April 13, 2010

# 36-303: Sampling, Surveys and Society
# Exam 2 Solutions — Tue Apr 13, 2010

1. [20 pts] *Multiple Choice (4 parts).*

   (a) [5 pts] *The* bias *between* $\overline{Y}_R$ *and* $\overline{Y}_{pop}$*, due to missing responses, is*

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

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

   [i.] All other things being equal, the more people who would *not respond*, the smaller the bias due to missing responses.

   ii. All other things being equal, the bigger the difference between mean number of meetings 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 $\overline{Y}_R$.

   iv. All of the above statements are true.

   (b) [5 pts] *In one-stage clustered sampling, the ICC* $\rho$ *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.

   (c) [5 pts] *Some respondents are never presented with the last page of the survey, due to a bug in the website software. This is most likely an example of*

   i. Unit nonresponse, Missing At Random (MAR)

   [ii.] Item nonresponse, Missing Completely At Random (MCAR)

   iii. Unit nonresponse, Missing Completely At Random (MCAR)

   iv. Item nonresponse, Missing At Random (MAR)

   (d) [5 pts] *If* 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 **up** and the variance of the cluster sample mean $\overline{y}_{cl}$ will go **down**.

   [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**.

   iii. Both variances will go **up**.

   iv. Both variances will go **down**.

2. [20 pts] *Stratified Sampling (3 parts).*

| | *Stratum 1* *Alpha* | *Stratum 2* *Berea* | *Stratum 3* *Rural Area* |
|---|---|---|---|
| *Number of households* | $N_1 = 155$ | $N_2 = 62$ | $N_3 = 93$ |
| *Stratum sample size* | $n_1 = 20$ | $n_2 = 8$ | $n_3 = 12$ |
| *TV viewing time, hr/wk* | 35 28 26 41<br>43 29 32 37<br>36 25 29 31<br>39 38 40 45<br>28 27 35 34 | 27 4 49 10<br>25 41 25 30 | 8 15 21 7<br>14 30 20 11<br>12 32 34 24 |
| *Stratum Statistics* | $\overline{y}_1 = 33.9000$<br>$s_1^2 = 35.358$ | $\overline{y}_2 = 25.125$<br>$s_2^2 = 232.411$ | $\overline{y}_3 = 19.000$<br>$s_3^2 = 87.636$ |

*Note:* The given value of $\overline{y}_2$ is incorrect, given the data in the table (the correct value was 26.375). However, I asked everyone to use the given value.

(a) [6 pts] *Compute $\overline{y}_{st}$ and $\overline{y}_{srs}$. If they should be the same, explain why. If they should be different, explain why.*

We first calculate the stratum weights,

$$W_1 = N_1/N = 155/310 = 0.5; \quad W_2 = N_2/N = 62/310 = 0.2; \quad W_3 = N_3/N = 93/310 = 0.3.$$

Then we calculate

$$\overline{y}_{st} = W_1\overline{y}_1 + W_2\overline{y}_2 + W_3\overline{y}_3 = (0.5)(33.9) + (0.2)(25.125) + (0.3)(19.000) = \boxed{27.675}$$

The value for $\overline{y}_{srs}$ is the same, because this design is self-weighted[1]. You can verify that by calculating $\overline{y}_{srs}$ explicitly.

(b) [6 pts] *Compute $SE(\overline{y}_{st})$ and use this to create an approximate 95% confidence interval for the mean household TV viewing time (hrs/wk) in Jefferson County.*

We need the sampling fractions

$$f_1 = n_1/N_1 = 20/155 = 0.129; \quad f_2 = n_2/N_2 = 8/62 = 0.129; \quad f_3 = n_3/N_3 = 12/93 = 0.129$$

The variance is then

$$\begin{aligned} SE(\overline{y}_{st})^2 &= W_1(1-f_1)s_1^2/n_1 + W_2(1-f_2)s_2^2/n_2 + W_3(1-f_3)s_3^2/n_3 \\ &= (0.5)(1-0.129)(35.358/20) + (0.3)(1-0.129)(232.411/8) \\ &\quad + (0.2)(1-0.129)(87.636/12) = 9.633 \end{aligned}$$

[1] A stratified design with the same sampling fraction $f$ in each stratum is self-weighted. See answer to part (b) also.

and the standard error is

$$SE(\overline{y}_{st}) = \sqrt{9.633} = \boxed{3.104},$$

so the 95% CI is about $27.675 \pm 2 \times 3.104$, or $\boxed{(21.467, 33.883)}$.

(c) [8 pts] *If we were to treat this as an SRS rather than a stratified sample, the sample variance would be* $s^2 = 83.658$.

i. *Estimate the DEFF (design effect) for this design.*

ii. *Was it a good idea to do the stratified sample rather than an SRS? Why or why not?*

The DEFF is

$$DEFF = \frac{\text{Var}(\overline{y}_{st})}{\text{Var}(\overline{y}_{srs})} = \frac{9.633}{83.658} = \boxed{0.12}$$

Since $\text{Var}(\overline{y}_{st})$ is almost 10 times smaller than $\text{Var}(\overline{y}_{srs})$, it is
definitely better to do the stratified design.

*Note:* $Var(\overline{y}_{srs})$ is actually wrong here. It should have been $Var(\overline{y}_{srs}) = 11.136$, instead of 83.658. In that case, DEFF = 9.633/11.136 = 0.87. This is still a win for the stratified sampling design.

3. [18 pts] *Approval Rating of Sen. Diane Feinstein, Part I (3 parts).*

*Approximately equal numbers of Democrats (342) and Republicans (298) were contacted by phone; 267 of the Democrats and 98 of the Republicans approved of the job Sen. Feinstein was doing.*

*According to the Ballot Access News (http://www.ballot-access.org/2008/120108.html), there were 7,683,495 registered Democrats and 5,428,052 registered Republicans in California at approximately the same time as the Field Poll was being taken.*

(a) [6 pts] *Compute* $\hat{p}_{dem}$*, the proportion of Democrats who approve of Sen. Feinstein, and its standard error,* $\text{SE}(\hat{p}_{dem})$.

$$\begin{aligned} \hat{p}_{dem} &= 267/342 = \boxed{0.78} \\ SE(\hat{p}_{dem}) &= \sqrt{(1-f)\hat{p}_{dem}(1-\hat{p}_{dem})/(n_{dem}-1)} \\ &= \sqrt{(1-342/7683495)(0.78)(1-0.78)/(342-1)} \\ &= \boxed{0.022} \end{aligned}$$

(b) [6 pts] *Do the same for Republicans: compute* $\hat{p}_{rep}$ *and* $\text{SE}(\hat{p}_{rep})$.

$$\begin{aligned} \hat{p}_{rep} &= 98/298 = \boxed{0.33} \\ SE(\hat{p}_{rep}) &= \sqrt{(1-f)\hat{p}_{rep}(1-\hat{p}_{rep})/(n_{rep}-1)} \\ &= \sqrt{(1-298/5428052)(0.33)(1-0.33)/(298-1)} \\ &= \boxed{0.027} \end{aligned}$$

(c) [6 pts] *Using the SE for the difference,* $\sqrt{\text{SE}(\hat{p}_{dem})^2 + \text{SE}(\hat{p}_{rep})^2}$*, compute an approximate 95% CI for the difference between the true proportions of Democrats and Republicans who approve of Senator Feinstein.*

The combined SE is

$$\sqrt{\text{SE}(\hat{p}_{dem})^2 + \text{SE}(\hat{p}_{rep})^2} = \sqrt{(0.022)^2 + (0.027)^2} = 0.035$$

and the difference vetween the two $\hat{p}$'s is

$$\hat{p}_{dem} - \hat{p}_{rep} = 0.78 - 0.33 = 0.45$$

so the CI is $0.45 \pm 2 \times 0.035$, or $\boxed{(0.38,0.52)}$.

4. [24 pts] *Approval Rating of Sen. Diane Feinstein, Part II (4 parts).*

(a) [6 pts] *Combining Democrats and Republicans, estimate the overall proportion of voters in California who approve.*

$\hat{p} = (267 + 98)/(342 + 298) = \boxed{0.57}$

(b) [6 pts] *Using political party registration to define two post-strata, compute post stratification weights to apply to this survey.*

Note that the total population size is $N = N_{dem} + N_{rep} = 13,111,547$, and the total sample size is $n = n_{dem} + n_{rep} = 640$.

Using this information, the weights are then

For Democ. $i$'s: $w_i = (N_{dem}/N)/(n_{dem}/n) = (7,683,495/13,111,547)/(342/640) = \boxed{1.097}$

For Repub. $i$'s: $w_i = (N_{rep}/N)/(n_{rep}/n) = (5,428,052/13,111,547)/(298/640) = \boxed{0.889}$

(c) [6 pts] *Compute the weighted proportion of voters in California who approve of Sen. Feinstein's job.*

$$\overline{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i} = \frac{(1.097)(267 \times 1 + 75 \times 0) + (0.889)(98 \times 1 + 200 \times 0)}{(1.097)(342) + (0.889)(298)} = \boxed{0.59}$$

since each $y_i$ is either 1 (approve) or 0 (disapprove).

(d) [6 pts] *Use the values of the weights to explain the difference between your answers in parts (a) and (c).*

The weight for Democratic respondents, 1.097, is larger than the weight for Republican respondents, 0.889 (to make the Democratic and Republican contributions to $\overline{y}_w$ proportional to their population counts). Since 1.097>0.889, the favorable Democratic responses will count a bit more than the disfavorable Republican responses, leading to a higher estimated proportion of voters approving of Sen. Feinstein.

5. [18 pts] *Data Collection Methods (3 parts.) In each situation below, choose the best mode of data collection, and give a reason why.*

> Note: Below I have given the intended mode of data collection and "model" reasons for those answers. You might come up with other reasons for these answers. You might even come up with different modes of data collection. Your answer and/or reason will be counted correct, if different from mine, **as long as** you provide a **clear and convincing reason**. The grader's decisions (or mine, if he asks for my help) will be final.

(a) [6 pts] *Your market research firm has been asked to conduct a survey in Pittsburgh to see who would purchase a new, sweet, fruity carbonated soft drink.*

- *Which mode of data collection (circle the best answer)?*
  i. Telephone survey.
  ii. Face to Face interviews.
  iii. Internet (web-based) survey.
- *Why?*
  Here are some possible considerations:
  - An Internet survey would not reach enough of the target population (almost anyone buys sodas, but not everyone has consistent internet access). This could introduce some bias into the survey, since we would not have opinions of people who for economic or other reasons did not have internet access.
  - A telephone survey would reach most households and consumers. This is a good, not too expensive choice. A drawback of this choice is that you can't have consumers actually try to the new soda; and this is information consumers might need before deciding whether they would purchase it.
  - Face to face interviews are expensive but you can have consumers try the product directly, rather than asking them to respond to a completely hypothetical question.

(b) [6 pts] *The Governor of your state has asked you to design a survey to assess the views of residents of the state toward public transportation and to evaluate how likely they would be to use alternative modes of transportation.*

- *Which mode of data collection (circle the best answer)?*
  i. Telephone survey.
  ii. Face to Face interviews.
  iii. Internet (web-based) survey.
- *Why?*
  Here are some possible considerations:
  - An Internet survey would not reach enough of the target population (almost anyone has to consider modes of transportation, but not everyone has consistent internet access). This could introduce some bias into the survey, since we would not have opinions of people who for economic or other reasons might not have internet access (and this might be correlated with transportation choices).
  - Face to face interviews are expensive and there is almost no information on this topic that you could provide the respondent with in a face to face setting, that you could not also provide on the phone.
  - A telephone survey would reach most people in the state (good coverage of the target population by the sampling frame). This is a good, not too expensive choice.

(c) [6 pts] *The ASA wishes to do a survey of its members across the United States on their views toward certification of statisticians (statisticians would need pass a test in order to get a license to practice statistics, like doctors, some engineering professions, etc.).*

- *Which mode of data collection (circle the best answer)?*
  i. Telephone survey.
  ii. Face to Face interviews.
  iii. Internet (web-based) survey.
- *Why?*
  Here are some possible considerations:
  - Face to face interviews would be prohibitively expensive for a nationwide survey.
  - A telephone survey would reach most members of the ASA. This is a good choice, but...
  - An Internet survey would be much cheaper than a face-to-face survey, and also cheaper than a telephone survey. If there is a great deal of text to consider (e.g. a description of the certification process) it is better presented on a web page than verbally over the phone. Moreover, since most statisticians are white collar workers with ready access to the internet, few statisticians would be unable to complete the survey for lack of access to the survey website.