36-303: Sampling, Surveys and Society
Exam 2
Solutions

1. *[24 pts]* Multiple Choice (4 parts). For each part, circle the roman numeral of the one best answer.

   (a) *[6 pts] Which of the following is* not *a usual part of post-survey processing?*

   ii. Sample size calculation

   (b) *[6 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*

   ii. 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**.

   (c) *[6 pts] Which of the following is* not *one of the recommended things to work on, to reduce the tendency of survey subjects to not respond?*

   ii. Choice of stratified or cluster sampling.

   (d) *[6 pts] Weights can be calculated and applied to individual observations for a variety of reasons. Circle the reason below that is* not *appropriate.*

   iv. All of the above *are* appropriate reasons to compute weights!

2. *[26 pts]* Cluster sampling. *A survey is conducted to find out the proportion of cell phone users in a certain city. From a population of 2500 residential blocks, 10 are sampled at random without replacement, and each person residing on that block is asked whether they use a cell phone. We will assume there are exactly 40 people living in each block*[1]*. This yields the following table of data:*

[1] This is a reasonable approximation as long as all the blocks have close to 40 residents.

Block h	*Total # of People* M	*# of Cell Phone Users* c_h	*Proportion of Cell Phone Users* p_h
1	*40*	*10*	*0.25*
2	*40*	*8*	*0.20*
3	*40*	*16*	*0.40*
4	*40*	*15*	*0.38*
5	*40*	*24*	*0.60*
6	*40*	*17*	*0.42*
7	*40*	*12*	*0.30*
8	*40*	*13*	*0.32*
9	*40*	*16*	*0.40*
10	*40*	*13*	*0.32*
Total	*400*	*144*	

(a) [6 pts] This is an example of one-stage clustered sampling. Circle one word in each pair of choices in the following sentences:

- *The primary sampling units (psu's) are the* **(blocks , residents)**.
- *The secondary sampling units (ssu's) are the* **(blocks, residents)**.

(b) [4 pts] Ignoring the clustering and treating this as an SRS of 400 residents, estimate the proportion of cell phone users and its standard error.

Under SRS with replacement,

$$\begin{aligned} \hat{p}_{SRS} &= 144/400 = \boxed{0.36} \\ \text{Var}_{SRS}(\hat{p}) &= (1 - n/N)\hat{p}(1-\hat{p})/n \\ &= (1 - 40 * 10/(40 * 2500)) * (0.36) * (1 - 0.36)/400 = 0.000573696 \\ \text{SE}_{SRS}(\hat{p}) &= \sqrt{0.000573696} = \boxed{0.02395195} \end{aligned}$$

(c) [6 pts] Now re-estimate the proportion of cell phone users and its standard error, using appropriate cluster sampling methods. Hint: to reduce calculation, use the fact that in the table above, $s^2_{p_h} = 0.01214333$. **NOTE: on the exam, I wrote 0.1214333 by mistake, but corrected it in class.**

Under single stage clustering with equal cluster sizes, the sample is self-weighting, so we know without any calculation that $\boxed{\hat{p}_{CL} = \hat{p}_{SRS} = 0.36}$ again. Of course you can obtain that by averaging the numbers in the p_h column in the data table as well (with rounding error in the table, you might end up with 0.359 if you do the average of the p_h's by hand).

For the variance, we treat clusters as sampling units and cluster means as the unit-level data. So

$$\begin{aligned} \text{Var}_{CL}(\hat{p}) &= (1 - n/N)s^2_{p_h}/n \\ &= (1 - 10/2500) * (0.01214333)/10 = 0.001209476 \\ \text{SE}(\hat{p}) &= \sqrt{0.001209476} = \boxed{0.03477752} \end{aligned}$$

(d) [6 pts] Calculate the design effect DEFF for this design.

$$\text{DEFF} = \frac{\text{Var}_{CL}(\hat{p})}{\text{Var}_{SRS}(\hat{p})} = \frac{0.001209476}{0.000573696} = \boxed{2.108218}$$

(e) [4 pts] Calculate ρ, the correlation between responses from residents on the same block.

$$\text{DEFF} = 1 + (M-1)\rho$$

so

$$\rho = (\text{DEFF} - 1)/(M-1) = (2.108218 - 1)/(10-1) = \boxed{0.1231353}$$

3. *[24 pts]* Response rates and missing data. *You are completing a telephone survey of an SRS of 1000 members of a much larger professional organization, regarding their level of involvement in support of the organization. Currently the response rate is 80%, with 52.5% of those responding saying they attend every monthly meeting of the local chapter of the organization, and 47.5% saying they do not.*

(a) [8 pts] Does it seem likely that the 200 (20% of 1000) who did not respond to your survey are missing completely at random (MCAR, ignorable missingness) or missing not at random (MNAR, non-ignorable missingness)? Choose MCAR or MNAR and briefly *explain your reasoning.*

It seems likely that these 200 respondents are MNAR. The survey is about *participation*, and *nonresponse is a form of nonparticipation*. Therefore I expect a lower proportion of the 200 nonrespondents to say that they attend every monthly meeting, than the 800 who did respond/participate in the survey. Since not responding is related to the response we would have seen, the missing data is MNAR.

(b) [8 pts] Which of the following is more likely to be correct (circle one):

ii. An unbiased estimate of the population proportion that attends every meeting would probably be less than 52.5%.

(c) [8 pts] To get the sample size nearer to the target of n = 1000, you could either

- *Ask the organization to pay for you to call a new SRS of size 250, hoping that* (0.80)(250) = 200 *people choose to respond* (cost: $200 because it just involves routine single calls to each new phone number)*; or*
- *Ask the organization to pay for you to followup with the 200 in your original sample who didn't respond yet, to try to get their responses* (cost: $800 because it involves repeated call-backs until each of the 200 non-respondents either responds or refuses).

Which do you choose, and why (briefly)*?*

Although it is more expensive, I prefer following up the 200 non-respondents. We want an unbiased estimate of participation, and the people we get in the 2nd sample are again likely to be the participators, rather than the nonparticipators/nonresponders. Since the 200 are likely MNAR it is important to convert them to responders, rather than looking for other easy-responders in

a new sample of 250. We will never find out how nonrespondents would have responded, if we only collect data from those who find it easy to respond, we will just get more data for a biased estimate that way.

4. *[26 pts]* Imputation methods.

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

      i. *[5 pts] Explain briefly how mean imputation works.*
      For each missing value, we fill in the average of all of the respondents who did provide a response to that question.

      ii. *[4 pts] Under what assumption (MCAR, MAR, MNAR) is mean imputation OK? (Choose one, no explanation needed.)*

      MCAR.

      iii. *[4 pts] Identify a possible problem with mean imputation.*
      The mean value of everyone else may not even be an admissible response. For example if the question is how many times a week do you go to the gym, admissible answers are whole numbers. However the average over all the responders might well not be a whole number.

   (b) *Another method of imputation for missing responses is* hot-deck imputation*.*

      i. *[5 pts] Explain briefly how hot-deck imputation works.*
      In hot deck imputation we find someone who (a) did respond to the question that is missing for this respondent; and (b) looks like this respondent in every other way (has similar demographics, similar responses to other questions, etc.). We then copy the response from the person who had the response to the person who didn't.

      ii. *[4 pts] Under what assumption (MCAR, MAR, MNAR) is hot-deck imputation OK? (Choose one, no explanation needed.)*

      MAR (MCAR too, but MAR is a better answer).

      iii. *[4 pts] Identify a possible problem with hot-deck imputation.*
      Here are two possible problems:
      - There might not be a good "matching" person among the other respondents in the survey.
      - The less good the match, the less likely it is that that the two people would have responded the same way. Even if the match is very good, there's no guarantee that the two people would have responded the same way.