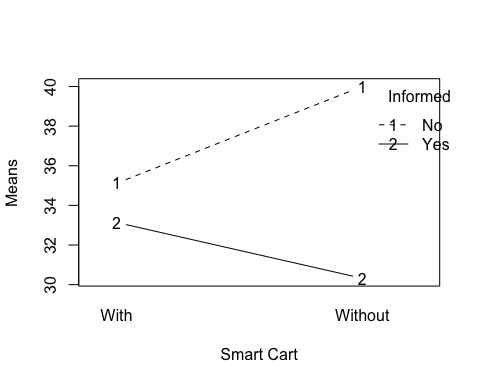
HW6.R

Yufan Ye

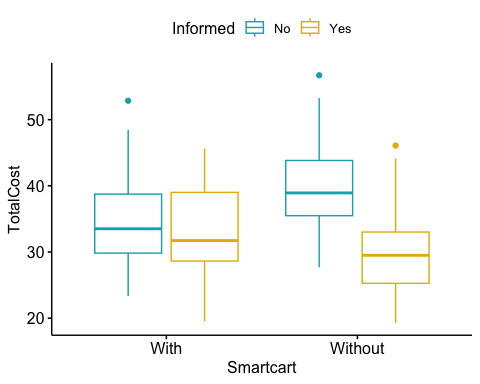
2022-12-12

#13.11  
#a)  
## Features: For subjects who were informed of their budget, average amount   
## of money spent is larger for those with real-time feedback than those  
## without. However, for subjects without informed budget, average amount of   
## money spent is smaller for those real-time feedback than those without.  
dfsmart <- read.csv('EX13-11SMART2.csv')  
interaction.plot(dfsmart$Smartcart,dfsmart$Informed,dfsmart$TotalCost,  
 type=c('b'),trace.label = 'Informed',  
 xlab='Smart Cart',ylab='Means',)  
  
library("ggpubr")

## Loading required package: ggplot2



ggboxplot(dfsmart, x = "Smartcart", y = "TotalCost", color = "Informed",  
 palette = c("#00AFBB", "#E7B800"))



#b)  
##Note: if factor is numeric, use df$name\_factor <- as.factor(df$name)  
## F(Smartcart)=1.30, F(Informed)=37.06, F(Interaction)=16.18. Degress of freedom of main  
## effects and interaction are all 1. P(Smartcart)=0.256, P(Informed)=0.000, P(Interaction)=0.000  
smartanova <- aov(TotalCost ~ Smartcart\*Informed, data = dfsmart)  
summary(smartanova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Smartcart 1 58 57.9 1.30 0.256   
## Informed 1 1651 1651.1 37.06 6.22e-09 \*\*\*  
## Smartcart:Informed 1 721 720.8 16.18 8.32e-05 \*\*\*  
## Residuals 190 8466 44.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

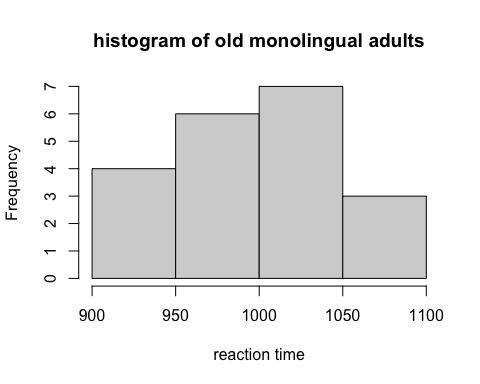
##Alternative method:  
##smartmodel<-lm(TotalCost~Smartcart+Informed+Smartcart:Informed, data=dfsmart)  
##smartanova2<-anova(smartmodel)  
##smartanova2  
  
#c)  
##Based on the output, F statistic for real-time feedback only is 1.30 with a P value of   
## 0.256<0.05, which indicated that the main effect of smart cart is not statistically   
## significant. F statistic for informed budget only is 37.06 with a P value close to 0.000<0.05,  
## which means that the main effect of informing the subjects of their budget is statistically significant.  
## The interaction beteen real-time feedback smartcart and informing the budget has a F statistic of   
## 16.18 with a P value close to 0.000<0.05, which means that the interaction between real-time feedback smartcart and informing budget is statistically significant  
  
#13.22  
#a)  
## It is reasonable to pool the variance.  
## The largest sd is 54.38, and the smallest sd is 42.04. sd max/sd min = 54.38/42.04 = 1.29 < 2  
dfbiling<-read.csv('EX13-22BILING.csv')  
dfoldmo<-dfbiling[dfbiling$Age=='Old'& dfbiling$Ling=='Mono',]  
dfoldbi<-dfbiling[dfbiling$Age=='Old'& dfbiling$Ling=='Bi',]  
dfyoungmo<-dfbiling[dfbiling$Age=='Young'& dfbiling$Ling=='Mono',]  
dfyoungbi<-dfbiling[dfbiling$Age=='Young'& dfbiling$Ling=='Bi',]  
tbbiling<-matrix(c(length(dfoldmo$Time),mean(dfoldmo$Time),sd(dfoldmo$Time),  
 length(dfoldbi$Time),mean(dfoldbi$Time),sd(dfoldbi$Time),  
 length(dfyoungmo$Time),mean(dfyoungmo$Time),sd(dfyoungmo$Time),  
 length(dfyoungbi$Time),mean(dfyoungbi$Time),sd(dfyoungbi$Time)),  
 ncol=3,byrow=TRUE)  
colnames(tbbiling)<-c('Sample size','Means','Standard deviation')  
rownames(tbbiling)<-c('Old\_Mono','Old\_Bi','Young\_Mono','Young\_Bi')  
tbbiling<-as.table(tbbiling)  
tbbiling

## Sample size Means Standard deviation  
## Old\_Mono 20.00000 996.85000 53.59524  
## Old\_Bi 20.00000 919.30000 54.38663  
## Young\_Mono 20.00000 820.70000 47.38543  
## Young\_Bi 20.00000 785.65000 42.04293

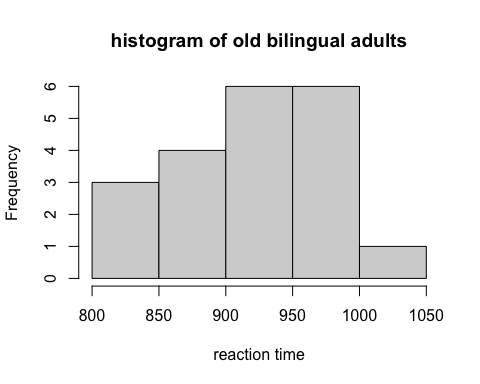
sdratio=54.38/42.04  
sdratio

## [1] 1.29353

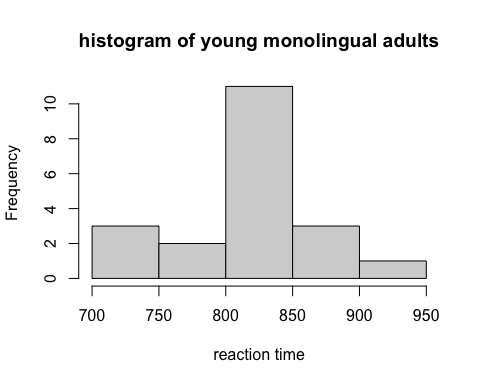
#b)  
## Yes based on the distribution of histograms of the four groups, we can confirm that the samples are approximately normal.  
oldmohist<-hist(dfoldmo$Time,xlab='reaction time',main='histogram of old monolingual adults')



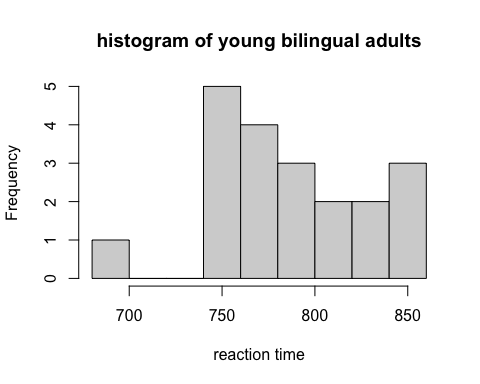
oldbihist<-hist(dfoldbi$Time,xlab='reaction time',main='histogram of old bilingual adults')



youngmohist<-hist(dfyoungmo$Time,xlab='reaction time',main='histogram of young monolingual adults')



youngbihist<-hist(dfyoungbi$Time,xlab='reaction time',main='histogram of young bilingual adults')



#13.23  
dfbiling2<-read.csv('EX13-23BILING.csv')  
#a)  
## If lingualism helps with brain functions as we age, we would expect to see the differences in reaction time between young adults and older adults   
## to be different in monolingual and bilingual groups. In other words, if bilingualism helps brian functioning, the decrease of brian function with aging will be   
## smaller for the bilingual group than the monolingual group, which is reflected by the interaction plot. Based on the interaction plot,  
## the two lines are not parallel, and the reaction time of bilingual group is smaller than monolingual group for both the young and older adults.  
interaction.plot(dfbiling2$Age,dfbiling2$Ling,dfbiling2$Time,  
 type=c('b'),trace.label = 'Bilingual or monolingual',  
 xlab='Age',ylab='Means of reaction time',)



#b)  
## F(age)=195.007,P(age)=0.000,df(age)=1  
## F(ling)=25.761,P(ling)=0.000,df(ling)=1  
## F(interaction)=3.67, P(interaction)=0.059, df(interaction)=1  
bilingmodel<-lm(Time~Age+Ling+Age:Ling, data=dfbiling2)  
bilinganova<-anova(bilingmodel)  
bilinganova

## Analysis of Variance Table  
##   
## Response: Time  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 479880 479880 195.007 < 2.2e-16 \*\*\*  
## Ling 1 63394 63394 25.761 2.667e-06 \*\*\*  
## Age:Ling 1 9031 9031 3.670 0.05916 .   
## Residuals 76 187024 2461   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#c)  
## Based on the results in part b, p value of both age and lingualism are smaller than  
## 0.05, indicating statistically significant main effects of both age(old or young), and lingualism (bilingual or monolingual)  
## on reaction time to the task. However, the p value of interaction between age and lingualism is 0.058 > 0.05, which  
## indicates a insignificant interaction of age and lingualism.   
  
#14.48  
dfgpahi<-read.csv('EX14-048GPAHI.csv')  
#a)  
## Based on the results, p(SATM)=0.00137<0.05, p(SATCR)=0.912>0.05, p(SATM+SATCR)=0.00028<0.05.   
## The null hypothesis that B1SATM=B2SATCR=0 was rejected.  
## The null hypothesis that B1SATM=0 was rejected, while we failed to reject B2SATCR=0.   
## The fitting model is log(odds)=-5.847+0.00957\*SATM-0.000265\*SATCR  
gpalogit <- glm(HIGPA ~ SATM + SATCR, data = dfgpahi, family = "binomial")  
summary(gpalogit)

##   
## Call:  
## glm(formula = HIGPA ~ SATM + SATCR, family = "binomial", data = dfgpahi)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8950 -1.0189 -0.6433 1.1187 1.7871   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.8470398 1.6093331 -3.633 0.00028 \*\*\*  
## SATM 0.0095754 0.0029907 3.202 0.00137 \*\*   
## SATCR -0.0002658 0.0024148 -0.110 0.91236   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 207.92 on 149 degrees of freedom  
## Residual deviance: 191.80 on 147 degrees of freedom  
## AIC: 197.8  
##   
## Number of Fisher Scoring iterations: 4

#b)  
## We are 95% confident that B1 of SATM falls between (0.00371, 0.01254)  
## We are 95% confident that B2 of SATCR falls between (-0.00499,0.00446)  
confint.default(gpalogit)

## 2.5 % 97.5 %  
## (Intercept) -9.001274674 -2.692805025  
## SATM 0.003713639 0.015437107  
## SATCR -0.004998629 0.004467055

#c)  
## Based on parts a and b, I concluded that the fitting model using SATM and SATCR as factor to predict HIGPA is log(odds)=-5.847+0.00957\*SATM-0.000265\*SATCR  
## with SATM as a statistically significant explanatory variable and SATCR as a insignificant predicting factor.  
## We are 95% confident that B1 of SATM falls between (0.00371, 0.01254) and 95% confident that B2 of SATCR falls between (-0.00499,0.00446)  
## Given that SATCR is insignificant, and the fact that its confident interval ranges from negative and positive numbers, it would be reasonable to drop SATCR as  
## a predicting factor in the fitting model for HIGPA  
  
#14.50  
#a)  
## When using sex as a explanatory variable to predict HIGPA, the fitting model when the   
## student is a male is log(odds)=0.1699-0.3240\*Sex\_male, meaning when the student is a male, the log(odd) of getting   
## a high GPA decreases by approximately 0.3240. And the fitting model when the student is a female  
## is log(odds)=0.1699-0.3240\*Sex\_female=0.1699-0.3240\*0=0.1699.   
## However, because the p value of this model is 0.334>0.05, model using sex as a single explanatory factor would not be considered as significant in predicting HIGPA   
dfgpahi2<-read.csv('EX14-050GPAHI.csv')  
dfgpahi2[dfgpahi2$sex==1,]$sex<-'M'  
dfgpahi2[dfgpahi2$sex==2,]$sex<-'F'  
dfgpahi2$sex<-as.factor(dfgpahi2$sex)  
dfgpahi2$HIGPA<-as.factor(dfgpahi2$HIGPA)  
gpalogitgender <- glm(HIGPA ~ sex, data = dfgpahi2, family = "binomial")  
summary(gpalogitgender)

##   
## Call:  
## glm(formula = HIGPA ~ sex, family = "binomial", data = dfgpahi2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.250 -1.113 -1.113 1.244 1.244   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.1699 0.2613 0.650 0.516  
## sexM -0.3240 0.3354 -0.966 0.334  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 207.92 on 149 degrees of freedom  
## Residual deviance: 206.98 on 148 degrees of freedom  
## AIC: 210.98  
##   
## Number of Fisher Scoring iterations: 3

#b)  
## When using sex and SAT scores as explanatory variables to predict HIGPA, the fitting model when the   
## student is a male is log(odds)=-7.345-1.233\*Sex\_male-0.0011\*SATCR+0.0140\*SATM.Only the p value of SATCR is larger than 0.05,  
## indicating that the null hypothesis that B2SATCR=0 failed to be rejected.  
gpalogit2 <- glm(HIGPA ~ sex+SATCR+SATM, data = dfgpahi2, family = "binomial")  
summary(gpalogit2)

##   
## Call:  
## glm(formula = HIGPA ~ sex + SATCR + SATM, family = "binomial",   
## data = dfgpahi2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9518 -0.9854 -0.4002 1.0098 1.8443   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.345220 1.757122 -4.180 2.91e-05 \*\*\*  
## sexM -1.233426 0.424515 -2.905 0.00367 \*\*   
## SATCR -0.001189 0.002505 -0.474 0.63516   
## SATM 0.014017 0.003507 3.997 6.41e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 207.92 on 149 degrees of freedom  
## Residual deviance: 182.59 on 146 degrees of freedom  
## AIC: 190.59  
##   
## Number of Fisher Scoring iterations: 4

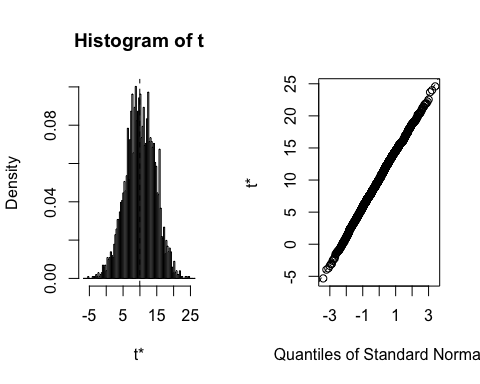
#c)  
# Based on parts a and b, I concluded that the fitting model using sed, SATM and SATCR as factor to predict HIGPA is log(odds)=-7.345-1.233\*Sex\_male-0.0011\*SATCR+0.0140\*SATM.  
## when the student is a male. The fitting model when the student is a female is log(odds)=-7.345-0.0011\*SATCR+0.0140\*SATM.  
## with both sex and SATM as statistically significant explanatory variables and SATCR as a insignificant predicting factor.  
## Thus, SATCR should be drop in this fitting model.   
  
#16.18  
dfdrp<- read.csv('EX16-18DRP.csv')  
#install.packages('boot', dep=TRUE)  
library(boot)  
#a)  
## bootstrap standard error = 4.3627  
fc1<-function(data,i){  
 d2<-data[i,]  
 return((mean(d2$drp[d2$group=='Treat']))-  
 (mean(d2$drp[d2$group=='Control'])))  
 }  
  
set.seed(12345)  
bootmeandiff<-boot(dfdrp,fc1,R=3000)  
summary(bootmeandiff)

## R original bootBias bootSE bootMed  
## 1 3000 9.9545 0.16429 4.3627 10.094

bootmeandiff

##   
## ORDINARY NONPARAMETRIC BOOTSTRAP  
##   
##   
## Call:  
## boot(data = dfdrp, statistic = fc1, R = 3000)  
##   
##   
## Bootstrap Statistics :  
## original bias std. error  
## t1\* 9.954451 0.1642901 4.362739

#b)  
## Yes, a bootstrap t confidence interval is appropriate  
## The 95% confidence interval is (1.239,18.341)  
plot(bootmeandiff)



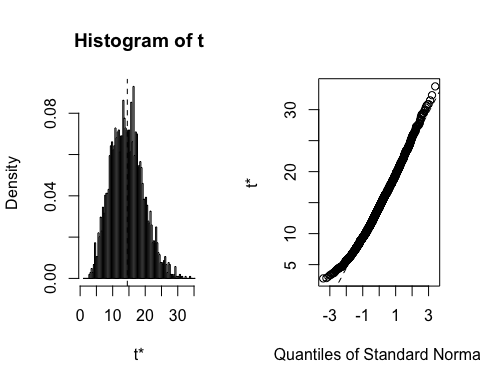
boot.ci(bootmeandiff,type='norm')

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 3000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = bootmeandiff, type = "norm")  
##   
## Intervals :   
## Level Normal   
## 95% ( 1.239, 18.341 )   
## Calculations and Intervals on Original Scale

#c)  
## The confidence interval using two-sample t test from page 441 is (1.233,18.675), which is   
## close to the bootstrap 95% confidence interval (1.239, 18.341)  
  
#16.24  
dftv<-read.csv('EX16-24TVTIME.csv')  
#a)  
## Based on the plot of bootmeantv, the bootstrap distribution of   
## mean of traditional tv watching is reasonably normal.   
## The bias is -0.014, which is small compared to the observed mean 14.5.  
fc2<-function(data,i){  
 d2<-data[i,]  
 return (mean(d2$Time[!is.na(d2$Time)]))  
 }  
set.seed(123456)  
bootmeantv<-boot(dftv,fc2,R=3000)  
bootmeantv

##   
## ORDINARY NONPARAMETRIC BOOTSTRAP  
##   
##   
## Call:  
## boot(data = dftv, statistic = fc2, R = 3000)  
##   
##   
## Bootstrap Statistics :  
## original bias std. error  
## t1\* 14.5 -0.01407011 5.024223

plot(bootmeantv)



#b)  
## The 95% bootstrap confidence interval is (4.67, 24.36)  
boot.ci(bootmeantv,type='norm')

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 3000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = bootmeantv, type = "norm")  
##   
## Intervals :   
## Level Normal   
## 95% ( 4.67, 24.36 )   
## Calculations and Intervals on Original Scale

#c)  
## 95% t confidence interval is (2.081, 26.918). The t CI has a larger  
## range compared to the 95% bootstrap CI (4.67, 24.36). The difference   
## between the these two CIs is approximately 2 for the low and upper tails, respectively.   
t.test(dftv$Time[!is.na(dftv$Time)])

##   
## One Sample t-test  
##   
## data: dftv$Time[!is.na(dftv$Time)]  
## t = 2.761, df = 7, p-value = 0.02806  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 2.081702 26.918298  
## sample estimates:  
## mean of x   
## 14.5