Variable Types and Sampling Methods

Lecture 1

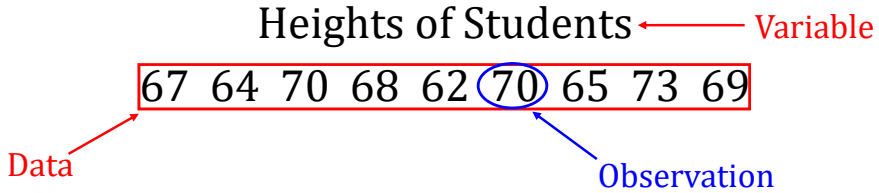
January 10, 2018

Four Stages of Statistics

- Data Collection
  - Variable Types and Sampling Methods
  - Surveys
  - Observational Studies
  - Experiments
- Displaying and Summarizing Data
- Probability
- Inference

Terminology

- **Experimental Unit:** an object (person, thing, event, etc.) upon which we collect information
- **Variable:** a characteristic of an experimental unit that differs from object to object
- **Data:** collection of observed values of a variable
- **Observation:** an individual value from a set of data



Variable Types

- **Categorical:** responses are categories with a finite number of possibilities
  - **Nominal:** ordering of categories does not matter
    - Examples: Gender, marital status, color
  - **Ordinal:** ordering of categories does matter
    - Examples: Final grade, restaurant rating, clothing sizes
- **Quantitative:** a variable where the data are real numbers and numerical operations make sense to perform
  - Examples: Height, weight, temperature, age, time

Summarizing Data

- **Categorical Data**
  - **Count:** number of observations in a category
  - **Proportion:** count in a category divided by total number of observations
  - **Percentage:** proportion as decimal times 100%
- **Quantitative Data**
  - **Mean (Average):** sum of observations divided by total number of observations
  - **Median**
  - **Standard Deviation**
  - **Variance**

Five Variable Situations

- There are typically five variable situations we encounter in statistics:
  - One Categorical
  - One Quantitative
  - Two Categorical
  - One Categorical and One Quantitative
  - Two Quantitative

  } Comparisons
- When two variables are involved, one is usually considered **explanatory** and the other is the **response**.

Example #1: Identifying Variable Situations

- **Scenario:** "The average IQ of Ivy League graduates is 142."
- **Question:** Which of the five variable situations is reflected in this scenario.
- **Answer:** _____
  - Variable(s): _____

Example #2: Identifying Variable Situations

- **Scenario:** "10% of the world's population is left-handed."
- **Question:** Which of the five variable situations is reflected in this statement?
- **Answer:** _____
  - Variable(s): _____

Example #3: Identifying Variable Situations

- **Scenario:** "New York City property values increase as houses get closer to Lower Manhattan."
- **Question:** Which of the five variable situations is reflected in this statement?
- **Answer:** _____
  - Variable(s): _____
    _____

Example #4: Identifying Variable Situations

- **Scenario:** "Study finds that mothers who smoke during pregnancy give birth to babies who weigh less compared to babies whose mothers did not smoke."
- **Question:** Which of the five variable situations is reflected in this statement?
- **Answer:** _____
  - Variable(s): _____

Example #5: Identifying Variable Situations

- **Scenario:** "Democrats are more likely to support gun control than Republicans."
- **Question:** Which of the five variable situations is reflected in this statement?
- **Answer:** _____
  - Variable(s): _____

Terminology

- **Population:** any <u>complete</u> collection of people or objects that a statistician is interested in
- **Parameter:** value that describes a characteristic of a population

- <u>Problem</u>: Parameters are usually unknown.
  - <u>Reason #1</u>: Knowing a parameter requires us to know every member of the population
  - <u>Reason #2</u>: Populations often too large to get response from every member
  - <u>Reason #3</u>: Impractical to examine every subject

Terminology

- **Sample:** set of units selected from a population that a statistician analyzes to better understand the population
- **Statistic:** value calculated from a sample that serves as an estimate of a parameter

- Solution: Take a sample from the population.
  - Sample represents population
  - Calculate statistic (proportion or mean)
  - Use statistic to approximate the parameter
    - Never perfect, but usually close

Example #6: Identifying Parts of a Study

- **Scenario:** 500 people were selected from a list of registered voters in Allegheny County. 40% of those sampled were registered Independents.
- **Task:** Identify the population, parameter, sample, and statistic.
- **Answer:**
  - Population: _____
  - Sample: _____
  - Parameter: _____
    _____
  - Statistic: _____
    _____

Sampling Methods

- Many different ways a sample can be selected from a population:
  - Simple Random Sample
  - Stratified Random Sample
  - Systematic Sample
  - Convenience Sample
  - Voluntary Sample

Simple Random Sample

- **Simple Random Sample:** subset of a population where every member has an equal chance of being chosen
  - Observations taken randomly and without replacement

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50

Stratified Random Sample

- **Stratified Random Sample:** collected by dividing population into separate groups (strata) and then drawing simple random samples from each stratum
  - Often done in proportion with population

M1	M2	M3	M4	M5	M6
M7	M8	M9	M10	M11	M12
M13	M14	M15	M16	M17	M18
M19	M20	M21	M22	M23	M24
M25	M26	M27	M28	M29	M30

$\frac{6}{30} = 20\%$ of males chosen

F1	F2	F3	F4
F5	F6	F7	F8
F9	F10	F11	F12
F13	F14	F15	F16
F17	F18	F19	F20

$\frac{4}{20} = 20\%$ of females chosen

Systematic Sample

- **Systematic Sample:** a sample collected from a population according to some pre-specified rule
  - Select every fifth person or every tenth person or every hundredth person, etc.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50

Convenience Sample

- **Convenience Sample:** subjects selected for the sample were the easiest to access
  - Subjects chosen by researcher/statistician

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50

Voluntary Sample

- **Voluntary Sample:** everyone in the population has the opportunity to participate, but sample consists of only those who choose to take part
  - Often consists of people with strong interest in topic
  - Sample chosen by viewers- not statistician

Willing Participants

2	3	8
10	12	19
21	22	23
27	30	34
41	44	49

People Unwilling to Participate

1	4	5	6	7	9	11
13	14	15	16	17	18	20
24	25	26	28	29	31	32
33	35	36	37	38	39	40
42	43	45	46	47	48	50

Example #7: Sampling Methods

- **Scenario:** Gallup wants to gauge Trump's approval rating. They randomly sample 1000 people from a list of voters in the 2016 election.
- **Question:** What type of sampling method was used?
- **Answer:** _____
  - Each person has _____

Example #8: Sampling Methods

- **Scenario:** Small city wants to know if three school districts of about the same size are performing equally well on standardized tests. Randomly sample 40 students from each school and collect their test scores.
- **Question:** What type of sampling method was used?
- **Answer:** _____
  - First _____
  - Take _____

Example #9: Sampling Methods

- **Scenario:** *The Voice* puts up a Twitter poll at the end of each episode asking people to send out a tweet to vote for the singer they believe should be saved.
- **Question:** What type of sampling method was used?
- **Answer:** _____
  - Every member of population (viewing audience) _____
  - Sample chosen _____

Example #10: Sampling Methods

- **Scenario:** Statistics student doing a project interviews people at a local mall about if they visited a clothing store while in the mall. The student records responses for the first 100 people willing to answer.
- **Question:** What type of sampling method was used?
- **Answer:** _____
  - Statistician _____
  - Subjects chosen because they were _____

Example #11: Sampling Methods

- **Scenario:** Every fourth person in line at a movie theater is taken aside and asked what movie they are going to see.
- **Question:** What type of sampling method was used?
- **Answer:** _____
  - People chosen according to _____

Errors in Data Collection

- **Sampling Error:** refers to the difference between a statistic and the parameter due to natural fluctuations in the population
  - Deviation between what we expect to happen and what the sample actually provided
- **Nonsampling Error:** occur due to mistakes that occur during the sampling process
  - **Data Acquisition Error:** recording data incorrectly in a spreadsheet or making incorrect measurements
  - **Nonresponse Bias:** a person selected to take part does not respond
  - **Selection Bias:** some people in the population cannot be selected for inclusion in the sample

Example #12: Identifying Errors

- **Scenario:** Residents of Oakland are left off of the list of potential candidates for jury duty.
- **Question:** What type of error is this?
- **Answer:** _____
  - Some members left out of _____
  - Cannot be _____

Example #13: Identifying Errors

- **Scenario:** 20 marbles are selected from a bucket of red and black marbles with half of each color. None are red.
- **Question:** What type of error is this?
- **Answer:** _____
  - Expect _____ and _____
  - No mistake was made- just an _____

Example #14: Identifying Errors

- **Scenario:** A patient is chosen to try a new weight loss drug, but does not show up for the follow-up appointment.
- **Question:** What type of error is this?
- **Answer:** _____
  - Person chosen _____
  - _____ this observation

Example #15: Identifying Errors

- **Scenario:** A doctor lists a patient's height as 622 inches.
- **Question:** What type of error is this?
- **Answer:** _____
  - Height listed _____: _____

Summary

- **Variable Types:** Categorical or quantitative
- **Five Variable Situations**
- **Population:** Entire collection, parameter
- **Sample:** Subset of population, statistic
- **Sampling Methods:** Simple random, stratified, systematic, convenience, voluntary
- **Types of Errors:**
  - **Sampling Error**
  - **Nonsampling Errors:** Data acquisition, nonresponse bias, selection bias