Module 2: Environmental Sampling

2.3 Stratified Random Sampling

Stratified Random Sampling

- Stratified random sampling involves splitting the population into sections, or strata, and choosing a random sample from each stratum.
- It is appropriate when population units are more similar within each strata than they are across strata.

Stratified Random Sampling

- Populations of people are often stratified by age, sex, geographic location, political party, or other important variables.
- Environmental samples are often stratified by land type, terrain, geography, geology, land use, zones of contamination, and so forth.

Stratified Random Sampling

- Advantages of stratification:
  - You can calculate separate estimates of the parameters for each stratum. If the strata are different from one another on the characteristic under study (contamination for example), you may make different management decisions for different strata.

Stratified Random Sampling

- Advantages of stratification:
  - Different strata can be sampled more or less intensively depending on study goals and population characteristics. For example, areas expected to be more variable should be sampled more intensively. Or possibly areas suspected of more contamination should be sampled more intensively.
  - The standard error of the mean will be smaller than for SRS, particularly if the strata are quite different from one another.

Module 2.3 5

Stratified Random Sampling

- Disadvantages of stratification:
  - Usually make decisions on stratification before the study is carried out and these choices may turn out to be incorrect
  - Stratification can complicate later data use
  - Data analysis is more complicated

Module 2.3 6

Notation

- K = Number of strata
- N_i = size of the i^{th} strata population
- N = $\sum_{i=1}^{K} N_i$ = size of total population
- n_i = size of the i^{th} strata sample
- n = $\sum_{i=1}^{K} n_i$ = size of the total sample

Module 2.3

Sample Statistics from Stratified Random Sampling

- μ_i = population mean of the i^{th} strata
- $\overline{y}_i$ = sample mean of the i^{th} strata
- s_i = sample standard deviation of the i^{th} strata
- The sample strata mean and standard deviation are calculated in the normal way

Module 2.3

Sample Statistics from Stratified Random Sampling

- $\overline{y}_i$ has sample standard error

$$SE(\overline{y}_i) = \sqrt{\left(\frac{s_i^2}{n_i}\right)\left(1 - \frac{n_i}{N_i}\right)}$$

Module 2.3

9

Sample Statistics from Stratified Random Sampling

The overall mean is a weighted mean,

$$\overline{y}_s = \sum_{i=1}^{K} \frac{N_i \overline{y}_i}{N} = \sum_{i=1}^{K} w_i \overline{y}_i$$

where the weights w_i are the proportion of the population in the i^{th} strata = N_i/N

Module 2.3

10

Sample Statistics from Stratified Random Sampling

$\overline{y}_s$ has sample standard error

$$S\hat{E}(\overline{y}_s) = \sqrt{\sum_{i=1}^{K}\left(\frac{N_i}{N}\right)^2\left(\frac{s_i^2}{n_i}\right)\left(1-\frac{n_i}{N_i}\right)}$$

Module 2.3

1

Sample Statistics from Stratified Random Sampling

- An approximate 100(1-α)% Confidence Interval for μ is

$$\overline{y}_s \pm Z_{\alpha/2}\, S\hat{E}(\overline{y}_s)$$

Module 2.3

2

Example

- Let's do an example
- Three strata

	N_i	n_i	w_i	Y bar_i	s_i^2
Strata 1	1000	10	0.0625	5.3	1.6
Strata 2	5000	50	0.3125	2.8	1.1
Strata 3	10000	100	0.625	1.4	0.8
	16000	160			

Module 2.3

Overall Mean

$$\overline{y}_s = \sum_{i=1}^{K} w_i \overline{y}_i$$

	w_i	Y bar_i	w_i*Y bar_i
Strata 1	0.0625	5.3	0.33125
Strata 2	0.3125	2.8	0.875
Strata 3	0.625	1.4	0.875
		Overall mean	2.08125

Module 2.3

14

Standard Error of the Overall Mean

$$S\hat{E}(\overline{y}_s) = \sqrt{\sum_{i=1}^{K}\left(\frac{N_i}{N}\right)^2\left(\frac{s_i^2}{n_i}\right)\left(1-\frac{n_i}{N_i}\right)}$$

	w_i	w_i^2	N_i	n_i	s_i^2	s_i^2/n_i	$1-n_i/N_i$	Product
Strata 1	0.0625	0.0039	1000	10	1.6	0.16	0.99	0.0006
Strata 2	0.3125	0.0977	5000	50	1.1	0.022	0.99	0.0021
Strata 3	0.625	0.3906	10000	100	0.8	0.008	0.99	0.0031
							Sum =	0.0058

- Standard Error of overall mean = sqrt(0.0058) = 0.076416

Module 2.3

95% Confidence Interval on the True Overall Mean

$$\overline{y}_s \pm Z_{\alpha/2} S\hat{E}(\overline{y}_s)$$

- $\overline{y}_s =$ 5 2.08125
- $Z_{\alpha/2} =$ 1.96
- $S\hat{E}(\overline{y}_s) =$ 0.076416

Module 2.3

95% Confidence Interval on the True Overall Mean

- = 2.08125 +/- 1.96*0.076416
- =(1.93, 2.23)

- So, 95% of confidence intervals constructed this way from repeated sampling would contain the true mean μ

Module 2.3