Degrees of Freedom | 6.6 | 4.12 | 17.24 | 2.05 | 5.82 | 1.39 |
| Critical Value | 1.91 | 2.11 | 1.74 | 2.87 | 1.95 | 4.03 |
| Result (Accept / Reject Null Hypothesis) | Reject | Accept | Accept | Accept | Accept | Accept |

**Interpretation:** For Dharwad district, we can reject the null hypothesis for **D1 only**. For D2 and D3, the p-values are 0.09 and 0.07 respectively. This we accept the null hypothesis at 95% confidence. However the t-values for **D2** (2.25 > 2.11) and **D3** (1.9 > 1.74) are greater than critical values, thus we may reject the null hypothesis at lower confidence of 91% and 93% respectively..