Stats 10 Lab 5

Name: Brandon Truong
UID: 705326387

Section 1

```
flint <-read.csv('~/UCLA Coursework/STATS 10/flint_2015.csv',
header=TRUE)
```

a)

P0 = the proportion of dangerous lead levels in flint

H0: p0 = 10% vs H1: p0 > 10%

One-sided test

b)

```
> n <- nrow(flint)
> dangerous_lead_indicator <- (flint$Pb >= 15)
> p_hat <- mean(dangerous_lead_indicator)
> sd_sample <-sqrt(p_hat*(1-p_hat)/n)
```

c)

```
> p_null <- 0.10
> se_null <- sqrt(p_null*(1-p_null)/n)
> z_stat <- (p_hat-p_null)/se_null
> print(z_stat)
[1] 1.848714
```

d)

```
> #H1: p > 0.10
> p_value <- 1-pnorm(z_stat,sd=1,mean=0)
> print(p_value)
[1] 0.03224953
```

e)

We reject the null hypothesis since we obtain a p-value of 0.0322, which tells us that if in fact the true population proportion of dangerous lead levels is 0.10, the probability of getting a random sample where the sample proportion is greater than 0.10 or higher would be 0.0322, which is statistically significant.

e)

We should tell the EPA that households in Flint need remediation action to be taken since we rejected the null hypothesis in favor of the alternative.

g)

Our results do not change since the we get roughly the same values, although with a bit of variance due to the prop.test continuity correction.

```
> library(mosaic)
```

```
> prop.test(x=sum(dangerous_lead_indicator), n=n, p=0.10,
alternative="greater")

      1-sample proportions test with continuity correction

data:  sum(dangerous_lead_indicator) out of n
X-squared = 3.1579, df = 1, p-value = 0.03778
alternative hypothesis: true p is greater than 0.1
95 percent confidence interval:
 0.101559 1.000000
sample estimates:
        p
0.1238447

> c(p_hat,p_value)
[1] 0.12384473 0.03224953
```

h)

```
> prop.test(x=sum(dangerous_lead_indicator), n=n, p=0.10,
alternative="greater", conf.level = 0.99)

      1-sample proportions test with continuity correction

data:  sum(dangerous_lead_indicator) out of n
X-squared = 3.1579, df = 1, p-value = 0.03778
alternative hypothesis: true p is greater than 0.1
99 percent confidence interval:
 0.09376523 1.00000000
sample estimates:
        p
0.1238447
```

Section 2

a)

H0: phat1 - phat2 = 0 vs H1: phat1 - phat2 != 0

Two sided test

b)

```
> flint_north <- flint[flint$Region =="North",]
> n_north <- nrow(flint_north)
> flint_south <- flint[flint$Region == "South",]
> n_south <- nrow(flint_south)

> p_hat_north <- mean(flint_north$Pb>=15)
> p_hat_south <- mean(flint_south$Pb>=15)
```

```
> p_hat_pooled <- mean(flint$Pb >= 15)

> SE <- sqrt(p_hat_pooled*(1-p_hat_pooled) * (1/n_north + 1/n_south))
> z_stat <- (p_hat_north-p_hat_south-0)/SE
> print(z_stat)
[1] 3.572283
```

c)

```
> p_value <- (1-pnorm(abs(z_stat), mean=0,sd=1)) * 2
> print(p_value)
[1] 0.0003538831
```

d)

We reject the null since our p-value is statistically significant if the significance level $\alpha = 0.05$.

e)

```
> library(mosaic)
> x_north <- sum(flint_north$Pb>=15)
> x_south <- sum(flint_south$Pb>=15)
> prop.test(x=c(x_north,x_south), n=c(n_north,n_south),
alternative="two.sided")

    2-sample test for equality of proportions with continuity correction

data:  c(x_north, x_south) out of c(n_north, n_south)
X-squared = 11.845, df = 1, p-value = 0.0005781
alternative hypothesis: two.sided
95 percent confidence interval:
 0.04196839 0.16052203
sample estimates:
   prop 1    prop 2
0.1762452 0.0750000
```

Section 3

a)

```
Ho: mu =  40 Ha mu != 40, Two sided Test
```

b)

```
> xbar <- mean(flint$Cu)
> s <- sd(flint$Cu)
```

c)

```
> n <- nrow(flint)
> SE = s/sqrt(n)
```

d)

```
> t_stat <- (xbar-40)/SE
> p_value <- (1-pt(abs(t_stat),df=n-1))*2
> print(p_value)
[1] 0.01123183
```

e)

We fail to reject the null hypothesis if the significance level α = 0.01, since our p-value is higher than the significance level. We don't have sufficient evidence to suggest that the alternative is plausible.

```
#Do not reject
```

f)

```
> library(mosaic)
> t.test(flint$Cu, mu=40, alt="two.sided")

    One Sample t-test

data:  flint$Cu
t = 2.5441, df = 540, p-value = 0.01123
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 43.32285 65.83920
sample estimates:
mean of x
 54.58102
```

Section Bonus Credit

a)

We can check if there is an association between the two variables by looking at b, which is the slope and thus shows that there is a relationship if it isn't 0. We can prove this by finding correlation coefficient r and r-squared

b)

H0: β = 0 (no relationship)

Ha: β != 0 (linear relationship)

c)

```
p-value: < 2.2e-16


soil<-read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/soil_
complete.txt", header=TRUE)

>linear_model <- lm(soil$lead ~ soil$zinc)
>summary(linear_model)
Call:
lm(formula = soil$lead ~ soil$zinc)

Residuals:
      Min    1Q  Median   3Q    Max
-79.853 -12.945  -1.646  15.339 104.200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.367688   4.344268   3.998 9.92e-05 ***
soil$zinc    0.289523   0.007296  39.681  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.24 on 153 degrees of freedom
Multiple R-squared:  0.9114,    Adjusted R-squared:  0.9109
F-statistic:  1575 on 1 and 153 DF,  p-value: < 2.2e-16

d)
We get p-value < 2.2e-16, which means that the data is statistically
significant with a significance level α = 0.05. Thus we reject the null
hypothesis, suggesting that there is a linear relationship within the data.
```