IV. Missing data-nonresponse

- Occurs in almost all surveys, even “compulsory” ones
  - Labor Force survey in Norway, quarterly, 20% nonresponse
- Perceived to have increased in recent years
- Besides sampling error, the most important source of error in sample survey
- Nonresponse is important to consider because of
  - **Bias** (will almost always result in bias)
  - Increased uncertainty in the estimates
  - Increased cost

- Nonresponse is the failure to obtain complete observations on the survey sample
- Unit nonresponse: unit (person or household) in the sample does not respond
  - Can be very high proportion, can be as much as 70% in postal surveys
  - 30% is not uncommon in telephone surveys
  - 50% in the Norwegian Consumer Expenditure survey, up from about 30% twenty years ago
- Item nonresponse: observations on some items are missing for unit in sample
- Standard treatment: *Weighting* for unit nonresponse, *imputation* for item nonresponse

Sources of unit nonresponse

- Non-contact: failure to locate/identify sample unit or to contact sample unit
- Refusal: sample unit refuses to participate
- Inability to respond: sample unit unable to participate, e.g. due to ill health, language problems
- Other: e.g. accidental loss of data/questionnaire

Sources of item nonresponse

- Respondent:
  - answer not known
  - refusal (sensitive or irrelevant question)
  - accidental skip
- Interviewer:
  - does not ask the question
  - does not record response
- Processing
  - (Illogical) response rejected at editing
- Amounts
  - some variables only 1-2%
  - Often highest for financial variables, e.g. total household income may have 20% missing data

Missing data mechanisms

- Basic question about missing data mechanism (response mechanism): Does the probability that data are missing depend on observed and/or unobserved data values ?
- Different models for analysis with missing data rely on different assumptions about the missing data mechanism
- Let for a unit (person or household) in the population:
  - Y : the study variable with value y
  - x : the values of the auxiliary variables
  - $R = 1$ if unit responds when selected in the sample and $R= 0$ if nonresponse
- Assume the units respond independently of each other

3 types of missing data mechanism

- *MCAR. Missing completely at random*. Probability of nonresponse is independent of *Y* and *x*
  - $P(R = 0| y,x) = P(R = 0)$
  - The observed values of *Y* form a random subsample of the sampled values of *Y*
- *MAR. Missing at random.* Probability of nonresponse depends on *x*, but not on *Y*.
  - $P(R = 0| y,x) = P(R = 0|x)$
  - The observed values of *Y* form random samples within subclasses defined by *x*
- *MNAR. Missing not at random*. Probability of nonresponse depends on *Y* and possibly *x* as well.

In this case the response mechanism is *nonignorable.*

General definition of missing data mechanisms

- Suppose we have p Y variables with values for the whole population denoted by $\mathbf{y}$
- Let $\mathbf{x}$ be the values of auxiliary variables known for the whole population.
- Let $\mathbf{y}_{\text{obs}}$ be the observed values of $\mathbf{y}$, $\mathbf{y}_{\text{unob}}$ be the y-values in the sample that are unobserved including missing values and Y values outside the sample.
- Let $\boldsymbol{R}$ be the set of all response indicators for all p Y variables in the sample.
- MCAR: $P(\boldsymbol{R} = \boldsymbol{r} | \mathbf{y}, \mathbf{x}) = P(\boldsymbol{R} = \boldsymbol{r})$
- MAR: $P(\boldsymbol{R} = \boldsymbol{r} | \mathbf{y}, \mathbf{x}) = P(\boldsymbol{R} = \boldsymbol{r} | \mathbf{x}, \mathbf{y}_{\text{obs}})$
- MNAR: $P(\boldsymbol{R} = \boldsymbol{r} | \mathbf{y}, \mathbf{x}) = P(\boldsymbol{R} = \boldsymbol{r} | \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{unob}}, \mathbf{x})$

- The response rate is the most widely reported quality indicator; does not fully capture the potential bias.
- Three examples to illustrate how nonresponse can lead to very misleading statistical analysis, even when the response rate is high.
  - In all cases: MNAR response mechanism
- In two of the examples: How to correct for nonresponse

1. Classical example, response rates 81-85%

- Political polling before the American presidential election in 1948
  - Democratic candidate: Truman
  - Republican candidate: Dewey
  - Instute: Roper
  - Surveys : July, August, September, October
  - Election: November

	July	August	Sept	Oct	Election
Truman	37.8	37.0	35.2	40.4	**49**
Dewey	55.5	52.4	57.0	53.4	**45**
Others	6.7	10.5	7.7	6.2	6
Sample size	3011	3490	3490	3500	
responses	2510	2951	2936	2841	
Nonresponse (Percentage)	501 (18.6)	539 (15.4)	554 (15.9)	659 (18.8)	

- Bias: Larger nonresponse rate among the economically poorer groups
- Compensating for nonresponse – MNAR model:
  - the probability of response dependent on which candidate the person will vote for, within in each socio-economic group
- Gives Truman 51%
- Method: Imputation, estimate 93-99% will vote for Truman in the nonresponse group
- MAR model on socio-economic groups:
  - estimate = 41%

2. Election survey in Norway 2009

- Sample: 2944 persons
- Number of responses: 1782
- Estimate the voting proportion
- Of the 1782, 1506 said they voted in Parliament election: 84.5%
- Margin of error: 1.7%
- True voting proportion = 76.4%
- Estimate 84.5% is biased because higher nonresponse rate among nonvoters. The response mechanism is MNAR
- The response sample is not like the nonresponse group (typically the case)

$$\text{Margin of error}: 2 \cdot SE = 2\sqrt{\frac{0.845 \cdot 0.155}{1782}} = 2 \cdot 0.00857 \; = 0.017$$

3. Estimation of the number of households in Norway in 1992

- Data from the Consumer Expenditure survey in 1992
- Sample: 1698 persons age 15+, self-weighting
- Estimation of the number of one-person households and the total number of households
- Norway has a register of families, know the family size for each person

Fam size	Household size						Non-response	
	1	2	3	4	5+	Total	Nr.	%
1	83	48	20	9	2	162	153	**48.6**
2	9	177	37	4	3	230	160	**41.0**
3	10	25	131	40	6	212	91	**30.0**
4	2	13	37	231	17	300	123	**29.1**
5+	1	4	4	17	181	207	60	**22.5**
Total	**105**	267	229	301	209	1111	587	**34.6**

Size of population as of 1.1.93: N = 4 131 874

Standard estimate for the number of one-person households:

$$\frac{105}{1111} \cdot 4{,}131{,}874 = 390\ 501$$

A post survey of the census from 1990: 626 000

Underestimates “enormously” the number of one-person households

Because: Nonresponse among one-person families is much higher then for larger family size

Correcting for nonresponse

1. Response model-MNAR: Probability of response depends on household size and place of residence (rural or urban)

   - MAR model on family size removes only about 50% of the bias- see later

2. Population model : Probability of household size depends only on family size

3. Use this model to derive

P(Household size =1|family size =x) for x=1,2,3,4,5+,

and estimate these probabilities

4. Estimated number of one-person households = function of these estimated probabilities

- The table on p.14 gives you estimated probability of different household sizes given family size for those who **respond**

- For example, the estimated probability of household size 1 given family size 1 turns out to be **0.60** while the observed 83/162 =**0.512** is for respondents only.

- Standard estimates and model-based estimates

	Standard	Model-based
Household size = 1	391,000	595,000
Total	1,599,000	1,765,000

The model-based method

$Y_i =$ size of the household for person $i, i = 1, \ldots, N$

$x_i =$ size of the family for person $i, i = 1, \ldots, N$

Population model : $P(Y_i = y)$ depends only on $x_i : P(Y_i = y \mid x_i)$

$R_i = 1/0$ if person i respond/does not respond

Logistic response model dependent on y_i and place of resident

$$
\begin{aligned}
&P(Y_i = y \mid x_i) \\
&= P(Y_i = y \mid x_i, R_i = 1) P(R_i = 1 \mid x_i) \\
&+ P(Y_i = y \mid x_i, R_i = 0) P(R_i = 0 \mid x_i)
\end{aligned}
$$

$H_1 = \text{total number of one-person households}$

$\text{Let } Z_i = 1 \text{ if person } i \text{ "belongs" to a one-person household}$

$$H_1 = \sum_{i=1}^{N} Z_i \Rightarrow$$

$$E(H_1) = \sum_{i=1}^{N} P(Z_i = 1 | x_i) = \sum_{i=1}^{N} P(Y_i = 1 | x_i)$$

$$\hat{H}_1 = \sum_{i=1}^{N} \hat{P}(Y = 1 | x_i) = \sum_{x=1}^{5+} N_x \hat{P}(Y = 1 | x)$$

$N_x = \text{number of persons in the population with registered family size } x$

Family size x	Number of families	Number of persons N_x
1	793,869	793,839
2	408,440	816,880
3	261,527	784,581
4	266,504	1,066,016
5+	127,653	670,528
Total	1,857,993	4,131,874

$\hat{P}(Y=1|x)$ (In parenthesis the observed rate from table on p. 14, $\hat{P}(Y=1|x,R=1)$)

In percentages:

Fam. size *x*	1	2	3	4	5+
$\hat{P}(Y=1\|x)$	60.01 (51.23)	5.27 (3.91)	7.53 (4.72)	1.06 (0.67)	0.84 (0.48)

Nonignorable nonresponse! Probability of response depends on variable of interest, household size

$$\hat{H}_1 = \sum_{x=1}^{5+} N_x \hat{P}(Y=1|x)$$

$$= 793{,}869 \cdot 0.6001 + 816{,}880 \cdot 0.0527 + \ldots + 670{,}528 \cdot 0.0084$$

$$= 595{,}462$$

Effect of nonresponse

Fixed population model of nonresponse:

U = finite population of N units

$$R_i = \begin{cases} 1 \text{ if unit } i \text{ does/would respond} \\ 0 \text{ if not} \end{cases}$$

$i = 1, \ldots, N$

R_i's are fixed, not random

$U_R = \{i \in U : R_i = 1\} = \text{responding subpopulation}$

$U_M = \{i \in U : R_i = 0\} = \text{nonresponding subpopulation}$

$N_R = \text{size of } U_R$

$N_M = \text{size of } U_M$

Bias of standard estimator

Simple random sample of size *n*

Response sample : $s_r = s \cap U_r$, size n_r

Estimate the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

Population means of U_R and U_M : $\bar{Y}_R$ and $\bar{Y}_M$

$$\bar{Y} = \frac{N_R \bar{Y}_R + N_M \bar{Y}_M}{N} = q_R \bar{Y}_R + (1 - q_R)\bar{Y}_M$$

$q_R = N_R / N =$ expected response rate

Standard estimator :

observed sample mean : $\bar{y}_r = \frac{1}{n_r} \sum_{i \in s_r} y_i$

Given n_r : the response sample s_r is a random sample from U_R

$$\Rightarrow E(\bar{y}_r) = \bar{Y}_R$$

$$\Rightarrow \text{Bias} = E(\bar{y}_r) - \bar{Y} = \bar{Y}_R - \bar{Y}$$
$$= \bar{Y}_R - q_R \bar{Y}_R - (1 - q_R)\bar{Y}_M = (1 - q_R)(\bar{Y}_R - \bar{Y}_M)$$

No bias if either $q_R = 1$ or $\bar{Y}_R = \bar{Y}_M$

Nonresponse unrelated to y

Mean square error:

$$E(\bar{y}_r - \bar{Y})^2 = Var(\bar{y}_r) + [E(\bar{y}_r) - \bar{Y}]^2$$

$$\approx \left(1 - \frac{q_R n}{N_R}\right) \frac{\sigma_R^2}{q_R n} + (1 - q_R)^2 (\bar{Y}_R - \bar{Y}_M)^2$$

We notice that even if there is no bias, the uncertainty increases because of smaller sample size

Expected sample size decreases from n to $q_R n$

For example, if we want a sample of 1000 units and we know q_R: n = 1000/q_R

If expected response rate is 60% : need n= 1000/0.60= 1667

$$\text{Bias} = E(\bar{y}_r) - \bar{Y} = (1 - q_R)(\bar{Y}_R - \bar{Y}_M)$$

Possible consequenses of nonresponse:

1. Bias is independent of *n*, can not be reduced by increasing *n*
2. Bias increases with increasing nonresponse rate (1-q_R)
3. Bias increases when $|\bar{Y}_R - \bar{Y}_M|$ increases
4. If $\bar{Y}_R = \bar{Y}_M$: *ignorable* nonresponse mechanism

Unrealistic to assume $\bar{Y}_R = \bar{Y}_M$,

But within smaller subpopulations it may not be unreasonable,

especially if the variable used to partition the population is highly correlated with *y*

Called: poststratification

Widely used tool to correct for nonresponse when MAR is a reasonable model for the response mechanism

Estimation methods for reducing the effect of nonresponse

- Handling nonresponse:
  - Reduce the *size* of nonresponse, especially by callbacks
  - Reduce the *effect* of nonresponse, by estimating the bias and correcting the original estimator designed for a full sample
- Estimation methods:
  - Weighting, especially for unit nonresponse
  - Imputation, especially for item nonresponse

Weighting

Basic idea:

- Some parts of the population are underrepresented in the response sample
- Weigh these parts up to compensate for underrepresentation
- Population-based
  - Reduces sampling error
  - Adjusts for unit nonresponse

Example – age standardized mortality

We have a random sample of 10,000 subjects from a population of 2,000,000, age 40-69 with 40% nonresponse. It turns out that there are different response rates for the age groups 40-49, 50-59, 60-69. Results:

Age group	Population	Sample	Response sample	Non-response	No of deaths	Mortality rate
40-49	1 200 000	6000	3000	50%	25	0.008333
50-59	600 000	3000	2200	26,7%	90	0.040909
60-69	200 000	1000	800	20%	200	0.2500
Total	2 000 000	10000	6000	40%	315	0.0525

- Crude mortality rate based on the sample is 315/6000 = 0.0525 = 52.5 per 1000 subjects
- Direct unweighted estimate of the number of deaths: 2,000,000x0.0525 = 105000
- Weighted estimate of the number of deaths in the population:

  1200000 x 0.008333 + 600000 x 0.040909+200000 x 0.2500 = 10000 + 24545+50000 = 84545
- Mortality rate, age adjusted: 84545/2000000 = 0.0423 = 42.3 per 1000 subjects
- Weighted estimate corrects for
  - Sample is not representative for age distribution
  - Different Nonresponse rates
- Example of poststratification

Poststratification

1. Stratify using variables that partition the population in homogeneous groups
2. Stratify according to varying response rates

H poststrata. For poststratum *h*, U_{Rh} is the responding substratum and U_{Mh} is the nonresponding substratum

$$q_h = \text{responserate in postratum} h$$

$$W_h = N_h / N, \text{where } N_h \text{ is the size of poststratum } h$$

$$\bar{Y}_{Rh} = \text{mean in responsestratum} h$$

$$\bar{Y}_{Mh} = \text{mean in nonresponse stratum} h$$

Simple random sample and $\bar{y}_r$ estimating $\bar{Y}$

$$E(\bar{y}_r)-\bar{Y}$$

$$=\frac{1}{q_R}\sum_{h=1}^{H}\bar{Y}_{Rh}W_h(q_h-q_R)+\sum_{h=1}^{H}(1-q_h)W_h(\bar{Y}_{Rh}-\bar{Y}_{Mh})$$

1. component: Bias because of different response rates in the poststrata, can be estimated

2. component can not be estimated if response and nonresponse means are different

Poststratification estimates the first component

Choose poststrata such that most of the bias is in the first component

$\Rightarrow q_h$ should vary as much as possible, and $\bar{Y}_{Rh} \approx \bar{Y}_{Mh}$

First component $:\bar{Y}_R - \sum_{h=1}^{H} W_h \bar{Y}_{Rh}$

Observed mean from poststratum $h : \bar{y}_h$

$\Rightarrow$ unbiased estimator for this component $: \bar{y}_r - \sum_{h=1}^{H} W_h \bar{y}_h$

$\Rightarrow$ adjusted estimator :

$$\hat{\bar{y}}_{post} = \bar{y}_r - (\bar{y}_r - \sum_h W_h \bar{y}_h) = \sum_{h=1}^{H} W_h \bar{y}_h$$

$= \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h ,$ and for the total $\hat{t}_{post} = \sum_{h=1}^{H} N_h \bar{y}_h$

The poststratified estimator

Weights for each observation in poststraum h :

$N_h / n_{rh}, n_{rh}$ is the size of the response sample in postratum h

Estimating no. one-person households

Poststrata: Fam. size $x = h$	1	2	3	4	5+
Observed rate with household size 1, $\bar{z}_h$	0.5123 (.6001)	0.0391 (0.0527)	0.0472 (0.0753)	0.0067 (0.0106)	0.0048 (0.0084)

$z_i = 1$ if household size is 1

$$\hat{t}_{post} = \sum_{h=1}^{5+} N_h \bar{z}_h = 486{,}032$$

Compared to

1) *unweighted estimate = 390,501*

2) *Modelbased estimate = 595,462* (nonignorable nonresponse)

Poststratification reduces the bias about 50%

SE and approximate 95% confidence interval based on poststratified estimator

$$\hat{t}_{post} = \sum_h N_h \bar{y}_h$$

$$\bar{y}_h = \frac{1}{n_{rh}} \sum_{i \in s_{rh}} y_i \text{, the mean in poststratum } h \text{ in response sample}$$

The sample variance in poststratum h: $s_h^2 = \frac{1}{n_{rh} - 1} \sum_{i \in s_{rh}} (y_i - \bar{y}_h)^2$

Conditional on the sample sizes of the respondents in the poststrata, n_{rh}, design-based SE (use SE from stratified estimator)

$$\hat{V}(\hat{t}_{post}) = \sum_h N_h^2 (1 - \frac{n_{rh}}{N_h}) \frac{s_h^2}{n_{rh}} \quad and \quad SE_{post} = SE(\hat{t}_{post}) = \sqrt{\hat{V}(\hat{t}_{post})}$$

Approximate 95% CI for t: $\hat{t}_{post} \pm 1.96 \cdot SE_{post}$

Simulation of nonresponse and poststratification

- Example: California schools API (Academic performance index)
- Assume a SRS of *n*=500 and the following MAR model: For schooltypes E, M, H: response rates of 30, 80 and 90 percent
- Simulate nonresponse and derive the poststratified estimator of the mean of API in 2000, with schooltypes as poststratifier.
- Compare the poststratified estimator with the sample mean
- SE of the poststratified estimator
- Simulation to estimate coverage of 95% confidence interval

R code: Simulation of MAR nonresponse and poststratification

```
> x=apipop$stype
>  y=apipop$api00
>  make123  = function(x)
+ {
+    x=as.factor(x)
+    levels_x = levels(x)
+    x=as.numeric(x)
+    attr(x,"levels") = levels_x
+    x
+ }
> s=sample(1:6194,500)
> pstrata=make123(x[s])
# poststratum 1 = E, poststratum 2 = H, poststratum 3 = M

> s1=s[pstrata==1]
> s2=s[pstrata==2]
> s3=s[pstrata==3]
```

Simulating nonresponse

```
> length(s1)
[1] 358
> length(s2)
[1] 60
> length(s3)
[1] 82

> n1r=0.3*length(s1)
> n2r=0.9*length(s2)
> n3r=0.8*length(s3)
> s1r=sample(s1,n1r)
> s2r=sample(s2,n2r)
> s3r=sample(s3,n3r)
> length(s1r)
[1] 107
> length(s2r)
[1] 54
> length(s3r)
[1] 65
```

Poststratified estimator

```
> y1mean=mean(y[s1r])
> y2mean=mean(y[s2r])
> y3mean=mean(y[s3r])
> y1mean
[1] 654.5327
> y2mean
[1] 638.3704
> y3mean
[1] 632.5077

> N1=4421
> N2=755
> N3=1018
 Ypost=(y1mean*N1+y2mean*N2+y3mean*N3)/6194
> Ypost
[1] 648.9428
```

SE and 95% CI of poststratified estimator

```
> n1=length(s1r)
> n2=length(s2r)
> n3=length(s3r)
> n=n1+n2+n3

> y1r=y[s1r]
> y2r=y[s2r]
> y3r=y[s3r]

> varest1=N1^2*var(y1r)*(N1-n1)/(N1*n1)
> varest2=N2^2*var(y2r)*(N2-n2)/(N2*n2)
> varest3=N3^2*var(y3r)*(N3n3)/(N3*n3)
> se=sqrt(varest1+varest2+varest3)
> semean=se/6194
> semean
[1] 9.191489
> CI=Ypost+qnorm(c(0.025,0.975))*semean
> CI
[1] 630.9278 666.9578 = (630.9, 667.0), true value =664.7
```

Sample mean and CI based on sample mean

```
> z=c(y1r,y2r,y3r)
> mean(z)
[1] 644.3363

 var(z)
[1] 15857.16
> sesrs=sqrt(var(z)*(6194-n)/(6194*n))
> sesrs
[1] 8.222185
> CIsrs=mean(z)+qnorm(c(0.025,0.975))*sesrs
> CIsrs
[1] 628.2211 660.4515

# misses the true value, because the sample mean is biased
```

R-function for simulating MAR nonresponse and for estimating true confidence level of approximate 95% CI, based on poststratification and sample mean

- R-function: simpostmean=function(b,n,N,N1,N2,N3,r1,r2,r3)
  - b = number of simulations
  - n = sample size of SRS
  - N= population size
  - N1= size of postratum 1
  - N2= size of postratum 2
  - N3 = size of postratum 3
  - r1= response rate for poststratum 1 (E)
  - r2= response rate for poststratum 2 (H)
  - r3 = response rate for poststratum 3 (M)

```
simpostmean=function(b,n,N,N1,N2,N3,r1,r2,r3)
{
Ypost=numeric(b)
se=numeric(b)
zbar=numeric(b)
sesrs=numeric(b)
for(k in 1:b){
s=sample(1:N,n)
pstrata=make123(x[s])
s1=s[pstrata==1]
s2=s[pstrata==2]
s3=s[pstrata==3]
n1r=r1*length(s1)
n2r=r2*length(s2)
n3r=r3*length(s3)
s1r=sample(s1,n1r)
s2r=sample(s2,n2r)
s3r=sample(s3,n3r)
y1[k]=mean(y[s1r])
y2[k]=mean(y[s2r])
y3[k]=mean(y[s3r])
Ypost[k]=(y1[k]*N1+y2[k]*N2+y3[k]*N3)/N
n1=length(s1r)
n2=length(s2r)
n3=length(s3r)
m=n1+n2+n3
y1r=y[s1r]
y2r=y[s2r]
y3r=y[s3r]
varest1=N1^2*var(y1r)*(N1-n1)/(N1*n1)
varest2=N2^2*var(y2r)*(N2-n2)/(N2*n2)
varest3=N3^2*var(y3r)*(N3-n3)/(N3*n3)
se[k]=(sqrt(varest1+varest2+varest3))/N
z=c(y1r,y2r,y3r)
zbar[k]=mean(z)
sesrs[k]=sqrt(var(z)*(N-m)/(N*m))
}
covmean=sum(mean(y)<zbar+1.96*sesrs)-sum(mean(y)<zbar-1.96*sesrs)
covmean=covmean/b
covpost=sum(mean(y)<Ypost+1.96*se)-sum(mean(y)<Ypost-1.96*se)
covpost=covpost/b
list(covpost=covpost,covmean=covmean)
}
```

R-code for simulations to estimate true confidence level of 95% CI, based on poststratification

```
x=apipop$stype
y=apipop$api00

 make123  = function(x)
{
   x=as.factor(x)
   levels_x = levels(x)
   x=as.numeric(x)
   attr(x,"levels") = levels_x
   x
}

 # write out r –code for the function:
 simpostmean=function(b,n,N,N1,N2,N3,r1,r2,r3)
```

Some cases:

n=500, r1=0,3, r2=0.9, r3=0.8:

```
simpostmean(10000,500,6194,4421,755,1018,0.3,0.9,0.8)
$covpost
[1] 0.9514

$covmean
[1] 0.8727
```

n=500, r1=0,5, r2=0.5, r3=0.5:

```
simpostmean(10000,500,6194,4421,755,1018,0.5,0.5,0.5)
$covpost
[1] 0.9436

$covmean
[1] 0.9448
```

Estimated (design-based) confidence level of the approximate 95% CI for poststratification and sample mean, based on 10000 simulations of SRS

n	r1	r2	r3	Conf.level post	Conf. level.mean
200	0.3	0.8	0.9	0.9477	**0.9191**
200	0.5	0.5	0.5	0.9483	0.9486
500	0.3	0.8	0.9	0.9514	**0.8727**
500	0.7	0.2	0.5	0.9450	0.9363
500	0.5	0.5	0.5	0.9436	0.9448
1000	0.3	0.8	0.9	0.9498	**0.7892**
2000	0.3	0.8	0.9	0.9514	**0.6025**
2000	0.6	0.6	0.6	0.9496	0.9517

Poststratified CI has correct coverage in general. The sample mean based CI only works when response rates are the same in all poststrata: response sample is a SRS.

Calibration methods

Consider weighting methods which satisfy **calibration constraints**

Design-based approach: Start with H-T estimator $\hat{t}_{HT} = \sum_{i \in s} (1/\pi_i) y_i$

Design weights: $d_i = 1/\pi_i$. Response sample s_r

Auxiliary information with known totals:

$$t_{x1} = \sum_{i=1}^{N} x_{1i}, t_{x2} = \sum_{i=1}^{N} x_{2i}, \ldots, t_{x\mathrm{k}} = \sum_{i=1}^{N} x_{\mathrm{k}i}$$

Final survey weights w_i satisfy the calibration constraints:

$$\sum_{i \in s_r} w_i x_{1i} = t_{x1}, \sum_{i \in s_r} w_i x_{2i} = t_{x2}, \ldots, \sum_{i \in s_r} w_i x_{\mathrm{k}i} = t_{x\mathrm{k}}$$

Calibrated estimator of *y*-total: $\hat{t}_{cal} = \sum_{i \in s_r} w_i y_i$

Choose the calibrated weights such that the “distance” between d_i and w_i is minimized

Poststratification is an example of calibration

H poststrata with sizes N_h, h= 1,…,H

Define auxiliary variable x_h

$$x_{hi} = \begin{cases} 1 \text{ if unit } i \in \text{poststratum } h \\ 0 \text{ otherwise} \end{cases}$$

$$t_h = \sum_{i=1}^{N} x_{hi} = N_h$$

Response sample in poststratum h: s_{rh} of size n$_{rh}$

Final calibrated estimator: $\hat{t}_{cal} = \sum_{h=1}^{H} \sum_{i \in s_{rh}} w_i y_i$

H calibration constraints:

$$\sum_{i \in s_r} w_i x_{hi} = N_h, \ h = 1, \ldots, H$$

$$\sum_{i \in s_{rh}} w_i = N_h, \; h = 1, \ldots, H \qquad (*)$$

Poststratified estimator:

$$\hat{t}_{post} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} N_h \frac{1}{n_{rh}} \sum_{i \in s_{rh}} y_i$$

$$= \sum_{h=1}^{H} \sum_{i \in s_{rh}} \frac{N_h}{n_{rh}} y_i$$

The weights are $w_i = N_h / n_{rh}$ for $i \in s_{rh}$

Satisfy the calibration constraints (*)

Other weights may also.

Why calibrate?

- Ensures that weighted estimates agree with given “benchmarks”, e.g. N_h
- Typically reduces nonreponse bias if nonresponse is related to the calibration variables
- Improve efficiency for variables related to the calibration variables

Imputation

- Mostly used for item nonresponse, but can also be used for unit nonresponse
- Item nonresponse creates problem even when the nonresponse happens at random, leaves us with few complete cases
- Imputation: filling in for each missing data value by predicting the missing values
- For a given variable *y*, for estimating population total or mean, use estimator constructed for the full sample, based on the observed and imputed data:
- Imputation based estimator
- Need proper variance estimates
- Also want to produce complete data sets that allow for standard statistical analysis
  - right variation in the data vs. variance estimation

Regression-based imputation methods

Regression imputation

Assume a regression model for Y given x,

where x is available also for the nonresponse group

f.ex. $E(Y\mid x)=\beta x,\ Var(Y\mid x)=\sigma^2 x$

Estimate β from the response sample s_r with

$\hat{\beta}_r = \sum_{i\in s_r} Y_i / \sum_{i\in s_r} x_i,$

and for all $i \in$ nonresponse group, predict y_i with

$y_i^* = \hat{\beta}_r x_i$

Problem: Not enough variation to account for the variability in the nonresponse group

Residual regression imputation

Since $Var\{(Y_i - \beta x_i)/\sqrt{x_i}\} = \sigma^2$

Standardized observed residuals : $e_i = (y_i - \hat{\beta}_r x_i)/\sqrt{x_i}$

For $i \in s - s_r$, draw the value e_i^* at random from the set of standardized observed residuals in the response sample $\{e_j : j \in s_r\}$

Imputed *y*-value is given by:

$$y_i^* = \hat{\beta}_r x_i + e_i^* \sqrt{x_i}$$

If the model assumption also includes a distributional assumption, say normality:

Draw imputed values from the estimated $N(\hat{\beta}_r x_i, \hat{\sigma}_r^2 x_i)$

Underlying assumption on the response mechanism:

Missing at random (*MAR*): Probability of response for unit *i* may depend on x_i, while independent of y_i

If basic full sample estimator is the ratio estimator,

$$\hat{T}_R = X \frac{\sum_s Y_i}{\sum_s x_i},$$

then the imputation - based estimator becomes

$$\hat{T}_{R,I} = X \frac{\sum_{s_r} Y_i + \sum_{s-s_r} Y_i^*}{\sum_s x_i}$$

Standard imputation methods, much used in National Statistical Institutes

(i) *Mean* imputation: $y_i^* = \bar{y}_r$

Within poststrata: poststratification

(ii) *Hot-deck* imputation (typically within poststrata):

y_i^* is drawn at random from the observed y values,

with replacement

(iii) *Nearest neighbour* imputation: Find a *donor* in the response sample based on closeness of auxiliary variables

R-code for hot-deck imputation

California schools, API for 2000

```
> y=apipop$api00
#SRS of size 200
> s=sample(6194,200)
> #70% response rate
> #response sample
> nr=0.7*200
> sr=sample(s,nr)
> #hot-deck imputation for s-sr
> simp=sample(sr,200-nr,replace=TRUE)
> simp
 [1] 5137 3870 1595 3990   73 3766  477 3873 2253 2758
3906 3339 5774 3002 3339
[16] 1930 1052  471 2253 5932  759 4343  841  890 4508
1615 3589  749 2758 3747
[31] 3306 1082 3156 1329  544 4465  471 2758 4175 2458
4569 2109 5183 5183 1919
[46] 1900 2063 5189 5137 2792 5118  974 5442 1796 3990
4343  477 3012  890 5118
> #imputed values
> yimp=y[simp]
> yobs=y[sr]
> #Total imputed sample
> ystar=c(yobs,yimp)
> mean(ystar)
[1] 653.57
```

MNAR: Nonignorable nonresponse
How to proceed

The response probabilities are assumed to depend on variable of interest

$P_\psi(R_i = 1 | x_i, y_i)$ is modeled, $f_\psi(r_i | x_i, y_i)$

Population model for Y_i given x_i: $f_\theta(y_i | x_i)$

Joint distribution of Y_i and R_i:

$$f_{\theta,\psi}(y_i, r_i | x_i) = f_\theta(y_i | x_i) f_\psi(r_i | x_i, y_i)$$

Conditional distribution of Y_i given nonresponse, R_i=0

$$f_{\theta,\psi}(y_i | x_i, R_i = 0)$$

$$f_{\theta,\psi}(y_i \mid x_i, R_i = 0) = f_\theta(y_i \mid x_i)P_\psi(R_i = 0 \mid y_i, x_i) / P_{\theta,\psi}(R_i = 0 \mid x_i)$$

where

$$P_{\theta,\psi}(R_i = 0 \mid x_i) = \int f_\theta(y_i \mid x_i)P_\psi(R_i = 0 \mid y_i, x_i)dy_i$$

$$\text{Maximum likelihood estimates}: \hat{\theta}, \hat{\psi}$$

Likelihood function, independence between (Y_i, R_i):

$$l(\theta,\psi) = \prod_{i \in s_r} f(y_{obs,i} \mid x_i)P_\psi(R_i = 1 \mid x_i, y_{obs.i}) \prod_{i \in s - s_r} P_{\theta,\psi}(R_i = 0 \mid x_i)$$

Note: Likelihood function could be quite flat in *y*, numerical difficulties for finding maximum.

$$\text{Imputed values}: y_i^* = E_{\hat{\theta},\hat{\psi}}(Y_i \mid x_i, R_i = 0)$$

or drawing a value from the estimated conditional distribution

Remarks on MNAR models

- Model assumptions cannot be supported by the data alone.
    - Example: If observed *y*-distribution is skewed it could be we have MCAR and a skewed population distribution of *Y*.
- Assumptions in an MNAR model cannot be verified without a specific population model for the Y_i's.
- Must use subject matter knowledge for the missing data mechanism

Illustration of an MNAR model

Binomial case $P(Y_i = 1) = \theta$

$$P(R_i = 1 \mid y_i) = \begin{cases} \psi & \text{if } y_i = 0 \\ 2\psi & \text{if } y_i = 1 \end{cases}, \text{ where } \psi \le 1/2$$

$$P(Y_i = 1 \mid R_i = 0)$$

$$= \frac{P(Y_i = 1)P(R_i = 0 \mid Y_i = 1)}{P(Y_i = 1)P(R_i = 0 \mid Y_i = 1) + P(Y_i = 0)P(R_i = 0 \mid Y_i = 0)}$$

$$= \frac{\theta(1-2\psi)}{\theta(1-2\psi) + (1-\theta)(1-\psi)} = \frac{\theta(1-2\psi)}{1-\psi(1+\theta)}$$

We note that $P(Y_i = 1 \mid R_i = 0) = 0 \quad \text{if } \psi = 1/2$

Maximum likelihood estimates

Let $v = \sum_{s_r} y_i$, the number of "successes"in the responsesample

and n_r is the size of s_r.

Proportion of successesin responsesample: $\hat{p} = v / n_r$

In case of no missing data: $\hat{p} = \hat{\theta}$

Let the responserate be $\hat{r} = n_r / n$

Likelihood function

$$l(\theta,\psi) = \prod_{i \in s_r} f_\theta(y_{obs,i}) \prod_{i \in s_r} P_\psi(R_i = 1 \mid y_{obs.i}) \prod_{i \in s - s_r} P_{\theta,\psi}(R_i = 0)$$

$$= \theta^v (1-\theta)^{n_r - v} (2\psi)^v \psi^{n_r - v} (1 - \psi - \theta\psi)^{n - n_r}$$

and

$$\log l(\theta,\psi) = v \log\theta + (n_r - v)\log(1-\theta) + v\log 2 + n_r \log\psi$$

$$+ (n - n_r)\log(1 - \psi - \theta\psi)$$

Likelihood equations:

$$\text{(I)}\ \ \partial \log l / \partial \psi = 0 \Leftrightarrow \frac{n_r}{\psi} - (1+\theta)\frac{n-n_r}{1-\psi(1+\theta)} = 0$$

$$\Leftrightarrow \psi = \frac{n_r}{n(1+\theta)} = \frac{\hat{r}}{1+\theta}$$

$$\text{(II)}\ \partial \log \mathrm{l}/\partial \theta = 0 \Leftrightarrow \frac{v}{\theta} - \frac{n_r - v}{1-\theta} - \psi\,\frac{n-n_r}{1-\psi(1+\theta)} = 0$$

$$\overset{\text{(I)}}{\Leftrightarrow} \frac{v}{\theta} - \frac{n_r - v}{1-\theta} - \frac{n_r}{n(1+\theta)} \cdot \frac{n-n_r}{1-(n_r/n)} = 0 \Leftrightarrow \frac{v}{\theta} - \frac{n_r - v}{1-\theta} - \frac{n_r}{1+\theta} = 0$$

$$\Leftrightarrow \theta = \frac{v}{2n_r - v} = \frac{\hat{p}}{2-\hat{p}}$$

$$\hat{\theta} = \frac{\hat{p}}{2 - \hat{p}}$$

$$\hat{\psi} = \begin{cases} \dfrac{\hat{r}}{(1+\hat{\theta})} = \dfrac{1}{2} \cdot \hat{r}(2-\hat{p}) \text{ if } \hat{r}(2-\hat{p}) < 1 \\ 1/2 \text{ otherwise} \end{cases}$$

Note: $E(\hat{p}) = E(\overline{Y}_r) = E(Y_i \mid R_i = 1) = P(Y_i = 1 \mid R_i = 1)$

$$= \frac{\theta \cdot 2\psi}{\psi + \theta\psi)} = \frac{2\theta}{\theta + 1} \quad (< \theta)$$

A reasonable estimate would satisfy

$$\hat{p} = \frac{2\hat{\theta}}{\hat{\theta} + 1}, \text{ that is} : \hat{\theta} = \frac{\hat{p}}{2 - \hat{p}} = MLE$$

$$y_i^* = E_{\hat{\theta},\hat{\psi}}(Y_i \mid x_i, r_i = 0) = P_{\hat{\theta},\hat{\psi}}(Y_i = 1 \mid R_i = 0)$$

$$= \frac{\hat{\theta}(1-2\hat{\psi})}{1-\hat{\psi}(1+\hat{\theta})} = \frac{\hat{\theta}-\hat{r}\hat{p}}{1-\hat{r}} \text{ if } \hat{r}(2-\hat{p}) < 1$$

$y_i^* = 0$ otherwise

Estimate the total number of successes in the population, $T = \sum_{i=1}^{N} Y_i$

Basic estimator without nonresponse: $\hat{T} = N \cdot \overline{Y}_s$

Imputation-based estimate:

$$\hat{t}_I = N\frac{1}{n}(\sum_{i \in s_r} y_i + \sum_{i \in s - s_r} y_i^*)$$

$$= N\frac{1}{n}\{n_r \hat{p} + (n-n_r)\frac{\hat{\theta}-\hat{r}\hat{p}}{1-\hat{r}}\} = N\{\hat{r}\hat{p} + (1-\hat{r})\frac{\hat{\theta}-\hat{r}\hat{p}}{1-\hat{r}}\} = N \cdot \hat{\theta}$$

Variance estimation in the presence of imputed values

Consider simplest possible case

- Simple random sample
- MCAR: Random nonresponse
- No auxiliary information

Two possible imputation methods

(i) *mean* imputation : $y_i^* = \bar{y}_r$

(ii) *hot-deck* imputation :

y_i^* is drawn at random from the observed y values, with replacement

- Mean imputation can not be used if the completed data set shall reflect expected variation in the nonresponse group

- Look at standard analysis based on the completed sample: observed and imputed data

$$\text{Problem : Estimate } \bar{Y}$$

$$\bar{y}_s = \text{sample mean if the whole sample } s \text{ is observed}$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum\nolimits_{i \in s} (y_i - \bar{y}_s)^2$$

Large *n*, *N*-*n*

Standard 95% confidence interval :

$$\text{CI} : \bar{y}_s \pm 1.96\hat{\sigma}\sqrt{\frac{1}{n} - \frac{1}{N}}$$

With nonresponse:

Standard CI based on the completed data set with observed and imputed values:

$$\text{CI}^* : \bar{y}_s^* \pm 1.96\hat{\sigma}_*\sqrt{\frac{1}{n} - \frac{1}{N}}$$

$\bar{y}_s^*, \hat{\sigma}_*^2 : \bar{y}_s, \hat{\sigma}^2$ based on the completed sample with observed and imputed values

Coverage with mean imputation

$$W_r = \frac{\bar{y}_r - \bar{Y}}{\hat{\sigma}_r \sqrt{\frac{1}{n_r} - \frac{1}{N}}} \sim N(0,1) \quad \text{approximat ely}$$

$$\hat{\sigma}_*^2 = \frac{n_r - 1}{n - 1} \hat{\sigma}_r^2 \quad \text{such that} \qquad CI^* \approx \bar{y}_r \pm 1.96 \hat{r} \hat{\sigma}_r \sqrt{\frac{1}{n_r} - \frac{1}{N}}$$

Confidence level: $C_* = P(\bar{Y} \in CI_*) = P(|W_r| \leq \hat{r} 1.96)$

Nonresponse (%)	0	10	20	30	40	50
Confidence level	0.95	0.922	0.883	0.830	0.760	0.673

Coverage with hot-deck imputation

$$E(\bar{y}_s^*) = \bar{Y}$$

$$Var(\bar{y}_s^*) \approx \sigma^2 \frac{1}{n}\left[(1-\hat{r}) + \frac{1}{\hat{r}}\right]$$

$$E(\hat{\sigma}_*^2) \approx \sigma^2$$

$$W_* = \frac{\bar{y}_s^* - \bar{Y}}{\hat{\sigma}_* \sqrt{\frac{1}{n}\{(1-\hat{r}) + \frac{1}{\hat{r}}\}}} \sim N(0,1) \quad \text{approximately}$$

Confidence level:

$$C_* \approx P(|W_*| \leq 1.96 \sqrt{\frac{1}{n}} / \sqrt{\frac{1}{n}\{(1-\hat{r}) + \frac{1}{\hat{r}}\}}$$

$$\approx P(|W_*| \leq 1.96 / \sqrt{1 + \frac{1}{\hat{r}} - \hat{r}})$$

Confidence levels of CI*:

Nonresponse (%)	Mean imputation	Hot-deck imputation
0	0.95	0.95
10	0.922	0.925
20	0.883	0.896
30	0.830	0.864
40	0.760	0.826
50	0.673	0.785

One possible solution: Multiple imputation

m repeated hot-deck imputations for each missing value:

m completed samples

$\bar{y}_s^*(i), \hat{\sigma}_*^2(i)$ for $i = 1, \ldots, m$

$\bar{y}_s, \hat{\sigma}^2$ based on the m completed samples

Averages: $\bar{\bar{y}}_s^* = \sum_{i=1}^{m} \bar{y}_s^*(i)/m$ and $\bar{\sigma}_*^2 = \sum_{i=1}^{m} \hat{\sigma}_*^2(i)/m$

A "direct" standard CI:

$$\mathrm{CI}^* : \bar{\bar{y}}_s^* \pm 1.96 \bar{\sigma}_* \sqrt{\frac{1}{n} - \frac{1}{N}}$$

Also too short

$\bar{\sigma}_* \sqrt{\frac{1}{n} - \frac{1}{N}}$ measures the variation only *within* the samples

It is necessary to include a measure of variation *between* the *m* samples; to measure the uncertainty due to imputation

$$B_* = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{y}_s^*(i) - \bar{\bar{y}}_s^*)^2$$

f_{mis} = the nonresponse rate

Replace $\hat{\sigma}_*^2$ with: $V_* = \bar{\sigma}_*^2 (\frac{1}{n} - \frac{1}{N}) + (\frac{1}{1-f_{mis}} + \frac{1}{m}) B_*$

and corresponding 95% CI: $\bar{\bar{y}}_s^* \pm 1.96 \sqrt{V_*}$

- If imputations are based on a Bayesian model, drawing imputed values from the posterior distribution given nonresponse, use 1 instead of 1/(1-f_{mis})