# 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



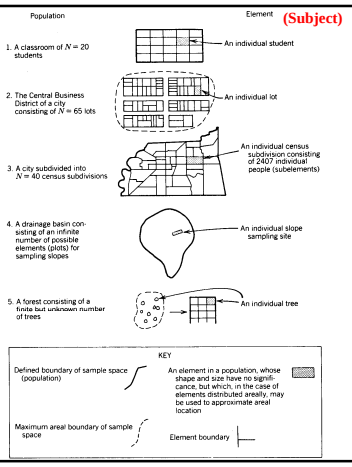
## Team Activity #1: Identify population and sample

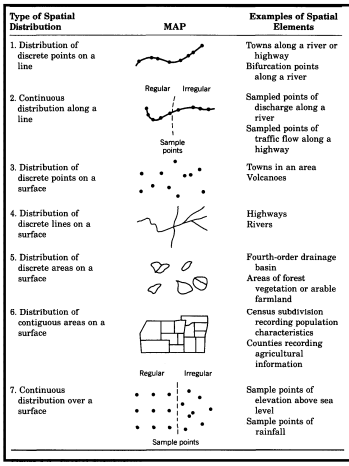
- Take 3 minutes
- If I want to study vegetation in a drainage basin, what is the "sample"?

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




*Spatial elements /subjects*

*Point*

*Line*

*Area*



## 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 address, 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



## Sampling Design

- Probability sampling methods
  - (Simple) random sampling
    - All individual subjects initially have equal chances 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

| Line/Col. | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| 1 | 10480 | 15011 | 01536 | 02011 | 81647 | 91646 | 69179 | 14194 |
| 2 | 22368 | 46573 | 25595 | 85393 | 30995 | 89198 | 27982 | 53402 |
| 3 | 24130 | 48360 | 22527 | 97265 | 76393 | 64809 | 15179 | 24830 |
| 4 | 42167 | 93093 | 06243 | 61680 | 07856 | 16376 | 39440 | 53537 |
| 5 | 37570 | 39975 | 81837 | 16656 | 06121 | 91782 | 60468 | 81305 |
| 6 | 77921 | 06907 | 11008 | 42751 | 27756 | 53498 | 18602 | 70659 |
| 7 | 99562 | 72905 | 56420 | 69994 | 98872 | 31016 | 71194 | 18738 |
| 8 | 96301 | 91977 | 05463 | 07972 | 18876 | 20922 | 94595 | 56869 |
| 9 | 89579 | 14342 | 63661 | 10281 | 17453 | 18103 | 57740 | 84378 |
| 10 | 85475 | 36857 | 53342 | 53988 | 53060 | 59533 | 38867 | 62300 |
| 11 | 28918 | 69578 | 88231 | 33276 | 70997 | 79936 | 56865 | 05859 |
| 12 | 63553 | 40961 | 48235 | 03427 | 49626 | 69445 | 18663 | 72695 |
| 13 | 09429 | 93969 | 52636 | 92737 | 88974 | 33488 | 36320 | 17617 |
| 14 | 10365 | 61129 | 87529 | 85689 | 48237 | 52267 | 67689 | 93394 |
| 15 | 07119 | 97336 | 71048 | 08178 | 77233 | 13916 | 47564 | 81056 |
| 16 | 51085 | 12765 | 51821 | 51259 | 77452 | 16308 | 60756 | 92144 |
| 17 | 02368 | 21382 | 52404 | 60268 | 89368 | 19885 | 55322 | 44819 |
| 18 | 01011 | 54092 | 33362 | 94904 | 31273 | 04146 | 18594 | 29852 |
| 19 | 52162 | 53916 | 46369 | 58586 | 23216 | 14513 | 83149 | 98736 |
| 20 | 07056 | 97628 | 33787 | 09998 | 42698 | 06691 | 76988 | 13602 |

Source: Abridged from William H. Beyer, ed., *Handbook of Tables for Probability and Statistics*, 2nd ed., © The Chemical Rubber Co., 1968, Used by permission of the Chemical Rubber Co.

## Team Activity #6

If we want to take a simple random sample of 100 students from the 10,000 students at a university, how do we utilize the random number table to select the sample?

## Sampling Design

- Probability sampling methods
  - (Simple) random sampling
  - Systematic random sampling
    - k=N/n, skip number
    - Select a subject at random from the first k subjects
    - Select every $k^{th}$ subject listed after that one
    - But, two adjacent subjects can never appear at the same time in the sample
    - Bias may occur if data is cyclical, and the cycle =k
      - Example?



## Sampling Design

- Probability sampling methods
  - (Simple) random sampling
  - Systematic random sampling
  - Stratified random sampling
    - Divide the population into separate groups (strata)
      - Race, gender, rural vs. urban, education level, occupation…
      - We are interested in comparing these strata
    - Select a simple random sample from each stratum
    - Proportional vs. disproportional



## 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:
      - Facilitate interviews by reducing travel time
      - Spatial autocorrelation
  - Multistage sampling
    - Combination of various sampling methods

*Need a complete list of population*



## Team Activity #7

Take 3 minutes

According to 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 #8
## Spatial Sampling Design

Take 5 minutes:

If we want to sample trees in the Adirondacks to study the impact of climate change, which sampling design is the best method to generate a *statistically* and *spatially* representative sample?

A) Simple random sample
B) Systematic random sampling
C) Stratified random sampling
D) Clustered random sampling
E) Multi-stage sampling



## Sampling Design

- Sampling geographic distribution
  - Spatially and statistically random

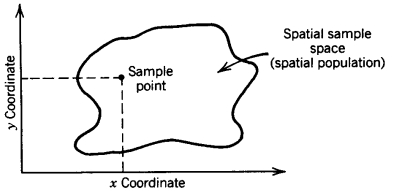
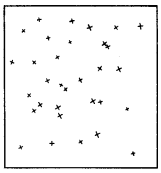



**Figure 5.2** Location of sample point by random $x$ and $y$ coordinates.



Simple random design



## 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



## Sampling Design

- Sampling error and bias
  - Errors that occurs when we use a statistic based on sample to predict the value of a population parameter
    - Approval rating: 63-68% by different polling agencies; population 66% (unknown)
  - Random sampling: ±3% (margin of error)
  - Nonprobability sampling: sampling bias
    - E.g. internet survey



## Sampling Design

- Response Bias
  - Due to the way a question is asked or worded
  - The order of questions
  - Incorrect response (characteristics of interviewees (race); lying)
  - E.g. CBS News poll in 2006
    - favor for a gasoline tax, 12%
    - Favor for a gasoline tax to reduce dependence on foreign oil , 55%
    - Favor a gasoline tax to help reduce global warming, 59%
- Nonresponse bias: missing data
  - Unreachable; refuse to participate; fail to answer Qs



## Sampling Design

- Sample size (n)
  - No rule can be provided
  - Cover the variability
  - 3% of population
  - > 30 obs
  - Finance and time may dictate sample size
  - Pilot study



## Unit of Analysis

- Casual definition: The type of things which we are collecting information about
- Common units of analysis:
  - Person (most common), family/household
  - Organization, firm, school…
  - Country, state, county, city…
  - Tree, bird, car…
  - …

## 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: Nominal

- Problem: Suppose you are interested in measuring "religion", but, a person is both Protestant and Jewish
- Solution:
  - Design a better survey that can cope with this. E.g. add category for "Prot/Jewish", or "multi-religion"
  - "Destroy" information by forcing the person to choose (or by choosing for them)
- E.g. 2000 census race



## Measurement Scale: Ordinal

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

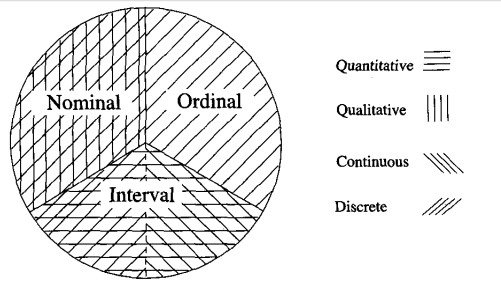


## Measurement: Interval/Ratio

- A set of numerical values, also "quantitative data"
  - Homogeneous; Ordered
  - Measured in comparable units, meaningful "distance"
- E.g. # of children a person has, age, income…
- May be "discrete" or "continuous"
  - 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 spread sheet 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: $age^2$
- Create a dummy variable based on attitude on capital punishment



**Exhibit 1.1: A Tiny Set of Data**

| | | |
|---|---|---|
| **Respondent #1:**<br>1.) Male<br>2.) 42<br>3.) White<br>4.) High-school diploma<br>5.) Strgly support | **Respondent #2:**<br>1.) Male<br>2.) 75<br>3.) Other<br>4.) College degree<br>5.) Oppose | **Respondent #3:**<br>1.) Female<br>2.) 20<br>3.) White<br>4.) Some high school<br>5.) Support |
| **Respondent #4:**<br>1.) Male<br>2.) 56<br>3.) Black<br>4.) Advanced degree<br>5.) Strgly oppose | **Respondent #5:**<br>1.) Female<br>2.) 33<br>3.) White<br>4.) College degree<br>5.) No answer | **Respondent #6:**<br>1.) Female<br>2.) 63<br>3.) Black<br>4.) High-school diploma<br>5.) Strgly oppose |



**Exhibit 1.3: A Filled-In Dataset**

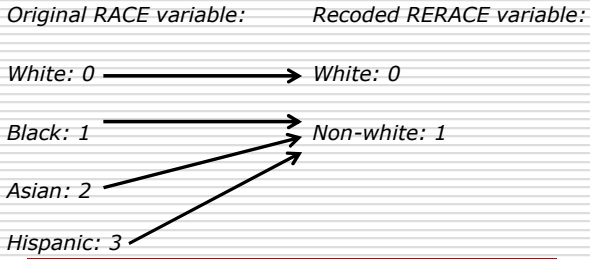
| | SEX | AGE | RACE | DEGREE | CAPPUN |
|---|---|---|---|---|---|
| 1 | 0 | 42 | 0 | 1 | 0 |
| 2 | | 75 | 2 | 3 | 2 |
| 3 | 1 | 20 | 0 | 0 | 1 |
| 4 | 0 | 56 | 1 | 4 | 3 |
| 5 | 1 | 33 | 0 | 3 | . |
| 6 | 1 | 63 | 1 | 1 | 3 |

## Data Manipulation

- Recoding
  - Example
    - Race: combining different minorities into one group
    - Age: creating age groups
    - From ratio to categorical, not vise versa
      - collect interval/ratio to begin with
- Always recode into a different variable; keep the original data

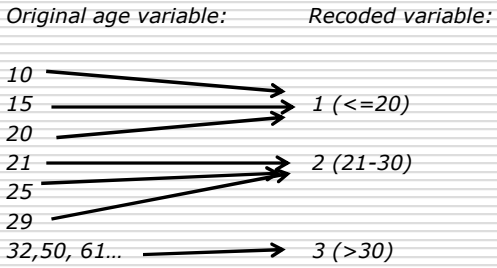


## Recoding the RACE Variable






## Recoding Variables (Ratio → ordinal)






## Data Manipulation

- Recoding
- Creating an index
  - Combining several similar variables
  - Need to be measured on the same scale
  - Housing facility index:
    - Heating: 1 yes, 0 no
    - Tap water: 1 yes, 0 no
    - Gas/electricity for cooking: 1 yes, 0 no
    - Index=heating+tap water+ cooking fuel (0-3)



## 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