Programming Assignment 2

Andrew Shaw

2025-09-04

Problem 1

The dataset contains a large proportion of data where the respondents noted never smoking an e-cig and therefore did not skipped Q37. Given their responses of never having smoked an e-cig and the codebook's label that .S is considered a "Legitimate Skip", those respondents labeled ("02", ".S") were re-coded as not having smoked an e-cig in the past 30 days (0 for the dummy variable) for the purposes of analysis. The data were cleaned using a two-step process after examining all combinations of the variables using cross-tab tables to look for invalid combinations. In total 111 observations were dropped that were logically inconsistent or invalid leaving a total of 18907 out of 19018 observations.

```
df <- nyts2019[c("StudentLoginID", "Q34", "Q37")]

## check for values and combinations
ct1 <- dplyr::count(df, Q34, Q37)

## drop combinations of .N and .Z
# 30(N,.N), 3(.N, 00), 1(.N,02), 12 (.Z,.Z)
df_1 <- df |>
  mutate(
    days = if_else((Q34 == "02" & Q37 == ".S"), 0, as.numeric(Q37))) |>
  filter(Q34 == "01" | Q34 == "02")
# Check for any remaining invalid combinations before applying dummy
ct2 <- dplyr::count(df_1, Q34, days, Q37)

## drop remaining invalid combinations
# 1(02,06), 2(02,.N), 1(01,99), 1(01,55), 1(01,.S), 59(01,.N)
df_dummy <- df_1 |>
  filter(Q34 %in% c("01", "02") & days %in% 0:30) |>
  filter(!(Q34 == "02" & Q37 == "06")) |>
  mutate(
    dummy = if_else(days %in% 1:30, 1, 0)
  )
dplyr::count(df_dummy, Q34, dummy)
```

```
##   Q34 dummy     n
## 1  01     0  2720
## 2  01     1  3627
## 3  02     0 12560
```

Problem 2 Make up an expected value for the percentage of students that use e-cig?

18%

Problem 3 What is your best estimate of the sample mean of this created variable?

Best estimate before looking at the data is 20% based off of total observations and proportion responses labeled "1".

Problem 4 What are possible reasons this measure could be biased?

The survey could be biased as the survey relies on students' responses who may not have been truthful in responses. As discovered in the cleaning process, there could have been some students who misinterpreted the questions and marked incorrect responses. The survey may not cover all schools in all regions so it it's not a reflection of the entire population. The coding for ".S" could be inaccurate being labeled as "Legitimate Skip".

Problem 5 Take your expectation from problem 2 and write a hypothesis on whether the actual value is the same as your expectation.

$$H_0 : p = 0.18$$

$$H_a : p \neq 0.18$$

Problem 6 Test whether you should reject or not the hypothesis from part 5.

```
x <- sum(df_dummy$dummy)
n <- nrow(df_dummy)

binom.test(x, n, p = 0.18, alternative = "two.sided",
              conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  x and n
## number of successes = 3627, number of trials = 18907, p-value =
## 2.636e-05
## alternative hypothesis: true probability of success is not equal to 0.18
## 95 percent confidence interval:
##  0.1862426 0.1975198
## sample estimates:
## probability of success
##              0.1918337
```

```r
p <- x/n
se_0 <- sqrt((0.18*(1-0.18))/n)

z <- (p - 0.18)/se_0

p_val <- 2*(1-pnorm(abs(z)))
p_val
```

```
## [1] 2.281976e-05
```

```r
se_p <- sqrt((p*(1-p))/n)

lower <- p - 1.96*se_p
upper <- p + 1.96*se_p

cat("95% CI: [",lower,", ",upper,"]")
```

```
## 95% CI: [ 0.1862212 ,  0.1974462 ]
```

Because the p-value is below the 0.05 significance level, the null hypothesis that the true proportion is 18% is rejected.

Problem 7: Compute a 95% confidence interval for the estimate from problem 3.

```r
se_g <- sqrt((0.20*(1-0.20))/n)

lower_g <- 0.20 - 1.96*se_g
upper_g <- 0.20 + 1.96*se_g


cat("95% CI: [",lower_g,", ",upper_g,"]")
```

```
## 95% CI: [ 0.1942983 ,  0.2057017 ]
```

Adjusting the standard error for the guesstimate of 0.20, the confidence interval at 95% is [0.1942, 0.2057]. The interval based on the estimate overlaps slightly with the confidence interval from the observed data.