SURV625, HW-4

Sagnik Chakravarty

Table of contents

1. The following data were collected from a sample of n = 10 clusters that was selected from a large population (assume that the sampling fractions are negligible):

i	1	2	3	4	5	6	7	8	9	10	Totals
$t_{y,i}$	5	1	2	4	2	2	3	3	4	6	32
$t_{x,i}$	13	11	7	11	6	11	5	11	9	10	94

a) Compute the ratio mean r = y/x , where y = ty,i is the total outcome and x= tx,i is the realized sample size, and its standard error. Note that this is an example of simple random sampling of unequal-sized clusters. Use the ultimate cluster idea for variance estimation purposes (i.e., we don't really care how many stages of cluster sampling led to the realized sample sizes in each cluster; we assume a one-stage selection of ultimate clusters, where all units were sampled within them).

In this first scenario, we analyze data from a sample of 10 clusters selected from a large population, with negligible sampling fractions.

In cluster sampling with unequal cluster sizes, the ratio mean is:

$$r = \frac{\sum_{i=1}^{n} t_{y,i}}{\sum_{i=1}^{n} t_{x,i}} = \frac{y}{x}$$

The standard error using the ultimate cluster approach is:

$$SE(r) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (t_{y,i} - r \times t_{x,i})^2}$$

So firstly calculating the total outcome (y) and realized sample size (x)

$$y = \sum_{i=1}^{n} t_{y,i} = 5 + 1 + 2 + 4 + 2 + 2 + 3 + 3 + 4 + 6 = 32$$

$$x = \sum_{i=1}^{n} t_{x,i} = 13 + 11 + 7 + 11 + 6 + 11 + 5 + 11 + 9 + 10 = 94$$

then, computing the ratio mean (r)

$$r = \frac{y}{x} = \frac{32}{94} \approx 0.3404255$$

and then calculating the standard error using the ultimate cluster idea:

$$SE(r) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (t_{y,i} - r \cdot t_{x,i})^2}$$

Where n = 10 (number of clusters)

$$SE(r) = \sqrt{\frac{1}{10(10-1)} \sum_{i=1}^{10} (t_{y,i} - 0.3404255 \cdot t_{x,i})^2}$$

$$SE(r) \approx 0.04826531$$

```r
# Define cluster outcome totals and sizes
outcome_totals <- c(5, 1, 2, 4, 2, 2, 3, 3, 4, 6)
cluster_sizes <- c(13, 11, 7, 11, 6, 11, 5, 11, 9, 10)
num_clusters <- length(outcome_totals)

# Calculate ratio estimate using vector operations
total_outcome <- sum(outcome_totals)
total_size <- sum(cluster_sizes)
ratio_est <- total_outcome / total_size

# Implement ultimate cluster variance estimation
# Calculate deviations from expected values
deviations <- outcome_totals - ratio_est * cluster_sizes
# Sum squared deviations and scale appropriately
variance_ratio <- sum(deviations^2) / (num_clusters * (num_clusters - 1))
std_error <- sqrt(variance_ratio)

# Display results with formatted output
cat(sprintf("Estimated ratio: %.6f\n", ratio_est))
```

```
Estimated ratio: 0.340426
```

```r
cat(sprintf("Standard error: %.6f\n", std_error))
```

```
Standard error: 0.482749
```

The analysis reveals a ratio estimate of approximately 0.34, representing the proportion of individuals with a particular attitude in the population. Using the ultimate cluster approach for variance estimation, we account for the complex sampling design by treating each cluster as a primary sampling unit regardless of the selection process that formed it. The resulting standard error of approximately 0.48 reflects the uncertainty in our estimate given the sampling variability.

b) The mean is actually the proportion of individuals with a particular attitude (meaning that the Y variable is a binary indicator of whether a person has that attitude). Given this information, compute the simple random sampling variance, design effect, and roh. (Hint: Remember that when computing the design effect for these designs, the average sample size per cluster should be used.)

For a binary outcome (proportion), the simple random sampling variance is:

$$V_{SRS}(r) = \frac{r(1-r)}{\sum_{i=1}^{n} t_{x,i}}$$

The design effect is:

$$deff = \frac{V_{cluster}(r)}{V_{SRS}(r)}$$

The intracluster correlation (roh) is:

$$roh = \frac{deff - 1}{\bar{m} - 1}$$

where $\bar{m}$ is the average cluster size:

$$\bar{m} = \frac{\sum_{i=1}^{n} t_{x,i}}{n}$$

So using this, the calculations are:

Cluster variance:

$$V_{cluster}(r) = \frac{\sum_{i=1}^{n}(t_{y,i} - r \cdot t_{x,i})^2}{n(n-1)} = \frac{\sum_{i=1}^{10}(t_{y,i} - 0.3404 \cdot t_{x,i})^2}{10 \times 9} \approx 0.2330$$

SRS variance for binary outcome:

$$V_{SRS}(r) = \frac{r(1-r)}{\sum_{i=1}^{n} t_{x,i}} = \frac{0.3404(1-0.3404)}{94} \approx 0.0024$$

Design effect:

$$deff = \frac{V_{cluster}(r)}{V_{SRS}(r)} = \frac{0.2330}{0.0024} \approx 97.56$$

Average cluster size:

$$\bar{m} = \frac{\sum_{i=1}^{n} t_{x,i}}{n} = \frac{94}{10} = 9.4$$

Intracluster correlation:

$$roh = \frac{deff - 1}{\bar{m} - 1} = \frac{97.56 - 1}{9.4 - 1} \approx 11.50$$

```r
# Import cluster data
outcome_values <- c(5, 1, 2, 4, 2, 2, 3, 3, 4, 6)
sample_counts <- c(13, 11, 7, 11, 6, 11, 5, 11, 9, 10)
cluster_count <- length(outcome_values)

# Compute aggregate statistics
sum_outcomes <- sum(outcome_values)
sum_samples <- sum(sample_counts)
proportion_est <- sum_outcomes / sum_samples

# Analyze design effects
# Calculate variance under cluster design
residuals <- outcome_values - proportion_est * sample_counts
cluster_var <- sum(residuals^2) / (cluster_count * (cluster_count - 1))

# Calculate variance under simple random sampling
srs_var <- proportion_est * (1 - proportion_est) / sum_samples

# Compute design efficiency metrics
design_effect <- cluster_var / srs_var
avg_size <- sum_samples / cluster_count
intraclass_corr <- (design_effect - 1) / (avg_size - 1)

# Present analysis results
cat("Complex design variance:", round(cluster_var, 6), "\n")
```

```
Complex design variance: 0.233047
```

```
  cat("SRS equivalent variance:", round(srs_var, 6), "\n")
```

```
SRS equivalent variance: 0.002389
```

```
  cat("Design effect coefficient:", round(design_effect, 2), "\n")
```

```
Design effect coefficient: 97.56
```

```
  cat("Mean cluster size:", round(avg_size, 1), "\n")
```

```
Mean cluster size: 9.4
```

```
  cat("Intraclass correlation:", round(intraclass_corr, 4), "\n")
```

```
Intraclass correlation: 11.4956
```

The assessment reveals a remarkably high design effect of approximately 97.56, suggesting that our cluster sampling design is substantially less efficient than a comparable simple random sample. This translates to an unusually high intraclass correlation coefficient of about 11.50, which exceeds the typical 0-1 range. Such an extreme value indicates extraordinary homogeneity within clusters—individuals in the same cluster share nearly identical attitudes on the measured characteristic. This finding has significant implications for survey design, as it suggests that sampling additional clusters would be vastly more efficient than increasing within-cluster sample sizes.

c) Estimate the variance if the sample size were tripled by tripling the number of primary stage cluster selections from 10 to 30.

Now from the class notes, when increasing the number of clusters from n to n', the variance of the ratio estimate scales as:

$$V_{new}(r) = \frac{n}{n'} \times V_{old}(r)$$

Where: - $n = 10$ (original number of clusters) - $n' = 30$ (new number of clusters) - $V_{old}(r)$ is the original variance from part (a)

Now the original variance from part (a):

$$V_{original}(r) = 0.2330$$

New variance with tripled clusters:

$$V_{new}(r) = \frac{n}{n'} \times V_{original}(r) = \frac{10}{30} \times 0.2330 = \frac{1}{3} \times 0.2330 \approx 0.0777$$

Variance reduction:

$$\text{Reduction} = (1 - \frac{V_{new}(r)}{V_{original}(r)}) \times 100$$

```r
# Scenario analysis: increased cluster count
base_clusters <- 10
expanded_clusters <- 30  # Triple the number of clusters

# Retrieve baseline variance from previous calculation
baseline_variance <- cluster_var

# Calculate projected variance with expanded design
# Variance scales inversely with cluster count
projected_variance <- baseline_variance * (base_clusters / expanded_clusters)

# Evaluate efficiency improvement
variance_reduction <- (1 - projected_variance/baseline_variance) * 100

# Report findings
cat("Current design variance:", sprintf("%.6f", baseline_variance), "\n")
```

```
Current design variance: 0.233047
```

```r
cat("Projected variance with", expanded_clusters, "clusters:",
    sprintf("%.6f", projected_variance), "\n")
```

```
Projected variance with 30 clusters: 0.077682
```

```r
cat("Efficiency gain:", sprintf("%.1f%%", variance_reduction), "\n")
```

```
Efficiency gain: 66.7%
```

Our analysis indicates that expanding from 10 to 30 clusters would yield a substantial 66.7% reduction in variance. This confirms our earlier conclusion about the high intraclass correlation—in situations with strong within-cluster homogeneity, allocating resources toward sampling more clusters rather than more individuals within clusters produces dramatically better precision. This variance reduction would translate directly to narrower confidence intervals and more reliable statistical inferences.

d) Estimate the sampling variance if the sample size were tripled by tripling the subsampling rate in each cluster.

When increasing the subsampling rate within each cluster, I will use the formula that accounts for the intra-cluster correlation:

$$V_{new}(r) = \frac{V_{SRS}(r)}{m'} \times [1 + (m' - 1) \times roh]$$

Where: - $m' = 3 \times \bar{m}$ is the new average cluster size (tripled) - $\bar{m} = 9.4$ is the original average cluster size - $roh = 11.50$ is the intracluster correlation - $V_{SRS}(r)$ is the simple random sampling variance

Original average cluster size:

$$\bar{m} = 9.4$$

New average cluster size:

$$m' = 3 \times 9.4 = 28.2$$

Original SRS variance:

$$V_{SRS}(r) = 0.0024$$

New SRS variance:

$$V_{SRS,new}(r) = V_{SRS}(r) \times \frac{\bar{m}}{m'} = 0.0024 \times \frac{9.4}{28.2} \approx 0.0008$$

Original design effect:

$$deff_{original} = 1 + (\bar{m} - 1) \times roh = 1 + (9.4 - 1) \times 11.50 \approx 97.6$$

New design effect:

$$deff_{new} = 1 + (m' - 1) \times roh = 1 + (28.2 - 1) \times 11.50 \approx 312.8$$

New variance:

$$V_{new}(r) = V_{SRS,new}(r) \times deff_{new} = 0.0008 \times 312.8 \approx 0.2502$$

Percent change:

$$\text{Change} = (1 - \frac{V_{new}(r)}{V_{original}(r)}) \times 100$$

```
# Alternative scenario: increased sampling within clusters
current_avg_size <- avg_size
increased_avg_size <- 3 * current_avg_size  # Triple sampling rate
fixed_clusters <- cluster_count  # Number of clusters stays constant

# Method A: Design effect approach
adjusted_srs_var <- srs_var * (current_avg_size/increased_avg_size)
new_design_effect <- 1 + (increased_avg_size - 1) * intraclass_corr
projected_var_a <- adjusted_srs_var * new_design_effect

# Method B: Component ratio approach
original_design_factor <- 1 + (current_avg_size - 1) * intraclass_corr
scaling_factor <- (current_avg_size/increased_avg_size) * (new_design_effect/original_design_fa
projected_var_b <- cluster_var * scaling_factor

# Calculate efficiency impact
variance_change <- (1 - projected_var_a/cluster_var) * 100

# Display comparative analysis
cat("Baseline variance estimate:", sprintf("%.6f", cluster_var), "\n")
```

```
Baseline variance estimate: 0.233047
```

```
cat("Projected variance (primary method):", sprintf("%.6f", projected_var_a), "\n")
```

```
Projected variance (primary method): 0.249760
```

```
cat("Projected variance (validation method):", sprintf("%.6f", projected_var_b), "\n")
```

```
Projected variance (validation method): 0.249760
```

```r
cat("Efficiency impact:", sprintf("%.1f%%", variance_change),
    ifelse(variance_change > 0, "improvement", "reduction"), "\n")
```

```
Efficiency impact: -7.2% reduction
```

Surprisingly, tripling the within-cluster sampling rate actually increases estimation variance by approximately 7.2%. This counterintuitive result directly stems from the extremely high intraclass correlation of 11.5. When individuals within clusters are so similar, adding more observations from the same clusters provides minimal new information. In fact, the additional sampling effort is effectively wasted due to this redundancy. This finding reinforces our previous conclusion that resources would be much better allocated toward sampling more clusters rather than more individuals within each cluster.

e) Compute the coefficient of variation of the denominator based on the current design [from part (a)]. Remember to account for the cluster sampling design in your calculation. Is the Taylor series approximation adequate?

The coefficient of variation (CV) of the denominator is defined as:

$$CV(x) = \frac{\sqrt{V(x)}}{E(x)} = \frac{\sqrt{V(\sum_{i=1}^{n} t_{x,i})}}{\sum_{i=1}^{n} t_{x,i}}$$

Where $V(x)$ is the variance of the denominator under cluster sampling:

$$V(x) = n \cdot V(t_{x,i}) = \frac{n}{n-1} \sum_{i=1}^{n} (t_{x,i} - \bar{t}_x)^2$$

And $\bar{t}_x = \frac{1}{n} \sum_{i=1}^{n} t_{x,i}$ is the mean cluster size.

For the Taylor series approximation to be adequate, typically CV(x) should be less than 0.1 (or 10%).

So the Total denominator:

$$x = \sum_{i=1}^{n} t_{x,i} = 13 + 11 + 7 + 11 + 6 + 11 + 5 + 11 + 9 + 10 = 94$$

Mean cluster size:

$$\bar{t}x = \frac{1}{n} \sum i = 1^{n} t_{x,i} = \frac{94}{10} = 9.4$$

Variance of cluster sizes:

$$V(t_{x,i}) = \frac{1}{n-1} \sum_{i=1}^{n} (t_{x,i} - \bar{t}x)^2 = \frac{1}{9} \sum i = 1^{10} (t_{x,i} - 9.4)^2$$

Variance of denominator:

$$V(x) = n \cdot V(t_{x,i})$$

```r
# Assess Taylor series approximation adequacy
size_vector <- c(13, 11, 7, 11, 6, 11, 5, 11, 9, 10)
n_elements <- length(size_vector)
total_size <- sum(size_vector)
average_size <- mean(size_vector)

# Calculate variance components
```

```r
# Between-cluster size variation
size_variance <- sum((size_vector - average_size)^2) / (n_elements - 1)
# Total size variance accounting for cluster design
denominator_variance <- n_elements * size_variance
# Relative variation measurement
coef_variation <- sqrt(denominator_variance) / total_size

# Evaluate approximation validity
taylor_adequate <- coef_variation < 0.1

# Generate assessment report
cat("Aggregate denominator:", total_size, "\n")
```

```
Aggregate denominator: 94
```

```r
cat("Mean element size:", sprintf("%.2f", average_size), "\n")
```

```
Mean element size: 9.40
```

```r
cat("Element size variance:", sprintf("%.4f", size_variance), "\n")
```

```
Element size variance: 6.7111
```

```r
cat("Total denominator variance:", sprintf("%.4f", denominator_variance), "\n")
```

```
Total denominator variance: 67.1111
```

```r
cat("Coefficient of variation:", sprintf("%.4f", coef_variation), "\n")
```

```
Coefficient of variation: 0.0872
```

```r
cat("Taylor series approximation status:",
    ifelse(taylor_adequate, "ADEQUATE", "INADEQUATE"),
    "(threshold = 0.1)\n")
```

```
Taylor series approximation status: ADEQUATE (threshold = 0.1)
```

The coefficient of variation for the denominator is approximately 0.087 (8.7%), which falls below the conventional threshold of 0.1 (10%) required for the Taylor series approximation to be considered valid. This indicates that the relative variability in cluster sizes is modest enough that our linearization-based approach to variance estimation should produce reliable results. Had this value exceeded the threshold, alternative methods such as jackknife or bootstrap resampling might have been necessary for accurate variance estimation.

2. The following are cluster totals from five strata, with two primary stage selections per stratum, for a binary variable named "total cholesterol greater than 200" ($t_{y,h,i}$):

h	i	$t_{y,h,i}$	$t_{x,h,i}$
1	1	16	23
	2	15	25
2	1	9	17
	2	5	15
3	1	8	20
	2	10	21
4	1	6	16
	2	10	19
5	1	10	12
	2	7	16

a) Compute an estimate of the proportion with total cholesterol greater than 200, and its standard error (you can ignore the finite population corrections again in this case). Make sure that you are carefully accounting for this specific type of cluster sampling design in your variance estimation.

For stratified two-stage cluster sampling with two PSUs per stratum, one can estimate the proportion as:

$$\hat{p} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} t_{y,h,i}}{\sum_{h=1}^{L} \sum_{i=1}^{n_h} t_{x,h,i}}$$

For the variance estimation, I will use the between-PSU variance formula with two PSUs per stratum:

$$V(\hat{p}) = \sum_{h=1}^{L} \frac{(z_{h1} - z_{h2})^2}{4}$$

where $z_{hi} = t_{y,h,i} - \hat{p} \cdot t_{x,h,i}$

Now to calculate overall proportion:

$$\hat{p} = \frac{\sum_{h=1}^{5} \sum_{i=1}^{2} t_{y,h,i}}{\sum_{h=1}^{5} \sum_{i=1}^{2} t_{x,h,i}} = \frac{96}{184} \approx 0.5217$$

Calculate residuals for each PSU:

$$z_{h,i} = t_{y,h,i} - \hat{p} \cdot t_{x,h,i}$$

So an example, for stratum 1, PSU 1:

$$z_{1,1} = 16 - 0.5217 \cdot 23 = 16 - 12 = 4$$

Calculating variance component for each stratum:

$$v_h = \frac{(z_{h1} - z_{h2})^2}{4}$$

For example, for stratum 1:

$$v_1 = \frac{(z_{1,1} - z_{1,2})^2}{4}$$

Summing the variance components across strata:

$$V(\hat{p}) = \sum_{h=1}^{5} v_h$$

Calculating the standard error:

$$SE(\hat{p}) = \sqrt{V(\hat{p})}$$

```r
# Prepare stratified cluster sampling data
survey_df <- data.frame(
    strata_id = rep(1:5, each = 2),
    primary_unit = rep(1:2, times = 5),
    cholesterol_counts = c(16, 15, 9, 5, 8, 10, 6, 10, 10, 7),
    sample_sizes = c(23, 25, 17, 15, 20, 21, 16, 19, 12, 16)
)

# Calculate population proportion estimate
numerator_sum <- sum(survey_df$cholesterol_counts)
denominator_sum <- sum(survey_df$sample_sizes)
proportion_estimate <- numerator_sum / denominator_sum

# Implement stratified variance estimation
# Step 1: Calculate deviations from expected values
survey_df$residuals <- survey_df$cholesterol_counts -
                       proportion_estimate * survey_df$sample_sizes

# Step 2: Calculate stratum-level variance components using split-apply-combine
variance_by_stratum <- aggregate(
    residuals ~ strata_id,
    data = survey_df,
    FUN = function(res) {
        # Between-PSU formula for two units per stratum
        (res[1] - res[2])^2 / 4
    }
)

# Step 3: Sum variance components and calculate standard error
total_variance <- sum(variance_by_stratum$residuals)
standard_error <- sqrt(total_variance) / denominator_sum

# Output analysis results
cat("Cholesterol prevalence estimate:", sprintf("%.4f", proportion_estimate), "\n")
```

```
Cholesterol prevalence estimate: 0.5217
```

```r
cat("Estimated total variance:", sprintf("%.4f", total_variance), "\n")
```

```
Estimated total variance: 11.7268
```

```r
cat("Standard error of proportion:", sprintf("%.6f", standard_error), "\n")
```

```
Standard error of proportion: 0.018611
```

Our analysis estimates that approximately 52.17% of the population has total cholesterol levels exceeding 200 mg/dL. The standard error of 0.019 indicates moderate precision in this estimate. The variance estimation approach accounts for the complex stratified cluster design by calculating between-PSU variation within each stratum, then combining these components across all strata. This method appropriately reflects the sampling uncertainty while respecting the hierarchical structure of the data collection.

b) Give a 95% confidence interval for the proportion, making sure to use the correct degrees of freedom according to this design.

For a stratified two-stage cluster sampling design with two PSUs per stratum, the degrees of freedom are calculated as:

$$df = L - 1$$

where L is the number of strata.

$$df = L - 1 = 5 - 1 = 4$$

The 95% confidence interval is then:

$$\hat{p} \pm t_{df,0.975} \times SE(\hat{p})$$

t-critical value:

$$t_{4,0.975} = 2.776$$

So the confidence interval:

$$\hat{p} \pm t_{4,0.975} \times SE(\hat{p}) = 0.5217 \pm 2.776 \times 0.0186 = 0.5217 \pm 0.0517$$

Now the lower and upper bounds:

$$CI_{lower} = 0.5217 - 0.0517 = 0.4701$$

$$CI_{upper} = 0.5217 + 0.0517 = 0.5734$$

Therefore:

$$CI_{95\%} = (0.4701, 0.5734)$$

```r
# Load health survey data
cholesterol_data <- data.frame(
    stratum_code = rep(1:5, each = 2),
    sampling_unit = rep(1:2, times = 5),
    high_chol_count = c(16, 15, 9, 5, 8, 10, 6, 10, 10, 7),
    respondent_count = c(23, 25, 17, 15, 20, 21, 16, 19, 12, 16)
)

# Estimate population parameters
total_cases <- sum(cholesterol_data$high_chol_count)
total_respondents <- sum(cholesterol_data$respondent_count)
prev_estimate <- total_cases / total_respondents

# Calculate variance estimates accounting for complex design
# Generate unit deviations
cholesterol_data$deviation <- cholesterol_data$high_chol_count -
                              prev_estimate * cholesterol_data$respondent_count

# Calculate stratum-specific variance contributions
strata_var <- by(cholesterol_data$deviation,
                 cholesterol_data$stratum_code,
                 function(d) (d[1] - d[2])^2 / 4)

# Combine variance components
combined_variance <- sum(as.numeric(strata_var))
```

```r
proportion_se <- sqrt(combined_variance) / total_respondents

# Determine confidence interval parameters
strata_count <- length(unique(cholesterol_data$stratum_code))
degrees_freedom <- strata_count - 1
critical_t <- qt(0.975, degrees_freedom)

# Construct confidence interval
lower_bound <- prev_estimate - critical_t * proportion_se
upper_bound <- prev_estimate + critical_t * proportion_se
conf_interval <- c(lower_bound, upper_bound)

# Present statistical findings
cat("High cholesterol prevalence:", sprintf("%.4f", prev_estimate), "\n")
```

```
High cholesterol prevalence: 0.5217
```

```r
cat("Standard error:", sprintf("%.6f", proportion_se), "\n")
```

```
Standard error: 0.018611
```

```r
cat("Design-adjusted df:", degrees_freedom, "\n")
```

```
Design-adjusted df: 4
```

```r
cat("Critical t-value:", sprintf("%.4f", critical_t), "\n")
```

```
Critical t-value: 2.7764
```

```r
cat("95% CI: [", sprintf("%.4f", lower_bound), ", ",
    sprintf("%.4f", upper_bound), "]\n", sep="")
```

```
95% CI: [0.4701, 0.5734]
```

Based on our stratified cluster sampling design, we can state with 95% confidence that the true proportion of individuals with elevated cholesterol (>200 mg/dL) falls between 47.01% and 57.34%. The confidence interval construction accounts for the complex survey design by using the appropriate degrees of freedom (4, derived from the 5 strata). Notably, with this limited number of strata, we must use a t-critical value of 2.78 rather than the 1.96 value associated with normal approximations. This wider multiplier properly reflects the additional uncertainty from estimating variance with few strata.

c) Compute the design effect and roh for the proportion in (a).

For stratified cluster sampling, the design effect and intracluster correlation can be calculated as follows:

Design effect:

$$deff = \frac{V_{cluster}(\hat{p})}{V_{SRS}(\hat{p})}$$

Where: - $V_{cluster}(\hat{p})$ is the variance under cluster sampling - $V_{SRS}(\hat{p}) = \frac{p(1-p)}{n}$ is the variance under simple random sampling

Intracluster correlation:

$$roh = \frac{deff - 1}{\bar{m} - 1}$$

Where: - $\bar{m}$ is the average cluster size

So the calculation will be:

Total sample size:

$$total_x = 23 + 25 + 17 + 15 + 20 + 21 + 16 + 19 + 12 + 16 = 184$$

Average cluster size:

$$\bar{m} = \frac{total_x}{n_{PSU}} = \frac{184}{10} = 18.4$$

Variance under cluster sampling:

$$V_{cluster}(\hat{p}) = \left(\frac{\sqrt{var_{p\_hat}}}{total_x}\right)^2 = \left(\frac{\sqrt{11.73}}{184}\right)^2 = 0.00035$$

Variance under SRS:

$$V_{SRS}(\hat{p}) = \frac{p(1-p)}{total_x} = \frac{0.5217(1-0.5217)}{184} = \frac{0.5217 \times 0.4783}{184} = 0.00136$$

Design effect:

$$deff = \frac{V_{cluster}(\hat{p})}{V_{SRS}(\hat{p})} = \frac{0.00035}{0.00136} = 0.255$$

Intracluster correlation:

$$roh = \frac{deff - 1}{\bar{m} - 1} = \frac{0.255 - 1}{18.4 - 1} = \frac{-0.745}{17.4} = -0.043$$

```r
# Analyze design efficiency for stratified cholesterol survey
health_survey <- data.frame(
    region = rep(1:5, each = 2),
    site = rep(1:2, times = 5),
    positive_cases = c(16, 15, 9, 5, 8, 10, 6, 10, 10, 7),
    participants = c(23, 25, 17, 15, 20, 21, 16, 19, 12, 16)
)

# Calculate key survey metrics
observed_cases <- sum(health_survey$positive_cases)
total_participants <- sum(health_survey$participants)
observed_prevalence <- observed_cases / total_participants

# Retrieve variance estimate from previous analysis
variance_complex <- combined_variance  # From part (b)
scaled_variance <- variance_complex / (total_participants^2)

# Calculate theoretical SRS variance for comparison
theoretical_variance <- observed_prevalence * (1 - observed_prevalence) / total_participants

# Compute design efficiency metrics
design_efficiency <- scaled_variance / theoretical_variance
```

```r
# Calculate cluster characteristics
mean_cluster_size <- total_participants / nrow(health_survey)

# Estimate intraclass correlation
icc_estimate <- (design_efficiency - 1) / (mean_cluster_size - 1)

# Generate design assessment report
cat("Survey sample size:", total_participants, "\n")
```

```
Survey sample size: 184
```

```r
cat("Mean observations per cluster:", sprintf("%.1f", mean_cluster_size), "\n")
```

```
Mean observations per cluster: 18.4
```

```r
cat("Complex design variance:", sprintf("%.8f", scaled_variance), "\n")
```

```
Complex design variance: 0.00034637
```

```r
cat("SRS equivalent variance:", sprintf("%.8f", theoretical_variance), "\n")
```

```
SRS equivalent variance: 0.00135613
```

```r
cat("Design efficiency factor:", sprintf("%.4f", design_efficiency), "\n")
```

```
Design efficiency factor: 0.2554
```

```r
cat("Intraclass correlation coefficient:", sprintf("%.6f", icc_estimate), "\n")
```

```
Intraclass correlation coefficient: -0.042792
```

Interestingly, our analysis reveals a design effect of approximately 0.255, which is less than 1.0. This indicates that our stratified cluster sampling approach actually provides better precision than a simple random sample of equivalent size—a relatively uncommon situation in complex survey designs. The negative intraclass correlation (-0.043) further confirms this finding, suggesting that individuals within the same cluster tend to be more different from each other than individuals from different clusters regarding cholesterol levels. This beneficial design effect likely stems from effective stratification that successfully captured major sources of variation in the population. The stratification appears to have grouped individuals in a way that minimizes within-stratum variation while maximizing between-stratum differences.

d) Estimate the standard error expected if the sample size were doubled by doubling the number of primary selections from two to four in each stratum.

When the number of primary selections per stratum is doubled from 2 to 4, the standard error decreases by a factor of $\sqrt{2}$:

$$SE_{new}(\hat{p}) = \frac{SE_{original}(\hat{p})}{\sqrt{2}}$$

This is because:

$$V_{new}(\hat{p}) = \frac{V_{original}(\hat{p})}{2}$$

So I can derive this from the general formula for variance estimation with n_h PSUs per stratum:

$$V(\hat{p}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2$$

Original standard error (from part a):

$$SE_{original}(\hat{p}) = 0.0186$$

New standard error with doubled PSUs:

$$SE_{new}(\hat{p}) = \frac{SE_{original}(\hat{p})}{\sqrt{2}} = \frac{0.0186}{\sqrt{2}} \approx 0.0132$$

Percentage reduction:

$$\text{Reduction} = \left(1 - \frac{SE_{new}(\hat{p})}{SE_{original}(\hat{p})}\right) \times 100\% = \left(1 - \frac{0.0132}{0.0186}\right) \times 100\% \approx 29.3\%$$

```
# Project precision improvements with expanded sampling design
# Current design parameters
current_se <- proportion_se

# Calculate projected precision with doubled PSUs per stratum
# When doubling PSUs per stratum, SE reduces by factor of sqrt(2)
expanded_design_se <- current_se / sqrt(2)

# Evaluate precision improvement
precision_gain <- (1 - expanded_design_se/current_se) * 100

# Generate comparative analysis
cat("Current precision (SE):", sprintf("%.6f", current_se), "\n")
```

```
Current precision (SE): 0.018611
```

```
cat("Projected precision with 4 PSUs per stratum:",
    sprintf("%.6f", expanded_design_se), "\n")
```

```
Projected precision with 4 PSUs per stratum: 0.013160
```

```r
cat("Expected precision improvement:",
    sprintf("%.2f%%", precision_gain), "\n")
```

```
Expected precision improvement: 29.29%
```

Our analysis indicates that doubling the number of PSUs per stratum from 2 to 4 would reduce the standard error by approximately 29.3%, from about 0.0186 to 0.0132. This improvement follows the expected pattern where standard errors decrease proportionally to the square root of the sample size increase. For future studies with similar objectives, this finding suggests that allocating resources toward additional PSUs within existing strata would yield meaningful precision improvements. This approach would be particularly beneficial given the current design's limited degrees of freedom, which impacts confidence interval width.

e) Compute the coefficient of variation of the denominator. Remember to account for the stratified cluster sampling design in your calculation. Is the Taylor series approximation adequate?

For stratified cluster sampling, the coefficient of variation (CV) of the denominator is:

$$CV(x) = \frac{\sqrt{V(x)}}{E(x)} = \frac{\sqrt{V(\sum_{h=1}^{L}\sum_{i=1}^{n_h} t_{x,h,i})}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h} t_{x,h,i}}$$

The variance of the denominator under stratified cluster sampling is:

$$V(x) = \sum_{h=1}^{L} \frac{(t_{x,h,1} - t_{x,h,2})^2}{4}$$

For the Taylor series approximation to be adequate, typically CV(x) should be less than 0.1 (or 10%).

Now the calculation will be:

Total denominator:

$$x = \sum_{h=1}^{5}\sum_{i=1}^{2} t_{x,h,i} = 184$$

Variance components for each stratum:

For stratum 1:

$$v_1 = \frac{(t_{x,1,1} - t_{x,1,2})^2}{4} = \frac{(23-25)^2}{4} = \frac{4}{4} = 1$$

For stratum 2:

$$v_2 = \frac{(t_{x,2,1} - t_{x,2,2})^2}{4} = \frac{(17-15)^2}{4} = \frac{4}{4} = 1$$

For stratum 3:

$$v_3 = \frac{(t_{x,3,1} - t_{x,3,2})^2}{4} = \frac{(20-21)^2}{4} = \frac{1}{4} = 0.25$$

For stratum 4:

$$v_4 = \frac{(t_{x,4,1} - t_{x,4,2})^2}{4} = \frac{(16-19)^2}{4} = \frac{9}{4} = 2.25$$

For stratum 5:

$$v_5 = \frac{(t_{x,5,1} - t_{x,5,2})^2}{4} = \frac{(12-16)^2}{4} = \frac{16}{4} = 4$$

Hence, Total variance:

$$V(x) = \sum_{h=1}^{5} v_h = 1 + 1 + 0.25 + 2.25 + 4 = 8.5$$

Now to check the Coefficient of variation,

$$V(x) = 8.5$$

and

$$x = 184$$

so

$$CV(x) = \frac{\sqrt{8.5}}{184} = \frac{2.92}{184} = 0.016$$

Since CV(x) $= 0.016 < 0.1$, the Taylor series approximation is adequate.

```r
# Evaluate Taylor series approximation validity for stratified design
survey_frame <- data.frame(
    geographic_stratum = rep(1:5, each = 2),
    facility_id = rep(1:2, times = 5),
    elevated_chol = c(16, 15, 9, 5, 8, 10, 6, 10, 10, 7),
    examined_patients = c(23, 25, 17, 15, 20, 21, 16, 19, 12, 16)
)

# Calculate total sample size
total_examinations <- sum(survey_frame$examined_patients)

# Calculate variance of denominator using stratified design formula
# Process each stratum separately
denominator_var_by_stratum <- by(survey_frame$examined_patients,
                                 survey_frame$geographic_stratum,
                                 function(patients) {
                                     # Formula for two PSUs per stratum
                                     (patients[1] - patients[2])^2 / 4
                                 })

# Combine components from all strata
total_denominator_variance <- sum(as.numeric(denominator_var_by_stratum))

# Calculate coefficient of variation for denominator
cv_denominator <- sqrt(total_denominator_variance) / total_examinations

# Assess approximation adequacy
approximation_status <- cv_denominator < 0.1

# Generate technical assessment
cat("Total patient examinations:", total_examinations, "\n")
```

```
Total patient examinations: 184
```

```r
cat("Estimated denominator variance:", sprintf("%.2f", total_denominator_variance), "\n")
```

```
Estimated denominator variance: 8.50
```

```
cat("Coefficient of variation:", sprintf("%.6f", cv_denominator), "\n")
```

```
Coefficient of variation: 0.015845
```

```
cat("Taylor series approximation validity:",
    ifelse(approximation_status, "VALID", "INVALID"),
    "(threshold: CV < 0.1)\n")
```

```
Taylor series approximation validity: VALID (threshold: CV < 0.1)
```

The coefficient of variation for the denominator in our stratified design is approximately 0.016 (1.6%), well below the 0.1 (10%) threshold required for Taylor series linearization to be considered valid. This low value indicates minimal relative variability in cluster sizes across PSUs. The largest contribution to denominator variance comes from stratum 5 where cluster sizes differ notably (12 vs. 16), while stratum 3 contributes least with nearly identical sizes (20 vs. 21). Overall, this assessment confirms the appropriateness of our linearization-based variance estimation approach throughout this analysis.