April 17, 2012

36-303: Sampling, Surveys and Society
Exam 2 — SOLUTIONS

1. [20 pts] *Multiple Choice (4 parts). For each part, circle the roman numeral of the one best answer.*

   (a) [5 pts] Let Y_i be the number of monthly neighborhood watch meetings attended by the i^{th} resident in a neighborhood, in the last year. You are going to conduct a survey, using a SRS to estimate $\overline{Y}_{pop}$, the population mean number of meetings attended by neighborhood residents. Among the N_R residents who *would respond* to your survey, the mean number of meetings is $\overline{Y}_R$, and among the N_M number of residents who would *not respond*, the mean number of meetings attended is $\overline{Y}_M$. In class we showed that the *bias* between $\overline{Y}_R$ and $\overline{Y}_{pop}$, due to missing responses, is

   $$\overline{Y}_R - \overline{Y}_{pop} = \frac{N_M}{N}(\overline{Y}_R - \overline{Y}_M)$$

   where $N = N_R + N_M$.

   Which statement below is false (or, circle iv. if all are OK)?

   [i.] The more people who respond to the survey, the smaller the bias due to missing responses.
   ii. The bigger the difference between mean number of meetinags attended by nonresponders, vs the mean number attended by responders, the bigger the bias due to missing responses.
   [iii.] The larger your SRS, the better you can estimate $\overline{Y}_{POP}$.
   iv. All of the above statements are true.

   NOTE: (iii) was the intended answer but due to poor editing on my part, either (i) or (iii) are acceptable answers. –BJ

   (b) [5 pts] Suppose we divide a sampling frame into groups, which we may treat as either strata for stratified sampling, or clusters for cluster sampling. If we make the groups so that *observations* **within groups** *are more* **similar** *to each other*, and *observations* **between groups** *are more* **different** *from each other*, then, all other things being equal, we expect

   [i.] The variance of the stratified sample mean $\overline{y}_{st}$ will go **down** and the variance of the cluster sample mean $\overline{y}_{cl}$ will go **up**.
   ii. The variance of the stratified sample mean $\overline{y}_{st}$ will go **up** and the variance of the cluster sample mean $\overline{y}_{cl}$ will go **down**.
   iii. Both variances will go **down**.
   iv. Both variances will go **up**.

   (c) [5 pts] In one-stage clustered sampling, the ICC ρ measures

   i. The correlation between observations in different clusters.
   ii. The correlation between the cluster means of different clusters.
   [iii.] The correlation between observations in the same cluster.
   iv. The correlation between the cluster mean and the individual observations in the cluster.

   (d) [5 pts] Which of the following is *not* a usual part of post-survey processing?

   i. Coding short-answer or text data
   ii. Variance calculation
   iii. Imputation
   iv. Checking post-strata and building weights if needed
   [v.] All of the above *are* usually part of post-survey processing!

2. [24 pts] *Stratified sampling…*

Stratum Number (h)	Employee Type	Population Size	No. of Surveys Distributed	Number of Responses	Number of Yes's	$\bar{y}_h$
1	Faculty	1374	500	232	167	0.72
2	Classified Staff	1960	653	514	459	0.89
3	Administrative Staff	252	74	67	58	0.87
4	Academic Professional	95	95	86	75	0.87
		3681	1322	899	759	

(a) [5 pts] $\bar{y}_{srs}$ and $\sqrt{\text{Var}(\bar{y}_{srs})}$, using $\hat{p}(1-\hat{p})$…

- $\bar{y}_{srs} = 759/899 = \boxed{0.84}$
- $\text{Var}(\bar{y}_{srs}) = (1-f)s_y^2/n = (1-n/N)s_y^2/n = (1-899/3681)(0.84)(1-0.84)/899 = 0.0001129876$, so the SE is $\sqrt{0.0001129876} = \boxed{0.0106}$

(b) [9 pts] Now treat this as a pre-stratified sample, and calculate $\bar{y}_{st}$ and $\sqrt{\text{Var}(\bar{y}_{st})}$, the stratified estimate and SE of the proportion of faculty and staff that responded "Yes". *[HINT: Use the number of responses as the sample size within each stratum; and use the same idea as in part (2a) for the needed variance calculations.]*

- Stratum weights and sampling fractions are

h	1	2	3	4
N_h	1374	1960	252	95
n_h	232	514	67	86
$W_h = N_h/N$	0.37	0.53	0.07	0.03
$f_h = n_h/N_h$	0.17	0.26	0.27	0.91

  where $N = N_1 + N_2 + N_3 + N_4 = 3681$.
- $\bar{y}_{st} = \sum_{h=1}^{H} W_h\bar{y}_h = (0.37)(0.72) + (0.53)(0.89) + (0.07)(0.87) + (0.03)(0.87) = \boxed{0.8251}$
- $\text{Var}(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2(1-f_h)s_h^2/n_h = (0.37)^2(1-0.17)(0.2016)/232 + (0.53)^2(1-0.26)(0.0979)/514 + (0.07)^2(1-0.27)(0.1131)/67 + (0.03)^2(1-0.91)(0.1131)/86 = 0.0001444743$, so the SE is $\sqrt{0.0001444743} = \boxed{0.0120}$.

(c) [4 pts] Calcuate the DEFF for the stratified design and comment on whether it was worthwhile to stratify.

- DEFF $= \text{Var}(\bar{y}_{st})/\text{Var}(\bar{y}_{srs}) = 0.0001444743/0.0001129876 = \boxed{1.28}$
- In this case it was $\boxed{\textit{not worth it}}$ to stratify since the DEFF is bigger than 1 (and when stratification is working it is more efficient than SRS, i.e. DEFF<1!).

(d) [6 pts] Clearly not everyone who received a survey responded.

- The response rates in each stratum are as follows:

h	1	2	3	4
n_{resp}	232	514	67	86
$n_{sampled}$	500	653	74	95
resp rate $= n_{resp}/n_{sampled}$	$\boxed{0.46}$	$\boxed{0.79}$	$\boxed{0.91}$	$\boxed{0.91}$

- I would expect $\boxed{\text{more nonresponse bias in the SRS estimates}}$. (A reason does not have to be given but my reasoning is this: The response rates differ dramatically across strata and so do the proportions of employees who would want the school closed over winter break in the future. When the response rate is related to the response, we expect to find nonresponse bias.)

3. [18 pts] The city council of a suburb of Chicago wants to know the proportion of eligible voters that oppose having a Chicago garbage incinerator opened in that suburb. They randomly select 100 residential phone numbers from the suburb's telephone book (which contains 3,000 residential numbers in all). Each selected residence is then called and asked for (a) the total number of eligible voters in that household and (b) the number of voters in the household opposed to the incinerator. A total of 157 voters were surveyed; of these, 23 refused to answer the question. Of the remaining 134 voters, 113 opposed to the incinerator, so the council estimates the proportion opposed as

$$\hat{p} = 112/134 = 0.83582$$

with

$$\text{Var}(\hat{p}) = 0.83582(1 - 0.83582)/134 = 0.00102$$

The use of 112 instead of 113 in these formulas was just a typo and not relevant to the statistical issues that this problem was inteded to address.

(a) [6 pts]

- The **target population** is registered voters in the suburb.
- The **sampling frame** is households listed in the phone book. *Or you could say:*
  - *The phone book, or*
  - *Voters in households listed in the phone book.*

(b) [6 pts]

- The psu's (primary sampling units) are the households.
- The ssu's (secondary sampling units) are the registered voters in each household.

(c) [6 pts]

- The the estimates $\hat{p}$ and $\text{Var}(\hat{p})$ given in the problem statement are not valid.
- REASON(S): *(only one is needed)*
  - A cluster sample is being treated as an SRS. The estimate of $\hat{p}$ is approximately correct, but the variance will likely be too small (since, likely, DEFF>1). *(this is the better answer)*
  - There may be nonresponse bias since we do not know why 23 of the voters refused to answer.

4. [18 pts] Give at least one strength and one weakness of each of the following strategies for identifying and correcting such coding errors.

- [6 pts] Examining the distribution of the data as it was entered, and identifying and eliminating any outliers.
  - **Strength:** Fast, doesn't take much effort, will identify observations which seem too high or too low relative to the rest of the data (like 1750 in the example).
  - **Weakness:** Will probably not identify observations that are miscoded and in the middle of the distribution (like 15,799 or 17,588 in the example).
- [6 pts] Drawing a 10% sample of the 20,000 responses, and identifying and correcting any errors found.
  - **Strength:** Fast, doesn't take much effort, will find any kind of miscoding in the sample (whether it is an outlier like 1750 or a non-outlier case like 15,799 or 17,588).
  - **Weakness:** 90% of the data will not be checked for errors!
- [6 pts] Examining all 20,000 responses and correcting any errors found.
  - **Strength:** Every coding error will be found and corrected.
  - **Weakness:** Takes a great deal of time and effort!

5. [20 pts] *Imputation methods.*

(a) One method of imputation for missing responses to individual survey items is *mean imputation*.

i. [4 pts] Explain briefly how mean imputation works.

If the response to question #17 is missing, fill in the mean of all the responses to #17 that you have from other survey participants who did answer that question.

ii. [3 pts] This approach only works under MCAR.

iii. [3 pts] Identify a possible problem with mean imputation.
Only one answer is required:

- *MCAR might not be true*
- *Whatever the mean is, it probably isn't what this person would have responded.*

(b) Another method of imputation for missing responses is *regression imputation*.

i. [4 pts] Explain briefly how regression imputation works.

If the response to question #17 is missing, build a regression equation for answers to #17 using demographic information, answers to other survey questions, etc., using those survey participants that did answer #17. Then use the regression equation with this person's demographic data & answers to other survey questions, to predict what this person would have answered to #17.

ii. [3 pts] This works under MAR.

iii. [3 pts] Identify a possible problem with regression imputation.

MAR means, in this case, that all people with the same demographic informaion and the same answers to other survey questions would be nonresponders by chance only, and not because of what they would have said. If we don't get the right variables into the regression equation, then we can't argue that MAR holds.

Other answers may be acceptable here, but the above is the best answer.

Note: A typo on the test mixed had "hot-deck imputation" instead of "regression imputation" for one part of this question. But the correction should have been announced during the exam.