Last name, first name: ______________________________. Student #: ________________

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

[55] 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

(a) [16] 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)

Solution:

[16] (a) (i) $\hat{\tau}_1 = N\bar{x}_1$= 850x150/40 = 3187.5, $\bar{x}_1$= 3.75, [4]

(ii) $\hat{\mu}_{x_2} = \bar{x}_2$= 29500/40 = 737.5 [4]

(iii) $\hat{\mu}_y = \bar{y}$= 3540/40 =88.5 [4]

(iv) $\hat{R}_{y/x_1} = r_1 = \bar{y}/\bar{x}_1$= 3540/150 = 0.12 = 23.6 [4]

[15] (b) $\hat{R}_{y/x_2} = r_2 = \bar{y}/\bar{x}_2$= 3540/29500 = 0.12 = 12% [5]

$$\hat{V}ar(r_2) = \frac{N-n}{N}\frac{1}{n\bar{x}_2}S_{r_2}^2 = \frac{850-40}{850}\frac{1}{40(737.5)^2}291.446 = 1.27656\text{x}10^{-5}, [5]$$

$$S_{r_2}^2 = \frac{1}{n-1}\Sigma(y_i - r_2 x_{2i})^2 = \frac{1}{n-1}[\Sigma y_i^2 - 2r_2\Sigma x_{2i}y_i + r_2^2\Sigma x_{2i}^2] =$$

= 1/39(330,000 - 2x0.12x2,677,640 + 0.12^2x22,500,000) = 291.446,

$\hat{S}D(r_2)$ =0.003573, B_{r_2} = 2x0.003573 = 0.007146=0.72%,

CI = 12% ± 0. 72% = [11.28%, 12.72%]. [5]

(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.

(d) [9] 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?

Solution:

[15] (c) Regression estimator is the best choice, because the medical expenditures are correlated with family size. [5]

$$\mu_{x_1} = 3000/850 = 3.5294 \approx 3.53,\ \text{b} = \frac{\Sigma y x_1 - n\bar{x}_1\bar{y}}{\Sigma x_1^2 - n\bar{x}_1^2} = \frac{14250 - 40x3.75x88.5}{650 - 40x3.75^2} = 11.14$$

$\hat{\mu}_{yL} = \bar{y} - b(\bar{x}_1 - \mu_{x_1})$ = 88.5-11.14x(3.75 - 3.53) = 86.05, [7]

$\hat{\tau}_y = N\hat{\mu}_{yL}$ = 850x86.05 = 73142.5. (**3**)

[9] (d) $B_r = 1\% = 0.01$, D = $(B_r\mu_{x_2}/2)^2$ = $(0.01\text{x}737.5/2)^2$ = 13.60, [2]

$$\hat{n} = \frac{N\hat{\sigma}_{r_2}^2}{ND + \hat{\sigma}_{r_2}^2} = \frac{850 \times 291.446}{850 \times 13.60 + 291.446} = 20.9 = 21 \text{ [5]}$$

((*N*-1)*D* can also be used)

The sample size should be < 40, because for n = 40, B = 0.72% < 1%. [2]

[45] 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.
I	12 16 0 6 2 2	6.333	40.667
II	4 15 9 0 8 4 4	6.286	23.571
III	6 8 0 10 7 5 0	5.143	14.810

(a) [7] Explain what is the population used in this example. What is the population size? Explain what is the sample design used here.
(b) [7] Assuming that test marks were out of 50 points, estimate the **percentage** of points lost due to grammatical errors, in the course.
(c) [11] 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)**

Solutions:
[7] (a) Population consists of test papers in all three tests (not of students!) **[3]**.
Population size N = 290 **[2]**. Sample design is stratified sampling with tests as strata **[2]**.
[7] (b) Percentage of points lost in the course is the same as the % of points lost per test paper, and is twice the number of points lost per paper (% out of 50 points):
[if it is not clear, look at this: % of points lost =100x(total # of points lost)/(total # of points) = 100x(total # of points lost)/(50x290) =2x(total # of points lost)/290 = 2μ]

$2\hat{\mu} = 2\bar{y}_{st} = 2\Sigma N_i \bar{y}_i / N =$ 2x(120x6.333+100x6.286+70x5.143)/290 = 2x1748.57/290 = 2x6.03 = 12.06% points lost. [if not multiplied by 2, give 4 points]
[11] (c) Find the proportion of tests with some points lost (> 0) in each sample:
$\hat{p} = \Sigma N_i \hat{p}_i / N =$ (120x5/6+100x6/7+70x5/7)/290=0.813=81.3% **[6]**
$\hat{V}ar(\hat{p}) = \Sigma(\frac{N_i}{N})^2 \frac{N_i - n_i}{N_i} \frac{\hat{p}_i \hat{q}_i}{n_i - 1} = (\frac{120}{290})^2 \frac{120-6}{120} \frac{1}{5} \frac{5}{6} \frac{1}{6} + (\frac{100}{290})^2 \frac{100-7}{100} \frac{1}{6} \frac{6}{7} \frac{1}{7} + (\frac{70}{290})^2 \frac{70-7}{70} \frac{1}{6} \frac{5}{7} \frac{2}{7}$ =8.5588x10^{-3},
$B_p = 2 \times \sqrt{8.5588 \times 10^{-3}} = 0.185 = 18.5\%$. **[5]**

(d) [9] 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? (Be careful with D)
(e) [8] Would you consider using simple random sampling instead of stratified sampling with (i) proportional, (ii) optimal allocation, in this problem? Explain.
(f) [3] Can you estimate the number of points lost per student in the course from the data? Why or why not?

Solutions:
[9] (d) $n = N\frac{\Sigma N_i\sigma_i^2}{N^2D+\Sigma N_i\sigma_i^2}, \Sigma N_i\sigma_i^2 =$ 120x40.667+100x23.571+70x14.810=8273.84,

$B = 2\sqrt{Var(2\hat{\mu})} = 2\sqrt{2^2Var(\hat{\mu})} = 4\sqrt{Var(\hat{\mu})}$ so that D = $(B/4)^2 = (2/4)^2 = 0.25$. Or, simpler, 2% error bound on the percentage lost is just 1 point error bound on the points lost, i.e., D = $(1/2)^2 = 0.25$.

$$n = \frac{290 \times 8273.84}{290^2 \times 0.25 + 8273.84} = 81.89$$, or n = 82. **[5]**

Allocation: $n_1 = \frac{120}{290} \times 82 = 34$, $n_2 = \frac{100}{290} \times 82 = 28$, $n_3 = \frac{70}{290} \times 82 = 20$. **[4]**

[8] (e) (i) It seems from the sample that mean values over strata (tests) are similar, so that using SRS instead of stratified sampling with proportional allocation would not increase the error of estimation, that is, SRS may be used **[4]** [if the student assumes that differences between strata means are significant, and SRS should not be used, this answer can be accepted]
(ii) It seems that standard deviations over strata are different, so that stratified sampling with optimal allocation should be better than SRS. **[4]** [it should not be doubt in this case]
[3] (f) You cannot, because the sample is from the population of tests, not population of students. Some students wrote one test only, some two, and some wrote three tests.