



MACHAKOS UNIVERSITY

University Examinations 2018/2019

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCE

FOURTH YEAR SECOND SEMESTER SUPPLEMENTARY EXAMINATION FOR

BACHELOR OF SCIENCE IN STATISTICS AND PROGRAMMING

SST 408: DESIGN AND ANALYSIS OF SAMPLE SURVEY II

DATE: 24/9/2019

TIME: 11:00 – 1:00 PM

INSTRUCTIONS: Attempt Question ONE and any other TWO questions.

QUESTION ONE (30 MARKS)

- a) Differentiate between Cluster, Multiphase and Multistage Sampling. (6 marks)
- b) List any five reasons why we use Multiphase design. (5 marks)
- c) Give any two reasons why unit nonresponse might occur in a household survey. (2 marks)
- d) Name any three ways of reducing non-sampling errors. (4 marks)
- e) Differentiate between Hot Deck and Cold deck as a way of dealing with non-response. (2 marks)
- f) A research was conducted to estimate prevalence of psychiatric disorders. Participants in a phase I sample were given the General Health Questionnaire (GHQ) and classified into three strata based on their GHQ score. The stratification was used to take a stratified random sample of 250 persons for the phase II sample; the Composite International Diagnostic Interview (CIDI), considered to be a more accurate diagnostic tool, was administered to each person in the phase II sample. The CIDI score was used to classify

the phase II sample members as having at least one psychiatric disorder (case) or having no psychiatric disorder (non-case). The results are given in the following table. The counts of cases and non-cases are from the phase II sample.

Stratum	n_h	m_h	Non-case	Case
GHQ ≤ 3 (low)	1049	60	33	27
GHQ=4, 5 (medium)	237	48	14	34
GHQ ≥ 6 (high)	272	142	23	119
Total	1558	250		

- i) Calculate the Phase II sampling weight $w_i^{(2)}$ for each stratum. (3 marks)
 - ii) Use the two-phase sample to estimate the percentage of persons with at least one psychiatric disorder, along with its standard error. Since we do not know the population size N , use a relative phase I weight of $w_i^{(1)} = 1$ (3 marks)
- g) A survey was conducted by means of a self-administered mail questionnaire. Twelve hundred questionnaires were sent to a randomly selected sample of house-dwellers. Follow-up thank you notes were sent a week later. In total, 329 questionnaires were returned, representing a response rate of 27%. This was deemed satisfactory since many mail surveyors consider a 15 to 20% response rate to be good return.
- i) Do you agree that the response rate of 27% is satisfactory? (2 marks)
 - ii) Suppose the investigators came to you for statistical advice on analyzing these data and designing a follow-up survey. What would you tell them? (3 marks)

QUESTION TWO (20 MARKS)

- a) Discuss a multi-phase design using a well labeled flowchart. (4 marks)
- b) Discuss the two types of non-response. (2 marks)
- c) In a wildlife survey, give two scenarios that could lead to missing data. (2 marks)
- d) An email has been sent to you with the subject line "Joint Statistical meetings 2019 Participants Survey". In order to improve the quality of the survey, what are the possible sources of error and how can you help mitigate them. (4 marks)

- e) What assumptions will you as an investigator make when using Capture-Recapture method in fish estimation? (3 marks)
- f) Describe the variables used in the equations below under two phase sampling. (5 marks)

$$V(t_y^{(1)}) = N^2 \left(1 - \frac{n^{(1)}}{N} \right) \frac{\hat{S}_y^2}{n^{(1)}}$$

$$\hat{S}_y^2 = \frac{1}{n^{(2)} - 1} \sum_{i \in S^{(2)}} (y_i - \bar{y}^{(2)})^2$$

QUESTION THREE (20 MARKS)

- a) Recall the factors that influence response rate in a survey. (4 marks)
- b) Justify any two ways to deal with non-response under imputation. (2 marks)
- c) Suppose that all calls are made during Monday through Friday evenings. Each respondent is asked whether he or she was at home at the time of the interview, on each of the four preceding weeknights. Respondent i replies that she was home k_i times of the 4 nights. It is then assumed that the probability of response is proportional to the number of nights at home during interviewing hours, so the probability of response is estimated by

$$\hat{\phi}_i = (k_i + 1)/5. \text{ Let } \hat{\bar{y}}_{HPS} = \frac{\sum_{i \in s} w_i y_i / \hat{\phi}_i}{\sum_{i \in s} w_i / \hat{\phi}_i}.$$

- i) Under what circumstances would you expect the method to reduce bias due to nonresponse? (4 marks)
- ii) What assumptions must be made for the estimator to be approximately unbiased? (3 marks)
- iii) What are some potential drawbacks of the method for use in practice? (4 marks)
- iv) How does it adjust for persons who were not at home during any of the five nights, or who refused to participate in the survey? (3 marks)

QUESTION FOUR (20 MARKS)

- a) State three disadvantages and advantages of Multiphase design (6 marks)
- b) Identify the different ways to deal with Non-response. (4 marks)
- c) Data mining in statistics are used to discover relationships among variables in very large data sets. A company, for example, has database of all its financial transactions. A very small fraction of these transactions involve fraud, but fraudulent transactions are expensive for the company. Discovering whether a transaction is fraudulent requires an investigation, so the company can only determine whether transactions are fraudulent for a small sample. Discuss how two-phase sampling might be used to improve prediction of fraudulent transactions. (5 marks)
- d) A study conducted using two-phase sampling with ratio estimation to estimate the density of nesting birds. The phase 1 sample, selected from the 2130 plots in the region of interest, is conducted using a rapid search method involving bird sightings to obtain an approximate count of birds in each phase 1 plot. Then, a subsample of 12 of the phase 1 sample plots are surveyed using an intensive method, a surveyor visits a plot for several hours over a period of days and searches for nests and other indications of territorial males in the plot. In the setting $x_i = \text{number of nests counted in plot } i$ using the rapid method, and $y_i = \text{number of nests counted in plot } i$ using the intensive method. (use $n^{(1)} = 201$). Estimate the total number of nests using the two-phase ratio estimator. (5 marks)

QUESTION FIVE (20 MARKS)

- a) John apply capture-recapture methods to estimate the number of Blackjack weed plants in a tract in Karatina. In some years, Blackjack weed plants do not produce aboveground parts; in addition, if they are in dense vegetation and are not flowering, they are difficult to observe. Thus, a census of observed plants in any given year is likely an undercount. From the first two years of observation, 15 plants were observed in year 1 but not in year 2, 12 plants were observed in year 2 but not in year 1, and 33 plants were observed in

both years. Estimate the total number of plants in the tract, along with a 95% CI. (5 marks)

- b) A study relied with KWS to collect fecal samples from lions. Each sample was genotyped, and the number of distinct individuals was found in each of 2001 and 2002. In 2001, 311 unique genotypes were obtained (134 males and 177 females). In 2002, the procedure was repeated and 239 unique genotypes were obtained (106 males and 133 females). 165 of the individuals sampled in 2001 were also sampled in 2002.
- i) Fifty-six lions in the area in 2001 had also been followed with radio transmitters; 36 of these lions were represented in the 311 genotypes from the 2001 feces samples. Estimate the number of bears in 2001, along with a 95% CI. (5 marks)
 - ii) In 2002, 57 lions had radio transmitters, and 28 of them were among the 239 genotypes from the 2002 feces samples. Estimate the number of lions in 2002, along with a 95% CI. (5 marks)
 - iii) Estimate the number of lions, along with a 95% CI, treating the samples from 2001 and 2002 as independent SRSs (and ignore the radio transmitter data). What assumptions are needed to use capture-recapture estimators of population size? (5 marks)