**INTRODUCTION TO SAMPLING**

**(D-Lab Workshop)**

**1. BASIC SAMPLING CONCEPTS**

**Population of interest: what we want to talk about (the target population)**

Define UNITS in SPACE and TIME  
 This is the population that we want the results of our survey to represent

But it may not be feasible to sample from the whole population

**Survey population: what it is practical or realistic to survey**

Excludes some elements of the population of interest -too difficult or costly to reach -examples: homeless people, jails, army barracks

Should be clear on the difference between the survey population and the target population

**Frame: basis for sampling**

Operationalizes the concept of the survey population: Lists, maps, OR a procedure to include all units of the survey population **Coverage errors - potentially a problem**

Defects of the frame or of the implementation of the sampling plan

-bad or incomplete or out-of-date lists (few lists are perfect) -erroneous identification of physical boundaries in area samples

**These errors may not even be detected**

**Characteristics of a probability sample (required for statistical inference)**

1. Each member of the population has a chance to be selected

2. A random method of selection is used

3. We know the probability of selection for each unit

(at least in comparison with other selected units)

**Sampling error - the basic idea** Replicability of results: What would happen if I had selected other areas  
 or other individuals? How likely would I have been to get the same answers?

Pattern of variability of the variable of interest affects the size of the confidence intervals

Sample size also affects the size of the confidence intervals

Sampling variance for SRS = (element variance) / (sample size)

Design effect: DEFF = variance (complex sample) / variance (SRS of the same size)

**Non-response - usually a BIG problem** Results in self-selection, instead of random selection from the population

Affects our confidence in extrapolating to the population from which the sample was drawn

May (or may not) result in incorrect inferences about the population

**2. BASIC METHODS FOR SELECTING SAMPLES**

**A. PREPARE THE SAMPLING FRAME**

**Ideal situation**

Every unit in the survey population appears:

**once** (none missing)

**only once** (no duplicates)

**mixed with nothing else** (no ineligible units)

**General solutions to frame problems**

**Correct the frame**

eliminate blanks, duplicates, ineligibles

find missing elements

**Redefine survey population** to fit the available frame—for example:

exclude homeless people

exclude group quarters

exclude non-telephone households

**Ignore the problem**, and perhaps try to adjust later (with **weights**)

**B. SAMPLING TECHNIQUES**

**1. Simple random sampling (SRS)**

**How: use a random number table (or computer)** **HANDOUT**

**WITH versus WITHOUT replacement**

WITH replacement: can select same element twice (or more)

WITHOUT replacement: usually better; more information per selection

**If blanks or ineligible units are on the list:** **DIRECTORY EXAMPLE**

If selected, simply discard (= screening)

Do NOT take next unit on the list (increases its probability of selection)

Increase sample size to compensate for expected blanks

**2. Systematic sampling (random start, then fixed interval)**

**Why not SRS?**

convenience; condition of the frame (e.g., files on shelves in a room)

avoid duplicate selection

**may want to preserve a certain order**

**Watch out for patterns in a list**

periodic patterns that match the selection interval

monotonic trend throughout the list

Low random start different from high random start

**Some remedies for undesirable patterns**

divide list into sublists, with a separate random start for each

draw several systematic samples from a list:

if interval was 100, draw 10 samples with interval of 1000

**But remember that some ordering can be deliberate -- called “implicit stratification”**

**How to select a systematic sample**: **WORKSHEET**

Interval = (total on the list) / (desired number of selections)

Use a random start (RS) <= interval

**If interval is a whole number, easy**

selections: RS, RS+interval, RS+2int, RS+3int, ...

**If the interval is a fraction (as it usually is)**, 3 common solutions:

**a. Round the interval down**; and select a few more cases.

For example, if you want approximately 100 out of 1025,

the exact interval would be 1025/100 = 10.25

Just round the interval down from 10.25 to 10,

and select 1/10 of 1025, which gives either 102 or 103,

depending on the random start.

**b. Eliminate some cases at random from full list BEFORE selection,**

to allow using an interval that is a whole number.

For example, eliminate 25 at random from 1025, then select 100/1000.

f = (1000/1025) \* (100/1000) = 100 / 1025

Do not eliminate extras at random AFTER selection;

This does not result in exactly equal probability of selection.

(See Kish, pp. 115-116)

**c. Use a fractional interval, then truncate**.

Selections:

Truncate(RS), Truncate(RS+interval), Truncate(RS+2\*interval) ...

**This is the easiest method to use, when selecting by hand.**

(Computer programs may use intervals, selection ranges, and unit numbers with many decimal places, without truncating, but it is difficult and unnecessary to do this by hand. In any case, apply one method consistently for each application.)

**3. Sampling with Probability Proportional to Size (PPS)** : **WORKSHEET – hand version**

Each unit or record has a **measure of size** (MOS)

-can be an estimate or an adjusted MOS

-sometimes rounded to a multiple of n (the sample size)

Calculate the cumulation of the MOS across records,

and **get the selection range for each record or unit.**

**Selection methods:**

**a) Random selection with replacement (sometimes used)**

Pick *n* random numbers between 1 and **Sum(MOS)**.

Select the unit into whose selection range each random number falls.

If a unit is picked twice, weight double or take 2 subsamples.

**b)** **Systematic selection (generally used)**

**interval = Sum(MOS) / (number of units to select)**

If the interval is a whole number, it is easy to apply

(If the work is going to be done by hand, it is sometimes worth adjusting the MOS before selection, in order to get an interval that is a whole number.)

**If the interval is a fraction, use the procedure for fractional intervals.**

**-By hand, use the truncation method.**

**-Computer programs will use lots of decimal places**

**(≥ number of decimals in the interval) .**

**PPS selection issues PPS EXERCISE**

**If the MOS for a unit is bigger than the interval,**

allow for multiple selections and subsamples

or set the unit aside as a separate stratum

**If a unit is too small** (< min MOS) to allow subsampling (in a 2-stage design),

it must be linked with one or more other units.

This can be done before selection (by hand)

or after the selection is done by an algorithm (see Kish, pp. 244-245).

Note that this is a common problem for units with MOS=0.

For example, a block that previously had no households could have new construction by the time the sample is implemented. But a unit with MOS=0 has NO chance to be selected, unless it is grouped with another unit.

**3. CLUSTER SAMPLING**

**1. Meaning of Clustering**

Desired elements are **in groups**

-e.g., students in classrooms, people in cities

-Sample **only some of the groups**

Members of selected groups **must also represent members of unselected groups**

**Compare this to element sampling**, in which people are **sampled directly**

-**without limiting selection of people to the selected groups**

**2. Reasons For and Against Clustering**

**FOR** clustering

-**COST**: time and travel expense; e.g., area sample of state, or even of a city

-easier to standardize procedures, fewer sites

-avoid listing the entire population; do only for the last stage of sampling

**AGAINST** clustering

-**less new info** than you might expect, if clusters are relatively homogeneous

**-increase in sampling error, compared to SRS of same size**

-also harder to calculate sampling errors

-clusters are usually created for some other purpose; e.g., classrooms or clinics

-may not be suitable for what we want to do

**3. Understanding the Effect of Clustering**

**Intra-cluster correlation (roh)–rate of homogeneity (ranges from 0 to 1)--Kish pp. 161-164**

**DEFF = 1 + roh(b-1), where *b* is the average size of the clusters**

**Extreme cases:**

-All people in the same cluster **give the same answer, but different** from other clusters

-effective sample size = number of clusters (groups**); roh = 1**

-M**ean answer** **within each cluster is the same** as the overall mean

-as if people were assigned at random to groups; **roh = 0**

**-**could take the whole sample from just one cluster and get the correct answer

Usually the situation is in between, but **we only know after the fact,**

**and each variable is different**. **CLUSTER SIZE EXERCISE**

We can **estimate *roh*** ahead of time based on other studies with similar variables.

This is necessary if we try to optimize our sample design.

*Roh* is actually calculated after the fact as (DEFF-1) / (b-1)

where DEFF is calculated by a program to compute complex standard errors.

**4. STRATIFICATION**

**A. Meaning of Stratification HMO PATIENT PROBLEM**

Divide the sampling frame into parts (strata),

then draw a sample from **EVERY** stratum.

Difference between strata and clusters (even if they use the same type of units)

-Strata: select a sample **from EVERY stratum**

-Clusters: select a sample **ONLY from the selected clusters**

**B. Reasons to Stratify**

1) **Ensure adequate coverage of some parts of population**

-use the SAME sampling fraction in all strata

-each stratum will have its "fair share" in sample

2) **Want to oversample some parts of the population**

-use a DIFFERENT sampling fraction in various strata

-will need to use weights to compensate for different probabilities of selection

-note that prior information on strata is needed

-otherwise must screen (e.g., for age or race)

3) **Hope to reduce sampling error**

-combined variance is the (weighted) sum of stratum variances

-depends on homogeneity of strata on the variables being estimated

-try to capture all variation by stratum definitions

-hard to do, especially for more than one variable

-look for stratifiers correlated with variables to be estimated

-e.g., affluence of area, for some health variables

-check results of other studies for good stratifiers

-**hope** to offset (partially) the effect of clustering

**C. Methods of Stratification**

**EXPLICIT stratification**

-Actually divide the frame into separate files or lists

-**Necessary if using different sampling fractions**

**IMPLICIT stratification**

-Sort the file or list by the stratifying variable(s)

-then select units by systematic random sampling

-More practical than sampling separately from many strata

-But be careful of the effect of the random start, if the selection interval is large. **COMBINE the two methods (very common)**

-Divide the frame into major strata, based on one or more variables

-Sort each stratum list by 1 or 2 other stratifying variable(s)

-then select systematically within each major stratum

**5. REQUIRED SAMPLE SIZE**

**Understand the two separate problems:**

**A. How many cases you want to end up with**

**B. How big the sample needs to be to end up with *A* cases**

**Problem *B* is the bigger problem*.***

**A. Figure Out How Many Cases You Want To End Up With**

**Decide what STATISTIC you must estimate - and for WHAT GROUP(S)**

Multi-purpose surveys (with many variables) can be a problem

Focus on the hardest statistic to estimate (biggest variance) or some compromise

**Must also consider desired estimates for subpopulations (like regions or small areas)**

**The smallest required subgroup will affect the required overall *n***

**This is the reason why big surveys have big samples**

Might require disproportionate sampling, to get enough of some small group(s)

For estimating percentages, see the table of 95% confidence intervals **HANDOUT**

For calculating precision, power, and differences, can use Stata

**B. Figure Out How Big The Sample Needs To Be, To End Up With *N* Cases**

(This is usually the part that causes problems for researchers.)

**Adjust the desired *n* for various estimated factors**

**Occupancy rate (OR)**

In a HH sample, proportion of HHs not vacant

In a phone sample, proportion of numbers belonging to HHs

**Eligibility rate (ER)** Proportion of sample units (usually households) that will have one or more members of

the survey population **For studies using screening, this estimate is VERY IMPORTANT Response rate: (RR)** Make a realistic estimate of the response rate you anticipate.

**Divide the desired *n* by the product of the factors**

For example: for OR = .95, ER = .90, RR = .70

If desired completed n was calculated to be 1,000,

**Required sample size = 1,000 / (.95 \* .90 \* .70) = 1,671**

**(Always check the result: 1,671 \* .95 \* .90 \*.70 = 1,000) WORKSHEET**

**Allow for uncertainty: calculate best-case/worst-case sample size to select**

For example, for a study involving screening in which only 12-15% of the households will have an eligible person, and you want to end up with 200 completed cases, you might project the following situation, assuming occupancy rates of 94-96% and response rates of 45-50%:

**Best case**

OR = .96, ER = .15, RR = .50

Required sample size = 200 / (.96 \* .15 \* .50) = 2,778

**Worst case**

OR = .94, ER = .12, RR = .45

Required sample size = 200 / (.94 \* .12 \* .45) = 3,940

**In this situation, you should select a large enough sample for the worst-case scenario. Then subselect at random enough sample for the best-case scenario, and start the survey with those sample selections. But be prepared to add more selections as needed, up to the sample size required for your worst-case scenario.**

**Source of information on these factors:**

**Field outcome reports for previous studies HANDOUT**

**C. Administer the Sample by Using Replicates or a Reserve Sample**

**In the example above, begin field work with at least 2,778 sample units, but be prepared to put up to about 3,940 into the field.**

In this situation, you could think of designing a sample of 4,000 households, which you would then divide at random into 16 replicates of 250 each. You would begin fieldwork with 11 of the replicates, for an initial sample of 2,750. The other 5 replicates would be used as needed. (If necessary, a partial replicate could also be used, by dividing a full replicate into random parts.)

**The various adjustment factors would be monitored as field work proceeds**, to see if field results are getting closer to the best-case or to the worst-case scenario. As soon as it became clear that the best-case scenario was not going to materialize, you would start putting in some of the reserve sample.

**Remember that you must complete fieldwork on all of the replicates or reserves that you put into the field. If you just stop doing callbacks in a replicate when you reach a target number of completes, you will only get the easy cases in that replicate, and your results may be biased.** So be careful not to put more replicates or reserve cases into the field than you can work thoroughly with your usual procedures.

**6. Sampling Rare Populations**

**Select a large random sample (to get a few rare population members)**

-With or without screening

-Gold standard, but very expensive

**Oversample selected areas or strata**

-Relies on ability to identify areas or strata with greater proportions

-Needs weights to compensate, but still a probability sample

**Specialized lists (if available)**

-Example: membership lists of ethnic clubs or associations

-Example: surname lists from the phone book

-Unknown bias if relied on exclusively

-Can be combined with a more general frame

**Snowball sampling**

-Start with a random sample

-Rare group members give referrals to others

**-**Can calculate probabilities of selection and get weights

-Start with a convenience sample

-Sample members give referrals to others

-Unknown bias

**Respondent-driven sampling (RDS)**

-More sophisticated version of snowball sampling

-Start with a convenience sample (the “seed”)

-Issue about 6 tickets to each seed, to recruit 6 others eligible for the study

-The new recruits do the same – up to 3 or 4 cycles

-Relies on a good network of all members of the target group

-Common problems

-Seeds unlikely to recruit members of another subgroup of the same

target population

-Chain can die out too soon – need to re-seed

-If respondents are paid, some referrals will try to refer relatives

or friends previously referred by others

-Tight administrative control is necessary.

(See article in *Survey Practice* on RDS experience and problems)

**Suggested Readings**

Robert M. **Groves**, et al., *Survey Methodology*, 2nd edition, Hoboken, NJ: John Wiley and Sons,

2009. [Best current summary of survey methodology; includes sections on sampling and weighting]

See especially pp. 97-138 on sampling.

Leslie **Kish**, *Survey Sampling*. New York: John Wiley and Sons, 1965, 1995.

[Comprehensive work on sampling, with many examples and illustrations; a basic reference for survey samplers]

Thomas **Piazza,** “Fundamentals of Applied Sampling,” chapter 5 in the *Handbook of Survey Research,* 2nd edition, edited by Peter V. Marsden and James D. Wright. Bingley, U.K.: Emerald Group Publishing, 2010.

[Basic introduction to survey sampling]