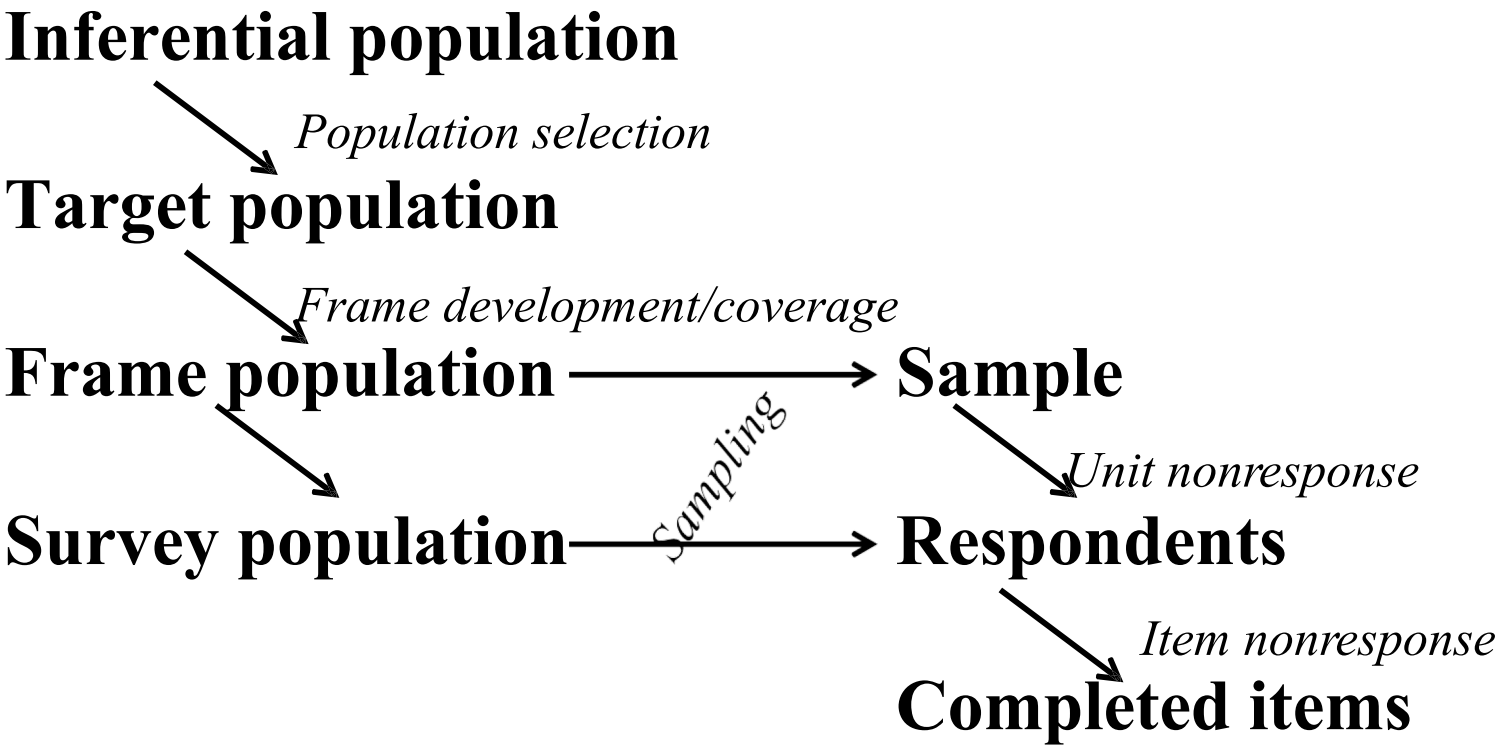
Statistical Issues in Design and Analysis of Surveys

Alan M. Zaslavsky
Department of Health Care Policy
Harvard Medical School

Three designs for research

- Experiments
  - Treatment is applied under control of experimenter, using randomization
- Surveys
  - Emphasis on ***representation*** of a population, through random ***sampling***
- Observational studies (quasi-experiments)
  - Analyze survey/administrative data like an experiment

Populations and samples
(and what can get lost along the way)



Frames

- Explicit: units known in advance
  - List (list of units)
  - Area (geographical division)
- Implicit: sampling process defines the units
  - RDD (Random Digit Dialing)
  - Sampling from patient stream

Sample design

- Key to statistics of survey design
- Representative of population
  - Sampling procedures
  - Actual and potential achieved samples
- Randomization (of sample selection) as the basis for statistical inference
  - "Design-based inference"
- Design to be efficient for questions of interest
  - Greatest precision within accepable cost

Sample designs

- Well-defined probability rules
  - Know "enough" about probability of each possible sample.
  - P(inclusion)>0 for each element in frame
- Motivated by survey objectives
  - Tradeoffs of cost versus precision
- Appropriate to structure of the population

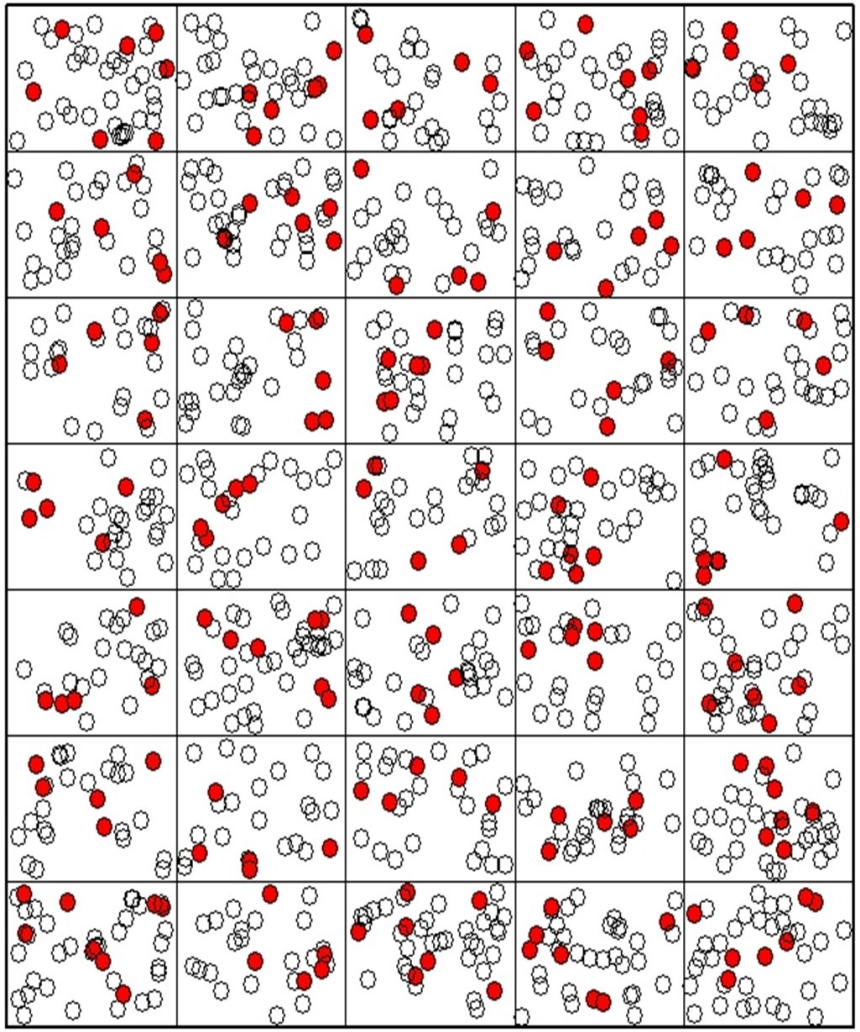
Sample designs: simple alternatives

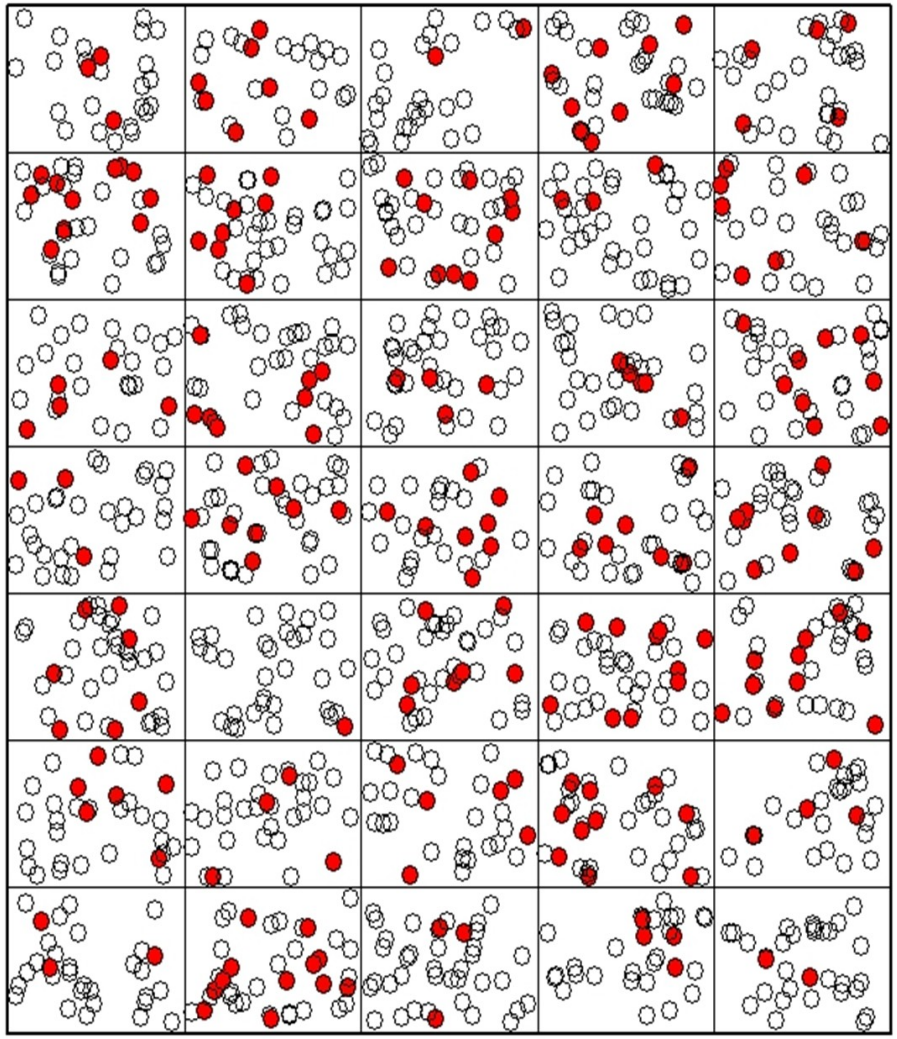
- Simple random sampling
  - Fixed sample size
  - Every such sample has equal probability
  - Implies equal P(inclusion) for each element
- Systematic sampling
  - Select units at fixed interval from random start
  - Usually roughly equivalent to SRS

Sample design: stratified sampling

- Divide population into parts (strata)
- Independent sampling in each stratum
  - Often assumed (by default) to be SRS
- Efficiently allocate sample to strata
  - Proportional allocation: sample matches population distribution
  - Vary sampling rates (disproportionate)
    - E.g. case-control design
  - Different designs/methods in different strata?

Stratified & simple random sampling

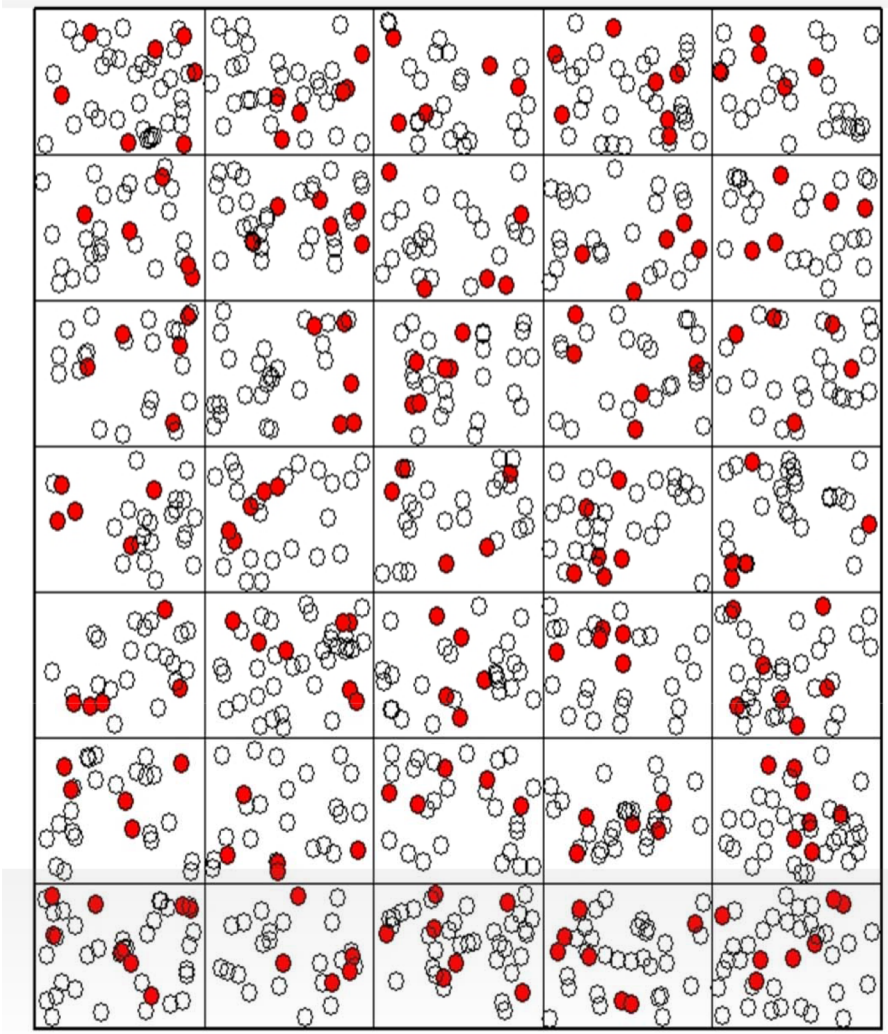


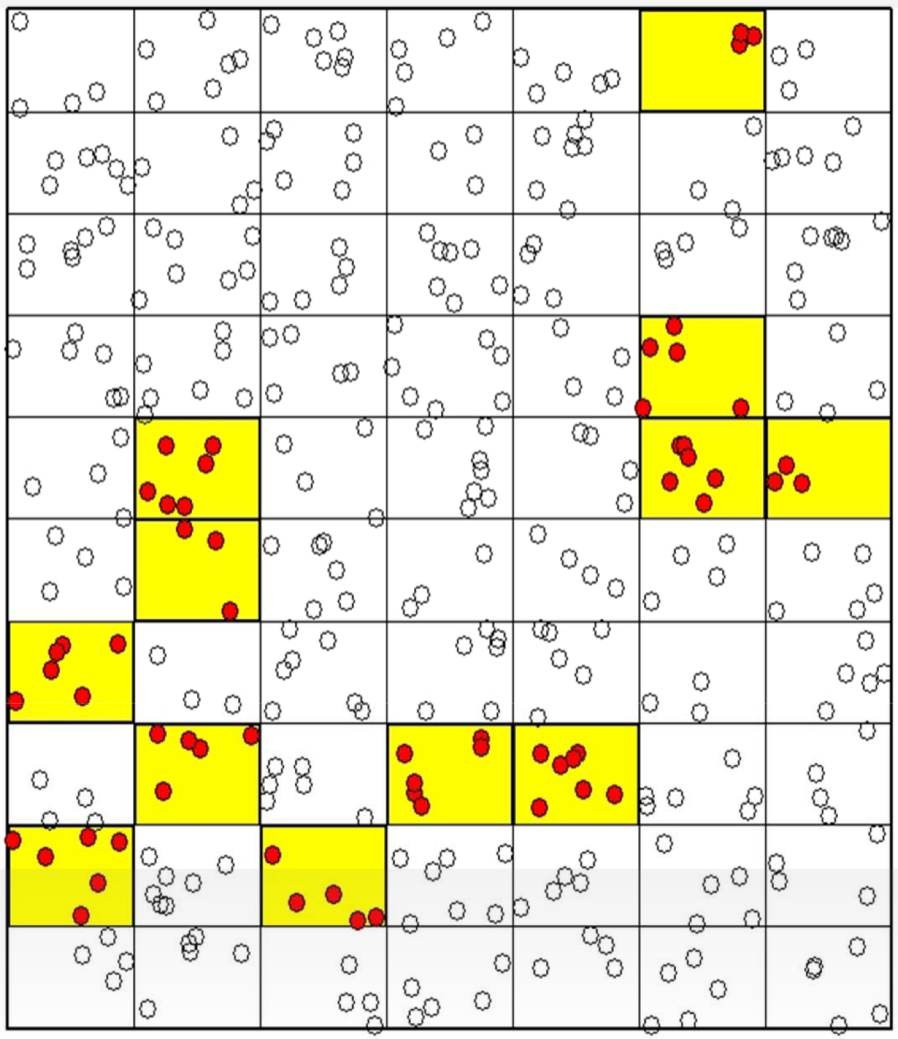


Sample design: Cluster sampling

- Elements are sampled in groups rather than separately
  - Sample city blocks within a community
  - Sample all patients of selected doctors
- Due to practical or cost constraints
  - Reflects structure of the population
- Tends to ***increase*** variability of estimates
  - Relative to SRS

Stratified & Cluster sampling





Clustering scenario

- Purpose of survey: estimate mean $/patient visit
  - Mixed population of pediatricians, neurosurgeons
- Two designs:
  - Researcher A: choose 9 patients independently
  - Researcher B: choose 3 doctors, 3 patients/doc
- Which design will show greater variation?
  - If each of the researchers replicates design a week later, whose results will be more consistent?

Simple random sample (SRS)

NS1	Ped1	Ped2	Ped3	NS4
$6000	*$200*	*$300*	*$600*	**$3000**
$5000	**$100**	**$500**	*$400*	*$3000*
$4000	*$400*	*$200*	*$200*	*$5000*

Ped4	NS2	**NS3**	Ped5	Ped6
$100	*$9000*	**$3000**	*$700*	**$300**
$100	*$5000*	*$4000*	**$200**	**$200**
$300	*$8000*	**$4000**	*$400*	*$100*

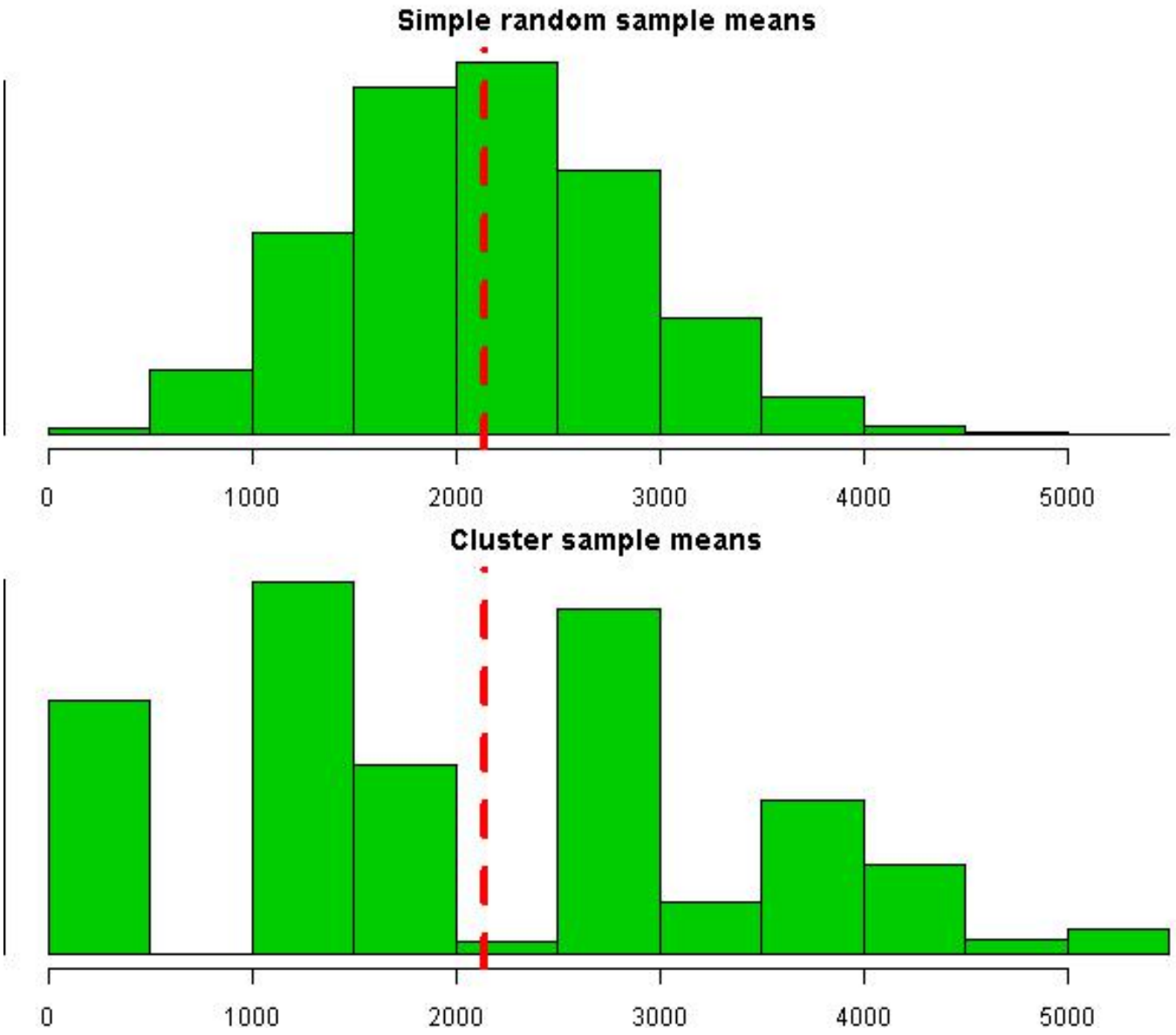
Estimated mean = $1233

Cluster sample

NS1	Ped1	Ped2	Ped3	**NS4**
$6000	**$200**	*$300*	*$600*	**$3000**
$5000	**$100**	*$500*	*$400*	**$3000**
$4000	**$400**	*$200*	*$200*	**$5000**

Ped4	**NS2**	NS3	Ped5	Ped6
$100	**$9000**	*$3000*	*$700*	*$300*
$100	**$5000**	*$4000*	*$200*	*$200*
$300	**$8000**	*$4000*	*$400*	*$100*

Estimated mean = $3744



Distributions of SRS and cluster sample means (10,000 randomly simulated draws for each)

Red line represents population mean (parameter).

More sample design options

- Multistage sampling
  - 2 or more stages of sampling
  - Sample clusters (primary sampling units = PSU)
  - … and then sample within clusters (secondary sampling units = SSU)
- Probability proportional to size (PPS)
  - unequal probability of selection for each PSU
  - typically, larger units more likely to be selected
  - Like continuously disproportiona stratification
- Double (two-phase) sampling

(Classical) Survey analysis

- "Design-based" approach to inference
  - Variation is represented by randomness of sampling, not randomness of population
- Nonparametric approach to inference
  - Targets are parameters of finite population
- Emphasis on robustness, objectivity
  - Minimal reliance on model specification
  - Unbiased (approximately) estimation

    [BUT: This is all changing]

Practical implications of complex sample designs

- Estimators for i.i.d. designs often will not yield consistent estimates of parameters
  - Disproportionate sample has different distribution than population
- Initial focus on estimation of totals
  - Simplest estimation problem
  - Most parameters are directly or indirectly functions of totals

Unbiased point estimation of totals

- Suppose P(inclusion of unit *k*)=π_k
  - Determined by sample design
  - $E(I_k) = \pi_k$ where $I_k$is 0/1 inclusion indicator
- Expectation of <u>weighted</u> sample total of variable *y* is $\mathrm{E}\left[\sum w_k I_k y_k\right] = \sum w_k \pi_k y_k$
- If $w_k = \frac{1}{\pi_k}$ this reduces to $\sum y_k$
  - So “inverse probability of selection” weighted estimator is unbiased for population total

Weighting scenario

- Population = 500 children, 4500 adults
- Sample = 50 children (asthma rate = 10/50=20%)
  + 50 adults (asthma rate = 2/50 = 4%)
  - Weights = 1/(50/500)=500/50=10 for children
  - Weights = 1/(50/4500)=4500/50=90 for adults
- Unweighted rate= (10+2)/(50+50)=12%
- Weighted rate = (10×10+90×2)/(10×50+90×50)
  = 280/5000 = 5.6%

Weighted point estimates

- Replace totals in population parameter formula with weighted totals from sample
  - Known as Π estimator, expansion estimator, Horvitz-Thompson estimator
  - Requires only one variable added to dataset
- Weights might incorporate adjustments of various kinds
  - Nonresponse weights
  - Poststratification or calibration weights
    - Bring the weighted sample closer to known population values

Variance estimation

- Requires formulas specific to design
  - Not just marginal inclusion probabilities
- Simple random sampling variance estimator:

$$\text{Var } \bar{y} = \frac{(1-f)S^2}{n}$$

$f` =$ sampling rate (for finite population inference)

$S^2 =$ individual level variance of *y*

n= sample size

Optimal design

- Use prior information to make designs efficient for desired estimands
- Example: allocation of sample across strata for estimation of population mean or total
  - Optimal sample proportional to
    $$N_h S_h / \sqrt{\{C_h\}}$$
  - where N_h =stratum population,
  - S_h = standard deviation in stratum,
  - C_h = unit cost in stratum,

Practical analysis of complex surveys

- Specify survey design information
  - Weights
  - Clusters (PSU)
  - Strata
- Special analytical methods required
  - Nonsurvey weights have different interpretations
  - Survey weights with nonsurvey software typically gives:
    - correct point estimate (approximately)
    - invalid variance estimates and statistical tests

Software for analysis of complex surveys

- SAS: SURVEY* procedures(limited set)
- Stata: svy--- (large list of procedures – econometrics focus)
- R (survey package)
- SUDAAN: greatest flexibility in specification of design, many procedures
- SPSS
- WesVar, EpiInfo, etc.

SAS template

- PROC SURVEYMEANS < options > < statistic-keywords > ; BY variables ;
- CLASS variables ;
  CLUSTER variables ;
- STRATA variables < / option > ;
  VAR variables ;
- RATIO variables ;
  DOMAIN variables ;
  WEIGHT variable ;

SUDAAN template

- PROC LOGISTIC DATA=file DESIGN=designtype
  •<options> <keywords>;
- WEIGHT variables;
  NEST variables;
  SUBGROUP variables;
  LEVELS levels;
  SUBPOPN expression;
- REFLEVEL variables = reference levels;
  MODEL dependent var = independent vars;

R design specification

- svydesign(ids,
- probs=...,
- strata = ...,
- variables = ...,
- fpc=...,
- data = ...,
- nest = ...,
- weights=...,
- pps=FALSE)

Stata design specification parameters

- **`strata(varname)`**
  - Stratum identifier variable.
- **`psu(varname)`**
  - PSU (cluster) identifier variable.
- **`fpc(varname)`**
  - Finite population correction variable.
- **`[pweight=varname]`**
  - Sampling weight variable.p

Analysis of complex surveys

- Most difficult with combination of complex sample design, complex estimands
  - e.g. coefficients of complex models
  - Compromises if required combination not handled by software
  - Resampling methods may help
- Need to puzzle out design features
  - Cryptic or cursory description in documentation
  - Some details might be confidential! (real PSUs)
    - Design effects; Pseudo-PSUs and Pseudo-strata

Survey stats: Analysis concepts

- Methods with missing data (weighting and imputation)
- Privacy and confidentiality issues
- Calibration/adjustment/postratification
  - Incorporate information from nonsurvey sources
  - Reduce variance, improve consistency
- Small-area estimation
  - Develop estimates for areas where survey alone has too little sample for estimation
  - Integrate regression predictions, survey estimates
- Multilevel models
  - When “clusters” are meaningful units
- Systematic approaches for sampling and nonsampling error
- New methods for variance estimation
  - Resampling: bootstrap, jackknife, etc.

Implementation innovations

- Integration with new data sources
  - Administrative records, “big data”
  - Apply to missing data imputation & weighting, calibration, estimation/correction of survey bias
  - Potential cost saving and reduced burden
- Methods for rare or elusive populations
- Multi-modal response
  - Integration of internet response
- Opinion survey panels

Conclusions

- Use appropriate statistical tools
  - Efficient design
  - Analysis methods
- Many opportunities for new methods and applications