

**MAT3377**

**Fall 2010**

**Summary formulas for SRS and stratified sampling**

## Simple Random Sampling

$N$  = Population size

$n$  = Sample size

Number of sample of size  $n$  without replacement =  $\binom{N}{n}$ .

Population:  $Y_1, \dots, Y_N$ .

Sample:  $y_1, \dots, y_n$ .

Population mean (average)

$$\bar{Y} = \frac{Y_1 + \dots + Y_N}{N} = \frac{\sum_{i=1}^N Y_i}{N}.$$

Sample mean (average)

$$\bar{y} = \frac{y_1 + \dots + y_n}{n} = \frac{\sum_{i=1}^n y_i}{n}$$

Population variance

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

Sometimes we use

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

We have

$$\frac{N}{N-1} \sigma^2 = S^2.$$

We have

$$E(y_i) = \bar{Y} \quad \text{and} \quad \text{Var}(y_i) = \sigma^2$$

and for  $i \neq j$ ,

$$Cov(y_i, y_j) = -\frac{\sigma^2}{N-1}.$$

### Estimation and Precision for the population mean $\bar{Y}$

Estimation:

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

The accurate formula for precision (need to know either  $S^2$  or  $\sigma^2$ )

$$Var(\bar{y}) = \frac{S^2}{n} \left( \frac{N-n}{N} \right) = \frac{\sigma^2}{n} \left( \frac{N-n}{N-1} \right)$$

An estimate for the precision is

$$var(\bar{y}) = \frac{s^2}{n} \left( \frac{N-n}{N} \right).$$

A bound for estimating  $\bar{Y}$  is

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

### Estimation and Precision for the population proportion $p$

$p$  = Proportion of units in a certain category in the population

Estimation:

$\hat{p}$  = proportion of units in a certain category in the sample

The accurate formula for precision (need to know  $p$ )

$$Var(\hat{p}) = \frac{pq}{n} \left( \frac{N-n}{N-1} \right)$$

An estimate for the precision of this estimate is

$$var(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1} \left( \frac{N-n}{N} \right)$$

A bound for estimating  $p$  is

$$B = 2\sqrt{\text{var}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1} \left(\frac{N-n}{N}\right)}$$

### Estimation and Precision for $A = Np$

$A = Np$  = Number of units in a certain category in the population.

$\hat{A} = N\hat{p} = a$  = Number of units in a certain category in the sample.

The notations  $A$  and  $a$  are not in the textbook.

### Estimates and their variances

$$\hat{A} = a = N\hat{p}.$$

The accurate formula

$$\text{Var}(\hat{A}) = \text{Var}(a) = N^2 \frac{pq}{n} \left(\frac{N-n}{N-1}\right)$$

An estimate for  $\text{Var}(\hat{A})$  is

$$\text{var}(a) = N^2 \frac{\hat{p}\hat{q}}{n-1} \left(\frac{N-n}{N}\right).$$

A bound for estimating  $a = N\hat{p}$  is

$$B = 2N\sqrt{\text{var}(\hat{p})} = 2N\sqrt{\frac{\hat{p}\hat{q}}{n-1} \left(\frac{N-n}{N}\right)}.$$

### Sample size.

To estimate  $\mu$  and  $\tau$  we use

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

where  $D = \frac{B^2}{4}$  for  $\mu$  and  $D = \frac{B^2}{4N^2}$  for  $\tau$ . If  $\sigma^2$  is not known use  $s^2$  if there is a prior experimentation otherwise use  $\sigma \approx R/4$  where  $R$  is the range.

To estimate  $p$  or  $A = Np$  use

$$n = \frac{Npq}{(N-1)D + pq}$$

where  $D = \frac{B^2}{4}$  for  $p$  and  $D = \frac{B^2}{4N^2}$  for  $A$ . If we have no idea about  $p$  then  $n < 1/(4D)$ .

# Stratified sampling

Notations:

$N$  = The population size

$N_i$  = The size of the stratum  $i$  in the population

$n_i$  = The size of stratum  $i$  in the sample

$$W_i = \frac{N_i}{N}, w_i = \frac{n_i}{n}$$

$\bar{Y}_i$  = Mean of the stratum  $i$  in the population

$\bar{y}_i$  = Mean of the stratum  $i$  in the sample

$L$  = Number of strata

$$N = N_1 + \dots + N_L$$

## Estimation for mean and total

$$\hat{\mu} = \sum_{i=1}^L W_i \bar{y}_i \quad \text{and} \quad \hat{\tau} = N \hat{\mu} = \sum_{i=1}^L N_i \bar{y}_i$$

## Variance and bound on Error

Accurate formula:

$$Var(\hat{\mu}) = \sum_{i=1}^L W_i^2 \frac{S_i^2}{n_i} \left( \frac{N_i - n_i}{N_i} \right) = \sum_{i=1}^L W_i^2 \frac{\sigma_i^2}{n_i} \left( \frac{N_i - n_i}{N_i - 1} \right)$$

Estimate:

$$var(\hat{\mu}) = \sum_{i=1}^L W_i^2 \frac{s_i^2}{n_i} \left( \frac{N_i - n_i}{N_i} \right).$$

Accurate formula:

$$Var(\hat{\tau}) = N^2 \sum_{i=1}^L W_i^2 \frac{S_i^2}{n_i} \left( \frac{N_i - n_i}{N_i} \right) = N^2 \sum_{i=1}^L W_i^2 \frac{\sigma_i^2}{n_i} \left( \frac{N_i - n_i}{N_i - 1} \right)$$

Estimate:

$$var(\hat{\tau}) = N^2 \sum_{i=1}^L W_i^2 \frac{s_i^2}{n_i} \left( \frac{N_i - n_i}{N_i} \right)$$

## Estimation for proportion ( $p$ ) and total ( $a$ )

$$\hat{p} = \sum_{i=1}^L W_i \hat{p}_i \quad \text{and} \quad \hat{A} = N \hat{p}$$

### Variance and bound on Error

Accurate formula

$$Var(\hat{p}) = \sum_{i=1}^L W_i^2 \frac{p_i q_i}{n_i} \left( \frac{N_i - n_i}{N_i - 1} \right)$$

Estimate

$$var(\hat{p}) = \sum_{i=1}^L W_i^2 \frac{p_i q_i}{n_i - 1} \left( \frac{N_i - n_i}{N_i} \right)$$

### Sample size

$$n = \frac{\sum_{i=1}^L N_i^2 \frac{S_i^2}{w_i}}{N^2 D + \sum_{i=1}^L N_i S_i^2}$$

where

$$D = \frac{B^2}{4}$$

if our goal is to estimate  $\mu$  and

$$D = \frac{B^2}{4N^2}$$

if our goal is to estimate  $\tau$ .  $w_i$  must be calculated based on the allocation as follows:

### Optimum allocation.

If the cost function is

$$C = c_0 + \sum_{i=1}^L c_i n_i$$

then

$$w_i = \frac{n_i}{n} = \frac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^L N_i S_i / \sqrt{c_i}} = \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^L W_i S_i / \sqrt{c_i}}.$$

In this case we can write

$$n = \frac{\left( \sum_{i=1}^L N_i S_i / \sqrt{c_i} \right) \left( \sum_{i=1}^L N_i S_i \sqrt{c_i} \right)}{N^2 D + \sum_{i=1}^L N_i S_i^2}$$

### Proportional allocation

$$w_i = W_i.$$

If the cost function is given and total cost is known we can use optimum allocation and calculate  $n$  from

$$n = \frac{(c - c_0) \sum_{i=1}^L N_i S_i / \sqrt{c_i}}{\sum_{i=1}^L N_i S_i \sqrt{c_i}}.$$

### Stratification

If Density is given need to have:

$$\int_{y_0}^{y_1} \sqrt{f(y)} dy = \dots = \int_{y_{L-1}}^{y_L} \sqrt{f(y)} dy.$$

If a table of frequencies are given Partition the population with multiples of

$$\sum_{i=1}^L \sqrt{f_i} / L.$$

## Ratio, Regression and Difference Estimation

1. Estimating population Ratio  $R = \frac{\tau_X}{\tau_Y}$ .

$$\hat{R} = r = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}.$$

$$\hat{V}_r = var(r) = \frac{1-f}{n\mu_X^2} s_r^2$$

where

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

Ratio estimator of total  $\hat{\tau}_y$ .

$$\hat{\tau}_y = r\tau_X.$$

$$var(\hat{\tau}_y) = \hat{V}(\hat{\tau}_Y) = \frac{N(N-n)}{n} s_r^2.$$

Ratio estimator of total  $\hat{\mu}_y$ .

$$\hat{\mu}_y = r\mu_X.$$

$$\text{var}(\hat{\mu}_y) = \hat{V}(\hat{\mu}_Y) = \left( \frac{N-n}{Nn} \right) S_r^2.$$

**Sample size.** Take

$$n = \frac{nS^2}{ND + S^2}$$

where (i)  $D = \frac{B^2\mu_x^2}{4}$  when we want to estimate  $R$ ,

(ii)  $D = \frac{B^2}{4}$  when we want to estimate  $\mu_y$

and

(iii)  $D = \frac{B^2}{4N^2}$  when we want to estimate  $\tau_y$

**Systematic Sample.**

$n$  = sample size,  $k$  = number of possible systematic samples,  $N = nk$ .

$Y_{ij} = j^{\text{th}}$  unit in  $i^{\text{th}}$  sample.

$$\bar{Y}_i = \frac{1}{n} \sum_{j=1}^n Y_{ij}, i = 1, 2, \dots, k.$$

$$E(\bar{y}_{sys}) = \bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^n Y_{ij}.$$

$$\text{Var}(\bar{y}_{sys}) = \frac{1}{k} \sum_{i=1}^k (\bar{Y}_i - \bar{Y}_{..})^2.$$

$$SST = \sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2.$$

$$SSB = n \sum_{i=1}^k (\bar{Y}_i - \bar{Y}_{..})^2.$$

$$SSW = \sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2.$$

$$SST = SSB + SSW$$

$$MSB = SSB/(k-1), S^2 = MST = SST/(N-1), MSW = SSW/k(n-1) = S_{WSYS}^2.$$

$$\text{Var}(\bar{y}_{sys}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_{WSYS}^2$$

$$\text{Var}(\bar{y}_{sys}) = \frac{S^2}{n} \left( \frac{N-1}{N} \right) [1 + \rho_W(n-1)].$$

$$\rho_W = \frac{2 \sum_{i=1}^k \sum_{j < u} (Y_{ij} - \bar{Y}_{..})(Y_{iu} - \bar{Y}_{..})}{(n-1)(N-1)S^2}.$$

$$\rho_W = \frac{nSSB - SST}{SST(n-1)} = \frac{n(k-1)MSB - (kn-1)MST}{(kn-1)(n-1)MST} \approx \frac{MSB - MST}{(n-1)MST}.$$

Estimating the variance of  $\bar{y}_{sys}$  by the successive differences.

$$\hat{V}(\bar{y}_{sys}) = \frac{N-n}{Nn} \frac{1}{2n_d} \sum_{i=1}^{n_d} d_i^2$$

where  $n_d = n - 1$ .

**Repeated systematic sampling.** If we have  $n_s$  repeated systematic samples then

$$\hat{\mu} = \frac{1}{n_s} \sum_{i=1}^{n_s} \bar{y}_i,$$

and

$$\hat{V}(\hat{\mu}) = \frac{N-n}{N} \frac{1}{n_s(n_s-1)} \sum_{i=1}^{n_s} (\bar{y}_i - \hat{\mu})^2.$$

**single stage cluster sampling.**

$$\hat{\mu} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$$

where

$y_i$  = total of the  $i^{th}$  cluster in the sample

$m_i$  = number of units in the  $i^{th}$  cluster in the sample.

We have

$$\hat{V}(\hat{\mu}) = \frac{1-f}{n\bar{M}^2} \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}..m_i)^2$$

where

$$\bar{M} = \frac{\sum_{i=1}^N M_i}{N}.$$