Solution ark #6.
Survey nonresponse

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

October 28, 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.

   Solution: With MCAR, the response sample is a random sample
   $\hat{p} = 1190/1403 = 0.848, \quad \widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\frac{N-n}{N}} = 0.00959$
   95% confidence interval:$\hat{p} \pm 1.96\widehat{SE}(\hat{p}) = 0.848 \pm 0.019 = [0.829; 0.867]$

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?

   Solution: The estimate and CI have large bias from the "true" value. Nonresponse can be ignored if $\bar{y}_r$ (sample mean for the respondents) approximately unbiased for the population mean. It does not hold in our case, thus, the nonresponse is not MCAR.

   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.

Group	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.

Solution:

$$\hat{p}_{pst} = \frac{1}{N}\hat{t}_{pst} = \sum_{h=1}^{3} \frac{N_h}{N}\bar{y}_{hR}$$

$\bar{y}_{1R} = 1060/1192 = 0.8893$
$\bar{y}_{2R} = 57/115 = 0.4957$
$\bar{y}_{3R} = 73/96 = 0.7604$
$\Longrightarrow \hat{t}_{pst} = 2667953$ and $\hat{p}_{pst} = 2667953/3259957 = 0.818$
The bias was reduced compared to (1), however the estimator is still biased.

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

Solution: "The post-stratified estimator approximately unbiased if within a post-stratum:

- the response variable y_i constant
- the response propensities ϕ_i the same for every unit
- the response y_i uncorrelated with the response propensity ϕ_i"(Lecture notes).

It is not working in our case. To cope the problem we need to use extra variables to construct the better classes to reduce the bias.

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.

Solution: Number of sampling units in the groups will be now:

- Group 1: $132 + 1060 + 850 = 2042$
- Group 2: $115 + 550 = 665$
- Group 3: $96 + 197 = 293$

Weighting class adjustment estimate for the proportion of voters will be:

$$\hat{p}_{wc} = \sum_{h=1}^{3} \frac{n_h}{n}\bar{y}_{hR}$$

$\hat{p}_{wc} = (2042 * 0.8893 + 665 * 0.4957 + 293 * 0.7604)/3000 = 2368/3000 = 0.789.$

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

Solution: "The models for weighting adjustments for nonresponse are strong: In each weighting cell/poststratum, the respondents and nonrespondents are assumed to be similar, or each individual in a weighting class is assumed equally likely to respond to the survey or have a response propensity that is uncorrelated with y" (Lohr, 2019, p.342). "The weighting class adjustment estimator approximately unbiased if within a class:

- the response variable y_i constant
- the response propensities ϕ_i the same for every unit
- the response y_i uncorrelated with the response propensity ϕ_i"(Lecture notes).

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(missing completely at random) 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.

Solution:
Hot-deck imputation: imputation made using the same dataset.
For example, we take an SRSWR of size 10 from the known responses. With a seed value in R = 2020 we impute 450, 460, 250, 600, 600, 500, 780, 460, 600, 250 so that we get values for all 20 persons
Average income than

$$\bar{y}_{imp} = \frac{1}{n}\left(\sum_{i\in\overline{1;10}} y_i + \sum_{i\in\overline{11;20}} y_{imp;i}\right),$$

where y_i - responded values, $y_{imp;i}$ - imputed values.
$\bar{y}_{imp} = 495.5$

$$\widehat{SE}(\bar{y}_{imp}) = (1-f)\frac{s^2}{n}$$

.
Since population is large $f \to 0$, then we obtain $\widehat{SE}(\bar{y}_{imp}) = 33.75$

$$CI_{imp} = \bar{y}_{imp} \pm 1.96 * \widehat{SE}(\bar{y}_{imp}) = [429.35; 561.65]$$

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

Solution:
Making the same calculations but for only the response sample without imputation we get:
$\bar{y} = 496, \quad \widehat{SE}(\bar{y}) = 46.43 \quad CI_{resp} = [405.00; 587.01]$
Comparing thee two intervals we can see that CI_{imp} is much shorter than CI_{resp}

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:

	Sample	Number of	Hours of TV	
GPA	size	respondents	$\bar{y}_{cR}$	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?

Solution: Find $\bar{y}_R$ using only the respondent set and provide a standard error of your estimate.

$$\bar{y}_R = \frac{1}{n_R}\sum_{c=1}^{3} n_{cR}\bar{y}_{cR} = \frac{1}{150}[66(32) + 58(41) + 26(54)] = 39.3\cdot$$

$$\widehat{SE}(\bar{y}_R) = \sqrt{\left(1 - \frac{150}{2\,000}\right)\frac{s_R^2}{150}}, \quad s_R^2 = \frac{1}{n_R - 1}\left(\sum_{c=1}^{3}\sum_{i=1}^{n_{cR}} y_i^2 - n_R\bar{y}_R^2\right),$$

$$\begin{aligned}\sum_{c=1}^{3}\sum_{i=1}^{n_{cR}} y_i^2 &= \sum_{c=1}^{3}(n_{cR} - 1)s_{cR}^2 + \sum_{c=1}^{3} n_{cR}\bar{y}_{cR}^2 \\ &= [65(15^2) + 57(19^2) + 25(25^2)] + [66(32^2) + 58(41^2) + 26(54^2)] = 291\,725\cdot\end{aligned}$$

- $s_R^2 = \frac{1}{149}[291\,725 - 150(39.3^2)] = 403.6$
- $\widehat{SE}(\bar{y}_R) = 1.58$

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?

Solution: Perform a χ^2 test for the null hypothesis that the three GPA groups have the same response rates, that is, $H_0 : \phi_c = 0.75$.

GPA	Respondents	Non respondents	Total
3.00-4.00	66	9	75
2.00-2.99	58	14	72
Below 2.00	26	27	53
Total	150	50	200

$$\begin{aligned}
\chi^2 &= \sum_c \frac{(\text{observed}_c - \text{expected}_c)^2}{\text{expected}_c} \\
&= \frac{[66 - 0.75(75)]^2}{0.75(75)} + \frac{[9 - 0.25(75)]^2}{0.25(75)} + \cdots + \frac{[27 - 0.25(53)]^2}{0.25(53)} \\
&= 1.69 + 5.07 + 0.30 + 0.89 + 4.76 + 14.27 = 26.97\cdot
\end{aligned}$$

The χ^2 text statistics is 26.97. The p-value for the area to the right of $\chi^2 = 26.97$ under a χ^2 distribution with degree-of-freedom df $= 2$ is $1.4 \times 10^{-6} << \alpha = 0.05$. Thus, there is a strong evidence against the null hypothesis that the three groups have the same response rates. The hypothesis test indicates that the nonresponse is not MCAR, because response rates appear to be related to GPA. We do not know whether the nonresponse is MAR, or whether is it 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?

Solution: The ANOVA table is given as follows:

Source	df	Sum of squares, estimated	Mean square, est.
Between groups	$C - 1$	$\widehat{SSB} = \sum_{c=1}^{3}\sum_{i=1}^{n_{cR}}(\bar{y}_{cR} - \bar{y}_R)^2$	$\widehat{MSB}$
Within groups	$n_R - C$	$\widehat{SSW} = \sum_{c=1}^{3}\sum_{i=1}^{n_{cR}}(y_i - \bar{y}_{cR})^2$	$\widehat{MSW}$
Total, about mean	$n_R - 1$	$\widehat{SST} = \sum_{c=1}^{3}\sum_{i=1}^{n_{cR}}(y_i - \bar{y}_R)^2$	$\widehat{MST} = s_R^2$

$$\widehat{SSB} = \sum_{c=1}^{3}\sum_{i=1}^{n_{cR}}(\bar{y}_{cR} - \bar{y}_R)^2 = \sum_{c=1}^{3} n_{cR}(\bar{y}_{cR} - \bar{y}_R)^2 = 9\,303.1$$

$$\widehat{MST} = \frac{1}{n_R - 1}\sum_{c=1}^{3}\sum_{i=1}^{n_{cR}}(y_i - \bar{y}_R)^2 = s_R^2 = 403.6$$

$$\widehat{SSW} = 149 s_R^2 - 9\,303.1 = 50\,833.3, \quad \widehat{MSW} = \frac{50\,833.3}{147} = 345.8$$

Source	df	$\hat{SS}$	$\hat{MS}$	F	p-value
Between groups	2	9 303.1	4 651.5	13.4	4.5×10^{-6}
Within groups	147	50 833.3	345.8		
Total, about mean	149	60 136.4			

There is a strong evidence against the null hypothesis that the three GPA groups have the same mean level of TV viewing. Based on the F-test here and the χ^2 test in (2), both the nonresponse rate and the TV viewing seems to be related to GPA. Therefore, it would be reasonable to use GPA groups for weighting class adjustment, or poststratification.

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?

Solution: Use the GPA classification for weighting class adjustment, and find $\bar{y}_{wc}$.

$$\bar{y}_{wc} = \frac{\sum_{c=1}^{3} \sum_{i=1}^{n_{cR}} d_i \hat{\phi}_c^{-1} y_i}{\sum_{c=1}^{3} \sum_{i=1}^{n_{cR}} d_i \hat{\phi}_c^{-1}} = \frac{\sum_{c=1}^{3} n_{cR} \hat{\phi}_c^{-1} \bar{y}_{cR}}{\sum_{c=1}^{3} n_{cR} \hat{\phi}_c^{-1}},$$

where $d_i = 2\,000/200 = 10$ and $\hat{\phi}_c = n_{cR}/n_c$.

$$\bar{y}_{wc} = \frac{66(75/66)32 + 58(72/58)41 + 26(53/26)54}{66(75/66) + 58(72/58) + 26(53/26)} = \frac{8\,214}{200} = 41.07\cdot$$

Here, the weights after the weighting class adjustment are given by $w_{i;wc} = d_i \hat{\phi}^{-1} = 10 n_c / n_{cR}$, $i \in S_{cR}$. We have $w_{i;wc} = 11.36$, $w_{i;wc} = 12.41$, and $w_{i;wc} = 20.38$, for the units in classes $c = 1$, $c = 2$, and $c = 3$, respectively.

The estimated average time of TV viewing is higher after the weighting class adjustment, that is, $\bar{y}_{wc} = 41.07 > \bar{y}_R = 39.3$. This is due to that $\bar{y}_{cR}$ is the highest for the class where the nonresponse rate is the highest.

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.

Solution: Let $N_1 = 700$, $N_2 = 800$, and $N_3 = 500$. Use a poststratified estimator to estimate $\bar{Y}$.

$$\bar{y}_{post} = \frac{\sum_{h=1}^{3} \sum_{i=1}^{n_{hR}} d_i \hat{\phi}_h^{-1} y_i}{\sum_{h=1}^{3} \sum_{i=1}^{n_{hR}} d_i \hat{\phi}_h^{-1}} = \frac{\sum_{h=1}^{3} N_h \bar{y}_{hR}}{\sum_{h=1}^{3} N_h},$$

where $d_i = 2\,000/200 = 10$, and

$$\hat{\phi}_h = \frac{\hat{N}_{hR}}{N_h} = \frac{\sum_{i=1}^{S_{hR}} d_i}{N_h} = \frac{n_{hR}(N/n)}{N_h}\cdot$$

.

$$\bar{y}_{post} = \frac{700(32) + 800(41) + 500(54)}{700 + 800 + 500} = \frac{82\,200}{2\,000} = 41.1\cdot$$

Here, the poststratified weights are given by

$$w_{i;post} = d_i \hat{\phi}_h^{-1} = d_i \frac{N_h}{n_{hR}(N/n)} = \frac{N_h}{n_{hR}},\ i \in S_{hR},$$

We have $w_{i;post} = 10.61$, $w_{i;post} = 13.79$, and $w_{i;post} = 19.23$, for the units in poststrata $h = 1$, $h = 2$, and $h = 3$, respectively.

Note that $\sum_{i \in S_{hR}} w_{i;post} = N_h$, that is, $66(10.61) = 700$, $58(13.79) = 800$, and $26(19.23) = 500$.

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? ASA survey of certification designation for its members. $n = N = 18\,609$ and $n_R = 5\,001$.

Solution:

	Percentage of ASA members in 1994 (p_c)
PhD	0.55
Master's	0.38
Other/unknown	0.07
Industry	0.29
Academia	0.34
Government	0.11
Other/unknown	0.26

Find response rates for the subclasses of ASA membership.

	n_c	n_{cR}	$\phi_c\,(\%)$
PhD	10 235	3 036	30
Master's	7 071	1 640	23
Other/Unknown	1 303	325	25
Industry	5 397	1 809	34
Academia	6 327	2 221	35
Government	2 047	880	43
Other/Unknown	4 838	91	2

Here, $n_c = 18\,609 p_c$, $\phi_c = n_{cR}/n_c$, and n_{cR} are obtained from the survey data.
The response rates are quite low. The nonresponse does not appear to be MCAR, as it differs by degree and by type of employment. It may not be sure that it is MAR either.

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?

Solution: The number of respondents by cells is given as follows:

	Industry	Academia	Other	
PhD	798	1 787	451	3 036
non-PhD	1 011	434	520	1 965
	1 809	2 221	971	5 001

Using $w_i = 18\,609/5\,001 = 3.72$ as initial weights, we obtain initial estimates of population totals by cells as follows:

	Industry	Academia	Other	
PhD	2 969.4	6 649.5	1 678.2	11 297.1
non-PhD	3 762.0	1 614.9	1 934.9	7 311.9
	6 731.4	8 264.5	3 613.1	18 609.0

Adjusting rows by the population counts 10 235 and 8 374 for PhD and non-PhD, respectively, we obtain

	Industry	Academia	Other	
PhD	2 690.2	6 024.4	1 520.4	10 235
non-PhD	4 308.5	1 849.5	2216.0	8 374
	6 998.7	7 873.9	3 736.4	18 609

Now adjusting columns by the population counts 5 397, 6 327, and 6 885 for industry, academia and other, respectively, we obtain

	Industry	Academia	Other	
PhD	2 074.6	4 840.8	2 801.6	9 717.0
non-PhD	3 322.4	1 486.2	4 083.4	8 892.0
	5 397.0	6 327.0	6 885.0	18 609.0

After iterative raking, the final cell estimates are given by

	Industry	Academia	Other	
PhD	2 239.2	4 980.6	3 015.2	10 235.0
non-PhD	3 157.8	1 346.4	3 869.8	8 374.0
	5 397.0	6 327.0	6 885.0	18 609.0

Final weights by cells are given as follows:

	Industry	Academia	Other
PhD	2.81	2.79	6.69
non-PhD	3.12	3.10	7.44

- The weights in the "Other" employment category are dramatically increased by raking.
- The raking weights are used to estimate population quantities.
- For raking, we assume that the response probabilities depend only on the marginal counts, not on the cell counts. Given marginal counts, the response y_i is uncorrelated with the response propensity ϕ_i.

3. Can you conclude from this survey that a majority of the ASA membership opposed certification in 1994? Why, or why not?

Solution:

	Without weights	With raking weights
No response	0.2	0.3
Yes	26.4	25.8
Possibly	22.3	22.3
No opinion	5.4	5.4
Unlikely	6.7	6.9
No	39.0	39.3

- We can not rely on the conclusion that the majority of the ASA members are against certification as the nonresponse rates are very high.