Matthew Meeks

STAT 3010

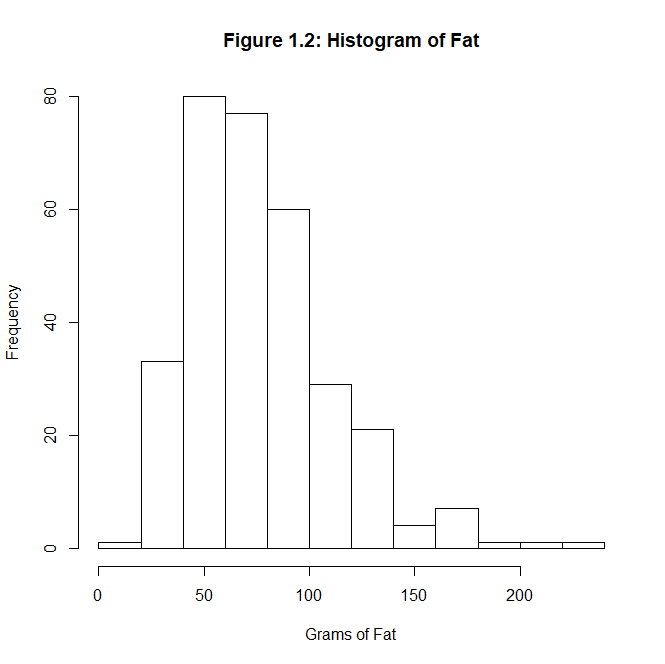
Homework 2

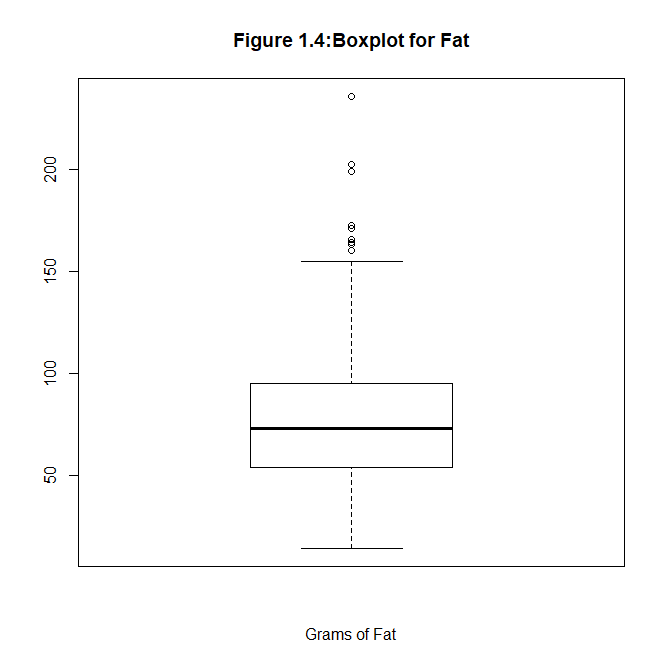
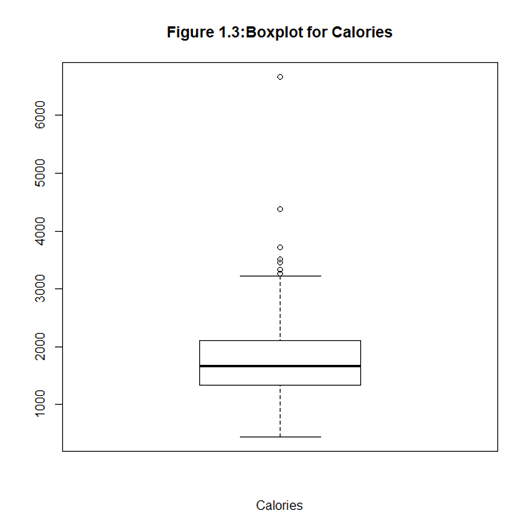
October 7, 2018



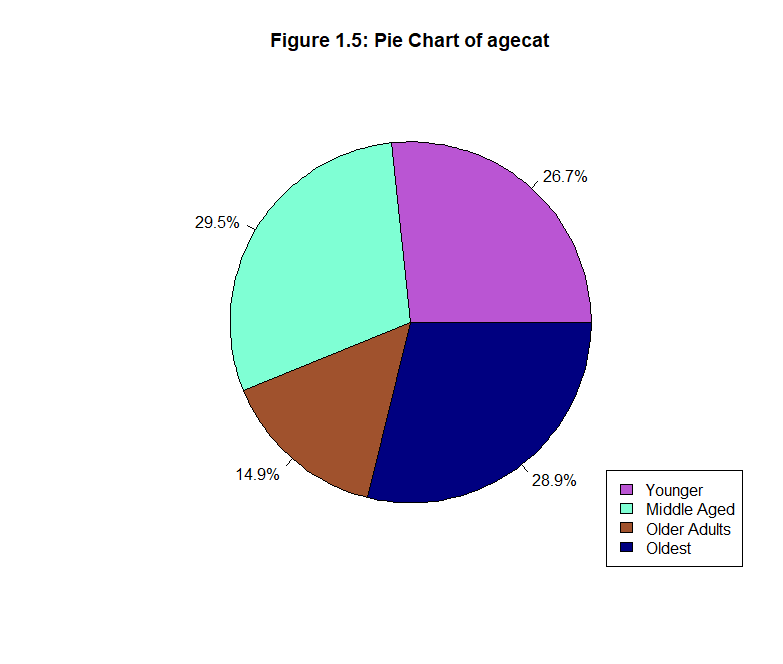


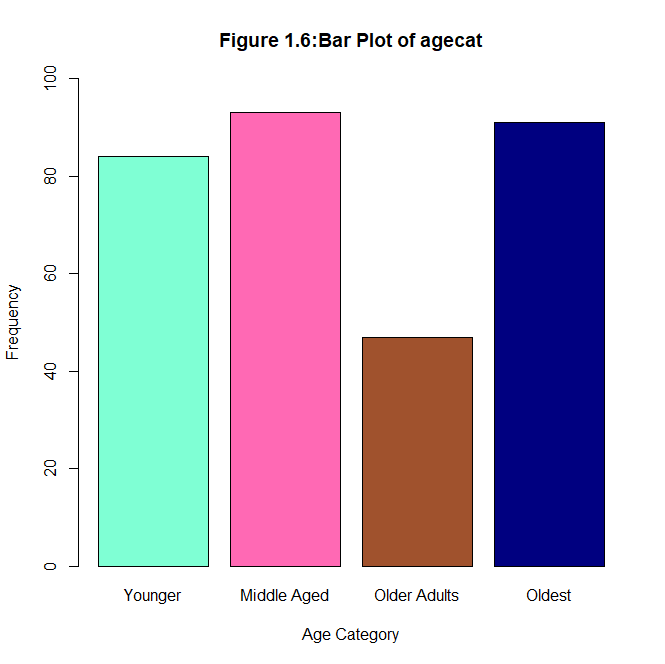






1. Both variables, Fat and Calories, are positively skewed, as seen in the histograms, figures 1.1 and 1.2. The boxplots, figures 1.3 and 1.4, also show that both have whiskers skewed positively. Given this, for Calories the appropriate measurement of central tendency is the median (1666.80), and the appropriate measurement of dispersion is the IQR (762.40). Also, for Fat the appropriate measurement of central tendency is the median (72.90 grams), and the appropriate measurement of dispersion is the IQR (41.30 grams).
2. 





Given that agecat, is a categorical variable, the appropriate measurement to describe the distribution of this variable is the mode. The mode of the variable agecat is Middle Aged (n=93).



Table 1.5 shows the frequency counts of gender by smoking status. Table 1.5 shows that 144 females have never smoked.



Table 1.6 shows the grand total percentages of the entire sample that each stratum represents (e.g., females who never smoked). 45.71% of the entire sample is women who have never smoked.



Table 1.7 shows the row total percentage that each stratum represents, so it shows what percent of each gender (the row variable) is one of the three smoking status types. 52.75% of all females in this sample have never smoked.



Table 1.8 shows the column total percentage that each stratum represents, so it shows what percent of each smoking type (the column variable) is male or female. 91.72% of the entire never smoked group is comprised of females.

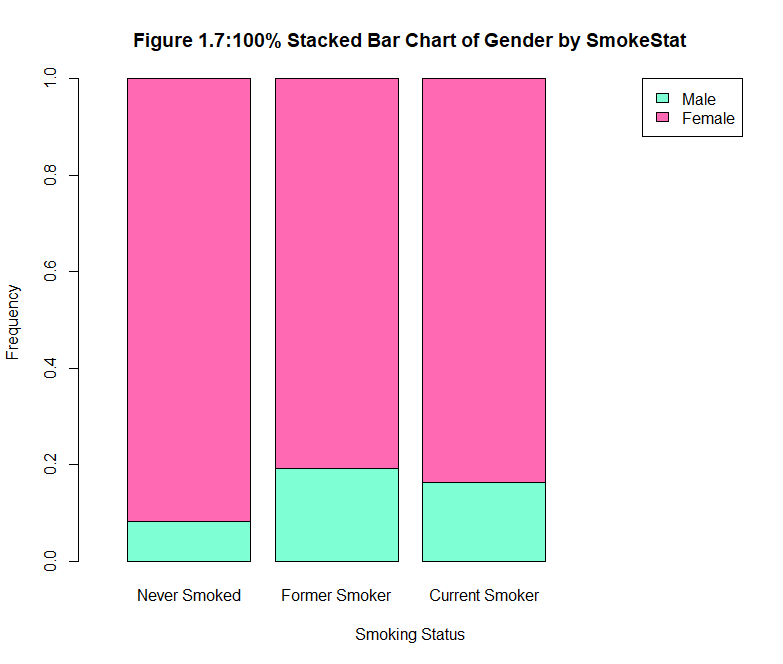


Figure 1.7 shows that females make up the largest portion of each of the smoking status groups. However, the former smoker and current smoker groups should be noted, because males make up close to a fifth of the proportion of each, even though males are only approximately 13% of the entire sample. The data shows a trend of higher smoking rates amongst men, but further analyses would be required for a conclusion.



Table 1.9 shows the 95% confidence interval for a random sample (n=75) of calories.

I hypothesize that age and calories will be negatively correlated, such that younger individuals will have higher calorie counts than older individuals. Age will be the explanatory variable and calories will be the response variable.

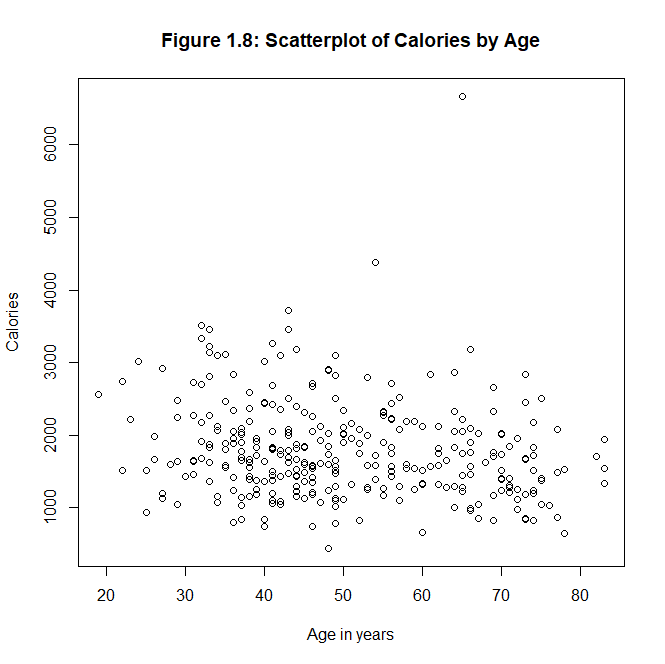


Figure 1.8 is a scatterplot that shows there is a trend in the data. Although there is no line of best fit, such a line would show a slight declining slope. Higher calorie scores are grouped towards the younger age in years side of the x axis. This scatterplot supports the hypothesis.

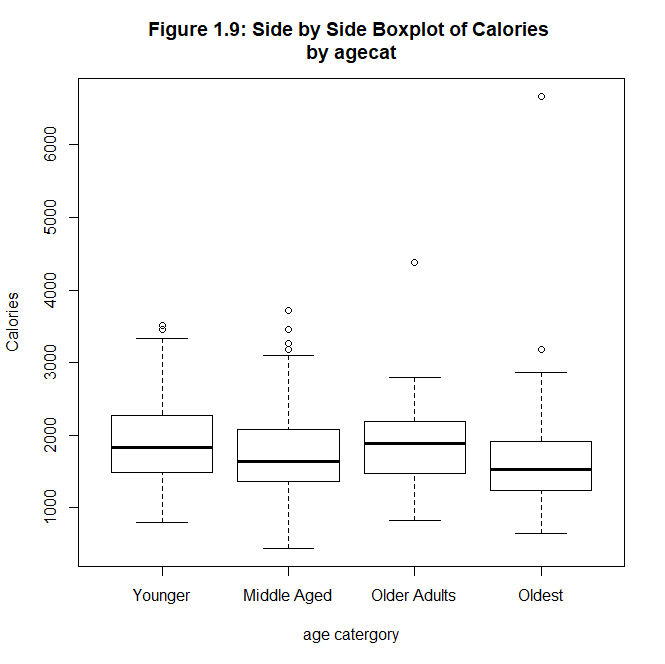


Figure 1.9 shows a side-by-side boxplot of calories by age category. The median of each group is illustrated by the bold line in each boxplot. The younger group has a slightly lower median than the older adults, but its upper quartile (Q3) is the greatest (in calories) than any other age category’s boxplot. This supports the hypothesis still, because the general trend is still that age and calories are negatively correlated. However, it is less supporting than the aforementioned scatterplot.



Table 2.0 gives the results of a stratified analysis of calories by age category. The youngest group has the highest caloric intake, but the Older Adults group is the second highest. There seems to be something unique about this older adults group, such that they go against the trend of declining caloric intake over time (years). The Older Adults group has a smaller sample (n=47) than all the other groups, which all have samples of greater than 80 observations. This smaller sample size may explain why the Older Adults group does not fit the trend the other groups follow.



Table 2.1 shows the stratified confidence intervals of calories by age category at a 95% confidence level. The trend of declining caloric intake over time (years) is still seen here. Once again, the Older Adults goes against the trend. However, it should be noted the relatively large size of the Older Adults group’s margin of error, when compared to the other groups. This large margin of error is likely due to the small sample size from this random sample (n=75). To achieve a 95% confidence level for the Older Adults group’s interval, a large margin of error was needed. This tells me that this data was highly spread, so a larger sample may reveal a truer depiction of the mean calories for the Older Adults group. This still supports the hypothesis, but this support is contingent on further investigation of the Older Adults group.

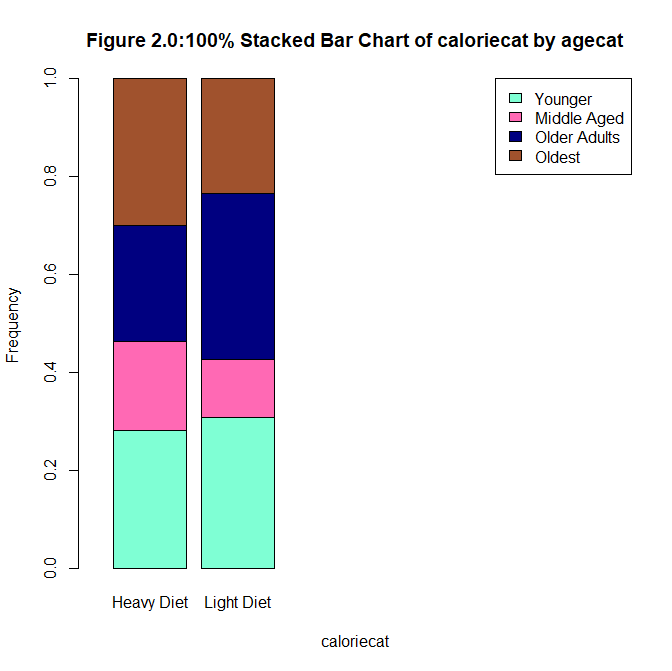


Figure 2.0 is a 100% stacked bar chart of calorie category by age category. The calorie category is binary and was created by splitting calories into “Heavy” or “Light” diets. A light diet is equal to or less than 1700 calories, while a heavy diet is equal to or greater than 1701 calories. Given the mean (1796.7) and median (1666.8) of calories, 1700 was chosen as the cutoff point to achieve two approximately even groups. This 100% stacked bar chart gives a conflicting trend of the data. In this figure, the Older Adults and Younger group have the largest portions of the light diet, while the Oldest and Younger groups represent the largest portions of the heavy diet. This is data goes against the trend of all the other analytic tools, so it is not supportive of the hypothesis. The opposite trend here may be accounted for by small stratums within each diet group.

In conclusion, my hypothesis was significantly supported. 4 out of 5 analytic tools were supportive of the hypothesis. With larger samples, I believe my hypothesis would be supported as statistically significant at a 95% confidence level. The trend of age (in years) being negatively correlated with calories consumed is supported from this analysis.

Appendix

#Importing data set

CancerRisk<- read.csv("F:\\CancerRisk.csv")

#descriptive statistics for quantitative variables, do IQR manually in R or excel and format tables in excel

summary(CancerRisk$Age)

round(summary(CancerRisk$Age),2)

round(sd(CancerRisk$Age),2)

#IQR manually computed

round(summary(CancerRisk$Fat),2)

round(sd(CancerRisk$Fat),2)

#IQR manually computed

round(summary(CancerRisk$Calories),2)

round(sd(CancerRisk$Calories),2)

#IQR manually computed

#creating histograms and boxplots for Calories and Fat

hist(CancerRisk$Calories, main="Figure 1.1: Histogram of Calories", xlab="Calories")

hist(CancerRisk$Fat, main="Figure 1.2: Histogram of Fat", xlab="Grams of Fat")

boxplot(CancerRisk$Calories, main = "Figure 1.3:Boxplot for Calories", xlab = "Calories")

boxplot(CancerRisk$Fat, main = "Figure 1.4:Boxplot for Fat", xlab = "Grams of Fat")

#Creating the categorical variable *agecat*

CancerRisk$agecat[CancerRisk$Age < 40] <- "Younger"

CancerRisk$agecat[CancerRisk$Age >= 40 & CancerRisk$Age <50] <- "Middle Aged"

CancerRisk$agecat[CancerRisk$Age >= 50 & CancerRisk$Age <60] <- "Older Adults"

CancerRisk$agecat[CancerRisk$Age >= 60] <- "Oldest"

table(CancerRisk$agecat)

#reordering the categorical variable agecat

agecat <- ordered(CancerRisk$agecat,c("Younger","Middle Aged","Older Adults","Oldest"))

#creating a pie chart for variable agecat

table(agecat)

agecat\_labels <- round(prop.table(table(agecat))\*100, 1)

agecat\_labels <- paste(agecat\_labels, "%", sep="")

colors <- c("mediumorchid","aquamarine","sienna","navyblue")

pie(table(agecat), main="Figure 1.5: Pie Chart of agecat",col=colors,

labels=agecat\_labels)

legend("bottomright", c("Younger","Middle Aged","Older Adults","Oldest"),fill=colors,bg="white")

#creating an ordered bar chart of agecat

barplot(table(agecat),xlab="Age Category",names.arg=c("Younger","Middle Aged","Older Adults","Oldest"),

main="Figure 1.6:Bar Plot of agecat",col=c("aquamarine","hotpink","sienna","navyblue"),

ylim=c(0,100),ylab="Frequency")

#creating the 4 contingency tables for gender by SmokeStat, format in excel

table(CancerRisk$Gender,CancerRisk$SmokeStat)

prop.table(table(CancerRisk$Gender,CancerRisk$SmokeStat))

prop.table(table(CancerRisk$Gender,CancerRisk$SmokeStat),1)

prop.table(table(CancerRisk$Gender,CancerRisk$SmokeStat),2)

#100% Stacked Bar Chart

barplot(prop.table(table(CancerRisk$Gender,CancerRisk$SmokeStat),2), xlab="Smoking Status",

names.arg=c("Never Smoked","Former Smoker","Current Smoker"), main="Figure 1.7:100% Stacked Bar Chart of Gender by SmokeStat",

col=c("aquamarine","hotpink"),ylab="Frequency",xlim=c(0,5))

legend("topright",c("Male","Female"),fill=c("aquamarine","hotpink"))

#Creating the random sample of 75 observations

set.seed(44522)

samp <- CancerRisk[sample(1:nrow(CancerRisk),75,replace=FALSE),]

#Creating the confidence interval function in the universe of R

CI<-function(x,alpha=0.05){

n<-sum(!is.na(x))

con<-(1-alpha)\*100

me<-qt(1-alpha/2,n-1)\*sd(x,na.rm=T)/sqrt(n)

lclm<-round(mean(x,na.rm=T)-me,digits=3)

uclm<-round(mean(x,na.rm=T)+me,digits=3)

mean<-round(mean(x,na.rm=T),digits=3)

{limits<-data.frame(cbind(variable=deparse(substitute(x)),n,

c.level=con,

mean,me=round(me,digits=3),lclm,uclm))}

print(limits)

rm(n,con,lclm,uclm,mean)}

#Generating a 95% confidence interval from the random sample newly created, format table in excel

CI(samp$Calories,alpha=0.05)

#scatterplot

plot(CancerRisk$Age,CancerRisk$Calories,main="Figure 1.8: Scatterplot of Calories by Age",xlab="Age in years",ylab="Calories")

#side by side boxplot

boxplot(CancerRisk$Calories~agecat,horizontal=F,main="Figure 1.9: Side by Side Boxplot of Calories

by agecat",xlab="age catergory",ylab="Calories")

#stratified analysis

aggregate(CancerRisk$Calories~CancerRisk$agecat,FUN=summary)

#stratified confidence intervals

aggregate(samp$Calories~samp$agecat,FUN=CI)

#making calories a categorical variable

CancerRisk$caloriecat[CancerRisk$Calories <= 1700] <- "light diet"

CancerRisk$caloriecat[CancerRisk$Calories >= 1701] <- "heavy diet"

table(CancerRisk$caloriecat)

#100% stacked bar chart of caloriecat by agecat

Barplot(prop.table(table(CancerRisk$caloriecat,CancerRisk$agecat),2), xlab=”agecat”, names.arg=c(“Younger”, “Middle Aged”, “Older Adults”, “Oldest”), main= “Figure 2.0: 100% Stacked Bar Chart of caloriecat by agecat”, col=c(”aquamarine”,”hotpink”),ylab=”Frequency”,xlim=c(0,5)) legend(“topright”,names.arg=c(“Light Diet”,”Heavy Diet”), fill=c(“aquamarine”,”hotpink”))