Last name, first name: . Student #:

STA 304H1 S WINTER 2013, Second Term-test, March 22 (20%)
Duration: 1h. Allowed: hand-calculator, aid-sheet, one sided or two sided, with theoretical formulas, as posted on Portal. Test has 4 pages.

[45] 1) An SRS of 40 families was selected from a community of 850 households as a preliminary survey for a big study. 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) [15] 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, (iv) the average weekly medical expenses per person, and (v) the average weekly net income per person.
(b) [12] 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:
[15] (a) (i) $\hat{\tau}_1 = N\bar{x}_1 = 850\text{x}150/40 = 3187.5,\ \bar{x}_1 = 3.75,$ [3]
(ii) $\hat{\mu}_{x_2} = \bar{x}_2 = 29500/40 = 737.5$ [3]
(iii) $\hat{\mu}_y = \bar{y} = 3540/40 = 88.5$ [3]
(iv) $\hat{R}_{y/x_1} = r_1 = \bar{y}/\bar{x}_1 = 3540/150 = 23.6$ [3]
(v) $\hat{R}_{x_2/x_1} = r_{2|1} = \bar{x}_2/\bar{x}_1 = 29500/150 = 196.67.$ [3]
[12] (b) $\hat{R}_{y/x_2} = r_2 = \bar{y}/\bar{x}_2 = 3540/29500 = 0.12 = 12\%$ [4]

$$\hat{Var}(r_2) = \frac{N-n}{N}\frac{1}{n(\bar{x}_2)^2}S_{r_2}^2 = \frac{850-40}{850}\frac{1}{40(737.5)^2}291.446 = 1.27656\text{x}10^{-5},$$ [4]

$$S_{r_2}^2 = \frac{1}{n-1}\Sigma(y_i - r_2 x_{2i})^2 = \frac{1}{n-1}[\Sigma y_i^2 - 2r_2\Sigma x_{2i}y_i + r_2^2\Sigma x_{2i}^2] =$$
$= 1/39(330{,}000 - 2\text{x}0.12\text{x}2{,}677{,}640 + 0.12^2\text{x}22{,}500{,}000) = 291.446,$
$\hat{SD}(r_2) = 0.003573,\ B_{r_2} = 2\text{x}0.003573 = 0.007146 = 0.72\%,$
CI = $12\% \pm 0.72\% = [11.28\%, 12.72\%]$. [4]

(c) [10] 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) [5] 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%. (use the information obtained from the given sample) Should this sample size be less than 40, or greater than 40? Explain.
(e) [3] Can you place a bound on the error of estimation in (a) (v)? Just explain.

Solution:

[10] (c) Regression estimator is the best choice, because the medical expenditures are correlated with family size. [3]

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

$\hat{\mu}_{yL} = \bar{y} - b(\bar{x}_1 - \mu_{x_1}) = 88.5\text{-}11.14\text{x}(3.75 - 3.53) = 86.05$, [5]

$\hat{\tau}_y = N\hat{\mu}_{yL} = 850\text{x}86.05 = 73{,}142.5$. [2]

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

$$\hat{n} = \frac{N\hat{\sigma}_{r_2}^2}{ND + \hat{\sigma}_{r_2}^2} = \frac{850 \times 291.446}{850 \times 13.60 + 291.446} = 20.9 = 21$$ [3]

($(N\text{-}1)D$ can also be used)

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

[3] (e) We cannot, because the sum $\sum x_1 x_2$ is not given in the data. We need it to estimate variance of $\hat{R}_{x_2/x_1} = r_{2|1} = \bar{x}_2/\bar{x}_1$.[3]

[55] 2) In order to estimate the total inventory of its products being held this year, a car company conducts a proportional stratified sample of its dealers, with these dealers being stratified according to their inventory held in the previous year. For a total sample size of n = 100, the following data for the current inventory was obtained

Stratum (last year inventory)	Number of dealers	Sample results $\bar{y}_i$
I [50 – 150)	400	105
II [150-250)	1000	180
III [250 – 450)	540	370
IV [450 – 650)	60	590

(a) [5] Find the allocation of the sample used.

(b) [10] From this stratified sample, estimate the mean current inventory μ and the total current inventory.

(c) [14] Can you estimate the variance of the estimator $\hat{\mu}$ used in (b) using the sample? Is there any other information in the above table that might help to estimate that variance? Explain and use it to estimate the variance (you may ignore the finite population corrections). **(continued)**

Solutions:

[5] (a) Proportional allocation: $n_i = \frac{N_i}{N} n$, N = 400+1000+540+60 = 2000, or

n_1 = 100x400/2000 = 20,
n_2 = 100x1000/2000 = 50,
n_3 = 100x540/2000 = 27,
n_4 = 100x60/2000 = 3. **[5]**

[10] (b) $\hat{\mu} = \Sigma \frac{N_i}{N} \bar{y}_i$ = 0.2x105 + 0.5x180 + 0.27x370 + 0.03x590 = 228.6 **[6]**

$\hat{\tau} = N\hat{\mu}$ = 2000x228.6 = 457,200. **[4]**

[14] (c) Strata sample variances are missing from the table, so the variance of the sample mean cannot be estimated, using sample results only. **[2]**
Strata ranges from the last year inventory might be used to estimate the strata standard deviations for this year inventory, assuming they remain similar, even the mean values seem to be shifted up. Using ranges and the formula $\sigma_i = R_i / 4$ we obtain

$\sigma_1 = \sigma_2 = 25, \sigma_3 = \sigma_4 = 50$ **. [4]**

Then by ignoring the finite population corrections, we obtain (for proportional allocation)

$Var(\hat{\mu}) = \Sigma W_i^2 \sigma_i^2 / n_i = \Sigma W_i \sigma_i^2 / n = (0.2 \times 25^2 + 0.5 \times 25^2 + 0.27 \times 50^2 + 0.03 \times 50^2) / 100 =$ 11.875. **[8]**

(d) [7] Ignoring the sampling cost, do you expect that the stratified sample with the proportional allocation will produce significantly better results than an SRS of the same size? Explain.

(e) [14] (i) If the cost of sampling any one dealer is $20 (time spent on contact and telephone cost), and the total money that can be spent on sampling is $2000 (ignore presamping costs), calculate the allocation of the sample which would minimize the error bound, using information obtained from the last year. (ii) Do you expect that this allocation from (i) will produce better results than one already used in the study? Explain, but also justify using some calculation.

(f) [5] (i) What was the main goal of this study? Do you think the stratification used in this study is advantageous for the goal of the study? Explain. (ii) Propose some other stratification that may be convenient in a different way.

Solutions:

[7] (d) It seems from the sample that the strata means are quite different (it is also obvious from the strata limits), which means that proportional allocation should produce an estimator with smaller variance than an SRS. **[7]**

[14] (e) (i) The total sample size is n = 2000/20 = 100 **[2]**, due to equal costs of sampling from each stratum; we then can use the Neyman allocation, for given sample size. Relative allocation is then $\omega_i = N_i\sigma_i / \Sigma N_i\sigma_i$, and the allocation is $n_i = \omega_i n$, where we may use the strata ranges to estimate strata standard deviations (see (c)) Then $\sigma_1 = \sigma_2 = 25, \sigma_3 = \sigma_4 = 50$ **[2].**

$\Sigma N_i\sigma_i$ = 400x25 + 1000x25 + 540x50 +60x50 = 65000,

n_1 = 100x4x25/650 = 15.38=15,

n_2 = 100x10x25/650 = 38.46=39,

n_3 = 100x5.4x50/650 = 41.54=42,

n_4 = 100 – (15+39+42) = 4. **[6]**

[s.size values should be rounded reasonably so that the total is 100]

(ii) The used allocation is proportional, and this one calculated here is optimal. Optimal allocation is more efficient if strata standard deviations are different. **[2]** We can find the difference by $Var(\bar{y}_{STR,PR}) - Var(\bar{y}_{STR,NEY}) \approx \frac{1}{100}\sum_{i=1}^{4} W_i(\sigma_i - \bar{\sigma})^2$, where

$\bar{\sigma} = \sum_{i=1}^{4} W_i\sigma_i$ = 0.2x25 + 0.5x25 + 0.27x50 + 0.03x50 = 0.7x25 + 0.3x50 = 32.5

$\frac{1}{100}\sum_{i=1}^{4} W_i(\sigma_i - \bar{\sigma})^2$ =0.01x[0.7x(25-32.5)^2 + 0.3x(50-32.5)^2] = 1.3125. **[2]**

Comparing to the value $\hat{Var}(\bar{y}_{STR,PR})$= 11.875 obtained in (c), $\hat{Var}(\bar{y}_{STR,NEY})$= 11.875 - 1.3125 = 10.5625, is not a big improvement.

[5] (f) (i) The main goal of the study was to estimate the mean and total inventory for this year. **[1]** The stratification is very good, because the strata are homogeneous (stratified by the variable correlated with the variable of interest), that is, strata variances are small, and there are large differences between mean values in the strata. **[2]**

(ii) The other possible stratification would be by region, which might be convenient if the dealers should be visited, or if the goal of the study is also to estimate regional differences between dealers. **[2]**