

Simulation

Hans Gerritsen

10/2/2019

Simple simulations to test HH estimator

These simulations are based on the examples in Mary Chrisman's presentation: "3) design-based univariate estimation.pptx". The intention is to explore what happens if you adjust your selection probability for estimation purposes (after you have completed your sampling). E.g. in the case where logbook data are available for the current year (but only once the year is over and sampling is complete).

We have a data frame 5 vessels and we base our (unequal probability sampling) selection probability (UP-Sprobs) on the landings in year y minus 1

```
df1 <- data.frame(Vessel=1:5,
                  LandingsYminus1=c(1,2,3,4,50))
# make UPSprobs proportional to landings in year -1
df1$UPSprobs <- df1$LandingsYminus1/sum(df1$LandingsYminus1)
```

We want to estimate the landings in year y by sampling the vessels. First assume that the landings in year y are the same as in the previous year

```
df1$Y <- df1$LandingsYminus1
df1
```

```
##   Vessel LandingsYminus1  UPSprobs  Y
## 1      1              1 0.01666667  1
## 2      2              2 0.03333333  2
## 3      3              3 0.05000000  3
## 4      4              4 0.06666667  4
## 5      5             50 0.83333333 50
```

Then the true total landings is the sum of Y: 60

Simulation with perfect knowledge

Now simulate repeatedly taking 2 samples

```
#number of samples to take
n <- 2

# blank output for HH estimate
hh <- NULL
# repeat 1000 times
for(i in 1:1000){
  # take the sample with replacement in proportion to UPSprobs
  j <- base::sample(1:nrow(df1),n,TRUE,df1$UPSprobs)
  # df2 is the sample data
  df2 <- df1[j,]
  # hh estimator
```

```

hh1 <- (1/n) * sum(df2$Y/df2$UPSprobs)
# add result to the output vector
hh <- c(hh, hh1)
}
mean(hh); sd(hh)

```

```
## [1] 60
```

```
## [1] 0
```

Because the UPSprobs are exactly proportional to Y, we get it exactly right all the time with a mean of 60 and a standard error of 0

Simulation with imperfect knowledge

Now we make the landings in year Y only not exactly the same as year Y-1 but still correlated. The small vessels double their landings and the single large vessel halves the landings.

```

df1$Y <- c(2,4,6,8,30)
df1

```

```

##   Vessel LandingsYminus1  UPSprobs  Y
## 1      1                1 0.01666667 2
## 2      2                2 0.03333333 4
## 3      3                3 0.05000000 6
## 4      4                4 0.06666667 8
## 5      5               50 0.83333333 30

```

```
sum(df1$Y)
```

```
## [1] 50
```

Now the true total landings is: 50

Adjusting the selection probabilities post-hoc

Let's see what happens if we estimate the landings using the actual selection probabilities used in the sampling design (UPSprobs) and also check what happens if we adjust these probabilities based on the actual landings in year Y (for the estimation but not for the sampling design).

```

df1$UPSprobsAdjusted <- df1$Y/sum(df1$Y)
df1

```

```

##   Vessel LandingsYminus1  UPSprobs  Y UPSprobsAdjusted
## 1      1                1 0.01666667 2          0.04
## 2      2                2 0.03333333 4          0.08
## 3      3                3 0.05000000 6          0.12
## 4      4                4 0.06666667 8          0.16
## 5      5               50 0.83333333 30          0.60

```

Now simulate again

```

# blank output for HH estimate
hh <- hhadj <- NULL
# repeat 1000 times
for(i in 1:1000){
  # take the sample with replacement in proportion to UPSprobs
  j <- base::sample(1:nrow(df1),n,TRUE,df1$UPSprobs)
  # df2 is the sample data
  df2 <- df1[j,]
  # hh estimator
  hh1 <- (1/n) * sum(df2$Y/df2$UPSprobs)
  # add result to the output vector
  hh <- c(hh,hh1)
  # hh estimator based on adjusted probabilities
  hh1 <- (1/n) * sum(df2$Y/df2$UPSprobsAdjusted)
  # add result to the output vector
  hhadj <- c(hhadj,hh1)
}
mean(hh); sd(hh)

```

```
## [1] 49.902
```

```
## [1] 21.73096
```

```
mean(hhadj); sd(hhadj)
```

```
## [1] 50
```

```
## [1] 0
```

So we seem to get an unbiased answer using the original sampling design and HH estimator. But if we adjust the probabilities (weights) post-hoc to the true population values, we get the perfect answer without bias and zero standard error.

I would imagine that this means that if you have census data of landings, which you think correlates with what you want to estimate (e.g. discards). Then you can design your sampling, based on landings in year-1 but once you have your logbooks for the current year, you can adjust these probabilities for your estimation and presumably get a better answer. This is probably a form of post-stratification, where the vessels become strata.