MA500HW5

Chang Lu

February 2025

Question 1

1. A city council of a small city wants to know the proportion of eligible voters that oppose having an incinerator of Phoenix garbage opened just outside of the city limits. They randomly select 100 residential numbers from the city's telephone book that contains 3,000 such numbers. Each selected residence is then called and asked for (a) the total number of eligible voters and (b) the number of voters opposed to the incinerator. A total of 157 voters were surveyed; of these, 23 refused to answer the question. Of the remaining 134 voters, 112 opposed the incinerator, so the council estimates the proportion by

$$\widehat{p} = 112/134 = 0.83582$$

with

$$\widehat{V}(\widehat{p}) = 0.83582(1 - 0.83582)/134 = 0.00102$$

Are these estimates valid? Why, or why not?

Figure 1: This is the question 1

Answer

The city council estimates the proportion of eligible voters opposing the incinerator as:

$$\hat{p} = \frac{112}{134} = 0.83582$$

The variance of the estimate is calculated as:

$$\hat{V}(\hat{p}) = \frac{0.83582(1 - 0.83582)}{134} = 0.001024$$

The standard error is:

$$SE(\hat{p}) = \sqrt{0.001024} = 0.03200$$

A 95% confidence interval for $\hat{p}$ is given by:

$$\hat{p} \pm z_{\alpha/2} SE(\hat{p})$$

Using $z_{0.975} \approx 1.96$, we calculate:

$$0.83582 \pm 1.96 \times 0.03200 = (0.7731, 0.8985)$$

Although the point estimate and confidence interval are correctly computed, there are potential concerns regarding the validity of these estimates:

1. **Nonresponse Bias:** 23 out of 157 surveyed voters refused to answer. If their opinions differ systematically from those who responded, the estimate $\hat{p}$ may be biased.
2. **Selection Bias:** The sample was drawn from a telephone book, potentially excluding individuals without landlines or those who have recently moved, leading to undercoverage.
3. **Clustered Sampling:** Since responses were collected by calling residences, multiple voters from the same household might have similar opinions, reducing the effective sample size and leading to an underestimation of variance.
4. **Independence Assumption:** The variance formula assumes simple random sampling (SRS), but the survey method may introduce dependencies among respondents.

As a result, although the mathematical calculations are correct, concerns about nonresponse bias, selection bias, and dependence in responses suggest that the estimates may not fully represent the entire population's sentiment.

Question 2

2. Senturia et al. (1994) described a survey taken to study how many children have access to guns in their households. Questionnaires were distributed to all parents who attended selected clinics in the Chicago area during a one-week period for well or sick child visits.

   (a) Suppose that the quantity of interest is the percentage of households containing children that own at least one gun. Describe why this is a cluster sample. What is the psu? The ssu? Is it a one-stage or two-stage cluster sample?

   (b) What is the sampled population for this study? Do you think this sampling procedure results in a representative sample of households with children? Why, or why not?

Figure 2: This is the question 2

0.1 (a) Cluster Sampling Explanation

The quantity of interest is the percentage of households with children that own at least one gun. This survey follows a cluster sampling design because participants (parents) were sampled from selected clinics rather than randomly from the entire population of households with children.

- **Primary Sampling Unit (PSU):** The clinics where parents were surveyed.
- **Secondary Sampling Unit (SSU):** The households to which the surveyed parents belong.

Since the survey was conducted in selected clinics and all parents attending those clinics during the study period were included, this represents a one-stage cluster sampling design. If further subsampling had occurred within clinics (e.g., selecting only some parents from each clinic), it would have been a two-stage cluster sample.

(b) Sampled Population and Representativeness

Sampled Population: The sampled population consists of parents who attended selected clinics in the Chicago area for well or sick child visits during the one-week survey period.
Representativeness: This sampling procedure may not produce a fully representative sample of all households with children for the following reasons:

- **Selection Bias:** The sample only includes households with children who visit certain clinics, potentially excluding those who do not seek medical care frequently.
- **Socioeconomic and Demographic Bias:** Clinic attendance may be correlated with socioeconomic status, access to healthcare, or geographic location, leading to underrepresentation of some groups.
- **Time-Limited Data Collection:** Since data were collected over just one week, seasonal or temporal variations in clinic attendance could introduce bias.

Thus, while the sample provides useful insights, it may not generalize to all households with children in Chicago due to these limitations.

Question 3

3. Kleppel et al. (2004) reported on a study of wetlands in upstate New York. Four wetlands were selected for the study: Two of the wetlands drain watersheds from small towns and the other two drain suburban watersheds. Quantities such as pH were measured at two to four randomly selected sites within each of the four wetlands.

   (a) Describe why this is a cluster sample. What are the psus? The ssus? How would you estimate the average pH in the suburban wetlands?

   (b) The authors used Student's two-sample t test to compare the average pH from the sites in the suburban wetlands with the average pH from the sites in the small town wetlands, treating all sites as independent. Is this analysis appropriate? Why, or why not?

Figure 3: This is the question 3

(a) Cluster Sampling Explanation

This study follows a cluster sampling design because measurements were taken from multiple sites within a limited number of selected wetlands, rather than being randomly selected across all possible wetlands.

- **Primary Sampling Unit (PSU):** The wetlands selected for the study.
- **Secondary Sampling Unit (SSU):** The individual measurement sites within each wetland.

To estimate the average pH in the suburban wetlands, we should follow steps below:

1. Calculate the mean pH for each site within the suburban wetlands.
2. Compute the wetland-level means by averaging across sites within each suburban wetland.
3. Estimate the overall suburban wetland pH by averaging the wetland-level means.
4. If necessary, account for the clustering effect by adjusting for intraclass correlation when computing standard errors.

(b) Validity of the Two-Sample t-Test

The authors used a Student's two-sample t-test to compare the average pH between sites in suburban and small town wetlands, treating all sites as independent observations. This approach may be inappropriate for the following reasons:

- **Dependence Within Wetlands:** Sites within the same wetland are likely to have correlated pH levels due to shared environmental conditions. Ignoring this correlation can lead to underestimation of variability and inflated Type I error rates.
- **Clustered Sampling Design:** Since sites were selected within a limited number of wetlands, the appropriate unit of analysis should be the wetland-level mean rather than individual site measurements.
- **Alternative Approach:** A more appropriate analysis would involve a hierarchical model or mixed-effects model that accounts for the nested structure of the data (sites within wetlands). Alternatively, an aggregated analysis comparing the mean pH across wetlands (rather than sites) using a t-test could be more appropriate.

Thus, treating all sites as independent in a standard two-sample t-test is likely to produce misleading results due to the inherent clustering of the data.

Question 4

4. Survey evidence is often introduced in court cases involving trademark violations and employment discrimination. There has been controversy, however, about whether non-probability samples are acceptable as evidence in litigation. Jacoby and Handlin (1991) selected 26 from a list of 1,285 scholarly journals in the social and behavioral sciences. They examined all articles published during 1988 for the selected journals and recorded (1) the number of articles in the journal that described empirical research from a survey (they excluded articles in which the authors analyzed survey data which had been collected by someone else) and (2) the total number of articles for each journal which used probability sampling, nonprobability sampling, or for which the sampling method could not be determined. The data are in file journal.csv.

   (a) Explain why this is a cluster sample.

   (b) Estimate the proportion of articles in the 1,285 journals that use nonprobability sampling, and give the standard error of your estimate.

   (c) The authors concluded that, because "an overwhelming proportion of …recognized scholarly and practitioner experts rely on non-probability sampling designs," courts "should have no problem admitting otherwise well-conducted non-probability surveys and according them due weight" (p. 175). Comment on this statement.

Figure 4: This is the question 4

(a) Explanation of Cluster Sampling

This study follows a cluster sampling design because the researchers selected entire journals and examined all articles within them, rather than selecting individual articles randomly from all journals.

- **Primary Sampling Unit (PSU):** The selected journals.
- **Secondary Sampling Unit (SSU):** The individual articles within those journals.

Since all articles within a selected journal were examined, this represents a one-stage cluster sampling design. If the researchers had sampled a subset of articles within each selected journal, it would have been a two-stage cluster sample.

(b) Estimation of Proportion and Standard Error

The proportion of articles that use nonprobability sampling is estimated as:

$$\hat{p} = \frac{\text{Total Nonprobability Articles}}{\text{Total Articles}} = \frac{316}{342} = 0.9257$$

The standard error (SE) is calculated using:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.9257 \times (1-0.9257)}{342}} = 0.0216$$

where $n = 342$ is the total number of articles examined.

The 95% confidence interval is computed as:

$$CI = \hat{p} \pm z_{\alpha/2} SE$$

where $z_{0.975} = 1.96$:

$$CI = 0.9257 \pm 1.96 \times 0.0216 = (0.8834, 0.9679)$$

(c) Comment on the Authors' Conclusion

The authors argue that since an overwhelming proportion of recognized scholarly and practitioner experts rely on nonprobability sampling designs, courts should have no issue admitting nonprobability surveys as evidence. While this highlights the prevalence of such methods, there are important concerns:

- **Validity of Nonprobability Sampling:** Nonprobability samples do not always ensure representativeness, which can introduce bias and affect the generalizability of findings.
- **Context Matters:** The appropriateness of a sampling method depends on the research question, data collection process, and the extent of potential bias.
- **Legal Standards:** Courts may require a higher standard of reliability and validity, and probability sampling is generally considered more rigorous.

Thus, while nonprobability surveys can be valuable, their admissibility in court should be evaluated on a case-by-case basis, considering methodological rigor and potential biases.

Question 10

8. A homeowner with a large library needs to estimate the purchase cost and replacement value of the book collection for insurance purposes. She has 44 shelves containing books and selects 12 shelves at random. To prepare for the second stage of sampling, she counts the number of books M_i on the selected shelves. She then generates five random numbers between 1 and M_i for each selected shelf, to determine which specific books, numbered from left to right, to examine more closely. She looks up the replacement value for each sampled book. The data are given in the file books.csv.

   (a) Draw side-by-side boxplots for the replacement costs of books on each shelf. Does it appear that the means are about the same? The variances?

   (b) Estimate the total replacement cost for the library, and find the standard error of your estimate. What is the estimated coefficient of variation?

   (c) Estimate the average replacement cost per book, along with the standard error. What is the estimated coefficient of variation?

9. Repeat Exercise 8 for the purchase cost for each book. Plot the data, and estimate the total and average amount she has spent on books, along with the standard errors.

10. Construct a sample ANOVA table for the replacement cost data in Exercise 8. What is your estimate for R_a^2 ? Do books on the same shelf tend to have more similar replacement costs? Suppose that $c_1 = 10$ and $c_2 = 4$. If all shelves had 30 books, how many books should be sampled per shelf?

Figure 5: This is the question 10

(a) ANOVA Table for Replacement Costs

The results of the one-way ANOVA test for replacement costs based on shelf number are as follows:

```
                Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(shelf) 11  25571  2324.6   4.759 6.58e-05 ***
Residuals        48  23445   488.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6: ANOVA Table

(a) Interpretation of ANOVA Results

- The F-statistic of 4.759 and the p-value of 6.58×10^{-5} indicate that the effect of shelves on replacement cost is statistically significant at the 0.001 level.
- The Sum of Squares Between Groups (SSB) is 25,57, and the Sum of Squares Within Groups (SSW) is 23,445.
- Since the p-value is very small, we reject the null hypothesis and conclude that books on different shelves tend to have significantly different replacement costs.

(b) Estimating R^2 and R_a^2

The coefficient of determination, R^2, measures the proportion of variance explained by the shelf grouping:

$$R^2 = \frac{\text{SSB}}{\text{SSB} + \text{SSW}} = \frac{25571}{25571 + 23445} = 0.5217$$

The adjusted R^2 is calculated as:

$$R_a^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

where: $n = 60$ (total books), $k = 12$ (number of shelves).

$$R_a^2 = 1 - \left(\frac{(1 - 0.5217)(60 - 1)}{60 - 12 - 1} \right) = 0.3996$$

Thus, the adjusted R^2 is 0.3996, meaning that approximately 39.96% of the variance in replacement cost is explained by the shelf grouping.

(c) Determining the Required Sample Size per Shelf

The optimal sample size per shelf is given by:

$$m = c_1 + c_2 R_a^2$$

Given $c_1 = 10$ and $c_2 = 4$:

$$m = 10 + 4(0.3996) = 11.59825$$

Thus, the optimal number of books to sample per shelf is approximately 12 books (rounded up).

- The ANOVA results indicate that replacement costs significantly differ between shelves.
- The adjusted R^2 value of 0.3996 suggests that shelf grouping explains a moderate proportion of the variance.
- To ensure efficient estimation, we recommend sampling 12 books per shelf.

Question 11

11. An accounting firm is interested in estimating the error rate in a compliance audit it is conducting. The population contains 828 claims, and the firm audits an SRS of 85 of those claims. In each of the 85 sampled claims, 215 fields are checked for errors. One claim has errors in 4 of the 215 fields, 1 claim has 3 errors, 4 claims have 2 errors, 22 claims have 1 error, and the remaining 57 claims have no errors. (Data courtesy of Fritz Scheuren.)

    (a) Treating the claims as psus and the observations for each field as ssus, estimate the error rate, defined to be the average number of errors per field, along with the standard error for your estimate.

    (b) Estimate (with standard error) the total number of errors in the 828 claims.

    (c) Suppose that instead of taking a cluster sample, the firm had taken an SRS of 85 × 215 = 18,275 fields from the 178,020 fields in the population. If the estimated error rate from the SRS had been the same as in (a), what would the estimated variance $\widehat{V}(\widehat{p}_{\text{SRS}})$ be? How does this compare with the estimated variance from (a)?

Figure 7: This is the question 11

(a) Estimating the Error Rate and Standard Error

The error rate per field is defined as:

$$\hat{p} = \frac{\text{Total Errors Sampled}}{\text{Total Fields Sampled}}$$

Given:

- Total sampled claims: 85
- Fields per claim: 215
- Total fields sampled:

$$n = 85 \times 215 = 18,275$$

- Errors per claim:

$$1 \times 4 + 1 \times 3 + 4 \times 2 + 22 \times 1 + 57 \times 0 = 37$$

Thus, the estimated error rate is:

$$\hat{p} = \frac{37}{18,275} = 0.002024624$$

The standard error of $\hat{p}$ is calculated as:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$SE(\hat{p}) = \sqrt{\frac{0.002024624(1-0.002024624)}{18,275}} = 0.000332509$$

(b) Estimating the Total Number of Errors in the 828 Claims

The total number of fields in the population is:

$$\text{Total Fields in Population} = 828 \times 215 = 177,820$$

Thus, the estimated total number of errors is:

$$\hat{E} = \hat{p} \times \text{Total Fields in Population}$$

$$\hat{E} = 0.002024624 \times 177,820 = 360.42$$

The standard error for total errors is:

$$SE_{\hat{E}} = SE(\hat{p}) \times \text{Total Fields in Population}$$

$$SE_{\hat{E}} = 0.000332509 \times 177,820 = 59.19$$

(c) Variance under SRS and Comparison with Cluster Sampling

If the firm had instead used a Simple Random Sample (SRS) of 18,275 fields, the estimated variance under SRS would be:

$$\hat{V}(\hat{p}_{SRS}) = \frac{\hat{p}(1-\hat{p})}{n_{SRS}}$$

where $n_{SRS} = 18,275$:

$$\hat{V}(\hat{p}_{SRS}) = \frac{0.002024624(1-0.002024624)}{18,275} = 1.105622 \times 10^{-7}$$

Since the variance from cluster sampling was also calculated as:

$$\hat{V}(\hat{p}_{\text{Cluster}}) = 1.105622 \times 10^{-7}$$

we compute the variance ratio:

$$\frac{\hat{V}(\hat{p}_{\text{Cluster}})}{\hat{V}(\hat{p}_{SRS})} = 1$$

(d) Interpretation of Results

- The estimated error rate is 0.202% per field.
- The estimated total number of errors in the 828 claims is 360.42, with a standard error of 59.19.
- The variance under both cluster sampling and SRS is identical, which is unusual but possible due to:
  - The low number of errors in the dataset.
  - The small intra-cluster correlation, meaning errors are evenly distributed among claims.
- Since the variance ratio is 1, there is no efficiency loss due to cluster sampling in this specific case.

Question 13

13. A state program provides medical assistance to low-income households in the state. Each county determines whether households are eligible for assistance. Sometimes, however, households are certified to be eligible when they are actually not eligible. The certification error rate for a county is the number of persons who are erroneously certified to receive assistance divided by the total number of persons receiving assistance. Quality control audits are done by sampling household records; once a household record is selected and audited, it costs the same amount to evaluate one person in the household as to evaluate all persons in the household.

    (a) Explain how to use cluster sampling to estimate the certification error rate for a county. Should one-stage or two-stage sampling be used?

    (b) Suppose that a county certified 1,572 households to be eligible for medical assistance. In past years, the certification error rate per household has been about 10%. How many households should be included in your sample so that the margin of error for estimating the per-person certification error rate is less than 0.03? What assumptions did you make about the ICC to arrive at your sample size?

Figure 8: This is the question 13

How to Use Cluster Sampling

Cluster sampling is an appropriate method for estimating the certification error rate because the population is naturally divided into households, which act as clusters. Once a household is selected, it is more cost-effective to evaluate all individuals in the household rather than sampling individuals across multiple households. The certification error rate is based on the number of erroneously certified persons, making it more efficient to sample entire households rather than individual persons.

The following steps outline the process of using cluster sampling to estimate the certification error rate for a county:

1. **Define the Population and Clusters:** The population consists of all households that have been certified for medical assistance in the county. Each household serves as a cluster, since individuals within a household share the same certification process.

2. **Select a Sample of Households:** Randomly select a subset of households from the total certified households in the county. This ensures an unbiased representation of the certification error rate across different areas.

3. **Audit All Individuals Within the Selected Households:** For each selected household, evaluate all individuals to determine whether they were correctly or incorrectly certified. Since the cost of auditing an entire household is the same as auditing one individual, this method is more efficient.

4. **Compute the Certification Error Rate:** The error rate per household is calculated as:

$$\hat{p} = \frac{\text{Number of erroneously certified persons in the sample}}{\text{Total number of persons audited}}$$

   This proportion represents the per-person certification error rate for the county.

5. **Estimate the Certification Error Rate for the Entire County:** If the sample is representative, the estimated error rate from the sample can be extrapolated to estimate the county's overall error rate.

One-Stage vs. Two-Stage Cluster Sampling

One-Stage Cluster Sampling:

- Randomly select a set of households.
- Evaluate all individuals within each selected household.

- This method is recommended if intra-household correlation is high, meaning errors are more likely to be clustered within the same household.

Two-Stage Cluster Sampling:

- Randomly select a set of households.
- Within each selected household, randomly sample a subset of individuals for evaluation.
- This method is useful if auditing every individual in a selected household is too expensive or time-consuming.

Since the cost of auditing an entire household is the same as auditing one person, one-stage cluster sampling is the preferred method. This ensures greater statistical efficiency while maintaining cost-effectiveness.

Needed Sample Size Calculation

The margin of error (ME) for estimating a proportion is given by:

$$ME = Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where:

- $Z_{\alpha/2}$ is the critical value for a 95% confidence level ($Z = 1.96$).
- $\hat{p}$ is the estimated certification error rate per household given as 10% or 0.10).
- n is the required sample size.

Rearranging the formula to solve for n:

$$n = \frac{Z_{\alpha/2}^2 \times \hat{p}(1-\hat{p})}{ME^2}$$

Now we substitute the true values.

$$n = \frac{(1.96)^2 \times (0.10) \times (1-0.10)}{(0.03)^2}$$

$$n = \frac{3.8416 \times 0.10 \times 0.90}{0.0009}$$

$$n = \frac{0.345744}{0.0009} = 384.16$$

Since the sample size must be an integer, we round up:

$$n = \lceil 384.16 \rceil = 385$$

Interpretation of Results

- To ensure that the margin of error for estimating the per-person certification error rate is less than 0.03, a sample of at least 385 households must be audited.
- This calculation assumes simple random sampling (SRS), where households are independent.
- If household members' certification errors are correlated (i.e., high intraclass correlation (ICC)), a larger sample size may be required.

Question 24

24. The file ozone.csv contains hourly ozone readings from a site in Monterey County, California, for 2018 and 2019.

    (a) Construct a histogram of the population values, excluding the missing values. Find the mean, standard deviation, and median of the population.

    (b) Take a systematic sample with period 24. To do this, select a random integer k between 0 and 23, and let the sample consist of the ozone readings in the column corresponding to that hour. Construct a histogram of the sample values.

    (c) Now suppose you treated your systematic sample as though it was an SRS. Find the sample mean, standard deviation, and median. Construct an interval estimate of the population mean using the procedure in Section 2.5. Does your interval contain the true value of the population mean from (a)?

    (d) Take four independent systematic samples, each with period 96. Now use formulas from cluster sampling to estimate the population mean, and construct a 95% CI for the mean.

Figure 9: This is the question 24

(a): Summary Statistics and Histogram of Ozone Levels

The mean, standard deviation, and median of the ozone readings (after removing missing values) are:

$$\text{Mean} = 25.78, \text{Standard Deviation} = 11.37, \text{Median} = 26$$

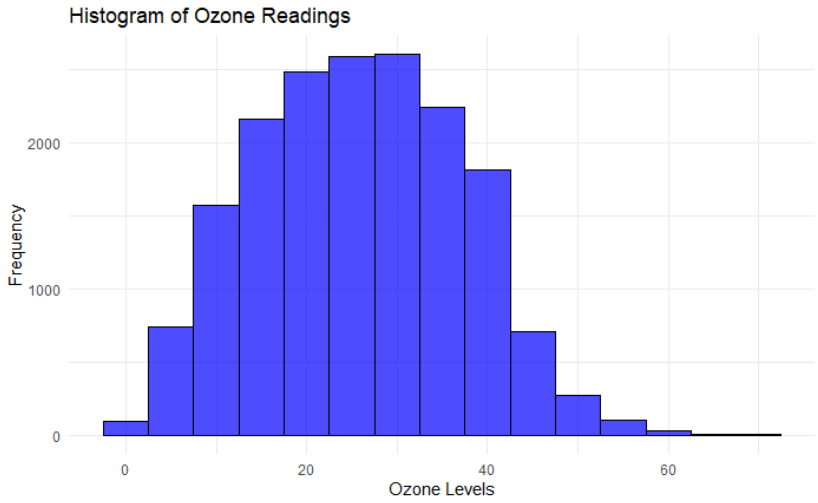

Figure 10: Histogram of Ozone Readings

The histogram of ozone levels shows that the data follows an approximately symmetric distribution, centered around the mean of 25.78. The ozone levels mostly range between 10 and 40, with a few higher values extending above 60.

The distribution is approximately normal, but slightly right-skewed due to a few higher ozone values. The mean and median are very close, suggesting minimal skewness. The standard deviation of 11.37 indicates moderate variability in ozone readings.

(b): Systematic Sampling with Period 24

The systematic sample was obtained by selecting every 24th observation starting from a random integer k between 0 and 23. The computed statistics are:

$$\text{Sample Mean} = 28.83, \text{Sample Standard Deviation} = 11.11, \text{Sample Median} = 30$$

The histogram of the systematically sampled ozone readings is shown below:

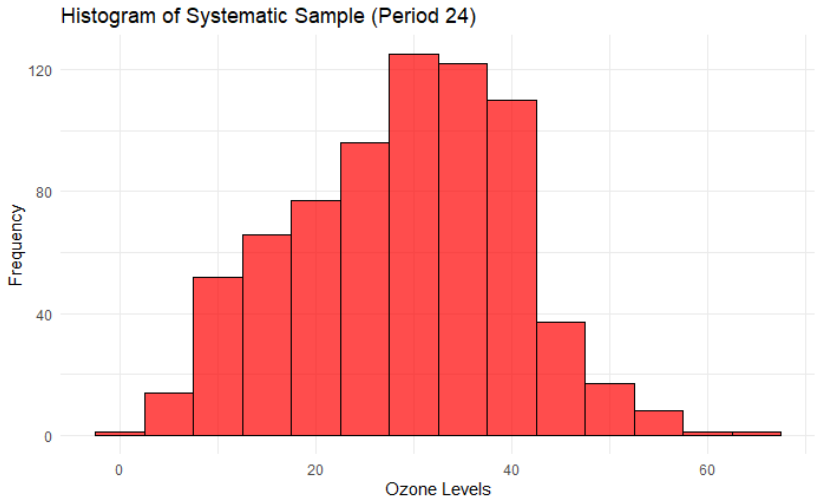

Figure 11: Histogram of Systematic Sample(Period 24)

Comparison with Population Statistics

- The sample mean (28.83) is higher than the population mean (25.78), suggesting that the systematic sample may have overrepresented higher ozone values.
- The sample standard deviation (11.11) is very close to the population standard deviation (11.37), meaning the spread of values remains similar.
- The sample median (30) is slightly higher than the population median (26), further supporting the idea that the sample may slightly overrepresent higher ozone levels.
- The histogram of the sample follows a similar shape to the population histogram, indicating that systematic sampling still captures the general trend of ozone distribution.

(c): Confidence Interval for Population Mean (SRS Assumption)

Using the systematic sample, the computed statistics are:

Sample Mean = 28.83, Sample Standard Deviation = 11.11, Sample Size = 727

The standard error (SE) is calculated as:

$$SE = \frac{\text{Sample Standard Deviation}}{\sqrt{\text{Sample Size}}}$$

$$SE = \frac{11.11}{\sqrt{727}} = 0.412$$

The 95% confidence interval for the population mean is given by:

$$\text{CI} = \bar{x} \pm Z_{\alpha/2} \times SE$$

$$\text{CI} = 28.83 \pm (1.96 \times 0.412)$$

$$\text{CI} = (28.02, 29.64)$$

Does the Interval Contain the Population Mean?

- The true population mean from part (a) is 25.78.
- The confidence interval is (28.02, 29.64).
- Since 25.78 is not within the interval, the sample does not provide an accurate estimate under the assumption of SRS.

In a word, the systematic sample overestimated the population mean, suggesting possible sampling bias. Systematic sampling may not behave like an SRS in this case and the confidence interval is too high, indicating that systematic sampling may have oversampled ozone values from certain times.

(d): Confidence Interval for Population Mean (Cluster Sampling)

Using four independent systematic samples (period 96), the computed statistics are:

Sample Mean = 26.36, Sample Standard Deviation = 11.39, Sample Size = 728

The standard error (SE) is calculated as:

$$SE = \frac{\text{Sample Standard Deviation}}{\sqrt{\text{Sample Size}}}$$

$$SE = \frac{11.39}{\sqrt{728}} = 0.422$$

The 95% confidence interval for the population mean is given by:

$$\text{CI} = \bar{x} \pm Z_{\alpha/2} \times SE$$

$$\text{CI} = 26.36 \pm (1.96 \times 0.422)$$

$$\text{CI} = (25.54, 27.19)$$

- The true population mean from part (a) is 25.78.
- The confidence interval is (25.54, 27.19).
- Since 25.78 is within the interval, the estimate is accurate.

In a word, the sample mean (26.36) is very close to the true mean (25.78), suggesting good accuracy. The cluster sampling improved estimation compared to part (c), where the CI did not contain the population mean. The confidence interval correctly includes the population mean, making this a more reliable estimation method.

Question 25

25. The ICC is defined as the Pearson correlation coefficient for the $NM(M-1)$ pairs (y_{ij}, y_{ik}) for i between 1 and N and $j \neq k$:

$$\text{ICC} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\bar{y}_{\mathscr{U}})(y_{ik}-\bar{y}_{\mathscr{U}})}{(NM-1)(M-1)\,S^2}. \tag{5.50}$$

Show that the above definition is equivalent to (5.10). Hint: First show that

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\bar{y}_{\mathscr{U}})(y_{ik}-\bar{y}_{\mathscr{U}}) + \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{\mathscr{U}})^2 = M(\text{SSB}).$$

Figure 12: This is question 25

Answer

The problem requires us to show that the ICC definition:

$$ICC = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}(y_{ij}-\bar{y}_u)(y_{ik}-\bar{y}_u)}{(NM-1)(M-1)S^2}$$

is equivalent to the ANOVA-based formula:

$$ICC = 1 - \frac{M}{M-1}\frac{\text{SSW}}{\text{SSTO}}.$$

To establish this equivalence, we introduce the key components of the ANOVA decomposition and show how they relate to the ICC formula.

1. Sum of Squares Total (SSTO) The total variability in the dataset is captured by the Sum of Squares Total (SSTO), defined as:

$$SSTO = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_u)^2.$$

Here: N is the number of clusters. M is the number of observations per cluster. y_{ij} represents an observation from the ith cluster and jth element. $\bar{y}_u$ is the grand mean, computed as:

$$\bar{y}_u = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}.$$

$SSTO$ measures the total variation in the dataset.

2. Sum of Squares Between (SSB) The between-cluster variability is captured by the Sum of Squares Between (SSB), defined as:

$$SSB = \sum_{i=1}^{N} M(\bar{y}_i - \bar{y}_u)^2.$$

where: $\bar{y}_i$ is the mean of cluster i, computed as:

$$\bar{y}_i = \frac{1}{M} \sum_{j=1}^{M} y_{ij}.$$

SSB quantifies the variation between different clusters.

3. Sum of Squares Within (SSW) The within-cluster variability is captured by the Sum of Squares Within (SSW), defined as:

$$SSW = \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2.$$

SSW measures how much individual observations deviate from their cluster mean, which represents the error variance within clusters.

4. ANOVA Relationship Between SSTO, SSB, and SSW The total variance is decomposed into between-cluster and within-cluster variance:

$$SSTO = SSB + SSW.$$

Now we make the proof of the equation:

Step 1: Express the Numerator in the ICC Formula

The numerator of the given ICC formula is:

$$\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k \neq j} (y_{ij} - \bar{y}_u)(y_{ik} - \bar{y}_u).$$

Using summation identities, we decompose this as:

$$\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k \neq j} (y_{ij} - \bar{y}_u)(y_{ik} - \bar{y}_u) = \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_u)^2 - \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2.$$

Recognizing the ANOVA decomposition, we have the following.

$$\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_u)^2 = SSTO,$$

$$\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_i)^2 = SSW,$$

we obtain:

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}(y_{ij}-\bar{y}_u)(y_{ik}-\bar{y}_u) = M\cdot SSB.$$

Step 2: Derive the ICC Expression

Rewriting the given ICC formula using this result:

$$ICC = \frac{MSSB}{(NM-1)(M-1)S^2}.$$

Using the total sum of squares decomposition:

$$SSTO = SSB + SSW,$$

we rewrite:

$$ICC = \frac{M(SSTO-SSW)}{(NM-1)(M-1)S^2}.$$

Dividing through by $SSTO$:

$$ICC = 1 - \frac{M}{M-1}\frac{SSW}{SSTO}.$$

which matches Equation (5.10) from the problem statement.

The **intraclass** (sometimes called **intracluster**) **correlation coefficient** (ICC) tells us how similar elements in the same cluster are. It provides a **measure of homogeneity** within the clusters. The ICC is defined to be the Pearson correlation coefficient for the $NM(M-1)$ pairs (y_{ij}, y_{ik}) for i between 1 and N and $j \neq k$ (see Exercise 25) and can be written in terms of the population ANOVA table quantities as

$$\text{ICC} = 1 - \frac{M}{M-1}\frac{\text{SSW}}{\text{SSTO}}. \tag{5.10}$$

Figure 13: Equation 5.10

Now, we have successfully shown that the ICC definition:

$$ICC = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}(y_{ij}-\bar{y}_u)(y_{ik}-\bar{y}_u)}{(NM-1)(M-1)S^2}$$

is equivalent to the ANOVA-based formula:

$$ICC = 1 - \frac{M}{M-1}\frac{\text{SSW}}{\text{SSTO}}.$$

This derivation confirms that ICC measures the proportion of variance due to between-cluster variability relative to total variance, making it an essential statistic in hierarchical modeling and reliability analysis.