UNIVERSITY OF LONDON

GOLDSMITHS' COLLEGE

B. Sc. Examination 2003

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) Define the *sampling distribution* of an estimator e of a population parameter θ. Write down, using the expectation operator E[ ] with respect to this distribution, (i) the variance (ii) the mean square error (MSE) (iii) the bias of this estimator. What is the relationship between these last three quantities? [7]

(b) If the population is $Y_1 = 2$, $Y_2 = 5$, $Y_3 = 7$, $Y_4 = 3$, $Y_5 = 1$, $Y_6 = 6$, $Y_7 = 8$, $Y_8 = 4$, and a simple random sample is drawn consisting of the units labelled 4 and 7, calculate (i) $\bar{y}$ (ii) $\bar{Y}$ (iii) S^2 (iv) s^2. [7]

(c) Hence find (i) Var$[\bar{y}]$ (ii) $v[\bar{y}]$ (iii) whether the nominal 95% confidence interval for $\bar{Y}$, based on a normal approximation to the sampling distribution of $\bar{y}$, covers the true value. [6]

Question 2 A simple random sample of size n is to be drawn from a population size N. A subclass or domain C of the population has size A, mean $\bar{Y}_C$, and variance S_C^2.

(a) State but do not prove a theorem which gives the mean and variance of the sampling distribution of the sample mean $\bar{y}_C$, conditional on exactly a ($a > 0$) units from the domain being selected. [3]

(b) Hence give the mean and variance of a suitable estimator of the subclass *total* when A is known. [3]

(c) State the estimator $\hat{A}$ of A, which is unbiased under the unconditional distribution, and hence state the estimator $\hat{T}_C$, say, of the subclass total when A is not known. [3]

(d) Show that the bias of $\hat{T}_C$ is $(\hat{A} - A)\bar{Y}_C$ under the *conditional* distribution, and hence find the conditional mean square error of $\hat{T}_C$. [5]

(e) Hence show that this estimator is always worse (in the sense of conditional mean square error) than the estimator in (b) when $A < \hat{A}$, and worse when $A > \hat{A}$ provided that the coefficient of variation for C satisfies the inequality

$$\frac{S_C^2}{\bar{Y}_C^2} < \frac{a}{(1 - a/A)} \frac{(A - \hat{A})}{(A + \hat{A})}. \qquad [6]$$

Question 3 A finite population of size N is partitioned into H strata of sizes $N_1, \ldots, N_H$. A stratified random sample of size n is to be drawn, with allocation $n_1, \ldots, n_H$ to be determined. The target for estimation is the finite population mean $\bar{Y}$, and the estimator to be used is $\bar{y}_{\text{st}} = \sum_{h=1}^{H} W_h \bar{y}_h$, where $W_h = N_h/N$, $h = 1, \ldots, H$ and $\bar{y}_h$, $h = 1, \ldots, H$ are the sample stratum means. The respective costs of sampling per unit are $c_1, \ldots, c_H$. The stratum variances are $S_1^2, \ldots, S_H^2$.

(a) Show that $\bar{y}_{st}$ is unbiased and

$$\mathrm{Var}[\bar{y}_{\mathrm{st}}] = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2,$$

quoting but not proving any theorems on simple random sampling, and mentioning explicitly which properties of stratified random sampling you use. [8]

(b) Show that in order to minimise $\mathrm{Var}[\bar{y}_{\mathrm{st}}]$ subject to fixed overall cost $C = \sum_{h=1}^{H} c_h n_h$, the optimal allocation is

$$n_h \propto \frac{N_h S_h}{\sqrt{c_h}}, \qquad h = 1, \ldots, H.$$

[9]

(c) Describe the circumstances under which stratified random sampling with optimal allocation performs *worse* than simple random sampling for the same overall sample size. [3]

Question 4 Assuming that the population mean $\bar{X}$ of a covariate X is known, the linear regression estimator

$$\bar{y}_{\mathrm{lr}} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x}),$$

where $\hat{\beta} = s_{XY}/s_X^2$, is proposed as an estimator of the population mean $\bar{Y}$ based on a simple random sample size n, with sample means $\bar{y}$, $\bar{x}$ and sample variances s_Y^2, s_X^2 of variates Y and X respectively, and where s_{XY} is the sample covariance.

(a) Draw a scatter diagram to motivate the use of this estimator. [8]

(b) Show that the exact bias of $\bar{y}_{\mathrm{lr}}$ is $-\mathrm{Cov}[\hat{\beta}, \bar{x}]$. [2]

(c) By obtaining an approximate expression for $\hat{\beta}$ in terms of B, where $B = S_{XY}/S_X^2$ is the population regression coefficient, show that the leading term in the bias is

$$\frac{1}{S_X^2}\left(B\,\mathrm{Cov}[\bar{x}, s_X^2] - \mathrm{Cov}[\bar{x}, s_{XY}]\right),$$

and that the variance (or the mean square error) is given approximately by

$$\mathrm{Var}[\bar{y}_{\mathrm{lr}}] \approx \frac{1-f}{n}(S_Y^2 - B^2 S_X^2). \qquad [7]$$

(d) Hence show that the linear regression estimator is always better than the sample mean for large samples. What can you say about small samples? [3]

Question 5 An estimator e of a population parameter θ is said to be *calibrated with respect to a covariate* X if $e = \theta$ for all samples (with positive probability) whenever $Y_i \propto X_i$, $i = 1, \ldots, N$.

(a) Show that both ratio and regression estimators of the population mean $\bar{Y}$ are calibrated with respect to the covariate X. [4]

(b) Show further that the sample mean is not calibrated with respect to X, but is calibrated with respect to a constant covariate. [4]

(c) Which of the ratio and regression estimators are calibrated with respect to a constant? Justify your answer. [4]

(d) Show that the mean of ratios estimator $\bar{X}\frac{1}{n}\sum_{i=1}^{n} y_i/x_i$ is also calibrated with respect to X when the sample size is fixed, but that it is only unbiased under an unequal probability sampling design if the single inclusion probabilities π_i, $i = 1, \ldots, N$ satisfy $\pi_i \propto X_i > 0$, $i = 1, \ldots, N$, in which case the estimator becomes the Horvitz-Thompson estimator. [6]

(e) Show further that the Hajek-Basu estimator

$$\hat{\bar{Y}}_{HB} = \frac{\sum_{i \in s} Y_i/\pi_i}{\sum_{i \in s} 1/\pi_i}$$

is calibrated with respect to a constant, whether the sample size is fixed or not. [2]

Question 6 (a) Use Waterman's algorithm and the sequence of random digits

81350325872483590467515226216975872861 50

starting from the left as economically as possible, to draw a simple random sample of 2 from a population of size 6, labelling the units $1, 2, \ldots, 6$. What is the probability of your selected sample? [6]

(b) The 6 units have size measures 4,4,4,4,1,1 respectively. Find the initial tableau and replacement probabilities for Chao's scheme, and use the scheme to draw a sample of size 2, again using the random digits starting from the left as economically as possible. What is the probability of your selected sample? [14]