

# Lecture 19 Double and Two Phase Sampling

---

- ◆ Introduction
- ◆ Ratio Estimator
- ◆ Regression Estimator (Very Brief)
- ◆ Stratification and Adjusting for Non Response
- ◆ Cluster Sampling (Very Brief)
- ◆ Ecological Examples
- ◆ Summary Remarks
- ◆ Sampling Rare and Clustered Populations:  
Adaptive Sampling (Very Brief)

## Improving Sampling Designs with Auxiliary Information

```
graph TD; A[Improving Sampling Designs with Auxiliary Information] --> B[Regression Methods]; A --> C[Stratified Random Sampling]; B --> D[Double or Two-Phase Sampling]; C --> D;
```

### Regression Methods

Use relationship between  $y$  and  $x$  in **linear regression models**.

### Stratified Random Sampling

Use auxiliary variable (s) to set up roughly **homogeneous groups** or **strata**

### Double or Two-Phase Sampling

Practical Advantages as sometimes not enough information on Frame to use either directly without Two-phase sampling

# Double Sampling

**Inexpensive Sampling Technique-Large Sample**

**Expensive Sampling Technique-Small Subsample  
from the larger Sample.**

**Example: Fish Sampling**

**Large sample: Make measurements that are quick and inexpensive like length weight etc**

**Small Subsample: Make more detailed time consuming and expensive measurements (blood, DNA etc) to get better information but at a higher cost.**

# Double Sampling: General Idea

---

## Two Phases

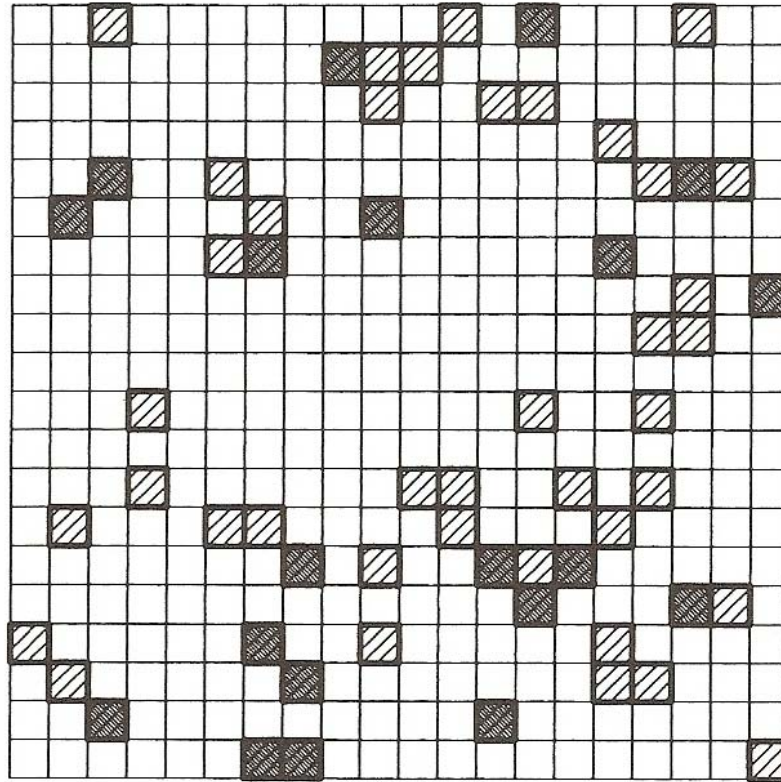
**Phase 1-** Take a large sample and collect some auxiliary data (say  $x$  if generalising a ratio or regression estimate)

**Phase 2-** Take a smaller sample usually a subsample of the first phase sample and collect some additional data on the variable of interest (say  $y$  if generalising a ratio or regression estimate)

# Double Sampling: General Idea

158

DOUBLE OR TWO-PHASE SAMPLING



**Figure 14.1.** Double sample. The variable of interest and an auxiliary variable are both recorded on the double-shaded (dark) units. On single-shaded units, only the auxiliary variable is observed.

# Double Sampling: General Idea

---

May be used to enable:

ratio estimation,

regression estimation,

stratification,

many other procedures.

We will start with ratio and regression estimators

# **Regression Methods Review**

**Regression Methods  
Improve Precision if Auxiliary  
Variable x is available**

**Linear Regression thru Origin**

Ratio Estimator (7)  
Estimators in Text  
Model

$$y_i = \beta x_i + \varepsilon_i$$

$$\hat{\mu}_r = r\mu_x = \left[\frac{\bar{y}}{\bar{x}}\right]\mu_x$$

Regression thru origin with errors increasing  
with x. Discussed in class and text.

$$\hat{\tau}_r = N\hat{\mu}_r$$

**Standard Linear Regression**

Regression Estimator (8)  
Estimators in Text  
Model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\hat{\mu}_L = a + b\mu_x$$

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x})$$

a and b standard least squares estimators  
of intercept and slope

$$\hat{\tau}_L = N\hat{\mu}_L$$

$$\hat{\mu} = \bar{y}$$

$$\hat{\tau} = N\bar{y}$$

If we ignore the x's then we lose precision!!.

## Linear Regression thru Origin

Ratio Estimator (Ch 7)

Model  $y_i = \beta x_i + \varepsilon_i$

Important Results

$$\hat{\mu}_r = r\mu_x = \left[\frac{\bar{y}}{\bar{x}}\right]\mu_x$$

$\hat{\mu}_r = \bar{y}\left[\frac{\mu_x}{\bar{x}}\right]$  This shows the motivation for the estimator

$$\hat{\tau}_r = N\hat{\mu}_r$$

$$\hat{Var}(\hat{\mu}_r) = \left[\frac{N-n}{N}\right]\frac{s_r^2}{n}$$

$$\hat{Var}(\hat{\tau}_r) = N^2\hat{Var}(\hat{\mu}_r)$$

$$s_r^2 = \frac{1}{n-1} \sum_1^n (y_i - rx_i)^2$$

$$\hat{\mu}_r \pm t_{n-1}(\alpha/2)\sqrt{\hat{Var}(\hat{\mu}_r)}$$

Key Point – Need  $\mu_x$

## Standard Linear Regression

Regression Estimator Popn Mean  
Estimators in Text  
Model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\hat{\mu}_L = a + b\mu_x$$

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x})$$

Least Squares Estimates

$$a = \bar{y} - \bar{b}x$$

$$b = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2}$$

## Standard Linear Regression

Regression Estimator Popn Mean

Variance Calculation

Model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x})$$

$$\text{Var}(\hat{\mu}_L) = \frac{N - n}{N} \frac{s_e^2}{n}$$

$$s_e^2 = \frac{\sum_{i=1}^n (y_i - a - bx_i)^2}{n - 2}$$

Key Point – Need  $\mu_x$

## Standard Linear Regression

Regression Estimator Popn Total  
Estimators in Text  
Model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x})$$

$$\hat{\tau}_L = N\hat{\mu}_L$$

$$\text{Var}(\hat{\tau}_L) = N^2 \text{Var}(\hat{\mu}_L)$$

# Good Example of the Ratio Estimator

- $y$ -Yield of crop,  $x$  –Size of farm
- $y$  small then  $x$  small so no intercept needed (Regression thru the origin example)
- $y$  is linearly related to  $x$ .
- All the  $x$ 's on the frame available from a prior Survey

## Linear Regression thru Origin

Ratio Estimator (Ch 7.1)

Updated to Two Phase (Ch 14.1)

$$\hat{\tau}_r = r \tau_x = \left[ \frac{\bar{y}}{\bar{x}} \right] \tau_x$$

Two Phase Version

$$\hat{\tau}_r = r \hat{\tau}_x = \hat{\tau}_x \left[ \frac{\bar{y}}{\bar{x}} \right]$$

Phase1 - Sample  $n'$

$$\text{Estimate } \hat{\tau}_x = \frac{N}{n'} \sum_1^{n'} x_i$$

Phase2 - Sample  $n$  from  $n'$

$$\text{Estimate } \left[ \frac{\bar{y}}{\bar{x}} \right]$$

## Linear Regression thru Origin

Two Phase (Ch 14.1)

Variance Components

$$\hat{\tau}_r = r\tau_x = \left[\frac{\bar{y}}{\bar{x}}\right]\tau_x$$

$$\text{var}(\hat{\tau}_r) \approx N(N - n') \frac{\sigma^2}{n'} + N^2 \frac{(n' - n)}{n'} \frac{\sigma_r^2}{n}$$

$$\text{var}(\hat{\tau}_r) \approx N(N - n') \frac{s^2}{n'} + N^2 \frac{(n' - n)}{n'} \frac{s_r^2}{n}$$

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

$$s_r^2 = \frac{\sum_{i=1}^n (y_i - rx_i)^2}{n - 1}$$

## Linear Regression thru Origin

Two Phase (Ch 14.1)

Variance Special Cases

$$\hat{\tau}_r = r\tau_x = \left[\frac{\bar{y}}{\bar{x}}\right]\tau_x$$

$$\text{var}(\hat{\tau}_r) \approx N(N - n') \frac{\sigma^2}{n'} + N^2 \frac{(n' - n)}{n'} \frac{\sigma_r^2}{n}$$

$n' \rightarrow N$  Standard Ratio Estimator Variance

$$\text{var}(\hat{\tau}_r) \approx N(N - n) \frac{\sigma_r^2}{n}$$

$n \rightarrow n'$  Standard Estimator Variance with sample of  $n'$

$$\text{var}(\hat{\tau}_r) \approx N(N - n') \frac{\sigma^2}{n'}$$

## Linear Regression thru Origin

Ratio Estimator (Ch 7.1)

Updated to Two Phase (Ch 14.1)

Two Phase Example - Aerial Survey P159 Example 1.

$$\hat{\tau}_r = r \hat{\tau}_x = \hat{\tau}_x \left[ \frac{\bar{y}}{\bar{x}} \right]$$

Phase 1 - Sample  $n' = 20$

$$\text{Estimate } \hat{\tau}_x = \frac{N}{n'} \sum_1^{n'} x_i = \frac{100}{20} 240 = 1200$$

Phase 2 - Sample  $n = 5$  from  $n' = 20$

$$\text{Estimate } \left[ \frac{\bar{y}}{\bar{x}} \right] = \frac{70/5}{56/5} = 1.25$$

$$\hat{\tau}_r = \hat{\tau}_x \left[ \frac{\bar{y}}{\bar{x}} \right] = 1200 \times 1.25 = 1500$$

## Linear Regression thru Origin

Ratio Estimator (Ch 7.1)

Updated to Two Phase (Ch 14.1)

Two Phase Example - Aerial Survey P159 Example 1.

$$\hat{\tau}_x = \frac{N}{n'} \sum_1^{n'} x_i = \frac{100}{20} 240 = 1200$$

Moose in Population

$$\hat{\tau}_r = \hat{\tau}_x \left[ \frac{\bar{y}}{\bar{x}} \right] = 1200 \times 1.25 = 1500$$

Aerial Detection Probability

$$1200/1500 = 0.8.$$

## Linear Regression thru Origin

Ratio Estimator (Ch 7.1)

Updated to Two Phase (Ch 14.1)

## Optimal Allocation 14.2 P 160

### Cost Function

$$C = c'n' + cn$$

Minimise  $\text{var}(\hat{\tau}_r)$  subject to fixed cost

$$\frac{n}{n'} = \sqrt{\left[\frac{c'}{c}\right] \left[\frac{\sigma_r^2}{\sigma^2 - \sigma_r^2}\right]}$$

$$n' = \frac{C}{c' + c \sqrt{\left[\frac{c'}{c}\right] \left[\frac{\sigma_r^2}{\sigma^2 - \sigma_r^2}\right]}}$$

## Linear Regression thru Origin

Ratio Estimator (Ch 7.1)

Updated to Two Phase (Ch 14.1)

### Example Optimal Allocation

$$C = 12,000, c' = 100, c = 900, \sigma_r^2 = 1, \sigma^2 = 5$$

$$\frac{n}{n'} = \sqrt{\left[\frac{c'}{c}\right]\left[\frac{\sigma_r^2}{\sigma^2 - \sigma_r^2}\right]} = \sqrt{\left[\frac{100}{900}\right]\left[\frac{1}{5-1}\right]} = \frac{1}{6}$$

$$n_{opt}' = \frac{C}{c' + c \sqrt{\left[\frac{c'}{c}\right]\left[\frac{\sigma_r^2}{\sigma^2 - \sigma_r^2}\right]}} = \frac{12000}{100 + 900 \times \frac{1}{6}}$$

$$n_{opt}' = 48 \text{ and } n_{opt} = 8$$

# **Standard Linear Regression**

Regression Estimator Ch 8

Two Phase Regression Estimator

## Standard Linear Regression

Regression Estimator Ch 8

Two Phase Regression Estimator

Not in Text

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x})$$

Two Phase Version

$$\hat{\mu}_L = \bar{y} + b(\hat{\mu}_x - \bar{x})$$

Phase 1

$$\hat{\mu}_x$$

Phase 2

$$\bar{y}, b, \bar{x}$$

## Stratification

Basic Estimators Ch 11

Two Phase Stratified Estimator 14.3

Phase 1  $n'$  units

$$w_h = \frac{n'_h}{n'}$$

Phase 2  $n_h$  out of  $n'_h$  units

measure the  $y_h$ 's.

Then we have

$$\hat{y}_d = \sum_{h=1}^L w_h \bar{y}_h$$

## **Two Phase Stratification Approach**

Use Ideas to account  
for Non Response in a Mail Survey

Phase 1 – Mail Survey

Response and Nonresponse Strata

Phase 2 – Followup Telephone Survey of Nonrespondents

## Two Phase Stratification Approach

Use Ideas to account  
for Non Response in a Mail Survey

$$\hat{y}_d = \sum_{h=1}^2 w_h \bar{y}_h$$

$$\hat{y}_d = \frac{n_1'}{n} \bar{y}_1 + \frac{n_2'}{n} \bar{y}_2$$

$\bar{y}_1$  – Mail Survey

$\bar{y}_2$  – Telephone Survey of Nonrespondents  
to Mail Survey

## Two Phase Cluster Sampling Idea

$$\hat{\tau}_r = \sum_{i=1}^n y_i \frac{\sum_{i=1}^N M_i}{\sum_{i=1}^n M_i}$$

$\sum_{i=1}^N M_i$  could be estimated by double sampling!

# Double Sampling

**Animal Sampling and Detectability (More in ST 506)**

**Inexpensive Sampling -Large Sample ( $n'$ )**

**Simple animal counts**

**Expensive Sampling -Small Subsample ( $n$ )**

**More detailed sampling so that detection probability can be estimated**

# Double Sampling

## **Illustrations:**

### **Aerial Surveys (already seen)**

**Aerial counts on all plots, ground counts on a subsample of plots.**

### **Larissa Bailey Salamander Study.**

**Counts on all plots and detailed capture recapture studies on a small sample of plots to estimate detection probability.**

### **Electrofishing in Streams**

**Single pass all streams and multipass in a small sample of streams.**

# Double Sampling Summary

Double Sampling is a very useful technique in Animal sampling

Conceptual  
Equation

$$\hat{N} = \frac{C}{\hat{p}}$$

Get C all plots  
Cheap

Estimate p only on some  
plots  
Expensive

# Double Sampling Summary

**Double Sampling is a very useful technique. Here are key results based on the paper referenced.**

## Salamanders

- 100 plots count salamanders
- Subsample 7 out of 100 plots for a detailed mark recapture study to estimate detection probability

# Double Sampling Summary

**Double Sampling is a very useful technique. Here are key results based on the paper referenced.**

## Stream Survey

- 100 streams one pass
- Subsample 35 out of 100 streams for a  
3 pass removal study to estimate detection  
probability

# Double Sampling

---

**Reference:** Pollock ,K. H., Nichols, J. D., Simons, T.R., Farnsworth, G.L., Bailey, L.L., and Sauer, J.R. (2002). The design of large scale wildlife monitoring studies. *Environmetrics*, 13: 105-119.

# Double Sampling Summary

**Double Sampling is a very useful technique. Can be applied to make:**

**Ratio Estimators**

**Regression estimators**

**Stratification**

**More Practical!!!**

# Sampling Rare Events

---

- ◆ Pilot Survey For Stratification then Full Survey ( A Kind of Double Sampling)
- ◆ Network Sampling (Ch 15)
- ◆ Adaptive Sampling (Ch 23-26)

# SAMPLING RARE CLUSTERED ANIMAL POPULATIONS: CONVENTIONAL DESIGNS

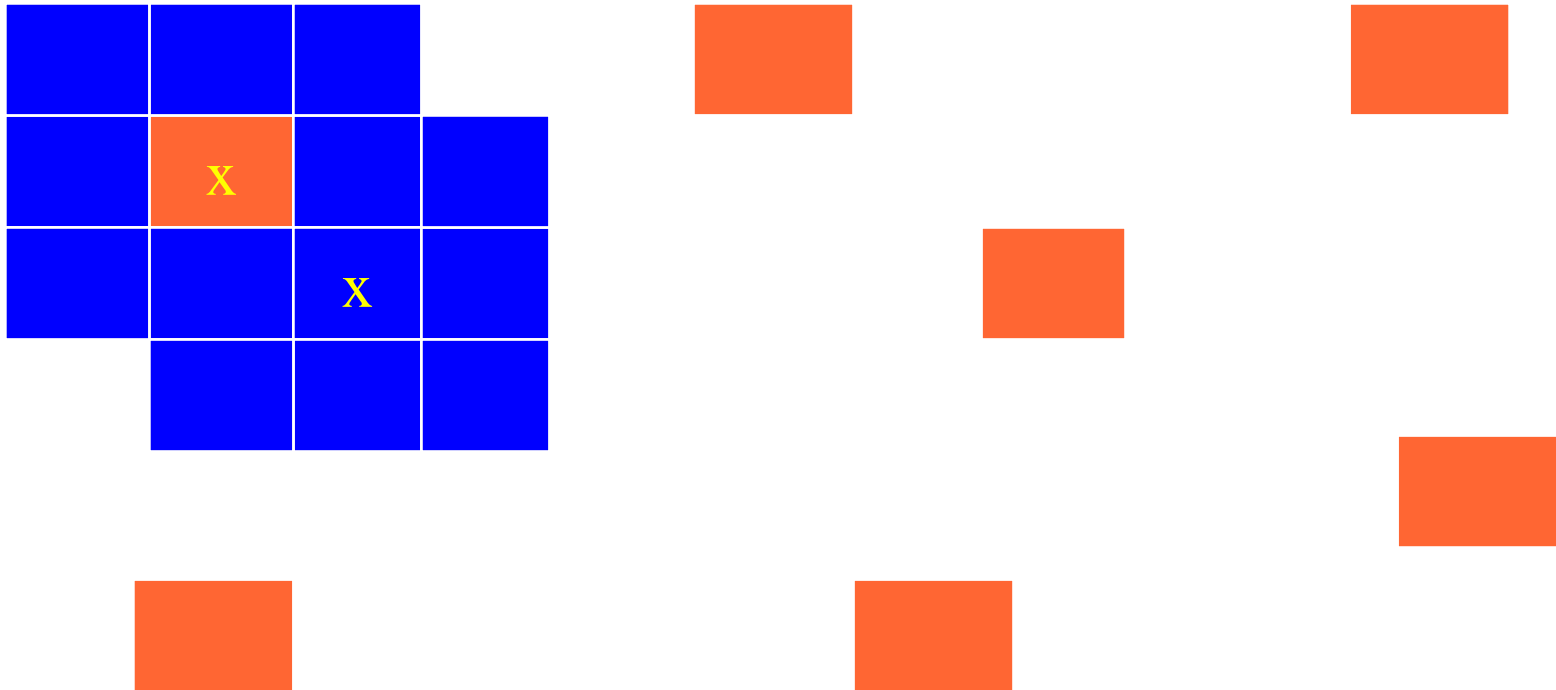
- For a **rare spatially clustered population**, the researcher may wish to add neighbouring sites to the sample, once high abundance of the species has finally been encountered.
- In **conventional sampling designs** like stratified random sampling, the probabilities of inclusion are chosen prior to the sample being drawn. This keeps results very simple and these designs are very widely used and useful for many situations.
- Therefore the **sample is fixed in advance** and a researcher cannot respond without biasing the results if they detect high species abundance in one location and wish to take additional sites there.

# SAMPLING RARE CLUSTERED ANIMAL POPULATIONS :ADAPTIVE SAMPLING

- **Adaptive sampling** refers to sampling designs where the inclusion probabilities may depend on values observed earlier in the survey. Therefore if they detect high species abundance in one location and wish to take additional samples there they may do so without biasing the results.
- In **adaptive cluster sampling**. An initial sample of plots are chosen at random (perhaps within strata) and then additional adjoining plots are chosen if some threshold density of animals is reached. Notice the much better coverage of diffuse rare clusters in an efficient way. **See figure from Thompson book.**
- It is more difficult to calculate inclusion probabilities but if the animals are very clustered and rare can be a huge gain in precision.
- **Strip** or **line transects** can also be extended to allow adaptive sampling.

# Adaptive Sampling: Conceptual

---



Sample Initial Orange Squares(7): If in a dense area (x) then sample blue squares around it.

# General Sampling References

**Williams et al. (2002) Chapter 5 and then Chapters 12-20 on detection issues.**

**Thompson, S. K. (1992). Sampling .John Wiley and Sons, New York, NY.**

**Thompson, S. K. and Seber, G.A.F. (1996). Adaptive Sampling .John Wiley and Sons, New York, NY.**

Not Used Here

---

# Double Sampling Detail

$$\hat{D}_h = \frac{\bar{C}_h}{\hat{\beta}_h}$$

$$\bar{C}_h = \sum_{j=1}^{n'} C_{hj} / n' \quad \hat{\beta}_h = \sum_{j=1}^n \hat{\beta}_{hj} / n$$

$$\text{Var}(\hat{D}_h) = (D_h)^2 \left\{ \frac{\tau_c^2}{n'} + \frac{\tau_\beta^2}{n} \right\}$$

with  $\tau$  the cv.

$$C = c_c n' + c_\beta n$$

$$\frac{n}{n'} = \left\{ \tau_\beta^2 c_c / \tau_c^2 c_\beta \right\}$$

# Double Sampling Detail

$$\frac{n'}{n} = \{\tau_{\beta}^2 c_c / \tau_c^2 c_{\beta}\} = (\delta / \gamma^2)^{1/2}$$

with  $\delta = c_c / c_{\beta}$  and  $\gamma = \tau_c / \tau_{\beta}$

Salamanders

$$\delta = 1/16 \text{ and } \gamma = 4 \quad \frac{n'}{n} = 1/16 = 0.066.$$

Subsample 7 out of 100 plots

Stream Survey

$$\delta = 1/2 \text{ and } \gamma = 2 \quad \frac{n'}{n} = 0.35$$

Subsample 35 out of 100 streams