KISII UNIVERSITY

UNIVERSITY EXAMINATIONS

FOURTH YEAR EXAMINATION FOR THE AWARD OF THE DEGREE OF BACHELOR OF SCIENCE IN MATHEMATICS AND EDUCATION
FIRST SEMESTER 2022/2023
[SEPTEMBER-DECEMBER, 2022]

STAT 446: SAMPLING METHODS II

STREAM: Y4S1 **TIME: 2 HOURS**

DAY: MONDAY, 9:00 – 11:00 AM **DATE: 19/12/2022**

INSTRUCTIONS

1. ***Do not write anything on this question paper.***
2. ***Answer question ONE and any other TWO questions.***

SECTION A: ANSWER ALL QUESTIONS (30 MARKS)

1.Let X be a sample from a population P and consider testing hypotheses $H0 : P = P0$ versus $H1 : P = P1$, where Pj is a known population with probability density fj with respect to a σ-finite measure ν, $j = 0, 1$. Let $\beta(P)$ be the power function of a UMP (uniformly most powerful) test of size $\alpha \in (0, 1)$. Show that $\alpha < \beta(P1)$ unless $P0 = P1$. Let X be a sample from a population P and consider testing hypotheses $H0 : P = P0$ versus $H1 : P = P1$, where Pj is a known population with probability density fj with respect to a σ-finite measure ν, $j = 0, 1$. Let $\beta(P)$ be the power function of a UMP (uniformly most powerful) test of size $\alpha \in (0, 1)$. Show that $\alpha < \beta(P1)$ unless $P0 = P1$. (5marks)

2.Let Y1, Y2, . . . , Yn constitute a random sample of size n from the continuous parent population

$fY (y; \theta) = (1 + \theta y)/2, - 1 < y < +1; - 1 < \theta < +1.$

(a) Derive the general structure of the rejection region R of the most powerful (MP) test of size α for testing H0 : θ = 0 versus H1 : θ = 0.50. Is this MP test also the uniformly most powerful (UMP) test of H0 : θ = 0 versus H1 : θ > 0? (2marks)

(b) When n = 1, find the critical value of the MP test statistic so that pr(Type I error) = 0.05 when using the MP test developed in part (a) to test H0 : θ = 0 versus H1 : θ = 0.50. (2marks)

(c) When n = 1, what is the numerical value of the power of the MP test of size α = 0.05 developed in part (b) to test H0 : θ = 0 versus H1 : θ = 0.50 when, in fact, θ = 0.50? (2marks)

3.Let Y1, Y2, . . . , Yn constitute a random sample of size n (≥ 2) from an N(μ, σ^2) population.
Consider using $\bar{Y}$ = n^{-1} $\sum_{i=1}^{n} Yi$ to test H0 : μ = 0 versus HA : μ > 0 when α = 0.025. If σ^2 is known to be equal to 1, what is the minimum sample size n∗ required so that β, the probability of a Type II error, is no more than 0.16 when μ = 0.50? (6marks)

4.The survival time T (in years) for patients who have had quadruple bypass surgery (QBPS) is assumed to have the distribution
fT (t; θ) = $\theta^{-2}te^{-t/\theta}$, 0 < t < +∞, θ > 0.
Suppose that survival times t1, t2, . . . , tn are recorded for n randomly selected patients who have had QBPS. In other words, t1, t2, . . . , tn are the realizations of a random sample T1, T2, . . . , Tn of size n from fT (t; θ).
If $\bar{t}$ = n^{-1} $\sum_{i=1}^{n} ti$ = 2.40 when n = 40, what is the P-value for the likelihood ratio test of H0 : θ = 1 versus H1 : θ ≠ 1? (5marks)

5.To estimate the average number $\bar{Y}$ of people per household in a given country,
we carry out a two-stage sampling design:
- 1st stage: Random sampling with replacement of m = 4 villages among M = 400 proportional to size. The size of a village is the number of households it has. Thus, for each of the four independent selections, a village is selected with a probability proportional to its size.
- 2nd stage: Simple random sampling of ni households among Ni if village i is selected.

The data are presented in the table 1 below.
$\bar{\bar{Y}}$ i is the mean number of people per household in village i, according to the sample.
The total number of households in the country is N = 10 000.
a). i)What is the selection probability pi for each of the four villages selected? (The selection probability is the probability a village has of being selected at the time of each of the four independent selections successively done under the same conditions.) (2marks)

Table 1

i	Ni	$\bar{\bar{Y}}$ i
1	20	5.25
2	23	5.50
3	25	4.50
4	18	5.00

ii) Calculate Pr(i /∈ S) as a function of (1 − pi). Deduce the inclusion probability πi = Pr(i ∈ S) as a function of pi. Examine the case where pi is small. (2marks)

b). What is the expression of $\bar{Y}$ (true value) and what is its unbiased estimator? (2marks)

c). Estimate the variance of this estimator. What interest do we have in using sampling with replacement at the 1st stage? (2marks)

SECTION B :ANSWER ANY 2 QUESTIONS (40 MARKS)

6.In a Comparison of two designs with two stages,
a population U with N individuals is divided into M primary units Ui (i = 1, ..., M) of size Ni. We are interested in the total Y of a variable taking the values yi,k (k ∈ Ui), and we denote

Yi = $\sum_{k\epsilon Ui} yi,k$ $\bar{Y}$ i = $\frac{Yi}{Ni}$

, Y =$\sum_{i=1}^{M} Yi$

$s_{2,1}^{2} = \frac{1}{Ni-1} \sum_{k\epsilon Ui} (yi,k - \bar{Y})^2$

$s_{T}^{2} = \frac{1}{M-1} \sum_{i=1}^{M} (Yi - \frac{Y}{M})^2$

$s_{N}^{2} = \frac{1}{M-1} \sum_{i=1}^{M} (Ni - \frac{N}{M})^2$

.

i). a) We select using simple random sampling (without replacement) m primary units, forming a sample S. Calculate the expected value and the variance of the estimator
$\hat{N}$ (S) = $\sum_{i\epsilon S} Ni$ (3marks)

b) In each primary unit of S, we select (by simple random sampling without replacement) a sample of secondary units at a rate f2. This rate is independent of S (strategy A). Calculate f2 so that the final sample size has an expected value $\bar{n}$ fixed in advance (we assume that f2 × Ni is an integer). (3marks)

c) What unbiased estimator $\hat{Y}$ of Y , as a linear function of $\bar{\bar{Y}}$i, do we propose? What is its variance? What does it become if s_2^2 (denoted s_2^2) for all i? (3marks)

d) For m sufficiently large, give a 95% confidence interval for the total size n of the final sample. (3marks)

(ii). We now examine another two-stage sampling design (strategy B). The sample of primary units is selected as previously done, but in each primary unit selected in the first stage, we select using simple random sampling without replacement a sample of size ni = f2Ni, with
f2 = f2(S) = $\bar{n}$/ $\hat{N}$(S)

a) Show that the sample is of fixed size (to be determined), and that for all i of S, the estimator $\hat{Y}$= Ni $\bar{\bar{Y}}$i estimates Yi without bias. Show that $\hat{Y}$defined from i.c.above is always unbiased. (3marks)

b) Calculate the variance of $\hat{Y}$ assuming that $s_{2,i}^2 = s_2^2$ for all i. (3marks)

(iii). Compare the two strategies A and B, under the conditions that we specified. Can we say that one is indisputably better than the other? (2marks)

7.In a list of N individuals, we are interested in a variable y. The individuals are identified by their order on the list, so their order goes from 1 (for the first) to N (for the last). We use systematic sampling with interval h to select n individuals from the list. We assume that: h = N/n ∈ N.

a). Show that everything happens as if we selected a unique cluster of individuals from a population pre-divided into clusters. We will specify what the clusters are, what their size is, and how many there are in the population. (2marks)

b). We henceforth use the following notation:
yi,k = value of y for the kth record counted in cluster number i.
We denote $\bar{Y}$ i as the mean of the yi,k calculated from all the individuals of cluster number i.

i) What is the unbiased estimator $\bar{\bar{Y}}$of the true mean $\bar{Y}$? Show that $\bar{\bar{Y}}$ is effectively unbiased. (3marks)

ii) What is the expression of its true variance, as a function of $\bar{Y}$ i, $\bar{Y}$ and h? (3marks)

iii) How do we estimate this variance without bias? (2marks)

c.(i) Considering the natural splitting of the population into h clusters, write the analysis equation for the variance by noting: (3marks)

$S_i^2=\frac{1}{n-1}\sum_{k=1}^{n}(yi,k-\bar{Y})^2$

ii) Show that if N is large, and if we denote: (3marks)

$s_w^2 = \frac{1}{h(n-1)} \sum_{i=1}^{h} \sum_{k=1}^{n} (yi,k - \bar{Y}i)^2$

then we have:

var($\bar{\bar{Y}}$) ≈ S_y^2 -$\frac{h(n-1)}{N} S_W^2$

iii) Show that systematic sampling is more precise than simple random sampling if and only if: $S_y^2 < s_w^2$

by considering N as very large with respect to n. (2marks)

iv) In order for this condition to be satisfied, it is necessary to ensure that s_w^2 is 'large'. How does this affect yi,k? How do we proceed in order that, in practice, this is indeed the case? (2marks)

8.Ear infections are quite common in infants. To assess whether ear infections in infants tend to occur in both ears rather than in just one ear in a certain U.S. area, the following statistical model is proposed.

For a random sample of n infants whose parents reside in this U.S. area, suppose, for i = 1, 2, . . . , n, that the random variable Xi = 0 with probability (1 − π) if the i-th infant does not have an ear infection, that Xi=1 with probability π(1−θ) if the i-th infant has an ear infection in only one ear, and that Xi = 2 with probability πθ if the i-th infant has ear infections in both ears. Here, π(0 < π < 1) is the probability that an infant has an infection in at least one ear; that is, π is the prevalence in this U.S. area of children with an infection in at least one ear. And, since

$$\text{pr}(X_i = 2 | X_i \geq 1) = \frac{\text{pr}\,[(X_i = 2) \cap (X_i \geq 1)]}{\text{pr}(X_i \geq 1)}$$

$$= \frac{\text{pr}(X_i = 2)}{\text{pr}(X_i \geq 1)}$$

$$= \frac{\pi\theta}{\pi}$$

$$= \theta,$$

it follows that θ(0 < θ < 1) is the conditional probability that an infant has ear infections in both ears given that this infant has at least one ear that is infected.

(a) Show that a score test statistic $\hat{s}$ for testing H0 : θ = θ0 versus (4marks)
H1 : θ ≠ θ0 can be written in the form

$$\hat{s} = \frac{(\bar{\theta} - \theta 0)^2}{\widehat{vo}\,(\bar{\theta})}$$

,
where $\widehat{vo}(\bar{\theta})$ is the estimated variance of $\bar{\theta}$ under the null hypothesis
H0 : θ = θ0.

(b) Suppose that n = 100, that n0 = 20 is the number of infants with no ear infections, that n1 = 35 is the number of infants with an ear infection in only one ear, and that n2 = 45 is the number of infants with ear infections in both ears. Use these data and the score test developed in part (a) to test H0 : θ = 0.50 versus H1 : θ ≠ 0.50 at the α = 0.025 significance level. (4marks)

Do these data provide statistical evidence that it is more likely than not that an infant in this U.S. region will have both ears infected once that infant develops an ear infection?

(c) Assuming that π = 0.80, provide a reasonable value for the smallest value of n, say n*, required so that the power of a one-sided score test of H0 : θ = 0.50 versus H1 : θ > 0.50 at the α = 0.025 level is at least 0.90 when, in fact, θ = 0.60. (4marks)

d)For the one factor ANOVA(Analysis of Variance) model, show that if the model is balanced we have $\hat{\mu} = \bar{Y}$. (8 marks)