STATS 10 Assignment 5

Your Name

2025-03-12

```
knitr::opts_chunk$set(echo = TRUE)
```

#Part I Exercise 1 (a)

```
pawnee <- read.csv("pawnee.csv", header = TRUE)
head(pawnee)
```

	ID <int>	Latitude <dbl>	Longitude <dbl>	Arsenic <dbl>	Sulfur <dbl>	New_hlth_issue <chr>
1	1	41.09414	-85.60974	0	0	N
2	2	41.09054	-85.70344	0	130	N
3	3	41.08601	-85.71996	4	170	N
4	4	41.08100	-85.75415	0	0	Y
5	5	41.07435	-85.70043	0	0	N
6	6	41.07399	-85.71788	0	0	N

6 rows

```
dim(pawnee)
```

```
## [1] 541   6
```

b.

```
set.seed(1337)
sample_pawnee <- pawnee[sample(nrow(pawnee), 30), ]
head(sample_pawnee)
```

	ID <int>	Latitude <dbl>	Longitude <dbl>	Arsenic <dbl>	Sulfur <dbl>	New_hlth_issue <chr>
147	147	41.03971	-85.72783	2	100	N
49	49	41.06113	-85.65553	0	0	Y
210	210	41.03178	-85.64253	0	0	N
356	356	41.01178	-85.66516	0	0	N
425	425	41.00096	-85.72899	0	0	N
239	239	41.02772	-85.72901	0	0	N

6 rows

c.

```
prop_sample <- mean(sample_pawnee$New_hlth_issue == "Y")
prop_population <- mean(pawnee$New_hlth_issue == "Y")
prop_sample
```

```
## [1] 0.2
```

```
prop_population
```

```
## [1] 0.2920518
```

d.

```
n <- 30
p_hat <- prop_sample
se <- sqrt(p_hat * (1 - p_hat) / n)

z_90 <- qnorm(0.95)
ci90 <- c(p_hat - z_90 * se, p_hat + z_90 * se)

z_95 <- qnorm(0.975)
ci95 <- c(p_hat - z_95 * se, p_hat + z_95 * se)

z_99 <- qnorm(0.995)
ci99 <- c(p_hat - z_99 * se, p_hat + z_99 * se)

ci90
```

```
## [1] 0.07987688 0.32012312
```

```
ci95
```

```
## [1] 0.05686447 0.34313553
```

```
ci99
```

```
## [1] 0.01188802 0.38811198
```

Exercise 2. (a) Answer: Null hypothesis (H_0): The proportion of dangerous lead levels is less than or equal to 10% (i.e., p<=0.10). Alternative hypothesis (H_a): The proportion is greater than 10% (i.e., p>0.10) This is a one-sided test.

b.

```
flint <- read.csv("flint.csv", header = TRUE)
n <- dim(flint)[1]
p_null <- 0.1

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

```
## [1] 0.04436229
```

```
sample_sd
```

```
## [1] 0.008852277
```

c.

```
# Standard error using the null hypothesis proportion p0 = 0.10
se_flint <- sqrt(p_null * (1 - p_null) / n)

# Calculate the z-value
z_value <- (p_hat - p_null) / se_flint
z_value
```

```
## [1] -4.313667
```

d.

```
# For a one-sided test (p > 0.10)
p_value <- 1 - pnorm(z_value)
p_value
```

```
## [1] 0.999992
```

e. Decision at α = 0.05

```
p_value <= 0.5
```

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

Our p-value is larger than 0.05 so we fail to reject the null hypothesis.

f. Determining whether to be told to the EPA since remediation action is required to be taken by the EPA if greater than 10% of households in Flint contain dangerous lead levels.

Since we failed to reject the null hypothesis, we have no evidence to suggest that more than 10% of homes in Flint have a dangerous lead level. Therefore, we would tell the EPA that they are not required to take remediation action.

#Part II

Exercise 1 (a)

- Null Hypothesis (H_0): $p = p_0 = 0.48$, the proportion of site users who get their world news on the site has not changed since 2013.
- Alternative Hypothesis (H_a): $p \neq p_0$, the proportion of site users who get their world news on the site has changed since 2013.

The conditions for using the z-test are satisfied since: 1. The sample is randomly selected 2. $n \times p_0 = 3625 \times 0.48 = 1740 > 10$ & $n(1 - p_0) = 1885 > 10$ 3. Population N is large as $N \geq 10n = 10 \times 3625 = 36250$ We calculate the sample proportion ($\hat{p}$) as $1830/3625 \approx 0.5047$, the standard error (SE) as $\sqrt{(p_0 (1-p_0)/n)} \approx 0.0083$, and the z-test statistic as $(\hat{p} - p_0) / SE \approx 2.99$

We calculated a z-score of approximately 2.99. For the two-tailed test (H_a: $p \neq p_0$), we assess the p-value = $P(Z \leq -|z|) + P(Z \geq |z|) = 2(1 - 0.9986) = 0.0028$ The p-value was calculated to be approximately 0.0028.

We reject the null hypothesis (H_0) since the p-value ≈ 0.0028 is less than the significance level (α) of 0.05. There is statistical evidence to suggest that the proportion of site users who get their world news on the site has changed since 2013.

b.

- Calculating 95% confidence interval: $\hat{p} \pm z^* \times SE = 0.5047 \pm 1.96 \times 0.0083 = 0.5047 \pm 0.016268 =$ (0.489, 0.521)
- Interpretation: The confidence interval provides a range of plausible values for the population proportion based on the sample data. Since this confidence interval does not include the 2013 proportion of 0.48, it suggests that there has been a significant change in the proportion of site users getting their news from the site since 2013 with the true population proportion being higher in 2018. Further, the interval suggests that if we were to take many samples of the same size from the population, about 95% of those samples would result in a sample proportion between 0.489 and 0.521, further indicating an increase from the 2013 proportion. In summary, the confidence interval does not contain the 2013 value of 0.48 and lies entirely above it, which is consistent with the hypothesis test result indicating a significant increase from the 2013 proportion.

Exercise 2 –

Type I Error: Concluding that the proportion of voters in 2018 is higher than 50% when, in reality, it is 50% (or not higher). This is a false positive. Type II Error: Failing to detect that the proportion is higher when it actually is (i.e., not rejecting H_0 when the true proportion is above 50%). This is a false negative.

Exercise 3

a.

```
# Data for 2016
n1 <- 3103
x1 <- 2087
p1 <- x1 / n1

# Data for 2017
n2 <- 2988
x2 <- 1930
p2 <- x2 / n2

# Pooled proportion (under H0: p1 = p2)
p_pool <- (x1 + x2) / (n1 + n2)

# Standard error using the pooled proportion
se_diff <- sqrt(p_pool * (1 - p_pool) * (1/n1 + 1/n2))

# Calculate the z-statistic
z_college <- (p1 - p2) / se_diff
z_college
```

```
## [1] 2.194808
```

```
# Two-tailed p-value
p_value_college <- 2 * (1 - pnorm(abs(z_college)))
p_value_college
```

```
## [1] 0.02817737
```

Interpretation: Since the p-value is less than 0.05, we conclude there is a significant change between 2016 and 2017.

b.

```
# Standard error for the difference (using separate variances)
se_diff_np <- sqrt((p1 * (1 - p1) / n1) + (p2 * (1 - p2) / n2))

# 95% Confidence Interval for the difference (p1 - p2)
ci_diff_lower <- (p1 - p2) - 1.96 * se_diff_np
ci_diff_upper <- (p1 - p2) + 1.96 * se_diff_np
ci_diff <- c(ci_diff_lower, ci_diff_upper)
ci_diff
```

```
## [1] 0.002852873 0.050462979
```

Since the 95% confidence interval does not include 0, it supports the conclusion that there is a significant difference between the two years.