

Week 11--IES 612.doc

IES 612 Spring/Winter 2009

Background

- Two questions:
1. Are there other types of sampling methods besides Simple Random Sampling, which we've always assumed up to this point?
 2. Do different sampling methods have any impact on the statistical side of things, ie inferences such as estimation and hypothesis testing?

Recall the following.

1. Sample statistics, such as \bar{y} , s , and p , have distributions known as **SAMPLING DISTRIBUTIONS**.
2. **SAMPLING DISTRIBUTIONS** have **CENTER aka**
SPREAD aka
SHAPE aka
3. We know that if we take a large Random Sample (ie a Simple Random Sample) from a population with mean = μ_{popln} , and variance = σ^2_{popln} , then the sampling distribution of the sample average, \bar{y} , will have a distribution with:

CENTER =

SPREAD =

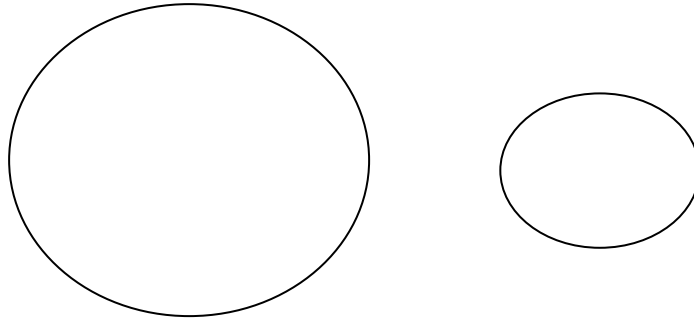
SHAPE =

- Two answers:
1. There are **MANY** different sampling methods, each with advantages and disadvantages.
 2. Different sampling methods will result in different sampling distributions of statistics, with the primary difference in the **standard error** of the statistic.

GOAL

Estimate population **parameters** based on information in a sample, so that:

1. Represent population and yield most precise information (lowest cost, shortest time).
2. But recognize that they will be influenced by sample size and popln variability.



TYPICAL PARAMETERS OF INTEREST

- μ : Population mean (μ = mean Dioxin level at a particular location)
- σ : Population standard deviation (σ = sdev of air contamination levels in Butler Co.)
- π : Population proportion (π = proportion of honey bees that died last winter)
- $\tau = N\mu$: Population total (τ = total income of Miami graduates their first year out of school)
- $\psi = N$: Population number of observations ($\psi = N$ = Ohio white-tailed deer population size)

WHY SAMPLE AND NOT CENSUS?

1. Cost (sample relative to population census)
2. Speed (to conduct the study)
3. Accuracy (more precision)

EXAMPLES OF SAMPLES

1. USEPA Clean Watersheds Needs Survey
[<http://www.census.gov/hhes/www/housing.html>]
2. Current Population Survey [http://www.bls.census.gov/cps/cpsmain.htm]
3. National Health and Nutrition Examination Survey
[<http://www.cdc.gov/nchs/about/major/nhanes/growthcharts/charts.htm>]
4. Bureau of Labor Statistics [http://www.bls.gov]
5. Built into the census [http://www.census.gov/hhes/www/housing.html]

PRINCIPAL STEPS IN TAKING A SAMPLE/SURVEY

1. Objectives (e.g. injuries among workers in Ohio Nursing Homes)
2. Population defined (e.g. all Nursing Homes in Ohio)
3. Data to be collected (variables measured) - e.g. number of injured workers
4. Precision desired
5. Methods of measurement - choice of instrument
6. Frame - list of sampling units in population (e.g. NHs in Ohio)
7. Selection of the sample - sampling plan (e.g. SRS)
8. Pretest/pilot test study (e.g. cognitive surveys)
9. Organization of fieldwork
 - a. Training of interviewers
 - b. Supervision of data collection
 - c. Data quality checks
 - d. Dealing with non-response - reminder cards
10. Analysis - edit, range checks, quality assurance, estimation
11. Future studies

TYPES OF SAMPLING

1. Probability sampling

each sample element has a known probability of selection (our focus)

2. Non-probability sampling

Haphazard/volunteers/convenience (may be useful results but doesn't support inference based on sampling theory).

SAMPLE SURVEY DESIGNS - PROCEDURES FOR SELECTING A SAMPLE

Simple Random Sampling (SRS)
Systematic Sampling
Stratified RS
Cluster Sampling
Capture-Recapture Sampling
Randomized Response Sampling

VOCABULARY/NOTATION

Frame = list of all elements in the population
N = number of elements in the population ("population size")
n = number of elements selected for inclusion in the sample ("sample size")
Element = object on which measurement is taken
Population = collection of all objects
Sample = subset of the population
Sampling Units = non-overlapping collection of elements (e.g. voter, household)
Parameter = numeric characteristic of the population (μ , π , τ , and ψ or N)
Statistic = numeric characteristic of the sample (\bar{y} , $\hat{\mu}$, p , $\hat{\pi}$, $\hat{\tau}$, and $\hat{\Psi}$ or \hat{N})
Population distribution = distribution of measurement/variable in the population
Sample distribution = distribution of measurement/variable in the sample
Sampling distribution = distribution of a statistic/estimator over repeated samples

PROPERTIES OF STATISTICS/ESTIMATORS

Unbiased Statistic/Estimator (mean of the sampling distribution = parameter it estimates)

Recall if n large, distn of \bar{y} is _____. So \bar{y} is _____.

Recall if n large, distn of p or $\hat{\pi}$ is _____. So p or $\hat{\pi}$ is _____.

Efficient Statistic/Estimator (smaller variance relative to other estimators)

If population is symmetric, popln mean and median are _____. So both _____ and _____ are unbiased estimates of popln mean.

HOWEVER, the _____ is more efficient than the _____ at estimating the popln mean.

Simple Random Sampling (SRS)

Defn: SRS = sample of n elements from a population of size N constructed so that every sample of size n has the same probability of being selected from the population.

PROCEDURE

1. Obtain population frame
2. Identify every element in the frame with a unique identifier, usually a number, could be as simple as numbering elements from 1 to N
3. **RANDOMLY** select n distinct items from the frame, eg if numbered from 1 to N , then select n random numbers from $[1, \dots, N]$

Example:

Suppose we wanted to select a SRS of 10 residents from a nursing home with 120 residents. Obtain a list of population residents and number them 1 to 120.

Using R

```
> residents = 1:120
> sample(residents, size=10) * note that replace = FALSE is the default
[1] 56 119 43 3 75 108 48 93 24 116
```

Our random sample of 10 residents are those labeled: **3, 24, 43, 48, 56, 75, 93, 108, 116, 119**

Using SAS

PROC PLAN gives one way to select this sample (more to come with SURVEYSELECT)

```
options nocenter nodate;
proc plan;
  factors resds=10 of 120;
run;
```

```
The PLAN Procedure
Factor      Select      Levels      Order
resds       10          120        Random
----- resds-----
87 95 107 3 24 22 21 2 13 106
```

So, select residents associated with labels: **2, 3, 13, 21, 22, 24, 87, 95, 106, 107**

NOTES/COMMENTS

1. SRS Advantages:

- i. easy to conduct and analyze
- ii. easy to explain

2. SRS Disadvantages:

- i. other designs may be more efficient (require larger "n" to achieve same precision)
- ii. if population stratified, then sample may not be representative
- iii. might not be time or cost efficient to collect

ESTIMATING PARAMETERS IN A SRS

1. ESTIMATING THE POPULATION MEAN (μ) IN A SRS

a. Point Estimate of μ is \bar{y}

b. Standard Error of PE = $se(\bar{y}) = \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$.

c. CI for μ : approximate $1 - \alpha$ design-based confidence interval for μ is

$$\bar{y} \pm z_{\frac{\alpha}{2}} se(\bar{y}) \quad \text{or} \quad \bar{y} \pm t_{\left(\frac{\alpha}{2}, n-1\right)} se(\bar{y}).$$

NOTES/COMMENTS

1. $(N-n)/N$ = finite population correction factor (fpc). Needed because sampling without replacement from a small population.
2. $fpc \approx 1$ if $n < 5\%$ of N .
3. The approx. $\frac{1}{2}$ width of a 95% CI, $2se(\bar{y})$, is sometimes referred to as B = bound on the error of estimation.

NOTES/COMMENTS (Continued)

4. How large a sample is needed so that the margin of error from an approximate 95% confidence interval for μ is no larger than, say, E units?

$$n \geq \frac{4N\sigma_{\text{popln}}^2}{4\sigma_{\text{popln}}^2 + (N-1)E^2}$$

where the population variance can be approximated as $\hat{\sigma}_{\text{popln}}^2 \approx \frac{\text{Range}^2}{26.5}$

2. ESTIMATING THE POPULATION PROPORTION IN A SRS (p or π) IN A SRS

- Point Estimate of π or p is \hat{p} Or $\hat{\pi}$, the sample proportion
- Standard Error of PE = $se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N} \right)}$.
- CI for μ : approximate $1 - \alpha$ design-based confidence interval for π or p is

$$\hat{p} \pm z_{\frac{\alpha}{2}} se(\hat{p}).$$

3. ESTIMATING THE POPULATION TOTAL ($\tau = N\mu$) IN A SRS

- Point Estimate of τ is $\hat{\tau} = N\bar{y}$
- Standard Error of PE = $se(\hat{\tau}) = se(N\bar{y}) = Nse(\bar{y}) = N\sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$.
- CI for μ : approximate $1 - \alpha$ design-based confidence interval for τ is

$$N\bar{y} \pm z_{\frac{\alpha}{2}} Nse(\bar{y}).$$

Example (SMO, p. 52, #4.7) Water consumption

GOAL #1: Estimate mean daily water consumption per household for a community of 10,000 residents.

SAMPLE: SRS $n = 100$ from $N = 10000$ household water meters in a community

SAMPLE RESULTS:

$$n = 100$$

$$\bar{y} = 12.5$$

$$s^2 = 1,252$$

ADDITIONAL RESULTS:

$$se(\bar{y}) = \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)} = \sqrt{\frac{1,252}{100} \left(\frac{10,000-100}{10,000} \right)} = \sqrt{12.3948} = 3.5206$$

Approx 95% CI for μ = mean household water usage is:

$$12.5 \pm 1.96 * 3.5206 \quad \text{or} \quad 12.5 \pm 2 * 3.5206 \quad \text{or} \quad 12.5 \pm 7.04 \quad \text{or} \quad [5.46, 19.54]$$

Bound (aka Bound on the Error of Estimation) = $B = 2 * 3.5206 = 7.04$

GOAL #2: How large a sample size would be required to estimate μ with a $B = 5$?

$$n \geq \frac{4N\sigma_{popln}^2}{4\sigma_{popln}^2 + (N-1)E^2} = \frac{4(10,000)(1,252)}{4(1,252) + (10,000-1)5^2} = 196.41 \text{ so use } n = 197$$

Note that we used the observed sample variance as our estimate of σ_{popln}^2 .

GOAL #3: Estimate TOTAL daily water consumption in the community

$$\hat{\tau} = N\bar{y} = 10,000 (12.5) = 125,000 \quad \text{and} \quad se(\hat{\tau}) = Nse(\bar{y}) = 10,000(3.5206) = 35,206$$

Approx 95% CI for τ = total household daily water usage is:

$$125,000 \pm 2 * 35,206 \quad \text{or} \quad 125,000 \pm 70,412 \quad \text{or} \quad [54,600, 195,400]$$

Bound (aka Bound on the Error of Estimation) = $B = 2 * 35,206 = 70,412$

Example: Soil contamination in Austria (Source: Majer et al. 2002)

City and site information for Austrian Soil Contamination Study.	
City	Available sites
1. Untertiefenbach	16
2. Reisenberg	12
3. Feistritz	13
4. Mitterberghütten	16
5. Ramingstein	14
6. Bleiberg	13
7. Arnoldstein	12

Let's first select a SRS of $n = 16$ from the 96 total sites.

Using R

```
> sites = c(paste("UT", 1:16, sep=""), paste("RB", 1:12, sep=""),
+           paste("FW", 1:13, sep=""), paste("MB", 1:16, sep=""),
+           paste("RN", 1:14, sep=""), paste("BB", 1:13, sep=""),
+           paste("AN", 1:12, sep=""))
> sites
 [1] "UT1"  "UT2"  "UT3"  "UT4"  "UT5"  "UT6"  "UT7"  "UT8"  "UT9"  "UT10"
[11] "UT11" "UT12" "UT13" "UT14" "UT15" "UT16" "RB1"  "RB2"  "RB3"  "RB4"
[21] "RB5"  "RB6"  "RB7"  "RB8"  "RB9"  "RB10" "RB11" "RB12" "FW1"  "FW2"
[31] "FW3"  "FW4"  "FW5"  "FW6"  "FW7"  "FW8"  "FW9"  "FW10" "FW11" "FW12"
[41] "FW13" "MB1"  "MB2"  "MB3"  "MB4"  "MB5"  "MB6"  "MB7"  "MB8"  "MB9"
[51] "MB10" "MB11" "MB12" "MB13" "MB14" "MB15" "MB16" "RN1"  "RN2"  "RN3"
[61] "RN4"  "RN5"  "RN6"  "RN7"  "RN8"  "RN9"  "RN10" "RN11" "RN12" "RN13"
[71] "RN14" "BB1"  "BB2"  "BB3"  "BB4"  "BB5"  "BB6"  "BB7"  "BB8"  "BB9"
[81] "BB10" "BB11" "BB12" "BB13" "AN1"  "AN2"  "AN3"  "AN4"  "AN5"  "AN6"
[91] "AN7"  "AN8"  "AN9"  "AN10" "AN11" "AN12"
> sample(sites, size=16)
 [1] "FW7"  "BB7"  "AN1"  "MB5"  "BB13" "BB9"  "BB8"  "RN13" "BB6"  "AN11"
[11] "MB12" "RN4"  "RN5"  "RB4"  "RB3"  "MB7"
```

Our simple random sample of 16 sites would: "FW7" "BB7" "AN1" "MB5" "BB13" "BB9" "BB8" "RN13" "BB6" "AN11" "MB12" "RN4" "RN5" "RB4" "RB3" "MB7"

Using SAS

SAS program to obtain a SRS of $n = 16$ sites.

```
* SAS code to build SRS;
data soilsrs;
attrib city length=$17;
input city $ siteID $ @@;
datalines;
Untertiefenbach UT01    Untertiefenbach UT02
Untertiefenbach UT03    Untertiefenbach UT04
    ...
Untertiefenbach UT15    Untertiefenbach UT16
```

```

Reisenberg RB01      Reisenberg RB02      Reisenberg RB03
      ...
Arnoldstein AN10    Arnoldstein AN11    Arnoldstein AN12
;
proc surveystest data=soilsrs method=srs  sampsize=16
                    seed=62656      out=sample81;
proc print data=sample81;

```

Edited SAS output of SRS for Soil Contamination Study

```

                The SAS System
                The SURVEYSELECT Procedure
Selection Method   Simple Random Sampling
Input Data Set      EX81
Random Number Seed 62656
Sample Size        16
Output Data Set    SAMPLE81

Obs   city      site      Obs   city      site
  1  Untertiefenbach  UT02   2  Untertiefenbach  UT04
  3  Reisenberg      RB05   4  Reisenberg      RB10
  5  Reisenberg      RB12   6  Feistritz       FW05
  7  Mitterberghuetten MB08   8  Ramingstein     RN02
  9  Ramingstein     RN03  10  Bleiberg        BB01
 11  Bleiberg        BB02  12  Bleiberg        BB04
 13  Bleiberg        BB08  14  Bleiberg        BB09
 15  Arnoldstein     AN05  16  Arnoldstein     AN11

```

Let's assume we've selected the sites and obtain the Cr concentrations from the 16 sites. Below is the table of resulting sample in which the Cr values have been log transformed to yield a more symmetric population.

Log-chromium (Cr) concentrations in soil from $n = 16$ Austrian sites. Sites taken from population above.

Site	UT7	UT14	RB5	RB11	FW2	FW9	MB9	MB10
log(Cr)	3.761	3.807	4.317	2.833	3.258	4.205	3.611	2.303

Site	RN1	RN7	BB8	BB11	BB12	AN2	AN5	AN12
log(Cr)	3.434	3.367	3.970	2.565	3.091	3.611	3.466	3.434

Now let's use R and SAS to obtain our estimate of the mean log(Cr), its se, and an approx 95% confidence interval.

```

> Cr *Note: Data entered via importing Excel File
  Site log.Cr.
1  UT7  3.761
2  UT14 3.807
3  RB5  4.317
4  RB11 2.833
5  FW2  3.258
6  FW9  4.205
7  MB9  3.611
8  MB10 2.303
9  RN1  3.434
10 RN7  3.367
11 BB8  3.970
12 BB11 2.565
13 BB12 3.091
14 AN2  3.611
15 AN5  3.466
16 AN12 3.434
> avglogCr = mean(Cr$log.Cr.)
> sdlogCr = sd(Cr$log.Cr.)
> avglogCr
[1] 3.439563
> sdlogCr
[1] 0.5488963
> bound=2*sdlogCr*sqrt((96-16)/96)/sqrt(16)
> bound
[1] 0.2505358
> seAvglogCr = sdlogCr*sqrt((96-16)/96)/sqrt(16)
> seAvglogCr
[1] 0.1252679
> avglogCr + qt(0.025, 15, lower.tail = FALSE)*c(-seAvglogCr, +seAvglogCr)
[1] 3.172560 3.706565

```

GOAL #1: Estimate mean $\log(\text{Cr})$ level for the $N = 96$ sites.

RESULTS:

Point estimate is $\bar{y} = 3.439563 = 3.4396$

$se(\bar{y}) = 0.1253$

Approx 95% CI for $\mu = \text{mean } \log(\text{Cr})$ level is [3.1726 , 3.7066]

Bound (aka Bound on the Error of Estimation) = $B = 0.2505$

Using SAS

SAS code to estimate population mean μ with Soil Contamination Data from finite population of size $N = 96$.

```
data soil96;
attrib city length=$17;
input city $ siteID $ YlogCr @@;
datalines;
Untertiefenbach   UT07 3.761   Untertiefenbach   UT14 3.807
Reisenberg        RB05 4.317   Reisenberg        RB11 2.833
Feistritz         FW02 3.258   Feistritz         FW09 4.205
Mitterberghuetten MB09 3.611   Mitterberghuetten MB10 2.303
Ramingstein       RN01 3.434   Ramingstein       RN07 3.367
Bleiberg          BB08 3.970   Bleiberg          BB11 2.565
Bleiberg          BB12 3.091   Arnoldstein       AN02 3.611
Arnoldstein       AN05 3.466   Arnoldstein       AN12 3.434
;
proc surveymeans total=96 mean df clm;
var YlogCr;
```

SAS output of estimate of mean soil contamination.

The SAS System					
The SURVEYMEANS Procedure					
Data Summary					
Number of Observations 16					
Variable	DF	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95%
YlogCr	15	3.439563	0.125268	3.172560	3.706565

Systematic Sampling (1-in-k Systematic Sample)

Defn: Systematic Sample = from a frame arranged in random order, select every k^{th} element, after randomly selecting from the first k elements.

PROCEDURE

1. Select the first sample element in a random fashion from among the first k elements in the frame
2. take every k^{th} element from the frame, where $k \leq N/n$

ESTIMATING PARAMETERS IN A SRS

Estimation of population quantities such as μ , τ , or π proceeds in a manner similar to that for simple random sampling.

If the population and frame are arranged randomly so that no underlying periodicity or other form of ordered structure is present, then the variance estimators and standard errors from the SRS case will also be approximately valid for systemic samples.

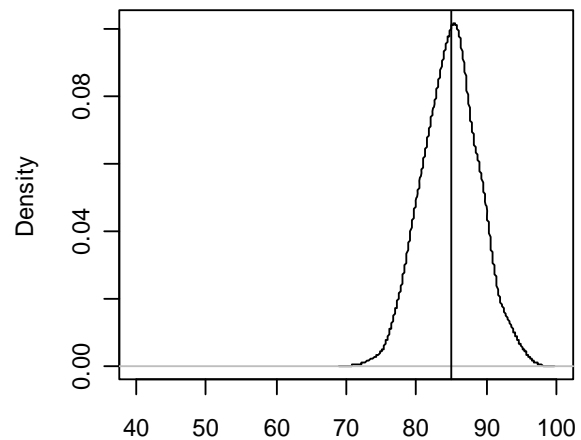
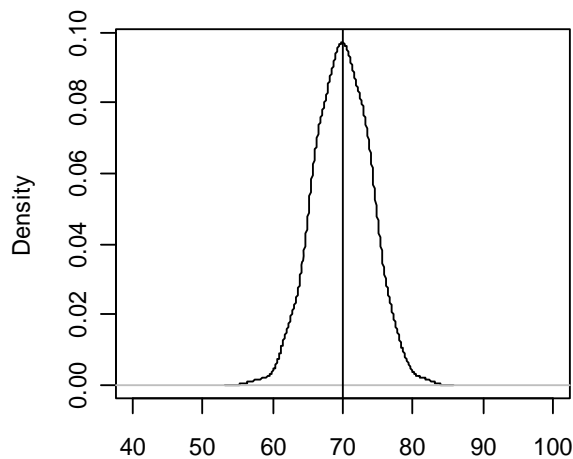
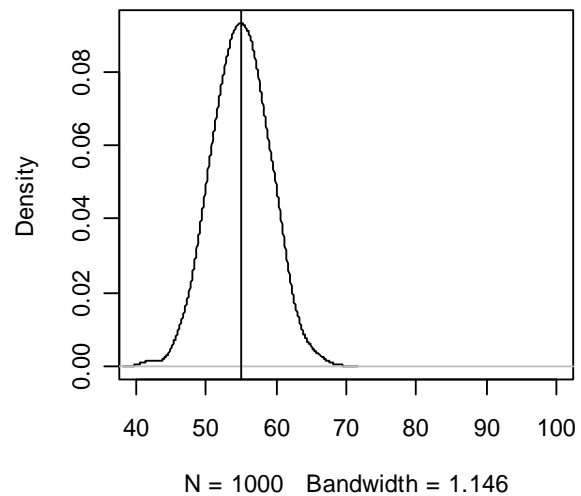
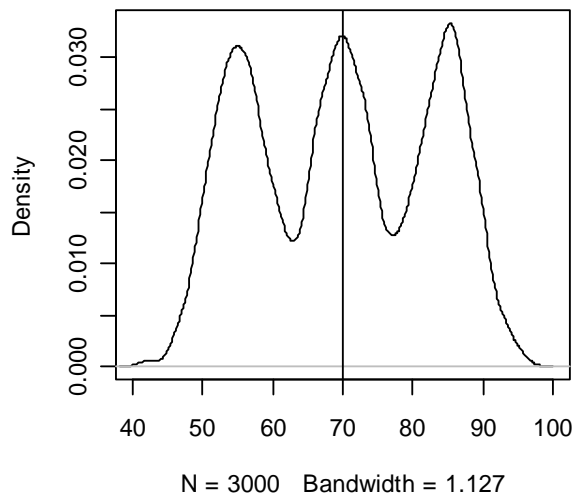
NOTES/COMMENTS

1. **WARNING:** Although systematic samples are easy to implement, they can be detrimentally affected by hidden periodicities in the frame.
2. PROC SURVEYSELECT can be used to draw a systematic sample by using the `method=sys` option in the procedure call. If `method=sys` is specified, use of the option `samprate=f` will specify the sampling interval as $k = 1/f$. One can also use the `sampsize=` option to select n , forcing PROC SURVEYSELECT to set k equal to N/n .

Stratified Random Sampling

Defn: Stratified Random Sample = assuming the population elements can be grouped into STRATA (groups in which the measurements are homogeneous), simple random samples are selected from each of the STRATA.

EXAMPLE



```
> par(mfrow=c(2,2))
> hgts = c(rnorm(1000, 55, 4), rnorm(1000, 70, 4), rnorm(1000, 85, 4))
> plot(density(hgts, bw="sj"), xlim=c(40,100), main="")
> abline(v=70)
> plot(density(hgts[1:1000], bw="sj"), xlim=c(40,100), main="")
> abline(v=55)
> plot(density(hgts[1001:2000], bw="sj"), xlim=c(40,100), main="")
> abline(v=70)
> plot(density(hgts[2001:3000], bw="sj"), xlim=c(40,100), main="")
> abline(v=85)
```

VOCABULARY/NOTATION

Stratum = a group of homogeneous elements of the population

H = number of strata in the population

N_h = number of elements in the h^{th} stratum, $h = 1, 2, \dots, H$

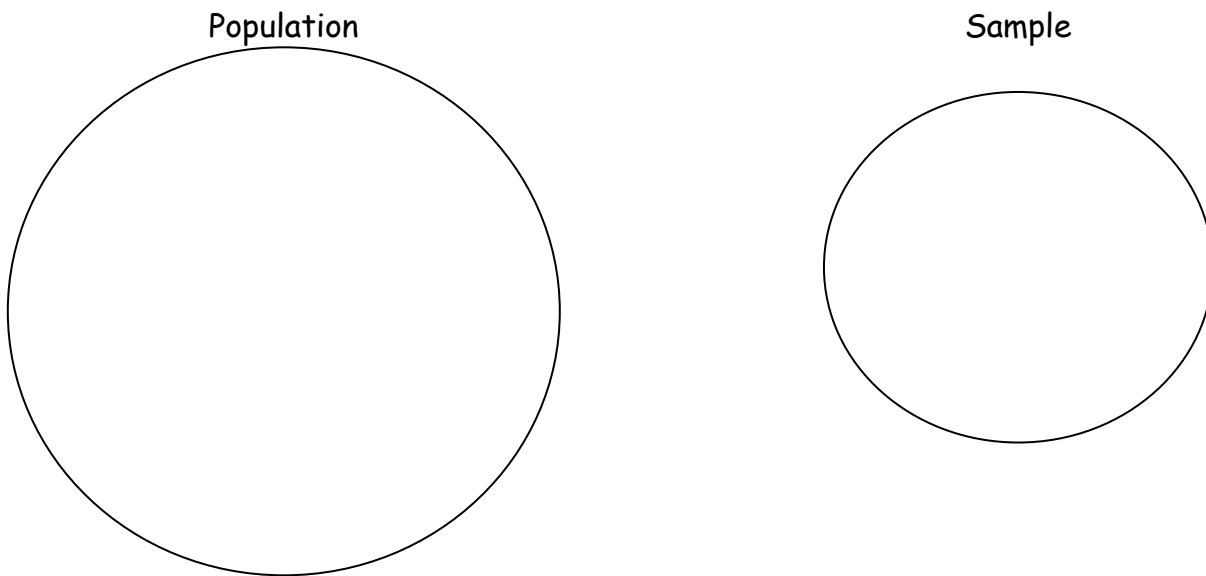
μ_h = mean measurement in the h^{th} stratum, $h = 1, 2, \dots, H$

$\tau_h = N_h \mu_h$ = total of measurements in the h^{th} stratum, $h = 1, 2, \dots, H$

n_h = number of sample elements selected from the h^{th} stratum, $h = 1, 2, \dots, H$

N = total population size = $\sum_{h=1}^H N_h$

n = total sample size = $\sum_{h=1}^H n_h$



PROCEDURE

Take a SRS of n_h from N_h elements in each of the H strata

FACTS CONCERNING THE PARAMETERS OF INTEREST (μ and τ)

$$\tau = \sum_{h=1}^H \tau_h \quad \mu_h = \frac{\tau_h}{N_h}, \quad h = 1, \dots, H$$

$$\mu = \frac{\tau}{N} = \frac{\sum_{h=1}^H \tau_h}{N} = \frac{\sum_{h=1}^H N_h \mu_h}{N} = \frac{1}{N} \sum_{h=1}^H N_h \mu_h = \sum_{h=1}^H \frac{N_h}{N} \mu_h = \sum_{h=1}^H W_h \mu_h$$

ESTIMATING PARAMETERS IN A STRATIFIED RANDOM SAMPLE

1. ESTIMATING THE PARAMETERS

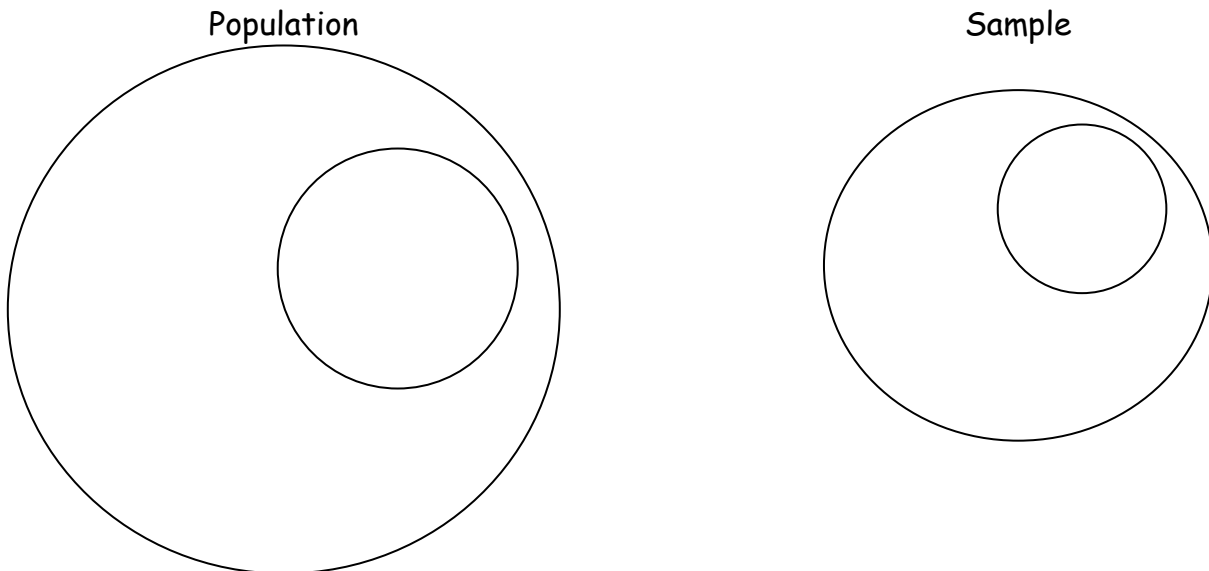
a. ESTIMATING A STRATUM MEAN (μ_h) UNDER STRATIFIED RS

i. Point Estimate of μ_h is $\hat{\mu}_h = \bar{y}_h = \frac{\sum_{h=1}^{n_h} y_{ih}}{n_h}$

ii. Standard Error of $\hat{\mu}_h (= \bar{y}_h) = se(\bar{y}_h) = \sqrt{\frac{s_h^2}{n_h} \left(\frac{N_h - n_h}{N_h} \right)}$.

iii. CI for μ_h : approximate $1 - \alpha$ design-based confidence interval for μ_h is

$$\bar{y}_h \pm z_{\frac{\alpha}{2}} se(\bar{y}_h) \quad \text{or} \quad \bar{y}_h \pm t_{(\frac{\alpha}{2}, n_h - 1)} se(\bar{y}_h).$$



b. ESTIMATING A POPULATION MEAN (μ) UNDER STRATIFIED RS

i. Point Estimate of μ is $\hat{\mu} = \sum_{h=1}^H W_h \bar{y}_h = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h$

ii. Standard Error of $\hat{\mu} = se(\hat{\mu}) = \sqrt{\text{Var}(\hat{\mu})} = \sqrt{\sum_{h=1}^H \text{Var}(W_h \bar{y}_h)}$

$$= \sqrt{\sum_{h=1}^H W_h^2 \text{Var}(\bar{y}_h)} = \sqrt{\sum_{h=1}^H W_h^2 \frac{s_h^2}{n_h} \left(\frac{N_h - n_h}{N_h} \right)}$$

$$= \sqrt{\sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{s_h^2}{n_h} \left(\frac{N_h - n_h}{N_h} \right)}$$

$$= \sqrt{\sum_{h=1}^H \frac{N_h^2}{N^2} \frac{s_h^2}{n_h} \left(\frac{N_h - n_h}{N_h} \right)}$$

iii. CI for μ : approximate $1 - \alpha$ design-based confidence interval for μ is

$$\hat{\mu} \pm z_{\frac{\alpha}{2}} se(\hat{\mu}) \quad \text{or} \quad \hat{\mu} \pm t_{\left(\frac{\alpha}{2}, df\right)} se(\hat{\mu}) \quad \text{where } df = \sum_{h=1}^H (n_h - 1).$$

c. ESTIMATING THE POPULATION TOTAL (τ) UNDER STRATIFIED RS

i. Point Estimate of τ is $\hat{\tau} = N\hat{\mu} = N \sum_{h=1}^H W_h \bar{y}_h = \sum_{h=1}^H N_h \bar{y}_h$

ii. Standard Error of $\hat{\tau} = \hat{\tau} = se(\hat{\tau}) = Nse(\hat{\mu})$.

iii. CI for τ : approximate $1 - \alpha$ design-based confidence interval for τ is

$$\hat{\tau} \pm z_{\frac{\alpha}{2}} se(\hat{\tau}) = N\hat{\mu} \pm z_{\frac{\alpha}{2}} Nse(\hat{\mu})$$

or

$$\hat{\tau} \pm t_{\left(\frac{\alpha}{2}, df\right)} se(\hat{\tau}) = N\hat{\mu} \pm t_{\left(\frac{\alpha}{2}, df\right)} Nse(\hat{\mu}) \quad \text{where } df = \sum_{h=1}^H (n_h - 1).$$

NOTES/COMMENTS

1. One advantage of Stratified Random Sampling is that estimates of parameters in the different STRATA are available. Note, however, that these estimates will lack the precision of the population parameters, since the strata sample sizes will be much smaller!
2. How large a sample is needed so that the margin of error from an approximate 95% confidence interval for μ is no larger than, say, E units?

$$n \geq \frac{\sum_{h=1}^H \frac{N_h^2 \sigma_h^2}{w_h}}{\sum_{h=1}^H N_h^2 \sigma_h^2 + \frac{1}{4} N^2 E^2} \quad \text{where } w_h = \frac{n_h}{n}.$$

3. One commonly used method of allocating sample sizes to the strata is one known as **PROPORTIONAL ALLOCATION**. Under proportional allocation, the sample size for the h^{th} stratum is in proportion to the size of the stratum, N_h .

$$n_h = n \frac{N_h}{N}, \quad h = 1, 2, \dots, H.$$

4. How large a sample is needed so that the margin of error from an approximate 95% confidence interval for μ is no larger than, say, E units, WHEN THE SAMPLE SIZES ARE ALLOCATED PROPORTIONALLY?

$$n \geq \frac{\sum_{h=1}^H N_h \sigma_h^2}{\frac{1}{N} \sum_{h=1}^H N_h \sigma_h^2 + \frac{1}{4} N E^2}.$$

5. Stratified sampling and blocking

Stratified sampling is closely related to the concept of *blocking* in experimental design. In general, blocking is advocated as a means for decreasing variation: one first subdivides the experimental units into relatively homogeneous subgroups based on some *blocking factor*, and next randomly assigns units within blocks to the experimental/treatment levels. The blocking factor is a known source of variability that affects the population in a specific, deterministic manner. When left unrecognized, it will contribute to experimental error in the observations, but by recognizing and incorporating blocking factors into the design, we reduce error

variance and improve our ability to detect true differences among the treatment effects.

Examples

Using R

To use R to obtain Stratified Random Samples, we need only use the "sample" function within each strata.

Using SAS

PROC SURVEYSELECT. Note the construction of a special SAS data set, `allocate`, used to specify the unbalanced allocation scheme via the `sampsize=` option. SAS requires that the variable `_NSIZE_` be used to contain the n_h s. The stratified sample is generated via use of the `strata` statement in PROC SURVEYSELECT, which defines the stratification variable. (Both the source data set and the allocation data set must be sorted prior to use of `strata`.)

SAS program to obtain a stratified random sample of size $n = 25$.

```
data insect;
input forest siteID @@;
datalines;
1 00 1 01 1 02 1 03 1 04 1 05 1 06 1 07
1 08 1 09 1 10 1 11 1 12 1 13 1 14 1 15
...
4 88 4 89 4 90 4 91 4 92 4 93 4 94 4 95
4 96 4 97 4 98 4 99

proc sort data=insect;
  by forest;

data allocate;
input forest _NSIZE_ @@;
datalines;
  1 6      2 6      3 6      4 7

proc sort data=allocate;
  by forest;

proc surveyselect data=insect method=srs sampsize=allocate
  seed=100956 out=sample25;
  strata forest;

proc print data=sample25;
```

SAS output of stratified random sample for Insect Abundance Study.

The SAS System					
The SURVEYSELECT Procedure					
Selection Method		Simple Random Sampling			
Strata Variable		forest			
Input Data Set		EX84			
Random Number Seed		100956			
Sample Size Data Set		ALLOCATE			
Number of Strata		4			
Total Sample Size		25			
Output Data Set		SAMPLE84			

Obs	forest	site ID	Obs	forest	site ID
1	1	07	2	1	26
3	1	30	4	1	36
5	1	64	6	1	83
7	2	00	8	2	12
9	2	28	10	2	45
11	2	87	12	2	96
13	3	26	14	3	29
15	3	50	16	3	53
17	3	65	18	3	78
19	4	14	20	4	25
21	4	27	22	4	38
23	4	58	24	4	78
25	4	95			

SAS program to estimate population mean μ via stratified random sample with Phosphorus Concentration Data.

```

data phosphor;
input depth $ Yconc @@;
  if depth='surf' then theta=3940000/12;
  if depth='inte' then theta=3200000/10;
  if depth='bot' then theta=2500000/8;
datalines;
surf 1.1 surf 1.2 surf 1.4 surf 1.5 surf 1.7 surf 1.9
surf 2.1 surf 2.1 surf 2.3 surf 2.5 surf 3.0 surf 3.1
inte 2.9 inte 3.0 inte 3.3 inte 3.6 inte 3.7 inte 4.0
inte 4.1 inte 4.4 inte 4.5 inte 4.8
bot 3.4 bot 3.9 bot 5.1 bot 5.3 bot 5.7 bot 5.8
bot 6.0 bot 6.9
data StrTotal;
input depth $ _total_ @@;
datalines;
surf 3940000 inte 3200000 bot 2500000
;
proc surveymeans data=phosphor total=StrTotal mean df clm;
stratum depth ;
var Yconc;
weight theta;

```

SAS output for estimating μ via stratified random sample with Phosphorus Concentration Data

The SAS System					
The SURVEYMEANS Procedure					
Variable	DF	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95%
Yconc	27	3.450147	0.146356	3.149850	3.750444