### Sampling: a way of studying a subset of the population but still ensuring “generalizability” (vs. census – study of entire population) – does the study have external validity?

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Example</th>
<th>Another Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit of Observation</td>
<td>Level of social life being studied – individuals or groups of individuals</td>
<td>Child (DV=days playing outside per week, IV=living on cul-de-sac or not)</td>
<td>Neighborhood (DV=number of kids playing outside, IV=share of streets that are cul-de-sacs)</td>
</tr>
<tr>
<td>Population (Target population)</td>
<td>The set of individuals or other elements to which study findings will be generalized; sample is drawn from the population</td>
<td>Sacramento area children, e.g. Charlie, Lucy, Linus, Patty, Violet, etc.</td>
<td>Neighborhoods in the Sacramento region, e.g. Midtown, Natomas, Land Park, Laguna West</td>
</tr>
<tr>
<td>Sampling Frame</td>
<td>List of all individuals or other elements; used for drawing sample</td>
<td>Public school rolls Phone listings Marketing list of households with children</td>
<td>List of neighborhoods List of cities in Sacramento region</td>
</tr>
<tr>
<td>Enumeration units: Units that contain one or more elements; useful if sampling frame not available for elements themselves</td>
<td>Households (contain children)</td>
<td>Cities (contain neighborhoods)</td>
<td></td>
</tr>
<tr>
<td>Sample</td>
<td>Individuals or elements selected for the sample</td>
<td>Charlie, Patty</td>
<td>Natomas, Laguna West</td>
</tr>
</tbody>
</table>

### Generalizability

<table>
<thead>
<tr>
<th>Generalizability Type</th>
<th>Definition</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Generalizability</td>
<td>From sample to population; depends on “sampling error”</td>
<td>Results for sample of children the same as for all children of Sacramento region?</td>
<td>Results for sample of neighborhoods in Sacramento region the same as for all neighborhoods in Sacramento region?</td>
</tr>
<tr>
<td>Cross-Population Generalizability</td>
<td>From one population to another</td>
<td>Results for individuals in Sacramento region the same as for other regions?</td>
<td>Results for neighborhoods in Sacramento region the same as for cities elsewhere in California? The U.S.?</td>
</tr>
</tbody>
</table>

**Sampling error:** Any difference between characteristics of a sample and the characteristics of population from which it was drawn:
- Random sampling error: inherent in process of sampling
- Systematic sampling error: minimizable with good sampling plan!
**Representative sample:** A sample that “looks like” the population from which it was selected in all respects that are potentially relevant to the study. How do you know you have one?

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
<th>Use</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability sampling, i.e. random</td>
<td>Every element in the population has a none non-zero probability of being selected; sampling involves random selection (equal chance)</td>
<td>Allows for use of inferential statistics; quantitative, hypothesis-testing studies</td>
<td>Systematic bias, non-respondents</td>
</tr>
<tr>
<td>Non-probability sampling, i.e. non-random</td>
<td>Do not know in advance how likely that any element of the population will be selected for the sample; non-random selection (not equal chance)</td>
<td>Allows for use of descriptive statistics only; qualitative, exploratory studies</td>
<td>Not representative, so can’t generalize</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability sampling</td>
<td>Simple random sampling</td>
<td>Sample chosen strictly by chance</td>
<td>Random digit dialing from phone listings for Sacramento region</td>
</tr>
<tr>
<td></td>
<td>Systematic random sampling</td>
<td>Select first element randomly, the select every nth element</td>
<td>Every 1000\textsuperscript{th} number in the phone listings</td>
</tr>
<tr>
<td></td>
<td>Stratified random sampling – proportionate</td>
<td>Population sorted into strata according to key characteristic; random sampling within strata; sample for each stratum in proportion to its size in population</td>
<td>Households split into those living on cul-de-sacs, those not; random sample drawn from each group; if 1 in 10 households live on cul-de-sac, then 1 cul-de-sac household for every 9 non-cul-de-sac households</td>
</tr>
<tr>
<td></td>
<td>Stratified random sampling – disproportionate or nonproportionate</td>
<td>Population sorted into strata according to key characteristic; random sampling within strata; sample for each stratum intentionally not in proportion to its size in population</td>
<td>Households split into those living on cul-de-sacs, those not; random sample drawn from each group; 500 households randomly selected in each group, even though only 1 in 10 household live on cul-de-sacs in region</td>
</tr>
<tr>
<td></td>
<td>Cluster sampling</td>
<td>Cluster is naturally occurring grouping of elements; draw a random sample of clusters, then draw random sample of elements within cluster</td>
<td>Neighborhoods randomly selected, households within these neighborhoods randomly selected. Often used when data collection done in person.</td>
</tr>
<tr>
<td></td>
<td>Matched-pairs sampling</td>
<td>Categorize sampling frame into two groups based on key characteristic. Take random sample of first group, then find elements in the second group that match those in the first.</td>
<td>Households split into those living on cul-de-sacs, those not; random sample drawn from cul-de-sacs group; researchers finds households in non-cul-de-sacs group that match. See methods for controlling for spuriousness.</td>
</tr>
<tr>
<td>Type</td>
<td>Method</td>
<td>Definition</td>
<td>Example</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Non-probability sampling</td>
<td>Availability sampling</td>
<td>Elements selected because they’re easy to find</td>
<td>Friends and co-workers of the research team who have children used as sample; OR families at the Farmer’s Market asked to participate</td>
</tr>
<tr>
<td></td>
<td>Quota sampling (proportional or non-proportional)</td>
<td>Quota set to ensure that sample represents certain characteristics in proportion to their prevalence in population</td>
<td>Families at the Farmer’s Market asked to participate until 10 families on cul-de-sacs and 90 not on cul-de-sacs recruited</td>
</tr>
<tr>
<td></td>
<td>Purposive sampling</td>
<td>Sample includes individuals with particular (rare) traits targeted by the researcher</td>
<td>Households with 2 or more kids and no cars are targeted to understand how they meet their mobility needs</td>
</tr>
<tr>
<td></td>
<td>Expert sampling</td>
<td>Sample includes individuals particularly knowledgeable about issues under study – “key informants”</td>
<td>20 public health officials who work on programs to get children to be more physically active</td>
</tr>
<tr>
<td></td>
<td>Snowball sampling</td>
<td>Start with initial sample, ask these individuals to recommend other individuals, and so on; for hard-to-reach populations</td>
<td>Find a few families with children that live on cul-de-sacs; ask them if they have neighbors with children or other friends who live on cul-de-sacs, and so on</td>
</tr>
</tbody>
</table>

**Inferential Statistics Basics for Probability Sampling:**

- Sample statistic: statistic (e.g. mean) computed from sample data
- Population parameter: true value for statistic (e.g. mean) for population (we don’t know this)
- Sampling error: population parameter – sample statistic (we don’t know this)
- Confidence interval: interval in which we can be confident that true value lies, based on sample statistic and its standard error (95% confidence interval is statistic +/- 2 standard errors)

**General rules:**

- Larger sample means more confidence in representativeness, less sampling error, narrower confidence intervals
- More homogeneous population means more confidence in representativeness, less sampling error, narrower confidence intervals
- What matters is sample size, not share of population sampled!

**Other things to think about in sampling...**

<table>
<thead>
<tr>
<th>Goal</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensuring that the independent variable varies</td>
<td>Sample includes children who live on cul-de-sacs and children who don’t</td>
<td>Sample includes neighborhoods that have lots of cul-de-sacs and neighborhoods that have few cul-de-sacs</td>
</tr>
<tr>
<td>Ensuring that the control variables don’t vary</td>
<td>Sample includes only children from moderate-income households</td>
<td>Sample includes only neighborhoods with average income in the moderate range</td>
</tr>
</tbody>
</table>
Prof. Handy’s Cul-de-Sac Sampling Plan:

The household survey will provide the primary data for testing the hypothesis. We expect the design of the sampling plan for the survey and the design of the survey instrument itself to be particularly challenging for this study. The target population for this study is children living in houses located on cul-de-sacs and through streets in the Sacramento region. Because no sampling frame exists for this population, we will use a multi-stage cluster sampling strategy. First, residential neighborhoods throughout the region will be defined based on major roadways and other geographic features. Census data will be used to eliminate neighborhoods built before 1950 because of the infrequent use of cul-de-sacs in residential developments before this time. From the remaining post-1950 neighborhoods, a random sample of neighborhoods will be selected. Within these neighborhoods, cul-de-sacs will be identified using the 2000 Census street network and the capabilities of geographic information systems (GIS). A random sample of the cul-de-sacs within each neighborhood will be chosen. For each cul-de-sac in the sample, a segment of a nearby through street (defined as a street that links arterial streets and carries significant levels of through traffic) of otherwise similar characteristics and similar length will be chosen. This approach creates matched pairs of streets. Next, all addresses on the sample street pairs will be compiled to create a sample of households. The sample of streets will be used in the field observations, the sample of households will be used in the household survey, and a sub-sample of the sample of households will be used in the in-depth interviews, as described below.