Stratified Random Sampling

- Sometimes in survey sampling certain amount of information is known about the elements of the population to be studied.
- For instance, information may be available on the geographical location of the area, e.g. if it is an inner city, a suburban or a rural area.
- Census information will provide a wealth of other information about the area, for instance, its population at the previous census, its rate of population change, the proportion of its population employed in manufacturing, or the proportion of its population with different origins.
- *Supplementary* information of this type can be used either at the design stage to improve the sample design, or at the analysis stage to improve the sample estimators, or both.

- The essence of **stratification** is the classification of the population into sub-populations, or *strata*, based on some supplementary information, and then the selection of separate samples from each of the strata.
- The benefits of stratification derive from the fact that the sample sizes in the strata are controlled by the sampler, rather than being randomly determined by the sampling process.
- Often the strata sample sizes are made proportional to the strata population sizes: this is known as proportionate stratification.
- The division of the total sample between the strata does not, however, have to be restricted to a proportionate allocation; disproportionate stratification is also possible.
- See a visual demonstration about *Stratified Sampling*:

```
library(animation)

sample.strat(col = c("bisque", "white"))
```

NOTATIONS:

- It is added a subscript h to existing symbols to denote the corresponding quantities in stratum h.
- Thus, N_h is the population size and n_h is the sample size in stratum h, with $N = \sum_{h=1}^{L} N_h$ and $n = \sum_{h=1}^{L} n_h$ being respectively the total population and sample sizes.
- $f_h = \frac{n_h}{N_h}$ is the sampling fraction in stratum h; and $\bar{Y}_h$ and $\bar{y}_h$ are the population mean and sample mean in stratum h.
- σ_h^2 and s_h^2 are the population element variance and sample element variance in stratum h.
- It is useful to add a new letter $W_h = \frac{N_h}{N}$ for the proportion of the population in stratum h, in such way that $\sum_{h=1}^{L} W_h = 1$.

- Given simple random sampling within strata, the results from *SRS* can be applied to each stratum separately. Hence, $\bar{y}_h$ are unbiased for the $\bar{Y}_h$ and their variance and standard errors can be estimated according to the *SRS* formulae.

- The new problem presented by stratified sampling is how to combine the strata sample means to produce an estimator of $\bar{Y}$ and how to estimate the variance of this estimator.

- Note that $\bar{Y}$ can be expressed as

$$\bar{Y} = \sum_{h=1}^{L} N_h \cdot \frac{\bar{Y}_h}{N} = \sum_{h=1}^{L} W_h \cdot \bar{Y}_h.$$

- Hence, an unbiased estimator is

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \cdot \bar{y}_h$$

(with subscript *st* for stratified).

- With *SRS* within strata

$$Var(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \cdot Var(\bar{y}_h) = \sum_{h=1}^{L} W_h^2 \cdot \left(\frac{N_h - n_h}{N_h - 1}\right) \cdot \frac{\sigma_h^2}{n_h}$$

- We can estimate $Var(\bar{y}_{st})$ by

$$\widehat{Var}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \left(\frac{N_h - n_h}{N_h}\right) \left(\frac{s_h^2}{n_h}\right)$$

- Given the total variability in the population, the gain in precision by using a proportionate stratified sample rather than an *SRS*, is greater the more heterogeneous are the strata means or, equivalently, the more homogeneous are the element values within the strata.

- See a **function** to calculate stratified sampling programmed in `R`.

```r
# Estimate population mean, mu, given SRS's drawn per stratum


str.mu.est <- function(N.vec,y,details="no") {
   # N.vec is a vector of the stratum sizes
   # y is a list object with each component being
   # a stratum sample
   N.ttl    <- sum(N.vec)
   n.vec    <- unlist(lapply(y,length))
   fpc      <- (N.vec-n.vec)/N.vec
   ybar     <- unlist(lapply(y,mean))
   yvar     <- unlist(lapply(y,var))
   mu.hat   <- sum(N.vec*ybar)/N.ttl
   var.mu   <- sum(N.vec^2*fpc*yvar/n.vec)/(N.ttl^2)
```

```
   Conf.Int <- 2*sqrt(var.mu)
   if(details=="no") {
        cbind(mu.hat,var.mu,Conf.Int)}
   else{
        cbind(mu.hat,var.mu,Conf.Int,n.vec,ybar,yvar)}
 }


# NOTES:
# unlist: given a list structure x, it produces a vector
# which contains all the individual components of x.


# lapply(X,FUN) returns a list of the same length
# as X, each element of which  is the result of
# applying a function to the corresponding element of X
```

EXAMPLE 5.1–5.2, p. 101–102 from Scheaffer et al. (1990):

An advertising firm, interested in determining how much to emphasize television advertising in a certain county, decides to conduct a sample survey to estimate the average number of hours per week that households, within the county, watch TV. The county contains two towns (*A* and *B)* and a rural area. Town A is built around a factory, and most households contain factory workers with school-aged children. Town B is a exclusive residential city and contains older residents with few children at home. There are 155 households in town *A*, 62 in Town *B* and 93 in the rural area. In this case, we have $N_1 = 155$, $N_2 = 62$ and $N_3 = 93$ with $N = 310$.

Suppose we take stratified survey sampling. The advertising firm has enough time and money to interview $n = 40$ households and decides to select random samples of size $n_1 = 20$ from town *A*, $n_2 = 8$ from Town *B* and $n_3 = 12$ in the rural area. Results are

A				B				R			
35	28	26	41	27	4	49	10	8	15	21	7
43	29	32	37	18	41	25	30	14	30	20	11
36	25	29	31					12	32	34	24
39	38	40	45								
28	27	35	34								

Application of R:

```
N.size <- c(155,62,93)

str1 <- c(35,43,36,39,28,28,29,25,38,27,26,32,29,40,35,41,37,31,45,34)

str2 <- c(27,15,4,41,49,25,10,30)

str3 <- c(8,14,12,12,15,30,32,21,20,34,7,11,24)

x <- list(townA=str1,townB=str2,rural=str3)


str.mu.est(N.vec=N.size, y=x, details="yes")
```

Choice of Strata

- Two conditions need to be fulfilled for stratification:
  - The population proportions in the strata W_h need to be known.
  - It has to be possible to draw separate samples from each stratum.
- There must be at least one selection sampled from each stratum; otherwise it would not be possible to calculate an unbiased estimator of the overall population mean.
- If the sample is also to provide a standard error estimate, there must be at least two selections per stratum.
- For gains in precision of the overall estimates, the strata should be formed to be as internally homogeneous in terms of the survey variables as possible.

SELECTION OF THE SAMPLE SIZE

- Suppose we specify that estimate $\bar{y}_{st}$ should lie within ϵ units of the population mean with a confidence of 95%:

$$2\sqrt{Var(\bar{y}_{st})} = \epsilon \Longrightarrow Var(\bar{y}_{st}) = \frac{\epsilon^2}{4}$$

  But we cannot solve this equation without any information about the relations among $n_1, \ldots, n_L$ and n.

- We denote the fraction w_h as the number n_h of observations allocated in stratum h as a part of total sample size n.

  Hence, $n_h = nw_h$, where $h = 1, \ldots, L$.

- Solving previous equation it is obtained that

$$n = \frac{\sum_{h=1}^{L} N_h^2 \sigma_h^2 / w_h}{N^2 D + \sum_{h=1}^{L} N_h \sigma_h^2}$$

where w_h is the fraction of observations allocated to stratum h, σ_h^2 is the population variance for stratum h and $D = \epsilon^2/4$.

- We must introduce previous knowledge about each σ_h. One common option is (Chebyshev's inequality) to relate them with the range of observations: the range is roughly 4 o 6 times standard deviations.

ALLOCATION OF THE SAMPLE

- The essence of sampling is to provide estimators with small variances at the lowest possible cost.
- After the sample size n is chosen it is necessary to divide it in $n_1, \ldots, n_L$ samples per stratum.
- The best allocation scheme is affected by three factors:
  - The total number of elements in each stratum.
  - The variability of observations within each stratum.
  - The cost of obtaining an observation from each stratum.

DISPROPORTIONATE STRATIFICATION

- Proportionate stratification produces simple estimators and it guarantees that the estimators are no less precise than those obtained from a simple random sample of the same size. There are, however, situations in which a disproportionate allocation is helpful.

- One purpose of disproportionate stratification is to achieve an allocation that maximizes the precision of the estimator of the population mean within the available resources.

- The optimum allocation is to make the sampling fraction in a stratum ($f_h = \frac{n_h}{N_h}$) proportional to the standard deviation in that stratum and inversely proportional to the square root of the cost of including an element from that stratum in the sample:

  $f_h \propto \frac{\sigma_h}{\sqrt{c_h}}$, where c_h is the cost per sample element in stratum h.

- Hence, more heterogeneous strata and strata where costs are lower should be sampled at higher rates.

- Often the costs do not differ between strata, so that the optimum allocation reduces to $f_h \propto \sigma_h$ the so-called **Neyman allocation**.

- It is necessary to estimate the stratum element variances and costs on which the allocation is based.

- Unlike the situation with proportionate stratification, a disproportionate allocation can produce less precise estimators than the same-sized simple random sample.

- Disproportionate allocation is valuable when the survey aims to make comparisons between the stratum estimates rather than to aggregate them into an overall estimate.

- The approximate allocation that minimizes cost for a fixed value of $Var(\bar{y}_{st})$ or minimizes $Var(\bar{y}_{st})$ for a fixed cost is:

$$n_h = n \cdot \left(\frac{\frac{N_h \sigma_h}{\sqrt{c_h}}}{\sum_{k=1}^{L} \frac{N_k \sigma_k}{\sqrt{c_k}}} \right)$$

where c_h is the cost of obtaining a single observation from the stratum h.

- By substituting this expressions in the optimal global sample size it is obtained that

$$n = \frac{\left(\sum_{k=1}^{L} \frac{N_k \sigma_k}{\sqrt{c_k}} \right) \cdot \left(\sum_{h=1}^{L} N_h \sigma_h \cdot \sqrt{c_h} \right)}{N^2 D + \sum_{h=1}^{L} N_h \sigma_h^2}$$

EXAMPLE from Scheaffer et al. (1990), p. 106

In the example of the television advertising survey to estimate the average number of hours per week that households, a prior survey suggests that the stratum variances are approximately $\sigma_1^2 \approx 25$, $\sigma_2^2 \approx 225$, and $\sigma_3^2 \approx 100$. We wish to estimate the population mean by using $\bar{y}_{st}$. We must choose, then, the sample size to obtain a bound on the error of estimation equal to 2 hours if the allocation fractions are identical in the three strata.

Solution

We obtain

$$\mathbf{n} = \frac{\sum_{i=1}^{3} N_i^2 \sigma_i^2 / w_i}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2} = \frac{6991275}{96100 + 27125} = 56.7$$

Then $n \approx 57$ with $n_i = n \cdot w_i = 57 \cdot \frac{1}{3} = 27$ in each stratum.

See a **function** to make a selection of the sample size in stratified sampling programmed in R:

```
strat.sample <- function(N.vec, sigma.vec, epsilon, proportion=F) {

# Sample in proportional way or in uniform way per stratum

N.tot <- sum(N.vec)

if(proportion==F){ wt.vec <- rep(1/length(N.vec),length(N.vec)) }

else  { wt.vec <- N.vec/N.tot }


numer <- sum(((N.vec*sigma.vec)^2)/wt.vec)

den <- sum(N.vec*(sigma.vec^2))

D <- epsilon^2/4

n <- ceiling(numer/(D*(N.tot^2) + den))

n.vec <- round(n*wt.vec,1)

cat("  Total n =",n,"\n", " n by stratum =",n.vec,"\n")

}
```

POST-STRATIFICATION

- Sometimes it is not possible to place observations into their correct strata until **after** the sample is selected.

  For example, we may wish to stratify a public opinion poll by sex of the respondent. If the poll is conducted by sampling telephone numbers, then respondents cannot be placed into the male or female stratum until after they are contacted.

- Post-stratification is feasible when, given a population size equal to N, population sizes N_i for each strata are **known**.

- If N_i/N are known and $n_i \geq 20$ for each stratum, *stratification after selection* of the sample is nearly as accurate as stratified random sampling with proportional allocation.

- Another term has to be added to the variance of $\overline{y}_{st}$ (see Scheaffer et al. 1990), p. 130):

$$\frac{1}{n^2} \sum_{h=1}^{L} \left(1 - \frac{N_h}{N}\right) s_h^2$$

Example of stratified sampling with the library `survey`:

```
# SYNTHETIC DATA

mything <- rbind(matrix(rep("nc",530),530,1,byrow=TRUE),

matrix(rep("sc",270),270,1,byrow=TRUE))

mything <- cbind.data.frame(mything,c(rep(1,350),rep(2,150),

rep(3,50),rep(1,100),rep(2,150)),rnorm(800,100,10))


names(mything) <- c("state","region","income")

table(mything$region)


n_1 <- table(mything$region)[[1]]

n_2 <- table(mything$region)[[2]]

n_3 <- table(mything$region)[[3]]


library(sampling)

s <- strata(mything,"region", size=c(50,30,20), method="srswor")

strat_thing <- getdata(mything,s)  # I extract the observed data
```

```
strat_thing$popsize <- with(strat_thing,
ifelse(region=="1",n_1,ifelse(region=="2",n_2,n_3)))


strat_thing$myweights <- 1/strat_thing$Prob


# Export data to Stata format
library(foreign)
write.dta(strat_thing,"C:/QM/mydatastrat.dta")


library(survey)
dstrata <- svydesign(id=~1, weights=~myweights, fpc=~popsize,
strat=~region, data=strat_thing)
summary(dstrata)


svymean(~income, dstrata, na.rm=TRUE)
```

```
# means by strata

svyby(~income, ~region, dstrata, svymean)


svytotal(~income, dstrata, na.rm=TRUE)

svyvar(~income, dstrata, na.rm=TRUE)


svyquantile(~income, quantile=c(0.25,0.5,0.75), design=dstrata, na.rm=TRUE, ci=TRUE)

svyby(~income, ~region, dstrata, svyquantile, quantiles=0.5, ci=TRUE)


X11()

svyhist(~income, dstrata, main="Sample",col="pink")

X11()

svyboxplot(income ~ as.factor(region), dstrata, col="peachpuff")

X11()

plot(svysmooth(~income, design=dstrata))
```

Example of stratified sampling with the library Stata:

```
* Read the previous artificial data

use C:\QM\mydatastrat.dta

count

* Set the sampling design

svyset [pweight=myweights], strata(Stratum) fpc(popsize)

svydescribe


* Compute several statistics

svy: mean income

estat effects

svy: mean income, over(Stratum)


svy: total income

svy: total income, over(Stratum)
```

```
svy: tabulate region

svy: tabulate state

svy: tabulate region state, row se ci


* Compute box-plots

graph box income [pweight=myweights]

graph box income [pweight=myweights], over(Stratum)
```