Trent Bellinger

Stats 10 Lab 5

## Exercise 1

1a

Null Hypothesis: H0: p = 0.10
Alternative Hypothesis: Ha: p > 0.10
This is a one-sided test.

1b

```
flint <- read.csv('flint_201')

n <- nrow(flint)

dangerous_lead <- flint$Pb >= 15

phat <- mean(dangerous_lead)
phat
```

```
[1] 0.1238447
```

```
sample_sd <- sd(dangerous_lead)
sample_sd
```

```
[1] 0.3297092
```

The sample proportion of dangerous lead levels is about 12% with a sample standard deviation of about 0.33.

1c

```
p0 <- 0.10

se <- sqrt(p0*(1-p0)/n)
se
```

```
[1] 0.01289801
```

```
z_stat <- (phat - p0)/se
z_stat
```

```
[1] 1.848714
```

The standard error of sample proportions is about 0.013 with a z-value of about 1.85.

1d

```
p_value <- 1 - pnorm(z_stat)
p_value
```

```
[1] 0.03224953
```

The p-value associated with this test is about 0.0322.

1e

Yes we reject the null at a 0.05 significance level. This is because our calculated p value is 0.0322, which is less than the 0.05 significance.

1f

Conditions for CLT:
1. simple random sample
2. Large sample is satisfied because we have n*p0 > 10 and n*(1-p0) > 10
3. Large population is satisfied because the population of Flint is greater than 10*541 = 54100.

Since the validity conditions for CLT hold and we have rejected the null at 5% confidence we can be confident that the true population proportion of households with dangerous lead levels is greater than 10% in Flint. Yes, the EPA should be contacted.

```
# 1g

prop.test(x=sum(dangerous_lead), n=n, p=p0, alt='greater')

        1-sample proportions test with continuity correction

data:  sum(dangerous_lead) out of n, null probability p0
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 
```

The p value calculated by prop.test is slightly different than our hand calculated p value. This is due to the continuity correction done by prop.test. However as the calculated p value of 0.0378 is still under 0.05, which does not change our conclusion of contacting the EPA.

```
# 1h
prop.test(x=sum(dangerous_lead), n=n, p=p0, alt='greater', conf.level = 0.99)

        1-sample proportions test with continuity correction

data:  sum(dangerous_lead) out of n, null probability p0
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 
```

The 99% confidence interval of (0.094, 1.000) includes the null proportion of 0.10, as we are not confident at the 1% level.

## Exercise 2

2a

Null Hypothesis: H0: p_N = p_S
Alternative Hypothesis: Ha: p_N != p_S
(p_N is the proportion of dangerous lead levels in the North and p_S is the proportion of dangerous lead levels in the South)
This is a two-sided test because we have a not equals in the alternative.

2b

```
count_N <- sum(dangerous_lead & flint$Region == 'North')
count_N
```

```
[1] 46
```

The number of dangerous lead locations in the North is 46.

```
count_S <- sum(dangerous_lead) - count_N
count_S
```

```
[1] 21
```

The number of dangerous lead locations in the South is 21.

```
n_N = sum(flint$Region == 'North')
n_N
```

```
[1] 261
```

The total number of locations in the North is 261.

```
n_S = n - n_N
n_S
```

```
[1] 280
```

The total number of locations in the South is 280.

```
p_N <- count_N / n_N
p_N
```

```
[1] 0.1762452
```

The proportion of dangerous lead levels in the North is about 0.176.

```
p_S <- count_S / n_S
p_S
```

```
[1] 0.075
```

The proportion of dangerous lead levels in the South is 0.075.

```
se <- sqrt(phat*(1-phat)*(1/n_N + 1/n_S))
se
```

```
[1] 0.02834188
```

The standard error is about 0.028.

```
z_stat <- (p_N - p_S)/se
z_stat
```

```
[1] 3.572283
```

The z-statistic is about 3.57.

2c

```
p_value <- (1 - pnorm(z_stat))*2
p_value
```

```
[1] 0.0003538831
```

The p-value associated with this test is about 0.00035.

2d

Validity conditions:
1. Assume that the Flint dataset is made of random sample that are independent of each other and independent within sampling groups.
2. All sample proportions satisfy the large sample criterion.
3. Each population (North and South) is 10 times as large as its sample.

Since the validity conditions are satisfied and our p value is less than 5%, we reject the null at significance level 0.05. In the context of our research question, this says that the north of Flint and the South of Flint experience significantly different levels of dangerous lead.

2e

```
prop.test(x=c(count_N, count_S), n=c(n_N, n_S), alt='two.sided')

	2-sample test for equality of proportions with continuity correction

data:  c(count_N, count_S) out of c(n_N, n_S)
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 
```

The p values are slightly different due to the continuity correction. This correction does not change the end conclusion since the new p value of 0.00058 is still less than 0.05.