Statistics for Public Health Research

Center for Data Research and Analytics

Session 3:

Research Question-Driven Data Analysis

1

Outline

Part-1

- Recap of Session-2
- In-class GitHub exercise on
  - ✓Research Question
  - ✓Population
  - ✓Parameter
- Data requirement to answer a research question
- In-class practice

Part-2

- Introduction to R

2

Population

- A population is the **complete collection** of similar objects, items, or subject of interest
- Each object, item, or subject included in the population is the **study unit**
- In a research study usually, we *cannot reach* to each study unit in a population, we take a random sample (*To be discussed later*) to represent the population
- Study unit could be distinct that can be counted, non-distinct that cannot be counted

3

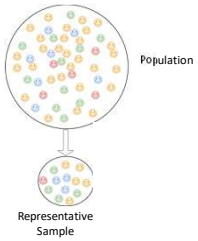
Parameter & Research Question

- A **parameter** is a **measurable characteristic** of a population of interest
- A Research Question is usually a WH type question that seeks information **about the parameter** of a **population of interest**

4

Sample and Estimator

- In Statistics, a sample is a representative part of a population
- We estimate (not determine) the parameter from a sample and **generalize** it to the population.
- The estimate of the parameter is called **estimator**



5

Sources of Bias in Research

- Sampling bias
- Measurement bias
- Due to inappropriate statistical method
- Confounding

6

Summary

- In research questions, we ask about the parameter of a population
- The population is may not accessible, so we use a random sample from the populations
- We estimate the parameter based on a sample and conclude about the population with the estimate
- Bias in research limits the generalizability of the finding from sample to the population
- Bias in research could be due to the sampling, measurement scale and inappropriate statistical method
- Know as much as possible the sources of bias in research

7

Activity-1 (5-Minutes)

- Please visit https://github.com/jaynal83/sphr/discussions/3
- Please write response to the forum post

  Write an example research question. Define population and parameter(s) from the research question. How can we get a random sample about the parameter of interest from the population? Please mention possible source of bias if there is any.
- If you don't have GitHub account, please create one now and then respond to the post

8

Session 3

Research Question-Driven Data Analysis

9

Quiz

- In a quantitative scientific article, we usually see a demographic table, **why do we need this table**?

10

Research Question-1

- What is the proportion of under 5 children suffers from anemia in Bangladesh in 2021?
- ***Population***:
- ***Parameter***:

11

Research Question-1

- What is the proportion of under 5 children suffers from anemia in Bangladesh in 2021?
- ***Population***: All live children <5 years in Bangladesh in 2021
- **Parameter**:

12

Research Question-1

- What is the proportion of under 5 children suffers from anemia in Bangladesh in 2021?
- ***Population***: All live children <5 years in Bangladesh in 2021
- ***Parameter***: **Proportion** of <5 children with anemia

13

Data Requirement

- Define **Anemia** to make it measurable and comparable
- Conduct a survey of **N** children whose age is <5 years
- Define a variable e.g., **`childID`** containing survey ID of each children
- Define a variable let's say **`anemia`** having two category, "*Yes*" and "*No*"
- Prepare a coding plan; **codebook** of the variables (see next slide)

14

Codebook or Data Dictionary

Name of Variable	Variable Label	Type	Possible Values	Value Label (if any)
`childID`	ID of surveyed child	Character of length 9 with a format `HxxxCyyyy`	`H001C0001`	
`anemia`	Status of Anemia	Numeric with single digit	`0 or 1`	`1 = Yes, 0 =No`

15

Data

childID	anemia
H001C0001	1
H002C0001	0
H002C0002	0
H003C0001	1
H004C0002	0
H005C0001	0

- Count 1's in anemia column
- Divide the count of 1's with total number of children surveyed to get the estimated proportion

What other variable do we need to appropriately answer our research question?

16

Extended Codebook

Name of Variable	Variable Label	Type	Possible Values	Value Label (if any)
childID	ID of surveyed child	Character of length 9 with a format HxxxCyyyy	H001C0001	
anemia	Status of Anemia	Numeric with single digit	0 or 1	1 = Yes, 0 =No
Age	Age of children in months	Numeric with two digits	Values between 1 to 60	
Nfammem	No. of member in the household	Numeric with two digits	Values between 1 to 20	
Medu	Mothers' education in schooling years	Numeric with two digits		
Fedu	Fathers' education in schooling years	Numeric with two digits		
Ruralurban	Place of resident	Numeric with single digit	0 or 1	1 = Urban 0 = Rural
famincome	Household income	Numeric with single digit	1, 2, 3, 4, 5	

To answer the research question appropriately, and to ensure the findings are generalizable to the population, we need some additional characteristics

17

How to Report

	Average (SD) Or Freq (%)
Age of Children (in months)	xx.x (yy)
No. of family members	
Mothers' education	
Father's education	
Family income (monthly) 0-10k 10k-20k 20k-35k 35k-60k 60k-above	
Place of resident Rural Urban	

This is a guiding template of a demographic table, where we usually report descriptive statistics.

The results presented in this table is to reflect the basic characteristics of the study population

18

Research Question-2

- What is the proportion of Covid-19 cases among fully vaccinated and unvaccinated adults in Bangladesh?
- Population:
- Parameter:
- Let's fill out the codebook now

19

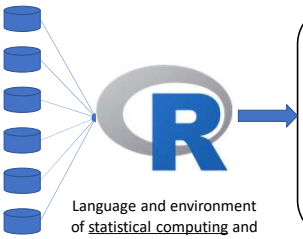
Codebook

Name of Variable	Variable Label	Type	Possible Values	Value Label (if any)
ID				
Vac_status				
Age				
Sex				
ruralurban				
Wealth				
casestatus				
Slum_nonslum				
occupation				
education				
workplace				

20

What is R?

Data

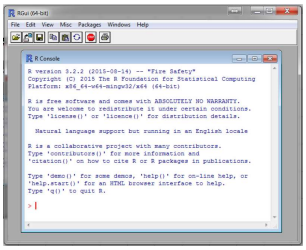


Language and environment of statistical computing and graphics

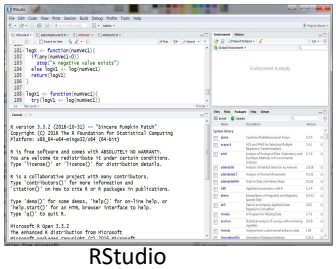
Graphs, Report and analytical dashboard

21

R vs. RStudio



R



RStudio

22

R Libraries & Repositories

- **Library:** A collection of customized pre-defined functions to perform tasks
- **Repository:** Collection of available libraries:
  - CRAN
  - CRAN (extras)
  - BioC software
  - BioC annotation
  - BioC experiment
  - BioC extra
  - R-Forge
  - rforge.net
  - Omegahat

23

Installing & Loading R Libraries

- **Net Installation**
  - Open R console
  - Connect the PC with internet
  - Type:
    `install.packages("pkgname",lib="location",dependencies=TRUE)`
  - `install.packages("ggplot2",dependencies = TRUE)`

24

Installing & Loading R Libraries

- **Installation**
  - Open R console
  - Connect the PC with internet
  - Type:
    `install.packages("pkgname", lib="location",dependencies =TRUE)`
  - `install.packages("ggplot2",dependencies = TRUE)`
- **Loading**
  - `library(ggplot2)`

25

Supported Data Files in R

- An Excel file (*.xls, *xlsx)
- **A Comma Separated Values (CSV) file**
- Plain text file with columns separated by tab or space
- Stata data file (*.dta)
- SPSS data file (*.sav)
- … and many others

26

Importing (or Reading) a Data File

- We have a csv file "`data_for_session_3.csv`" to import into R
- `read.csv()` is a typical function to import the csv file into R environment
- The complete code structure is:
  - `read.csv("path-to-the-csv-file", as.is=TRUE)`

27

Seeking Help in R

- To know how to use a function in R, we can check the help documentation by typing the following line into R console
- `help(read.csv)` or `?read.csv`
- Each of the help page is a standalone guide to use a function of interest
- The documentation of all functions in R follows same structure and easy to follow with the examples provided

28

Summary

- A research question is the guiding force to understand
  - Population
  - Parameter
  - Data requirement
- A research question also helps us to prepare appropriate data collection tools
- R is a language and environment for data analysis and visualization
- Importing a csv file into R environment for further work
- Seeking help in R to learn independently

29