Audrey Leung October 30, 2014

Part 1 – Multiple Choice

1) E
2) A
3) D
4) B
5) D
6) B
7) D
8) B
9) F

Part 2 – Data Analysis and Short Answer

1. Data Import and Error Checking: Using the GSS dataset.
   a. Examine the "agewed" variable (age when married).
      i. What are the value(s) of agewed, if any, that do not meaningfully correspond to ages?

```
> table(GSS$agewed)

  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
```

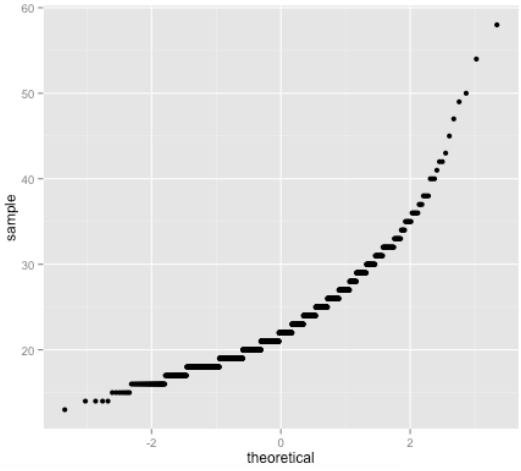
      After examining the table of values for the agewed variable, the values '0' and '99' did not meaningfully correspond to ages. '99' is a code that means that there was no data or the data was missing. '0' is code for never married. Since we are only interested in the age that people were married, neither of these codes would be relevant.

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

```
> mean(GSS$agewed, na.rm=TRUE)
[1] 22.79201
```

   I recoded '0' and '99' as NA. The mean of the agewed variable excluding 0 and 99 was 22.92.

2. Checking assumptions
   a. Produce a QQ plot for the agewed variable.



i. Using this plot information, is agewed normal and how precisely do you know?
Using the plot information, agewed is not normal. Since the data in the QQ plot is deviating from the diagonal, we know that the distribution is not symmetrical. If it is normal, the plot should be roughly a straight diagonal line. The plot suggests that the values are more clustered around the low-end of the scale since the plot line starts low and then rises exponentially. The QQ plot shows the deviation from normality, but not precisely by how much the data is positively skewed.

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

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

        Shapiro-Wilk normality test

data:  GSS$agewed
W = 0.8896, p-value < 2.2e-16
```

i. What is the precise null and alternative hypothesis for your test?
The null hypothesis is that the data has a normal distribution; the distribution of the sample is the same as a normal distribution.
The alternative hypothesis is that the data has a non-normal distribution; the distribution of the sample is not the same as a normal distribution.

ii. What is your p-value, and what is your specific conclusion?
My p-value is < 2.2e-16. Since the p-value < 0.05, this means our test is highly significant. Thus, we reject the null hypothesis and accept our alternative hypothesis; the data has a non-normal distribution.

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

```
> by(GSS$agewed, GSS$sex, var, na.rm=TRUE)
GSS$sex: Male
[1] 23.6843
------------------------------------------
GSS$sex: Female
[1] 24.29948
```

The variance of agewed for men is 23.68. The variance of agewed for women is 24.30.

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

```
> leveneTest(GSS$agewed,GSS$sex,na.rm=TRUE)
Levene's Test for Homogeneity of Variance (center = median: TRUE)
       Df F value Pr(>F)
group   1  0.9609 0.3272
      1200
```

i. What is the precise null and alternative hypothesis for this test?
The null hypothesis is that the variance for the agewed variable grouped by men is equal to the variance for the agewed variable grouped by women.
The alternative hypothesis is that the variance for the agewed variable grouped by men is not equal to the variance for the agewed variable grouped by men.

ii. What is your p-value, and what is your specific conclusion?
The p-value is 0.33. The variances were similar for males and females, F(1,1200) = 0.33. Since the test results in an insignificant p-value > 0.05, this means that we have equal variances between the male and female groups, and we fail to reject the null hypothesis.

Audrey Leung October 30, 2014

3. More on Plots and Visualizing Data

a. Pick one metric variable from the dataset (other than agewed) and any other variable of interest to you. Produce a plot or visualization that allows you to look at your two variables in an interesting or informative way. (Feel free to get creative and try some things with ggplot, but you are welcome to produce any type of plot or graphic that you want no matter how simple or complex)

i. Write a short paragraph describing what your chosen plot or graphic shows about your variables (be descriptive!).

This two-dimensional box-plot shows the TV hours watched vs. political views, for men and women. This plot shows that the inner quartile range of males watch less TV if they have more liberal political beliefs than conservative. Females who tend conservatively watch a wider range and higher upper quartile of TV hours than males who tend conservatively. Conversely, females who tend liberally watch a wider range and lower lower quartile of TV hours than males who tend liberally.