Solution for the R part of Lab 3

Ali

August 6, 2015

```r
library(foreign)
library(car)
load("GSS.Rdata")
# summary(GSS) #commented out for the report
```

1. Chi-squared test

```
##       married       widowed      divorced     separated never married 
##           795           165           213            40           286 
##            NA 
##             1
```

```r
GSS$marital[GSS$marital=="NA"] = NA
GSS$marital[GSS$marital=="NA"]
```

```
## [1] <NA>
## Levels: married widowed divorced separated never married NA
```

```r
summary(GSS$politics)
```

```
##      Liberal     Tend Lib     Moderate    Tend Cons Conservative 
##          193          193          527          248          282 
##         NA's 
##           57
```

```r
table(GSS$marital, GSS$politics)
```

```
##                
##                 Liberal Tend Lib Moderate Tend Cons Conservative
##   married            93       92      271       140          173
##   widowed            15       16       57        24           37
##   divorced           22       36       79        38           29
##   separated           7        3       22         6            1
##   never married      55       46       98        40           42
##   NA                  0        0        0         0            0
```

```r
cs = chisq.test(GSS$marital, GSS$politics)

cs$statistic
```

```
## X-squared 
##   44.2255
```

```r
# cs$stdres
# cs$expected

library(lsr)
```

```
## Warning: package 'lsr' was built under R version 3.1.3
```

```r
cramersV(cs$observed)
```

```
## [1] 0.08756363
```

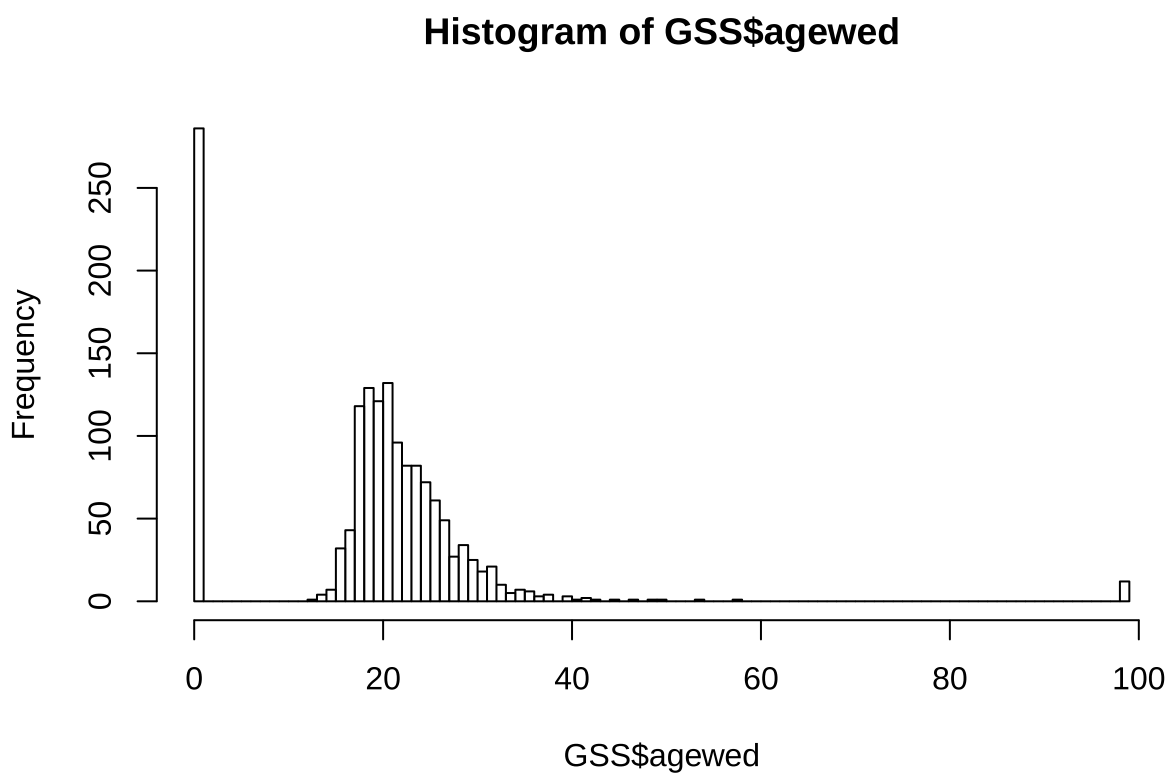
---

2. Correlation

```r
# Conduct an analysis to examine the
# relationship between agewed and tvhours.
summary(GSS$agewed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   18.00   21.00   19.06   24.00   99.00
```

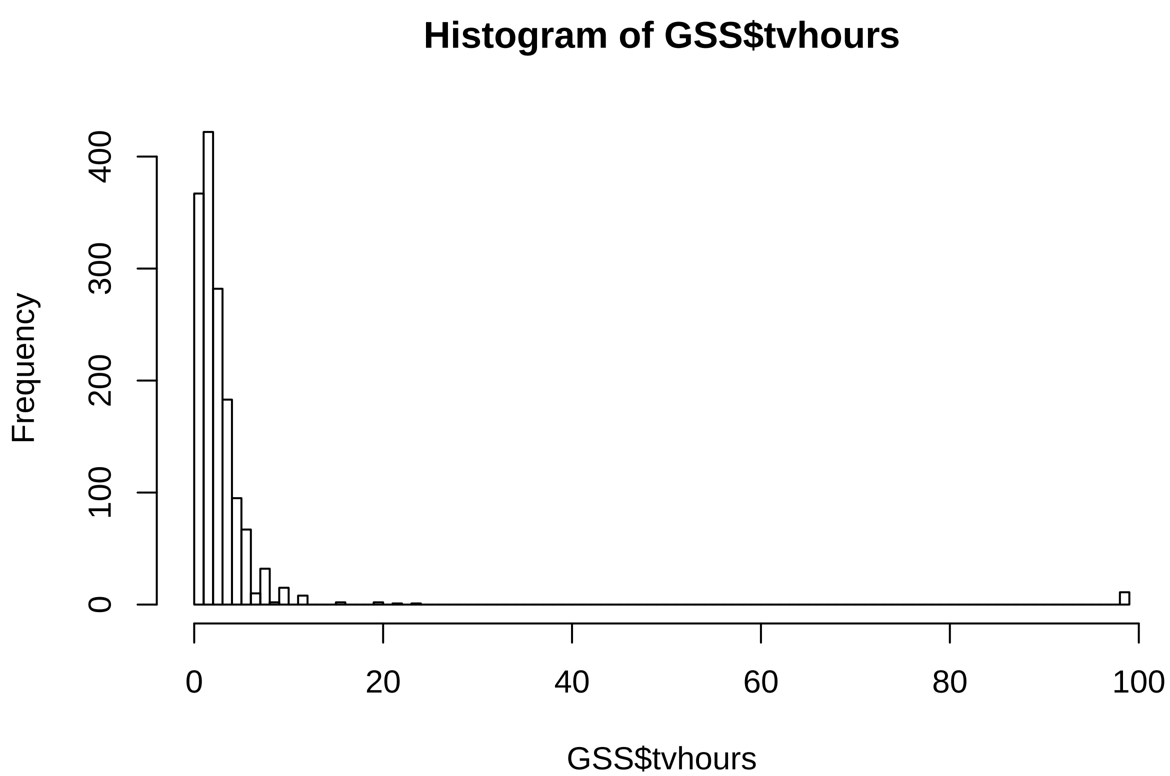
```r
hist(GSS$agewed, breaks= 100)
```



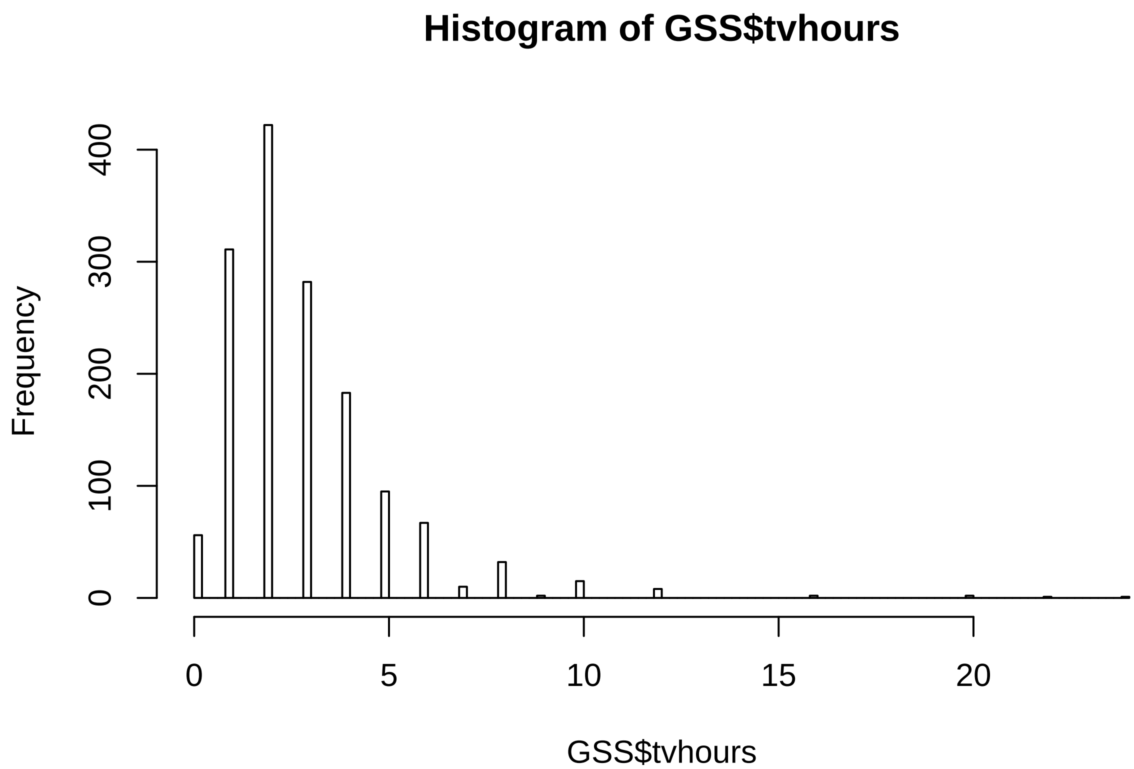
```r
GSS$agewed[GSS$agewed==0] = NA
GSS$agewed[GSS$agewed==99] = NA

# tv hours
summary(GSS$tvhours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   2.000   3.605   4.000  99.000
```

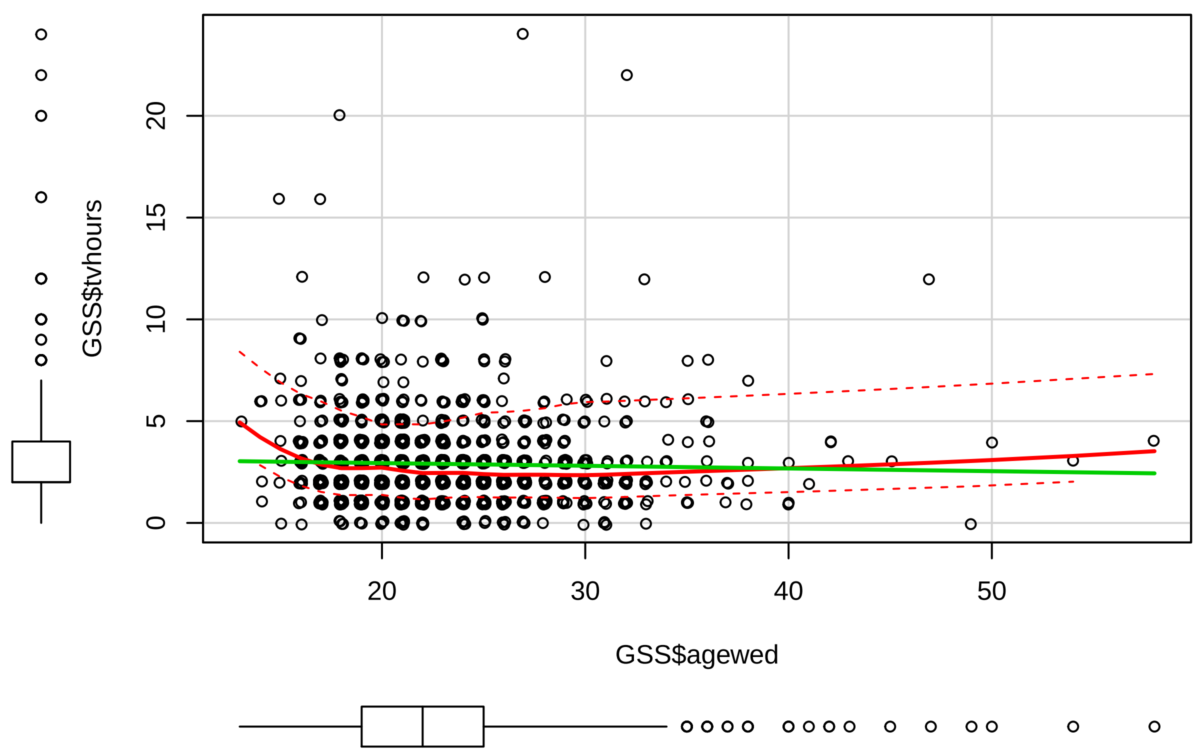
```r
hist(GSS$tvhours, breaks= 100)
```




```r
GSS$tvhours[GSS$tvhours==99] = NA
hist(GSS$tvhours, breaks= 100)
```




```
scatterplot(GSS$agewed, GSS$tvhours,jitter = list(x=.5,y=.5),lwd=2)
```




```
ct = cor.test(GSS$agewed, GSS$tvhours)
cat ( 'r = ', ct$estimate,'....' ,'p-value=', ct$p.value )
```

```
## r =  -0.02996096 .... p-value= 0.3009361
```

---

3. T-test

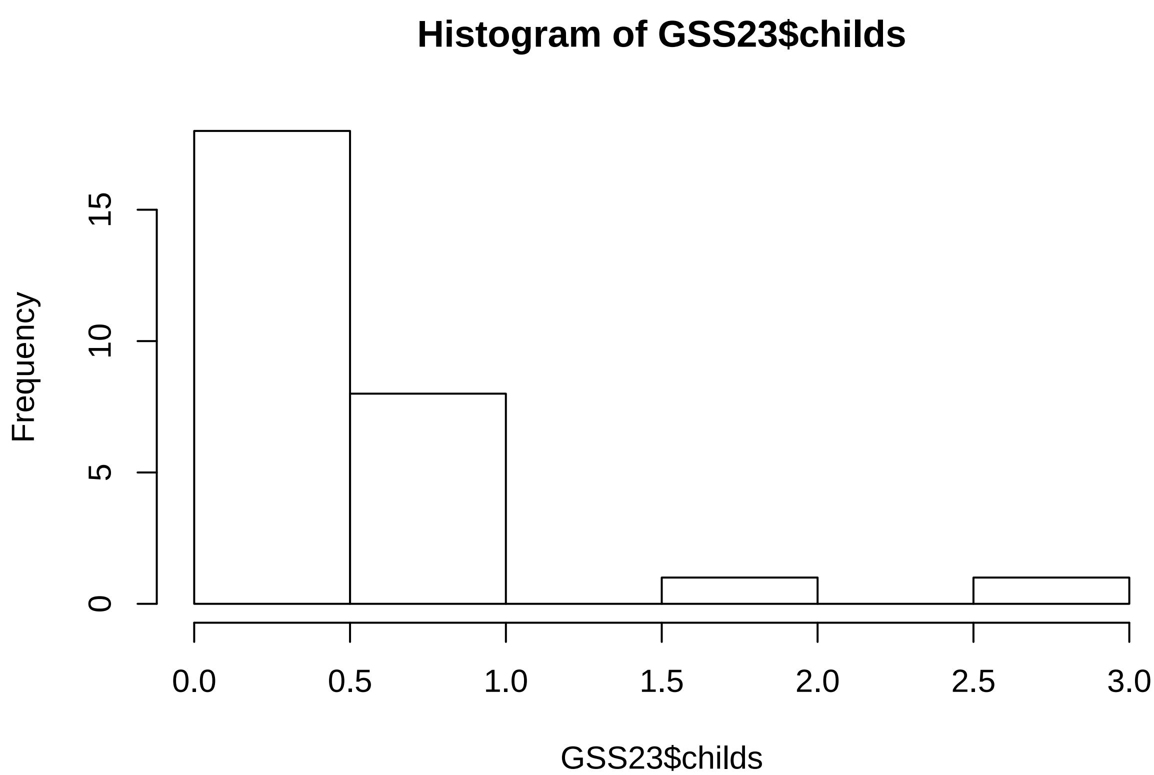
```
GSS23=GSS[GSS$age==23, ]
# summary(GSS23)

summary(GSS23$marital)
```

```
##       married       widowed      divorced     separated never married
##             8             0             0             0            20
##            NA
##             0
```

```
GSS23$married = ifelse(GSS23$marital == "married", 1, 0)

# percent of married 23-year olds
mean( as.numeric(GSS23$married, na.rm=T) )
```

```
## [1] 0.2857143
```

```
hist(GSS23$childs)
```



```r
wt = wilcox.test(childs ~ married, data = GSS23)
```

```
## Warning in wilcox.test.default(x = c(0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, :
## cannot compute exact p-value with ties
```

```r
wt$statistic
```

```
##  W
## 19
```

```r
wt$p.value
```

```
## [1] 0.0002656205
```

```r
# Using the formula in the book (ch.15)
# for effect size out of wilcox.test

rFromWilcox<-function(wilcoxModel, N){
  z<- qnorm(wilcoxModel$p.value/2)
    r<- z/ sqrt(N)
    cat(wilcoxModel$data.name, "Effect Size, r = ", r,"\n")
  r
}

N = dim(GSS23)[1]
rFW = rFromWilcox(wt,N)
```

```
## childs by married Effect Size, r =  -0.6891632
```

Something to note above Field's function is that the way he calculates the z-score from the p-value, we always get a negative z-score, so the result is always going to be negative. Here, we know for a fact that the effect is positive; for example, see `t.test(childs ~ married, data = GSS23)` to see why.). But we really just care about the size of `rFromWilcox` not its sign. Or, just square it be happy.

```r
cat('r = ', abs(rFW) , '........ r^2 = ',rFW^2)
```

```
## r =  0.6891632 ........ r^2 =  0.4749459
```

The above result is what we are looking for. But one could also do `cohens_d` or a regular effect size correlation while ignoring that we have been asked to do a wilcoxon rank-sum test.

```r
cor.test(GSS23$childs, GSS23$married)
```

```
##
##  Pearson's product-moment correlation
##
## data:  GSS23$childs and GSS23$married
## t = 4.7257, df = 26, p-value = 6.934e-05
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.4108605 0.8398404
## sample estimates:
##       cor
## 0.6797444
```

```
cohens_d <- function(x,y)
{
    # this function takes two vectors as input and compares their means
    #lets compute the pooled standard error
    lx = length(subset(x,!is.na(x)))
    ly = length(subset(y,!is.na(y)))
    #numerator of the pooled variance
    num = (lx-1)*var(x,na.rm=T) + (ly-1)*var(y,na.rm=T)
    pooled_var = num / (lx + ly -2)
    pooled_sd = sqrt(pooled_var)
    # cohen's d
    cd = abs(mean(x,na.rm=T)-mean(y,na.rm=T))/pooled_sd
    return(cd)
}


with(GSS23, cohens_d(childs[married==0] , childs[married==1]) )
```

```
## [1] 1.976885
```

---

4. ANOVA

```
#examine relationship between relig and agewed
summary(GSS$relig)
```

```
## Protestant   Catholic     Jewish       None      Other         DK
##        953        333         31        140         35          1
##         NA
##          7
```

```
GSS$relig[GSS$relig == "NA"] = NA
GSS$relig[GSS$relig == "DK"] = NA
GSS$relig = factor(GSS$relig)
summary(GSS$relig)
```

```
## Protestant   Catholic     Jewish       None      Other       NA's
##        953        333         31        140         35          8
```

```
aov_model = aov(agewed ~ relig, data = GSS)
summary.lm(aov_model)
```

```
##
## Call:
## aov(formula = agewed ~ relig, data = GSS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.634 -3.253 -1.253  2.366 35.747
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.2529     0.1772 125.553  < 2e-16 ***
## religCatholic   1.3811     0.3531   3.911 9.71e-05 ***
## religJewish     3.3993     1.0518   3.232  0.00126 **
## religNone       1.0629     0.5400   1.968  0.04927 *
## religOther      3.1702     0.9911   3.199  0.00142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.972 on 1191 degrees of freedom
##   (304 observations deleted due to missingness)
## Multiple R-squared:  0.02679,    Adjusted R-squared:  0.02352
## F-statistic: 8.197 on 4 and 1191 DF,  p-value: 1.606e-06
```

```
summary(aov_model)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## relig          4    811  202.64   8.197 1.61e-06 ***
## Residuals   1191  29444   24.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 304 observations deleted due to missingness
```

```
# remember that eta^2
#   in the anova lingo is: R^2
eta.sqr = summary.lm(aov_model)$r.squared
eta = sqrt(eta.sqr)
cat( ' eta^2 = ',eta.sqr ,'\n eta = ' ,sqrt(eta))
```

```
##  eta^2 =  0.02679156
##  eta =  0.4045754
```

```
omega.sq <- function(aovm){
    sum_stats <- summary(aovm)[[1]]
    SSm <- sum_stats[["Sum Sq"]][1]
    SSr <- sum_stats[["Sum Sq"]][2]
    DFm <- sum_stats[["Df"]][1]
    MSr <- sum_stats[["Mean Sq"]][2]
    SSt <- SSm + SSr
    `w^2` <- (SSm-DFm*MSr)/(SSt+MSr)
    return(`w^2`)
}
```

```r
omega.sqr = omega.sq(aov_model)
cat( ' omega^2 = ',omega.sqr ,'\n omega = ' ,sqrt(omega.sqr))
```

```
##  omega^2 =  0.02350381
##  omega =  0.1533095
```

```r
pairwise.t.test(GSS$agewed, GSS$relig, p.adjust.method="bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  GSS$agewed and GSS$relig
##
##          Protestant Catholic Jewish  None
## Catholic 0.00097    -        -       -
## Jewish   0.01263    0.62106  -       -
## None     0.49274    1.00000  0.43398 -
## Other    0.01417    0.80224  1.00000 0.55751
##
## P value adjustment method: bonferroni
```

```r
# significant comparisons: Catholic-Protestant (p=.00097), Jewish-Protestant (p=.01263),
# Other-Protestant (p=0.01417)
```