

Data 407 - Assignment 2:

Samira Almuallim, 62197256, Data 407 2025WT2

1. (2 marks) Which are true about the reasons for stratified sampling?

Answers: a, b, c, & d

- a. **Enhance representation** - ensures key subgroups are properly represented.
- b. **Reduce sampling bias** - especially when strata differ systematically.
- c. **Improve precision** - variance is reduced when strata are internally homogeneous.
- d. **Identify patterns and trends** - allows separate analysis within strata.
- e. **Make research easier and cheaper** - stratification **adds** cost and complexity.

Thus: the answers are all except E

Answers: a, b, c, & d

2. (2 marks) Which of the following is correct regarding the sampling of an SRS from each stratum?

Answer: a. We independently take an SRS from each stratum.

In stratified sampling - each stratum is sampled **independently** using SRS. Dependence would violate the design pre assumptions

3. (2 marks) Why would anyone ever take an SRS that is not stratified?

Answers: a, b, c, & d

- a. Stratification adds complexity and may not justify the gain in precision - **correct**

- **b.** We need auxiliary information to form strata - **correct**
- **c.** We must know which population members belong to each stratum - **correct**
- **d.** We must know both the size and membership of each stratum - **correct**
- **e.** SRS would produce better estimates - **IN-correct** stratification never worsens precision if done right

Thus: the answers are all except E

Answers: a, b, c, & d

4. (2 marks) Regarding stratified sampling, what is the variance of $\bar{y}_{\text{str}} = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h$?

Answer: d $\sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{N^2} \frac{S_h^2}{n_h}$

Explanation:

- Each stratum is sampled by SRS without replacement
- The var of the stratified mean is = the **weighted sum of within-stratum variances**, each with a finite population correction:

$$Var(\bar{y}_{\text{str}}) = \sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}.$$

5. (2 marks) In stratified sampling, is \bar{y}_{str} an unbiased estimate of \bar{y}_U ?

Answer: Yes

Each stratum sample mean \bar{y}_h is an unbiased estimator of the stratum mean, and the stratified mean is a weighted average of these unbiased estimators: $E(\bar{y}_{\text{str}}) = \bar{y}_U$.

Thus: **Ans: a. Yes**

6. (2 marks) Regarding stratified sampling, what is the variance of $\hat{p}_{\text{str}} = \sum_{h=1}^H \frac{N_h}{N} \hat{p}_h$?

Answer: C $\sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{N^2} \frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}$

Within each stratum, the variance of \hat{p}_h under SRS without replacement is $\left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}$. Weighting by $(N_h/N)^2$ and summing across strata gives the result.

Thus: **Ans: c.** $\sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{N^2} \frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}$

- See lec 7 notes p 10 for why its $n_h - 1$ and not n_h

7. (3 marks) Stratified sample of NYC food stores (Hayes, 2000)

Given strata information:

Stratum	Income level	N_h	n_h	\bar{y}_h	s_h^2
1	Low income	190	21	3.925	0.037^2
2	Middle income	407	14	3.938	0.052^2
3	Upper income	811	22	3.942	0.070^2

Total population size = $N = 190 + 407 + 811 = 1408$

a. (1.5 marks) 95% CI for the population total t

The stratified estimator of the population total is $\hat{t}_{\text{str}} = \sum_{h=1}^3 N_h \bar{y}_h$

Compute each contribution:

$$190(3.925) = 745.75$$

$$407(3.938) = 1602.77$$

$$811(3.942) = 3196.96$$

thus: $\hat{t}_{\text{str}} = 745.75 + 1602.77 + 3196.96 = 5545.48$

Variance of \hat{t}_{str}

For stratified sampling,

$$\text{Var}(\hat{t}_{\text{str}}) = \sum_{h=1}^3 N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

Now lets compute each stratum's variance term:

- **Stratum 1:** $190^2 \left(1 - \frac{21}{190}\right) \frac{0.037^2}{21} \approx 2.093$
- **Stratum 2:** $407^2 \left(1 - \frac{14}{407}\right) \frac{0.052^2}{14} \approx 30.893$
- **Stratum 3:** $811^2 \left(1 - \frac{22}{811}\right) \frac{0.070^2}{22} \approx 142.519$

Total variance: $\text{Var}(\hat{t}_{\text{str}}) \approx \text{sum of stratum 1 : 3} \approx 175.505$

Standard error:

$$SE(\hat{t}_{\text{str}}) = \sqrt{175.505} \approx 13.25$$

Using the normal approximation ($z_{0.975} = 1.96$):

$$\text{ME} = 1.96 \times 13.25 \approx 25.97$$

95% CI for t :

$$5545.48 \pm 25.97 = (5519.51, 5571.45)$$

b. (1.5 marks) 95% CI for the population mean \bar{y}_U

The stratified mean estimator is $\bar{y}_{\text{str}} = \frac{\hat{t}_{\text{str}}}{N} = \frac{5545.48}{1408} \approx 3.939$

Var of \bar{y}_{str}

$$\text{Var}(\bar{y}_{\text{str}}) = \frac{1}{N^2} \text{Var}(\hat{t}_{\text{str}}) = \frac{175.505}{1408^2} \approx 0.0000885$$

$$\text{St error: } SE(\bar{y}_{\text{str}}) \approx \sqrt{0.0000885} \approx 0.0094$$

$$\text{Margin of error: } 1.96 \times 0.0094 \approx 0.0184$$

95% CI for \bar{y}_U :

$$3.939 \pm 0.0184 = (3.9206, 3.9574)$$

Final Answers Summary

- **(a)** 95% CI for t :
 $(5519.51, 5571.45)$
- **(b)** 95% CI for \bar{y}_U :
 $(3.9206, 3.9574)$