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

Lecture 10:
Examples of Sampling Designs

Ramya Thinniyam

June 17, 2014

Example 1: Lipstick Preference

A marketing research firm (MRF) estimates the proportion of customers preferring a certain brand of lipstick by "randomly" selecting 100 women who came by their booth in a shopping mall. Of the 100 sampled, 65 women stated a preference for brand A.

a) Identify the following for this study:

- Target Population
- Sampled Population
- Variable
- Sampling Unit
- Observation Unit

b) What type of sampling design is it?

a) target population: all women/customers who use lipstick

Sample population: all women who came by MRF booth at the shopping mall

Variable: whether or not brand A is preferred or band preference

Sampling unit : a woman

Observation unit : a woman

b) " convenience sample" If women were truly randomly selected, could be considered an SRS with n = 100, N large, (but sampling frame is not available a prior!

Lipstick Preference - Questions (cont'd)

c) Assuming all is correct with this sampling method, estimate the true proportion of women preferring brand A, and place a bound on the error of estimation.

d) If a more accurate estimate has to be found, propose the minimal sample size required to estimate the true proportion within 3% of the true value. A reasonable guess is that the true proportion is within $\pm 10\%$ of the value obtained in c).

e) Did MRF select a simple random sample from the target population? Explain. Even if the sample may not be an SRS, could it still be reasonably representative of the target population?

f) How would you help MRF to improve their sampling strategy? Revise the sampling design.

c) proportion of women who prefer brand A = p

$\hat{p} = \frac{65}{100} = 0.65$

$se(\hat{p}) = \sqrt{\hat{V}(\hat{p})} = \sqrt{(1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$

$= \sqrt{\frac{0.65(0.35)}{99}} = 0.0479$

Bound on error of est based on 95% confidence

is: $1.96\, se(\hat{p}) = 1.96(0.0479) = 0.0940$

N large fpc $(1-\frac{n}{N}) \longrightarrow 1$ $\quad 1-\frac{n}{N} \approx 1$

d) $e=0.03$ $0.55 \le p \le 0.75$ N large $1-\frac{n}{N} \approx 1$ 95% confidence $Z_{\frac{\alpha}{2}} = 1.96$

Since no fpc required Sample size $= n_0$ maximize $S^2 = p(1-p)$ is maximized at $p = 0.55$

under constraint $0.55 \le p \le 0.75$ $n_0 = \left(\frac{Z_{\frac{\alpha}{2}} S^*}{e}\right)^2 = \frac{(1.96)^2(0.55)(0.45)}{(0.03)^2} = 1056.44 = 1057$

e) undercoverage : Not all women go to the mall

overcoverage : not all woman go to booth uses lipstick

f) random Select many mall and set many booth

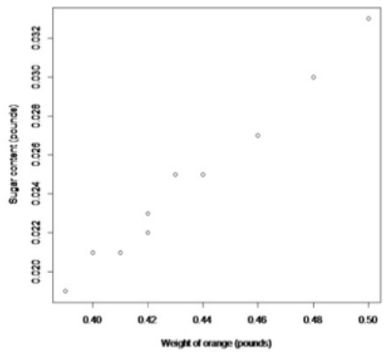
ask women (not only who comes to booth) anywhere in the mall

Example 2: Sugar Content in Oranges

A company wishes to make inferences about the sugar content of its oranges based on a truckload of oranges that has just arrived. The total number of oranges is unknown and tedious to count. However, the total weight of all the oranges is determined to be 1800 pounds, by first weighing the truck loaded then unloaded. A random sample of 10 oranges is taken and for each orange the weight (in pounds) and sugar content (in pounds) is measured. Below is a summary of the sample data:

y_i x_i $(resi)^2 = (y_i - r*x_i)^2$ $n=10$ N unknow

	Sugar Content (S)	Weight (W)	$(S - r * W)^2$
Total:	0.246	4.35	0.000053



$\hat{S}_i = r w_i$
↳ slope estimate

$S_i = R w_i + \varepsilon_i$ $r = \hat{R}$

$\bar{y} = \frac{0.246}{10} = 0.0246$

$\bar{w} = 0.435$

$S_r^2 = \frac{0.000053}{9}$

$1800 = T_w = T_x$

a) SRS formulas cannot be used to estimate total sugar content since N, the total # of orange is unknown

b) Ratio Estimation: 1. Sugar content and weight highly (+ve) correlated by common sense & scatterplot

2. Makes sense to fit regression through origin

3. N is unknown but could be estimated / not even needed for ratio est.

c) $\hat{N} = \frac{T_w}{\bar{w}} = \frac{1800}{0.435} = 4137.9310$ so approx 4138 oranges

Sugar Content in Oranges - Questions:

a) Explain why Simple Random Sampling formulas cannot be used to estimate the total sugar content in the truckload.

b) What type of estimation would be reasonable to use for this scenario? Justify.

c) Estimate the total number of oranges in the truckload.

d) Give a point estimate for the mean amount of sugar content per pound of orange.

e) The company has now been informed that that the truckload contained 3000 oranges. Find a 95% CI for the total sugar content in the truckload.

f) In the previous part, suppose you were to instead use SRS to find the 95% CI. Which CI (the one from this part or the previous part) would you expect to be wider. Justify.

d) Estimate $R = \frac{\bar{S}_u}{\bar{w}_u} = \frac{T_s}{T_w}$ $r = \frac{\bar{s}}{\bar{w}} = \frac{\hat{T}_s}{T_w} = \frac{0.246}{4.35} = 0.0566$ pounds of sugar per pound of orange

e) now $N = 3000$ known $\hat{T}_{sr} = r T_w = \frac{0.0246}{0.435}(1800) = 101.88$?

$SE(\hat{T}_{sr}) = \sqrt{N(1 - \frac{n}{N})(\frac{T_w}{\bar{w}})^2 \frac{S_r^2}{n}} = \sqrt{(1 - \frac{10}{3000})(\frac{1800}{0.435})^2 \frac{0.000053}{9(10)}} = 3.1701$

95% CI for T_s (95.58, 108.18) pounds (算錯)

f) Wider: SRS less precise

Example 3: Escalators in Subways

A city transportation system includes 30 subway stations, each containing 6 escalators, and is interested in the number of days that the escalators were down for repair in the past year. 3 subway stations were randomly selected and the maintenance records for all 6 escalators in each are examined. Below are the results:

Station	Number of Days Escalator is Down	Average	Sample Variance
1	4, 3, 7, 2, 11, 0	4.5	15.5
2	11, 4, 3, 1, 0, 2	3.5	15.5
3	0, 3, 6, 4, 3, 2	3	4

```
Analysis of Variance Table

Response: days
          Df Sum Sq Mean Sq F value Pr(>F)
station    2      7   3.500     0.3 0.7452
Residuals 15    175  11.667
```

Escalators in Subways - Questions:

a) What type of sampling design is used here. Identify the characteristics and parameters.

b) Estimate the mean downtime per escalator with a 95% CI.

c) Estimate the total number of days down for all escalators in the subway.

d) Compare the efficiency of this estimator with that of the SRS in this case. Decide which sampling design is better for this situation.

e) Estimate the Intracluster Correlation Coefficient (ICC). Is the value of ICC in accordance to what you found in the previous part? Explain. If the SRS is more efficient than this sampling design, suggest another sampling design that would give better precision.

a) One-Stage cluster Sample psu = Station psu = escalator

$N=30$ $n=3$ $M_i = m = 6$ (equal cluster sizes)

$M = \sum_{i=1}^{N} M_i = Nm = 30\times 6 = 180$ $\bar{M} = 6$ equal cluster sizes

b) $\hat{\bar{y}}_{unb} = \frac{1}{M}\frac{N}{n}\sum_{i\in s} M_i\bar{y}_i = \frac{1}{180}\frac{30}{3}\cdot 6(4.5+3.5+3) = 3.6667 = \bar{y}_c$

$se(\hat{\bar{y}}_{unb}) = \sqrt{\frac{1}{n\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{i\in s}(M_i\bar{y}_i - \bar{M}\hat{\bar{y}}_{unb})^2} = \sqrt{\frac{1}{3(6)^2}\left(1-\frac{3}{30}\right)\underbrace{\frac{1}{n-1}m\sum_{i\in s}(\bar{y}_i-\bar{y}_{unb})^2}_{MS\ station}}$

$= \sqrt{\frac{1}{3(6)^2}\left(1-\frac{3}{30}\right)(3.5)} = 0.1708$

c) $\hat{\tau}_{unb} = M\times\hat{\bar{y}}_{unb} = 180\times 3.6667 = 660.006$ total down days

d) using SRS: $se_{SRS}(\bar{y}) = \sqrt{\left(1-\frac{nm}{M}\right)\frac{s^2}{nm}} = \sqrt{\left(1-\frac{18}{180}\right)\frac{SSTO}{17(18)}}$

$= \sqrt{\left(1-\frac{8}{180}\right)\frac{182}{17(18)}}$ $SSTO = SS\ station + SS_{res} = 7+175 = 182$

$= 0.7316 \Rightarrow$ cluster is better

e) $\hat{ICC} = 1-\frac{m}{m-1}\frac{\hat{SSW}}{\hat{SSTO}}$ $\hat{SSW} = \hat{MSW}\times 30^{(N)}(5)^{(m-1)} = 11.667(150) = 1750.05$

$SS_{station} = \hat{SSB} = \hat{MSB}\times 29^{(N-1)} = 3.5\times 29 = 101.5$

$\hat{ICC} = 1-\frac{6}{5}\left(\frac{1750.05}{1750.05+101.5}\right) = -0.1342 < 0$

As expected. Since var of cluster < var of SRS $\Rightarrow$

within each cluster, it's heterogeneous

Example 4: Refereed Publications by Department

A university has 807 faculty members of which 102 work in Biological Sciences, 310 in Physical Sciences, 217 in Social Sciences, and 178 in Humanities. The university is interested in estimating the number of refereed publications of its faculty members. Some faculty were randomly selected from each department for a total of 50 faculty members in the sample. The number of refereed publications of each selected member was carefully investigated and then recorded. Below is the frequency table for the number of refereed publications in each department:

Number of Refereed Publications	Number of Faculty Members			
	Biological	Physical	Social	Humanities
0	1	10	9	8
1	2	2	0	2
2	0	0	1	0
3	1	1	0	1
4	0	2	2	0
5	2	1	0	0
6	0	1	1	0
7	1	0	0	0
8	0	2	0	0
Total	7	19	13	11

Refereed Publications by Department - Questions:

a) What type of sampling design was used? Identify its parameters. Discuss the merits of using this type of sampling design and its applications.

b) Estimate the total number of refereed publications by faculty members in this university, with a standard error.

c) Give a 95% CI for the proportion of faculty members in this university with no refereed publications.

d) Give a 95% CI for the percentage of faculty members in this university that have at least one refereed publication.

e) Aside from the estimation done above, what other inferences (using any statistical methods) can be made. Name the methods and specify what types of inferences / questions of interest could be answered for each.

a) STRS by department - easy to get data from each department. Compare var between different department

Summary Stats

	Bio	Phy	Soc	Hum
Average	3.14	2.10	1.23	0.45
Sample Var	6.81	8.21	4.36	0.87
Sample size	7	19	13	11

L=4 Strata $N_1=102$ $N_2=310$ $N_3=217$ $N_4=178$

c) $\hat{\tau}_{bio} = N_1\bar{y}_1 = 102(3.14) = 320.57$

$\widehat{Var}(\hat{\tau}_{bio}) = N_1^2(1-\frac{n_1}{N_1})\frac{S_1^2}{n_1} = (102)^2(1-\frac{7}{102})\frac{6.81}{7} = 9426.327$

$\hat{\tau}_{str} = \hat{\tau}_{bio} + \hat{\tau}_{phy} + \hat{\tau}_{soc} + \hat{\tau}_{hum} = 1321.189$ $\quad \widehat{Var}_{str} = \widehat{Var}(\hat{\tau}_{bio}) + \cdots = 65610.78$

P = P(no referred publications)

	Ni	ni	$\hat{p}_i$	$\frac{N_i}{n_i}\hat{p}_i$	$(1-\frac{n_i}{N_i})(\frac{N_i}{N})^2\frac{\hat{p}_i(1-\hat{p}_i)}{n_i-1}$
Bio	102	7	1/7	0.018	0.0003
Phy	310	9	10/19	0.202	0.0019
Soc	217	13	9/13	0.186	0.001
Hum	178	11	8/11	0.160	0.0009
total	807	50		0.567	0.0043

$\hat{p}_{str} = 0.567$ $\quad se(\hat{p}_{str}) = \sqrt{0.0043} = 0.0656$

95% CI for p: $0.567 \pm 1.96(0.0656) = (0.4384, 0.6956)$

95% CI for 1-p $(1-0.6956, 1-0.4384) = (0.3044, 0.6956) \Rightarrow$
$(30.44\%, 69.56\%)$