Introduction to Probability Sampling

Concepts, Practices and Pitfalls

Johnny Blair

July 2009

Considerations is sample design and sample selection

- Fundamental concepts of sampling
- Probability sampling within the survey process
  - Issues of implementation
  - Logic of some sample designs
  - Sources of error in sampling
- Some issues in sampling rare and special populations

Some reasons for sampling

- Exploratory
- Pretesting
- Testing relationships
- Estimate univariate characteristics of a population
  - Basic research
  - Information for policy decisions
  - Decisions for resource allocation

Estimation of population characteristics

- Average annual number of doctor visits by people 18-65 in the U.S.
- Percentage of voters in the U.S. supporting a public health insurance option
- Proportion of persons 18+ in California with on-line health records
- Mean number of sex partners annually of MSM's living in Los Angeles
- Total number of retail establishments that plan to purchase accounting software

The sampling process

- Population
- Unknown population parameter(s)
- Select a sample
- Take measurements on that sample
- Use sample measurements to estimate population parameter(s)

- *N*
- $\overline{Y}$
- *n/N*
- y_i
- $\overline{y}$

Sampling for good estimation

"….A *representative sample* is a sample which, for a specified set of variables, resembles the population…[in that] certain specified analyses….(computation of means, totals etc.) yields results…within acceptable limits set about the corresponding population values, …The mere statement or claim that a sample is representative of a population tells us nothing."

Stephan and McCarthy "Sampling Opinions: an analysis of survey procedures," 1968

Probability and non-probability samples

- Some non-probability samples
  - Convenience samples
  - Volunteer samples
  - Judgment
  - Quota
- Types of probability sample designs
  - Simple random sampling
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multi-stage samples

Probability samples

- Definition
  - Each population member has a known, non-zero chance of inclusion
  - Sample members are drawn with a random selection mechanism
- Characteristics
  - Statistical basis for estimating population characteristics
  - Estimates of precision of estimates (sampling error) are possible

Simple Random Samples

- Every population element has the same chance of selection, n/N
- Every sample of size n has the same chance of selection

Permits estimates of population parameters

Sample mean as estimate of population mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = est\,\bar{Y}$$

Sampling variance

- The sampling variance can be estimated from the sample itself

mean

$$v(\overline{y}) = \frac{(1-f)}{n} s^2$$

proportion

$$v(p) = \frac{(1-f)}{n-1} p(1-p)$$

fpc: *f=(1-n/N)*

Sample of CAPS employees

- Simple random sample
- Questions:
  - Household size (adults and children)
  - Total household income in 2008
  - Took a vacation in last 6 months
- Estimates:
  - Mean number of people in CAPS employee households
  - Average income
  - Proportion of households that took a vacation

Sampling variance, standard error, and confidence limits of estimated mean

- Sampling variance

$$v(\bar{y}) = \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 \Big/ n-1\right)$$

- Standard error

$$se(\bar{y}) = \sqrt{v(\bar{y})}$$

- Confidence limits

$$\bar{y} \pm 1.96(se(\bar{y}))$$

Sample size and the precision of sample estimates

For estimating a proportion, if the population variance, *p(1-p)* can be estimated (or maximized), then

$$v(p) = \frac{(1-f)}{n-1} p(1-p)$$

can be used to compute the necessary sample size, e.g.

$$n = \frac{p(1-p)}{v(p)}$$

Sample size and the precision of sample estimates, continued

And, substituting standard error for variance, for 95% confidence interval the required sample size is:

$$n = \frac{(1.96)^2}{se^2}(p(1-p))$$

Sample Sizes for 95% confidence intervals around estimates of proportions

Pre-survey estimate of *p*	Sample size for 95% confidence the post-survey estimate is within x% of the true proportion			
	1%	3%	5%	10%
• 1% or 99%	381	43	16	4
• 2% or 98%	753	84	31	8
• 5% or 95%	1,825	203	73	19
• 10% or 90%	3,458	385	139	35
• 20% or 80%	6,147	683	246	62
• 30% or 70%	8,068	897	323	81
• 40% or 60%	9,220	1,025	369	93
• 50%	9,604	1,068	385	97

Different ways to set the sample size

- Statistical methods
  - Based on targeted confidence intervals
  - Required hypothesis testing power
- Informal methods
  - According to previous (typical) practice
  - Based on similar research study
  - To achieve minimum cell sizes for subgroup analyses
  - Based on resource limitations

The sampling process

- Population
- Unknown population parameter(s)
- Select a sample
- Take measurements on that sample
- Use sample measurements to estimate population parameter(s)

- *N*
- $\overline{Y}$
- *n/N*
- y_i
- $\overline{y}$

probability sampling within the survey process

- Define target population
- Frame the population
- Select a sample design
- Decide on sample size
- Select the sample
- Collect the data
- Produce estimates
- Compute variance of the estimates

Sampling Error and Sampling Bias: Threats to the quality of the survey estimates

- Forms of Error
  - Nonsampling error
    - Measurement error
    - Frame error
  - Sampling error
    - sampling variance
    - **sampling bias**

Source of Bias

- **Coverage bias**
- **Nonresponse bias**

Some types of sample bias and why they matter

- Sampling bias
  - selection probability = 0
  - Unintended variable selection probabilities
- Coverage bias
  - data collection mode exclusion
  - language exclusion
- Nonresponse bias
  - non-contacts
  - refusals

ideally estimation is unbiased

An estimate is unbiased (for a particular process) if:

$$E(\bar{Y} - \bar{y}) = 0$$

Over all possible instances of the process {e.g. coverage, selection, response)

Factors affecting magnitude of bias

- Coverage bias

$$\overline{Y_C} - \overline{Y} = \frac{U}{N}(\overline{Y_C} - \overline{Y_U})$$

- Nonresponse bias

$$\overline{y_r} - \overline{y_s} = \frac{m_s}{n}(\overline{y_r} - \overline{y_m})$$

Total Survey Error

Mean square error:

$$MSE(\bar{y}) = \sum_r S_r^2 / n_r + (\sum_r B_r)^2$$

Total survey error [root mean square error]:

$$RMSE(\bar{y}) = \sqrt{\sum_r S_r^2 / n_r + (\sum_r B_r)^2}$$

The decision to participate and nonresponse bias

- Household survey
  - Topic: health behaviors
  - Estimate: mean number of doctor visits
  - Mode: phone
  - Decision to participate
    - Interviewer's manner
    - Sponsor
    - Available time
    - Topic

Defining a population

- The conceptual target population definition
  - All adults 18-65 in the U.S.
  - All families in California with children under 18
  - Males who have sex with other men
- Issues in developing an operational definition
  - The frame
  - Data collection mode

Developing an operational definition for a telephone survey

- Population: Males 18-55 who have sex with other men
- Frame
  - Geographic area
  - Households
  - Phone: landline, cell
- Screener
  - Age
  - Household
  - Defining behavior
    - Meaning of sex
    - Ever sex with male

Sample frames

- Frames are seldom developed for the purpose of sample surveys
- There may not be a frame that matches the population
- All frames have flaws
- Need to decide whether a frame is suitable

Overview of frame problems

- Omissions
- Ineligibles
- Multiples
- Clusters
- Information error

The need for other sample designs

- Deal with practical constraints
- Address analysis needs
- Accommodate data collection mode or needs
- Control costs
- Improve precision
- Nature of the population

Different approaches to determining the sample design and sample size

- Specified error for minimal cost
- Minimal error for specified cost

- Specify sample size, compute error
- Specify error, compute sample size

Simplified selection: systematic sampling

- Compute sampling interval: $I = N/n$
- Select random start: $1 \leq S \leq I$
- First selection = S
- Subsequent selections: $S+I$, $S+2I$, $S+3I$,……

Sampling to support analysis objectives

- Key objective:

  compare race/ethnic groups

  white non-Hispanic

  black non-Hispanic

  Hispanic

- Optimal design: equal numbers of each group
- Sample at different rates (sampling fractions), e.g.

white non-Hispanic	1/4
black non-Hispanic	1/2
Hispanic	take all
Others	take none

Accommodate data collection mode

- Telephone data collection issues
  - Unlisted numbers
  - Multiple phone lines
  - Landline and cell
  - Cell only
  - No phone
- Some sample design alternatives
  - RDD
  - Weighting in estimation
  - Data collection rule
  - Dual frame
  - Using interruptions in phone service as surrogate

Sampling to reduce costs

- Two-stage samples
  - Clusters
  - Individuals with in clusters

  Population: high school students

  1. Probability sample of schools
  2. Probability sample of classrooms
  3. Select all students within a sampled class
  4. Self-administered questionnaire

Sampling to increase precision

- Increase sample size
- Stratification
  - Proportional
  - Disproportional
    - Costs or population variances differ by strata
    - Strata are of direct interest

Sampling a rare population

- Some demographic groups
  - Hispanics age 65 or older
  - Blacks with household incomes > $150,000
- Groups defined by multiple factors
  - Asian MSM's residing in New York city
  - Females who are recent immigrants from Middle East
- Groups defined by relatively rare experiences or characteristics
  - MSM's
  - Gulf war veterans
  - Persons with AIDS
  - Patrons of sex clubs

Sampling to survey a rare population

- Multiple frames
  - Special frame to improve yield
  - General population frame to ensure coverage
- Stratification
  - Oversample known neighborhoods
  - Some sampling in remainder of geographic area
- Network sampling
  - Link population elements for reporting purposes
  - Specify reporting rules so that inclusion probabilities are known

Multiple frames

- Population: Patrons of sex clubs in a designated city
  - Club membership lists
  - Time-location sampling at clubs
  - General population screening in the city

Network sampling

- In most household sample surveys, respondents report only about themselves or their households
  - One-to-one reporting rule
  - Screening yield= eligible households / n

- In network sampling, households are permitted report about themselves and other households to which they are linked
  - One-to-many reporting rule
  - Screening yield=(eligible households + networked households)/n
  - Cost savings to identify E eligible households

  Network sampling can be used to:

  1. Count eligible households [estimate population size]
  2. Collect further information about eligible households

Network sampling

Population: Gulf War veterans

Mode: telephone

- Select a general population sample of households
- Screen each selected household for target population
- Ask each household about adult sibling households in the geographic area: presence of target population member- Yes/No
  - If Yes
    - Request contact information for sibling household
    - or simply ask about sibling household
    - Determine how many other households *could have* reported the identified sibling household

Disproportionate stratified sampling

- Divide the population into strata with differing prevalence rates for the rare population
- Sample strata with higher prevalence at higher sampling rates
- Appreciable gains in precision are possible if:
  - Some strata have substantially higher prevalence rates
  - The strata contain a large proportion of the rare population
  - The gain is also dependent on the cost of a full interview relative to a screening interview

Telephone survey of MSM's in selected cities

- Stratified telephone exchanges into groups based on prevalence estimates
  - Estimates based on multiple sources
  - Expected error in the prevalence estimates
- Set sample allocations to strata
  - Based primarily on expected cost per identified eligible during screening
  - Released sample in random replicates
  - Adjusted sample allocations based on actual costs per screened case

Sample weights: an overview

- Accounting for the sample survey design
  - Different probabilities of selection
  - Under-coverage
- Accounting for survey sample implementation
  - Nonresponse
  - Adjusting to known population distributions

Weighting to adjust for selection bias

- Telephone survey of adults 18-65 in households in NY
  - RDD covers all landlines
  - Cell frame
    - Representations in frame(s)
      - Households group A: 1 phone number
      - Households group B: 2 phone numbers
      - Households group C: 3 phone numbers
      - Households group D: 4+ phone numbers

Affect of frame multiplicity on selection: $P_{H_i} \propto \sum L_i$

Weight for frame multiplicity: $W_{H_i} = 1 \Big/ \sum L_{H_i}$

The decision to participate and nonresponse bias (2)

Are refusals for some reasons of more concern than others?

Reasons for cooperation

- Interviewer's manner
- Sponsor
- Available time
- **Topic**

The logic of weighting for nonresponse

- Groups over-represented in the sample
  - Persons 60 and older
  - Households with children
  - Persons with college or more
  - Whites
- Groups under-represented in the sample
  - Persons under age 30
  - Households without children
  - Persons with less than high school
  - Hispanics and blacks

- We want the sample to reflect the population demographics
- We have information to suggest some of these groups may differ on the measures

- Create weighting classes [cells] to which weights are applied
- Effectiveness depends on
  - The respondents in each weighting class being a random sample of all respondents
  - Reduction in bias exceeds increase in sampling variance

Good sampling practice: a summary

- Define population
- Maximize coverage
- Probability sampling
- Know when and how to weight
- Response rate: don’t let it become a sample of volunteers

Models in survey sampling inference: the example of Respondent Driven Sampling (RDS)

- Network sampling uses respondent reports of network size to establish known (though not error-free) probabilities of selection.
- Network samples can produce estimates of population parameters much like those from any probability sample

- Respondent Driven sampling also uses reports to identify members of a target population
- RDS estimates are based on a mathematical model, which relies on a set of assumptions:
  - Degree
  - Recruitment at random
  - Reciprocity
  - Convergence

Questions about model-based sampling inference

- What do we know about the assumptions the model depends on?
  - Are key assumptions strong or weak?
  - To what extent do the assumptions hold in a particular instance?
  - How robust is the model?
  - What protection do you have if they don't?
- Can you adjust for (or measure) the impact on your results?
- What sources of nonsampling error may affect estimates?

Survey Sampling and Methodology References

- Basic
  - Kalton, G. : Introduction to Sampling, Sage, 1983
  - Groves et al.: Survey Methodology, Wiley, 2004
- Advanced
  - Kish, L.: Survey Sampling, Wiley, 1965
- Specialized
  - Binson et al. "*Sampling in Surveys of Lesbian, Gay, and Bisexual People",* 2007