R Training Book for IKU

Table of contents

Preface **1**
Introduction 1
Objective 1
Way Forward 2
Collaborators Are Welcome 2

1 Introduction **5**
1.1 R 5
1.2 RStudio 5
1.3 Quarto 5

2 Data Wrangling **7**

3 Data Wrangling **9**

4 Data Wrangling **11**

5 Complex Sampling Design in NHMS **13**
5.1 Why Complex Sampling Design? 13
5.1.1 Benefits of Complex Sampling Design 13
5.1.2 Challenges in Implementing Complex Sampling Design 14
5.1.3 Example: Sampling Probability of a Sabahan 14
5.2 Conclusion 14
5.3 Practical 15
5.3.1 Setup Project 15
5.3.2 Analysis 15

References **17**

Preface

Introduction

The Institute for Public Health (IPH) (Malay: Institut Kesihatan Umum, IKU) is a research institution under the Ministry of Health Malaysia, primarily focusing on public health research. In its daily activities, software like SPSS and STATA plays a crucial role in data analysis. However, using these softwares results in significant operational costs for the institute due to the purchase of software licenses. Recognising this issue, IKU is committed to transitioning towards using open-source and free software such as R and Python. This shift reduces cost burdens and empowers IKU staff with more flexible and advanced tools for data analysis.

R is a practical programming language for statistical analysis and graphics production. Its open-source and free nature makes it the preferred choice for research in public health. Through this book, it is hoped that the data analysis skills among IKU staff will be enhanced, leading to improvements in the quality of IKU's research.

Objective

1. Introduce R software and RStudio.
2. Enhance statistical data analysis skills, including NHMS data.
3. Explore advanced features of R and RStudio in public health applications.

Way Forward

R offers capabilities that extend well beyond statistical analysis. As more IKU staff become proficient in R, we anticipate leveraging R's diverse project capabilities to benefit IKU significantly:

1. Shiny: Develop interactive dashboards for dynamic and near-real-time result presentation.
2. Quarto: Utilize this publishing system for expedited reports and paper production.
3. IKU-specific R packages: Create tailored R packages incorporating functions for tasks such as sample size calculation, importing data from REDCap via API, standardising analysis of NHMS data, and uniform reporting of NHMS findings.

This forward-looking approach aims to harness R's full potential to streamline and enhance IKU's research and reporting processes, making them more efficient and impactful.

Collaborators Are Welcome

In the spirit of open science and continuous improvement, individuals both within and outside IPH are invites for collaboration. Whether one is an author with insights to share, an editor with an eye for detail, or possesses constructive suggestions, these contribution can significantly enhance the utility and reach of this manual. It is particularly interested in contributions in the following areas:

- Content Enhancement: The addition of new chapters or sections covering unexplored areas of R, the introduction of advanced statistical techniques, or the expansion on the applications of R in public health research are welcomed.
- Technical Review: Contributors can help ensure the accuracy of code examples, update or optimize R scripts, and contribute towards a repository of R functions tailored for public health data analysis.
- Case Studies: The IPH appreciates the sharing of real-world applications of R in public health, especially those within the context of IKU's research projects. This could include case studies on data visualization, statistical analysis, or the development of interactive applications with Shiny.
- Educational Materials: There is a need for developing tutorials, exercises, or additional learning resources that complement the manual's content, thereby facilitating a deeper understanding of R programming among IPH staff.

How to Contribute:

Individuals interested in contributing or who have suggestions to improve this manual are encouraged not to hesitate in reaching out. Your input is invaluable in making this resource more comprehensive, accurate, and beneficial for all users.

Contact Information:

Ideas, proposals for collaboration, or any feedback should be emailed to Mohd Azmi Bin Suliman at the Centre for Non-communicable Diseases Research (CNCDR). The institute looks forward to hearing from contributors and exploring how collaboration can further advance public health research through the power of R programming.

Mohd Azmi Bin Suliman
Centre for Non-communicable Diseases Research (CNCDR)
February 2024

1 Introduction

1.1 R

1.2 RStudio

1.3 Quarto

2 Data Wrangling

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

3 Data Wrangling

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

4 Data Wrangling

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

5 Complex Sampling Design in NHMS

5.1 Why Complex Sampling Design?

Surveys are essential for understanding population characteristics, offering a more efficient and resource-friendly alternative to censuses. Censuses, aiming to collect data from every individual within a population, are historically resource-intensive. In contrast, surveys, whether conducted by governments or researchers, enable effective population inferences with less expenditure.

Simple random sampling, the traditional gold standard, offers a straightforward approach to population inference due to its readily implemented methodology and unbiased estimates. However, its effectiveness wanes in populations with high heterogeneity, potentially leading to underrepresentation of minority groups. This limitation necessitates the exploration of more intricate sampling designs, such as stratified sampling, despite their inherent complexities and potential for biased selection probabilities.

One of the significant advantages of complex sampling designs is their feasibility without a comprehensive population list, focusing instead on broader stratifications like specific localities, simplifying the sampling process.

5.1.1 Benefits of Complex Sampling Design

The National Health and Morbidity Survey (NHMS), conducted by the Institute for Public Health (IPH), benefits extensively from complex sampling designs, showcasing several advantages:

1. Cost Efficiency: By clustering samples within selected strata or areas, operational costs are notably reduced, obviating the need to cover extensive and potentially scattered geographical locations.

2. Enhanced Representativeness: Stratification techniques ensure the sample accurately reflects specific subgroups or geographic areas, improving the survey's overall representativeness and reliability.

3. Data Analysis Advantages: Complex sampling designs facilitate the adjustment of sampling weights, enabling the generation of accurate national or state-level estimates. Furthermore, they support comprehensive subgroup analyses, ensuring sufficient statistical power.

5.1.2 Challenges in Implementing Complex Sampling Design

Despite their benefits, complex sampling designs require meticulous planning and sophisticated analytical techniques. These designs necessitate accounting for factors like clustering and weighting, demanding specialised expertise for both the sample's design and subsequent data analysis.

5.1.3 Example: Sampling Probability of a Sabahan

Problem: Consider a hypothetical scenario within a diverse group of 100 people, composed of 60% Malay, 20% Chinese, 15% Indian, and an additional 5% from other ethnic backgrounds, including 1% Sabahan.

Answer: To calculate the probability of selecting at least one Sabahan in a 10-person sample, one might initially consider the likelihood of not choosing a Sabahan and subtract this figure from 1. With 99 of the 100 individuals not being Sabahan, the probability of not selecting a Sabahan in a single attempt is 99/100. Over 10 independent selections, this probability becomes (99/100)^10. Consequently, the probability of selecting at least one Sabahan is 1 - (99/100)^10, equating to approximately 9.56%. This calculation suggests a close to 10% chance that the sample will include at least one Sabahan.

5.2 Conclusion

Complex sampling designs present a pragmatic and efficient approach for conducting extensive surveys like the NHMS. They ensure a representative sample, optimise resource utilisation, and require careful planning and specialised statistical expertise for accurate population inferences.

5.3 Practical

5.3.1 Setup Project

1. Setup your project
2. Copy the NHMS dataset into the working directory
3. Create Quarto document
    - update the YAML metadata to make the document self-contained

```
---
title: "Sesi 4 - NHMS"
format:
  html:
    embed-resources: true
---
```

5.3.2 Analysis

5.3.2.1 Dataset Context

1. In this practical, the example was shown using NHMS NCD 2019's cholesterol dataset.
    - we will focus on known hypercholesterolaemia status (column `known_chol`) as the outcome
2. Import Dataset
    - On the Files pane, click on the spps .sav file
    - Select Import Dataset...

References