MIDS W203 Lab 2

Matthew Potts

March 1, 2016

Part 1. Multiple Choice

1. A
2. C
3. F
4. E
5. D
6. B
7. D
8. F
9. D
10. F

Part 2. Test Selection

11. B (The Levene's test is used to test the null hypothesis of equal variances)
12. A (Shapiro-Wilk test tests the null hypothesis that the variable is normally distributed)

Part 3. Data Analysis and Short Answer

Part 13a. Examine the "agewed" variable (age when married).

```r
# Load libraries, load GSS data.
library(ggplot2)
library(car)
load("GSS.Rdata")

# Look at a summary of the agewed variable
summary(GSS$agewed)
```

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

```r
# Look at the frequency of certain values, particularly 0.
stem(GSS$agewed)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 000000000000000000000000000000000000000000000000000000000000000000+206
##   0 |
##   1 | 34444
##    1 | 55555556666666666666666666666666666666677777777777777777777777777777+249
##    2 | 00000000000000000000000000000000000000000000000000000000000000000000+433
##    2 | 55555555555555555555555555555555555555555555555555555555555555555555+163
##    3 | 00000000000000000000000001111111111111111112222222222222222222223333
##    3 | 55555556666667778888
##    4 | 0001223
##    4 | 579
##    5 | 04
##    5 | 8
##    6 |
##    6 |
##    7 |
##    7 |
##    8 |
##    8 |
##    9 |
##    9 | 999999999999
```

```r
# Find weird values where agewed is greater than current age convert to NA.
weird <- GSS$id[GSS$agewed > GSS$age]
weird
```

```
##  [1]  215  359  361  565  595  609  848  853 1013 1222 1339 1581 1594
```

13a. i) There are some 0 values and some 99 values that do not make sense. There are also 13 values (mostly agewed = 99, and a few agewed = 26) where the agewed is greater than the age of the person.

13b. Recode values that do not meaningfully correspond to ages as NA.

```r
# Take the values of agewed that are greater than age and convert to NA.
for(i in weird) {
  GSS$agewed[GSS$id == i] <- NA
}

# Turn agewed = 0 to NAs
GSS$agewed[GSS$agewed == 0] <- NA

# Turn agewed = 99 to NAs
GSS$agewed[GSS$agewed == 99] <- NA
```

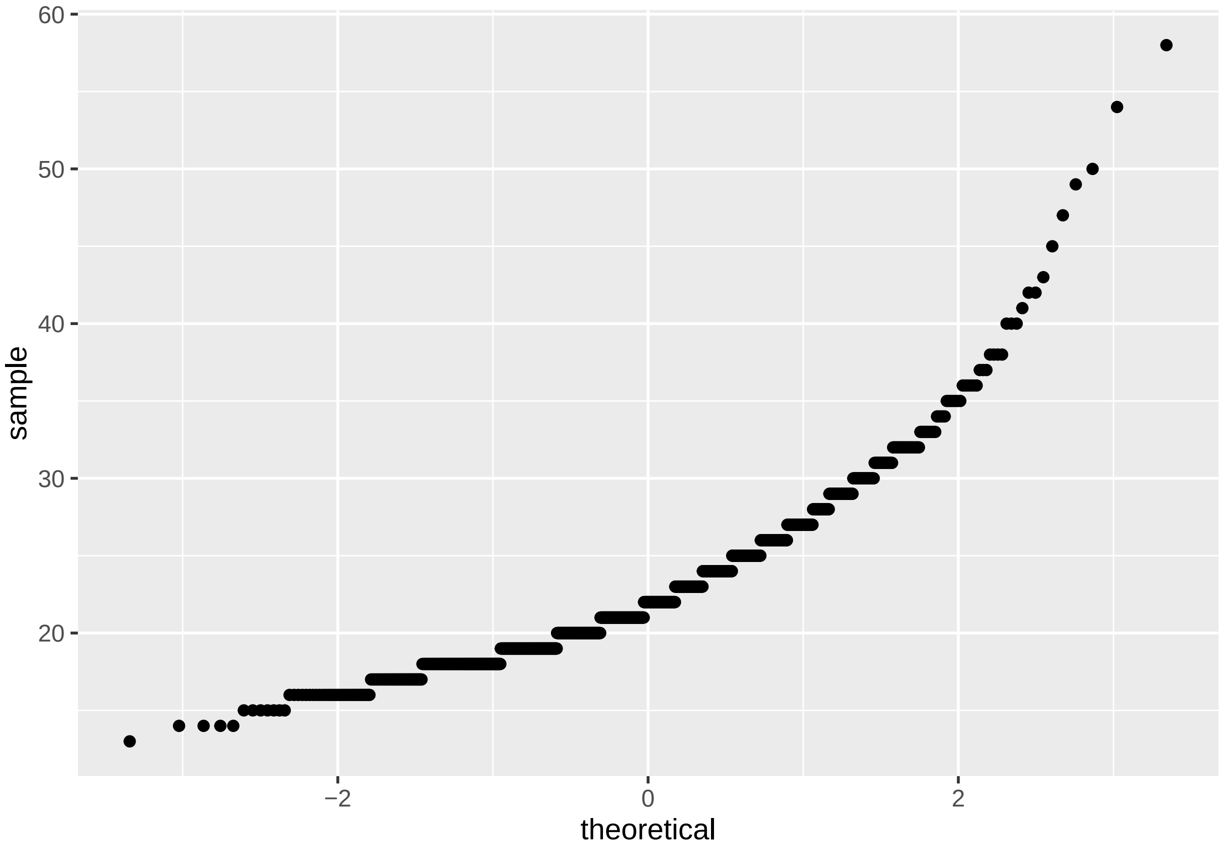
13b. i) Find the mean of the agewed variable.

```r
mean(GSS$agewed, na.rm = TRUE)
```

```
## [1] 22.78667
```

Part 14a. Produce a QQ plot for the agewed variable.

```r
suppressWarnings(qplot(sample = GSS$agewed, stat="qq", na.rm = TRUE))
```



14a. i) The plot shows that the variable agewed is not normal. A normal distribution would show a linear trend. The agewed variable does not show a linear trend in the QQ plot.

Part 14b. Perform a Shapiro-Wilk test on the agewed variable.

```r
shapiro.test(GSS$agewed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  GSS$agewed
## W = 0.88914, p-value < 2.2e-16
```

14b. i) The null hypothesis for the Shapiro-Wilk test is that the data is normally distributed. The alternate hypothesis is that the data in not normally distributed. **ii)** The p-value < 2.2e-16. This rejects the null at a low probability meaning that the data is most likely not normally distributed.

Part 14c. What is the variance of agewed for men? What is the variance of agewed for women?

```r
# Calculate variance of agewed for men
var(GSS$agewed[GSS$sex == "Male"], na.rm = TRUE)
```

```
## [1] 23.6843
```

```r
# Calculate variance of agewed for women
var(GSS$agewed[GSS$sex == "Female"], na.rm = TRUE)
```

```
## [1] 24.31918
```

Part 14d. Perform a Levene's test for the agewed variable grouped by men and women.

```r
leveneTest(GSS$agewed, GSS$sex)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  1.0006 0.3174
##       1198
```

14d. i) The null hypothesis for the Levene's Test is that the variance of the agewed variable for men and women is the same. The alternate hypothesis is that the variance of the agewed for men and women is not the same. **ii)** The probability value is 0.3174. This probability value is not statistically significant. The hypothesis that male and female have the same agewed variance is statistically plausible.

Part 15a. 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. The equation for a z-test is below.

$$z = (\mu - \mu_0)/(\sigma/\sqrt{N})$$

```r
# Perform a z-test to analyze the hypothesis.
z <- (mean(GSS$agewed, na.rm = TRUE) - 23)/(5/sqrt(length(na.omit(GSS$agewed))))
z
```

```
## [1] -1.478017
```

```r
# Calculate a p-value for a two-tailed test.
pvalue <- 2*(1-pnorm(abs(z)))
pvalue
```

```
## [1] 0.1394033
```

15a. i) The null hypothesis is that the agewed mean of the population and the sample is the same. The alternate hypothesis is that the agewed mean of the population and the sample is not the same. **ii)** The p-value is 0.1394. Assuming an α of 0.05 or 0.10, we cannot reject the null hypothesis that the sample mean equals population mean.