LAB 2

PART I. MULTIPLE CHOICE

1. a) bar graphs
2. c) H_0: $\mu = \mu_0$; H_a: $\mu > \mu_0$
3. a) Our alpha is less than .10.
4. e) a and d
5. e) Raise the variable to a power greater than 1
6. b) The standard deviation of Berkeley student ages is 2 years.
7. d) What is the probability of the data we observe, assuming that the null hypothesis is true?
8. f) Assuming your null hypothesis is actually true, and you were to repeat the experiment a large number of times, you would expect a type 1 error 4% of the time.
9. d) Independence of observations.
10. f) None of the above.

PART II. TEST SELECTION

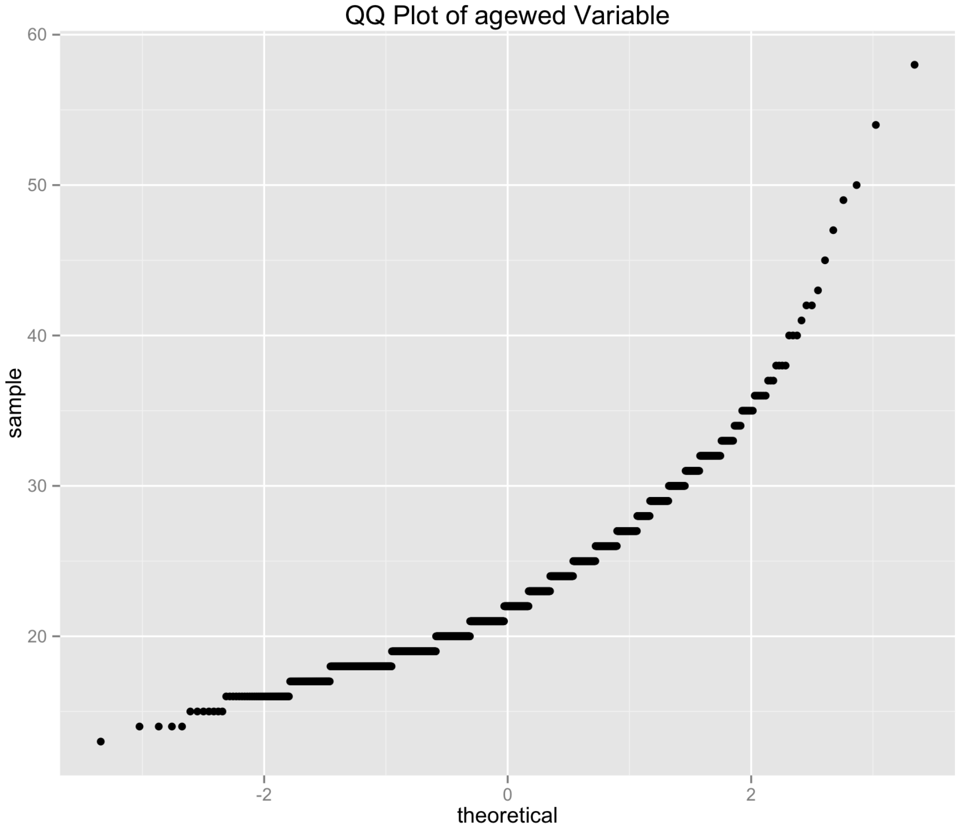
11. b) Levene's test
12. a) Shapiro-Wilk test

PART III. DATA ANALYSIS AND SHORT ANSWER

13. a) When we examine the agewed variable, we find a number of ages that may not meaningfully correspond to ages. We find 286 respondents coded with an agewed of 0. It is likely that these respondents have not married rather than that they married when they were born. We also have 12 respondents with an agewed of 99. This also seems unlikely given that the next highest agewed is 58. Perhaps respondents coded with the age 99 are married but refused to provide their age on the survey. We should replace both the 0's and 99's in the aged variable with NA's.

    b) The mean age of the agewed when excluding NA's is 22.79201 years.

14. a) Based on the QQ Plot (see below), the variable aged is not normal because a normal distribution has a straight line QQ plot that would perfectly cut from the bottom left corner to the top right corner of the graph. Instead, the QQ plot for the agewed variable is closer to a quadratic function, slowly increasing at the beginning and then speeding up towards the end.



b)

i. Null Hypothesis: The distribution of agewed is normal.
Alternative Hypothesis: The distribution of agewed is not normal.

ii. My p-value from the Shapiro-Wilk test is 2.2e-16 (well below .05). I reject my null hypothesis. I conclude that the distribution of the agewed variable is unlikely to be normal.

c) The variance of agewed for men is 23.6843 and 24.29948 for women.

d)

i. Null Hypothesis: The variances of agewed for men and for women are equal.
Alternative Hypothesis: The variances of agewed for men and for women are not equal.

ii. My p-value from Levene's test is .3272. I fail to reject the null hypothesis and conclude that the observed differences in the variation of agewed by sex are not statistically significant.

15. a) Per the prompt, I will assume a population standard deviation of 5.

i. Null Hypothesis: The mean age of marriage (agewed) is 23.
Alternative Hypothesis: The mean age of marriage (agewed) is not 23.
I will reject the null hypothesis at a significance level of 5% ($p<.05$).

ii. My p-value is 0.1492532. I fail to reject the null hypothesis. I conclude that the mean of my variable agewed is not significantly different than the hypothesized population mean of 23.

APPENDIX. R SCRIPT

```r
# Alex Smith
# W203 - Exploring and Analyzing Data
# Lab 2

# load packages, ggplot2 and car, install if don't have
library(ggplot2)
library(car)

# set working directory to make it easier to pull appropriate files
setwd("~/Documents/MIDS/Spring14/W203/Lab2")

# load the General Social Survey data in a data frame
load("GSS.Rdata")

# get a look at the data as a whole
head(GSS)

# examine the agewed variable (the age at which couples have married)
summary(GSS$agewed)
table(GSS$agewed)

# recode values of 0 and 99 in agewed variable to "NA"
GSS$agewed[GSS$agewed==0] <- NA
GSS$agewed[GSS$agewed==99] <- NA

# calculate the mean agewed value, excluding NA's
mean_agewed = mean(GSS$agewed, na.rm = TRUE)
mean_agewed

# create a QQ plot for the agewed variable
qq_agewed <- qplot(sample = GSS$agewed, stat = "qq") + ggtitle("QQ Plot of agewed Variable")
qq_agewed

# run a Shapiro-Wilk test to test for normality
shapiro.test(GSS$agewed)

# run a Shapiro-Wilk test, separating out men and women
# by(GSS$agewed,GSS$sex,shapiro.test)
# commented out because unnecessary for prompt

# calculate the variance of agewed by sex
by(GSS$agewed,GSS$sex,var,na.rm=TRUE)

# perform a Levene's test on the agewed variable by gender
leveneTest(GSS$agewed,GSS$sex)

# perform a z-test
# assume that the population standard deviation is 5
# null hypothesis: mean of agewed = 23
# alternative hypothesis: mean of agewed != 23
# calculate a z-score using the formula z = (theorized mean - calculated mean)/(sigma/sqrt(sample size))
theorized_mean = 23
calculated_mean = mean(GSS$agewed,na.rm=TRUE)
z_numerator = theorized_mean - calculated_mean
sigma = 5
sample_size = sum(GSS$agewed != "NA", na.rm = TRUE)
z_denominator = sigma/(sqrt(sample_size))

z_score = z_numerator/z_denominatorz

# use the z-score to calculate the probability of calculated average given the null hypothesis
p_value = (1-pnorm(z_score))*2
p_value
```