#3 and 4

Clayton Chan

2024-10-17

3a.

Group	Mean of Income	Variance of Income	Sample Size
Intact Family	9182.08	37256691.57	2777
Non-intact Family	7042.96	24563879.81	928

```
## [1] 8646.293
```

```
## [1] 34929285
```

$\bar{Y}_{..} = 8646.29, S_y^2 = 34929285$

3b.

i)

Since we are going to compare this to ANOVA and assume equal variance, we will use the pooled t-test. $t = \frac{\bar{Y_{.1}} - \bar{Y_{.2}}}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}$ where $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = \frac{2776*37256691.57+927*24563879.81}{2777+928-2} = 34079204$

$t = \frac{9182.08-7042.96}{\sqrt{34079204(\frac{1}{2777}+\frac{1}{928})}} = 9.66$

The t-value for testing $H_0 : \mu_{intact} = \mu_{non-intact}$ is 9.66

ii)

The 95% confidence interval is $(\bar{Y}_{1.} - \bar{Y}_{2.}) \pm t_{1-\frac{.05}{2},2777+928-2}\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})} = (9182.08 - 7042.96) \pm 1.96 * \sqrt{34079204(\frac{1}{2777}+\frac{1}{928})} = [1705.145, 2573.099]$

```
## [1] 1705.145 2573.099
```

We are 95% confident that the population mean income difference between intact families and non-intact families is contained within [1705.145,2573.099]

iii)

```
## [1] 3182780904
```

$SSB = n_1(\bar{Y_{1.}} - \bar{Y_{..}})^2 + n_2(\bar{Y_{2.}} - \bar{Y_{..}})^2 = 2777(9182.08 - 8646.29)^2 + 928(7042.96 - 8646.29)^2 = 3,182,780,904$

iv)

```
## [1] 126195292388
```

$SSW = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 = 2776 * 37256691.57 + 927 * 24563879.81 = 126,195,292,388$

v)

$F = \frac{\frac{SSB}{k-1}}{\frac{SSW}{N-k}} = \frac{\frac{3182780904}{2-1}}{\frac{126195292388}{3705-2}} = 93.39$

vi)

```
## [1] 0.02460062
```

$R^2 = \frac{SSB}{SST} = \frac{SSB}{SSB+SSW} = \frac{3182780904}{3182780904+126195292388} = 0.0246$ which means that 2.46% of the variance in income is explained by the variable nonint

4

parts a and b

```r
df2 <- read.csv('campusclimate.csv')
df2 <- df2[,c("Q10_A_5","classcomfort")]
table(df2$classcomfort)
```

```
##
##    1    2    3    4    5    6
##  910 2913 1172  328   52    3
```

```r
df2 <- df2[df2$classcomfort!=6,] #Drop level 6 in comfort as it's an unknown
#level not of interest
df2$classcomfort <- ifelse(df2$classcomfort == 1, "very comfortable",
                    ifelse(df2$classcomfort == 2,"comfortable",
                    ifelse(df2$classcomfort == 3,"somewhat","uncomfortable")))
df2$classcomfort <- factor(df2$classcomfort,
                           levels = c("very comfortable","comfortable",
                                      "somewhat","uncomfortable"),
                           ordered = TRUE)
#Recode level 1 as very comfortable level 2 as comfortable
#level 3 as somewhat and levels 4/5 uncomfortable
#Reorder levels appropriately
df2 <- df2[complete.cases(df2),] #Drop NA rows
df2 <- df2[df2$Q10_A_5!=3 & df2$Q10_A_5!=6,] # Drop levels 3 and 6 for Q10
df2$Q10_A_5 <- ifelse(df2$Q10_A_5==1 | df2$Q10_A_5==2,"agree","disagree")
#Recode levels 1 and 2
#as agree and levels 3 and 4 as disagree
```

c)

```
##
##                      agree disagree
##   very comfortable   558      194
##   comfortable       1283      948
##   somewhat           349      531
##   uncomfortable       95      222
```

```
##
## very comfortable      comfortable         somewhat    uncomfortable
##              752             2231              880              317
```

```
##
##    agree disagree
##     2285     1895
```

```
## [1] 281.0743
```

Let H_0:Perception of academic success and perception of comfort are independent and H_a :Perception of academic success and perception of comfort are dependent with $\alpha = 0.05$

$\chi^2 = \sum \frac{(Observed-Expected)^2}{Expected}, Expected = \frac{RowTotal*ColTotal}{OverallTotal}$

$\chi^2 = \frac{(558-752*2285/4180)^2}{752*2285/4180} + \frac{(194-752*1895/4180)^2}{752*1895/4180} ... + \frac{(222-317*1895/4180)^2}{317*1895/4180} = 281.07$

Based on our p-value, it is extremely small and < 0.05. Thus, we will reject the null hypothesis and conclude that perception of academic success and perception of comfort are dependent.

d)

```
##
##                          agree  disagree
##   very comfortable 0.7420213 0.2579787
##   comfortable      0.5750784 0.4249216
##   somewhat         0.3965909 0.6034091
##   uncomfortable    0.2996845 0.7003155
```

When a student perceives class comfort as very comfortable, 74.2% agree with performing academically as well as they could while 25.8% disagree with that statement

When a student perceives class comfort as comfortable, 57.5% agree with performing academically as well as they could while 42.5% disagree with that statement

When a student perceives class comfort as somewhat, 39.7% agree with performing academically as well as they could while 60.3% disagree with that statement

When a student perceives class comfort as uncomfortable, 30% agree with performing academically as well as they could while 70% disagree with that statement

We can see as perceived comfort levels decline, more and more students will disagree with performing academically as well as they could.

e)

```
## [1] 3.162625
```

We'll define the odds ratio as $\frac{\frac{P(Agree|Comfortable)}{P(Disagree|Comfortable)}}{\frac{P(Agree|Uncomfortable)}{P(Disagree|Uncomfortable)}} = \frac{\frac{.575}{.425}}{\frac{.3}{.7}} = 3.16$

In this context, the odds of one agreeing with performing academically as well as possible for those who have a comfortable perception is 3.16x more compared to those who have an uncomfortable perception

```r
knitr::opts_chunk$set(echo = FALSE)
library(kableExtra)
df <- read.csv("womenpowers.csv")
#sum(complete.cases(df)) Check for NA values there are none
ybars <- tapply(df$income,df$nonint,mean)
sy <- tapply(df$income,df$nonint,var)
n <- table(df$nonint)
anovatable <- matrix(c("Group","Mean of Income","Variance of Income","Sample Size",
                       "Intact Family",round(ybars[1],2),round(sy[1],2),n[1],
                       "Non-intact Family",round(ybars[2],2),round(sy[2],2),n[2]),nrow=3,byrow = TRUE)
kbl(anovatable) %>% kable_styling(latex_options = c("striped", "hold_position"))
y.. <- mean(df$income)
y..
var(df$income)
Sp2 <- unname((2776*sy[1]+927*sy[2])/(2777+928-2))
tstat <- (ybars[1]-ybars[2])/sqrt(Sp2*(1/2777+1/928))
L <- (ybars[1]-ybars[2])-qt(1-.05/2,2777+928-2)*sqrt(Sp2*(1/2777+1/928))
U <- (ybars[1]-ybars[2])+qt(1-.05/2,2777+928-2)*sqrt(Sp2*(1/2777+1/928))
unname(c(L,U))
SSB <- 2777*(ybars[1]-y..)^2+928*(ybars[2]-y..)^2
unname(SSB)
(SSW <- unname(2776*sy[1]+927*sy[2]))
Fstat <- (SSB/1)/(SSW/(nrow(df)-2))
r2 <- SSB/(SSB+SSW)
unname(r2)
df2 <- read.csv('campusclimate.csv')
df2 <- df2[,c("Q10_A_5","classcomfort")]
table(df2$classcomfort)
df2 <- df2[df2$classcomfort!=6,] #Drop level 6 in comfort as it's an unknown
#level not of interest
df2$classcomfort <- ifelse(df2$classcomfort == 1, "very comfortable",
                    ifelse(df2$classcomfort == 2,"comfortable",
                    ifelse(df2$classcomfort == 3,"somewhat","uncomfortable")))
df2$classcomfort <- factor(df2$classcomfort,
                           levels = c("very comfortable","comfortable",
                                      "somewhat","uncomfortable"),
                           ordered = TRUE)
#Recode level 1 as very comfortable level 2 as comfortable
#level 3 as somewhat and levels 4/5 uncomfortable
#Reorder levels appropriately
df2 <- df2[complete.cases(df2),] #Drop NA rows
df2 <- df2[df2$Q10_A_5!=3 & df2$Q10_A_5!=6,] # Drop levels 3 and 6 for Q10
df2$Q10_A_5 <- ifelse(df2$Q10_A_5==1 | df2$Q10_A_5==2,"agree","disagree")
#Recode levels 1 and 2
#as agree and levels 3 and 4 as disagree

O <- table(df2$classcomfort,df2$Q10_A_5)
```

```r
O
margin.table(O,1)
margin.table(O,2)
E <- as.matrix(margin.table(O,1))%*%t(as.matrix(margin.table(O,2)))/sum(O) #Expected Values
(X2 <- sum((O-E)^2/E))

prop.table(O,1)
probs <- prop.table(O,1)
comfodd <- probs[2,1]/probs[2,2]
uncomfodd <- probs[4,1]/probs[4,2]
comfodd/uncomfodd
```