90

Last name: Li First name: Yanfei Lisa Student #: 996051392

STA 304H1 F/1003H F SUMMER 2010, First Test, May 27 (20%)
Duration: 60 min. Allowed: hand-calculator, aid-sheet, one side, with theoretical formulas and definitions only.

23

[32] 1) The faculty senate at the University of Toronto wanted to know what proportion of students thought a foreign language should be required for everyone. With a help of the Department of Statistics a simple random sample of 500 students was selected from all students enrolled in statistical courses. A survey form was sent by e-mail to these 500 students. Answer in short the following questions:
(a) What is the population of interest to the faculty senate?
(b) What is the sampling frame?
(c) Describe the variable of interest. What type of the variable is the variable of interest?
(d) Discuss in short an extend to which each of the three types of bias would be likely to occur in this survey: (i) inadequate frame, (ii) selection bias, (iii) nonresponse bias. Which of these three types of bias do you think would be the most serious in this study? Explain.
(e) Do you expect that the obtained estimate would overestimate, or underestimate the parameter of interest? Explain.

(a). All current University of Toronto students
(b). The list of all students enrolled in statistical courses.
(c). The variable of interest whether the student thinks a foreign language should be required for everyone.

$$y(e_i)=\begin{cases}1 & \text{if yes}\\ 0 & \text{if no}\end{cases}$$ qualitative, categorical variable.

-7 (d). Inadequate frame is the ~~major~~ problem because the sampling population (students enrolled in stats courses) is far -1 less than the target population (all students in U of T), thus it's biased. ~~Selection bias~~ selection bias is also a problem since students -1 who like statistics tend to have a quantitative mind, not representative for all students. -1 Nonresponse bias is also a minor problem (Answer this question is easy).

-2 (e). Underestimate the parameter of interest. Because statisticians are ~~quant peop~~ less likely to learn a foreign language than students in humanities or linguistics.

36

[37] **2)** Distribution of family sizes in a recent census in Toronto was as follows:

Number of persons	1	2	3	4	5	6	7	Total
Proportion of families, %	25	32	17	16	7	2	1	100%

(a) What is the population in this question? What is the variable? What other variables might be of interest in this population? Name a few.
(b) Calculate the population mean and standard deviation from the distribution in (a).

Using a list of households, a random sample of 400 families from the population was selected and the following result was obtained:

Number of persons	1	2	3	4	5	6	Total
Number of families	105	140	75	62	14	4	400

(c) Estimate the population mean and calculate the exact error of estimation (use a result from (b)).
$Var(\bar{y}) =$
(d) Calculate a bound on the error of estimation in (c) using a result from (b) (don't use the sample). Does the error of estimation from (c) indicate something about the sample and the census? Discuss.
(e) If you were to estimate the average family size in Toronto with a bound on the error of estimation of at most 0.10, what would you suggest as the sample size? You have no other information except that the family size is at most 7 (a few exceptions may be ignored), and, obviously, at least 1.
(f) Can you estimate the total number of people living in Toronto using only the sample given above? Why, or why not? Make a reasonable guess about the missing information and then estimate that number.

(a). The population is all families in Toronto. The variable is the number of persons in a family. Other variables may include: salary per year for the family, number of family members who has college degree, whether the family has child under 18 ...

(b) $\mu = \frac{1}{100}(1\times25+2\times32+\cdots+7\times1) = 2.58.$

$\sigma = \sqrt{\frac{1}{100}(1^2\times25+2^2\times32+\cdots+7^2\times1)-2.58^2} = 1.387.$

(c). $\bar{y} = \frac{1}{400}(1\times105+\cdots+6\times4) = 2.38.$ $|\bar{y}-\mu| = |2.38-2.58| = 0.2$

(d) $Var(\bar{y}) = \frac{N-n}{N-1}\frac{\sigma_y^2}{n} \approx \frac{\sigma_y^2}{n} = \frac{1.9236}{400} = 0.004809.$ since N is large

$B_\mu = 2\sqrt{Var(\bar{y})} = 2\sqrt{0.004809} = 0.1387$

Since the error of estimation 0.2 is larger than B_μ, it indicates that the sample may be biased and it does not well represent the population.

(e) $B_u \le 0.1$. N is large so $n \approx \frac{\sigma_y^2}{D} = \frac{1.9236}{0.0025} = 769.44$

-1 $D = \left(\frac{0.1}{2}\right)^2 = 0.0025$

The sample size will be 770

(f) No, because the total number of families is not known.

Assume ~~the~~ there are 40,000 families in GTA.

Then $\hat{\tau} = \bar{y}N = 2.38 \times 40{,}000 = 95200$

31

[31] 3) There are 850 students in an introductory statistical course at U of T, and their names are listed in a file with identification numbers 1, 2, ... , 850. At the beginning of the course, the instructor wants to conduct a short test to estimate how the class is prepared for university.

(a) Use the table of random numbers given below to select an SRS (without replacement) of 5 students from the first 70 students in the file. Select the sample and explain your method.

(b) Select also another SRS of 5 students, using idents from 71 to 500. Select the sample and explain your method.

(c) If you combine these two samples in one sample of size 10, would it be an SRS from the population? Explain why, or why not.

(d) Even if the sample in (a) may not be, strictly speaking, an SRS from the population, can it be, in a way, representative of the population (ignore small sample size)? You may consider some additional assumptions about the file. Explain.

(e) If an SRS of size 50 was selected, and found that 40 students didn't have any previous statistical background, estimate the total number of students in the class without any previous background in statistics, and calculate the confidence interval for the estimate.

Table of random numbers:

92325 19474 23632 27889 47914 02584 37680 20801 72152 39339 34806 08903 25570
31624 76384 17403 53363 44167 64486 64758 75366 76554 31601 12614 33072 60332
01624 76384 97403 53363 44167 64486 64758 75366 76554 31601 12614 33072 19474
23632 27889 47914 02584 37680 20801 72152 39339 34806 08930 25570 33120 45732

(a). Use 2 digits and the 1st line. Assign random digits 01, 02, ⋯, 70 to students 1, 2, ⋯, 70 and ignore digits 71, ⋯, 99.

digits	92	32	51	94	74	23	63	22
Students ID	X	32	51	X	X	23	63	22

The sample is 22, 23, 32, 51, 63.

(b). Use 3 digits and the 2nd line. Assign random digits 071, 072, ⋯, 500 to students 71, 72, ⋯, 500. Ignore digits 001, 002, ⋯, 070 and 501 ~ 999.

digits	316	247	638	417	403	533	634	416
Student ID	316	247	X	417	403	X	X	416

The sample is 247, 316, 403, 416, 417.

(c). No. Because ~~there might be ones~~ it only has 500 students, far less than the population 850, and combining 2 SRS from different source is not SRS at all.

(d). It may be representative of the population if the listing order of the students has nothing to do with the students' academic behavior, for example the list is ~~in alpha~~ by last name. It's a biased sample and not representative of the population if the students are ordered by their ~~exam~~ academic records in statistics.

(e) $\hat{p} = \frac{40}{50} = 0.8$.

$\hat{\tau} = N\hat{p} = 850 \times 0.8 = 680$

$\widehat{Var}(\hat{p}) = \frac{N-n}{N}\frac{\hat{p}\hat{q}}{n-1} = \frac{850-50}{850} \times \frac{0.8\times0.2}{49} = 0.003073$

$B_\tau = 2N\sqrt{\widehat{Var}(\hat{p})} = 2\times850\times\sqrt{0.003073} = 94.24$

95% CI for τ is $\hat{\tau} \pm B_\tau = 680 \pm 94.24$

$= [585.76, 774.24]$

/

92

Last name, first name: Li Yanfei Lisa Student #: 996051392

STA 304H1 F SUMMER 2010, Second Term-test, June 10 (20%)
Duration: 1h. Allowed: hand-calculator, aid-sheet, one side, with theoretical formulas and definitions only

[60] 1) In the National Health Survey a community of 850 households was selected at the first stage. An SRS of 40 families was selected from the community at the second stage. The following table gives a summary of the results on the family size (x_1), weekly net family income (x_2), and weekly cost of medical expenditures (y) in the sample from the community.

$\sum x_1$	$\sum x_2$	$\sum y$	$\sum x_1^2$	$\sum x_2^2$	$\sum y^2$	$\sum x_1 y$	$\sum x_2 y$
150	29,500	3,540	650	22,500,000	330,000	14,250	2,677,640

$N=850$
$n=40$.

(a) [20] Estimate: (i) the total number of persons in the community, (ii) the average weekly net family income, (iii) the average weekly medical expenses per family and (iv) the average weekly medical expenses per person.

(b) [15] Estimate and give a 95% CI for the proportion of family income spent on medical expenses (be careful what is the proportion here).(continued)

(a). $\hat{\tau}_{x_1} = N\hat{\mu}_{x_1} = N\bar{x}_1 = 850 \times \frac{150}{40} = 3187.5$.

16

$\hat{\mu}_{x_2} = \bar{x}_2 = \frac{29500}{40} = 737.5$.

$\hat{\mu}_y = \bar{y} = \frac{3540}{40} = 88.5$.

$\hat{R} = \frac{\sum y}{\sum x_1} = \frac{3540}{150} = 23.6$.

(b). $r = \frac{\sum y}{\sum x_2} = \frac{3540}{29500} = 0.12 = 12\%$.

15

$\hat{Var}(r) = \frac{N-n}{N} \frac{1}{\bar{x}_2^2} \frac{s_r^2}{n}$. $s_r^2 = \frac{1}{39}[330000 - 0.24 \times 2677640 + 0.12^2 \times 22500000]$

$= \frac{850-40}{850} \times \frac{1}{737.5^2} \times \frac{291.446}{40}$ $= 291.446$

$= 0.000012765$.

$B_r = 2\sqrt{\hat{Var}(r)} = 0.007146 = 0.7146\%$.

95% CI : $12\% \pm 0.7146\% = [11.2854\%, 12.7146\%]$

(c) [15] If the total number of persons in the population is known to be 3000, estimate the total weekly medical expenses in the community. Use an estimator you consider is the best one in this situation. Explain your choice. SRS. ratio. regression. difference

(d) [10] Select the sample size (number of families) necessary to estimate the percentage of the family income spent on medical expenses with a bound on the error of estimation of 1%. Should this sample size be less than 40, or greater than 40?

(c) $\tau_{x_1} = 3000$

$\hat{\tau}_y = \hat{R}\tau_{x_1} = \frac{\sum y}{\sum x_1}\tau_{x_1} = \frac{3540}{150} \times 3000 = 70800$. This is ratio estimator.

7

~~Since~~ Theoretically regression estimator is the best estimator with the smallest variance. But in this case, total weekly medical expenses is proportional to total # of persons, ratio estimator and regression estimator are the same. And ratio estimator is easy to calculate.

(d) $D = \left(\frac{B_r \hat{\mu}_{x_2}}{2}\right)^2 = \left(\frac{0.01 \times 737.5}{2}\right)^2 = 13.598.$

9

$n = \frac{N\hat{\sigma}_r^2}{ND + \hat{\sigma}_r^2} = \frac{850 \times 291.446}{850 \times 13.598 + 291.446} = \frac{247729.1}{11849.746} = 20.9 = 21$

This sample size is smaller than 40, because from part (b) we know the error is 0.7146% when sample size is 40. If error bound is larger, sample size required will be smaller.

92

stra.

[50] 2) A course coordinator wishes to investigate the points lost by students due to grammatical errors, in a language course. Three tests were done by $N_1 = 120$, $N_2 = 100$, and $N_3 = 70$ students per test respectively (first test was held at the beginning of the course, second test before the drop-day (mid-term), and third test before the end of the course). A random sample of test papers from every test was selected and the following results for points lost were obtained:

	test	Points lost		Sample mean	Sample var.
$N_1=120$	I	12 16 0 6 2 2	6	6.333	40.667
100	II	4 15 9 0 8 4 4	7	6.286	23.571
70	III	6 8 0 10 7 5 0	7	5.143	14.810

[8] (a) Explain what is the population used in this example. What is the population size? Explain what is the sample design used here.
[8] (b) Assuming that test marks were out of 50 points, estimate the **percentage** of points lost due to grammatical errors, in the course.
[12] (c) Estimate the percentage of all test papers with some points lost due to grammatical errors, and place a bound on the error of that estimation. **(continued)**

(a). The population is the test paper submitted by students in all three tests. Population size $N = N_1+N_2+N_3 = 120+100+70 = 290$.
7
This sample used stratified sampling, and the population is stratified by three different ~~test~~. There are 3 strata, and select SRS from each of them.

(b). $\bar{y}_{str} = \sum \frac{N_i}{N}\bar{y}_i = \frac{120}{290} \times 6.333 + \frac{100}{290} \times 6.286 + \frac{70}{290} \times 5.143$
$= 2.621 + 2.168 + 1.241 = 6.03$.
7
$\% = \frac{\bar{y}_{str}}{50} = \frac{6.03}{50} = 12.06\%$ percentage of points lost due to grammatical errors.

(c). $\hat{p}_{str} = \sum \frac{N_i}{N}\hat{p}_i = \frac{120}{290} \times \frac{5}{6} + \frac{100}{290} \times \frac{6}{7} + \frac{70}{290} \times \frac{5}{7} = 0.8128 = 81.28\%$.
11
$\hat{Var}(\hat{p}_{str}) = \left(\frac{120}{290}\right)^2 \times \frac{120-6}{120} \times \frac{\frac{5}{6}\times\frac{1}{6}}{5} + \left(\frac{100}{290}\right)^2 \times \frac{100-7}{100} \times \frac{\frac{1}{7}\times\frac{6}{7}}{6} + \left(\frac{70}{290}\right)^2 \times \frac{70-7}{70} \times \frac{\frac{5}{7}\times\frac{2}{7}}{6}$
$= 0.00451843 + 0.002256788 + 0.00178359 = 0.008558808$

$B_p = 2\sqrt{\hat{Var}(\hat{p}_{str})} = 0.18503 = 18.503\%$

[10] (d) If the percentage in (b) has to be estimated with the bound on the error of 2%, and using proportional allocation, what would be the appropriate total sample size? Use the given sample as a presample. What would be the allocation?

[8] (e) Would you consider using simple random sampling instead of stratified sampling with (i) proportional, (ii) optimal allocation, in this problem? Explain.

[4] (f) Can you estimate the number of points lost per student in the course from the data? Why or why not?

(d) $B_Z = 2\sqrt{Var\left(\frac{\bar{y}_{str}}{50}\right)} = \frac{1}{25}\sqrt{Var(\bar{y}_{str})}$

9

$D = Var(\bar{y}_{str}) = (25B_Z)^2 = (25\times 0.02)^2 = 0.25.$

$n = \dfrac{\sum W_i \sigma_i^2}{D + \frac{1}{N}\sum W_i\sigma_i^2}$. $\hat{\sigma}_i^2 = s_i^2.$

$\sum W_i \hat{\sigma}_i^2 = \frac{120}{290}\times 40.667 + \frac{100}{290}\times 23.571 + \frac{70}{290}\times 14.810 = 28.53.$

$n = \dfrac{28.53}{0.25 + \frac{1}{290}\times 28.53} = 81.89 = 82.$

$w_1 = W_1 = \frac{120}{290} = 41.38\%.$ $w_2 = W_2 = \frac{100}{290} = 34.48\%.$

$w_3 = W_3 = \frac{70}{290} = 24.14\%.$

$n_1 = 82\times 41.38\% = 34.$ $n_2 = 82\times 34.48\% = 28.$

$n_3 = 82\times 24.14\% = 20.$

(e) Since the sample means of three strata are very close,

8

$Var(\bar{y}_{SRS}) \approx Var(\bar{y}_{str,prop})$. Both SRS and stratified sampling using proprotional allocation are good.

However the variances of the three strata are quite difference, $Var(\bar{y}_{SRS}) \approx Var(\bar{y}_{str,prop}) \geq Var(\bar{y}_{str,opt})$. I would consider using stratified sampling with optimal allocation.

(f) No. We don't know the total number of students in the course and the numbers of students attending each test is different.

3

4