Lab 2

Greg Ceccarelli

June 30, 2015

Part 1. Multiple Choice (5 points each)

1. a
2. c
3. f
4. e
5. d
6. b
7. b
8. c
9. d
10. f

Part 2. Test Selection (10 points)

Every other year, the General Social Survey collects responses to thousands of questions, covering a wide variety of topics. You will be using a subset of data from 1993, including a small number of variables.

This may be found in the file, GSS.Rdata.

Like any survey, GSS data creates additional concerns that would normally go into a statistical analysis. Surveys are usually weighted in order to compensate for over- or under-representation of subgroups. For this lab, however, you will be using unweighted data, which limits how well your findings generalize to the U.S. population. For the following problems, select the most appropriate statistical test to answer the question from the given choices.

11. b
12. a

Part 3: Data Analysis and Short Answer (50 points)

13. Data Import and Error Checking

    a. Examine the "agewed" variable (age when married). What are the value(s) of agewed, if any, that do not meaningfully correspond to ages?

RESPONSE: There are 286 instances where agewed = 0, 11 instances where agewed = 99 and age < 99, 2 instances where agewed = 26 and age = 25. All of these instances do not meaninfully correspond to ages (although it is technically feasible to wed at age 99 and I left that outlier in the dataset)

    b. Recode any value(s) that do not correspond to age as NA. What is the mean of the agewed variable?

RESPONSE: With NA values of column agewed removed, the mean of agewed is 22.85012

```r
#set up working envrionment
setwd('/Users/gdc/Documents/MIDS/DATASCI_W203/Assignments/Labs/Lab 2')
rm( list = ls() )

#load supplied R data file
load("GSS.Rdata")

#make copy of DF
GSS_copy <- GSS[,]

#check for when agewed might exceed age
GSS$agewed.spurious <- GSS$agewed > GSS$age

#view tabulated counts by value
#View(data.frame(table(GSS$agewed)))

#review counts of spurious agewed data
summary(GSS$agewed.spurious)
```

```
##    Mode   FALSE    TRUE    NA's
## logical    1487      13       0
```

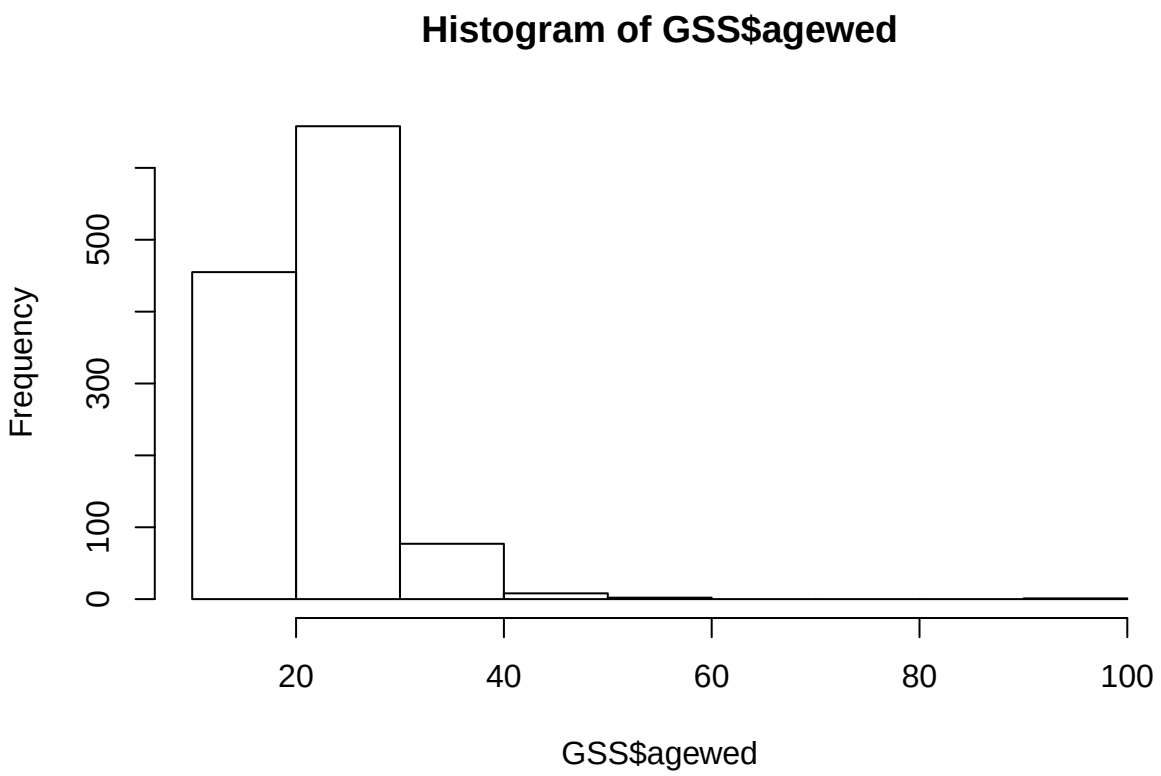
```r
#computer agewed mean prior to transformation
GSS.agewed.mean.before <-  mean(GSS$agewed)

#review the distinct agewed values that don't correspond with age
#View(data.frame(table(GSS$agewed[(GSS$agewed == 0 | (GSS$agewed.spurious == TRUE))])))

#update agewed variable in three instances (separated for clarity)
#NA when agewed = 0
GSS$agewed[GSS$agewed == 0] <- NA

#NA when agewed > wed -- keep one value of 99 (agewed == age)
GSS$agewed[GSS$agewed.spurious == TRUE] <- NA

#check to see how many variables were set to NA
GSS.na.summary <- is.na(GSS$agewed)
summary(GSS.na.summary)
```

```
##    Mode   FALSE    TRUE    NA's
## logical    1201     299       0
```

```r
#define function to omit row values where only certain columns are NA
data.complete <- function(data, desiredCols) {
  completeVec <- complete.cases(data[, desiredCols])
  return(data[completeVec, ])
}

#update dataset to remove NA values only where agedwed is NA, should remove 299 values
GSS <- data.complete(GSS, c("agewed"))

#double check
nrow(GSS_copy)-nrow(GSS)
```

```
## [1] 299
```

```
#compute mean
GSS.agewed.mean.after <- mean(GSS$agewed)

#review hist of updated agewed variable
hist(GSS$agewed)
```



14. Checking assumptions

a. Produce a QQ plot for the agewed variable. Using this plot information, is agewed normal and how do you know?

RESPONSE: A perfectly normal distribution produces a diagonal line in the qqplot. As we move along the X axis, agewed doesn't move very fast at all (and is not diagonal). Thus, agewed is not normally distributed

b. Perform a Shapiro-Wilk test on the agewed variable.
c. What is the null and alternative hypothesis for your test?

RESPONSE: The null hypothesis is that agewed is drawn from a normal distribution The alternative hypothesis is that agewed is drawn from a non-normal distribution

ii. What is your p-value, and what is your specific conclusion?

RESPONSE: In this case, the p-value is very small [2.2e-16]. Given it being far less than an alpha value = .05, it is very unlikely that we'd find this distribution of agewed variables under the assumption the population was normal. This is a very non-normal distribution.

c. What is the variance of agewed for men? What is the variance of agewed for women?

RESPONSE: Variances of agewed by gender

GSS$sex: Male [1] 34.99648

GSS$sex: Female [1] 24.31918

d. Perform a Levene's test for the agewed variable grouped by men and women.
e. What is the null and alternative hypothesis for this test?

RESPONSE: The null hypothesis is the assumption that the variances for agewed are equal between men and women The alternative hypothesis is how unlikely this difference of variances is to appear between men and women

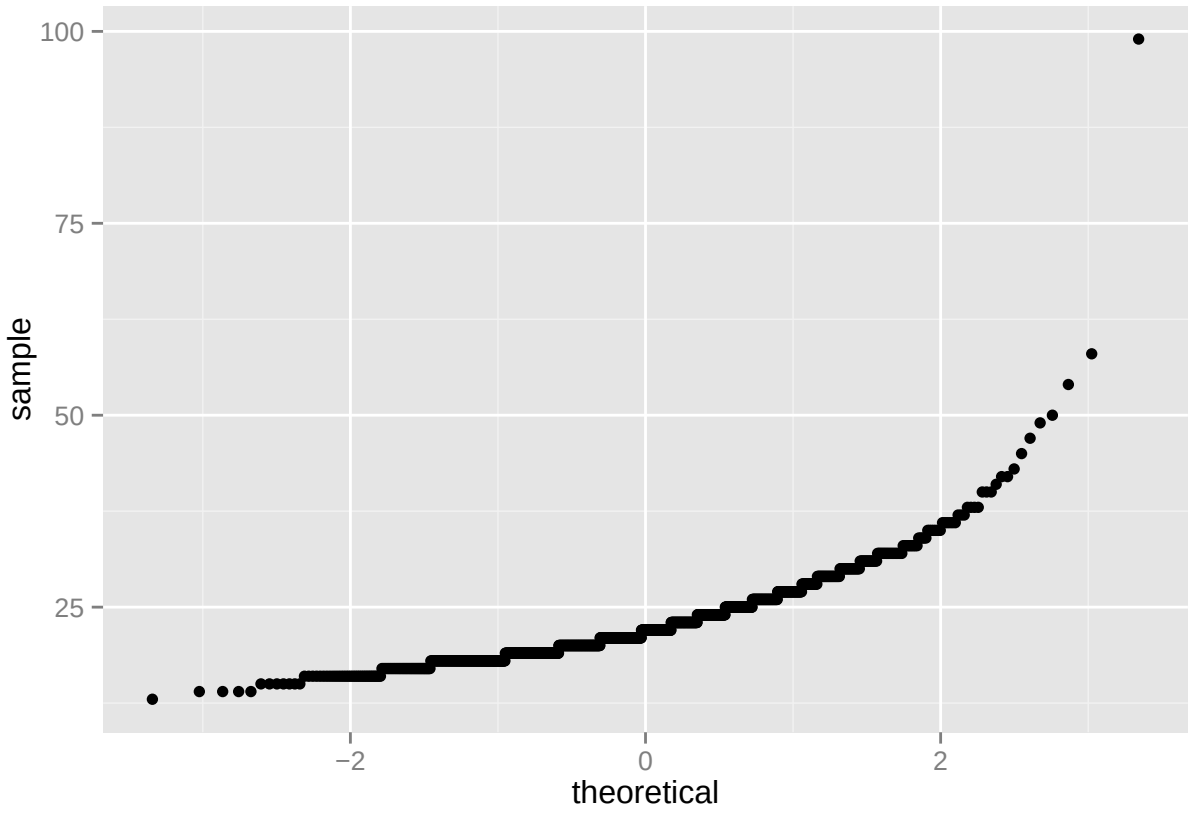
ii. What is your p-value, and what is your specific conclusion?

RESPONSE: In this case the p-value is .1418. Given it actually exceeds an alpha value of .05, this is actually not statistically significant - the hypothesis that these variables has the same variance for these two groups is actually statistically plausible.

```
#Load relevant packages
library(ggplot2)
library(car)
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:car':
##
##     logit
##
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

```
#perform a qqplot to visually inspect normality
qqplot = qplot(sample = GSS$agewed, stat="qq")
qqplot
```




```
#perform a shapiro wilk test shapiro.test(GSS$agewed)

#check gender variable
summary(GSS$sex) #great no NAs
```

```
##   Male Female
##    493    708
```

```
variances <- list(by(GSS$agewed, GSS$sex, var, na.rm = TRUE))
variances
```

```
## [[1]]
## GSS$sex: Male
## [1] 34.99648
## ------------------------------------------------------------
## GSS$sex: Female
## [1] 24.31918
```

```
#check to see if variances are similar enough
leveneTest(GSS$agewed, GSS$sex)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  2.1609 0.1418
##       1199
```

15.More hypothesis testing

a. Suppose we have a hypothesis that the age of marriage (agewed) in the population has a mean of exactly 23, with a standard deviation of 5 years

(you should assume this value is correct rather than estimating the standard deviation from the data). Perform a z-test to analyze this hypothesis.

i. What is the null and alternative hypothesis for this test?

RESPONSE: The null hypothesis is that the sample mean of agewed = 23 The alternative hypothesis is that the sample mean of agewed != 23 (conduct a two tailed test)

ii. What is your p-value, and what is your specific conclusion?

RESPONSE: In this case the value is .7338565. Given this exceeds an alpha value of .05, there is not enough evidence to reject the null hypothesis - it is not statistically significant. Thus, it is plausible that the mean of the population is 23 assuming a standard deviation of 5 years.

```
#Define function to compute p-vale
pval= function(data, mu, pop.sd, twotailed=TRUE) {
  len <- length(data)
  m <- mean(data)
  se <- pop.sd / sqrt(len)

  zscore <- (m-mu) / se
  pvalue <- (twotailed + 1 ) * pnorm(abs(zscore),0,1, lower.tail = FALSE)

  df <- data.frame(pvalue , (pvalue<=0.05))
  colnames(df) <- c("pvalue","reject")
  return(df)

  }

#sample 100 values from GSS$agewed vector
GSS.sample <- sample(GSS$agewed,100)

#run test
#pval(GSS.sample,23,5)
#commented out evaluation because it will differ between when I first ran the test and when its run to
```