# Homework_8_Lab2

*Natarajan Shankar*

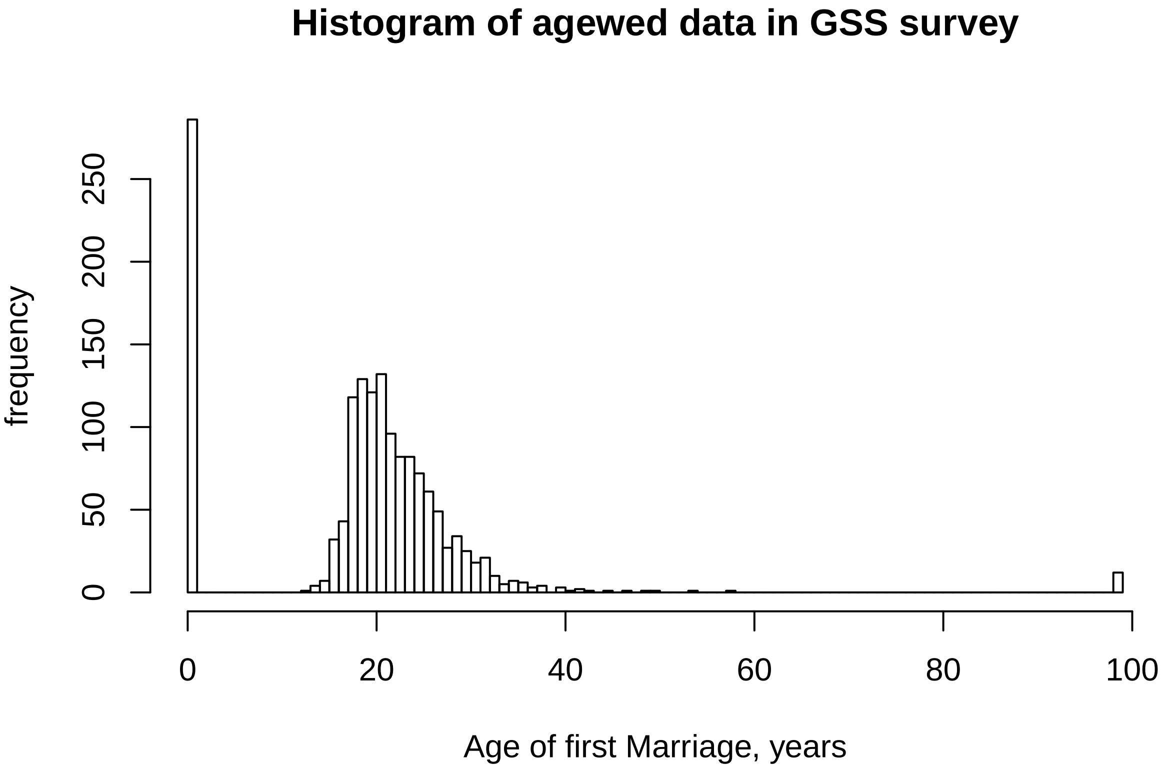
*June 29, 2016*

```r
library(ggplot2); library(plyr); library(car)

# Load the GSSS data into local storage
load("GSS.RData")

# Examine the agewed data to get a feel for what data has been gathered
# Also, confirm the number or rows in the GSS list,  there are 1500 rows and 47 columns
nrow(GSS); table(GSS$agewed)
```

```
## [1] 1500
```

```
##
##   0  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
## 286   1   4   7  32  43 118 129 121 132  96  82  82  72  61  49  27  34
##  30  31  32  33  34  35  36  37  38  40  41  42  43  45  47  49  50  54
##  25  18  21  10   5   7   6   3   4   3   1   2   1   1   1   1   1   1
##  58  99
##   1  12
```

```r
# Plot a histograme to understand the agewed data
hist(na.exclude(GSS$agewed), breaks =100, main = "Histogram of agewed data in GSS survey",
                  xlab = "Age of first Marriage, years", ylab ="frequency")
```



```
# Histogram shows a huge spike at agewed =0 and a small outlier mode at agewed = 99
# The data is not showing a NORMAL DISTRIBUTION
```

### Question 13, Part a

**There are 12 instances of an agewed value of 99, which seem spurious. There are 286 instances of an agewed value of 0, which legitimately indicates people never married (must these be treated separately?). There are instances of underage people (age below 18) - These will be disregarded these for they are likely spurious/illegal.**

1. 1 at age 13
2. 4 at age 14
3. 7 at age 15
4. 32 at age 16
5. 43 at age 17

**There are also instances where agewed is greater than age - treat these as errors in data collection**

```
# Create a new sanitized agewed column, leave the original agewed column in place
GSS$sanitizedAgewed <- GSS$agewed

# Recode the agewed values that do not meet the above requirements as NA, put into sanitizedAgewed colu
GSS$sanitizedAgewed[(GSS$agewed < 18) | (GSS$agewed > GSS$age) | (GSS$agewed == 99)] <- NA
```

### Agewed data has now been cleaned up and in the programmatic part of this work, is called sanitizedAgewed

### What is the mean of the agewed (sanitizedAgewed) variable?

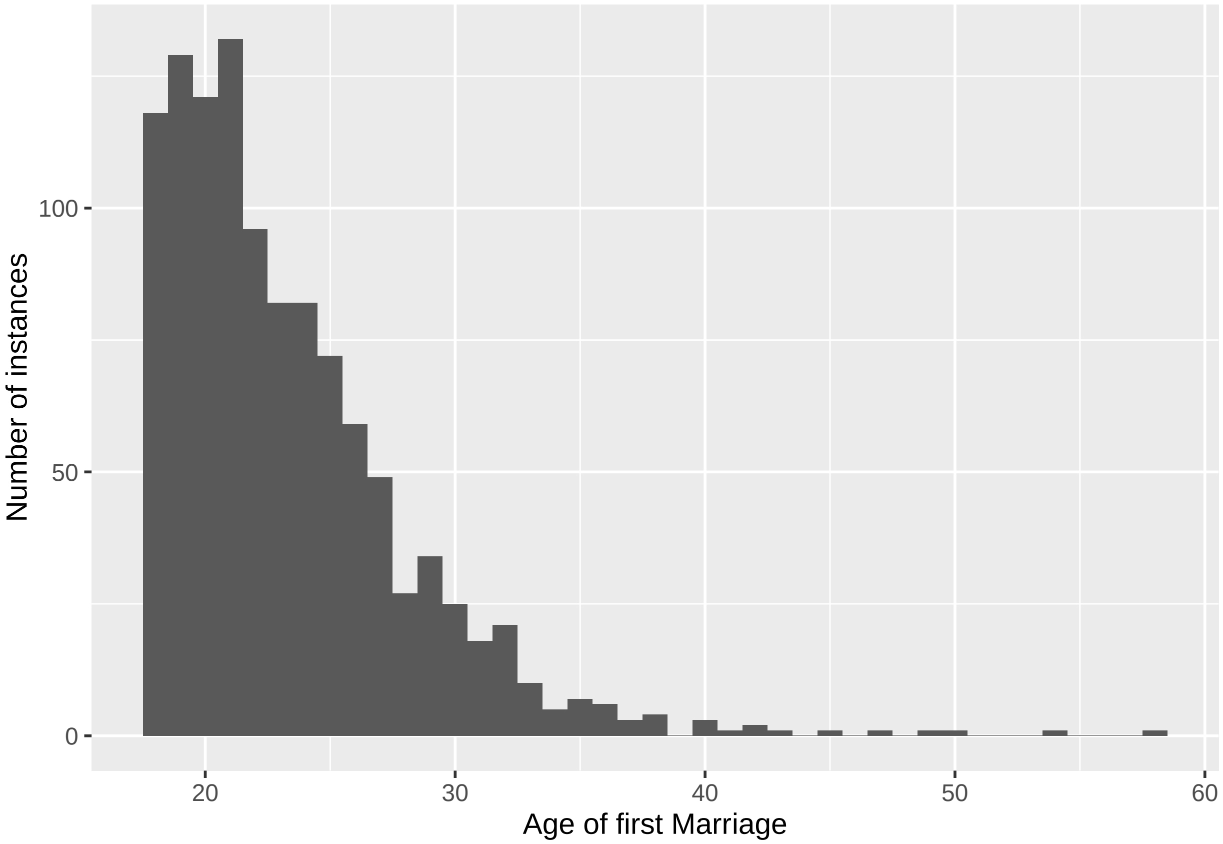
```
agewedMean <- mean(na.exclude(GSS$sanitizedAgewed))
agewedMean
```

```
## [1] 23.2947
```

### Question 13, Part b

### The mean of the re-coded agewed variable is: 23.294699

```
# Look at a histogram of sanitizedAgewed data to get a feel for the distribution of the re-coded data
agewedHistPlot <- ggplot(data.frame(na.exclude(GSS$sanitizedAgewed)),
                                    aes(na.exclude(GSS$sanitizedAgewed)))
agewedHistPlot <- agewedHistPlot + geom_histogram(binwidth=1) +
    labs(y = "Number of instances", x = "Age of first Marriage")
agewedHistPlot
```

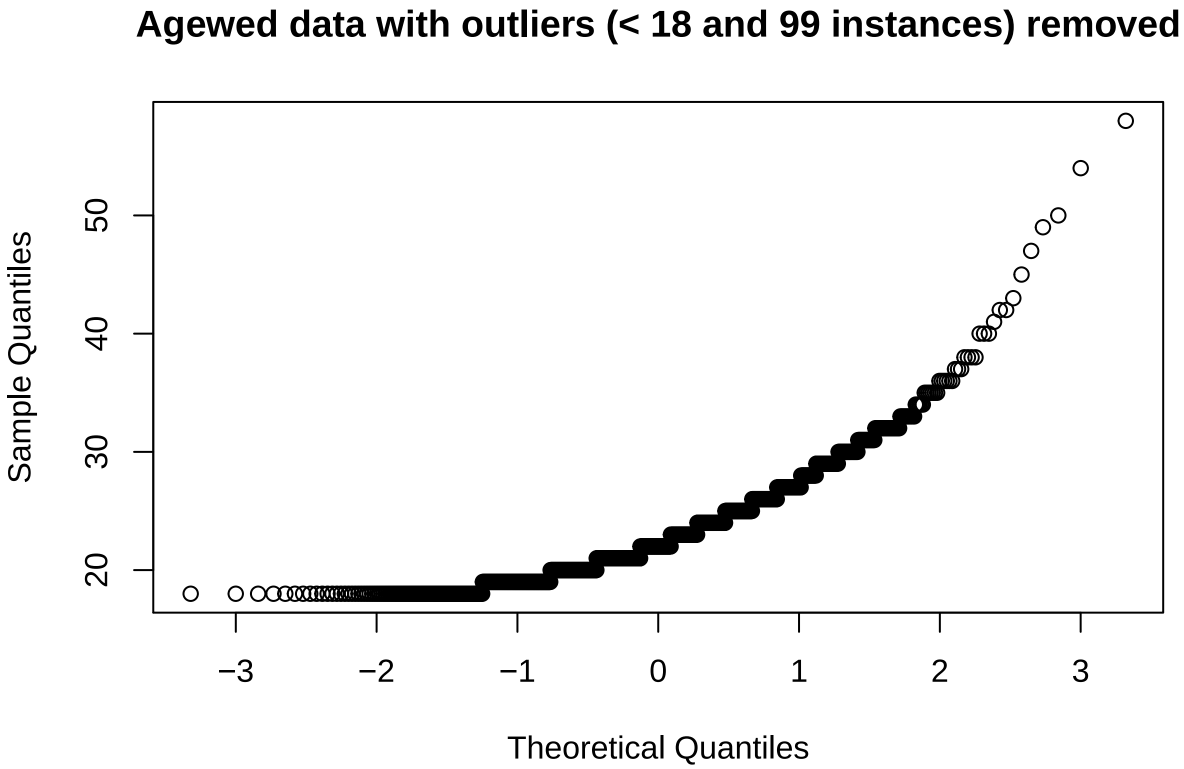


```
# The histogram shows a strong Postive skew.The data is still not showing the form of a Normal distribu
# Transformation (log) is likley need in order to achieve Normality
```

## Question 14, Part a

## Produce a Q-Q plot for the agewed variable

```
GSS$sanitizedAgewed <- GSS$agewed
GSS$sanitizedAgewed[(GSS$agewed < 18) | (GSS$agewed > GSS$age) | (GSS$agewed == 99)] <- NA
qqnorm(na.exclude(GSS$sanitizedAgewed), main = "Agewed data with outliers (< 18 and 99 instances) remove
```



when the (<18 and 99) spurious data and outliers are removed, the above graph shows a single segment for the data distribution, As expected (from the histogram view), the agewed variable in the QQ plot is not a Normal distribution. Because of the right skew of agewed, the cumulative observed value is right shifted. There is a significant skew in the collected data towards lower ages (< 40).

```
# Ensure that the sex column is made to match the agewed data. This sex data will be called upon later.
GSS$sexBackup <- GSS$sex
GSS$sex[(GSS$agewed < 18) | (GSS$agewed > GSS$age) | (GSS$agewed == 99)] <- NA
unique(GSS$sex)
```

```
## [1] Male   <NA>   Female
## Levels: Male Female
```

## Question 14, Part b

### Perform a Shapiro-Wilk test on the sanitized agewed variable

## The Null Hypothesis is that the "the samples come from a Normal distribution"

## The Alternative hypothesis is that "the samples do not come from a Normal distribution"

```r
shapiroWilkResult <- shapiro.test(na.exclude(GSS$sanitizedAgewed))
shapiroWilkResult
```

```
##
##  Shapiro-Wilk normality test
##
## data:  na.exclude(GSS$sanitizedAgewed)
## W = 0.8517, p-value < 2.2e-16
```

```r
shapiroWilkResult[[2]]
```

```
## [1] 4.655978e-31
```

## p-value from Shapiro Wilk test is $< 0.05$. CONCLUSION : Since the p-value $<=$ 0.05, the NULL hypothesis that the samples came from a Normal distribution is highly unlikely

## Question 14, Part c

## Separate the Men data from Wowen data in sanitizedAgewed, compute Variance

```r
dataForMen <- GSS[GSS$sex == "Male",]
menVariance <- var(na.exclude(dataForMen$sanitizedAgewed))
menVariance
```

```
## [1] 23.30283
```

```r
dataForWomen <- GSS[GSS$sex == "Female",]
womenVariance <- var(na.exclude(dataForWomen$sanitizedAgewed))
womenVariance
```

```
## [1] 22.80681
```

## Variance for men : 23.3028255

## Variance for Women: 22.8068137

```r
dataForMen <- GSS[GSS$sex == "Male",]
menVariance <- var(na.exclude(dataForMen$sanitizedAgewed))
menVariance
```

```
## [1] 23.30283
```

```r
dataForWomen <- GSS[GSS$sex == "Female",]
womenVariance <- var(na.exclude(dataForWomen$sanitizedAgewed))
womenVariance
```

```
## [1] 22.80681
```

## ALTERNATIVE WORK : Just out of curiosity, check whether the assumption to drop ages less that 18 has impacted the Variance. Redo variance with the agewed $< 18$ data back in

```r
GSS$testSanitizedAgewed <- GSS$agewed
GSS$testSanitizedAgewed[(GSS$agewed > GSS$age) | (GSS$agewed == 99)] <- NA
dataForMen <- GSS[GSS$sex == "Male",]
menVariance <- var(na.exclude(dataForMen$testSanitizedAgewed))
menVariance
```

```
## [1] 23.30283
```

```r
dataForWomen <- GSS[GSS$sex == "Female",]
womenVariance <- var(na.exclude(dataForWomen$testSanitizedAgewed))
womenVariance
```

```
## [1] 22.80681
```

The change in variance does not appear to show that the Variance data has been significantly affected by the decision to drop the agewed $< 18$ years data.

### Question 14 d

Perform a Levene's test for the agewed variable grouped by men and women

- The Null hypothesis is that the variances of agewed for Men and for Women are equal

- The Alternative hypothesis is that the variance in agewed for Men and Women are not equal

```
leveneResult <- leveneTest(na.exclude(GSS$sanitizedAgewed), na.exclude(GSS$sex))
leveneResult
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  2.4935 0.1146
##       1111
```

Because the p value 0.1145998, NA is greater than the alpha value of 0.05, the Variances in agewed between Man and Women are not significant to reject the Null hypothesis

ALTERNATIVE WORK: Just out of curiosity, check whether the assumption to drop ages less that 18 has impacted the Levene Test results. Redo variance with the agewed $< 18$ data back in. The expectation is that the variance will show a major change

```
GSS$sexBackup[(GSS$agewed > GSS$age) | (GSS$agewed == 99)] <- NA
leveneResult <- leveneTest(na.exclude(GSS$testSanitizedAgewed), na.exclude(GSS$sexBackup))
leveneResult
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    1  15.684 7.841e-05 ***
##       1484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**ALTERNATIVE WORK CONCLUSION: The exclusion of agewed < 18 has had a major impact. By inclusion of agewed < 18 data, the Levene Null hypothesis is rejected because the p-value of 7.841e-05 is less than 0.05. By non-inclusion of the agewed < 18 data, the p value (0.1146) shows that the Levene Null hypothesis cannot be rejected**

## Question 15 Part a

### - Null hypothesis is that mean = 23. SD is 5 years

### - The alternative hypothesis is that mean != 23

```
m0 <- 23 # The population mean
sigma <- 5 # The poupulation Standard Deviation
sampleSize <- 50 # Pick a sample size > 30

# Get a sample of size sampleSize from the sanitized agewed data
asample <- sample(na.exclude(GSS$sanitizedAgewed), size = sampleSize, replace = TRUE)

# Compute the mean of the sample
mu <- mean(asample)
mu
```

```
## [1] 22.94
```

```
# Compute  the Z value using standard formula and output
zvalue <- (mu - m0)/(5/sqrt(sampleSize))
zvalue
```

```
## [1] -0.08485281
```

```
# a two tailed test is being performed. Double the p value to cover both tails
pvalue <- 2*pnorm(zvalue)
pvalue
```

```
## [1] 0.9323784
```

### Question 15 Part b

**The p value is 0.9323784**

**Because the p-value is bigger than the chosen significance level of 0.05, the Null hypothesis cannot be rejected**

**ALTERNATIVE WORK: Just out of curiosity, check whether the assumption to drop agewed less than 18 has impacted the Z value calculation. Redo the Z calculation with the agewed $<$ 18 data back in. The expectation is that the z-value will show a major change**

```r
# Get a sample of size sampleSize from the sanitized agewed data
asample <- sample(na.exclude(GSS$testSanitizedAgewed), size = sampleSize, replace = TRUE)

# Compute the mean of the sample
mu <- mean(asample)
mu
```

```
## [1] 18.66
```

```r
# Compute  the Z value using standard formula and output
zvalue <- (mu - m0)/(5/sqrt(sampleSize))
zvalue
```

```
## [1] -6.137687
```

```r
# a two tailed test is being performed. Double the p value to cover both tails
pvalue <- 2*pnorm(zvalue)
pvalue
```

```
## [1] 8.37317e-10
```

**ALTERNATIVE WORK AND OVERALL CONCLUSION: The choice to exclude agewed $< 18$ seems to impact the z test results greatly. When excluded, the Null hypothesis cannot be rejected. When included, the Alternative hypothesis is more likely.**

**The decision to go one way or the other is one that will dependent upon the validity of the agewed data that is in the GSS survey. I originally theorized (agewed below 18 is illegal in most US states) that this data reflecting agewed $< 18$ is invalid.**

**I present data both ways for Shapiro-Wilk, Levene and Z-tests, the analysis contributed much to my learning.**