**Extra Credit Homework Submitted by:** Shivali Dalmia

**Step1.Loading the required libraries and dataset 'who' from tidyr package.**

library(tidyverse)

library(tidyr)

library(ggplot2)

View(who)

Table

Description automatically generated

Dataset definition:

A subset of data from the World Health Organization Global Tuberculosis Report and accompanying global populations with 7,240 rows.

|  |  |
| --- | --- |
| **Column name** | **Description** |
| country | Country Name |
| iso2 | 2-letter ISO country code |
| Iso3 | 3-letter ISO country code |
| Year | Year for which new cases were recorded |
| new\_sp\_m014 -new\_rel\_f65 | Count of TB cases recorded by for different age groups for both males and females.  e.g., new\_sp\_m014  new - stands for new cases  sp - code for method of diagnosis   * rel = relapse * sn = negative pulmonary smear * sp = positive pulmonary smear * ep = extrapulmonary   m - code for gender   * f = female * m = male   014 - code for age group   * 014 = 0-14 years * 1524 = 15-24 years * 2534 = 25 to 34 years * 3544 = 35 to 44 years * 4554 = 45 to 54 years * 5564 = 55 to 64 years * 65 = 65 years or older |

**Step2.Data exploration.**

This analysis is done using a subset for India country.

who\_IN <- filter(who,who$iso2 == "IN")

View(who\_IN) #Viewing new data frame

Table

Description automatically generated

There is total 34 rows for IN as country. Checking the summary statistics.

summary(who\_IN)

Text

Description automatically generated with medium confidence

From summary statistics we can see that there are 30 plus NA values for other diagnosis methods like 'sn','ep' etc except 'sp'. Hence, subsetting data for diagnosis method 'sp' for this analysis.

Text

Description automatically generated

who\_IN\_sp <- select(who\_IN,country,iso2,iso3,year,contains("sp"))

View(who\_IN\_sp)

Table, calendar

Description automatically generated

Visualizing new TB cases for different age groups from 1980-2013.

*# Function for plotting new cases for different age groups*

plotNewTBCases <- function(ageGrp) {

plot(who\_IN\_sp[[paste("new\_sp\_m",ageGrp,sep="")]]~who\_IN\_sp$year,

xlab="Year",

ylab="Count of Cases",

main=paste("TB cases for",ageGrp,"age group"),

type="l",

col="blue")

lines(who\_IN\_sp[[paste("new\_sp\_f",ageGrp,sep="")]]~who\_IN\_sp$year,col="red")

legend("topleft",c("Male","Female"),fill=c("blue","red"))

}

plotNewTBCases("014") #0-14 plotNewTBCases("1524") #15-24

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

plotNewTBCases("2534") #25-34 plotNewTBCases ("3544") #35-44

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

plotNewTBCases("4554") #45-54 plotNewTBCases("5564") #55-64

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

plotNewTBCases("65") #65 and above

Chart, line chart

Description automatically generated

Observations:

1.All the plots show visible trend from year 1995 to 2015 as from the year 1980 to 1994 the values are NA. Verifying this by calculating summary stats for years 1980-1994.

who\_IN\_SP\_1980\_1994 <- filter(who\_IN\_sp,between(year,1980,1994))

A picture containing text, indoor

Description automatically generatedView(who\_IN\_SP\_1980\_1994)

Table

Description automatically generatedsummary(who\_IN\_SP\_1980\_1994)

From summary we can see that all columns from new\_sp\_m014 to new\_sp\_f65 have NA values. Hence we remove these rows before performing the further analysis.

who\_mod <- filter(who\_IN\_sp,between(year,1995,2012))

View(who\_mod)

A picture containing text, wall, indoor

Description automatically generated

2.From the line plots for all age groups in Step1, we can say that as from 1995 to 2005 count of new cases increases for both males and females and then slight dip is observed.

Chart, line chart

Description automatically generated3. From the plot for 0-14 age group, it looks that the number of cases for females are higher than males over the years. Let's calculate the average number of new cases for both males and females.

colnames(who\_mod)

Text

Description automatically generatedsummary(select(who\_mod,new\_sp\_m014,new\_sp\_f014))

From summary stats we can see that the average number of new cases for females is 4625 which are nearly twice of the number of cases for males i.e., 2560.

4.Plot the average count of cases(average of cases from 1994-2012) for males across different age groups.

#Creating a new data frame with average number of cases for both males and females across all age groups

mean\_males <- c(mean(who\_mod$new\_sp\_m014,na.rm = TRUE),mean(who\_mod$new\_sp\_m1524,na.rm = TRUE),mean(who\_mod$new\_sp\_m2534,na.rm = TRUE),mean(who\_mod$new\_sp\_m3544,na.rm = TRUE),mean(who\_mod$new\_sp\_m4554,na.rm = TRUE),mean(who\_mod$new\_sp\_m5564,na.rm = TRUE),mean(who\_mod$new\_sp\_m65,na.rm = TRUE))

mean\_females <- c(mean(who\_mod$new\_sp\_f014,na.rm = TRUE),mean(who\_mod$new\_sp\_f1524,na.rm = TRUE),mean(who\_mod$new\_sp\_f2534,na.rm = TRUE),mean(who\_mod$new\_sp\_f3544,na.rm = TRUE),mean(who\_mod$new\_sp\_f4554,na.rm = TRUE),mean(who\_mod$new\_sp\_f5564,na.rm = TRUE),mean(who\_mod$new\_sp\_f65,na.rm = TRUE))

sd\_males <-c(sd(who\_mod$new\_sp\_m014,na.rm = TRUE),sd(who\_mod$new\_sp\_m1524,na.rm = TRUE),sd(who\_mod$new\_sp\_m2534,na.rm = TRUE),sd(who\_mod$new\_sp\_m3544,na.rm = TRUE),sd(who\_mod$new\_sp\_m4554,na.rm = TRUE),sd(who\_mod$new\_sp\_m5564,na.rm = TRUE),sd(who\_mod$new\_sp\_m65,na.rm = TRUE))

sd\_females <- c(sd(who\_mod$new\_sp\_f014,na.rm = TRUE),sd(who\_mod$new\_sp\_f1524,na.rm = TRUE),sd(who\_mod$new\_sp\_f2534,na.rm = TRUE),sd(who\_mod$new\_sp\_f3544,na.rm = TRUE),sd(who\_mod$new\_sp\_f4554,na.rm = TRUE),sd(who\_mod$new\_sp\_f5564,na.rm = TRUE),sd(who\_mod$new\_sp\_f65,na.rm = TRUE))

n\_males <- c(18,18,18,18,18,18,18)

n\_females <- c(18,18,18,18,18,18,18)

who\_avg\_mf <- data.frame("Age group" = c("0-14","15-24","25-34","35-44","45-54","55-64","65+"),

"Mean\_cases\_males" = mean\_males,

"Mean\_cases\_females" = mean\_females,

"SD\_cases\_males" = sd\_males,

"SD\_cases\_females" = sd\_females,

"N\_males" = n\_males,

"N\_females" = n\_females)

Table

Description automatically generatedView(who\_avg\_mf)

Visualizing average count of cases across all age groups for males.

boxplot(who\_avg\_mf$Mean\_cases\_males~who\_avg\_mf$Age.group,ylab="Count of cases",

xlab="Age groups",main="Males: Average cases v/s Age groups")

Chart, box and whisker chart

Description automatically generatedConclusion: From the above plots we can say that average number of cases vary for males across different age groups.

We can test this using hypothesis testing.

Visualizing average count of cases across all age groups for females.

boxplot(who\_avg\_mf$Mean\_cases\_females~who\_avg\_mf$Age.group,ylab="Count of cases",

xlab="Age groups",main="Females: Average cases v/s Age groups")

Chart, box and whisker chart

Description automatically generated

Conclusion: From the above plots we can say that average number of cases vary for females as well across different age groups.

We can test this claim using hypothesis testing.

**Step3.Questions to be tested using Hypothesis testing** (Reasoning explained in Step3)

Question1: The average number of new TB cases are different for different age groups for males.

Question2: The average number of new TB cases are different for different age groups for females.

Test the claim using ANOVA.

**Step4.Hypothesis testing using ANNOVA.**

**Testing claim:** The average number of new TB cases are different for different age groups for males

1. Question1: The average number of new TB cases are different for different age groups for males. Test using ANOVA test statistics.
2. Reason: As seen from the boxplot in step2 point 4 for males we see considerable difference in average number of new cases for different age groups. Hence, decided to confirm this claim using hypothesis testing.
3. Type of test: ANOVA analysis using F statistics.
4. Checking model assumptions

1.Independence: In this study the diagnosis of TB in single individual is independent of another individual. Hence observations are independent.

2.Normal approximation: By plotting the normal qqplot and histogram we can verify normality of data for all age groups.

3.Constant Variance: By plotting boxplots we can check for constant variance across all age groups.

Chart, histogram

Description automatically generated#0-14

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_m014,main="Noramlity check")

qqline(who\_mod$new\_sp\_m014,col = "blue")

hist(who\_mod$new\_sp\_m014,main="Skew check")

boxplot(who\_mod$new\_sp\_m014,main="Constant Variance Check")

Chart, histogram

Description automatically generated

#15-24

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_m1524,main="Noramlity check")

qqline(who\_mod$new\_sp\_m1524,col="blue")

hist(who\_mod$new\_sp\_m1524,main="Skew check")

boxplot(who\_mod$new\_sp\_m1524,main="Constant Variance Check")

Chart, histogram

Description automatically generated#25-34

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_m2534, main="Noramlity check")

qqline(who\_mod$new\_sp\_m2534, col = "blue")

hist(who\_mod$new\_sp\_m2534,main="Skew check")

boxplot(who\_mod$new\_sp\_m2534, main="Constant Variance Check" )

Chart, histogram

Description automatically generated

#35-44

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_m3544,main="Noramlity check")

qqline(who\_mod$new\_sp\_m3544, col = "blue")

hist(who\_mod$new\_sp\_m3544,main="Skew check")

boxplot(who\_mod$new\_sp\_m3544,main="Constant Variance Check")

Chart, histogram

Description automatically generated#45-54

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_m4554,main="Noramlity check")

qqline(who\_mod$new\_sp\_m4554, col = "blue")

hist(who\_mod$new\_sp\_m4554,main="Skew check")

boxplot(who\_mod$new\_sp\_m4554,main="Constant Variance Check")

Chart, histogram

Description automatically generated#55-64

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_m5564,main="Noramlity check")

qqline(who\_mod$new\_sp\_m5564,col = "blue")

hist(who\_mod$new\_sp\_m5564,main="Skew check")

boxplot(who\_mod$new\_sp\_m5564,main="Constant Variance Check")

Chart, histogram

Description automatically generated#65+

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_m65,main="Noramlity check")

qqline(who\_mod$new\_sp\_m65, col = "blue")

hist(who\_mod$new\_sp\_m65,main="Skew check")

boxplot(who\_mod$new\_sp\_m65,main="Constant Variance Check")

Conclusion: From all normality plots we can see a normal trend, histogram doesn’t show signs of strong skewness. Finally, from boxplots we can conclude constant variance.

1. Null and alternate hypothesis

H0: µm\_014 = µm\_1524 = µm\_2534 = µm\_3544 = µm\_4554 = µm\_5564 = µm\_65

HA: At least average number of cases is different for one age group for males.

1. Calculating f statistics. (Using data frame ‘who\_avg\_mf’)

Note: I have not used R’s inbuilt annova() to calculate values because dataset was not in the correct format for lm() model. So, calculated them using math formulas and then verified the result with online ANNOVA calculator (<https://goodcalculators.com/one-way-anova-calculator/>).

# Total sample size

(n<- sum(who\_avg\_mf$N\_males))

>126

# Total groups

(k <- length(who\_avg\_mf$Age.group))

>7

# Degree of freedom for groups, error, and total.

(dfg <- k-1)

>6

(dft <- n-1)

>125

(dfe <- dft-dfg)

> 119

#Determine mean of mean number of cases for all groups

t\_mean <- mean(who\_avg\_mf$Mean\_cases\_males)

> 35672.94

#Appending number of cases for all age groups for males

(no\_of\_sp\_m <- (c(who\_mod$new\_sp\_m014,who\_mod$new\_sp\_m1524,who\_mod$new\_sp\_m2534,

who\_mod$new\_sp\_m3544,who\_mod$new\_sp\_m4554,who\_mod$new\_sp\_m5564,

who\_mod$new\_sp\_m65)))

# Sum of squares total

(SST <- sum((no\_of\_sp\_m-t\_mean)^2))

> 133474512097

# Sum of squares between groups

(SSG <- sum(who\_avg\_mf$N\_males \* (who\_avg\_mf$Mean\_cases\_males - t\_mean)^2))

> 37615696704

# Sum of squares error

(SSE <- SST - SSG)

> 95858815393

# Mean square error

(MSE <- SSE/dfe)

> 805536264

# Mean square for groups

(MSG <- SSG/dfg)

> 6269282784

# F-statistics

(F <- MSG/MSE)

> 7.782744

# p-value

(round(pf(F,dfg,dfe,lower.tail = FALSE))) #Rounding off value 4.487155e-07 to 0

>0

1. Conclusion:

Table

Description automatically generated

(Table from online ANNOVA calculator: <https://goodcalculators.com/one-way-anova-calculator/>)

The p-value is 0 which is less than significance level of 0.05. Hence, we reject the null hypothesis.

We have enough evidence to say that average number of new cases for males are different for at least one of the age groups.

**Testing claim:** The average number of new TB cases are different for different age groups for females

1. Question1: The average number of new TB cases are different for different age groups for females. Test using ANOVA test statistics.
2. Reason: As seen from the boxplot in step2 point 4 for females we see considerable difference in average number of new cases for different age groups. Hence, decided to confirm this claim using hypothesis testing.
3. Type of test: ANOVA analysis using F statistics.
4. Checking model assumptions

1.Independence: In this study the diagnosis of TB in single individual is independent of another individual. Hence observations are independent.

2.Normal approximation: By plotting the normal qqplot and histogram we can verify normality of data for all age groups.

3.Constant Variance: By plotting boxplots we can check for constant variance across all age groups.

Chart, histogram

Description automatically generated#0-14

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_f014,main="Noramlity check")

qqline(who\_mod$new\_sp\_f014,col="blue")

hist(who\_mod$new\_sp\_f014,main="Skew check")

boxplot(who\_mod$new\_sp\_f014,main="Constant Variance Check")

Chart, histogram

Description automatically generated

#15-24

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_f1524,main="Noramlity check")

qqline(who\_mod$new\_sp\_f1524,col="blue")

hist(who\_mod$new\_sp\_f1524,main="Skew check")

boxplot(who\_mod$new\_sp\_f1524,main="Constant Variance Check")

Chart, histogram

Description automatically generated

#25-34

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_f2534,main="Noramlity check")

qqline(who\_mod$new\_sp\_f2534,col = "blue")

hist(who\_mod$new\_sp\_f2534,main="Skew check")

boxplot(who\_mod$new\_sp\_f2534,main="Constant Variance Check" )

Chart, histogram

Description automatically generated#35-44

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_f3544,main="Noramlity check")

qqline(who\_mod$new\_sp\_f3544, col = "blue")

hist(who\_mod$new\_sp\_f3544,main="Skew check")

boxplot(who\_mod$new\_sp\_f3544,main="Constant Variance Check")

Chart, histogram

Description automatically generated#45-54

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_f4554,main="Noramlity check")

qqline(who\_mod$new\_sp\_4554, col = "blue")

hist(who\_mod$new\_sp\_f4554,main="Skew check")

boxplot(who\_mod$new\_sp\_f4554,main="Constant Variance Check")

Chart, histogram

Description automatically generated#55-64

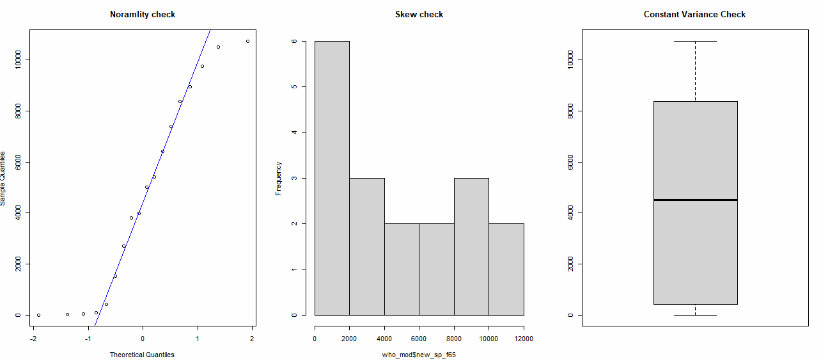
par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_f5564,main="Noramlity check")

qqline(who\_mod$new\_sp\_f5564,col = "blue")

hist(who\_mod$new\_sp\_f5564,main="Skew check")

boxplot(who\_mod$new\_sp\_f5564,main="Constant Variance Check")



#65+

par(mfrow=c(1,3))

qqnorm(who\_mod$new\_sp\_f65,main="Noramlity check")

qqline(who\_mod$new\_sp\_f65, col = "blue")

hist(who\_mod$new\_sp\_f65,main="Skew check")

boxplot(who\_mod$new\_sp\_f65,main="Constant Variance Check")

Conclusion: From all normality plots we can see a normal trend, histogram doesn’t show signs of strong skewness. Finally, from boxplots we can conclude constant variance.

1. Null and alternate hypothesis

H0: µf\_014 = µf\_1524 = µf\_2534 = µf\_3544 = µf\_4554 = µf\_5564 = µf\_65

HA: At least average number of cases is different for one age group for females.

1. Calculating f statistics. (Using data frame ‘who\_avg\_mf’)

Note: I have not used R’s inbuilt annova() to calculate values because dataset was not in the correct format for lm() model. So, calculated them using math formulas and then verified the result with online ANNOVA calculator (<https://goodcalculators.com/one-way-anova-calculator/>).

# Total sample size

(n<- sum(who\_avg\_mf$N\_females))

>126

# Total groups

(k <- length(who\_avg\_mf$Age.group))

>7

# Degree of freedom for groups, error, and total.

(dfg <- k-1)

>6

(dft <- n-1)

>125

(dfe <- dft-dfg)

> 119

#Determine mean of mean number of cases for all groups

t\_mean <- mean(who\_avg\_mf$Mean\_cases\_females)

> 15857.62

#Appending number of cases for all observations

(no\_of\_sp\_f <- (c(who\_mod$new\_sp\_f014,who\_mod$new\_sp\_f1524,who\_mod$new\_sp\_f2534,

who\_mod$new\_sp\_f3544,who\_mod$new\_sp\_f4554,who\_mod$new\_sp\_f5564,

who\_mod$new\_sp\_f65)))

# Sum of squares total

(SST <- sum((no\_of\_sp\_f-t\_mean)^2))

> 35132579340

# Sum of squares between groups

(SSG <- sum(who\_avg\_mf$N\_females \* (who\_avg\_mf$Mean\_cases\_females - t\_mean)^2))

> 13551496438

# Sum of squares error

(SSE <- SST - SSG)

> 21581082902

# Mean square error

(MSE <- SSE/dfe)

> 181353638

# Mean square for groups

(MSG <- SSG/dfg)

> 2258582740

# F-statistics

(F <- MSG/MSE)

> 12.45403

# p-value

(round(pf(F,dfg,dfe,lower.tail = FALSE))) #Rounding off value 7.457776e-11 to 0

>0

1. Conclusion:

Table

Description automatically generated

(Table from online ANNOVA calculator: <https://goodcalculators.com/one-way-anova-calculator/>)

The p-value is 0 which is less than significance level of 0.05. Hence, we reject the null hypothesis.

We have enough evidence to say that average number of new cases for females are different for at least one of the age groups.