Last name: ______________________ First name: ____________________ Student #:

STA 304H1F SUMMER 2016, Second Test, June 9 (20%)
Duration: 55 min. Allowed: nonprogrammable hand-calculator, aid-sheet, four pages, with theoretical formulas only, as posted on the web-site; the test contains 4 pages, please check.
You may use any back side, with clear indication of Question part. ***Your textual answers should be clear and short. Show the formula you are using, if any. Only answers on this test paper are counted.***

[45] 1) A city is divided into 600 blocks. A preliminary SRS of size 10 was selected from the city blocks and the following data was obtained (x - number of houses with finished basement, y - number of houses with rented basement, z – total number of houses in the block; it is assumed that an unfinished basement cannot be rented):

var	1	2	3	4	5	6	7	8	9	10	Σ (var)	Σ (var)^2
x	8	9	4	2	0	5	3	0	4	3	38	224
y	2	3	2	0	0	2	0	0	1	1	11	23
z	12	15	10	8	6	20	5	12	8	9	105	1283

$\Sigma xy = 68$

(a) Estimate the total number of houses in the city,
(b) Estimate the total number of houses with finished basement in the city,
(c) Estimate the total number of rented basements in the city,
(d) Estimate the proportion of houses with finished basement (out of all houses).

(continued)

Solutions:
From the table $\bar{x} = 3.8,\ \bar{y} = 1.1,\ \bar{z} = 10.5$.

(a) $\hat{\tau}_z = N\bar{z}$ = 600 x 10.5 = 6300, **[5]**

(b) $\hat{\tau}_x = N\ \bar{x}$ = 600 x 3.8 = 2280, **[5]**

(c) $\hat{\tau}_y = N\ \bar{y}$ = 600 x 1.1 = 660, **[5]**

(d) $\hat{R}_{x/z}$ = 38/105 = 0.362 = 36.2%, **[5]**

(e) Estimate the proportion of finished basements rented and the standard deviation of that estimator.

(f) It was found that the total number of houses in the city is 6200. Use this information to again estimate the total number of finished basements in the city. Do you expect this estimator be better than one in (b)? How can you check it? Just explain, don't do any calculation.

Solutions:

(e) $\hat{R}_{y/x}$ = 11/38 = 0.289 = 28.9%, **[5]**

$S_r^2 = \sum (y_i - rx_i)^2 /(n-1)$ = (23 – 2 x 0.289 x 68 + 0.289^2 x 224)/9 = 0.267, **[5]**

$\hat{V}ar(\hat{R}) = \frac{N-n}{N} S_r^2 /(n\overline{x}^2)$ = (600-10)/600 x 0.267/(10 x 3.8^2) = 1.82 x 10^{-3},

$\hat{S}d(\hat{R}) = \sqrt{1.82 \times 10^{-3}}$ = 0.043. **[5]**

(f) You may use the ratio estimator $\hat{\tau}_x = \hat{R}_{x/z}\tau_z$ = (38/105) x 6200 = 2243.8 = 2244. **[5]**

Yes, we expect this ratio estimator be better than one in (b), due to correlation between the number of basements and the block size (number of houses) **[3].** To check it, we just may calculate their variances. **[2]**

[55] 2) A certain city is divided into three service areas, the North, Southeast and Southwest, with 125,000, 75,000 and 50,000 households respectively (including apartments buildings and family homes). In the most recent survey, an SRS of households from each area was selected, and each selected household was interviewed. Among other variables, the number of years living in the household (length of stay, y), number of people living in the household, and whether the household is in apartment building, or a family home, were recorded. The results are summarized in the following table:

Area	Number of households	Sample size	Average Length of stay, $\bar{y}_i$, and standard deviation, S_i	Number of people in the sample	Family home, proportion $\hat{p}_i$
North	125,000	500	7.12 (1.81)	2,345	0.45
Southeast	75,000	300	10.24 (2.03)	1,020	0.62
Southwest	50,000	200	15.53 (2.15)	684	0.85
Total	250,000	1,000			

(you may assume that the population is … , where appropriate and convenient).

(a) Estimate (i) the average length of stay in the city, and (ii) the standard deviation of the estimator.
(b) Estimate (i) the total number of the people living in the city, and (ii) the average size of the household.
(c) Estimate (i) the total number of households living in a family home (not in an apartment building), and (ii) place a bound on the error of estimation.

(continued)

Solutions:
Use that the population is large, and then appropriate approximations.

(a) [12] (i) $\hat{\mu} = \sum W_i \bar{y}_i$ =0.5x7.12 + 0.3x10.24 + 0.2x15.53 = 9.738. **[6]**

(ii) $\hat{Var}(\hat{\mu}) = \sum W_i^2 \frac{S_i^2}{n_i} = 0.5^2 \frac{1.81^2}{500} + 0.3^2 \frac{2.03^2}{300} + 0.2^2 \frac{2.15^2}{200} = \frac{0.379882}{100}$ = 0.00379882.

$\hat{SD}(\hat{\mu}) = \sqrt{0.00379882}$ = 0.0616. **[6]**

(b) [8] (i) Use weighted mean of household sizes:

$\hat{\tau} = N\hat{\mu}_1 = 250{,}000 \times (0.5 \times \frac{2345}{500} + 0.3 \times \frac{1020}{300} + 0.2 \times \frac{684}{200})$ =1,012,250. **[4]**

(ii) $\hat{\mu}_1 = \frac{1{,}012{,}250}{250{,}000}$ = 4.049 (or first calculate $\hat{\mu}_1$, and then $\hat{\tau} = N\hat{\mu}_1$) **[4]**

(c) [12] (i) $\hat{\tau} = N\hat{p} = 250{,}000 \times (0.5 \times 0.45 + 0.3 \times 0.62 + 0.2 \times 0.85) =$ 145,250. **[6]**

(ii) $\hat{Var}(\hat{p}) = \sum W_i^2 \frac{\hat{p}_i \hat{q}_i}{n_i - 1} = 0.5^2 \frac{0.45 \times 0.55}{499} + 0.3^2 \frac{0.62 \times 0.38}{299} + 0.2^2 \frac{0.85 \times 0.15}{199} =$

$= \frac{0.02205}{100} = 0.0002205$, $\hat{SD}(\hat{\mu}) = \sqrt{0.0002205}$ =0.01485067,

$B_\tau = 2N \times \hat{SD}(\hat{\mu}) = 500{,}000 \times 0.01485067$ = 7425.34. **[6]**

(d) If the costs of sampling from each stratum were equal would the optimal allocation produce significantly better results in (a) than: (i) an SRS, (ii) proportional allocation, for the same sample size? Explain, without using calculation, but considering sample results. What about estimation in (c)?
(e) A new survey is planned. The costs of interviewing households from the North, Southeast and Southwest are $10, $15, and $20 respectively (differences mostly due to traveling). (i) What do you propose as *the minimal cost of a survey* that would estimate the average length of stay with a bound on the error of estimation of 0.2 years? Presampling costs may be ignored. (hint, if you want to do it faster: cost is required, not allocation) (ii) How much this would be more costly than the sampling reported in this question?
(f) (**bonus** [5]) Can you estimate the number of family homes in the city from the data? If answer is "Yes", explain why you can, and estimate it; if answer is "No", explain why not (what information is missing or unclear). The answer should be short and clear, don't write stories.

Solutions:

(d) [10] (i) Strata means are different (concluding from estimates from the presample), so the optimal allocation should produce better results than SRS. **[4]**
(ii) Standard deviations over strata (looking at presample) are close, so no significant improvement is expected from optimal allocation over proportional allocation. **[3]**

We should look at standard deviations $\sqrt{p_i q_i}$. There is a small difference between first two ($\sqrt{0.45\times 0.55} =$ 0.4975, $\sqrt{0.65\times 0.35} = 0.4770$), but greater that $\sqrt{0.85\times 0.15} = 0.3571$, so it likely that optimal allocation will perform better than proportional and SRS. Proportional will be better than SRS, because p_i are different. **[3]**
[if students concludes that the difference between standard deviations is not significant, it can be accepted that optimal is not better than the proportional allocation]

(e) [13] (i) For large population (here applicable), an approximate formula can be used to calculate minimal cost

$$C_{\min} = \sum n_i c_i = \frac{1}{D}(\sum W_i \sigma_i \sqrt{c_i})^2 = \frac{1}{(0.2/2)^2}(0.5\times 1.81\times\sqrt{10} + 0.3\times 2.03\times\sqrt{15} + 0.2\times 2.15\times\sqrt{20})^2$$

$$= \left(\frac{7.14353}{0.1}\right)^2 = \$5{,}103.\ \mathbf{[8]}$$

(ii) $C_{pre} = \sum n_i c_i = 10\times 500 + 15\times 300 + 20\times 200$ = $13,500.
$13,500 - $5,103 ≈ $8,400. **[5]**

(f) [5] Only a clear and strait answer is acceptable: Answer is NO, because it is possible that several households live in a single family home. This information is not clearly stated in the question.

If a student answers YES, he/she should state the assumption that in one family home lives one household only. In that case $\hat{\tau}_H = \hat{\tau} = 145{,}250$.