Smoking and Death Rates Analysis

510: Analytics II: Principles & Applications

Deepti Gupta

Madhu Shri Rajagopalan

Pallava Arasu Pari

November 15, 2018

## STAGE 1

Loading and Pre processing of the data

help(setwd)

## starting httpd help server ... done

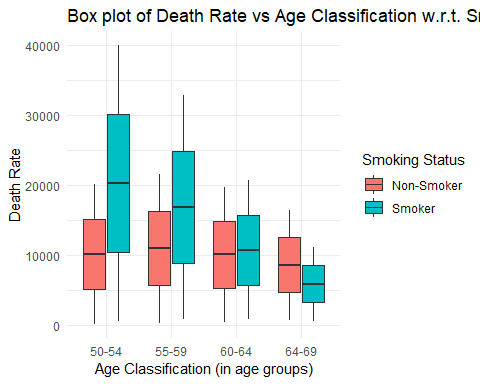
setwd("C:/Study/510-PrinciplesAndApplicationsII/Project")  
  
#install.packages("readxl")  
#install.packages("ggpubr")  
#install.packages("car")  
#install.packages("moments")  
#install.packages("ggplot2")  
#install.packages("FrF2")  
#install.packages("nlme")  
  
library(readxl)  
SmkData <- read\_excel("smoking.xlsx")  
  
summary(SmkData)  
str(SmkData)  
  
colnames(SmkData) <- c("Order", "SmokingStatus", "AgeClassification", "DeathStatus", "NoOfCases" )  
  
SmkData$SmokingStatus <- as.factor(SmkData$SmokingStatus)  
SmkData$AgeClassification <- as.factor(SmkData$AgeClassification)  
SmkData$DeathStatus <- as.factor(SmkData$DeathStatus)  
  
# For death status   
SmkData\_DS0 = data.frame(  
 AgeClassification = c(1,2,3,4,1,2,3,4),  
 SmokingStatus= as.factor(c(0,0,0,0,1,1,1,1)),   
 NoOfCases = c(20132,21671,19790,16499,39990,32894,20739,11197)  
)  
  
SmkData\_DS1 = data.frame(  
 AgeClassification = c(1,2,3,4,1,2,3,4),  
 SmokingStatus= as.factor(c(0,0,0,0,1,1,1,1)),   
 NoOfCases = c(204,394,488,766,647,857,855,643)  
)  
  
  
range01 <- function(x){(x-min(x))/(max(x)-min(x))}  
SmkData\_DS1$NoOfCases\_Scaled = range01(SmkData\_DS1$NoOfCases)

Data exploration

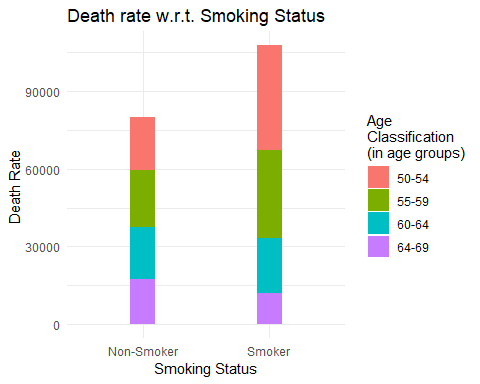
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.1

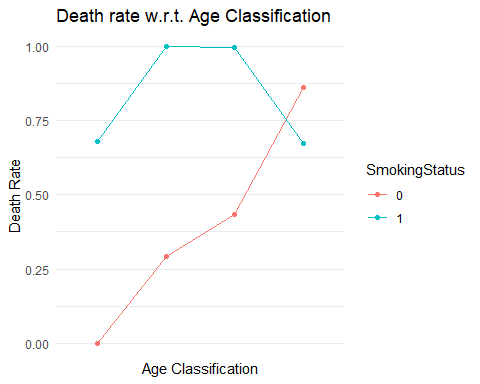
ggSmkData1 <- ggplot(SmkData, aes(x=AgeClassification, y=NoOfCases,fill = SmokingStatus))  
ggSmkData1 + geom\_boxplot(show.legend=TRUE) +  
 scale\_x\_discrete(labels = c("50-54","55-59","60-64", "64-69")) +  
 scale\_fill\_discrete(name ="Smoking Status", breaks = c(0,1), labels=c("Non-Smoker", "Smoker")) +  
 theme\_minimal() +  
 ggtitle("Box plot of Death Rate vs Age Classification w.r.t. Smoking Status") +  
 ylab("Death Rate") +  
 xlab("Age Classification (in age groups)")



ggSmkData2 <- ggplot(SmkData, aes(x=SmokingStatus, y = NoOfCases, fill = AgeClassification) )   
ggSmkData2 + geom\_bar(stat = "identity", width = 0.2)+   
 theme\_minimal() +   
 scale\_x\_discrete( labels = c("Non-Smoker","Smoker")) +  
 scale\_fill\_discrete(name ="Age\nClassification\n(in age groups)", breaks = c(1,2,3,4),   
 labels=c("50-54","55-59","60-64", "64-69")) +  
   
 ggtitle("Death rate w.r.t. Smoking Status") +   
 ylab("Death Rate ") +   
 xlab("Smoking Status")



ggSmkData3 <- ggplot(SmkData\_DS1, aes(x= AgeClassification, y = NoOfCases\_Scaled, color = SmokingStatus, group = as.numeric(SmokingStatus)))   
ggSmkData3 + geom\_line(stat = "identity") + geom\_point() +  
 scale\_x\_discrete( labels = c("50-54","55-59","60-64", "64-69")) +  
 theme\_minimal() +  
 ggtitle("Death rate w.r.t. Age Classification") +   
 ylab("Death Rate ") +   
 xlab("Age Classification")



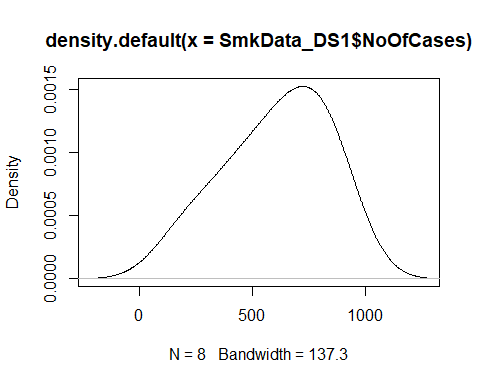
## STAGE 2

Anova and regressions

# Normality test  
shapiro.test(SmkData\_DS1$NoOfCases) #p-value = 0.5152

##   
## Shapiro-Wilk normality test  
##   
## data: SmkData\_DS1$NoOfCases  
## W = 0.92991, p-value = 0.5152

# Concludes that the distribution is normal  
  
  
# Density plot  
plot(density(SmkData\_DS1$NoOfCases))



# Skew check  
library(moments)  
agostino.test(SmkData\_DS1$NoOfCases) #p-value = 0.389

##   
## D'Agostino skewness test  
##   
## data: SmkData\_DS1$NoOfCases  
## skew = -0.51308, z = -0.86134, p-value = 0.389  
## alternative hypothesis: data have a skewness

# which means non-significant amount of skew  
  
  
# To check the assumption that variances are equal across groups or samples  
bartlett.test(SmkData\_DS1$NoOfCases, SmkData\_DS1$AgeClassification) #p-value = 0.7865

##   
## Bartlett test of homogeneity of variances  
##   
## data: SmkData\_DS1$NoOfCases and SmkData\_DS1$AgeClassification  
## Bartlett's K-squared = 1.061, df = 3, p-value = 0.7865

bartlett.test(SmkData\_DS1$NoOfCases, SmkData\_DS1$SmokingStatus) #p-value = 0.3112

##   
## Bartlett test of homogeneity of variances  
##   
## data: SmkData\_DS1$NoOfCases and SmkData\_DS1$SmokingStatus  
## Bartlett's K-squared = 1.0255, df = 1, p-value = 0.3112

# it concludes that the variances between the two different populations do not differ significantly, good to proceed  
  
  
  
  
# Anova Model  
  
library(car)

## Warning: package 'car' was built under R version 3.5.1

## Loading required package: carData

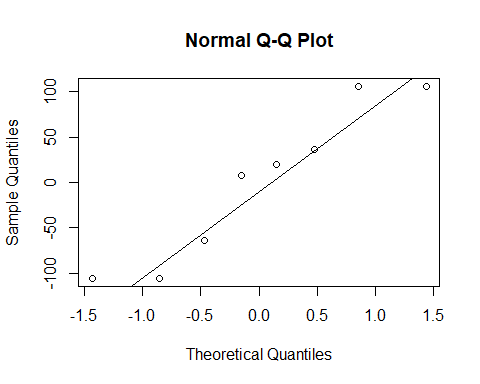
model <- aov(NoOfCases ~ AgeClassification \* SmokingStatus,data = SmkData\_DS1)  
summary(model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## AgeClassification 1 77969 77969 6.191 0.0676 .  
## SmokingStatus 1 165313 165313 13.126 0.0223 \*  
## AgeClassification:SmokingStatus 1 80461 80461 6.389 0.0648 .  
## Residuals 4 50377 12594   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## the model suggests that age classification is not signicant in predicting the death rate, the   
  
  
model\_lm <- lm(NoOfCases ~ AgeClassification \* SmokingStatus,data = SmkData\_DS1)  
summary(model\_lm)

##   
## Call:  
## lm(formula = NoOfCases ~ AgeClassification \* SmokingStatus, data = SmkData\_DS1)  
##   
## Residuals:  
## 1 2 3 4 5 6 7 8   
## 8.0 20.0 -64.0 36.0 -105.6 105.8 105.2 -105.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 18.00 137.45 0.131 0.9021   
## AgeClassification 178.00 50.19 3.547 0.0239 \*  
## SmokingStatus1 736.00 194.38 3.786 0.0193 \*  
## AgeClassification:SmokingStatus1 -179.40 70.98 -2.528 0.0648 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 112.2 on 4 degrees of freedom  
## Multiple R-squared: 0.8653, Adjusted R-squared: 0.7644   
## F-statistic: 8.568 on 3 and 4 DF, p-value: 0.03243

SumofSquareError <- sum( model\_lm$resid^2 )  
  
#Residual analysis  
qqnorm(model$residuals)  
qqline(model$residuals)



shapiro.test(model$residuals) #p-value = 0.2781

##   
## Shapiro-Wilk normality test  
##   
## data: model$residuals  
## W = 0.89815, p-value = 0.2781

# Residuals are normally distributed

## STAGE 3

Elective Analysis

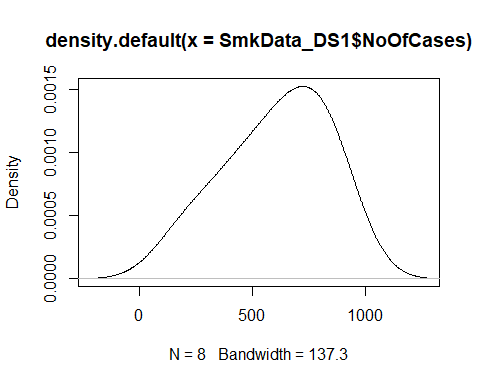
# ################################ TEST 1 ##########################################  
# CHI SQUARE  
# The chi-square test evaluates whether the two variables are independent or not  
# Test the hypothesis whether the death rate of men who smoke or not is independent on their age classification at .05 significance level.  
  
SmkData\_DS1\_ChiSq = data.frame( # With Death Rate = 1  
 row.names = c(1,2,3,4),  
 SmokingStatus0= c(204,394,488,766),   
 SmokingStatus1= c(647,857,855,643)  
)  
  
chisq.test(SmkData\_DS1\_ChiSq)

##   
## Pearson's Chi-squared test  
##   
## data: SmkData\_DS1\_ChiSq  
## X-squared = 254.84, df = 3, p-value < 2.2e-16

# Conclusion  
# p-value < 2.2e-16  
# As the p-value is less than the .05 significance level, we reject the null hypothesis and conclude that death rate of men who smoke or not is dependent on their age classification  
  
  
################################# TEST 2 ##########################################  
# Z-score  
  
View(SmkData\_DS1)  
   
mean\_SmkData\_DS1 = mean(SmkData\_DS1$NoOfCases)  
sd\_SmkData\_DS1 = sd(SmkData\_DS1$NoOfCases)  
  
z\_SmkData\_DS1 = round((SmkData\_DS1$NoOfCases - mean\_SmkData\_DS1) / sd\_SmkData\_DS1, 2)  
  
  
# Help text - Age classification (1=50-54, 2=55-59, 3=60-64, 4=65-69)  
# the z-score gives us the following interpretations  
  
# For people who smoked and lie in the first and youngest age group, 50-54 years have a death rate 0.17 standard deviations more than the average   
  
# For people who smoked and lie in the age group, 55-64 years have the highest death rate 1.08 and 1.07 standard deviations more than the average   
  
# For people who smoked and lie in the eldest age group, 65-69 years have the lowest death rate 0.16 standard deviations more than the average   
  
  
  
################################# TEST 3 ##########################################  
# distributions (Normal, not normal)  
  
# Normality test  
shapiro.test(SmkData\_DS1$NoOfCases)

##   
## Shapiro-Wilk normality test  
##   
## data: SmkData\_DS1$NoOfCases  
## W = 0.92991, p-value = 0.5152

# Concludes that the distribution is normal  
  
# Density plot  
plot(density(SmkData\_DS1$NoOfCases))



# Skew check  
library(moments)  
agostino.test(SmkData\_DS1$NoOfCases) #p-value = 0.389

##   
## D'Agostino skewness test  
##   
## data: SmkData\_DS1$NoOfCases  
## skew = -0.51308, z = -0.86134, p-value = 0.389  
## alternative hypothesis: data have a skewness

# which means non-significant amount of skew  
  
# To check the assumption that variances are equal across groups or samples  
bartlett.test(SmkData\_DS1$NoOfCases, SmkData\_DS1$AgeClassification) #p-value = 0.7865

##   
## Bartlett test of homogeneity of variances  
##   
## data: SmkData\_DS1$NoOfCases and SmkData\_DS1$AgeClassification  
## Bartlett's K-squared = 1.061, df = 3, p-value = 0.7865

bartlett.test(SmkData\_DS1$NoOfCases, SmkData\_DS1$SmokingStatus) #p-value = 0.3112

##   
## Bartlett test of homogeneity of variances  
##   
## data: SmkData\_DS1$NoOfCases and SmkData\_DS1$SmokingStatus  
## Bartlett's K-squared = 1.0255, df = 1, p-value = 0.3112

# it concludes that the variances between the two different populations do not differ significantly, good to proceed  
  
  
  
################################# TEST 4 ##########################################  
# Covariance  
  
cov(SmkData\_DS1$NoOfCases, as.numeric(SmkData\_DS1$SmokingStatus))

## [1] 82.14286

cov(SmkData\_DS1$NoOfCases, as.numeric(SmkData\_DS1$AgeClassification))

## [1] 126.1429

cor(SmkData\_DS1$NoOfCases, as.numeric(SmkData\_DS1$SmokingStatus))

## [1] 0.6647337

cor(SmkData\_DS1$NoOfCases, as.numeric(SmkData\_DS1$AgeClassification))

## [1] 0.4565155

## STAGE 4

F Test

############################ F Test ############################  
var.test(SmkData\_DS1\_ChiSq$SmokingStatus0,   
 SmkData\_DS1\_ChiSq$SmokingStatus1, ratio = 1,   
 alternative = c("two.sided"),  
 conf.level = 0.95)

##   
## F test to compare two variances  
##   
## data: SmkData\_DS1\_ChiSq$SmokingStatus0 and SmkData\_DS1\_ChiSq$SmokingStatus1  
## F = 3.689, num df = 3, denom df = 3, p-value = 0.3121  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.2389392 56.9555394  
## sample estimates:  
## ratio of variances   
## 3.689026

# F-test between 2 normal populations with hypothesis that variances of the 2 populations are equal.  
# Since the p-value = 0.3121, which is much higher than 0.05, the hypothesis that the variances of x and y are equal is accepted.  
  
  
############################ Mean Test ############################  
# Comparing the mean with   
tapply(SmkData\_DS1$NoOfCases, SmkData\_DS1$SmokingStatus, mean)

## 0 1   
## 463.0 750.5

# it suggests that non-smokers have higher chances of surviving than smokers  
  
tapply(SmkData\_DS1$NoOfCases, SmkData\_DS1$AgeClassification, mean)

## 1 2 3 4   
## 425.5 625.5 671.5 704.5

# There is a clear pattern here. It suggests that with increasing age, your chances of surviving decreasing  
  
  
  
############################ Coefficient ############################  
cor.test(SmkData\_DS1$NoOfCases, SmkData\_DS1$AgeClassification,   
 method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: SmkData\_DS1$NoOfCases and SmkData\_DS1$AgeClassification  
## t = 1.2568, df = 6, p-value = 0.2555  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.3658488 0.8785606  
## sample estimates:  
## cor   
## 0.4565155

cor.test(SmkData\_DS1$NoOfCases, as.numeric(SmkData\_DS1$SmokingStatus),   
 method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: SmkData\_DS1$NoOfCases and as.numeric(SmkData\_DS1$SmokingStatus)  
## t = 2.1795, df = 6, p-value = 0.07211  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.07513309 0.93257158  
## sample estimates:  
## cor   
## 0.6647337

############################ P values from z-score ############################  
p\_SmkData\_DS1 = round(2\*pnorm(-abs(z\_SmkData\_DS1)),2)