MACHAKOS UNIVERSITY

University Examinations 2018/2019

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCE

FOURTH YEAR SECOND SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (STATISTICS AND PROGRAMMING)

SST 408: DESIGN AND ANALYSIS OF SAMPLE SURVEY II

DATE: 8/5/2019 **TIME: 2:00 – 4:00 PM**

INSTRUCTION:

Attempt Question ONE and any other TWO questions.

QUESTION ONE (COMPULSORY) (30 MARKS)

a) Differentiate between multi-stage and multi-phase sampling. (2 marks)

b) Discuss a multi-phase design using a well labeled flowchart. (4 marks)

c) Suppose that a Simple Random Sample of 500 fish is caught from a lake; each is marked and released and the fish are allowed to remix with the other fish in the lake. A second sample of 300 fish has 120 marked fish. Estimate the total number of fish in the lake, along with a 95% CI. (4 marks)

d) For a two-phase sampling with ratio estimation, suppose the phase I sample is a Simple Random Sample of size $n^{(1)}$, and the phase II sample is an Simple Random Sample of fixed size $n^{(2)}$. Show that $P(Z_i = 1) = n^{(1)} \big/ N$, and $P(D_i = \big| Z) = Z_i n^{(2)} / n^{(1)}$.

(2 marks)

e) What design would you consider for sampling each of the following rare populations?

i) Alumni of your university who are currently working as engineers.

(2 marks)

ii) Households with children aged 18-36 months. (2 marks)

f) An application allows users to prohibit telemarketers from calling the registered numbers. A telephone survey conducted by customer care reported that. Respondents who had signed up for the registry were asked about their satisfaction with the application; 51% said they are still getting calls they thought the application was supposed to block. Discuss possible sources of measurement error in this survey.

(4 marks)

g) A survey was conducted on mothers, investigating infant feeding in a certain country. Questionnaires asking mothers about type of milk fed to the infant during each of the first six months and about socio economic variables were mailed to a sample of mothers of six-month-old infants. The number of questionnaires mailed increased from 1999 to 2004: In 1999, 56,894 questionnaires were mailed and 30,694 were returned. In 2004, 196,000 questionnaires were mailed and 89,640 were returned. Low-income families were oversampled in the survey design because they had the lowest response rate. Respondents were divided into subclasses defined by region, ethnic background, age and education; weights were computed using information from another official survey.

i) What are the advantages and drawbacks of oversampling the low-income families in this survey? What implicit model is adopted for nonresponse? (3 marks)

ii) Suppose the weighted counts estimated that about 53% of mothers had one child, while the official survey date indicated that about 43% of mothers had one child. Does the agreement of weighted counts with official statistics indicate that the weighting corrects the nonresponse bias? Explain.

(2 marks)

iii) Discuss the use of weighting in this survey. Can you think of any improvements?

(1 mark)

h) Discuss the two types of non-response. (2 marks)

i) Justify any two ways to deal with non-response under imputation. (2 marks)

QUESTION TWO (20 MARKS)

a) An investigator takes a two-phase sample to estimate the prevalence of diabetes in a population. In phase I, a Simple Random Sample of size 1000 is taken from the population of size 100,000, and each individual is asked demographic information and whether he or she has diabetes. It is known that some demographic groups are more at risk for diabetes than others; in addition, the self-report of diabetes may be inaccurate. Therefore, each individual in the phase II sample is given a medical exam to determine diabetes status. The phase I sample is divided into 4 strata. Stratum h has n_h observations in the phase I sample and m_h observations in the phase II sample. After the medical exam, r_h persons in stratum h of the phase II sample were determined to have diabetes.

Stratum	n_h	m_h	r_h
High risk group and reports diabetes	241	96	86
High risk group and does not report diabetes	113	45	17
Low risk group and reports diabetes	174	35	29
Low risk group and does not report diabetes	472	47	8

Estimate the total number of persons with diabetes in the population, along with its standard error. (5 marks)

b) A study 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)

c) Suppose you volunteered to be in an online panel for a survey. What are the possible sources of error in the survey (5 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)

QUESTION THREE (20 MARKS)

A two-phase survey to estimate the prevalence of psychiatric disorders of War veterans was conducted. A random sample of N=53,462 War veterans was administered with a 36-question survey on health. Respondents who scored below 72.2 on physical functioning subscale were defined as disabled (stratum 1, with $n_1 = 406$); respondents who scored above 72.2 were defined as not disabled (stratum 2, with $n_2 = 3047$). A random subsample of 111 veterans was taken from stratum 1, and a random subsample of 98 veterans was taken from stratum 2. The 209 veterans in the phase II sample were evaluated by psychiatrists; the counts in the table below give the number of veterans who are determined to have any alcohol related disorder (Alcohol), any sleep disorder (Sleep), or any psychiatric related disorder (Psych). A veteran can be in more than one of these categories.

Stratum	n_h	m_h	Alcohol	Sleep	Psych
Disabled	406	111	8	20	27
Not disabled	3047	98	10	17	12

a) Estimate the total number of War veterans with an alcohol related disorder, and give a 95% CI. (7 marks)

b) Estimate the total number of War veterans with a sleep related disorder, and give a 95% CI. (7 marks)

c) Estimate the total number of War veterans with any psychiatric disorder, and give a 95% CI. (6 marks)

QUESTION FOUR (20 MARKS)

a) Suppose the phase I sample is a Simple Random Sample of size $n^{(1)}$, and the phase II subsample is a Simple Random Sample of size $n^{(2)}$, with $n^{(2)} < n^{(1)}$. Show that

$$V(\hat{t}_y^{(2)}) = N^2\left(1-\frac{n^{(2)}}{N}\right)\frac{S_y{}^2}{n^{(2)}}$$ (7 marks)

b) Investigators in Machakos University used capture-recapture to estimate the number of fishers in Lake Victoria

i) In the first study, 7 fishers were captured between August 11, 1999 and January 31, 2000. Twelve fishers were captured between February 1 and February 19, 2000; of those 12, 4 had also been captured in the first sample. Give an estimate of the total number of fishers in the area, along with a 95% CI. (5 marks)

ii) In the second study, 16 fishers were captured between September 28, 2000 and October 31, 2000, and 19 fishers were captured between November 1 and November 17, 2000. Eleven of the 19 fishers in the second had also been caught in the first sample. Give an estimate of the total number of fishers in the area, along with a 95% CI. (5 marks)

iii) What assumptions are you making to calculate these estimates? What do these assumptions mean in terms of fishers behavior and 'catchability'? (3 marks)

QUESTION FIVE (20 MARKS)

Investigators selected a Simple Random Sample of 200 high school seniors from a population of 2000 for a survey of television-viewing habits, with an overall response rate of 75%. By checking school records, they were able to find the grade point average for the non-respondents, and classify the sample accordingly:

GPA	Sample Size	Number of Respondents	Hours of TV	
			$\bar{y}$	S_y
3.00-4.00	75	66	32	15
2.00-2.90	72	58	41	19
Below 2.00	53	26	54	25
Total	**200**	**150**		

i) What is the estimate for the average number of hours of TV watched per week if only respondents are analyzed? What is the standard error of the estimate? (5 marks)

ii) Perform a χ^2 test for the null hypothesis that the three GPA groups have the same response rates. What do you conclude? What do your results say about the type of missing data.
(2 marks)

iii) Perform a one-way ANOVA analysis to test the null hypothesis that the three GPA groups have the same mean level of television viewing. What do you conclude? Does your ANOVA analysis indicate that GPA would be a good variable for constructing weighting cells? Why, or why not? (2 marks)

iv) 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? (5 marks)

v) The population counts are 700 students with GPA between 3 and 4; 800 students with GPA between 2 and 3; and 500 students with GPA less than 2. Use these population counts to construct a post stratified estimate of the mean viewing time. (4 marks)

vi) What other methods might you use to adjust for the non-response? (1 mark)

vii) What other variable might be collected that could be used in nonresponse models? (1 mark)