Homework 1

Alex Schaefer

4/21/2020

1. **Systematic Sampling** Recall, the four necessary properties of a probability sample are as follows: Property 1 is about every individual in the population having a non-zero probability to end up in the sample; Property 2 says that sampling probabilities for every individual are known; Property 3 is about every pair of individuals in the population having a non-zero probability to end up in the sample; Property 4 says that pairwise sampling probabilities are known for every pair of individuals.

- (5 points) Imagine systematic sampling that involves taking a list of the population and choosing, for example, every 100th entry in the list. Which of the necessary properties of a probability sample does this procedure have? Explain.

In this case, it is hard to say whether the properties of a probability sample are present because we do not know whether the list is ordered in some way, and it doesn't appear as though the starting point is random.

Requirement 1: It is possible that the 1st item in the list does not have a chance of getting chosen if every 100 are picked NOT starting starting in a random place, making me question whether this method fills the first property.

Requirement 2: Because we seem to be starting with the top of the list and working our way down, but we don't know the size of the population or of the overall sample, and it isn't mentioned that we are starting in a randomly-selected place, it doesn't seem that we know the sampling probability of every individual unless we know where each person falls in the list (probably of 1 if they are person 100, 200, 300, etc, probability 0 if they are anyone else).

Requirement 3: Unless we sample every observation using this method, or at least go through the population twice, it seems to me that this criterion is not met because two observations right next to each other cannot be sampled.

Requirement 4: I do not think we know the sampling probability of any one observation (unless there is a randomly-assigned starting point and we know that our sample is large enough to get through the list of individuals one time), which means we don't know the pairwise sampling probabilities either.

- (5 points) Imagine systematic sampling with multiple random starts where we choose 5 random starting points) in 1, 2, . . ., 500 and take every 500th entry starting from each of the 5 random points). Which of the necessary properties of a probability sample does this procedure have? Explain.

This procedure appears to satisfy the conditions of each observation having a chance at getting into the sample, as well as every pair having a non-zero chance at being selected, because we are starting at a random place in the list and choosing multiple random starting places, so there is a shot for consecutive individuals to be drawn. In this case, because there is randomness in where the starting points are, it seems that we know that there is a 5/500 chance for every individual in the population to be drawn, and that we also know the pairwise sampling probabilities.

- (5 points) In what situations treating a systematic sample as if it were a simple random sample would give good results? Provide example(s and explain.

If we know that the individuals we are trying to sample are not arranged in a pattern that could affect our results, then the results of this type of procedure could be treated as though they were a SRS. If on the other hand, for example, we are drawing every 100th individual from a list of students ordered first by school (let's say the schools usually have around 100 students in them) and then by class ranking, and then try to make inferences about the grades, our systematic sampling might have yielded biased results because we might have selected students that were always at the high or low ends of their classes.

2. Using the California API data...

- (5 points) Draw a simple random sample of size 350 from the population file and estimate the total enrollment in California schools and its standard error.

```
set.seed(2000)
apipop$nschools <- nrow(apipop)
sample_rows <- sample(1:nrow(apipop), 350)
api_srs <- apipop[sample_rows,]
dim(api_srs)
```

```
## [1] 350  38
```

```
srsdesign <- svydesign(id=~snum, fpc=~nschools, data=api_srs)
```

The survey total and SE from this sample was:

```
svytotal(~enroll,design=srsdesign, na.rm=TRUE)
```

```
##          total     SE
## enroll 4050947 150434
```

- (10 points) Draw a stratified random sample of 200 elementary schools, 100 middle schools, and 50 high schools. Estimate the total enrollment in California schools and its standard error.

```
set.seed(2000)
strat_rows <- stratsample(apipop$stype,
                          counts=c(E=200,
                                   M=100,
                                   H=50))

samp <- apipop[strat_rows,]
dim(samp)
```

```
## [1] 350  38
```

```
summary(samp$stype)
```

```
##   E   H   M
## 200  50 100
```

```r
n <- summary(samp$stype)
summary(apipop$stype)
```

```
##    E    H    M
## 4421  755 1018
```

```r
N <- summary(apipop$stype)
n/N
```

```
##          E          H          M
## 0.04523863 0.06622517 0.09823183
```

```r
N/n
```

```
##      E      H      M
## 22.105 15.100 10.180
```

```r
unique(apistrat$pw)
```

```
## [1] 44.21 20.36 15.10
```

```r
samp$fpc <- samp$weights <- NA

samp$fpc[samp$stype == "E"] <- N[1]
samp$fpc[samp$stype == "H"] <- N[2]
samp$fpc[samp$stype == "M"] <- N[3]

samp$weights[samp$stype == "E"] <- (N/n)[1]
samp$weights[samp$stype == "H"] <- (N/n)[2]
samp$weights[samp$stype == "M"] <- (N/n)[3]

stratdesign <- svydesign(id = ~1,
                         strata = ~stype,
                         weights = ~weights,
                         fpc = ~fpc,
                         data = samp)
stratdesign
```

```
## Stratified Independent Sampling design
## svydesign(id = ~1, strata = ~stype, weights = ~weights, fpc = ~fpc,
##     data = samp)
```

** The survey total and SE from this sample was: **

```r
svytotal(~enroll,design=stratdesign, na.rm=TRUE)
```

```
##          total     SE
## enroll 3956030 103310
```

- (5 points) Why is the stratified estimate more precise?

There are a few reasons that the stratified sample should be more precise. A big one is that the schools within each stratum are more homogenous than the population of schools overall, and using a variance which is lower within each stratum than the population gives us a lower variance and greater precision overall.

3. Using the NHANES blood pressure data on the class website, give estimates and confidence intervals for:

```
nhanes<-read.dta("nhanesbp.dta",
         convert.dates = TRUE,
         convert.factors = TRUE,
         missing.type = FALSE,
         convert.underscore = FALSE,
         warn.missing.labels = TRUE)
#this dataset has 11140 obs of 45 vars.
# https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=Questionnaire

# all relevant variables must exist at the time of specifying survey deisgn, so creating this high/low
nhanes<-mutate(nhanes, sodium= ifelse(DR1TSODI>2400, 1, 0))

summary(nhanes$sodium)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   0.675   1.000   1.000
```

```
design<- svydesign(id=~SDMVPSU,
                   weight=~fouryearwt,
                   strata=~SDMVSTRA,
                   data=nhanes,
                   nest = FALSE,
                   check.strata = FALSE)
design
```

```
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (2) clusters.
## svydesign(id = ~SDMVPSU, weight = ~fouryearwt, strata = ~SDMVSTRA,
##     data = nhanes, nest = FALSE, check.strata = FALSE)
```

- (5 points) the mean and median systolic blood pressure in the population,

The mean systolic blood pressure is:

```
svymean(~BPXSAR, design=design, na.rm=TRUE)
```

```
##          mean     SE
## BPXSAR 122.16 0.3688
```

```r
confint(svymean(~BPXSAR, design=design, na.rm=TRUE))
```

```
##           2.5 %   97.5 %
## BPXSAR 121.4346 122.8802
```

And the median systolic blood pressure is:

```r
svyquantile(~BPXSAR, design, c(.5), na.rm=TRUE)
```

```
##        0.5
## BPXSAR 119
```

```r
confint(svyquantile(~BPXSAR, design, c(.5), na.rm=TRUE, ci=TRUE))
```

```
## Warning in qt(p, df = degf(design), lower.tail = lower.tail): NaNs produced

## Warning in qt(p, df = degf(design), lower.tail = lower.tail): NaNs produced

## Warning in qt(p, df = degf(design), lower.tail = lower.tail): NaNs produced

##            (lower upper)
## 0.5_BPXSAR    NaN    NaN
```

```r
# I am not sure why NAs were produced here! I specified that they should be removed.
```

- (5 points) the number of people in the population with hypertension,

The following is the estimated number of people in the population with hypertension, with CIs below. (BPQ020==1 TRUE)

```r
#yes-1
#no-2
#refused-7
#DK-9

svytotal(~nhanes$BPQ020==1, design=design, na.rm=TRUE)
```

```
##                              total      SE
## nhanes$BPQ020 == 1FALSE 158738469 7813448
## nhanes$BPQ020 == 1TRUE   64769246 3863281
```

```r
confint(svytotal(~nhanes$BPQ020==1, design=design, na.rm=TRUE))
```

```
##                              2.5 %    97.5 %
## nhanes$BPQ020 == 1FALSE 143424392 174052547
## nhanes$BPQ020 == 1TRUE   57197354  72341137
```

- (5 points) the mean blood pressure in men and in women,

The mean blood pressures for men (1) and women (2) are:

```r
svyby(~BPXSAR, ~RIAGENDR,  design=design, na.rm=TRUE, svymean)
```

```
##   RIAGENDR   BPXSAR        se
## 1        1 123.4577 0.4607068
## 2        2 120.9238 0.3725201
```

```r
confint(svyby(~BPXSAR, ~RIAGENDR,  design=design, na.rm=TRUE, svymean))
```

```
##      2.5 %   97.5 %
## 1 122.5547 124.3606
## 2 120.1937 121.6539
```

- (5 points) the mean and median daily sodium intake,

The mean daily sodium intake is:

```r
svymean(~DR1TSODI, design = design, na.rm=TRUE)
```

```
##            mean     SE
## DR1TSODI 3519.4 31.006
```

```r
confint(svymean(~DR1TSODI, design = design, na.rm=TRUE))
```

```
##             2.5 %   97.5 %
## DR1TSODI 3458.637 3580.177
```

The median daily sodium intake is:

```r
svyquantile(~DR1TSODI, design, c(.5), na.rm=TRUE)
```

```
##           0.5
## DR1TSODI 3200
```

```r
confint(svyquantile(~DR1TSODI, design, c(.5), na.rm=TRUE, ci=TRUE))
```

```
## Warning in qt(p, df = degf(design), lower.tail = lower.tail): NaNs produced

## Warning in qt(p, df = degf(design), lower.tail = lower.tail): NaNs produced

## Warning in qt(p, df = degf(design), lower.tail = lower.tail): NaNs produced

##              (lower upper)
## 0.5_DR1TSODI    NaN    NaN
```

- (5 points) the proportion of the population that exceeds 2.4g/day sodium intake.

The proportion of the population that axceeds 2.4g/day sodium is:

```r
svymean(~sodium, design = design, na.rm=TRUE)
```

```
##           mean    SE
## sodium 0.70463 0.009
```

```r
svyciprop(~I(sodium==1), design = design, na.rm=TRUE)
```

```
## Warning in qt((1 - level)/2, df = ddf): NaNs produced
```

```
##                       2.5% 97.5%
## I(sodium == 1) 0.705   NA    NA
```

4. Suppose you are interested in estimating job loss among restaurant workers during the COVID-19 pandemic in Seattle. Assume you have a list of restaurants and names and phone number of their employees from March 1, 2020. You would like to conduct telephone interview with individual restaurant workers to collect self-reported data on whether they have lost their restaurant job in the three months of March, April, and May. For your survey sample:

- (5 points) Would cluster sampling be useful in this survey? If so, explain why and describe the clusters; if not, explain why.

It seems that cluster sampling would not be well-applied here, because selecting clusters in terms of restaurant would likely lead to all or most employees within a restaurant giving similar responses (i.e., the whole restaurant either stayed afloat or laid off employees). Additionally, since we are doing a phone interview, there is no increased cost associated with using a method with greater precision, including SRS and stratified sampling.

- (5 points) What strata would you recommend? Would you recommend sampling some strata with higher probability?

I would want to know the research question before deciding on the strata and it would depend on what information we have about the restaurants/individuals. If we wanted to see what kinds of restaurants are most affected, I might stratify by restaurant type and price point. This would mean, for example, selecting from strata that represent low, medium, and high price points and each of a variety of food types (Thai, Indian, American, etc). Unless there are some strata that are very small or large or of a special interest, I would probably be okay with equal probabilities.

5. **Emergency preparedness survey** You are conducting a survey of emergency preparedness at a large Health Management Organization. One of the goals is to estimate what proportion of the medical staff (physicians and others that would be able to get to work in case of public transport shutdown.

- (5 points) Does this goal correspond to a population or a process inference? Why?

My assumption is that, because we are looking for a proportion of all staff (but not necessarily at how their experiences differ), that this would be a population inference. If we were trying to see whether doctors and nurses face different barriers to getting to work without public transit, I would think it were a process inference instead.

- (5 points) You can either send out a single questionnaire to all staff, or send out a questionnaire to about 10% of the staff and make follow-up phone calls for questionnaires that are not returned. What are disadvantages of each approach?

An advantage of the 10% sample with follow-up is that, assuming the follow-up is effective, it may produce a less-biased sample because the people who are too busy or careless or antisocial to respond to surveys will still be represented. However, the disadvantage is that the sample size will be small. The all-staff questionnaire may have the advantage of a larger sample size (and therefore smaller confidence intervals etc), but is more likely to be biased by nonresponse.

- (5 points) You choose to survey just a sample, and choose to stratify with just two strata: physicians and other staff. The HMO has 900 physicians and 9000 other staff. You sample 450 physicians and 450 other staff. What are the sampling probabilities in each stratum?

The sampling probability for physicians is 450/900, or 0.5, but the probability for other staff is 450/9000, or 0.05.

- (5 points) 350 physicians and 150 other staff say they would be able to get to work in case of public transport shutdown. Give unbiased estimates of the proportion in each stratum and the total proportion of staff who would be able to get to work in case of public transport shutdown.

The estimated proportion of doctors who could not make it to work is 0.78 and the estimated proportion of other staff who could not make it to work is 0.33. The total proportion of staff who might not make it to work is 0.37.

- (5 points) How would you explain to the managers that commissioned the study how the estimate was computed and why it wasn't just the number who said "yes" divided by the total number surveyed?

I would explain to the manager that each person who answered the survey counts for 20 other people and each doctor who answered the survey counts for only 2 other people, so a weighted proportion is needed to reflect that reality. I would then point out that it makes sense that the weighted population proportion estimate is closer to the "other staff" value than the doctor value.