King Fahd University of Petroleum and Minerals
College of Computing and Mathematics
Department of Mathematics

STAT 565 - Final Exam

AY 2025-2026 (Term 252)

Time Allowed: 180 Minutes

Name: ..... **ID number:** .....

- Textbook, notes, and mobile phones are not allowed in this exam.
- A separate formula sheet is provided.
- Write neatly and legibly. You may lose points for messy work.
- Show all your work. No points for answers without justification.

Question	Marks	Max Marks
1		
2		
3		
4		
5		
Total		

Question 1

a) Define the following terms:

(i) Systematic sample

(ii) Regression estimator

(iii) Self-weighting sample

(iv) Unit nonresponse

(v) Item nonresponse

(vi) Imputation

b) A city wants to estimate the average monthly electricity usage per household. A frame of apartment buildings is available, and for each building the number of apartments is known. The survey team first selects buildings, and then selects apartments within each chosen building.

Describe the target population, sampling frame, PSU, SSU, and whether this is a one-stage or two-stage cluster design. Then give one reason why probability-proportional-to-size (pps) sampling of buildings may be preferable to equal-probability sampling, and state one weighting-based remedy if some sampled households do not respond.

Question 2

A supermarket chain has $N = 4$ stores (PSUs), labeled A, B, C, D. Let M_i be the number of receipts in store i on a given day, and let t_i be the total daily sales (in thousand SAR). Suppose

Store i	A	B	C	D
M_i	10	20	30	100
t_i	11	20	24	245

Stores are sampled with probability proportional to M_i. Answer the following questions:

a) Suppose one store is selected and all receipts in that store are observed. What is the probability that store D is selected? If store D is selected, compute the estimated population total of sales.

b) Suppose 4 first-stage with replacement draws are made. For store C, compute the probability that it is selected at least once, and the expected number of times it is selected.

c) Under two-stage cluster sampling, suppose 4 first-stage with replacement draws are made, and then 8 receipts are selected independently by SRS from each chosen store. Find the sampling weight for a sampled receipt from store C. State whether the design is self-weighting.

d) Under one-stage cluster sampling without replacement, suppose that two stores are selected and the realized sample is $\{B, D\}$, with $\pi_B = 0.40, \quad \pi_D = 0.70, \quad \pi_{BD} = 0.25$. Compute the Horvitz-Thompson estimate of total sales. Then, using the Sen-Yates–Grundy variance estimator, estimate the variance of the HT estimator.

e) Consider two-stage cluster sampling without replacement and suppose stores B and D are selected first, with $\pi_B = 0.40, \quad \pi_D = 0.70$, and then 5 receipts are selected by SRS from each selected store. Find the sampling weights for a sampled receipt from store B and from store D. State whether the design is self-weighting.

Question 3

A finite population has $N = 5$ units with values (y_i, x_i) given below. Assume an SRSWOR of size $n = 4$ is selected, and $\bar{x}_U$ is known.

Unit i	1	2	3	4	5
y_i	3	4	6	6	7
x_i	1	2	3	4	5

Consider the following three estimators of the population mean: $\bar{y}, \quad \hat{\bar{y}}_r, \quad \hat{\bar{y}}_{\text{diff}}$

(a) Compute $\bar{y}_U$ and $\bar{x}_U$.

(b) List all possible SRSWOR samples of size 4. For each sample, compute $\bar{y}, \quad \bar{x}, \quad \hat{\bar{y}}_r, \quad \hat{\bar{y}}_{\text{diff}}$.

(c) Using the results of part (b), verify that $\bar{y}, \hat{\bar{y}}_{\text{diff}}$ are unbiased estimators and find the bias of $\hat{\bar{y}}_r$.

(d) Compute the MSE of each estimator and interpret the results.

#	Sample (size 4)	$\bar{y}$	$\bar{x}$	$\hat{\bar{y}}_r$	$\hat{\bar{y}}_{\text{diff}}$

Question 4

Consider a two-stage cluster sampling design with equal cluster sizes $M_i = M$ and equal within-cluster sample sizes $m_i = m$ for all clusters $i = 1, \ldots, N$. Assume the approximate variance of the estimator of the population mean can be written as

$$V(\hat{\bar{y}}) \approx \frac{1}{n}\left(\frac{B}{M} + \frac{W}{m} - \frac{W}{M}\right),$$

where B and W denote the between-cluster and within-cluster mean squares, respectively.

Suppose the cost function is

$$C = c_1 n + c_2 n m,$$

where c_1 is the cost per sampled cluster and c_2 is the cost per sampled unit within a cluster.

(a) Show that the value of m that minimizes the variance for fixed total cost C is

$$m_{\text{opt}} = \sqrt{\frac{c_1 M\, W}{c_2\big(B - W\big)}}.$$

(b) In a household expenditure survey, clusters correspond to neighborhoods. Suppose households within each neighborhood are very similar. Would you recommend sampling many households per neighborhood or few? Explain.

(c) Briefly explain why cluster sampling may still be preferred over simple random sampling even though it typically leads to higher variance.

Question 5

A simple random sample of $n = 240$ students is selected from a population of $N = 2400$ for a survey on weekly hours of online video viewing. The overall response rate is 70%. Using university records, the sampled students are classified by GPA as follows.

GPA group	Sample size	Respondents	Respondent mean $\bar{y}_h$	Respondent SD s_h
3.00–4.00	90	78	5	2
2.00–2.99	100	70	8	3
Below 2.00	50	20	12	4
Total	240	168		

(a) If only respondents are analyzed, find the estimate of the average number of hours per week and its standard error.

(b) Use the GPA classification to adjust the weights of the respondents in the sample. What is the weighting class estimate of the average viewing time?