Solutions to Ex #8

1. a) $E(y_i^*|y_{obs}) = \sum_{j\in s_r} y_j P(y_i^* = y_j) = \sum_{j\in s_r} y_j \cdot \frac{1}{n_r} = \underline{\bar{y}_r}$

$$\begin{aligned} Var(y_i^*|y_{obs}) &= E(y_i^* - \bar{y}_r)^2 \\ &= \sum_{j\in s_r} (y_j - \bar{y}_r)^2 \cdot P(y_i^* = y_j) \\ &= \sum_{j\in s_r} (y_j - \bar{y}_r)^2 \cdot \frac{1}{n_r} = \underline{\frac{n_r - 1}{n_r} \hat{\sigma}_r^2} \end{aligned}$$

b) $E(\bar{y}_s^*|y_{obs}) = \frac{1}{n}\left(\sum_{i\in s_r} y_i + \sum_{i\in s - s_r} E(y_i^*|y_{obs})\right)$

$$= \frac{1}{n}(n_r \cdot \bar{y}_r + (n - n_r)\bar{y}_r) = \underline{\bar{y}_r}$$

$$Var(\bar{y}_s^*|y_{obs}) = \frac{1}{n^2} Var\left(\sum_{i\in s_r} y_i + \sum_{i\in s_r} y_i^*|y_{obs}\right)$$

$$= \frac{1}{n^2} Var\left(\sum_{i\in s - s_r} y_i^*|y_{obs}\right)$$

$$= \frac{1}{n^2} \sum_{i\in s - s_r} Var(y_i^*|y_{obs}) = \frac{n - n_r}{n^2} \cdot \frac{n_r - 1}{n_r} \hat{\sigma}_r^2$$

c) $E(\bar{y}_s^*) = EE(\bar{y}_s^*|y_{obs}) = E(\bar{y}_r) = \bar{Y}$

since s_r is a random sample.

$$Var(\bar{y}_s^*) = E\left(\frac{n - n_r}{n^2} \cdot \frac{n_r - 1}{n_r} \hat{\sigma}_r^2\right) + Var(\bar{y}_r)$$

$$= \frac{n - n_r}{n^2} \cdot \frac{n_r - 1}{n_r} \sigma^2 + \sigma^2\left(\frac{1}{n_r} - \frac{1}{N}\right)$$

where $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \bar{Y})^2$ is the pop'n variance

Hence, $Var(\bar{y}_s^*) = \frac{\sigma^2}{n}\left((1 - \frac{n_r}{n})(1 - \frac{1}{n_r}) + \frac{n}{n_r}\right)$

$$\approx \frac{\sigma^2}{n}\left((1 - \hat{r}) + \frac{1}{\hat{r}}\right) \text{ where } \hat{r} = \frac{n_r}{n}.$$

2 a) With MCAR, the response sample is a random sample

$\hat{p} = 1190/1403 = .848$

$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \cdot \frac{N-n}{N}} = .00959$

95% confidence interval: $\hat{p} \pm 1.96 \cdot SE(\hat{p}) = .848 \pm .019$

$= (0.829, 0.867)$

The estimate and CI have large bias.

The nonresponse is not MCAR

b)

$\hat{p}_{post} = \hat{t}_{post}/N$ where $\hat{t}_{post} = \sum_{h=1}^{3} N_h \cdot \bar{y}_h$

$\bar{y}_1 = 1060/1192 = .8893$

$\bar{y}_2 = 57/115 = .4957$

$\bar{y}_3 = 73/96 = .7604$

$\Rightarrow \hat{t}_{post} = 2667953$ and $\hat{p}_{post} = 2667953/3259057$

$= .818$

The estimate is still unbiased. Poststratification is not enough, i.e., the response example is not representative for the nonresponse sample within each poststratum.

c) The basic estimator is the poststratified estimator. The imputed values for nonresponse in stratum h is equal to $\bar{y}_h$ or equivalent: 88.93% of the nonresponse sample in stratum 1 is imputed with value, 49.57% in stratum 2 ~~was~~ imputed values and 76.04% in stratum 3

d. The poststratified estimator is unbiased if response sample is representative of nonresponse sample within each post-stratum is not the case here. Need to assume that the probability of response depends on the value of y.

Solution to R-exercise 5 in Exercise 8

3a

```
> y=c(600,520,620,500,380,460,450,250,400,780)
> s=c(1,2,3,4,5,6,7,8,9,10)
> simp=sample(s,10,replace=TRUE)
> simp
 [1]  4 10  5 10  5  4  2  6  3  4
> yimp=y[simp]
> yimp
 [1] 500 780 380 780 380 500 520 460 620 500
> ycomp=c(y,yimp)
> ycomp
 [1] 600 520 620 500 380 460 450 250 400 780 500 780 380 780 380 500 520 460 620
[20] 500
> mean(ycomp)
[1] 519
> mean(y)
[1] 496

> var(ycomp)
[1] 20472.63
> v=var(ycomp)
> v=var(ycomp)/20
> se=sqrt(v)
> se
[1] 31.99424
> CIimp=mean(ycomp)+qnorm(c(0.025,0.975))*se
> CIimp
[1] 456.2924 581.7076
```

Standard 95% confidence interval, based on completed data set: (456.3, 581.7)

3b

```
> vresp=var(y)/10
> seresp=sqrt(vresp)
> seresp
[1] 46.43275
> CIresp=mean(y)+qnorm(c(0.025,0.975))*seresp
> CIresp
[1] 404.9935 587.0065
```

Standard 95% confidence interval based on response sample: (405.0, 587.0)
The confidence interval in 3a is much too short.

3c

<u>R-code for multiple imputation:</u>

```
>y=c(600,520,620,500,380,460,450,250,400,780)
> s=c(1,2,3,4,5,6,7,8,9,10)
>
> b=5
> n=20
> nmis=10
```

```
> m=5
> for(k in 1:b){
+ simp=sample(s,nmis,replace=TRUE)
+ yimp=y[simp]
+ ycomp=c(y,yimp)
+ ybar[k]=mean(ycomp)
+ var[k]=var(ycomp)/n
+ ymean=sum(ybar)/b
+ varimp1=var(ybar)*(1+1/m)
+ varimp2=var(ybar)*(n/(n-nmis)+1/m)
+ varbar=sum(var)/b
+ se1=sqrt(varbar+varimp1)
+ se2=sqrt(varbar+varimp2)
+ }
> CI_1=ymean+qnorm(c(0.025,0.975))*se1
> CI_2=ymean+qnorm(c(0.025,0.975))*se2
> CI_1
[1] 409.5318 566.4682
># This is the confidence interval using 1 in the combination formula: (409.5, 566.5)
> CI_2
[1] 396.3449 579.6551
># This is the confidence interval using 1/(1-f_mis) in the combination formula: (396,3, 579.7))
> se1
[1] 40.03551
> se2
[1] 46.76368
> ymean
[1] 488
```

We note that CI_2 and se2 are very similar to using the response sample in 3b, and therefore the correct way to combine the multiple hot-deck imputations. CI_1 is too narrow as is the confidence interval in 3a.