Introduction to Probabilistic Sampling

- In survey samples it is specified a *population*, whose data values are unknown but are regarded as **fixed**, *not random*. Although, the observed sample is random because it depends the random selection of individuals from this fixed population.

- Properties of a sampling method

  1. Every individual in the population must have a known and a nonzero probability of belonging to the sample ($\pi_i > 0$ for individual i). And π_i must be known for every individual who ends up in the sample.

  2. Every pair of individuals in the sample must have a known and a nonzero probability of belonging to the sample ($\pi_{ij} > 0$ for the pair of individuals (i, j)). And π_{ij} must be known for every pair that ends up in the sample.

Sampling weights I

- If we take a simple random sample of 3500 people from *Neverland* (with total population 35 million) then any person in *Neverland* has a chance of being sampled equal to $\pi_i = 3500/35000000 = 1/10000$ for every i.

- Then, each of the people we sample represents 10000 *Neverland* inhabitants.

- If 100 people of our sample are unemployed, we would expect then $100 \times 10000 = 1$ million unemployed in *Neverland*.

- An individual sampled with a sampling probability of π_i represents $1/\pi_i$ individuals in the population. This value is called the **sampling weight**.

Sampling weights II

- **Example**: Measure the income on a sample of **one** individual from a population of N individuals, where π_i might be different for each individual.

- The estimate ($\widehat{T}_{income}$) of the total income of the population (T_{income}) would be the income for that individual multiplied by the sampling weight

$$\widehat{T}_{income} = \frac{1}{\pi_i} \times income_i$$

- Not a good estimate, it is based on only one person, but it is will be *unbiased*: the expected value of the estimate will equal the true population total:

$$E\left(\widehat{T}_{income}\right) = \sum_{i=1}^{N} \frac{1}{\pi_i} \times income_i \cdot \pi_i = \sum_{i=1}^{N} income_i = income_T$$

The Horvitz-Thompson estimator

- The so called *Horvitz-Thompson* estimator of the population total is the foundation for many complex analysis

- If X_i is a measurement of variable X on person i , we write

$$\widetilde{X}_i = \frac{1}{\pi_i} X_i$$

- Given a sample of size n the Horvitz-Thompson estimator $\widehat{T}_X$ for the population total T_X of X is

$$\hat{T}_X = \sum_{i=1}^{n} \frac{1}{\pi_i} X_i = \sum_{i=1}^{n} \widetilde{X}_i$$

- The variance estimator is

$$\widehat{Var}\left(\hat{T}_X\right) = \sum_{i,j}\left(\frac{X_i X_j}{\pi_{ij}} - \frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\right)$$

- The formula applies to any design, however complicated, where π_i and π_{ij} are known for the sampled observations.

- The formula depends on the pairwise sampling probabilities π_{ij}, not just on the sampling weights: so the correlations in the sampling design enter the computations.

- See formal definitions and properties in Lohr (2007) p. 240–244.

Simple Random Sampling

- Simple random sampling (*SRS*) provides a natural starting point for a discussion of probability sampling methods. It is the simplest method and it underlies many of the more complex methods.

- Notations: Sample size is given by n and the population size by N.

- Formally defined: Simple random sampling is a sampling scheme with the property that any of the possible subsets of n distinct elements, from the population of N elements, is **equally** likely to be the chosen sample.

- Every element in the population has the same probability of being selected for the sample, and the joint probabilities of sets of elements being selected are equal.

EXAMPLE:

- Suppose that a survey is to be conducted in a high school to find out about the students' leisure habits. A list of the school's 1872 students is available, with the list being ordered by the students' identification numbers.

- Suppose that an *SRS* of $n = 250$ is required for the survey. How to draw this sample?

  - By a lottery method: an **urn**. Although conceptually simple, this method is cumbersome to execute and it depends on the assumption that the representative discs (one for each student) have been thoroughly mixed: it is seldom used.

  - By means of a **table of random numbers**. It is useful if you make it *by hand*: in this way is a tedious task, requiring a large selection of random numbers, most of which are nonproductive.

- There are two options:
  - Simple random sampling **with replacement**: an element can be selected more than once.

    ```
    sample(1:1872, size=200, replace=TRUE)
    ```

  - Simple random sampling **without replacement**: the sample must contain n distinct elements.

    ```
    sample(1:1872, size=200, replace=FALSE)
    ```

- Sampling without replacement gives **more precise estimators** than sampling with replacement
- Now, assume that we have responses from all those sampled (there are not problems of *non-response*)
- **Next step**: Summarize the individual responses to provide estimates of characteristics of interest for the population.

  For instance: average number of hours of television viewing per day and the proportion of students currently reading a novel.

See a visual demonstration about *SRS*:

```
library(animation)

sample.simple(nrow=10, ncol=10, size=15, p.col=c("blue", "red"), p.cex = c(1,3))
```

NOTATIONS:

- Capital letters are used for population values and parameters, and lower-case letters for sample values and estimators.
- $Y_1, Y_2, \ldots\ Y_N$ denote the values of the variable y (e.g., hours of television viewing) for the N elements in the population.
- $y_1, y_2, \ldots\ y_n$ are the values for the n sampled elements.
- In general, the value of variable y for the i-th element in the population is Y_i $(i = 1, 2, \ldots\ N)$, and that for the i-th element in the sample is y_i $(i = 1, 2, \ldots\ n)$.

- The **population mean** is given by

$$\bar{Y} = \sum_{i=1}^{N} \frac{Y_i}{N}$$

- and the **sample mean** by

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

- In survey sampling the **population variance** is defined as

$$\sigma^2 = \sum_{i=1}^{N} \frac{\left(Y_i - \bar{Y}\right)^2}{N}$$

- and the **sample variance** as

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

Suppose that we wish to estimate the mean number of hours of television viewing per day for all the students in the school: $\bar{Y}$

Question: how good the sample mean $\bar{y}$ is as an estimator of $\bar{Y}$?

On average, the estimator must be very closed to $\bar{Y}$ over repeated applications of the sampling method.

- Observe that the term *estimate* is used for a specific value, while *estimator* is used for the **rule** of procedure used for obtaining the estimate.
- In the example we obtain, an estimate of 2.20 hours of television viewing (computed by substituting the values obtained from the sampled students in the estimator).

NOTE:

Statistical theory provides a means of evaluating estimators but **NOT** estimates.

- Properties of sample estimators are derived theoretically by considering the pattern of results that would be generated by repeating the sampling procedure an *infinite* number of times.

- **Example**: suppose that drawing a *SRS* of 250 students from the 1872 students and then calculating the sample mean for each sample were carried out *infinite* times (with replacement).

- The resulting set of sample means would have a distribution, known as the *sampling distribution* of the mean.

- If the sample size is not too small ($n \approx 20$ is sufficient) the distribution of the means of each sample approximates the *normal distribution*, and the mean of this distribution is the population mean, $\bar{Y}$.

- Then it is said that the mean of the individual sample estimates $\bar{y}$ over an infinite number of samples is an **unbiased estimator** of $\bar{Y}$.

- Although the sampling distribution of $\bar{y}$ is centered on $\bar{Y}$, any one estimate will differ from $\bar{Y}$.

- To avoid confusion with the standard deviation of the element values, standard deviations of *sampling distributions* are known as *standard errors*.

- The variance (but it is **not useful** in practice) of a sample mean ($\bar{y}_0$) of a *SRS* of size n is given by

$$Var(\bar{y}_0) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} = \frac{N-n}{N-1} \cdot \frac{1}{n \cdot N} \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2$$

- See example 2.1, p. 29 (artificial example) from Lohr (2006)

  Suppose we have a population with eight elements (e.g. an almost extinct species of animal...) and we know the weight y_i for each of the $N = 8$ units of the whole population. We want to know the whole weight of the population.

Animal *i*	1	2	3	4	5	6	7	8
y_i	1	2	4	4	7	7	7	8

- We take a sample of size 4. How many samples of size 4, can be drawn without replacement from this population?

- There are $\binom{8}{4} = 70$ possible samples of size 4 that can be drawn without replacement from this population.

- We define as $P(S) = 1/70$ for each distinct subset of size 4 from the *population*.

```r
# Consider a vector of 'observations'
y <- c(1,2,4,4,7,7,7,8)

# Consider all possible samples of size 4. They are placed into a matrix with 70 rows.
allguys <- NULL
for(i in 1:5){
    for (j in (i+1):6) {
        for (k in (j+1):7){
            for(m in (k+1):8) {
                allguys <- rbind(allguys, c(i,j,k,m)) }}}}
dim(allguys)
# Observe how indices change in each row
allguys
# Other option
library(combinat)
cuales <- combn((1:8),4)
t(cuales)
```

```
# To estimate the whole population weight, we take two times the sum
# of the y values for each possible sample of indices.
alltotal <- apply(allguys, 1, function(i){2*sum(y[i])})

# It produces a vector of length 70: all possible
# estimators of the attribute 'total' based on the 70
# equally likely samples that could be drawn.
table(alltotal)

# Each of these values divided by 70 gives the probability
# of resulting estimate of total of y attributes.

# Observe that these  values are coincident
print(c("Mean:",mean(alltotal)),quote=F)
print(c("Sum:",sum(y)),quote=F)
```

- The formula $Var(\bar{y}_0)$ depends on $\frac{N-n}{N-1}$ and the sample size n.
- The $\frac{N-n}{N-1}$ term reflects the fact that the survey population is finite in size and that sampling is conducted without replacement.
- With an infinite population, or if sampling were conducted with replacement, the term is not included and the expressions are reduced to the familiar forms.
- The term indicates the gains of sampling without replacement over sampling with replacement.
- In many practical situations the populations are large and, even though the samples may also be large, the sampling fractions are small.

- In large populations, the difference between sampling with and without replacement is not important: even if the sample is drawn with replacement, the chance of selecting an element more than once is slight.

- If the sampling fraction n/N is small, $\frac{N-n}{N-1}$ is close to 1 and has a negligible effect on the standard error.

- The correction factor is commonly neglected (i.e., treated as 1) when the sampling fraction (n/N) is less than 1 in 20, or even 1 in 10.

- The larger the sample size n is, the smaller is $Var(\bar{y}_0)$.

- For large populations it is the sample size that is dominant in determining the precision of survey results.

- A sample of size 2000 drawn from a country with a population of 200 million yields about as precise results as a sample of the same size drawn from a small city of 40000 (assuming the element variances in the two populations are the same).

- The element variance in the population σ^2 is unknown in a practical application. Denote

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

As (see Scheaffer et al. (1990) in Appendix)

$$E(s^2) = \frac{N}{N-1}\sigma^2,$$

then,

$$\widehat{Var}(\bar{y}_0) = \frac{N-n}{N} \cdot \frac{s^2}{n} = \left(1 - \frac{n}{N}\right) \cdot \frac{s^2}{n}$$

- The factor $\left(1-\frac{n}{N}\right)$ is called the *finite population correction* (*fpc*) where n/N is the sampling fraction.
- It is easy to determine a confidence interval for the population mean, applying standard *Central Limit Theorem*.
- **Example**: suppose that the mean hours watching television per day for the 250 sampled students is $\bar{y}_0 = 2.192$ hours, with an element variance of $s^2 = 1.008$. Then a 95% confidence interval for $\bar{Y}$ is

$$2.192 \pm 1.96\sqrt{\left(1-\frac{250}{1872}\right)\frac{1.008}{250}} = 2.192 \pm 0.116$$

- That is, we are 95% confident that the interval from 2.076 to 2.308 contains the population mean.
- See some **functions** to illustrate the Central Limit Theorem and to compute confidence intervals programmed in `R`.

```r
# Using R to illustrate the Central Limit Theorem (from Venables)

N <- 10000

 graphics.off()

 par(mfrow = c(1,2), pty = "s")

 for(k in 1:20) {

    m <- (rowMeans(matrix(runif(N*k), N, k)) - 0.5)*sqrt(12*k)

    hist(m, breaks = "FD", xlim = c(-4,4), main = k,

            prob = TRUE, ylim = c(0,0.5), col = "lemonchiffon")

    pu <- par("usr")[1:2]

    x <- seq(pu[1], pu[2], len = 500)

    lines(x, dnorm(x), col = "red")

    qqnorm(m, ylim = c(-4,4), xlim = c(-4,4), pch = ".", col = "blue")

    abline(0, 1, col = "red")

    Sys.sleep(1)

  }
```

```
# Using the TeachingDemo library


library(TeachingDemos)

X11()

clt.examp()

X11()

clt.examp(5)

X11()

clt.examp(30)

X11()

clt.examp(50)
```

```r
srs.mu <- function(y,N,cuacua=0.95) {
        # Estimate the confidence interval for the mean
        # y is the sample values
        # N is the population’s size
        # cuacua is the desired quantile (e.g. 0.95)
        n <- length(y)
        ybar <- mean(y)
        z <- qnorm(1-(1-cuacua)/2)
        s2 <- var(y)
        var.ybar <- ((N-n)/N) * s2/n
        se.ybar <- sqrt(var.ybar)
        B1 <- ybar + z*se.ybar
        B2 <- ybar - z*se.ybar
        list(B2,B1)}
```

Sample size for estimation population means

- How large must be a sample? $\Rightarrow$ Observations cost money, time and efforts.

- The number of observations needed to estimate a population mean μ with a bound on the error of estimation of magnitude ε is obtained by solving this equation for n:

$$2\sqrt{Var(\bar{y})} = 2\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)} = \varepsilon$$

(as $z_{0.025} = 1.96$, we approximate this value by **2**, for a 95% of confidence).

- Hence, the sample size required to estimate μ with a bound on the error of estimation ε is

$$n = \frac{N \cdot s^2}{\frac{N}{4}\varepsilon^2 + s^2}$$

- Note that s^2 must be estimated previously by means of another argument.

EXAMPLE

- The average amount of money μ for a hospital's accounts receivable must be estimated. Although no prior data is available to estimate the population variance σ^2, it is known that most accounts lie within a €100 range. There are $N = 1000$ open accounts. Find the sample size needed to estimate μ with a bound on the error of estimation $\varepsilon =$ €3.

- First we estimate the population variance. Since the range is often approximately equal to four or six standard deviations ($4 \cdot s$ or $6 \cdot s$), depending of the normality of data (see the Chebyshev's inequality), then

$$s \approx \frac{\text{range}}{4} = \frac{100}{4} = 25$$

then $s^2 \approx 625$, so

$$n = \frac{N \cdot s^2}{\frac{N}{4}\varepsilon^2 + s^2} = \frac{1000 \cdot 625}{\frac{1000}{4} \cdot 9 + 625} = 217.39 \approx 217 \text{ or } 218 \text{ observations}$$

See a **function** to calculate sample sizes programmed in R:

```
     n.mu <- function(N,s2,epsilon) {
        # n to estimate the population mean
        D <- (epsilon^2)/4
        n <- (N*s2)/((N*D)+s2)
        n.round <- round(n)
        cbind(n.round)
     }

# Application
n.mu(N=1000,s2=625,epsilon=3)
```

- Many sample surveys are interested about a *population total*, e.g. when analyzing total accounts.

- The population total (the sum of all observations) in the population is denoted by the symbol τ. Hence

$$N\mu = \tau$$

- It is expected the estimator of τ to be N times the estimator of μ, then

  - **Estimator of population total**

$$\hat{\tau} = N\bar{y} = \frac{N\sum_{i=1}^{n} y_i}{n}$$

  - **Estimated variance of $\hat{\tau}$**

$$\widehat{Var}(\hat{\tau}) = \widehat{Var}(N\bar{y}) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Systematic Sampling

- The method of systematic sampling reduces the effort required for the sample selection.
- Systematic sampling is easy to apply, involving simply taking every k-th element after a random start.
- Example: suppose that a sample of 250 students is required from a school with 2000 students. The sampling fraction is 250/2000, or 1 in 8.
- A systematic sample of the required size, would then be obtained by taking a random number between 1 and 8, to determine the first student in the sample, and taking every eighth student thereafter.
- If the random number were 5, the selected students would be the fifth, thirteenth, twenty-first, and so on, on the list.
- When the sampling is not a simple integer, we can round the interval to an integer, with a resultant change in the sample size.

- Example: If fraction is 250/1872 or 1 in 7.488, a 1 in 7 sample would produce a sample of 267 or 268, while a 1 in 8 sample would produce a sample of 234.
- Other solution is to round the interval down to start with an element selected at random from the N elements in the population, and to proceed until the desired sample size has been achieved. Therefore, the list is treated as circular, and the last listing is followed by the first.
- Like *SRS*, systematic sampling gives each element in the population the same chance of being selected for the sample.
- It differs, however, from *SRS* in that the probabilities of different sets of elements being included in the sample are **not all equal**.
- In systematic sampling the sample mean is a reasonable estimator of the population mean. However, the unequal probabilities of sets of elements means that the *SRS* standard error formulae are **not directly applicable**.

- In order to estimate the standard error of estimators based on systematic samples. Sometimes it is reasonable to assume that the list is approximately randomly ordered, in which case the sample can be treated as if it were a simple random sample.

- Lists arranged in alphabetical order may often be reasonably treated in this way.

- Systematic sampling performs badly when the list is ordered in cycles of values of the survey variables and when the sampling interval coincides with a multiple of the length of the cycle.

- Systematic sampling is widely used in practice without excessive concern for the damaging effects of undetected cycles in the ordering of the list.

- See visual demonstration in:

  ```r
  library(animation)
  sample.system()
  ```

- See a **function** to calculate systematic sampling programmed in `R`:

```
systematic.sample <- function(n, N, initial=F){
 k <- floor(N/n)
     if(initial==F){
        initial <- sample(1:k,1)}
cat("Interval=", k, " Starting value=", initial, "\n")
# Put the origin in the value 'initial'
 shift <- (1:N) - initial
 # I search numbers who are multiple of number k
 # (equivalent to find the rest of a%%b=0)
 guy <- (1:N)[(shift %% k) == 0]
 return(guy)
}
```

Sampling with probabilities proportional to size

- Sometimes is advantageous to select sampling units with different probabilities (other than uniform).
- The method is called sampling with probabilities proportional to size or *pps sampling*.
- For a sample $y_1, y_2, \ldots, y_n$ from a population of size N, let π_i the probability that y_i appears in the sample.
- the pps estimator of μ only produces smaller variances than an standard *SRS* if the weights π_i are approximately proportional to the size of the y_i under investigation.

- In this case, the estimator of the population mean μ is

$$\hat{\mu}_{pps} = \frac{1}{Nn} \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

and the estimated variance of $\hat{\mu}$ is

$$\widehat{Var}(\hat{\mu}_{pps}) = \frac{1}{N^2 n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i} - N \cdot \hat{\mu}_{pps} \right)^2$$

- The best practical way to choose the weights π_i's is to chose them proportional to a known measurement that is highly correlated with y_i.

- See the library `pps` of `R` from:

`http://cran.r-project.org/web/packages/pps/index.html`

Example (from *Scheaffer* et al., p. 80):

An investigator wishes to estimate the average number of defects per keyboard on keyboards of electronic components manufactured for installation in computers. The keyboards contain varying numbers of components, and the investigator feels that the number of defects should be positively correlated with the number of components on a keyboard.

Thus, *pps* sampling is used with the probability of selecting anyone keyboard for the sample being proportional to the number of components on that keyboard. A sample of $n = 4$ keyboards is to be selected from the $N = 10$ keyboards of one day's production. The number of components on the 10 keyboards are, respectively: 10, 12, 22, 8, 16, 24, 9, 10, 8, 31.

After the sampling was completed, the number of defects found on the four keyboards were, respectively, 1, 3, 2 and 1. Estimate the average number of defects per keyboard, and place a bound on the error of estimation.

```r
data <- 1:10
N <- length(data)
n <- 4
weights <- c(10, 12, 22, 8, 16, 24, 9, 10, 8, 31)
probs <- weights/sum(weights)
who <- sample((1:N), n, prob=probs, replace=FALSE)
# In Scheaffer et al. the number of defects for each board were
yi <- c(1, 3, 2, 1)

m.pps <- (1/(N*n))*sum(yi/probs[who])
m.pps

var.pps <- (1/((N^2)*n*(n-1)))*sum(((yi/probs[who]) - (N*m.pps))^2)
err <- 2*sqrt(var.pps)
cat("interval 95% -> [",  m.pps-err, ";" , m.pps+err, "]","\n")
```

Consider a *SRS* based in the library `survey`:

`http://faculty.washington.edu/tlumley/survey/`

We consider this **example**

```
# Artificial Data

mydata <- rbind(matrix(rep("nc",165),165,1,byrow=TRUE),

matrix(rep("sc",70),70,1,byrow=TRUE))

mydata <- cbind.data.frame(mydata,c(rep(1,100),rep(2,50),rep(3,15),

rep(1,30),rep(2,40)),100*runif(235))

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

N <- dim(mydata)[[1]]

n <- 50


# Export data to a file with Stata format

library(foreign)

write.dta(mydata,"C:/QM/mydata.dta")
```

```r
# Selection of a sample
srs_rows <- sample(N,n)
srs <- mydata[srs_rows,]

library(survey)
srs$popsize <- N
dsrs <- svydesign(id=~1, fpc=~popsize, data=srs)
summary(dsrs)

svytotal(~income, dsrs, na.rm=TRUE)
svymean(~income, dsrs, na.rm=TRUE)
svyvar(~income, dsrs, na.rm=TRUE)
svyquantile(~income, quantile=c(0.25,0.5,0.75), design=dsrs, na.rm=TRUE, ci=TRUE)

par(mfrow=c(2,1))
svyhist(~income, dsrs, main="Survey weighted", col="red")
hist(mydata$income, main="Population", xlab="income", col="yellow", prob=TRUE)
```

Consider a *SRS* programmed with `Stata`:

```
* Read the previous artificial data

use C:\QM\mydata.dta

count



* Fix the seed of the randomization

set seed 666

* Take a sample equal to the 20\% of the population

sample 20

count



* Compute weights and the factor of population correction

gen pw = 235/47

gen fpc = 235
```

```
* Set the sampling design
svyset [pweight=pw], fpc(fpc)
svydescribe

* Compute several statistics
svy: mean income
svy: total income
svy linearized : tabulate region
svy linearized : tabulate state
svy: tabulate region state, row se ci format(%7.4f)

* Compute a box-plot
graph box income [pweight=pw]
```