title: "W203 Week 8 Lab 2" author: "Mohammad Jawad Habib" date: "March 1, 2016" output: pdf_document —

Part 1: Multiple Choice

1. a - Bar Graphs
2. c -

$$H_0 : \mu = \mu_0; H_a : \mu > \mu_0$$

3. f - none of the above
4. e - Type II error will go up, Power will go down
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. c - Assuming your null hypothesis is actually false, your p-value is likely to decrease as you increase your sample size
9. d - Independence of observations
10. f - None of the above

Part 2: Test Selection

11. b - Levene's Test
12. a - Shapiro-Wilk Test

Part 3: Data Analysis and Short Answer

```
require(knitr)
```

```
## Loading required package: knitr
```

```
setwd("~/Exploring and Analyzing Data/W203 Async/W203 Week 8/")

load("GSS.Rdata")
```

13. Data Import and Checking

13.a. Examine `agewed`

```
sort(unique(GSS$agewed))
```

```
##  [1]  0 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
## [24] 35 36 37 38 40 41 42 43 45 47 49 50 54 58 99
```

```r
sort(unique(GSS$agewed[GSS$agewed < 18]))
```

```
## [1]  0 13 14 15 16 17
```

```r
sort(unique(GSS$agewed[GSS$agewed > 58]))
```

```
## [1] 99
```

We can assume that `GSS$agewed == 0` and `GSS$agewed == 99` are not reasonable ages to get married. It would also seem that `GSS$agewed < 18` is not a reasonable age to get married. Marriageable age is 18 in most US states and most other countries, and parental consent is required in the US marry younger than that. However, let's assume that there were cases where people got married younger than 18 and consider those ages valid. It must be noted that in some backward countries like Pakistan, children are sometimes "married"" at very young ages, sometimes even at birth, without their consent. That's a horrible practice but we cannot fix that in this assignment. We can only hope that these people will come to their senses soon.

13.b. Recode unreasobale `agewed` as NA

```r
# As noted before, we will only recode 0 and 99
# Yes, we're predujiced against the very old when it comes to marriage
GSS$agewed[GSS$agewed == 0] <- NA
GSS$agewed[GSS$agewed == 99] <- NA

# Calculate the mean, ignoring NA
agewed.mean <- mean(GSS$agewed, na.rm = TRUE)
agewed.mean
```

```
## [1] 22.79201
```

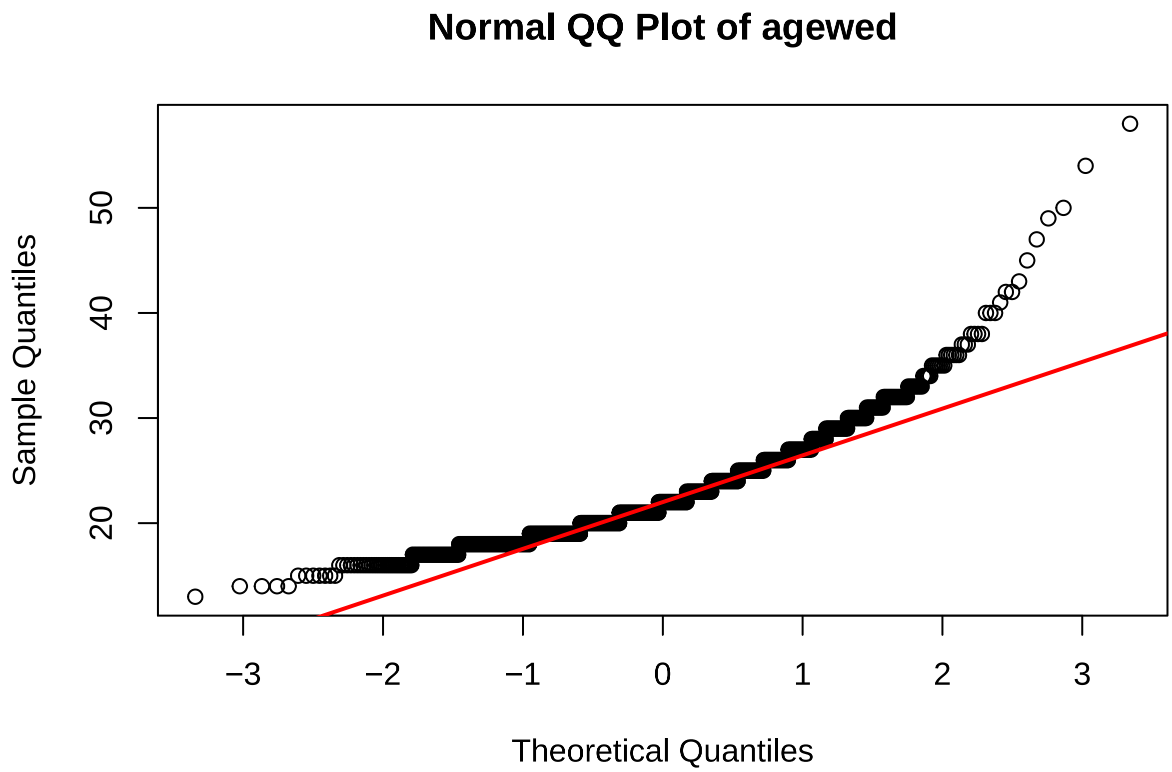
It can be seen that the mean `agewed` is 22.7920133 when `0` and `99` are recoded as `NA`.

14. Checking assumptions

14.a. Produce a QQ plot of `agewed`

We will use `qqnorm` to generate the qqplot and add a line that passes through first and third quantiles with `qqline`. The plot will ingore the values that we set to `NA`

```r
qqnorm(GSS$agewed, main = "Normal QQ Plot of agewed")
qqline(GSS$agewed, col = 2, lwd = 2)
```



As we can see from the graph, the `agewed` variable is not normally distributed. This is because for a normal distribution, the data points will closely follow the `qqline`.

14.b. Perform a Shapiro-Wilk Test

We will perform a Shapiro-Wilk test to determine if `agewed` is normally distributed. The **null hypothesis** in Shapiro-Wilk test is that the variable is **normally distributed**. A significant p-value in Shapiro-Wilk test tells us that the we can reject the null hypothesis of normal distribution for the variable.

```
shapres <- shapiro.test(GSS$agewed)
shapres
```

```
##
##  Shapiro-Wilk normality test
##
## data:  GSS$agewed
## W = 0.88959, p-value < 2.2e-16
```

We see from the above output that `p-value < 2.2e-16` which is significantly below p-value of 0.05. Therefore, we can reject the null hypothesis of normality. Here, I think the p-value is limited by `.Machine$double.eps` which happens to be `2.220446e-16` on my environment.

Formally, "The agewed, W = 0.89 and p-value < 2.2e-16, was significantly non-normal."

We can also use the `stat.desc` function from `pastecs` package to see if our variable fits a normal distribution.

```r
require(pastecs)
```

```
## Loading required package: pastecs
```

```
## Loading required package: boot
```

```r
statres <- stat.desc(GSS$agewed, basic = FALSE, norm = TRUE, p = 0.95)
statres
```

```
##       median         mean      SE.mean CI.mean.0.95          var
## 2.200000e+01 2.279201e+01 1.451678e-01 2.848106e-01 2.533056e+01
##      std.dev     coef.var     skewness     skew.2SE     kurtosis
## 5.032948e+00 2.208207e-01 1.653714e+00 1.171786e+01 5.340543e+00
##     kurt.2SE   normtest.W   normtest.p
## 1.893660e+01 8.895862e-01 1.816354e-28
```

The above output shows us `skew.2SE` and `kurt.2SE` which are skew and kurtosis divided by two standard errors. At a critical value of $p < 0.05$, we can compare these two values to 1.96/2 (0.98). As we see above both `skew.2SE` of 11.7178552 and `kurt.2SE` of 18.9365968 are greater than 0.98 and therefore we cannot accept that `agewed` is normally distributed.

We also get a `normtest.p` of `1.816354e-28` that is signinifcantly less than 0.05 and therefore, per Shapiro-Wilk test, we can reject the null hypothesis of normal distribution for `agewed`. As you can see, the `stat.desc` function was able to compute an actual p-value.

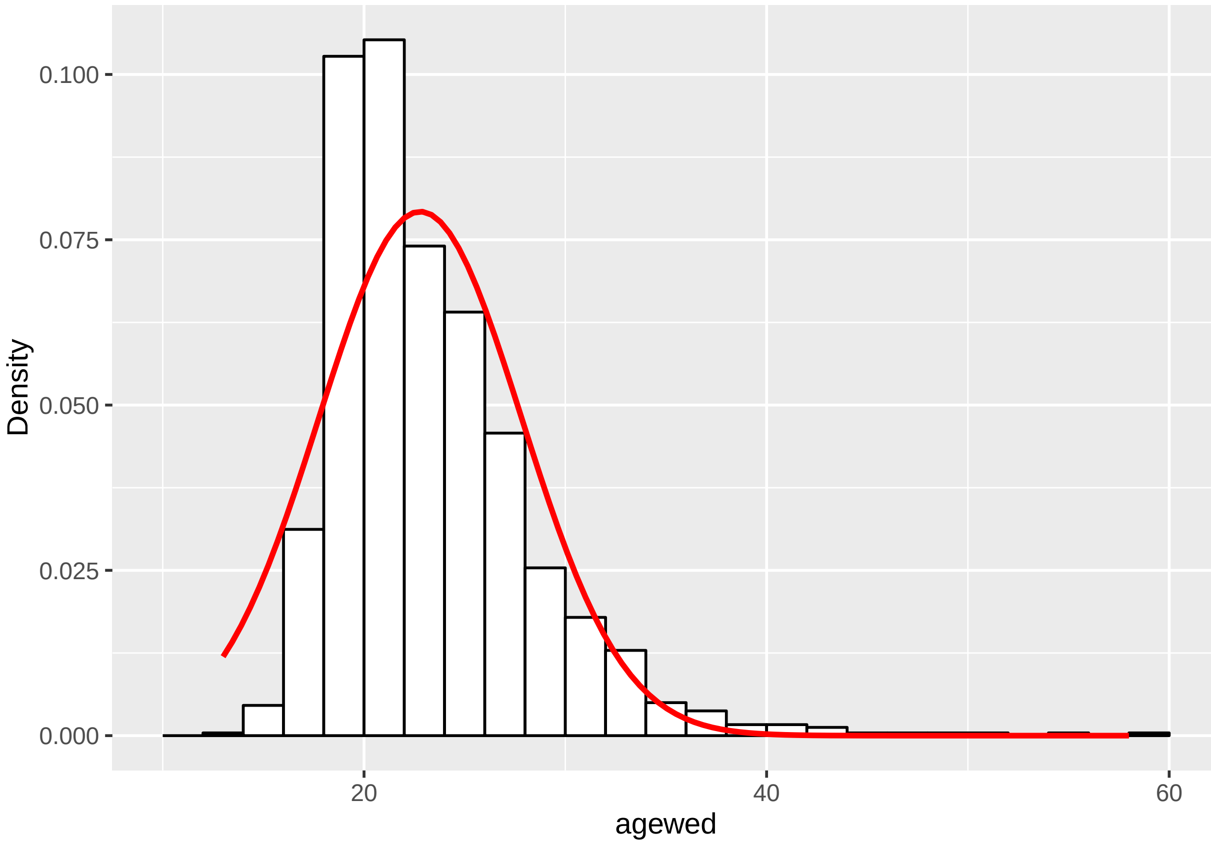
And finally, here's a quick histogram with a density curve for the visually minded.

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
agewed.hist <- ggplot(GSS, aes(agewed)) +
  geom_histogram(aes(y = ..density..), fill = "white", colour = "black", binwidth = 2) +
  labs(x = "agewed", y = "Density") +
  stat_function(fun = dnorm, args = list(mean = mean(GSS$agewed, na.rm = TRUE),
                                         sd = sd(GSS$agewed, na.rm = TRUE)), colour = "red", size = 1)
agewed.hist
```

```
## Warning: Removed 298 rows containing non-finite values (stat_bin).
```



As you can see that histogram is leptokurtic and has a positive skew.

14.c. What is the variance for `agewed` for men and women

Let's calculate variance for men.

```
agewed.var.men <- var(GSS$agewed[GSS$sex == "Male"], na.rm = TRUE)
agewed.var.men
```

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

The variance in `agewed`for men is 23.6843012.

Let's repeat the same process for women.

```
agewed.var.women <- var(GSS$agewed[GSS$sex == "Female"], na.rm = TRUE)
agewed.var.women
```

```
## [1] 24.29948
```

The variance in `agewed` for women is 24.2994815.

14.d. Perform a Levene's Test for `agewed`

14.d.i. Levene's test assumes homogeneity of variance in different groups under test. In our case the **null hypothesis** is that variance are equal in `agewed` for men and women.

We will use the `leveneTest` function from `car` package for our test.

```
require(car)
```

```
## Loading required package: car
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

```
lr <- leveneTest(y = GSS$agewed, group = GSS$sex)
lr
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  0.9609 0.3272
##       1200
```

14.d.ii. In `leveneTest` we can assume that the result is significant if the value shown under Pr(>F) is **less than** 0.05. Our result shows a Pr(>F) value of `0.3272` meaning that the result is non-significant. That is, we cannot reject the homogeneity of variances between `agewed` for men and women.

Formally, "For `agewed`, the variances were similar for men and women, 1200) = `0.3272`".

15. More hypothesis testing

Assumptions: Age of marriage in population has mean = 23 and sd = 5.

15.a.i. Null hypothesis is that mean = 23, and alternative hypothesis is that mean does not equal 23 (two-tailed test).

$$H_0 : \mu_0 = 23, H_a : \mu_a! = 23$$

15.a.ii. Let's calculate the p-value using a two-tailed test.

```
mu.0 <- 23 # population mean
sd.0 <- 5 # population standard deviation

z.value <- (mean(GSS$agewed, na.rm = TRUE) - mu.0) / (sd.0 / sqrt(length(GSS$agewed[!is.na(GSS$agewed)]
z.value
```

```
## [1] -1.442174
```

```r
p.value <- pnorm(-abs(z.value), lower.tail = TRUE) * 2
p.value
```

```
## [1] 0.1492532
```

```r
ifelse(p.value < 0.025, TRUE, FALSE) # Reject Null Hypothesis or not
```

```
## [1] FALSE
```

From the calculation above, we get a z-score of -1.4421744 which corresponds to a p-value of 0.1492532 for a two-tailed test. At a significance level of $p = 0.05$ we fail to reject the null hypothesis that population mean is 23.