UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination 2001

STATISTICS

ST53002A (ST311) Sampling Techniques

Duration: 2 hours 15 minutes

Date and time:

Answer FOUR questions, which carry 20 marks each. Full marks will be obtained by correct answers to FOUR questions.

Electronic calculators may be used. The make and model should be specified on the script. The calculator must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

NOTE: Full details of all calculations are to be shown; pre-programmed statistical tests and procedures on a calculator, apart from mean and standard deviation, must not be used.

WHITE, YEATS & SKIPWORTH: Tables for Statisticians to be provided.

THIS EXAMINATION PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Question 1 A multi-national company commissions a survey to find out how much its employees spend on food whilst at work e.g. at lunchtimes. Write an essay of not more than 400 words describing the principal steps in planning such a survey. [20]

Question 2 The UK national lottery selects six integers at random without replacement from 1 to 49 inclusive. Players buy tickets in advance and on each ticket select six of these integers of their choice. A prize is won if three or more integers match with the lottery draw, and the jackpot is won if all six integers match.

(a) What is the probability that a player buying a single ticket wins the jackpot? [3]

(b) Show that the chance of winning any prize on a single ticket is approximately 1 in 53.655. [6]

(c) By using the general addition law for two events, find the probability of winning any prize with two non-overlapping tickets. [7]

(d) Hence find the probability of winning any prize with eight non-overlapping tickets. [4]

Question 3 A simple random sample of $n = 6$ students were drawn from a register of $N = 282$, and each sampled student was asked to state their expenditure on travel to College (£) in the previous seven days. The results were in order:

$$0,\ 8.8,\ 10,\ 10.6,\ 15,\ 20$$

(a) Estimate the average weekly expenditure on travel to College for this population and give an approximate 90% confidence interval using the Student's t-distribution. [4]

(b) Estimate the probability that your estimator lies within 20% of the true value. [4]

(c) It was discovered on sampling that a student lived on the campus and hence had an expenditure of zero. Estimate and give a nominal 95% interval for the domain mean of interest, that is the average expenditure on travel for students who live off campus (you will need to use the conditional distribution given that there are 5 such people off the campus in the sample and an unknown number in the population). Is your variance estimator (conditionally) unbiased? [8]

(d) Estimate also the population total expenditure, giving an interval based on the unconditional distribution. Is your variance estimator unbiased? [4]

Question 4 (a) Show that for stratified random sampling of sample size 4 with allocation 2 to each of two strata of equal size 3 (i.e. total population size is 6),

$$\bar{y}_{\text{st}} = (\bar{y}_1 + \bar{y}_2)/2, \quad \text{Var}[\bar{y}_{\text{st}}] = (S_1^2 + S_2^2)/24, \quad v[\bar{y}_{\text{st}}] = (s_1^2 + s_2^2)/24,$$

where $s_h^2 = (y_{h1} - y_{h2})^2/2$, $h = 1, 2$ are the sample variances. [5]

(b) Hence give the full sampling distribution (based on 9 equally likely samples) of $\bar{y}_{\text{st}}$ and its variance estimator for a population in which the values are 0, 5, 10 in the first stratum and 0, 8, 12 in the second stratum. Calculate the coverage of a nominal 95% confidence interval for $\bar{Y}$ based on the Normal distribution. [15]

Question 5 (a) Give the formulae for the variance of the Horvitz-Thompson estimator

$$\hat{T}_{\text{HT}} = \sum_{i \in s} \frac{Y_i}{\pi_i}$$

of the population total T under an unequal probability sampling design with single inclusion probabilities π_i, $i = 1, \ldots, N$ and joint inclusion probabilities π_{ij}, $i, j = 1, \ldots, N$, $(i \neq j)$ if

(i) the sample size is not necessarily fixed

(ii) the sample size is fixed at n.

[6]

(b) The estimator $\hat{N} = \sum_{i \in s} 1/\pi_i$ is proposed when the population size N is unknown.

(i) Show that it is unbiased for N, provided $\pi_i > 0$, $i = 1, \ldots, N$. [2]

(ii) Write down the variance of this estimator when the sample size is fixed, showing that it is zero for an equal probability (*epsem*) design. Why is this result not useful? [6]

(iii) Name one other fixed sample size design for which this variance is zero. [2]

(iv) Show that the variance of $\hat{N}$ for a general design reduces to

$$\text{Var}[\hat{N}] = \sum_{i=1}^{N} 1/\pi_i + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \phi_{ij} - N^2,$$

where $\phi_{ij} = \pi_{ij}/(\pi_i \pi_j)$, $i, j = 1, \ldots, N$, $i \neq j$. [4]

Question 6 (a) State the two recursive formulae for the second order joint inclusion probabilities under Chao's scheme, which proceeds through stages $k = n, \ldots, N-1$, where n and N are the sample and population sizes respectively. [6]

(b) Evaluate Chao's design for a sample size two from a population with size measures 5,4,3,2,1 (in that order), checking that the joint inclusion probability of the two smallest units is 2/105. [14]