STA304/1003 H1 F - Summer 2014: Surveys, Sampling, and Observational Data

Lecture 4 - Part I: Simple Random Sampling (cont'd)

Ramya Thinniyam

May 27, 2014

Sample Size Estimation

Determine sample size needed for survey based on your expectations:

1. Specify tolerable error, e (how close you want estimate to be to parameter):
   $P(|\bar{y} - \bar{y}_U| \leq e) = 1 - \alpha$ can be other parameter as well, say $P(|\hat{t} - t| < e)$

   e called margin of error.
2. Determine N
3. Find an equation to relate tolerable error and sample size:
   $e = z_{\alpha/2}\sqrt{\left(1 - \frac{n}{N}\right)}\frac{S}{\sqrt{N}}$
4. Estimate unknown quantities:

   Be conservative: when in doubt, overestimate the variances $\longrightarrow$ wider CI

   - Find s from past research
   - For proportion, use $S^2 = \frac{1}{4}$ ie. $p = \frac{1}{2}$ for large populations (if no other information is given)
   - For proportion, $Var(y_i) = p(1-p) = p - p^2$ is a parabola so you can obtain its maximum if there are constraints on p
5. Solve for n
6. If calculated n is larger than you can afford, change expectations for the survey and retry

FORMULA FOR SAMPLE SIZE:

$$n = \frac{z_{\alpha/2}^2 S^2}{e^2 + \frac{z_{\alpha/2}^2 S^2}{N}} = \frac{n_0}{1 + \frac{n_0}{N}} \text{ ; where } n_0 = \left(\frac{z_{\alpha/2} S}{e}\right)^2$$

Examples: Sample Size Calculations

Use benchmarks of 95% confidence when not specified.

1. 'mydata' Example: estimate mean within 0.5 of its true value. Previous study yields a sample variance of 10.

   Estimate $\bar{y}_U$ with 0.5 of the true value.

   $e=0.5$, $\alpha=0.05$, $S^2=10$, $N=100$, plug in values, solve for n.

   $$n_0=\left(\frac{Z_{\frac{\alpha}{2}}S}{e}\right)^2=\left(\frac{1.96\cdot\sqrt{10}}{0.5}\right)^2$$

   $$\doteq 153.664$$

   $$n=\frac{n_0}{1+\frac{n_0}{N}}\doteq 60.5772$$

   $n=61$ (recommended)

2. Estimate the proportion of students who own a computer in a population with 1000 students, with a margin of error of 0.03. We know at least 80% in the population own computers.

   $N=1000$

   proportion $p\geq 0.8$

   $e=0.03$

   $\alpha=0.05$, $Z_{\frac{\alpha}{2}}=1.96$

   $$n_0=\left(\frac{Z_{\frac{\alpha}{2}}S}{e}\right)^2=\left(\frac{1.96S}{0.03}\right)^2$$

   $S=?$

   Be conservative, find maximum, for $S^2=(1-p)p$ under constraint $p\geq 0.8$

   If $p\geq 0.8$ then S^2 is maximized at $p=0.8$, $S^{2*}=0.8\cdot 0.2=0.16$, $S^*=0.4$

   So $n_0=\left(\frac{1.96\cdot 0.4}{0.03}\right)^2=682.9511$ $\Rightarrow n=\frac{n_0}{1+\frac{n_0}{N}}\doteq 405.8057$ $\boxed{n=406}$

Sampling Weights

$\pi_i = P(\textit{unit } i \textit{ is in sample})$: Inclusion Probability
(not restricted to SRS)

For any sampling design, the sampling weight is the reciprocal of the inclusion probability:

$$w_i = \frac{1}{\pi_i}$$

.

For SRS
$\sum_{i=1}^{n} W_i = N$

- Interpreted as the number of population units represented by i
- For SRS: $w_i = \frac{N}{n}$: each unit represents itself + $\frac{N}{n} - 1$ unsampled units in the population
- For SRS: all weights are the same (each unit in the sample represents the same number of units in the population), called a self-weighting sample.

different design has different weights

Large Sample CIs for a Location - Finite Populations

* Use Finite Population Correction (fpc) in variance estimates - $\boxed{\left(1-\frac{n}{N}\right)}$ *

For Mean:

$$\bar{y} \pm z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}} \quad \text{OR} \quad \bar{y} \pm z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

For Total:

$$\hat{t} \pm z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{N^2 S^2}{n}} \quad \text{OR} \quad \hat{t} \pm z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{N^2 s^2}{n}}$$

For Proportion:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}}$$

with replacement

For infinite populations or SRSWR, we don't have fpc. b/c in this case $1-\frac{n}{N} \rightarrow 1$ so all stuffs above are without fpc

Example: Local News Coverage

A major metropolitan newspaper selected a SRS of 1,600 readers from their list of 100,000 subscribers. They asked whether the paper should increase its coverage of local news (1='yes', 0='no').

$$y_i = \text{localcoverage}_i = \begin{cases} 1, & \text{if } i \text{ answered yes} \\ 0, & \text{o.w.} \end{cases}$$

```
length(localcoverage)
[1] 1600
> mean(localcoverage)
[1] 0.4
> sum(localcoverage)
[1] 640
```

$N = 100{,}000$ length(localcoverage) = length of vector $= n$

$n = 1600$

$\bar{y} = 0.4$

$\text{sum(localcoverage)} = \sum_{i=1}^{n} y_i = \#$ of sampled subscribers who said yes.

Let p = proportion of subscribers who wants more local news coverage.

$$\hat{p} = \frac{\sum y_i}{n} = \bar{y} = \frac{640}{1600} = 0.4$$

a) Find a 99% CI for the proportion of readers who would like more coverage of local news.

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}} = 0.4 \pm 2.58\sqrt{\left(1-\frac{1600}{100000}\right)\frac{0.4\times 0.6}{1599}} \doteq (0.3686, 0.4314)$$

b) Find a 99% CI for the percent of readers who would like more coverage of local news.

99% CI for $100p$: $(36.86, 43.14)$

c) Find a 99% CI for the proportion of readers who would do not want more local news coverage.

$(1-0.4314, 1-0.3686) = (0.5686, 0.6314)$

Comparing Two Means - Infinite Populations

$100(1-\alpha)\%$ Approximate CI for difference of two means:

$$\boxed{(\hat{\mu}_1 - \hat{\mu}_2) \pm z_{\alpha/2}\sqrt{V(\hat{\mu}_1) + V(\hat{\mu}_2) - 2Cov(\hat{\mu}_1, \hat{\mu}_2)}}$$

- If zero is in the interval, then there is no statistically significant difference between the two population means

- $100(1-\alpha)\%$ of all samples generate an interval that captures the true difference between the two means

Example: Lifetime of Lightbulbs — assume 2 pop independent Infinite pop. N_1, N_2 infinite

Random sample of 140 traditional light bulb lifetimes:
$\bar{y}_1 = 1348.2, s_1 = 22.65$, $n_1=140$

Random sample of 80 new technology light bulb lifetimes:
$\bar{y}_2 = 1387.7, s_2 = 23.06$, $n_2=80$

Find a 95% CI for the difference between the two mean lifetimes. Can the manufacturer of the new technology light bulbs claim their mean lifetime is better than that for the traditional ones?

95% CI for $(\mu_1-\mu_2)$: $(\bar{y}_1-\bar{y}_2) \pm z_{0.025}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ (no covariance b/c independence)

$= (1348.2-1387.7) \pm 1.96\sqrt{\frac{(22.65)^2}{140}+\frac{(23.06)^2}{80}} = (-45.7939, -33.2061)$ $\Rightarrow$ traditional is significantly lower

YES.

Comparing Two Independent Locations - Finite Populations

Approximate large sample $100(1-\alpha)\%$ CI for difference between

Two Means:

$$(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{s_1^2}{n_1} + \left(1-\frac{n_2}{N_2}\right)\frac{s_2^2}{n_2}}$$

Two Totals:

$$(\hat{t}_1 - \hat{t}_2) \pm z_{\alpha/2}\sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{N_1^2 S_1^2}{n_1} + \left(1-\frac{n_2}{N_2}\right)\frac{N_2^2 S_2^2}{n_2}}$$

Two Proportions:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} + \left(1-\frac{n_2}{N_2}\right)\frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1}}$$

- If zero is in the interval, then there is no statistically significant difference between the two population locations

2 proportions. both are finite

(1). F $N_1=352$, $n_1=88$, $\sum_{i=1}^{n_1} y_i = 22$

(2). M $N_2=315$, $n_2=105$, $\sum_{j=1}^{n_2} y_j = 21$

$\hat{p}_1 = 0.25$

$\hat{p}_2 = 0.2$

95% CI for $(p_1 - p_2)$:

Example: Two Proportions - N_1 and N_2 finite

A company wishes to increase the sales of their new product by advertising/targeting the correct consumers. They take a simple random sample in the male and female population and ask consumers who bought their product, “Do you like this product?”

Female Population: $N_1 = 352, n_1 = 88, 22$ answered ‘yes’

Male Population: $N_2 = 315, n_2 = 105, 21$ answered ‘yes’

Use a 95% CI to answer the question of interest. What would you recommend the company to do?

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1} + \left(1-\frac{n_2}{N_2}\right)\frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}}$$

$$0.25 \pm 1.96\sqrt{\left(1-\frac{88}{352}\right)\frac{0.25\cdot 0.75}{87} + \left(1-\frac{105}{352}\right)\frac{0.2\cdot 0.8}{104}} = (-0.0507, 0.1507)$$

Therefore, no significant difference between the true proportion of males & females who like the product.

Example: Two Means- N_1 finite, N_2 infinite

Compare a sample with a general type of population

A school has 500 children of which sample 25 are sampled: the mean number of pets per child is 1.32, $s = 0.3$. In the general population, 25 children are sampled: the mean number of pets per child is 1.08, $s = 0.5$. We wish to compare the mean number of pets in the school population with the general population.

do fpc to finite & do not do fcp to infinite

a) What assumptions must be made to use a CI?

(1) school (2) general

$N_1=500$, $n_1=25$, $\bar{y}_1=1.32$, $s_1=0.3$

$N_2=\infty$, $n_2=25$, $\bar{y}_2=1.08$, $s_2=0.5$

b) Find a 95% CI and make conclusions.

95% CI for $(\mu_1-\mu_2)$: $(1.32-1.08) \pm 1.96\sqrt{\left(1-\frac{25}{500}\right)\frac{0.3^2}{25}+\frac{0.5^2}{25}} = (0.0129, 0.4671)$

0 not inside, so the first pop has more pets. significant difference.