Introduction to Complex Sample Surveys

Michael Laviolette, PhD, MPH
Senior Management Analyst
Office of Health Statistics and Data Management
mlaviolette@dhhs.state.nh.us
December 16, 2009

Why surveys?

- As public health professionals, we need health data for assessment and evaluation
- Primary ways to gather data are experiments and observational studies
- Experiments impose a treatment on individuals to observe responses; observational studies do not
- Experiments are used in clinical trials and give most reliable evidence for causation
- For ethical and practical reasons, experiments are not usually feasible in public health, so we use observational studies (surveys)
- CDC surveys include BRFSS, PRAMS, NHANES, YRBS, NHIS

Some important national surveys

- American Community Survey (ACS)
  - Ongoing Census survey on community characteristics
- Current Population Survey (CPS)
  - Census and Bureau of Labor Statistics
  - Data on labor force, including unemployment rate
- National Health and Nutrition Examination Survey (NHANES)
  - Includes interviews and physical exams
- National Health Interview Survey (NHIS)
  - Large-scale household interview survey
  - Personal interviews with 35,000-40,000 households per year
- Pregnancy Risk Assessment Monitoring System (PRAMS)
  - Maternal attitudes and experiences
- Youth Risk Behavior Surveillance System (YRBSS)
  - Youth and young adults; includes school-based component
- Behavioral Risk Factor Surveillance System (BRFSS): today's focus

Behavioral Risk Factor Surveillance System (BRFSS)

- Ongoing state-based telephone survey to track health status and risk behaviors throughout U.S.
- Begun 1984 in 15 states; NH entered 1987
- Now in all 50 states, D.C., three territories
- Three parts
  - Core questions—common to all 54 areas
  - Optional modules on specific topics
    - NH uses diabetes and childhood asthma prevalence
  - State-added questions
    - NH asks town of residence and radon awareness
- NH version about 120 questions and 30 minutes

Survey terminology

- Usually we want information about some defined target *population*
  - e.g. adults resident in NH during 2009
- Population is a collection of *elements* (units on which data is collected)
  - Often persons, could be households, schools, or other unit
- Often practical constraints dictate modifications from target population
  - e.g. institutionalized individuals usually excluded
- When population is large, enumeration is not practical, so we survey a portion instead—called a *sample*
- Plan for selecting elements for sample is called sampling *design*

More terminology

- *Subpopulation*: defined subset of population (e.g. diabetics)
- *Domain*: partition of sample into mutually exclusive and exhaustive subpopulations based on a variable
  - e.g. gender; domains are male, female
- *Parameter*: characteristic of a population
  - Mean body mass index of NH adults
  - Proportion of NH adults with diagnosed diabetes
- *Statistic*: characteristic of a sample
  - Average BMI among sampled adults

Objectivity in survey design

- Many popular surveys are *biased*; that is, they systematically favor certain outcomes
  - Self-selected samples (Internet polls)
  - Convenience samples (mall surveys)
  - Judgment sampling
- Flaw in above is that samples are not probability-based
- In a *probability sample,* each element has a known nonzero probability of being selected into the sample

Advantages of probability sampling

- Well-defined theoretical framework
- Minimizes (but does not eliminate) bias
- Permits quantification of uncertainty of results
  - Uncertainty arises because different sets of elements in sample will yield different survey results
  - Primary tools are confidence intervals and significance tests

Common sampling designs

- Simple random sample
- Stratified sample
- Systematic sample
- Cluster sample
  - Single-stage
  - Multistage
- Combinations of above

Simple random sample (SRS)

- Need complete list of elements, called the *frame*
- Population contains N elements; want to choose SRS of n elements without replacement
- Suppose $N = 20$ and $n = 4$
  - Label each element 1 to 20
  - Use random number table or generator to select four distinct integers from 1 to 20
    - e.g. 9, 18, 4, 17
  - Elements with corresponding labels enter the sample (order not important)

Definition of SRS

- A sample of n elements from a finite population is a *simple random sample* (SRS) if every possible set of n unordered elements is equally likely to be chosen as the sample
- If $N = 20$ and $n = 4$, say, then all 4,845 possible sets are equally likely to become the sample
- If each *element* has an equal probability to enter the sample, the design is called *EPSEM* ("equal probability selection method")
- SRS is EPSEM, but not only EPSEM

Properties of SRS

- Advantages
  - Conceptually simplest probability sampling method
  - Building block for more complex designs
  - Benchmark for evaluating efficiency of more complex designs
- Disadvantages
  - Need the frame; not always feasible to get it
  - Sample may have wide geographic dispersion
  - If an important subpopulation is small, may not get enough of its elements with an SRS
- SRS is BRFSS design in Guam and U.S. Virgin Islands

Discrete data

- *Nominal* variables place each element in one of several categories with no inherent order
  - Gender (male or female)
  - Ever had colonoscopy (yes, no, refused, unknown)
  - Statistic of interest is proportion or percentage
- *Ordinal* variables place the categories in some meaningful order
  - Health status (excellent, very good, good, fair, poor)
  - Usually interested in proportions and percentages here as well

Continuous data

- *Continuous* variables consist of numbers for which computing a sum or average makes sense
  - Height, weight, BMI, not SSN
- Usually interested in means or totals
- Ratio scale for continuous data: zero is absolute
  - Zero signifies absence of quantity being measured
  - Appropriate comparisons are ratios and differences
- Interval scale for continuous data: zero is relative
  - Zero is an arbitrary point on scale, e.g. temperature in ^{o}C or ^{o}F
  - Appropriate comparisons are differences
  - Most practical problems focus on differences
- Continuous variables are often collapsed into ordinal categories
  - e.g. BMI (normal, overweight, obese)

Population parameters

Suppose we measure some continuous variable Y on a population with N elements $Y_1, Y_2, \ldots, Y_N$

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N}$$

Population mean

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N-1}$$

Population variance

Standard deviation = Square root of variance

Sample statistics

From the population we draw a simple random sample of n elements $y_1, y_2, \ldots, y_n$

$$\hat{\mu}_{SRS} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Sample mean (average)

$$\hat{\sigma}^2_{SRS} = s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

Sample variance

Variation of sample averages based on SRS drawn from infinite population

- The sample average $\bar{y}$ is an "unbiased" estimator of the population mean μ
  - Tends to neither overestimate nor underestimate
- The variance of $\bar{y}$ is σ^2/n
  - Variability decreases as sample size increases
- As n gets large, distribution of $\bar{y}$ over all possible samples tends to the "normal" bell curve (*Central Limit* effect)

Variation of sample averages from a finite population

The variance of the estimator $\bar{y}$ under simple random sampling without replacement is

$$\mathrm{var}_{SRS}(\bar{y}) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

n/N is the *sampling fraction*

$1 - n/N$ is the *finite population correction factor* (fpc)

The larger n is, the smaller the variation in $\bar{y}$

If n is small compared to N, fpc is negligible

Standard error of sample average

Substitute sample estimate s for unknown σ in variance formula; take square root

$$se_{SRS}(\bar{y}) = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$$

If n is large but still small compared to N, fpc is negligible and the sampling distribution of is approximately normal

Confidence interval

The range of values

$$\bar{y} \pm 1.96 \times se_{SRS}(\bar{y})$$

is a 95% *confidence interval* for the unknown population mean μ

(the values ±1.96 cut off the middle 95% of the standard normal curve)

Sample size selection

- Choose desired length of confidence interval and confidence level (usually 95%)
- Need estimate of population variance; can come from pilot study or guess
- With SRS, solve variance formula for required n

$$\text{var}_{SRS}(\bar{y}) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

Sample size with nominal data

- Suppose we are estimating a population proportion p and want the estimate to fall within some distance d from p
  - e.g. want to estimate proportion of binge drinkers within ±10 percentage points with 95% confidence
- Previous studies show p about 20%
- Required n is $(1.96/0.1)^2(0.2)(0.8) = 61.46 \cong 62$ (round up)
- Conservative approach assumes $p = 0.5$; this makes required n of 385
- Quick and dirty method; for 95% confidence $n = 1/d^2$
  - e.g. for $d = 0.04$, $n = 625$

Stratification

- Designed to increase precision (decrease variation) of estimates
- Using supplementary information, divide population into groups called *strata*
  - Strata are often geographic, but can be based on gender, SES, or other factor
- Within each stratum, elements should be as similar as possible
- For each stratum, determine sample size and take probability sample
  - *Stratified random sampling* takes SRS from each stratum
- Stratification strategies
  - Proportionate
  - Optimum
  - Disproportionate (BRFSS)
  - Balanced

Proportionate stratification

- Draw SRS from each stratum, using the same sampling fraction for all strata
  - Proportionate stratified random sampling is EPSEM
  - Precision of estimates from proportionate stratified sampling will never be worse than that from SRS
- Primary aim is to increase precision, though gain is often modest

Disproportionate stratification

- In *disproportionate* stratification, sampling fraction depends on stratum
- A higher sampling fraction (*oversampling*) is often used in small but important strata to obtain a large enough sample for precise estimation
  - e.g. LBW births in PRAMS
- Primary aim is to achieve small enough standard error to permit analysis of important subpopulations
- Elements must be assigned weights to restore original proportions and thus produce unbiased estimates
- Weighting process tends to increase standard errors compared to proportionate stratification

BRFSS sampling design

- Landlines in 41 states, including NH, are sampled using disproportionate stratified sampling (DSS)
- NH divided into 12 geographic strata
  - Manchester, Nashua, rest of Hillsborough County, other nine counties
- PSU's are household telephone numbers, obtained from industry database
- Within each geographic stratum, numbers are divided into blocks of 100 and further stratified into 'high-density' or 'medium-density,' depending on number of listed household numbers in block
  - NH BRFSS thus has 24 strata

BRFSS sampling design (2)

- Numbers are randomly selected and dialed by computer
- High-density strata are sampled at higher fraction for efficiency
- Interviews done by contractor, occasionally monitored by DHHS
- In 2008, rural areas (Coos County) were oversampled slightly
- Total NH sample size about 6000
- Funding issues may affect future sample sizes

Cell phones

- Through 2008, cell phones excluded
- Households with only cell phones now at about 20% in US and still growing
- Research indicates that cell-only households differ from those with landlines, resulting in biased estimates
  - Younger
  - Lower income
- Cell phones sampled in NH starting in 2009
  - Only core questions included
  - Additional questions specific to cell phones for safety
  - Statewide subsample; no geographic strata
- Cell phone sampling about four times as expensive per dialed number
  - By law, cell numbers must be dialed manually

Optimum allocation for stratified sampling

$$n_h \propto \frac{N_h \sigma_h}{\sqrt{c_h}}$$

- *Optimum allocation* divides a fixed total sample size n among H strata to achieve minimum standard error of estimated population mean or total
- Primary aim is to maximize precision of estimates within available resources
- Sample size in hth stratum is directly proportional to
  - Size of stratum population
  - Variation (standard deviation) in stratum
- Sample size in hth stratum is inversely proportional to square root of unit sampling cost in that stratum
- Optimal allocation oversamples heterogeneous strata and undersamples more expensive ones

Estimators of population mean using simple random sampling and stratified random sampling

$$\hat{\mu}_{SRS} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\hat{\mu}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h$$

Estimate in stratified sample is weighted mean of stratum averages

Standard errors of estimates of population mean

$$se(\hat{\mu}_{SRS}) = \sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$$

$$se(\hat{\mu}_{st}) = \sqrt{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)}$$

Different designs require different estimators!

Special software is needed for complex survey analysis

- The estimators $\hat{\mu}_{SRS}$ and $\hat{\mu}_{st}$ estimate the same quantity: population mean μ
- Standard errors are calculated differently for each design—in fact, the object of stratified sampling is to *reduce* the standard error
- Most statistical software calculates estimates as if the sample is an SRS from an infinite population
  - Basis for development of standard statistical methods: two-sample t test, simple and multiple regression, ANOVA, etc.
- In surveys, software must account for the sampling design, otherwise estimates will be incorrect

Cluster sampling

- Used to reduce costs or make sampling easier
  - Sampling from a wide geographic area is expensive
- Population is partitioned into *primary sampling units* (PSU); each PSU is divided into *secondary sampling units* (SSU)
- Probability sample of PSU's is taken
- Each selected PSU is enumerated (*single-stage design*), or a probability sample of SSU's is drawn from it (*multi-stage design*)

A prototype national multistage survey design

- Stage 1: Divide the U.S. into 2,007 PSU's
  - 428 with largest populations automatically sampled
  - Stratified sample of 326 chosen from remaining PSU's
- Stage 2: Divide each sampled PSU into blocks
  - Stratify blocks by demographics; take stratified sample of blocks
- Stage 3: Sort housing units in each block into clusters of four nearby units
  - Choose probability sample of clusters for interviews

Properties of cluster sampling

- Key principle in cluster sampling: for lowest standard error, clusters should be similar *to each other* but internally diverse
  - Opposite of stratification, where strata should be *different* from each other but similar internally
- Selected clusters need to represent unselected ones
- Cluster sampling almost always increases standard errors
  - Increase depends on variability between clusters and sample size per cluster
  - Cost depends on size and shape of PSU's
- Primary aim is usually to efficiently sample a widely dispersed area
- NHIS uses a multistage design

Stratification vs. Clustering

- Divide population into internally similar groups
- Draw sample of elements from each group
- Can get smaller standard error than with SRS of same total size
- Goal is to increase precision of estimates of entire population or important strata

- Divide population into internally diverse groups
- Draw sample of groups
- Usually less precise than SRS of same size
- Goal is to increase efficiency by sampling more in selected clusters to offset loss of precision

Issues in surveys

- Undercoverage: Eligible elements missing from frame
  - e.g. households with only cell phones
  - If missing elements differ from included, estimates will be biased
- Overcoverage: Ineligible elements included in frame
  - e.g. phone numbers for business, disconnected numbers
  - Element must be discarded if drawn and found ineligible
  - Main problem is added cost of calling replacement number

Problems in response

- Nonresponse
  - *Unit nonresponse*: Individual chosen for sample can't be contacted or refuses to cooperate
  - *Item nonresponse*: Respondent refuses to answer some questions; problem with sensitive topics (sexual behavior, drug use)
- Measurement error
  - If question is poorly worded, respondents will misunderstand it
  - Response bias: respondent does not answer truthfully (e.g. post-election polls)

Weighting

- If data is unweighted, all records count the same
  - Assumes all elements likely to be selected (only true if EPSEM)
  - Assumes undercoverage and nonresponse evenly distributed in population (usually not true!)
- To adjust for bias, each sampled element is assigned a weight
- BRFSS weights are designed to adjust for
  - Unequal probability of selection
  - Differences in demographics between sample and target population
- Weight interpreted as number of population elements represented by each sampled element
  - Total of weights = size of population
- After data collection, weights are further adjusted for undercoverage and nonresponse (poststratification)
  - Weights of respondents must usually be increased to compensate for refusals

General BRFSS weight

$$FINALWT = STRWT \times \frac{1}{NPH} \times NAD \times POSTSTRAT$$

- FINALWT = weight assigned to each respondent for analysis
- STRWT = stratum weight (probability adjustment)
- NPH = number of residential landlines in household
- NAD = number of adults in household
- POSTSTRAT = adjustment for undercoverage and nonresponse; forces sum of weights to population estimate for demographic category
  - Based on gender, race/ethnicity, age group

BRFSS weighting

- For NH 2008 BRFSS, weights calculated separately for Manchester and Nashua
- New ‘raking’ methodology to be introduced in 2010; designed to have sample match population with respect to
  - Age group by gender
  - Detailed race/ethnicity
  - Education
  - Marital status
  - Gender by race/ethnicity
  - Age group by race/ethnicity
  - Geography

Available software for complex sample surveys

- SAS survey procs (SURVEYLOGISTIC, SURVEYFREQ)
- SUDAAN ("SUrvey DAta ANalysis")
  - Runs as standalone or add-on to SAS
- SPSS "PASW Complex Samples" module
- Stata "svy" family of commands
- Epi Info
- R package "survey"
- Last two are free; all others commercial

Using survey software

- Generally need to specify which variables contain
  - Weights
  - Strata
  - PSU's
- Estimates of means and percentages adjusted for weight
- Standard error computed as appropriate for the sample design
  - Most popular method is linear approximation (Taylor method)
- Confidence intervals and hypothesis tests are constructed using appropriate standard errors
- In NH BRFSS, usually need only weights and strata
  - PSU is telephone number

Online resources

- CDC’s BRFSS home page
  - http://www.cdc.gov/brfss/
- BRFSS Operational and User’s Guide (2006)
  - ftp://ftp.cdc.gov/pub/Data/Brfss/userguide.pdf
- NH Health WRQS query system
  - http://nhhealthwrqs.org/
- NH BRFSS reports
  - http://www.dhhs.state.nh.us/DHHS/HSDM/behavioral-risk.htm
- Summary of survey analysis software
  - http://www.hcp.med.harvard.edu/statistics/survey-soft

Bibliography

- Cochran, W.G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.
- Hosmer, D.W. and Lemeshow S. (2000), *Applied Logistic Regression* (2nd ed.), New York: Wiley.
- Kalton, G. (1983), *Introduction to Survey Sampling*, Newbury Park, CA: Sage Publications.
- Kish, L. (1965), *Survey Sampling*, New York: Wiley.
- Korn, E.L. and Graubard, B.I. (1999), *Analysis of Health Surveys*, New York: Wiley.
- Lee, E.S., Forthofer, R.N., and Lorimer, R.J. (1989), *Analyzing Complex Survey Data*, Newbury Park, CA: Sage Publications.
- Moore, D.S. and McCabe, G.P. (2006), *Introduction to the Practice of Statistics* (5th ed.), New York: W.H. Freeman.
- Thompson, S.K. (2002), *Sampling* (2nd ed.), New York: Wiley.

Questions and comments

Michael Laviolette
Senior Management Analyst
mlaviolette@dhhs.state.nh.us
(603) 271-5688

Susan Knight
BRFSS Coordinator
sknight@dhhs.state.nh.us
(603) 271-4671