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

STA 322H1 F SUMMER 2008, Second Test, June 12 (20%)
Duration: 50min. Allowed: nonprogrammable hand-calculator, aid-sheet, one side, with theoretical formulas only; the test contains 4 pages, please check

[55] 1) A plant has 400 employees and is interested in effects of their sickness on the income and productivity during the last year. A preliminary SRS of n = 25 employees has been selected from the list and the number of sick leaves (x), and the number of lost working days due to sick leave (y) were recorded. Ten employees with at least one sick leave were found in the sample, and the following results were obtained for them (don't forget employees without sick leaves in the sample!).

employee	1	2	3	4	5	6	7	8	9	10
x	1	1	3	2	1	2	4	3	1	1
y	3	5	12	5	2	7	16	10	4	3

$\Sigma x = 19, \Sigma x^2 = 47, \Sigma y = 67, \Sigma y^2 = 637, \Sigma xy = 171$

(a) Estimate the total number of sick leaves during the last year and place a bound on the error of estimation.

(b) Estimate the percentage of lost working days per employee due to sick leaves during the last year, and place a bound on the error of estimation (last year = 365 days).

(c) Find the sample size necessary to achieve 5% error bound for the estimation of the percentage of employees without sick leaves using the results from the presample.

(continued)

Solutions:

[11] (a) $\bar{x} = 19/25 = 0.76$, $\hat{\tau}_x = 400 \times \bar{x} = 304$, **[6]**

$S^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1) = 1.357$, $\hat{V}ar(\hat{\tau}_x) = 400^2 \frac{S^2}{25}\frac{400-25}{400} = 8140$,

$B = 2\sqrt{\hat{V}ar(\hat{\tau}_x)} = 2\text{x}90.22 = 180.44$. **[5]**

[12] (b) $z = y/365$ - percentage of lost working days during the last year,

$\bar{y} = 67/25 = 2.68$, $S_y^2 = 19.06$

$\hat{\mu}_z = \hat{\mu}_y / 365 = 2.68/365 = 0.73\%$ **[7]**

$\hat{V}ar(\hat{\mu}_z) = \hat{V}ar(\bar{y})/365^2 = \frac{400-25}{400}\frac{19.06}{25}\frac{1}{365^2} = \frac{0.71475}{365^2}$,

$B_z = 2\sqrt{\hat{V}ar(\hat{\mu}_z)} = 0.00463 = 0.463\%$. **[5]**

[10] (c) $n = N\hat{p}\hat{q}/((N-1)D + \hat{p}\hat{q}) = 400\frac{10}{25}\frac{15}{25}/(399 \times (0.025)^2 + \frac{10}{25}\frac{15}{25}) = 196$. **[8]**

$D = (B/2)^2 = (0.05/2)^2 = 0.025^2$. **[2]**

(d) Estimate the average number of lost working days per sick leave. What kind of estimator are you using? Place a bound on the error of estimation.

(e) It is known that the total number of lost working days was 1020. Use that information to estimate the total number of sick leaves. Which estimator, from (a), or this one from (e) would you use? Justify.

(f) (bonus) It is yet possible to use another estimator in part (e). Explain. Which one of these two that are possible in (e) would you use? Explain without calculation.

Solutions:

[14] (d) Use ratio estimator: $\hat{R} = r = \frac{\Sigma y}{\Sigma x} = \frac{67}{19} = 3.53$, **[6]**

$S_r^2 = \sum (y_i - rx_i)^2 /(n-1)$ = (637 – 2 x 3.53 x 171 + 3.53 ^{2} x 47)/24 =0.6418 **[4]**

$\hat{V}ar(\hat{R}) = \frac{N-n}{N} \frac{S_r^2}{\hat{\mu}_x^2 n} = \frac{400-25}{400} \frac{0.6418}{0.76^2 25} = 0.04167$,

$B_r = 2\sqrt{\hat{V}ar(\hat{R})} = 0.4082$. **[4]**

[8] (e) Use ratio type estimator of the total: $\hat{\tau}_x = \frac{\Sigma x}{\Sigma y} \tau_y = \frac{19}{67} \times 1020 = 289.25$. **[5]**

The number of lost working days (y) and the number of sick leaves (x) are correlated, so that the ratio estimator should be better than the sample mean estimator from (a). **[3]** (actual estimation of variances, which is not required, might show different result, and then can be accepted as an answer)
If student uses regression estimator and have similar discussion, this should be also accepted.

[5] (f) (bonus) It is also possible to use the regression type estimator of the total. As the regression goes through the origin (the number of lost working days is roughly proportional to the number of sick leaves), the regression and ratio type estimators should be equally efficient. The ratio estimator is easier to calculate and then might be preferable. The student needs not to choose between them, if the answer is justified.
(if the student uses regression estimator in (e) and ratio estimator here, the answer is also acceptable).

[45] 2) A provincial park management is interested in the structure of the visitors to the camping area. The park has 650 camping places situated into two park areas, the smaller on the East (250 places), and the larger on the West (400 places).

(a) Explain in short how would you select a stratified sample from the population of the camping places (you should adjust your explanation to this particular study).

(b) All camping place were occupied. A stratified sample of n = 12 places was selected from the park. The number of people camping on the place (campers), and out of it the number of campers over 65, were recorded. The following results were obtained:

Area	W	E
Camping place	1 2 3 4 5 6 7	8 9 10 11 12
Campers (y)	6 4 5 7 8 5 4	3 2 4 4 6
Camp. over 65 (x)	2 0 0 2 1 2 1	0 2 0 2 0
Sample mean	y: 5.57; x: 1.143	y: 3.8; x: 0.8
Sample variance	x: 0.81	x: 1.2

(i) Estimate the total number of campers in the park, and

(ii) the total number of campers over 65 in the park.

(continued)

Solutions:

[6] (a) Use, for example, the proportional allocation of the sample size, and decide on the total sample size. Then select an SRS of camping places from 250 places at the east side using three random digits, and an SRS from 400 places at the west side using also three random digits.

[12] (b) (i) $\hat{\tau}_y = \Sigma N_i \overline{y}_i = 250 \times 3.8 + 400 \times 5.57 = 3178$. **[6]**

(ii) $\hat{\tau}_x = \Sigma N_i \overline{x}_i = 250 \times 0.8 + 400 \times 1.143 = 657.2$.**[6]**

(c) Estimate the total number of camping places with campers over 65, and place a bound on the error of estimation.

(d) What would be the size and allocation of a stratified sample selected to estimate the total number of campers over 65, with a bound on the error of 70 campers? Use the sample indicated in (b) as a presample and the optimal allocation (cost of sampling is the same for both areas).

Solutions:

[12] (c) $\hat{\tau}_z = N\hat{p} = \Sigma N_i \hat{p}_i = 250 \times 2/5 + 400 \times 5/7 = 385.7$ **[6]**

$$\hat{V}ar(\hat{\tau}_z) = \Sigma N_i^2 \tfrac{N_i - n_i}{N_i} \tfrac{\hat{p}_i \hat{q}_i}{n_i - 1} = 250 \times (250-5) \times \frac{\frac{2}{5} \times \frac{3}{5}}{4} + 400 \times (400-7) \times \frac{\frac{5}{7} \times \frac{2}{7}}{6} = 9021.94\,,$$

$B_\tau = 2\sqrt{9021.94} = 189.97$. **[6]**

[15] (d) Using Neyman allocation

$$n = \frac{(\Sigma N_i \sigma_i)^2}{N^2 D + \Sigma N_i \sigma_i^2}\,,\ n_i = n \frac{N_i \sigma_i}{\Sigma N_i \sigma_i}\,,\ \hat{\sigma}_i^2 = S_{x,i}^2\,,$$

$\Sigma N_i \hat{\sigma}_i^2 = 250 \times 1.2 + 400 \times 0.81 = 624$, **[3]**

$\Sigma N_i \hat{\sigma}_i = 250 \times \sqrt{1.2} + 400 \times \sqrt{0.81} = 633.86$, **[3]**

$N^2 D = 650^2 (\frac{70}{2 \times 650})^2 = (\frac{70}{2})^2 = 35^2$,

$$n = \frac{(633.86)^2}{35^2 + 624} = 217.3 = 218\,,$$ **[5]**

$$n_1 = 218 \times \frac{250 \times \sqrt{1.2}}{633.86} = 94.19 = 94, \quad n_2 = 218 - 94 = 124\,.$$ **[4]**