DO NOT ENTER ANY ANSWERS ON THIS YELLOW QUESTION PAPER. THIS YELLOW QUESTION PAPER WILL BE DESTROYED AFTER THE EXAM. ANY ANSWERS WRITTEN ON IT WILL NOT BE MARKED.

UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER 2011 EXAMINATIONS

PLEASE HAND IN

STA304H1 F / STA1003H F

Duration 2 hours

Examination Aids: Two sides handwritten notes (8 1/2 x 11) and one non-programmable calculator.

Answer all questions in examination booklets.

---

1. **(10 marks)** A auditor is confronted with a long list of accounts receivable for a firm. He must verify the amounts on 10% of these accounts and estimate the average difference between the audited and book values.

   (a) Suppose the accounts are arranged chronologically (according to their dates), with the older accounts tending to have smaller values. Would systematic or random sampling be preferred? Explain briefly.

   (b) Suppose the accounts are grouped by department, and then listed chronologically within departments. The older accounts again tend to have smaller values. Would systematic or random sampling be preferred? Explain briefly.

2. **(30 marks)** The consulting service at the Department of Statistics carried out a study in 2005-2006 to estimate the accumulated debt-load of graduating students in statistics and actuarial science. They selected a simple random sample of five 400-level fall and spring half-courses in STA and ACT, from the 25 400-level courses offered. For each of the sampled classes, a graduate student visited the class, and asked all the students present to fill in a form like the one on the next page.

Thank you for agreeing to participate in this survey. We are working for the Consulting Service of the Department of Statistics, on a study of student debt loads. Your answers will be kept confidential; only summary statistics will be used in a report for the Faculty of Arts and Science.

Student Number: (used for sampling verification only) _____

1. What year and month do you expect to graduate?

(Circle the most appropriate choice)

(a) May 2006 (b) November 2006 (c) May 2007 (d) Other _____

2. Do you expect to have any debt upon graduation? _____Yes _____No

3. If so, what do you estimate your accumulated debt to be? $_____

4. What is your current cumulative GPA? _____

(a) What type of sampling design did the Consulting Service use? Explain.

(b) Do you think the sample of students was representative of graduating students? Why or why not?

(c) The results were as follows:

Class	Number of students enrolled	Number of students sampled	sample mean debt load ($1000's)	sample SD debt load ($1000's)
ACT 451F	53	50	11	53
ACT 466S	20	19	13	55
STA 410F	25	10	7	31
STA 414S	55	25	2.5	1.1
STA 422F	15	10	8	16

From this the population mean debt load was estimated to be

$$\begin{aligned}\hat{\mu} &= (53 \times 11 + 20 \times 13 + 25 \times 7 + 55 \times 2.5 + 15 \times 8)/(53 + 20 + 25 + 55 + 15) \\ &= 7.6\$1000's = \$7600,\end{aligned}$$

and its estimated variance was $\widehat{V}(\hat{\mu}) = \$4576$.

i. What type of estimate is $\hat{\mu}$? Is this estimator unbiased for the population mean? Why or why not?

ii. What is the margin of error for the estimate $\hat{\mu}$?

(d) A plot of debt against GPA for the students in the sample showed an approximate linear relationship; and this enabled the investigators to compute an alternate es-

timate of the mean debt, using information on the GPA. What type of estimation did they likely use?

(e) Another survey will be held in 2012-13 to see how much debt load has increased. Suggest two improvements that could be implemented, to make the sampled population more representative of the target population, and give one sentence for each your suggestions, to explain why you think it would be an improvement.

3. **(15 marks)** A survey was undertaken to estimate the total value of the books in a private library. Two-stage cluster sampling was used: the primary sampling unit was shelf, and the secondary sampling unit was books. The survey selected $n = 12$ shelves by simple random sampling from all 44 shelves, and $m_i = 5$ books on each shelf, by simple random sampling, and determined the value of each of the sampled books by comparison to a rare books catalogue. The analysis of variance table below shows the between and within sums of squares for the data.

Source	df	Sum of Squares	Mean Square
between shelves	11	25571.0	2324.6
within shelves	48	23445.2	488.4
total	59	49016.2	

For this survey, is cluster sampling likely to be more efficient or less efficient than simple random sampling with the same number of sampled books? Explain.

4. **(25 marks)** Define any **four** of the following terms, and illustrate each with an example.

   (a) questionnaire bias

   (b) prospective study

   (c) sampling frame

   (d) non-response bias

   (e) respondent-driven sampling

   (f) probability proportional to size

   (g) case-control study

   (h) post-stratification

   (i) regression estimation

Answer ONE of the following two questions. Clearly indicate which question you have chosen: the other question will not be marked. The questions are of equal value.

5. **(20 marks)** Suppose we have L strata defined in a population, with $N_1, \ldots, N_L$ elements of the population in each of the L strata. We take a simple random sample of n_ℓ observations in each stratum.

(a) Show that

$$\hat{\tau}_{st} = \sum_{\ell=1}^{L} N_\ell \bar{y}_\ell$$

is an unbiased estimator of the population total τ, where

$$\bar{y}_\ell = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} y_{\ell j}$$

is the sample mean in each stratum.

(b) Suppose we also measure an auxiliary variable, with values x_{ij} in each stratum, and estimate the population total using the ratio estimator

$$\hat{\tau}_{rat} = \sum_{ell=1}^{L} N_\ell \frac{\bar{y}_\ell}{\bar{x}_\ell} \tau_{x,\ell},$$

where $\tau_{x,\ell}$ is the population total for the auxiliary variable in the ℓth stratum. Is this estimator unbiased? Explain briefly, without actually calculating $E(\hat{\tau}_{rat})$.

6. **(20 marks)** A recent study published in Social Science Quarterly by Zagorsky and Smith[1] concluded that the widely quoted 'fifteen pound' weight gain in freshman year was a media myth. From the information in the following excerpts from their paper, answer the questions that follow these excerpts.

> This research investigates freshman weight gain using data from the National Longitudinal Survey of Youth 1997 cohort (NLSY97) by employing descriptive statistics, linear regressions, simulations, and longitudinal analysis ...
> The NLSY97 is a nationally representative panel survey of nearly 9,000 people living in the United States in 1997 and born between 1980 and 1984 ...
> All descriptive tables and graphs use data adjusted for the sampling structure...
> Freshmen women gained slightly more than three pounds and men gained three and a half pounds, on average...
> The NLSY97 data show that between ages 18 and 19, women who attended college gained 3.55 pounds (mean) while women who never attended gained 3.54 pounds, a statistically insignificant difference...
> Table 4 shows how freshman weight gain varies across five factors: full-time versus part-time status; two-year versus four-year degree; private versus public institution; lived in a dormitory or elsewhere; and heavy drinking status (consuming six or more drinks on at least four days per month).

(a) How might the Bureau of Labor Statistics ensure that the NLSY97 is 'nationally representative'?

[1]Zagorsky, J.L. & Smith, P.K. (2011). The Freshman 15: a critical time for intervention or a media myth? *Soc. Sci. Qu.* **92**, 1389–1407

(b) What does it mean to say the descriptive tables use **data adjusted for the sampling structure**?

(c) Why did the authors study how weight gain varies across five factors? Which of these factors would you expect to have the largest potential influence on weight gain?

(d) What does the "statistically insignificant difference" tell us about the issue of interest?