36-303 Sampling, Surveys & Society
Homework 04 Solutions

April 8, 2010

1 Question 1 (Robert Groves Visit)

I put a sampling of answers, questions, observations, and comments at the end of this solution sheet.

2 Question 2 (taken from S. Lohr's book)

2.1 a)

The summary table provides the distinct error proportions for 5 groups of clusters.

$$\bar{y}_{cl} = \frac{1}{85}(0.01860 * 1 + 0.01395 * 1 + 0.009302 * 4 + 0.004561 * 22 + 0 * 57) = 0.002$$

now for the standard error of the estimate:

$$s^2_{\bar{y}_i} = \frac{1}{85-1} * [(0.01860 - 0.002)^2 + (0.01395 - 0.002)^2 + 4 * (0.009302 - 0.002)^2$$
$$+22 * (0.004561 - 0.002)^2 + 57 * (0 - 0.002)^2] = 1.195156 * 10^{-5}$$

$$var(\bar{y}_{cl}) = (1 - \frac{85}{828})\frac{1}{85} * s^2_{\bar{y}_i} = (1 - \frac{85}{828})\frac{1}{85} * 1.195156 * 10^{-5} = 1.26172 * 10^{-7}$$

$$se(\bar{y}_{cl}) = 0.000355$$

2.2 b)

an estimate of the total number of errors should be $178020 * 0.002 = 356$ based on our answer from part a, we could find a standard error estimate as

$$var(y_{\text{total}}) = var(178020 * \bar{y}_{cl}) = 178020^2 * var(\bar{y}_{cl})$$

so our standard error estimate should be $178020 * 0.000355 = 63.197$

2.3 c)

In this case we have to think our universe as composed from 'fields'. We have $N = 828 \times 215$ fields and our sample is composed from $n = 85 \times 215$ fields. Assuming the error rate is the same as in the previous case,

$$\hat{p}_{SRS} = \frac{\text{fields with errors}}{n} = \frac{37}{85 \times 215} = 0.002025 \tag{1}$$

The interesting thing happens when we compute the variance assuming that the sample is effectively a SRS *of fields*:

$$\hat{V}[\hat{p}_{SRS}] = \left(1 - \frac{85 \times 215}{828 \times 215}\right) \frac{\hat{p}_{SRS}(1 - \hat{p}_{SRS})}{85 \times 215} = 9.92 \times 10^{-8}$$

If we compare this estimate with the estimate obtained in part a), assuming cluster sampling,

$$\hat{V}[\hat{y}_{cl}] = 1.26172 \times 10^{-7} \tag{2}$$

we see that this last variance estimate is bigger than the one computed assuming SRS. This is a general phenomenon when we have clustered samples. To achieve the same error levels, a clustered sample must be bigger than a SRS. This example also illustrate the problems of analyzing clustered samples using SRS methods: the SE of the estimates will be underestimated. This is dangerous because we (and others) will think that our point estimates are better than they really are.

3 Question 3

Creating the dataset in R:

```
strata <- data.frame(expand.grid(Sex=factor(c('M','F')),
     College=factor(c('Eng','Lib'))),
  n_h = c(8,4,2,6),
  N_h = c(617,450,380,551),
  sam_w = NA,
  Pop_W = NA
)
strata$Pop_W <- strata$N_h / sum(strata$N_h)
strata$sam_w  <- strata$n_h / sum(strata$n_h)
HrsWk <- c(28,29,23,35,29,30,34,31,30,31,36,33,27,28,29,30,28,28,32,30)
data <- cbind(strata[rep(1:NROW(strata), strata$n_h),],HrsWk)
```

3.1 a)

The mean of 'Hrs/Wk' is

```
> mean(data$HrsWk)
[1] 30.05
```

Since this is SRS without replacement, an estimate of the standard error of the mean is

$$\hat{SE}[\bar{y}] = \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}}$$

where s^2 is the sample variance. Computing this,

```
> fpc <- 1 - sum(strata$n_h)/sum(strata$N_h)
> sqrt(fpc) * sqrt(var(data$HrsWk) / NROW(data))
[1] 0.6634416
```

3.2 b)

Computing the post-stratification weights,

```
data <- cbind(data,PSW = data$Pop_W/data$sam_w)
```

And the weighted mean using the post-stratification weights is

```
> weighted.mean(data$HrsWk, w = data$PSW)
[1] 29.9111
```

which is slightly lower than the one without using the population-level information.

3.3 c)

To estimate the SE using a first order Taylor series approximation we use the R procedure given in class (don't forget the finite population correction)

```
>tsv <- ts.variance(data$HrsWk, w = data$PSW)
> se.ts <- sqrt(fpc)*sqrt(tsv$var.ts)
> se.ts
[1] 0.6709889
```

4 Question 4

To estimate the SE using the Jackknife technique we use the R function given in class,

```
> stacked_strata <- data.frame(stratum = paste(data$Sex,
    data$College, sep='.'),data$N_h)
> jkv <- jk.variance(data$HrsWk, stacked_strata$stratum,
    unique(stacked_strata$data.N_h))
> jkv
$ybar.weighted
[1] 29.9111

$ybar.reps
[1] 29.9111

$var.jk
[1] 0.3419697
```

Then, the jacknife estimate of the SE is

```
> se.jk <- sqrt(fpc)*sqrt(jkv$var.jk)
> se.jk
[1] 0.5818476
```

5 Answers to Question 1

Here are a few observations from your notes on his visit...

Lecture in McConomy

- The question time at the public lecture had a few politically loaded questions. In particular, someone asked aobut prisoners being counted in the location where the prison was instead of their home community. Groves answered by pointing out that deciding where to count prisoners is not very straightforward, since the "home community" of the prisoner might be their last "permanent residence" before conviction, the location where they commited the crime or were convicted, the location they intended to go to when released from prison, etc.
- Groves also made the point that in order for the census to be taken seriously, it has to be politically neutral.
- The difference in cost between mailing in the census form (around \$0.44) vs having someone visit your home to collect the data (about \$60.00) leads one to wonder if there is a better way to raise mail-in response rates for the census.

Class Visit

- Concern about correcting errors on census forms; it might bias results towards "social norms". What imputation methods are actually used? How many forms require imputation? *[Note: I do not know the answers to these questions. However, since the 2010 mail-in form is quite short, there probably is less need/opportunity for imputation than in past years. Moreover, if very many questions are omitted by the respondent, Census sends an enumerator to visit your home to get the correct answers, rather than trying to impute. –BJ]*
- The use of the word "negro" as a possible racial designation on the census form intrigued several people. On the one hand, it is considered controversial in many parts of US society. On the other hand, many older African-Americans wrote in "negro" for description of their race in the 2000 census (which was part of the motivation for including it on the 2010 form). Using answers from the "fill in the blank" option on this question to determine what options to use in future census's was also interesting.
- Interest in census workers (enumerator) going out to count the homeless, and in hiring local people to be enumerators in areas where the government is distrusted.
  - Interest in how much real risk enumerators take in visiting homeless communities (especially those "Mercedes Benz" communities in Los Angeles...) *[Note: Of course these aren't the only homeless communities, and probably not the most dangerous... –BJ]*
  - How accurate are counts in communities that distrust the government?
  - Are enumerators harmed? What happens when trouble finds them?
- Intrigued by the efforts made to count *everyone*.
  - Why do they go to all that trouble to find and count the homeless, anyway?
  - Why do they go to all that trouble to go on mules through Indian reservations, or cross rivers and tundra in Alaska, to make sure every person is counted. This costs a lot of tax dollars!
  - Why do we try to count illegal aliens?

  [Note: it says we must count everyone living in the US, in Article I section 2 of the US Constitution –BJ]
- Intrigued by people who lost everything in the recession except their luxury car or Winnebago camper, and choose to live in their expensive vehicle rather than finding cheaper housing, etc. *[I suspect those people do not have jobs anymore (downsized from the recession) but own their cars free and clear. So, perhaps, getting rid of them would deprive them of the one remaining asset (both financial and emotional I suppose!) that they have. –BJ]*

- Interesting that Census is exploring the use of administrative records (drivers' licenses, voting records, tax forms, medical records, etc) to reduce the cost of the Census in future years.
- Why are dorms considered "group housing" rather than "apartments"?
- It sounds like the possibilities for double-counting dorm residents is still large. Resident assistants shouldn't just encourage everyone to fill out their form, they should also discuss with each resident whether they would be counted on their parents form...
- The varying compliance and response rates across the country (why can't everyone be like Iowa!?) was interesting.
- Intrigued by the enormous efforts to reduce unit nonresponse. Amazed at advertising campaign beforehand, amazed that 700,000 people need to be hired as enumerators to followup on those who did not mail in forms. Really amazed at the costs ($14.7 billion for the 2010 census). Maybe using the web as an alternative to mail-in form would also help?
- The issue of illegal aliens was interesting. On the one hand it appears that the Census knows where they are and they go and try to count them (using local workers to increase illegals' trust and response rate). On the other hand, if they know where they are, why doesn't the immigration service go find them and deport them? Apparently that's what they did relocating Japanese in the US during WWII (whether illegal aliens or not)? *[Strong laws exist—and have been strengthed by both congress and the courts in recent years—making it illegal for Census to share this type of information with other units in the government. The reason is that the Census count must be as accurate as possible, and if illegals (or others at risk of action against them by the government) believe that the Census will not "rat them out" then they are more likely to cooperate, and the census will be more accurate. These laws were violated in the case of Japanese-Americans during WWII –BJ]*
- Before Groves' visit I thought that the Census was not for me, because I am an international student. Now I know that I should fill out the form and send it in also.
- The variation in how (or whether!) a census is done, from one country in the world to the next, was very interesting.
  - Are governments always the ones who run censuses? How many countries haven't done a census in a long time? Should regular censuses be mandated (and enforced) by an international organization like the UN?
  - Interesting that Nordic countries do not run a US-style Census since they have so many records on every resident of the country that they can get the same info from looking at their records.

  - Interesting that countries with large nomadic populations have to adapt census methods to those nomadic ways. For example, everyone has to go to a certain location to be counted, on a "National Census Day", rather than being counted "where they are".

- How does the Census avoid double-counting? Would running an "expert system" on the census data help to identify duplicates? *[Some of the questions on the census form are designed to help detect duplicates, and/or people living in multiple places. In addition there are indeed statistical models and algorithms capable of estimtating the probability of a "match" between two filled-out census forms (or other survey forms); these can be used to detect or count the number of duplicate responses. –BJ]*

- Interesting that US Citizens who are not living in the US are not counted in the Census.

- It was probably wrong in past Censuses for the Census bureau to correct same-sex couples to be either heterosexual couples or not counted as couples. It is better that in the 2010 Census, same-sex couples will be counted as such.

- Intrigued by both (a) the difference between sampling (capture-recapture) adjustments; and "demographic analysis" adjustments to the US census. Strange that sampling is not allowed, but demographic analysis is. Sounds like this is all a political decision on the part of Republicans who would have been harmed by sampling adjustments in the past (since sampling tends to find more people in lower economic classes, who presumably would be more favorable to Democrats).

- In using capture-recapture sampling adjustments, you have to do another survey to get a second "list". But if they work so hard to get everyone into the Census, wouldn't the second, smaller survey be just a subset of the Census (and then capture-recapture wouldn't work)? *[Usually these post-census surveys involve other sources that the Census does not use, such as administrative records, local community records, etc. In addition, just by chance, some people missed in one count will be found in the other. Thus the survey doesn't completely overlap with the census list. –BJ]*

- Fascinated by the idea that the Census would use very tightly targeted marketing techniques to increase the response rate for the mail-in form.

- Would like to hear more about the quantitative work Groves did as a sociologist. *[http://www.psc.isr.umich.edu/people/cv/groves_robert_cv.pdf –BJ]*

- Interesting that Groves was a professor before becoming director of the census, and that he moved freely between statistics and sociology.

- Interesting that Census director is a presidential appointment, and that Groves himself was a controversial choice. *[He was controversial because he advocates capture-recapture sampling to adjust the census. However, for practical reasons* *(he became director after the design for the 2010 census was in place, for example), he could not implement that method for the 2010 census. Moreover, sampling has been struck down as an adjustment method by the Supreme Court, so it seems unlikely that we will see sampling anytime soon for adjusting the main census numbers. –BJ]*

- Intrigued that people are not "punished" for not sending in Census information. Understand that punishment would scare some people off and reduce the accuracy of the count, though.
- How large is the overall "error" associated with the census count? *[Although I don't know the exact answer, I do know that this error is estimated using the "demographic analysis" methods that Prof Fienberg talked about in class while I was away. –BJ]*
- The costs of assembling a "universal address list", to send out census forms to, seems huge. It would be better if the lists were assembled and paid for by local municipalities (who probably know the addresses better anyway).