EMBLEM
High Quality, High Impact Research

Cancer Research Methods Workshop

Mayur M. Desai, PhD, MPH

Mwanza, Tanzania

September 18-19, 2012

Yale SCHOOL OF PUBLIC HEALTH

Sampling Methods

Study sample

- Need to specify inclusion/exclusion criteria
- Depending on study, considerations could include:
  - individual characteristics (demographics, employment history, etc.)
  - clinical status
  - accessibility
  - study logistics
  - ethics
- Generalizability? → internal vs. external validity

Sampling

- Census study = if the entire population of interest is included in the study
- Most studies require sampling because collecting data on an entire population is generally not feasible
- Collect information on a subset that represents (or is similar to) the entire population

Sampling basics

- Target population or universe
  - the group one wishes to generalize to or make inferences about
- Sampling frame
  - the list of the target population from which the sample is drawn
  - operational definition of who, where, and when
- Sampling unit or element
  - the unit of observation for data collection

Advantages of sampling

- Time efficient
- Less costly
- Potentially more accurate (since it is more feasible to maintain quality control over a smaller number of subjects)

Disadvantages of sampling

- Potential bias in the selection of subjects, which may lead to error in interpretation of results and decrease in ability to generalize the results beyond the subjects actually studied

Assessing generalizability

- Is the sample drawn from the target population?
- Is the sampling scheme biased?
- Can we demonstrate that enrolled subjects are reasonably comparable to the population of interest with respect to important demographic/clinical characteristics?

2 types of sample designs

- Probability sample
  - each member of the population has a known, non-zero probability of being selected
  - the probability is not necessarily the same for all members
- Nonprobability sample
  - members are selected from the population in some nonrandom manner

Probability sample designs

- Simple random sample
- Systematic random sample
- Stratified sample
- Cluster sample

Simple random sample

- Purest form of probability sampling
- Every element in the population has an equal chance of being included in the sample (e.g., drawing numbers from a hat)

Systematic random sample

- Researcher selects a random starting point and then systematically selects subjects from the sampling frame at a specified interval (e.g., every *k*th unit is selected)
- Starting point and sampling interval (*k*) are based on the required sample size
  - e.g.: if sampling frame N=1,000 and sample size n=100, then k = 1,000/100 = 10

Stratified sample

- Used when there is a particular interest in making sure that:
  - certain groups will be included in the study, or
  - some groups will be sampled at a higher rate than others, or
  - there is geographic representativeness
- The population is divided into strata, or groups of units having certain characteristic(s) in common, and then a sample of units is drawn from each stratum (stratum-specific sampling may be proportionate or disproportionate)

Proportionate vs. disproportionate

- Proportionate
  - same sampling fraction in each stratum
- Disproportionate
  - different sampling fraction in each stratum

Cluster sample

- Primary purpose is to maximize the dispersion of the sample throughout the community in order to represent the diversity that exists while also minimizing costs (money and time)
- Geographic area is divided into “clusters”
- “Clusters” are first sampled from all of the clusters in a defined area (often using probability proportionate to size [PPS] design) and then data are collected on all or a subsample of units in each sampled cluster

Multistage probability sample

- A combination of these sampling methods can be used
  - for example, if conducting a large-scale community-based survey

Analytic issues

- Need to consider weighting and variance estimation issues when analyzing complex sample data (unequal probability of sample selection and clustering)
- Sample weight reflects the inverse of the selection probability – can be interpreted as the number of population members a given subject represents

Nonprobability sample designs

- Convenience sample
- Consecutive sample
- Quota sample
- Snowball sample

Convenience sample

- Subjects are selected because of their convenient accessibility to the investigator
- Subjects are chosen simply because they are the easiest to obtain for the study (easy, fast, and usually the least expensive and troublesome)
- Investigator makes no attempt (or only a limited attempt) to insure that the sample is an accurate representation of the population of interest
- Greatest potential for bias

Consecutive sample

- Strict version of convenience sampling – where every available subject is selected, i.e., the complete accessible population is studied
- This is the best choice of the nonprobability sampling techniques, since by studying everybody available, a good representation of the overall population is possible in a reasonable period of time

Quota sample

- Nonprobability equivalent of stratified sampling
- The researcher first identifies the strata of interest and then more targeted sampling is used to select the required number of subjects from each stratum

Snowball sample

- Used when the desired sample is hard to reach or the sample characteristic is rare
- Relies on referrals from initial subjects to generate additional subjects (also called chain-referral sampling)
- Reduced likelihood that the sample will represent the population of interest