Unit 1: Research Design, Sampling & Measurement

iRAT

☐ You have 10 minutes to do it
☐ Mark your answer on both answer sheet and question sheet
☐ Turn in the answer sheet

Team RAT (tRAT): 15 minutes

If you don’t find the correct answer on the first try, keep trying to earn partial credit. Here’s the scale:

- Right answer on 1st try = 10 points
- Right answer on 2nd try = 5 points
- Right answer on 3rd try = 3 points
- Right answer on 4th or 5th try = 0 points

- Record team score

Missed questions?

Does any team want to appeal?

Concepts

☐ Population
  - The total set of subjects of interest in a study
  - Infinite or finite (N) number of subjects

☐ Sample
  - The subset of the population

“Subjects” (elements): people, families, schools, cities, neighborhoods, regions, companies, cars, trees, birds...

Concepts

☐ Population
☐ Sample
☐ Parameters
  - Numerical summary of the population

☐ Statistics
  - Numerical summary of the sample
Team Activity #1: Identify population and sample

☐ Take 3 minutes

☐ Using census data for NYC, study if there is a spatial assimilation among the second generation of immigrants. In this study, the “population” is

A) All people in NYC  
B) All second gen. immigrants in NYC  
C) All people in the US  
D) All second gen. immigrants in the US  
E) None of the above

---

Team Activity #1: Identify population and sample

☐ Take 3 minutes

☐ Using trees in Adirondacks to study the impact of climate change on forests. In this study “population” refers to

A) All trees in Adirondacks  
B) All trees in NE of U.S.  
C) All forests in the U.S.  
D) All forests in the world  
E) None of the above

---

Team Activity #1: Identify population and sample

☐ Take 3 minutes

☐ If I want to study vegetation in a drainage basin, what is the “sample”?

A) Several sample sites in the drainage basin 
B) Several kinds of vegetation  
C) Several kinds of vegetation in a few sample sites  
D) All vegetation in the drainage basin  
E) None of the above

---

Spatial population and sample

How to draw a sample?

☐ You can draw many samples from a population

☐ The best sample

- Sample characteristics match population characteristics – sample “represents” population
  - But we often do not know pop characteristics
  - Thus need “scientific” sampling design
Team Activity #2
Sampling Design

Which of the following is a “scientific” sample?
A) A sample collected by a scholar with questionnaires distributed to people at street corners
B) A sample of students drawn randomly by a computer using their IDs
C) A large sample collected by a government agency by polling people on their website
D) A sample created by a professional polling agency through calling people aged 21-60.

Is a random survey a scientific survey?
- The Republican-led House voted to eliminate the American Community Survey. Daniel Webster, a first-term Republican congressman from Florida who sponsored the relevant legislation, argued

““We’re spending $70 per person to fill this out. That’s just not cost effective,” he continued, "especially since in the end this is not a scientific survey. It's a random survey.”

How to draw a sample?
- Random sample
  - One in which every individual in the population initially has the same chance of being included in the sample

Team Activity #3
Random sample

A simple random sample of size n is one in which:
A) every nth member is selected from the population
B) each possible sample of size n has the same chance of being selected
C) there must be exactly the same proportion of women in the sample as in the population
D) you keep sampling until you have a fixed number of people having various characteristics (e.g. males, females)

Team Activity #4:
sampling design

We want to find out UAlbany students’ views on budget cut in higher education. It is impossible to interview all students. How do we choose a sample of students to represent all UAlbany students?

How to draw a random sample?

Team Activity #4:
sampling design

which of the following is the best design?
A) Send an email to all UAlbany students, and those who reply make up the sample
B) Mail the questionnaire to all students, and ask them mail back the response, followed by postcard reminders
C) Hire a few students to interview 300 people at the campus center (300 is decided based on available financial resource); a small gift will be given to the interviewee for his/her time.
D) There are 50 departments/programs. Ask each program director to interview randomly selected 6 students in their programs, which makes up 300 responses.
Team Activity #5: Sampling Design

- To help DOT plan for additional bus services in the Capital District, we want to find out information on how residents use and view public transportation. Due to financial and time constraints, we intend to interview 1000 people in the region. How do we choose these 1000 people for questionnaire survey to get representative views?

Which is the best design?

A. Randomly choose 1000 numbers from the phone book, and call them and interview them over the phone; if people refuse, call more people to make up 1000 people

B. Send questionnaire to 1000 randomly selected addresses, followed by reminders

C. Randomly choose 20 neighborhoods throughout the region, and conduct street-corner interviews for 50 people in each neighborhood

D. Identify 50 bus stops on 5 major bus lines, hand out 100 questionnaires to bus riders at each stop

- How would you draw the sample?

Sampling Design

- Randomization:
  - Fairness, equal chance; ensuring adequate sample for inference
  - Probability sampling
    - Can specify the probability of any particular sample
    - Simple random sampling
    - Systematic random sampling
    - Stratified random sampling
    - Cluster random sampling
    - Multistage sampling

- Nonprobability sampling
  - Impossible to specify the probabilities, inferences are hence of unknown reliability
  - Often unrepresentative, leading to misleading conclusions
  - Volunteer sampling
    - E.g. internet poll, street corner interview

Probability sampling methods

- (Simple) random sampling
  - All individual subjects initially have equal chance of being selected
  - Any set of elements of size n has an equal chance of selection
  - Need a complete list of all subjects in the population
    - Called "Sampling frame"
  - Computer generated random number table

Table 2.1 Part of a Table of Random Numbers

<table>
<thead>
<tr>
<th>Line/Col</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15488</td>
<td>17011</td>
<td>01536</td>
<td>02011</td>
<td>81647</td>
<td>91648</td>
<td>60179</td>
<td>54194</td>
</tr>
<tr>
<td>2</td>
<td>22368</td>
<td>46573</td>
<td>25959</td>
<td>85393</td>
<td>30935</td>
<td>89198</td>
<td>27982</td>
<td>53402</td>
</tr>
<tr>
<td>3</td>
<td>34140</td>
<td>48360</td>
<td>22527</td>
<td>97265</td>
<td>76365</td>
<td>64809</td>
<td>15179</td>
<td>28430</td>
</tr>
<tr>
<td>4</td>
<td>42167</td>
<td>93090</td>
<td>06243</td>
<td>61980</td>
<td>07856</td>
<td>16376</td>
<td>39460</td>
<td>35337</td>
</tr>
<tr>
<td>5</td>
<td>57350</td>
<td>39975</td>
<td>81357</td>
<td>16656</td>
<td>06121</td>
<td>91782</td>
<td>60668</td>
<td>81305</td>
</tr>
<tr>
<td>6</td>
<td>77921</td>
<td>06907</td>
<td>11008</td>
<td>42571</td>
<td>27756</td>
<td>53498</td>
<td>18602</td>
<td>78059</td>
</tr>
<tr>
<td>7</td>
<td>99552</td>
<td>72905</td>
<td>56430</td>
<td>99904</td>
<td>98872</td>
<td>31616</td>
<td>71194</td>
<td>18738</td>
</tr>
<tr>
<td>8</td>
<td>96301</td>
<td>91977</td>
<td>00463</td>
<td>07972</td>
<td>18876</td>
<td>20922</td>
<td>94595</td>
<td>56669</td>
</tr>
<tr>
<td>9</td>
<td>89159</td>
<td>14342</td>
<td>63661</td>
<td>10281</td>
<td>17453</td>
<td>18303</td>
<td>57740</td>
<td>48728</td>
</tr>
<tr>
<td>10</td>
<td>85475</td>
<td>38657</td>
<td>53342</td>
<td>53088</td>
<td>53860</td>
<td>59553</td>
<td>38867</td>
<td>62300</td>
</tr>
<tr>
<td>11</td>
<td>28818</td>
<td>66378</td>
<td>82231</td>
<td>33776</td>
<td>70997</td>
<td>79956</td>
<td>58863</td>
<td>50559</td>
</tr>
<tr>
<td>12</td>
<td>63553</td>
<td>40961</td>
<td>48235</td>
<td>03427</td>
<td>49626</td>
<td>69445</td>
<td>18663</td>
<td>72695</td>
</tr>
<tr>
<td>13</td>
<td>04249</td>
<td>99969</td>
<td>53636</td>
<td>97377</td>
<td>89974</td>
<td>35488</td>
<td>36320</td>
<td>17647</td>
</tr>
<tr>
<td>14</td>
<td>10465</td>
<td>61219</td>
<td>87529</td>
<td>85039</td>
<td>48237</td>
<td>52267</td>
<td>67568</td>
<td>93398</td>
</tr>
<tr>
<td>15</td>
<td>07119</td>
<td>97336</td>
<td>71048</td>
<td>08178</td>
<td>77233</td>
<td>13916</td>
<td>47664</td>
<td>81056</td>
</tr>
<tr>
<td>16</td>
<td>51085</td>
<td>12705</td>
<td>58212</td>
<td>51229</td>
<td>77452</td>
<td>16308</td>
<td>40756</td>
<td>60144</td>
</tr>
<tr>
<td>17</td>
<td>02688</td>
<td>21382</td>
<td>54204</td>
<td>60268</td>
<td>89368</td>
<td>19885</td>
<td>55322</td>
<td>44819</td>
</tr>
<tr>
<td>18</td>
<td>01011</td>
<td>54092</td>
<td>51362</td>
<td>94004</td>
<td>31273</td>
<td>04146</td>
<td>18594</td>
<td>29852</td>
</tr>
<tr>
<td>19</td>
<td>52162</td>
<td>53916</td>
<td>40369</td>
<td>58366</td>
<td>32316</td>
<td>45133</td>
<td>83149</td>
<td>98736</td>
</tr>
<tr>
<td>20</td>
<td>07006</td>
<td>97628</td>
<td>37787</td>
<td>09098</td>
<td>42698</td>
<td>06991</td>
<td>76988</td>
<td>13602</td>
</tr>
</tbody>
</table>

Sampling Design

- Probability sampling methods
  - (Simple) random sampling
  - Systematic random sampling
  - Stratified random sampling
  - Cluster sampling
    - Divide the population into a large number of clusters
    - Randomly select clusters
    - Every subject in selected clusters is included in the sample
  - Geographical cluster:
    - Spatial autocorrelation
  - Multistage sampling
  - Combination of various sampling methods

---

Team Activity #7

Take 3 minutes

According to the 2010 US census, 78% of the population are white alone, 13% are black alone, 5% are Asian alone, 4% have other races or two or more races. I want to draw a sample of 10,000 people. If I randomly draw 7000 from whites, 1300 from blacks, 1000 Asians, 700 others. What kind of sampling design did I use?

A) Simple random sample
B) Systematic random sampling
C) Stratified random sampling
D) Clustered random sampling
E) None of the above
Team Activity #9:

If I use one of the "scientific" probability sampling designs to collect my data, which of the following best describe my sample?

A) it is the best sample you can possibly get
B) it has exactly same characteristics as the population
C) it is a scientific dataset with no errors/biases
D) it is possible that I cannot make inference on population
E) I can describe the population with high confidence
Team Activity #10: Questionnaire design and measurement

Take 5 minutes:
Suppose we have already chosen 1000 residents. Which of the following is the best way to measure their views on bus service?
A) In-depth interview with guided questions, record the conversation
B) Do you agree that there is a need for more bus services? Yes vs. No
C) On a scale of 1-5, with 1 indicating strongly oppose, 5 indicating strongly support, what is your position on adding more bus services?
D) If bus services are accessible to you, how many times would you use per week?

Measurement Scales: Nominal
- A set of "unordered" categories
  - Nominal = latin for "name" or "label"
  - Even if number is used to label (1=female, 2=male)
- Categories are "homogeneous"
  - All people in that category must have a commonality
  - "Mutually Exclusive"
    - People can't fit into more than one category
  - "Exhaustive"
    - There should be a category for everyone
    - Even if it is "none of the above" or "other"
- Often called "qualitative" data

Measurement Scales: Ordinal
- Similar to nominal, but "ordered" categories
- Don't specify "distance" between categories
- Ordinal Scales are:
  - Homogeneous
  - Mutually exclusive
  - Ordered

Measurement: Interval/Ratio
- A set of numerical values, also "quantitative data"
  - Measured in comparable units, meaningful "distance"
  - E.g. # of children a person has, age, income...
  - Can no longer subdivide the basic unit (e.g. number of children, integers)
  - Infinite possibility, infinite precision (hours of work: 41.2354566)
  - Interval: no natural zero (temperature in Celsius);
    - Compare by difference, not ratio
  - Ratio: has natural/intrinsic zero (income);
    - Compare by both difference and ratio

Different Ways of Classification

Note: Ordinal data are treated sometimes as qualitative and sometimes as quantitative.
Graded Team Activity:
Measurement scale

Take 5 minutes:
- Hand out in the folder

Measurement Scale

- Different scales:
  - Nominal
  - Ordinal
  - Interval/ratio

- Different statistical methods for different scales (gender vs. income)

- Quantitative variables can be treated as qualitative, but lose information:
  - e.g. age: <20, 21-30, 31-40...
  - Years of schooling: elementary, middle, high school...

From Design to Datasets

- Choose appropriate sampling methods, & n
- Choose appropriate measurement:
  - Choose an unit of analysis
  - Choose a measurement scale
- Take measurements on relevant subjects:
  - sets of measurements on a group of cases
- Data entry, creating a database
- Data is often organized in a spreadsheet format:
  - Rows contain all measurements on each subject
  - Columns reflect sets of measurements or “variables”

Individual Activity:
Creating a dataset in SPSS

- Take 10 minutes
- Create a tiny dataset, as shown in Exhibit 1.1 in the book
- Recode age into age group:
  - You can decide what kind of age groups to be created
  - Usually 5-, 10- year cohort
- Create a new variable: age²
- Create a dummy variable based on attitude on capital punishment

Exhibit 1.1: A Tiny Set of Data

<table>
<thead>
<tr>
<th>Respondent #1:</th>
<th>Respondent #2:</th>
<th>Respondent #3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.) Male</td>
<td>1.) Male</td>
<td>1.) Female</td>
</tr>
<tr>
<td>2.) 42</td>
<td>2.) 75</td>
<td>2.) 20</td>
</tr>
<tr>
<td>3.) White</td>
<td>3.) Other</td>
<td>3.) White</td>
</tr>
<tr>
<td>4.) High-school diploma</td>
<td>4.) College degree</td>
<td>4.) Some high school</td>
</tr>
<tr>
<td>5.) Strgly support</td>
<td>5.) Oppose</td>
<td>5.) Support</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Respondent #4:</th>
<th>Respondent #5:</th>
<th>Respondent #6:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.) Male</td>
<td>1.) Female</td>
<td>1.) Female</td>
</tr>
<tr>
<td>2.) 56</td>
<td>2.) 33</td>
<td>2.) 63</td>
</tr>
<tr>
<td>3.) Black</td>
<td>3.) White</td>
<td>3.) Black</td>
</tr>
<tr>
<td>4.) Advanced degree</td>
<td>4.) College degree</td>
<td>4.) High-school diploma</td>
</tr>
<tr>
<td>5.) Strgly oppose</td>
<td>5.) No answer</td>
<td>5.) Strgly oppose</td>
</tr>
</tbody>
</table>

Exhibit 1.3: A Filled-In Dataset

<table>
<thead>
<tr>
<th>SEX</th>
<th>AGE</th>
<th>RACE</th>
<th>DEGREE</th>
<th>CAPPUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>42</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>56</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>33</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>63</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Data Manipulation

- **Recoding**
  - Example
    - Race: combining different minorities into one group
    - Age: creating age groups
    - From ratio to categorical, not vice versa
  - Always recode into a different variable; keep the original data

- **Recoding the RACE Variable**
  - Original RACE variable: Recoded RERACE variable:
    - White: 0
    - Black: 1
    - Asian: 2
    - Hispanic: 3
    - Non-white: 1

- **Recoding Variables (Ratio → ordinal)**
  - Original age variable: Recoded variable:
    - 10
    - 15
    - 20
    - 21
    - 25
    - 29
    - 32,50, 61... → 3 (>30)

Team Activity:
Recoding; creating an index

- Two variables:
  - How happy are you? Codes: 0=Very happy, 1=Pretty happy, 2=Not at all happy
  - How satisfied are you with life? Codes: 0=Not at all satisfied, 1=Fairly satisfied, 2=Very satisfied
- Creating an index to measure both
  - How? What is the scale?

Summary

- **Sampling design**
  - Probability Sampling Methods
  - (Simple) random sampling
  - Systematic random sampling
  - Stratified random sampling
  - Cluster sampling
  - Multistage sampling
  - Nonprobability Sampling Methods
  - Sampling geographic data
  - Sampling error, sample size
- **Measurement scale**
  - Nominal, ordinal interval/ratio
- **Dataset creation, manipulation**