Stats C155 - Sampling Project

Albert Putranegoro

Summer 2024

1. Why did you choose to do this project and what was the research question that you were trying to answer?

Research Question: During the ninth week of classes, on Thursday between 11 AM - 12 PM, what proportion of individuals walking up Bruinwalk are wearing glasses?

I chose this project because I wanted to conduct an observational study on Bruinwalk, which is a high-traffic area on campus, making it an ideal location for systematic sampling over a short period of time. Glasses are a visible and easily identifiable characteristic, which makes data collection straightforward and minimizes subjective judgment in classification.

2. What was your target population, your sampling frame, and your sampling units? How did you go about enumerating them?

- **Target Population:** All individuals who walk up Bruinwalk during Thursday of ninth week of classes between 11AM - 12PM.
- **Sampling Frame:** Individuals who passed a designated observation point at the top of the Bruinwalk stairs between 11AM and 12PM.
- **Sampling Units:** Each individual walking up Bruinwalk during the observation period. (Every 3rd person walking up Bruinwalk)

To enumerate the individuals in my sampling frame, I used a systematic approach where I visually identified every 3rd person passing my observation point and recorded their glasses status (excluding sun glasses).

3. What was your sampling method? Did any issues come up that you had to deal with? Describe the strengths and weaknesses of your sampling plan. How could it have been improved?

Sampling Method:

- I used systematic sampling, selecting every 3rd individual walking up Bruinwalk during the observation period (11AM - 12PM). The person has to pass my observation point to be considered walking up to the end of Bruinwalk.
- For each selected person, I recorded whether they were wearing glasses (excluding sunglasses)

Issues:

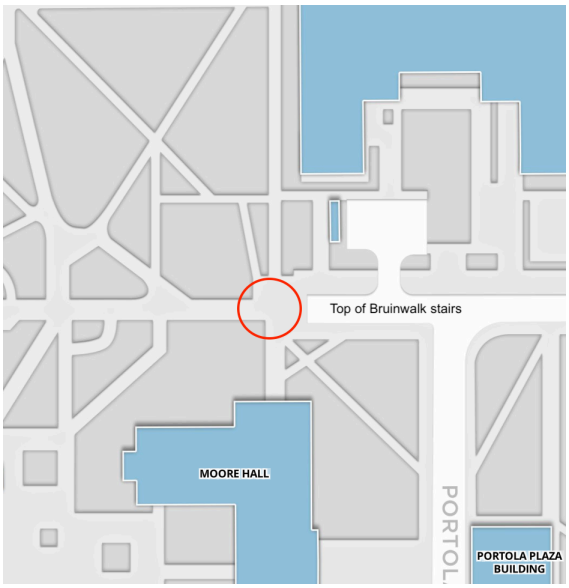
- In some cases, groups of individuals walked closely together or next to each other, making it difficult to consistently identify the exact 3rd person in the sequence.

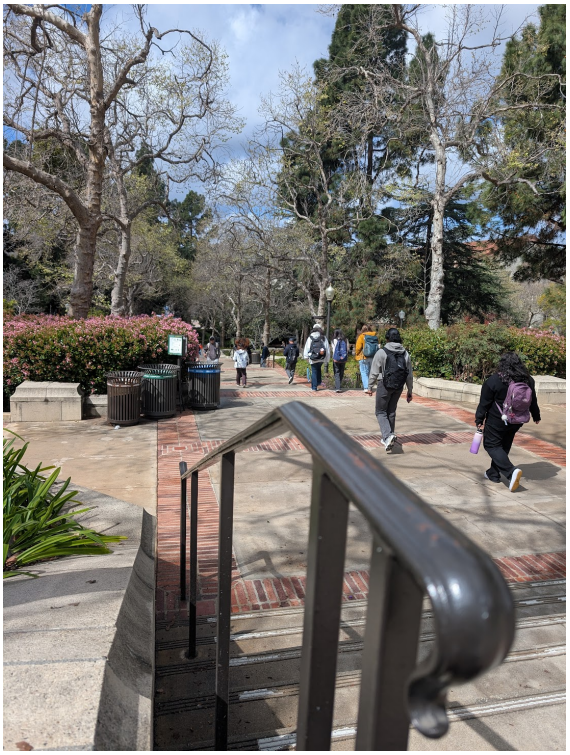
Strengths:

- Using systematic sampling made it efficient and practical to collect a good number of observations within an hour. By selecting every 3rd person, the observations were spread across the full hour, avoiding over-sampling specific clusters.
- The method was straightforward, no need for equipments or have to deal with not knowing if the individuals do not fit the criteria.

Weaknesses & Improvements:

- The findings are limited to a single one-hour period. Collecting data across multiple time slots or different days would provide a broader perspective.
- Instead of selecting every 3rd person from a fixed start point, using a randomized start within each minute could further reduce selection bias.





4. Please describe the method of analysis that you decided to use to analyze your observations. What formulas did you use and why?

1. Proportion Estimate

The proportion of individuals wearing glasses is estimated as:

$$\hat{p}_{sy} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- $y_i = 1$ if the individual is wearing glasses, 0 o.w
- n = Total sample size

This provides a point estimate of the proportion of individuals wearing glasses.

2. Variance of the Proportion Estimate

Since N is unknown, we use the following variance formula for systematic sampling:

$$\hat{V}(\hat{p}_{sy}) = \frac{\hat{p}_{sy}(1 - \hat{p}_{sy})}{n - 1}$$

- n = Total population size

This accounts for variance in systematic sampling, ensuring an accurate measure of estimation error. Ensures accuracy when N is unknown.

3. Confidence Interval for the Population Proportion

95% confidence interval:

$$\hat{p}_{sy} \pm Z^* \cdot \sqrt{\hat{V}(\hat{p}_{sy})}$$

- $Z^* = 1.96$ for a 95% confidence level
- $\sqrt{\hat{V}(\hat{p}_{sy})}$ = Standard error of the sample proportion

This interval quantifies the range in which the true population proportion is likely to fall with 95% confidence.

5. What is your estimate of the population parameter of interest. Please show your calculations. If you have used Stata or R to develop your estimates, please include a copy of the output.

$$\hat{p}_{sy} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- $\sum_{i=1}^{n} y_i$ = 53 (number of individuals wearing glasses)
- n = 181 (total sample size)

$$\hat{p}_{sy} = \frac{53}{181} = 0.2928177$$

```
53/181
```

```
## [1] 0.2928177
```

The estimated proportion of individuals wearing glasses is 0.2928 (29.28%).

6. Generate an estimate of the sampling variance and a confidence interval for your estimate (You may assume that you have drawn a large sample from an even larger population). If you have used Stata or R to develop your estimates, please include a copy of the output.

$$\hat{V}(\hat{p}_{sy}) = \frac{\hat{p}_{sy}(1 - \hat{p}_{sy})}{n - 1}$$

```
n <- 181
x <- 53

# Sample proportion
p_hat_sy <- x / n

# Variance of the sample proportion
variance_p_hat_sy <- (p_hat_sy * (1 - p_hat_sy)) / (n - 1)

# Standard error
se_p_hat_sy <- sqrt(variance_p_hat_sy)

# 95% confidence interval
z_critical <- 1.96
ci_lower <- p_hat_sy - z_critical * se_p_hat_sy
ci_upper <- p_hat_sy + z_critical * se_p_hat_sy

results_df <- data.frame(
  Estimate = c("Variance", "Lower CI (95%)", "Upper CI (95%)"),
  Value = round(c(variance_p_hat_sy, ci_lower, ci_upper), 4)
)

kable(results_df, col.names = c("Estimate", "Value"), caption = "Estimate of the sampling variance and a
```

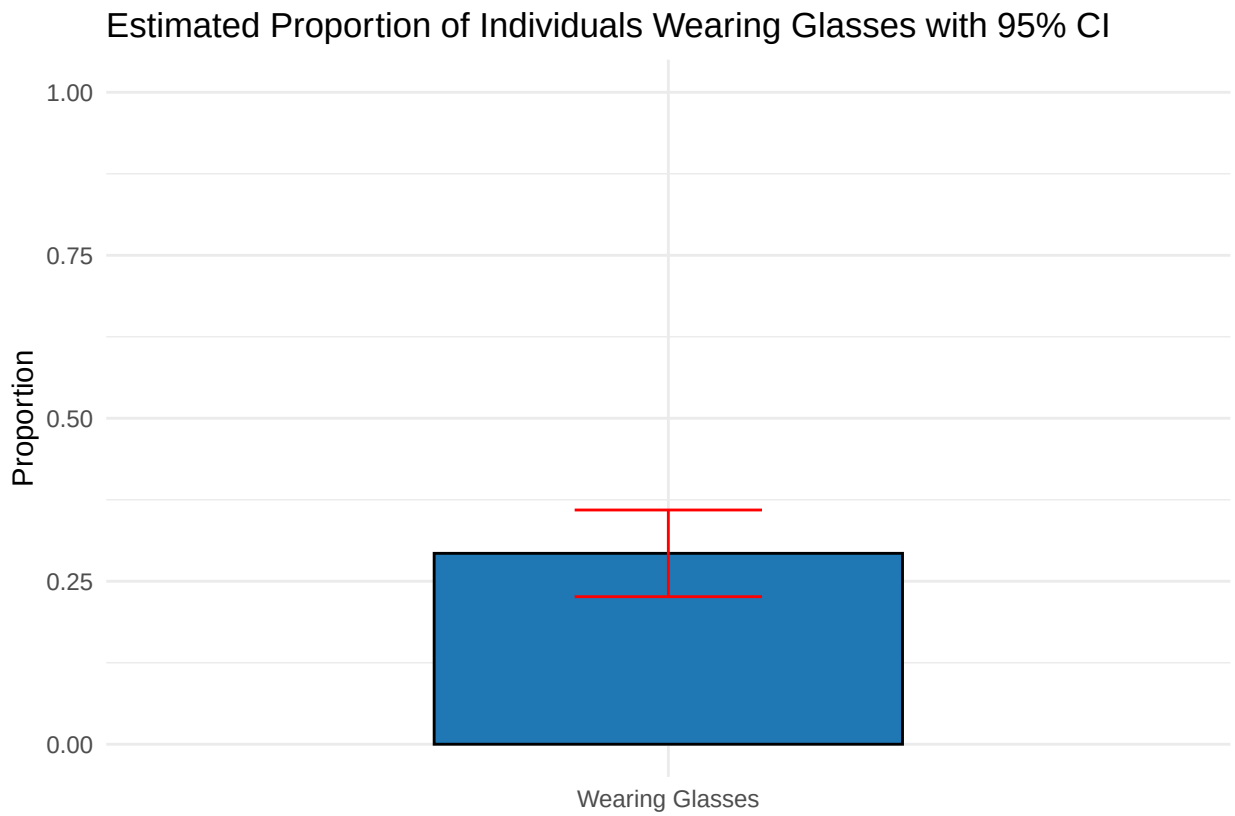
Table 1: Estimate of the sampling variance and a confidence interval

Estimate	Value
Variance	0.0012
Lower CI (95%)	0.2263
Upper CI (95%)	0.3593

7. Please clearly and concisely present your results using both a graph and a table. The graph is easily made in Stata or R if you wish. For templates of tables, take a look at a journal publishing research articles or a textbook for ideas on how to convey your findings.

Table 2: Results

Estimate	Value
Proportion Wearing Glasses	0.2928
Variance	0.0012
Standard Error	0.0339
Lower CI (95%)	0.2263
Upper CI (95%)	0.3593



8. Discuss your findings briefly. Please use your sample estimates to make an inference about the population parameter. Given your experiences in collecting the data, do you have any concerns about non-sampling error and other sources of bias? Discuss the strengths and limitations of your estimate.

Based on a systematic sample of 181 individuals walking up Bruinwalk, the estimated proportion of individuals wearing glasses, excluding sunglasses, is 0.2928 (29.28%), with a 95 percent confidence interval of (0.2263, 0.3593). This means that the true proportion of individuals wearing glasses during this time period is likely between 22.63 percent and 35.93 percent. The confidence interval is relatively wide, indicating some level of uncertainty in the estimate, but it still provides a reasonable approximation of the actual population proportion.

There are several potential sources of non-sampling error that may affect the accuracy of this estimate. One key limitation is time-specific bias, as data collection was done only between 11AM and 12PM on a single day. The proportion of individuals wearing glasses may vary at different times due to differences in demographics, such as workers, visitors, and other members of the public passing through Bruinwalk at different times of the day. Collecting data across multiple time slots or different days would provide broader findings.

Despite these limitations, the estimate has several strengths. Systematic sampling ensured an even spread of observations throughout the sampling period, avoiding clustering and ensuring a fair representation of all individuals passing through Bruinwalk. The sample size of 181 individuals is large enough to provide a reasonable estimate with an interpretable confidence interval.

Overall, while the estimate provides a reasonable approximation of the proportion of individuals wearing glasses on Bruinwalk, expanding the data collection across different time periods and refining aspects of the sampling process would improve accuracy and generalizability. The findings are useful within the specific study conditions but could be strengthened by addressing the potential sources of bias identified.

9. Please attach a copy of your sampling frame enumeration and your raw data to this report.

Table 3: Raw Count

Glasses	Count
Yes	53
No	128