Université d'Ottawa · University of Ottawa

Faculté des sciences
Mathématiques et de statistique

Faculty of Science
Mathematics and Statistics

Mid-term test/ MAT3377
Surveys and Sampling
Fall 2015.

Time : 80 Minutes **Professor : M. Zarepour**

Name : ______________________________

Student Number : ______________________________

Calculators are permitted. Write your answers in the space provided.

[15]1. Consider a stratified sampling with three strata and different cost for sampling per unit in each stratum. The previous results from a census concluded the following results :

i	N_i	S_i	C_i
1	120	3	\$9.00
2	700	4	\$28.00
3	40	5	\$42.00

where N_i, S_i and C_i are the population sizes of each stratum, their standard deviations and the cost to draw a sample from each stratum. To estimate the population mean with a margin of error of $B = 1$, find the appropriate sample size using the optimum allocation. How many samples do we need to draw from each stratum ?

Solution. We have

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

We have

$$\sum_{i=1}^{L} N_i S_i/\sqrt{C_i} = 690.01,$$

$$\sum_{i=1}^{L} N_i S_i \sqrt{C_i} = 17192.355,$$

and

$$\sum_{i=1}^{L} N_i S_i^2 = 13280.$$

Moreover

$$D = B^2/4 = 0.25.$$

This gives $n = 58.94 \approx 59$. Using optimum allocation, we have

$$n_i = n\frac{N_i S_i/\sqrt{C_i}}{\sum_{i=1}^{L} N_i S_i/\sqrt{C_i}}, i = 1, 2, 3.$$

This gives :

$$n_1 = 10, n_2 = 46, n_3 = 3.$$

[15] 2. The blood types for the students in a class of $n = 12$ students (an artificial population) is listed below

$$A, B, O, O, O, B, O, B, AB, B, O, O.$$

A sample of size 3 is chosen at random. Let $\hat{p}$ be the sample proportion of the students with blood type O in a sample of size $n = 3$. Find the probability distribution for $\hat{p}$, $E(\hat{p})$ and $Var(\hat{p})$. Confirm if this example verifies the formula we derive for $V(\hat{p})$.

We have

$$P(\hat{p} = 0) = \frac{\binom{6}{3}}{\binom{12}{3}} = \frac{2}{22},$$

$$P(\hat{p} = 1/3) = \frac{\binom{6}{1}\binom{6}{2}}{\binom{12}{3}} = \frac{9}{22},$$

$$P(\hat{p} = 2/3) = \frac{\binom{6}{2}\binom{6}{1}}{\binom{12}{3}} = \frac{9}{22}$$

and

$$P(\hat{p} = 1) = \frac{\binom{6}{3}}{\binom{12}{3}} = \frac{2}{22},$$

This gives

$$E(\hat{p}) = (0)(2/22) + (1/3)(9/22) + (1/3)(9/22) + 1(2/22) = 0.5 = 6/12 = 0.5.$$

Calculate

$$E(\hat{p}^2) = (0)^2(2/22) + (1/3)^2(9/22) + (1/3)^2(9/22) + 1^2(2/22) = 0.31818181.$$

This gives the exact Variance of

$$\hat{p} = E(\hat{p}^2) - (E(\hat{p}))^2 = 0.068181.$$

On the other hand

$$Var(\hat{p}) = \frac{pq}{n}\left(\frac{N-n}{N-1}\right) = \frac{(0.50)(0.5)}{3}(9/11) = 0.068181$$

and they are equal.

3. Quiz marks for a class of $N = 10$ students are as follows :

$$8, 5, 1, 6, 6, 6, 9, 4, 4, 7.$$

[10] (i) Find μ and σ^2, the mean and Variance of this finite population of size $N = 10$.

$$\mu == \bar{Y} = \sum_{i=1}^{N} Y_i = 5.6$$

and

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \bar{Y^2} - \bar{Y}^2 = 4.64.$$

[10] (ii) A student fails in this quiz if he scores less than 5. Calculate the exact Variance of the proportion of students failed in this quiz based on a sample of size $n = 3$ (sampling is performed without replacement).

$$Var(\hat{p}) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right) = \frac{(0.3)(0.7)}{3}(7/9) = 0.05444.$$

[10] (iii) A simple random sample of size $n = 3$ (without replacement) is selected from this class of $N = 10$ students. The observations are $5, 7$ and 4 Estimate the population total ($\tau = N\bar{Y}$) and place a bound on your estimate (use your estimated bound and imagine that the population values are not provided).

$$\hat{\tau} = N\bar{y} = 10(16/3) = 53.3333.$$

$$\hat{V}(\hat{\tau}) = N^2 \frac{s^2}{n} \left(\frac{N - n}{N} \right).$$

We have

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = 2.3333.$$

Therefore

$$\hat{V}(\hat{\tau}) = 100 \frac{2.3}{3} \left(\frac{10 - 3}{10} \right) = 54.44444.$$

The c.i. is

$$53.333 \pm 2\sqrt{54.4} = 53.33 \pm 14.76.$$

[10] 4. Explain with the help of a proper formula when stratified sampling can provide a more precise estimator compare to a simple random sampling. Comments should be concise and rigorous.

Read the lecture note about comparing SRS and stratified sampling.

[15] 5. Two hundred ($n = 200$) residents of a city of N=150,000 people in the city were asked whether they are in favor of increasing property taxes for the betterment of extracurricular activities in the schools. From 110 male in the sample only 50 were in favor of raising the property taxes while out of 90 female in the sample 63 were in favor of raising taxes. Estimate the proportion of residents who are in favor of the tax increase and place a bound on your error.

We have

$$n = 200, W_1 = W_2 = 0.5, n_1 = 110, n_2 = 90$$

and

$$\hat{p_1} = 50/110 = 5/11 = 0.45, \hat{p_2} = 63/90 = 5/11 = 0.7.$$

This gives

$$\hat{p} = \sum_{i=1}^{2} W_i \hat{p}_i = 0.577.$$

We have

$$\hat{V}(\hat{p}) = \sum_{i=1}^{2} W_i^2 \frac{\hat{p}_i \hat{q}_i}{n_i - 1} \left(\frac{N_i - n_i}{N_i} \right) = 0.00115651.$$

Therefore the c.i. is

$$0.577 \pm 0.068.$$

6. Label each of the following statements as true or False and briefly justify your answers.

[7.5] (i) In a population of size N with values $Y_1, \ldots, Y_N$ a single sample y is drawn. We have

$$E(y) = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad \text{and} \quad Var(y) = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2.$$

False. Note that

$$E(y_1) = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y}$$

but

$$Var(y_1) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sigma^2.$$

[7.5] (ii) In a sample of size n (with replacement) from the population with values of $Y_1, Y_2, \ldots, Y_N$, we have

$$E(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad \text{and} \quad Var(\bar{y}) = \frac{1}{nN}\sum_{i=1}^{N}(Y_i - \bar{Y})^2.$$

True.

Since sampling is with replacement

$$E(\bar{y}) = \bar{Y}$$

and

$$Var(\bar{y}) = \sigma^2/n$$

MAT3377
Fall 2015
Formula Sheet
Mid-term test

Simple Random Sampling

N = Population size
n = Sample size
Number of sample of size n without replacement = $\binom{N}{n}$.
Population : $Y_1, \ldots, Y_N$.
Sample : $y_1, \ldots, y_n$.

Population mean (average)

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

Sample mean (average)

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

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

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

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 :

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

The accurate formula for precision (need to know either S^2 or σ^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

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

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

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

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

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

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

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

Sample size.

To estimate μ and τ we use

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

where $D = \frac{B^2}{4}$ for μ and $D = \frac{B^2}{4N^2}$ for τ. If σ^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 :

$$N = \text{The population size}$$
$$N_i = \text{The size of the stratum i in the population}$$
$$n_i = \text{The size of stratum i in the sample}$$
$$W_i = \frac{N_i}{N}, w_i = \frac{n_i}{n}$$
$$\bar{Y}_i = \text{Mean of the stratum i in the population}$$
$$\bar{y}_i = \text{Mean of the stratum i in the sample}$$
$$L = \text{Number of strata}$$
$$N = N_1 + \cdots + N_L$$

Estimation for mean and total

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

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)

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

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 μ and

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

if our goal is to estimate τ. 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

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