BST5420 Sampling Theory and Survey Design Homework 3

Miao Cai[*]

2019-04-02

Due April 2, 2019

This is worth 51 points, and the recorded score will be the proportion correct times 10 (rounded to the nearest half point) in order to be on a 10 point scale.

1. (12 Points) A researcher developed a test designed to measure the degree of awareness of current public health events. She wants to estimate the average score that would be achieved on this test by all students in a certain high school. The administration at the school will not allow the experimenter to randomly select students from classes in session, but it will allow her to interrupt a small number of classes for the purpose of giving the test to every member of the class. Thus, the experimenter selects 25 classes at random from the 108 classes in session at a particular hour. The test is given to each member of the sampled classes, with results as shown in the table below. Estimate the mean score that would be achieved on the test by all students in the school. Also, give an approximate 95% confidence interval for the mean.

```
n_stu = c(31, 29, 25, 35, 15, 31, 22, 27, 25, 19, 30, 18, 21,
          40, 38, 28, 17, 22, 41, 32, 35, 19, 29, 18, 31)
scores = c(1590, 1510, 1490, 1610, 800, 1720, 1310, 1427, 1290,
           860, 1620, 710, 1140, 1980, 1990, 1420, 900, 1080,
           2010, 1740, 1750, 890, 1470, 910, 1740)
```

$$\hat{\mu} = \frac{\hat{\tau}}{NM}$$
$$\hat{\tau} = \frac{N}{n}\sum_{i\in S}\tau_i$$

$$\hat{V}(\hat{\mu}) = \frac{1-\frac{n}{N}}{n\bar{M}^2}\frac{1}{n-1}\sum_{i\in S}M_i^2(\bar{y_i}-\hat{\mu})^2$$

```
N = 108
n = length(scores)
y_bar = scores/n_stu
M_bar = sum(n_stu)/n
tau_hat = N/n*sum(scores)
mu_hat = tau_hat/(N*M_bar)
sr2 = sum(n_stu^2*(y_bar - mu_hat)^2)
V_mu = (1-n/N)*sr2/(n*(n-1)*M_bar^2)
MOE = 2*sqrt(V_mu)
```

The mean is 51.56 and the 95% confidence intervals is (50.21, 52.9).

[*]PhD student, Department of Epidemiology and Biostatistics, College for Public Health and Social Justice, Saint Louis University. Email address miao.cai@slu.edu

2. (21 Points) Read the paper: *Thorpe, Lorna E., et al. "Childhood obesity in New York City elementary school students." American Journal of Public Health 94.9 (2004): 1496-1500.* This can be obtained by searching "thorpe childhood obesity new york" at scholar.google.com.

The authors say "This height-and-weight survey was conducted with a stratified, multistage, probability sample of elementary public school children in New York City."

1. Describe what "stratification" means in this context. What are the strata?
2. Describe what multistage means in this context. Within each stratum, what is the design of the sampling plan?
3. Describe the levels of the sampling design from part (b). (For example, "four-stage cluster sample where a sample of BLANK psus is selected from BLANK; then from each BLANK a sample of BLANK from ...")
4. Describe what probability sample means in this context.
5. Describe one of the response variables in this study.
6. What is the population in this study?
7. Describe one parameter that is estimated (e.g., mean, total, proportion). Give its point estimate and margin of error.

1. There are three strata in this study: schools, grades, classes.
2. Multistage means to sample units in the order of the three strata.
3. Three-stage cluster sample were conducted:

- First, randomly sample 70 schools from all 736 schools,
- Second, systematically sample two grades from each of the 70 sampled schools,
- Third, randomly sample 2 classes from the sample schools and grades.

4. Probability sample here means some schools, grades, or classes refuse to respond, so there are issues with probability of responding, which is needed to be accounted for.
5. Height of the respondent (student).
6. The population in this study is all the students in the 736 elementary schools (all New York City public schools, excluding special education schools).
7. The proportion of primary students who were overweight ($\geq$ 85th percentile BMI) is estimated to be 43.2%, and the confidence interval is (39.3%, 47.3%).

3. (12 Points) Determine the sampling weights for the following sampling plans.

1. A simple random sample of size n from a population of size N.
2. A stratified random sample where for stratum $h(h = 1, 2, \ldots, H)$, n_h units are selected from the N_h items in.
3. A one-stage cluster sample with notation given on p. 169, where each cluster is selected with equal probability.
4. A one-stage cluster sample where the probability of selecting a cluster is proportional to the cluster size, and clusters are selected with replacement.

1. N/n
2. $\frac{1}{\frac{1}{H} * \frac{n_h}{N_h}} = \frac{H * N_h}{n_h}$
3. $\frac{1}{\frac{1}{N}} = N$
4. $1/\left(\frac{M_i}{\sum_{i=1}^{N} M_i}\right) = \frac{\sum_{i=1}^{N} M_i}{M_i}$

4. (6 Points) Read problem 2 on pp. 267-268.

1. Don't work out the problem as stated in the book, but rather, use R to select a sample of size 10 with replacement using the probabilities ψ_i. Show your R code.
2. Ignoring the probabilities ψ_i use R to select a sample of size 10 with replacement. Show your R code.

1.

```
psu = 1:25
psi = c(0.000110, 0.018556, 0.062998, 0.078216, 0.075245,
        0.073983, 0.076580, 0.038981, 0.040772, 0.022876,
        0.003721, 0.024917, 0.040654, 0.014804, 0.005577,
        0.070784, 0.069635, 0.034650, 0.069492, 0.036590,
        0.033853, 0.016959, 0.009066, 0.021795, 0.059186)
set.seed(123)
sample(psu, 10, TRUE, psi)
```

```
##  [1]  6 20 17 12  2  4  3 10  3 19
```

2.

```
set.seed(123)
sample(psu, 10, TRUE)
```

```
##  [1]  8 20 11 23 24  2 14 23 14 12
```