

SAMPLING DESIGN

- Simple Random Sample
- Systematic Sample
- Stratified Random Sample
- Cluster Sample
- Bad Sampling Methods
- Errors in Data Collection



MOTIVATION: SAMPLING PLANS

- **Scenario:** Allegheny County has 974 bridges that are at least 20 feet long. Our goal is to learn about the lengths of the bridges.
 - **Dataset:** Contains 974 rows, each with the municipality, rating, and length
- **Question:** Why is this dataset a population?
- **Answer:** _____ in Allegheny County is included
 - Allows us to calculate _____ for length
- **Scenario:** Suppose we only want to work with a subset of the data.
- **Question:** How can we be confident that the sample we obtain will give us accurate information about the population?
- **Answer:** Use a _____

PROBABILITY SAMPLING PLAN

- **Probability sampling plan:** a method of selecting a sample from a population that utilizes some form of random selection
 - A good plan will return a sample that is **representative** – the sample resembles the population across all characteristics
 - **Example:** If a population is half male and half female, the sample will have about half males and half females as well
 - **Example:** If the average age in a population is 40, then the average age in the sample is also around 40.
- Several different types
 - Simple random sample
 - Systematic sample
 - Stratified random sample
 - Cluster sample

SIMPLE RANDOM SAMPLE

4

- **Simple random sample:** a probability sampling plan where every member of the population has an equal chance of being selected for inclusion in the sample
 - Equivalent to putting every observation's name in a hat and randomly pulling out names without replacing them to obtain the sample

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

EXAMPLE: SAMPLING METHODS IN ACTION

5

- **Scenario:** Data contains every bridge in Allegheny County.
- **Task:** Use a simple random sample to estimate the average bridge length and compare it against the average length of all bridges
- **Strategy:** Use R to select a _____
 - Read in the dataset
 - Use the **sample** function to select observations and create a subset
 - Choose 50 observations
 - Average the bridge lengths in the subset to obtain the statistic
- **Parameter:** $\mu =$ _____
- **Statistic:** $\bar{x} =$ _____

OTHER PROBABILITY SAMPLING METHODS

6

- **Systematic sample:** members are sampled according to some predetermined rule by skipping a certain number of subjects and then sampling every n^{th} observation
- **Stratified random sample:** two step process where:
 - Population is first divided into groups of similar individuals (called **strata**)
 - Simple random sample is chosen from each stratum and combined to form the full sample
- **Cluster sample:** three step process where:
 - Population is first divided into groups (called **clusters**)
 - Simple random sample of the clusters is taken
 - Every member in the selected clusters becomes part of the sample

EXAMPLE: SYSTEMATIC SAMPLE

7

- **Scenario:** Data contains every bridge in Allegheny County.
- **Task:** Use a systematic sample to estimate the average bridge length.
- **Strategy:**
 - Select every _____ observation for inclusion in the sample
 - Includes observation numbers ____, ____, ____, ..., ____
 - Total of ____ observations
 - Average sampled observations
- **Parameter:** $\mu = 284.24$
- **Statistic:** $\bar{x} =$ _____

EXAMPLE: COMPARING STRATIFIED AND CLUSTER SAMPLING

8

- **Scenario:** Want to estimate support of Supreme Court for Democrats, Republicans, and Independents.
- **Question:** What sampling method should be used?
- **Answer:** _____
 - From a list of registered voters, randomly sample a subset of people from _____
 - _____ each person
 - Find _____ who support the Supreme Court for each affiliation and _____
 - Expect there to be _____

EXAMPLE: COMPARING STRATIFIED AND CLUSTER SAMPLING

9

- **Scenario:** Want to estimate the average water bill of households in a neighborhood with 20 streets by going door to door and asking residents what their bill was last month.
- **Question:** What sampling method should be used?
- **Answer:** _____
 - Streets in the same neighborhood should be _____ so a _____ random sample is not necessary
 - Randomly choose a _____ (say 3) and visit _____ on these streets
 - Better than a _____ random sample where a _____ from _____ streets might be selected

EXAMPLE: STRATIFIED RANDOM SAMPLE

10

- **Scenario:** Data contains every bridge in Allegheny County.
- **Task:** Use a stratified random sample to compare the average bridge length for those in good, fair, and poor condition.
- **Strategy:**
 - _____ the data into _____ subsets by _____
 - Randomly sample a _____ of bridges from each rating → Choose 25
 - Obtain the bridge length for each sampled observation and average the 25 bridges _____
- **Parameters:** $\mu_G = 222.76$, $\mu_F = 321.77$, $\mu_P = 147.41$
- **Statistics:** $\bar{x}_G = \underline{\hspace{2cm}}$, $\bar{x}_F = \underline{\hspace{2cm}}$, $\bar{x}_P = \underline{\hspace{2cm}}$

EXAMPLE: CLUSTER SAMPLE

11

- **Scenario:** Data contains every bridge in Allegheny County.
- **Task:** Use a cluster sample to estimate the average bridge length using municipality as the clusters.
- **Strategy:**
 - Identify _____ → _____ municipalities
 - Use R to select a _____ of clusters → Choose 5
 - Select _____ from these municipalities to create the sample
 - _____ observations in the selected clusters
- **Parameter:** $\mu = 284.24$
- **Statistic:** $\bar{x} = \underline{\hspace{2cm}}$

BIASED SAMPLING METHODS

12

- Some sampling methods are likely to be **biased** – obtained sample is not representative of the population
 - Some groups are overrepresented while others are underrepresented
- Two commonly used, often biased sampling methods include:
 - **Voluntary sample:** a sample obtained when large group of individuals is invited to participate and all responses are recorded as observations
 - Statistician does little to no work in collecting the data other than offering the opportunity to participate
 - Often performed as a survey
 - **Convenience sample:** a sample obtained from members of the population who were the easiest to access

EXAMPLE: VOLUNTARY SAMPLE

- **Scenario:** After a political debate, CNN wants to gauge if its viewers believed the Democratic or Republican candidate won the debate. They put a link to a poll on CNN.com across the scroll at the bottom of screen and invite the viewing audience to respond.
- **Question:** Why is this a voluntary sample?
- **Answer:** Respondents _____ and no sampling is done _____
- **Question:** Why is this likely to be biased?
- **Answer:**
 - Only viewers who _____ one way or another will take time to vote
 - CNN has a more _____ so sample is likely to vote for the _____ as the winner

EXAMPLE: CONVENIENCE SAMPLE

- **Scenario:** A new store is considering moving into a shopping mall. A representative for the store polls the first 100 shoppers she encountered on how likely they would be to shop at this store.
- **Question:** Why is this a convenience sample?
- **Answer:** Responses taken from people who were _____
- **Question:** Why is this sample likely to be biased?
- **Answer:**
 - Likely sampled groups who came to the mall _____ who have _____ opinions
 - Only _____ was used
 - Sample taken on _____ (many _____ people) is very different from sample taken on _____ (many _____ people and _____ parents)

ERRORS IN DATA COLLECTION

- **Sampling error:** refers to the difference between the statistic and parameter due to the observations selected for the sample
 - Occur in every sample because the statistic (almost) never equals the parameter
 - Despite the name, they are not mistakes and **occur naturally**
- **Nonsampling error:** occur due to mistakes that occur in the sampling process, either due to human error or other problems
 - **Data acquisition:** recording data or taking measurements incorrectly
 - **Nonresponse bias:** observation selected for the sample does not respond
 - **Selection bias:** some observations in the population cannot be chosen for inclusion in the sample

EXAMPLE: SAMPLING ERROR

- **Scenario:** The parameter for the length (i.e. average length of all 974 bridges) was 284.24, while the statistic from the simple random sample (i.e. average length of 50 bridges) was 296.72.
- **Question:** What is the sampling error?
- **Answer:** _____ (Difference between _____ and _____)
- **Question:** Should we be concerned about this difference?
- **Answer:** _____
 - Samples are _____ – only “missed” the parameter by _____

EXAMPLE: NONSAMPLING ERRORS

- **Task:** Match each of the following nonsampling errors with the corresponding type.
 - I. A simple random sample is chosen from a list of residents who have a driver's license.
 - II. A scale in a laboratory is not calibrated correctly and adds 3 grams to each measurement.
 - III. A subject ignores a phone call from a pollster during election season.
- **Answer:**
 - I. _____: Some members do not _____
 - II. _____: Measurements are _____
 - III. _____: Subject selected for sample _____