PLS 506
Mark T. Imperial, Ph.D.

Lecture Notes: Sampling

- Sampling
  - What is the rationale behind sampling
    - We seek knowledge about a whole class of similar objects or events (a population)
    - We cannot observe the whole population so we observe some of them (sample)
    - We then extend these findings to the whole population
  - Defining the *population* is a multi-step process
    - Clearly identify the target population.
    - Geographic and time boundaries
    - If the target is a group or organization, then the type and size must be determined
    - Then have to construct the *sampling frame* or the set of all cases from which the sample is actually selected. It is not a sample but rather is the operational definition of the population that provides the basis for sampling.
      - Denotes all of the cases from which to sample. You can either list all of the cases or provide a rule for defining membership
  - Sampling designs:
    - Ideally we would like to obtain a sample that would be representative of the target population.
      - To be *representative* means to provide a close approximation of certain characteristics of target groups
      - It is extremely unlikely that one will draw a perfectly representative sample
      - It is rarely possible to evaluate a specific sample in terms of its representativeness because the populations we study are not known in all respects
    - The quality of a sample must be judged in terms of the procedures that produced it.
    - The *sampling plan* refers to the part of the research plan that indicates how cases will be selected for observation
    - Sample size depends on
      - Heterogeneity of the population
      - Desired precision in terms of the confidence level (significance level) and confidence interval (margin of error)
      - Size of the population (to some extent, less important as population begins to exceed 100,000)
      - Type of sampling design
      - Number of breakdowns in the data analysis (need to make sure you have enough cases in each cell to run certain statistics)
      - Available resource

- *Probability sampling*
  - Scientifically more acceptable but not always feasible or economical
  - Probability sampling always involves the process of *random selection* at some point.
  - Random refers to a process which gives each case in the population an equal chance of being included in the sample.
  - If the selection process favors certain cases it then increases the likelihood that certain cases will be sampled, which can then bias the results.
  - Random involves being more than simply picking cases haphazardly. It requires mechanical or electronic aids to ensure that chance alone dictates selection
  - *Simple random sampling*: every combination of cases has an equal chance of being included in the sample. For this you need a complete list of the population and the random sample of the cases.
  - *Sampling error* is the amount a given statistic deviates from the population parameter it estimates.
    - Significance (confidence) level reflects the amount of evidence you want to ensure that you are correct with your conclusions. Social science typically uses a confidence level of .05 so that you are 95 percent certain of your results.
    - Confidence interval (or margin of error) is the range within which a population parameter is expected to lie with a specified level of confidence
    - Sample size is a function of the desired confidence level and confidence interval and to a lesser extent the size of the population examined. Once the population gets to be more than about 100,000, population size has little effect on the sample size.
  - *Stratified random sampling*:
    - Population is first divided into two or more mutually exclusive segments called strata based on the categories of relevant variables
    - Simple random samples are then drawn from each stratum
  - Cluster sampling
    - The population is broken into groups of cases called clusters and then a sample of clusters is selected at random.
    - The clusters generally consist of natural groupings like dormitories, counties, cities, census tracts, etc.
  - Systematic sampling
    - Start with a randomly selected case and the select every Kth case after that.
    - Sampling interval (K) is merely the ration of the number of cases in the population to the desired sample size
    - While its advantage is simplicity it can produce biased samples

- *Nonprobability sampling*
  - Case selection involves methods other than random selection
  - Without random selection there are several weaknesses
    - Cannot control for the investigator's bias in selecting units
    - Pattern of variability cannot be predicted from probability sampling theory making it impossible to calculate the sampling error or sample precision
  - *Convenience sampling*
    - Research simply selects the requisite number of cases that are conveniently available
    - Case selection is easy, quick, and inexpensive
  - *Purposive sampling*
    - Investigator relies on expert judgment to select units that are representative or typical of the population
    - Exit polls use specific precincts and geographic areas that are reliable predictors of overall returns
    - Unacceptable when precise and accurate generalizations are required
  - *Quota sampling*
    - Divides the population into relevant strata as in stratified sampling and the fractions within each stratum are estimated via external data like the census.
    - The total sample is then allocated among the strata in direct proportion to their estimated or actual size in the population.
    - Interviewers then speak to a fixed quota of respondents in each strata but to fill requirements they are free to choose anyone who meets quota requirements.
    - Interviewers free to self select respondents are subject to several kinds of bias. They will interview friends or avoid people that make them uncomfortable (a shy guy may not approach a pretty girl). They may also focus on heavily traveled areas like a shopping center to gather their quota quickly.
  - *Referral* or *snowball* sampling
    - Ask respondents to help identify the population. These are used when the population is unknown or contact information is missing.
    - In a snowball sample, you ask each respondent to identify others to interview and keep going until no other respondents are identified.
    - Snowball samples are common in network studies
  - Other creative ways to combine these random and nonprobability based designs