Exercise ark #6.
Survey nonresponse

Oğuz–Alper, Melike & Pekarskaya, Tatsiana, Statistics Norway

October 15, 2020

Exercise 1

We look at the Norwegian election survey from 1993. The sample consists of 3000 persons. 11 callbacks were used. The sample of 3000 is assumed to be a random sample. We shall use the data after two callbacks. The number of responses were 1403.

1. Of the 1403 persons, 1190 said they voted in the Parliament election 1993. Assume that the nonresponse is MCAR and compute an estimate and a 95% confidence interval for the proportion of voters in the population.

2. The true voting proportion was 0.755. Compare the estimate and confidence interval from (1) with the true proportion. What can you say about the MCAR assumption? To try to correct the bias estimation in part (1) we shall poststratify according to voting participation in the Parliament election in 1989. We use 3 groups:

   - Group 1: participating in the 1989 election: $N_1 = 2\ 510\ 669$
   - Group 2: not participating in the 1989 election: $N_2 = 508\ 288$
   - Group 3: new voters: $N_3 = 241\ 000$

   In the response sample we have the following result, where y = 1 indicate voting in 1993 election and y = 0 indicate not voting in 1993.

Stratum	1		2		3	
y	0	1	0	1	0	1
# persons	132	1060	58	57	23	73
Total	1192		115		96	

3. Find the post-stratified estimate for voting proportion in 1993 and compare with the estimate in part (1) and the true value 0.755.

4. Under which condition is post-stratified estimator is approximately unbiased? Is that the case in (3)? If not, how would you cope with this problem?

5. Assume that we got an additional information: nonresponses are distributed between the 3 groups as follows:

   - Group 1: 850

- Group 2: 550
- Group 3: 197

and they have the same voting proportion as in the response sample (the same voting proportion as in (3)).
Assume that there were no nonresponses and calculate weighting class adjustment estimate for the proportion of voters in this case.

6. Under which condition weighting class adjustment estimator(4) can help largely reduce nonresponse bias? (Bjørnstad, 2018)

Exercise 2

(R code available) We shall estimate the mean income in a large population and take a random sample of n = 20 persons. 10 persons responded with the following income (in 1000): 600, 520, 620, 500, 380, 460, 450, 250, 400 and 780. We assume MCAR nonresponse.

1. Use R to perform a hot-deck imputation for the nonresponse. Derive the standard 95% confidence interval of the mean income in the population, based on the completed data set with observed and the imputed values.

2. Use R to derive the standard 95% confidence interval of the mean income in the population, based on the response sample. (Bjørnstad, 2018)

Exercise 3

Investigators selected an SRS of 200 high school seniors from a population of 2 000 for a survey of television-viewing habits, with an overall response rate of 75%. By checking school records, they were able to find the grade point average for the nonrespondents, and classify the sample accordingly:

GPA	Sample size	Number of respondents	Hours of TV $\bar{y}_{cR}$	s_{cR}
3.00–4.00	75	66	32	15
2.00–2.99	72	58	41	19
Below 2.00	53	26	54	25
Total	200	150		

1. What is the estimate for the average number of hours of TV watched per week if only respondents are analyzed? What is the standard error of the estimate?

2. Perform a χ^2 test for the null hypothesis that the three GPA groups have the same response rates. What do you conclude? What do your results say about the type of missing data: Do you think the data are MCAR? MAR? Nonignorable?

3. Perform a one-way ANOVA analysis to test the null hypothesis that the three GPA groups have the same mean level of television viewing. What do you conclude? Does your ANOVA analysis indicate that GPA would be a good variable for constructing weighting cells? Why, or why not?

4. Use the GPA classification to adjust the weights of the respondents in the sample. What is the weighting class estimate of the average viewing time?

5. The population counts are 700 students with GPA between 3 and 4; 800 students with GPA between 2 and 3; and 500 students with GPA less than 2. Use these population counts to construct a poststratified estimate of the mean viewing time. (Lohr, 2019, p.358)

Exercise 4

(R code available) The American Statistical Association (ASA) studied whether it should offer a certification designation for its members, so that statisticians meeting the qualifications could be designated as "Certified Statisticians." In 1994, the ASA surveyed its membership about this issue, with data in file certify.dat. The survey was sent to all 18 609 members; 5 001 responses were obtained. Results from the survey were reported in the October 1994 issue of Amstat News.

Assume that in 1994, the ASA membership had the following characteristics: 55% have PhD's and 38% have Master's degrees; 29% work in industry, 34% work in academia, and 11% work in government. The cross-classification between education and workplace was unavailable.

1. What are the response rates for the various subclasses of ASA membership? Are the nonrespondents MCAR? Do you think they are MAR?

2. Use raking to adjust the weights for the six cells defined by education (PhD or non-PhD) and workplace (industry, academia, or other). Start with an initial weight of $18\,609/5\,001$ for each respondent. What assumptions must you make to use raking?

3. Can you conclude from this survey that a majority of the ASA membership opposed certification in 1994? Why, or why not? (Lohr, 2019, pp.359-360)