Case searching, coding and sampling

Su Li
University of California, Berkeley
School of Law
11/13/2014
• Where to find my data?
• How large my data should be
  --- data size for sufficient power analysis
  --- data size for representative sampling
Case searching and coding

Search for cases:
• Westlaw; Lexisnexis (keyword search)
• Case database:
  - Bureau of Justice Statistics (e.g. civil justice survey of state courts)
  - other organizations (e.g. Supreme Court Database; Lex Machina)
• Visit a local court or befriend with a public defender

Code cases:
• Pre-coded variables
• Code your own variables (establish your own codebook; create a spreadsheet of numbers, each number represents a value of a variable)
• How large my data should be
--- data size for sufficient power analysis
--- data size for representative sampling
Power analysis

• Power: the probability of detecting an effect, given that the effect is really there. (.8 level, .9 level)

• Why do power analysis: help determine sample size; use in grant proposal

• Limitations of power analysis: does not generalize well; may not suggest enough sample size; based on lots assumptions (best case scenario)
Sample Size and Quantitative Research

• Population vs. Sample

• Research Question: Do county demographic characteristics exhibit a statistical association with case-level outcomes of jury-tried tort cases?

To Answer the Example Research Question

• Dependent Variable:
  Plaintiff’s success in establishing defendant’s liability and damage award
• Independent variables:
  County level poverty rate, minority population composition
• Data to be collected: trial court cases and the county level demographic information
• A subset of jury-tried tort cases drawn from the 2001 Civil Justice Survey of State Courts
How many cases would be enough for this study?

• Non-statistical consideration
  - Availability of resources (Manpower; Budget etc)

• Statistical Consideration:
  - Level of precision
  - Confidence level
  - Degree of variability
  - Response rate
  - Types of test to be conducted
Strategies to determine sample size

- Using a census for smaller populations
- Using published tables
- Imitating a sample size of similar studies
- Applying formulas to calculate a sample size
General steps of determining sample size

• 1. Determine **Goals**
• 2. Determine desired **Precision of results**
• 3. Determine **Confidence level**
• 4. Estimate the degree of **Variability**
• 5. Estimate the **Response Rate**
• In addition: Plan tests to be conducted
Step 1: determine the goal

- What is the target population?
  - If small: use census
  - If not: sample

- Research design, concepts/attribute to measure

- Know your resources and limitation
Step 2: determine the level of precision (desired precision of results)

- “Sampling error“ or “margin of error” is the range in which the true value of the population is estimated to be
- E.g. The supporting rate of a presidential candidate is 60% with a margin of error 5%, i.e. the supporting rate in the POPULATION is between 55% and 65%
- This range is usually at 95% confidence level
- See handout for quick sample size table
Step 3: Determine Confidence level

• (Risk level) 90%, 95%, 99%

• According to central limit theorem, if we repeated sample from a population, the mean of the samples equals to the actual mean of the population.

• 95% confidence levels means 95 out of 100 samples of the population will have the true population value within range of precision.

• See handout for quick sample size table
Step 4: Estimate Degrees of Variability

- Depending upon the target population and attributes under consideration, the degree of variability varies considerably. The more heterogeneous a population is, the larger the sample size is required to get an optimum level of precision.

- Estimate variability--take a reasonable guess of the size of the smaller attribute or concept you’re trying to measure, rounding up if necessary.

- See handout for quick sample size table (get base sample size using level of precision and estimated variability)
Formulas for calculating sample sizes

• $n$ is sample size, $N$ is population size, $e$ is the level of precision

\[
n = \frac{N}{1 + N(e)^2}
\]

• If $N$ is small, $n$ can be adjusted to be even smaller

\[
n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}
\]
Step 5: Estimate response rate

- The base sample size is the number of responses you must get back when you conduct your survey.
- However, since not everyone will respond, you will need to increase your sample size, and perhaps the number of contacts you attempt to account for these non-responses.
- Final sample size = divide the base sample size by the percentage of response (estimated)
Non-respondent vs. respondent

• Research show the difference could be significant
• Additional sampling on non-respondents to check the difference between the two groups is useful.
• Non-respondents could mean the trial cases that are not included in searchable databases.
Final sample size

• n: sample size; N: population size; p: estimated variance; e: precision desired (3% or 5% or 7% etc)
• z: z score for confidence level (1.96 for 95% confidence, 1.64 for 90% confidence, 2.58 for 99% confidence)
• R=estimated response rate.

\[ n = \frac{p(1-p)}{z^2 + \frac{p(1-p)}{N}} \]
In addition: types of tests

• Univariate analysis (what we have been talking about so far on sampling) vs. bivariate analysis (double the size of the sample to ensure each group has reasonably big sample size)

• Analysis that involves multiple variables such as regression analysis (how many independent variables/covariates affects sample sizes)
Sampling Strategy

• Simple Random Sampling
• Stratified Random Sampling
• Clustered sampling
• Oversampling and weight
Further reading and references


• Israel, Glenn D. Determining Sample Size. University of Florida IFAS Extension