W3 Saving money using cluster sampling

3.1 Simple Complex Sampling – choosing entire clusters

A population

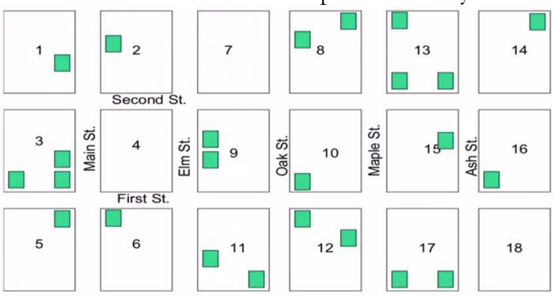
Target Population Elements:

101 Main St.
104 Main St.
107 Main St.
112 Main St.
115 Main St.
122 Main St.
129 Main St.
132 Main St.
201 Main St.
206 Main St.
. .
. .
.

$$\bar{Y} = \sum_{i=1}^{N} Y_i \Big/ N$$

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

Simple Random Sampling
• Doesn't have to include a sample from every block in the neighborhood



• If I don't have a list of addresses… then we build our "own list" ➔ list addresses by hand ➔ but that list creation activity incurs "costs"
• We can still get "list of blocks" that is used in Census (but doesn't list individual addresses often due to confidentiality/privacy issues) ➔ we have the CLUSTER but don't have the addresses

Cluster Sampling

$$\boxed{Var(\bar{y}) = \frac{(1-f)}{a} s_a^2}$$

• Populations often distributed geographically like this
  - cannot afford to create an element frame
  - cannot afford to visit n units drawn randomly from the entire area
• Cluster selections are used to reduce listing costs
  - select clusters and list elements only for selected clusters
• Clusters are used to reduce travel costs
• Clusters are often already listed
  - makes them "naturally occurring units"
  - seldom equal size

• Suppose we select an SRS of a = 10 classrooms from A=1000, and examine the immunization history of all b = 24 children in selected classrooms
• Here N = A x B = 1000 x 24 and n = a x b = 240
• We refer to the A classrooms as **primary sampling units** or **PSUs**
• For each of the (a = 10) selected PSU's, we record the number of children immunized:

$$\frac{9}{24}, \frac{11}{24}, \frac{13}{24}, \frac{15}{24}, \frac{16}{24}, \frac{17}{24}, \frac{18}{24}, \frac{20}{24}, \frac{20}{24}, \frac{21}{24}$$

• Adding the numerators, there are 160 immunized children
• The overall proportion immunized is p = 160/240 = 0.67

• Recall from SRS (without replacement selection n elements), the sample proportion was $p = \sum_{i=1} y_i \Big/ n$
• The estimated sampling variance is:

$$\text{var}(p) = (1-f)s^2/n = (1-f)\frac{p(1-p)}{n-1}$$

• But for an SRS of "a" equal sized clusters from "A", we have sampled clusters not elements
• Randomization occurs at the cluster level
• We have a P_a for each selected PSU
• In cluster sampling, treat the sample as an SRS of "a" units from "A":

$$\text{var}(p) = \frac{(1-f)}{a}s_a^2$$

• Where

$$s_a^2 = \sum_{\alpha=1}^{a}(p_\alpha - p)^2 \Big/ (a-1) \qquad f = a/A$$

• That is,

$$\text{var}(p) = \frac{(1-f)}{a}\frac{\sum_{\alpha=1}^{a}(p_\alpha - p)^2}{a-1}$$

$$se(p) = \sqrt{\text{var}(p)}$$

• For the illustration,

$$s_a^2 = \frac{1}{10-1}\left[\left(\frac{9}{24} - \frac{160}{240}\right)^2 + \left(\frac{11}{24} - \frac{160}{240}\right)^2 + \cdots\right]$$

$$= 0.02816$$

$$\text{var}(p) = (1-f)s_a^2/a = 0.002760$$

$$se(p) = \sqrt{\text{var}(p)} = 0.0525$$

• Of course, the standard error is then used in a confidence interval
• But there is an important adjustment to what we've done up till now
• Recall that we briefly introduced he idea of using the t-distribution rather than the normal in confidence intervals
• That is much more important here with cluster samples than for simple random samples of elements
• That's because the confidence interval is built on a standard error that depends on the number of random events in the sample
• The number of random events in simple cluster sampling is "a", not "n"
• Hence we need to be worried about not "n" degrees of freedom, but "a" degrees of freedom
• And "a" is much smaller than "n"
• As a result, we will use "t-statistic" instead of the "z"
• In particular, we will use

$$\left(p - t_{(1-\alpha/2,a-1)} \times se(p),\, p + t_{(1-\alpha/2,a-1)} \times se(p)\right)$$

• It's the same as p, and the same standard error, but the multiplier for the standard error is from the t-distribution
• It's because we have only a random events in the sample, not n – a much smaller number
• We need to use a larger multiplier for the confidence interval when the number of random events is smaller

3.2 Design Effects

• A question is how did the cluster sample compare to a simple random sample?
• Need to establish grounds for comparison
  - compare precision since both designs are unbiased, and yield the same mean on average
  - on what basis should the precision be compared?
  - usually equal sample size
  - And a comparison of sampling variances
• If the sample had instead been an SRS of n = 240 children from all schools, then p = 160/240

$$\text{var}_{SRS}(p) = (1-f)\frac{p(1-p)}{n-1} = 0.0009112$$

• Compared to cluster sampling, the estimated variance of p is considerably smaller for SRS
• A ratio quantifies the comparison:
Deff(p) = var(p) / var_SRS(p)
• By definition, the numerator sampling variance must have the same sample size as the denominator
• For the illustration,
Deff(p) = var(p) / var_SRS(p) = 0.002760 / 0.0009112 = 3.029
• The design effect may be used in several ways
• One is to recognize the following:
Var(p) = deff(p) * var_SRS(p)
• In other words, the cluster sampling variance is the SRS sampling variance, adjusted for the effect of clustering
• This expression can be used to help design new surveys – to be discussed in the next lecture
• The design effect is directly a function of differences between clusters compared to differences among elements

• If deff > 1, then clusters are more variable than elements
- But why? : Heterogeneity between implies homogeneity within – the more different clusters are from one another… the more similar are elements within clusters to one another
• Empirical results have revealed that deff depends on homogeneity within and the size of the clusters, say b
• The homogeneity is measured by the **intra-cluster correlation roh (rate of homogeneity)**
• The design effect is given by:
Deff(p) = 1 + (b-1) roh
• The intra-cluster correlation can be estimated from the design effect:
(24 elements per cluster selected = b)
Roh = deff(p) – 1 / (b-1) = 3.029 – 1 /(24-1) = 0.088
• roh is a property of the clusters and the variable under study
- the design effect is then also going to differ across variables
• roh is substantive, not statistical
• roh is nearly always positive
- elements in a cluster tend to resemble one another
• Source of roh
- environment
- self-selection
- interaction
• Alternatively, the actual sample size n = 240 in the cluster sample
• But an SRS is equally precise would only have to have
n_eff = 240 / 3.029 = 79
• Effective sample size
• Consider alternative outcomes for our sample of a = 10 classrooms
- homogeneity within, heterogeneity between

$$\frac{0}{24},\frac{0}{24},\frac{0}{24},\frac{16}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24},\frac{24}{24}$$

$$s_a^2 = 0.2222 \quad \text{var}\left(p\right) = 0.02178$$

deff = 23.90 | *n_eff* = 240 / 23.9 = 10 (why so small? Cuz within cluster similarity is so huge that there is no new information you get from additional samples from a cluster… you already know everything about that cluster with just the first sample of the cluster… thus need very little effective sample size due to this extreme homogeneity within clusters)
- heterogeneity within, homogeneity between

$$\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24},\frac{16}{24}$$

$$s_a^2 = 0.0 \quad \text{var}\left(p\right) = 0.0$$

$$deff = 0$$

$$n_{eff} = 240/0$$

$$roh = \frac{0-1}{24-1} = -0.043$$

• Consider an equal probability (epsem) sample of n = 2400 obtained from a one-stage sample of a = 60 equal-sized clusters each of size b = 40 selected by SRS
• In a journal article, describing survey results, for a key production, p = 0.40
Var(p) = 0.00021795| How would we estimated deff and roh?

1. **Compute the simple random sampling variance**

$$\text{var}_{SRS}\left(p\right) = \frac{p\left(1-p\right)}{n-1}$$

(Ignore the *fpc* – that is, or assume it is 1)

2. **Compute the design effect**

$$deff\left(p\right) = \frac{\text{var}\left(p\right)}{\text{var}_{SRS}\left(p\right)} = \frac{0.00021795}{\frac{p\left(1-p\right)}{\text{n - 1}}}$$

3. **Compute the intra-clustermogeneity *roh***

$$roh = \frac{deff\left(p\right)-1}{b-1} = \frac{deff\left(p\right)-1}{40-1} =$$

The SRS variance is

$$\mathrm{var}_{SRS}(p) = \frac{p(1-p)}{n} = \frac{0.4 \times 0.6}{2400} = 0.0001$$

Thus, the design effect is

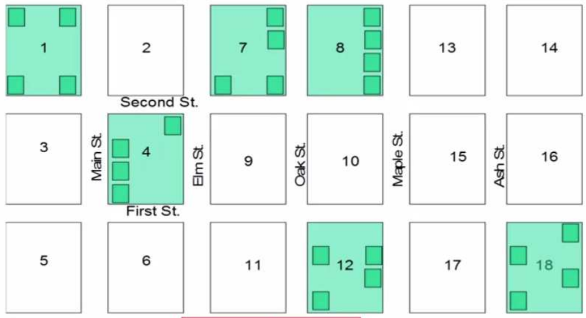
$$deff(p) = \frac{\mathrm{var}(p)}{\mathrm{var}_{SRS}(p)} = \frac{0.00021795}{0.0001} = 2.1795$$

And an estimate of intra-class correlation is

$$roh = \frac{deff(p) - 1}{b - 1} = \frac{2.1795 - 1}{40 - 1} = 0.03024$$

3.3 Two stage cluster sampling

• choose 6 blocks (clusters) [1st stage] and then subsample housing units within each [2nd stage]



$$\boxed{Var(\bar{y}) = \frac{(1-f)}{a} s_a^2}$$

• Suppose we select an SRS of a = 20 classrooms from A = 1,000, and examine the immunization history of only b = 12 children in selected classrooms
• Here again
N = A * B = 1000 * 24 and n = a*b = 240
• For each of the a = 20 selected PSU's, we record the number of children immunized:

$$\frac{4}{12}, \frac{5}{12}, \frac{5}{12}, \frac{6}{12}, \frac{6}{12}, \frac{6}{12}, \frac{7}{12}, \frac{8}{12}, \frac{8}{12}, \frac{8}{12},$$
$$\frac{8}{12}, \frac{8}{12}, \frac{8}{12}, \frac{9}{12}, \frac{9}{12}, \frac{10}{12}, \frac{10}{12}, \frac{11}{12}, \frac{12}{12}, \frac{12}{12}$$

• **Again, the overall proportion immunized is**
$p = 160/240 = 0.67$

• Again as in cluster sampling treat the sample as an SRS of a = 20 units from A = 240:

$$\mathrm{var}(p) = \frac{(1-f)}{a} s_a^2$$

• **Where**

$$s_a^2 = \sum_{\alpha=1}^{a} (p_\alpha - p)^2 \Big/ (a-1) \qquad f = \frac{a}{A} \times \frac{b}{B}$$

• **That is,**

$$\mathrm{var}(p) = \frac{(1-f)}{a} \frac{\sum_{\alpha=1}^{a} (p_\alpha - p)^2}{a-1}$$

$$se(p) = \sqrt{\mathrm{var}(p)}$$

• The design effect for two-stage sampling is the same as for simple cluster sampling:
Deff(p) = var(p) / var_SRS(p)
• Selecting many elements per cluster increases variances
• As noted before, even small values of roh can be magnified by large b since
Deff(p) = 1 + (b-1) roh
• One way to think about the design effect now is to see how it affects potentially the sampling variance

• Remember

var(p) = Deff(p) * var_SRS(p)

• If we keep the same sample size, then the SRS sampling variance does not change

• Then any change to the design effect is a change to the sampling variance

• Manipulation of sampling fractions between first and second stages, maintaining the overall sample size, reveals the nature of the design effect, and the effective sample size

• Sample a = 20 classrooms and b = 12:

Deff(p) = 1 + (12-1) * 0.088 = 1.97 | *n_eff* = 122

• Sample a = 30 classrooms and b = 8:

Deff(p) = 1 + (8-1) * 0.088 = 1.62 | *n_eff* = 148

• Sample a = 80 classrooms and b = 3:

Deff(p) = 1 + (3-1) * 0.088 = 1.18 | *n_eff* = 204

3.4 Designing 2-stage samples

• **Estimation**

$$\mathrm{var}_{(1)}(p) = \frac{(1-f)}{a} s_a^2$$

$$\mathrm{var}_{(1),SRS}(p) = (1-f)\frac{p(1-p)}{n_{(1)}-1}$$

$$deff_{(1)} = \frac{\mathrm{var}_{(1)}(p)}{\mathrm{var}_{(1),SRS}(p)}$$

$$roh = \frac{deff_{(1)}-1}{b_{(1)}-1}$$

• **Projection**

$$\mathrm{var}_{(2)}(p) = deff_{(2)} \times \mathrm{var}_{(2),SRS}(p)$$

$$\mathrm{var}_{(2),SRS}(p) = \frac{p(1-p)}{n_{(2)}}$$

$$deff_{(2)} = 1 + (b_{(2)}-1)roh$$

$$\boxed{roh}$$

CASE STUDIES

• (CASE A) Suppose the sample described in exercise 4 (with n = 2400 and a = 60) is to be repeated with a smaller sample of n = 1200 and in only a =30 equal sized clusters. Project (say what is expected) how large the sampling variance of p will be under this new design.

• (CASE B) Suppose the reduced size of n = 1200 is retained, but we want to consider a = 60 equal sized clusters. Project how large the sampling variance of p will be under this new design.

➔ For A), compute the simple random sampling variance

Var_SRS(p) = p(1-p) / (n-1)

➔ computed the design effect

Deff(p) = 1+(b-1) * roh = 1 + (40 – 1) roh

➔ compute the projected sampling variance

Var(p) = var_SRS(p) * deff(p) =

➔ For B), repeat the above steps, replacing b = 40 with b = 20

• For A), n = 1200 for a = 30 and b = 40

Var(p) = deff(p) * var_SRS(p)

Here deff(p) = 2.1796

Thus, using the design effect and new SRS variance, we can obtain var(p) under the new design

And ignoring the fpc,

var_SRS(p) = p(1-p) / n = 0.4 * 0.6 / 1200 = 0.0002

Then, var(p) = 2.1795 * 0.0002 = <u>0.0004358</u>

• For B), when a = 60 and b = 20 for n = 1200,

Deff(p) = 1+(20-1)(0.3024) = 1.575

Var(p) = 1.575 * 0.0002 = <u>0.0003150</u>

n	*a*	*b*	*deff*	*var(p)*
2400	60	40	2.1795	.000218
1200	30	40	2.1795	.000436
1200	60	20	1.5750	.000315

• Design effects, when projected, can also help us determine sample size in cluster sampling
• Cluster sampling increases variances by a factor
Deff(p) = 1 + (b-1)roh
Compared to SRS…
• Let's 'offset' this increase by increasing sample size by
deff(p) = 1 + (b-1)roh
• That is compute an **SRS sample size** and **inflate it by a design effect**
• For example, suppose, for our proportion p = 0.4 we want a 95% confidence interval (0.37, 0.43)
- this is margin of error of 0.03 …
- or a standard error of 0.015

• **Which for a proportion yields an SRS sample size**

$$n_{SRS} = \frac{S^2}{\left[se(p)\right]^2} = \frac{(0.4)(1-0.4)}{\left[0.015\right]^2} = 1066.67$$

• **If the cluster sample has *deff* = 2.1795, the sample size for the cluster sample would be**

$$n = n_{SRS} \times deff(p) = 1066.67 \times 2.1795 \approx 2{,}325$$

• We can take the variance projection one step further, and project what a 95% confidence interval
• For B), when a = 60 and b = 20 for n = 1200,
Deff(p) = 1 + (20-1)(0.03024) = 1.575
Var(p) = 1.575 * 0.0002 = 0.0003150
… the 95% confidence interval, using the 'Normal' distribution multiplier is

$$\left(0.4 - 1.96 \times \sqrt{0.000315}, 0.4 + 1.96 \times \sqrt{0.000315}\right)$$
$$(0.365, 0.435)$$

3.5 Unequal sized clusters

• Naturally occurring clusters tend of unequal in size
• Fixed sampling rates and unequal sized clusters result in variation in sample size

The problem

Hospital	B_a	**Hospital**	B_a
1	420	7	60
2	180	8	60
3	120	9	720
4	600	10	1860
5	240	11	1140
6	360	12	240

• An epsem sample of n=100 employees is desired from the N = 6000
- select a = 2 hospitals
- f = 100 / 6000 = 1/60
• First select SRS a = 2 (a rate of 1/6 cuz there are 12 hospitals in total)
- and then choose employees at the rate 1/10 within the selected hospitals
f = (1/6) * (1/10) = 1/60
• Suppose hospitals 2 and 6 are chosen
- subsampling at the rate of 1/10 yields sample size

N = (180+360) / 10 = 18+36 = 54
• If hospitals 2 and 10 were chosen, though,
N = (180 + 1860)/10 = 18+ 186 = 204
• subsample size varies
• sample administration becomes difficult
• variation in the overall sample size is undesirable
• since n is a random variable, y_bar = (1/n) sigma(1 to n)y_i, no longer applies
• we need to use a ratio estimator

$$r = \frac{\sum_{\alpha=1}^{a} y_\alpha}{\sum_{\alpha=1}^{a} x_\alpha} = \frac{y}{x}$$

• seeking to control $x = \sum_{\alpha=1}^{n} x_\alpha$
• controlled sample size provides administrative convenience in fieldwork
• also provides greater statistical efficiency of estimators
• several methods
  - select exactly b elements per cluster
  - probability proportionate to size (PPS)
• Suppose a = 2 and b = 50 employees per selected hospital are chosen
  - sample size is n = 100, and does not vary by which hospitals are chosen
• This design will on average across all possible samples over-represent employees in small hospitals
  - the probability of selection of small hospital employees is higher
• For example, for hospital #2, f=(1/6)(50/180) = 1/21.6
• While for hospital #10, f=(1/6)(50/1860)=1/223.2
• The variation in rates can be remedied through **weighting**

PPS (probability proportionate to size)
• Require a method that is
  - epsem
  - achieves equal sized subsamples in clusters
• again, consider a = 2 (2 hospitals) and b = 50 (50 employees from each cluster)
• In order to achieve epsem, the following must be the "selection equation":
f = 1/60 = P{alpha} * (50 / B_alpha)
P{alpha} = (1/60) * (B_alpha /50) = (B_alpha /3000)

Hospital	B_α	Cum. B_α
1	420	420
2	180	600
3	120	720
4	600	1320
5	240	1560
6	360	1920
7	60	1980
8	60	2040
9	720	2760
10	1860	4620
11	1140	5760
12	240	6000

• **Re-expressing,**

$$P\{\alpha\} = \frac{2 \cdot B_\alpha}{6000} = \frac{2 \cdot B_\alpha}{\sum_\alpha B_\alpha}$$

• **In general, this becomes, across two stages,**

$$f = P\{\alpha \; and \; \beta\} = \frac{a \cdot B_\alpha}{\sum_\alpha B_\alpha} \cdot \frac{b}{B_\alpha} = \frac{a \cdot b}{\sum_\alpha B_\alpha} = \frac{n}{N}$$

• Select Random Numbers (RNs) from 1 to 6000, say …
  - RN = 702
  - RN = 1744
• Find the first hospital with cumulative sum greater than or equal to the first RN
• Find the next hospital with sum greater than the second RN
• Theses choose hospitals 3 and 7

• Alternatively, select one RN from 1 to the interval 6000/2 = 3000 ➔ say RN = 702
• Find the selected hospital, as above
• Add the interval to the RN to obtain + 3000 = 3702
• Find the second hospital with this selection number, as above
• The RN yields hospitals 3 and 10

3.6 Subsample Size

Cost Model

• Projecting standard errors and confidence intervals for cluster sampling depends on b and deff
• estimating sample size for cluster sample sizes depends on b and deff
• that is, **knowing b and roh** leads to a **projected deff & sample size n**
• We know that as "b" goes up or down deff goes up or down
• And var(p) follows
• But we also have seen that as "b" goes up or down "a" goes down or up
• And as "a" goes down or up the cost of the data collection goes down or up
• There is a **cost-error trade-off** in cluster sample design
• Can we choose any set of "b" and "a" as long as we don't exceed budget?
• Or is there a choice, an optimum choice for "a" and "b" that gives us the best (minimum sampling variance) among all possible choices for the given budget?
• There is an "optimum" choice for "a" and "b"
• It can be obtained by minimizing the sampling variance for fixed cost (or vice versa)
• Cost model for two stage sampling:

$$C - C_0 = a\,c_a + a(b\,c_b)$$

• **$C - C_0$ is the budget available, after overhead costs are removed**

• **c_a is the cost per cluster**
• **c_a is dominated by travel and preparation costs**

• **c_b is the cost per observation within a cluster**
• **c_b is dominated by interviewing costs**

• There is corresponding "sampling variance" model for two stage sampling:

$$\mathrm{var}(p) = \frac{(1-f)\,p(1-p)}{ab-1}\left[1+(b-1)roh\right]$$

• As "a" goes up or down, the sampling variance goes up or down
• The relationship between "b" and sampling variance is more complicated
• The optimum subsample size for fixed cost C – C_0 can be found by a calculus or algebraic approach
• Finding "b" that minimizes the sampling variance
• The optimum b is:

$$b_{opt} = \sqrt{\frac{c_a}{c_b}\cdot\frac{1-roh}{roh}}$$

• **As c_a increases, b increases**
• **As c_b increases, b decreases**
• **As roh increases, b decrases**

• **For example, if *roh* = 0.01, then** $\frac{1-roh}{roh} = \frac{1-0.01}{0.01} = \frac{0.99}{0.01} = 99$

• **But if *roh* = 0.05, then** $\frac{1-roh}{roh} = \frac{1-0.05}{0.05} = \frac{0.95}{0.05} = 19$

• **More homogeneity within, take fewer observations within …**

• **What about *a*?**

• **Consider the cost model again:**

$$C - C_o = ac_a + (ab_{opt})c_b$$

• **Solve for *a*:**

$$a = \frac{C - C_o}{c_a + b_{opt}c_b}$$

• **For example, suppose c_a = \$65.40 and c_b = \$25**
• **If *roh* = 0.05 (for a single variable, or on average),**

$$b_{opt} = \sqrt{\frac{65.40}{25}\cdot\frac{1-0.05}{0.05}} = 7.05$$

• **And if we had $C - C_0$ = \$10,000, then**

$$a = \frac{C - C_o}{c_a + b_{opt}c_b} = \frac{\$10{,}000}{\$65.40 + 7.05\times\$25} = 41.38 \approx 41$$

• **We might in this case increase *b* to obtain an integer value for *a* that meets the budget exactly**