R Training Book for IKU

Table of contents

Preface

Objectives

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.

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. How sure are we, than when we randomly select 10 people from the group, at least one of the 10 people will be a 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.

5.1.4 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.

5.2 Practical

In complex survey analysis using the `survey::` package in R, it's crucial to account for the design aspects of the survey beyond just the outcome variables and covariates. This includes specifying:

Required Information/Specification	Common NHMS Variable Name
Cluster IDs (PSU)	EB ID
Strata	State.Strata, State.wt
Sampling Weight	ADW, weight_final, weight

5.2.1 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

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

5.2.2 Analysis

5.2.2.1 Setup

0. Understand the dataset context
    - In this practical, the example was shown using NHMS NCD 2019's cholesterol dataset.
    - Two outcome will be selected
        - Categorical Type: known hypercholesterolaemia status (column `known_chol`)
        - Numerical Type: capillary total cholesterol level (column `u303`)

1. Import Dataset
    - On the `Files` pane, click on the spps .sav file
    - Select `Import Dataset ...`
    - Copy the code into the r code chunk
    - add function `as_factor(_)` to convert labelled code

```{r}
#| output: false

library(tidyverse)
library(haven)

nhms19ds <- read_sav("nhms19ds.sav") %>%
  as_factor()

nhms19ds
```

> **Note**
>
> there are 40 columns in the dataset, hence the dataset is not shown here.

2. Briefly (or in detail, up to you), explore the dataset.
    - Identify the outcome variable
        - data type: numerical, character or factor?
        - any missing data
    - Identify the complex sampling related variable:
        - the cluster ids
        - the strata
        - the sampling weight

> **Tip**
>
> some packages and functions that offer a quick data exploration:
> - `skimr::` package: `skim(_)` function.
> - `summarytools::` package: `dfSummary(_)` function.

```{r}
#| eval: false

library(skimr)

nhms19ds %>%
  select(known_chol, u303) %>%
  skim()
```

Variable Name	Variable Label
state	[Final] State
strata_gp	[Final] Locality
A2101	[Final] Gender
A2104	Age (Numerical)
A2104_grp	[Final] Age Group - 16 groups
A2106_5grp	Ethnicity (5 groups)
A2107	Citizenship
A2108_3grp	[Final] Marital Status (3 groups)
A2109_4grp	[Final] Highest Education Level (5 groups)
A2221	If working, type of occupation
A2222_7grp	Employement status (7 groups)
A2222_5grp	[Final] Occupation (5 groups)
indvid	_no label_
hh_id	_no label_
state_st	PSU
ebid	EB ID - Cluster
wtfinal_ncd	Sampling Weight
c01	ever had total blood cholesterol level measured
c02	ever told have high cholesterol level
c03	when told to have high cholesterol

Variable Name	Variable Label
c03a	years since was told to have high cholesterol
c04a	on medication for past 2 week
c04b	advice for special low fat diet
c04c	advice to loose weight
c04d	advice to exercise
c05	treatment - herbal/TCM
c06	common place to receive treatment
u303	Total Cholesterol (mmol/L)
known_chol	_no label_
undiagnosed_chol	_no label_
total_chol	_no label_
bodyweight1	Body Weight (kg)
bodyweight2	Body Weight (kg)
bodyheight1	Body Height (cm)
bodyheight2	Body Height (cm)
wc2	Waist Circumference (cm)
wc1	Waist Circumference (cm)
weight	Body Weight (kg)
height	Body Height (cm)
wc	Waist Circumference (cm)

Table 5.1: Data summary

Name	Piped data
Number of rows	10472
Number of columns	2
Column type frequency:	
factor	2
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
known_chol	6	1.00	FALSE	2	No: 8451, Yes: 2015, N/A: 0
u303	594	0.94	FALSE	87	5: 448, 5.1: 383, 4.8: 378, 4.3: 373

Warning

- there are missing values in the outcome variable `known_chol`. while is it not a must to remove sample with no outcome, as the analysis will automatic remove sample with no outcome using `na.rm = T` parameter, it is advisable to remove any sample that do not have the outcome.
- the outcome variable of capillary total cholesterol was in categorical type. we need to convert it to numerical type

Tip

later in complex sampling design analysis, the analysis accept the variable outcome (i.e. the `known_chol`) variable in either numeric or factor type. but binary type is preferable

3. In this practical we will make some data wrangling
    - remove missing outcome
    - transform factor type to numerical binary type

````
```{r}
nhms19ds <- nhms19ds %>%
  as_factor() %>%
  filter(!is.na(known_chol)) %>%
  mutate(known_cholN = as.numeric(known_chol)-2,
         u303 = as.numeric(as.character(u303)))
```
````

> **Note**
>
> The variable `known_col` have there levels, which can be check using `levels(_)` function: `levels(nhms19ds$known_chol)`. When converted to numeric using `as.numeric(_)` function, the `known_chol` value was either 1 (correspond to NA), 2 (correspond to No) and 3 (correspond to Yes), thus the value need to minus 2, so that `No` is correspond to value `0` and `Yes` is correspond with value 1.
>
> the conversion can be check by looking at both the variable
>
> ```
> ```{r}
> #| eval: false
>
> nhms19ds %>%
>   select(known_chol, known_cholN)
> ```
> ```

4. Specifying the Complex Sampling Design
    - Add options at the top of Quarto file
    - These option is to handle in which if there is single PSU within strata or domains

```
```{r}
library(survey)

options(survey.lonely.psu = 'adjust',
        survey.adjust.domain.lonely = TRUE)
```
```

    - Unweighted Design
        - cluster ids set as 1 (i.e., no clustering)
        - weight as 1 (i.e., same probability)

```
```{r}
nhms_unwdsg <- svydesign(id = ~1,
                         weights = ~1,
                         data = nhms19ds)
```
```

- we can use function `summary(_)` to view our complex sample design

```
summary(nhms_unwdsg)
```

```
Independent Sampling design (with replacement)
svydesign(id = ~1, weights = ~1, data = nhms19ds)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       1       1       1       1       1
Data variables:
 [1] "state"         "strata_gp"      "A2101"         "A2104"
 [5] "A2104_grp"     "A2106_5grp"     "A2107"         "A2108_3grp"
 [9] "A2109_4grp"    "A2221"         "A2222_7grp"     "A2222_5grp"
[13] "indvid"        "hh_id"         "state_st"       "ebid"
[17] "wtfinal_ncd"     "c01"           "c02"           "c03"
[21] "c03a"          "c04a"          "c04b"           "c04c"
[25] "c04d"          "c05"           "c06"            "u303"
[29] "known_chol"    "undiagnosed_chol" "total_chol"     "bodyweight1"
[33] "bodyweight2"    "bodyheight1"    "bodyheight2"    "wc2"
[37] "wc1"           "weight"         "height"         "wc"
[41] "known_cholN"
```

- in unweighted design, the probability for sample range from 1 to 1.
- Weighted Design
  - cluster `id` set as the PSU (commonly the variable `ebid`)
  - `strata` set as the stratification. since most NHMS applied two stage of stratification, the strata must include both 1st stage and 2nd stage (commonly the variable `state_st`)
  - `weights` set as the sampling weight
  - Note that parameter `nest = T` to ensure that the cluster is nested within the specified strata

```
```{r}
nhms_surdsg <- svydesign(id = ~ebid,
                         strata = ~state_st,
                         weights = ~wtfinal_ncd,
                         data = nhms19ds,
                         nest = T)
```
```

- we can use function `summary(_)` to view our complex sample design

```
```{r}
options(width = 70) # the output width limit

summary(nhms_surdsg)
```
```

```
Stratified 1 - level Cluster Sampling design (with replacement)
With (475) clusters.
svydesign(id = ~ebid, strata = ~state_st, weights = ~wtfinal_ncd,
    data = nhms19ds, nest = T)
Probabilities:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
1.405e-05 3.608e-04 7.000e-04 2.850e-03 2.000e-03 1.200e-01
Stratum Sizes:
            1   2   3   4   5   6   7   8   9  10  11  12  13  14
obs       584 274 281 263 281 307 331 319 294 245 302 338 307 333
design.PSU  27  13  13  11  12  12  12  12  12  12  12  12  14  12
actual.PSU  27  13  13  11  12  12  12  12  12  12  12  12  14  12
          15  16  17  18  19  20  21  22  23  24  25  26  27  28 29
obs       317 265 258 294 898 224 301 341 405 429 388 358 504 420 99
design.PSU  16  11  12  12  53  11  11  13  20  19  16  14  25  19  4
actual.PSU  16  11  12  12  53  11  11  13  20  19  16  14  25  19  4
            30
obs         506
design.PSU   33
actual.PSU   33
Data variables:
 [1] "state"            "strata_gp"        "A2101"
 [4] "A2104"            "A2104_grp"        "A2106_5grp"
 [7] "A2107"            "A2108_3grp"       "A2109_4grp"
[10] "A2221"            "A2222_7grp"       "A2222_5grp"
[13] "indvid"           "hh_id"            "state_st"
[16] "ebid"             "wtfinal_ncd"      "c01"
[19] "c02"              "c03"              "c03a"
[22] "c04a"             "c04b"             "c04c"
[25] "c04d"             "c05"              "c06"
[28] "u303"            "known_chol"      "undiagnosed_chol"
[31] "total_chol"       "bodyweight1"      "bodyweight2"
[34] "bodyheight1"      "bodyheight2"      "wc2"
[37] "wc1"              "weight"           "height"
[40] "wc"               "known_cholN"
```

- in weighted design summary, several info were given
  - the sampling probabilities. in this dataset, each of the sample have probability from 0.00001 to 0.12
  - the number of strata, number of sample in each of the strata and number of PSU (EB) in each strata. in this dataset, there are total 30 strata (13 states + 3 federal territories, with each state have 2 locality urban and rural).

5.2.2.2 Count the unweighted sample

1. To count the number of sample, we will use function `svymean(_)` from `survey::`.
    - the outcome variable can be either factor type, or if it in numerical type, it must be binary 0-1 number.
    - to estimate the number of sample, we will use the unweighed design.
    - the `x =` parameter must be in formula form with ~ (tilde) symbol before the variable name, i.e. `~known_chol`.
2. this is if we want to use the original factor type.

````
```{r}
svytotal(x = ~known_chol,
         design = nhms_unwdsg,
         na.rm = T)
```
````

```
              total     SE
known_cholN/A     0  0.000
known_cholNo   8451 40.339
known_cholYes  2015 40.339
```

3. this is if we want to use the converted to binary 0-1 numerical type. noticed the output differences.

````
```{r}
svytotal(x = ~known_cholN,
         design = nhms_unwdsg,
         na.rm = T)
```
````

```
            total     SE
known_cholN  2015 40.339
```

> **Note**
>
> Note 1: noticed that parameter `na.rm =` were set as T (TRUE). this is so that any sample with missing at parameter `x =` (i.e. the `known_chol`) will be removed.
>
> Note 2: From this point forward, I'll use `known_cholN` variable (the binary 0-1 numerical type) as the outcome. You are feel free to use the original factor type, and explore as you wish.

5.2.2.3 Estimating the estimated population

1. to estimate total number of population that have the outcome (i.e., known_cholN), same formula is used, with changes at the design used, i.e. the weighted design

```
```{r}
svytotal(x = ~known_cholN,
         design = nhms_surdsg,
         na.rm = T)
```
```

```
              total     SE
known_cholN 2868124 103013
```

5.2.2.4 Estimating Prevalence

0. Estimating the prevalence using the function of svymean(_) from survey :: package.
    - if the outcome variable is factor type, both original factor type and converted numerical type can be used.
        - if original factor type is used, prevalence for both No and Yes will be estimated.
        - if the outcome have three or more levels, using original factor type is preferable.
        - when using the binary 0-1 numerical type (i.e., the known_cholN), svymean(_) will calculate prevalence by calculating how many 1 since 0 does not have value.

1. Using function svymean(_) to calculate

```
```{r}
svymean(x = ~known_cholN,
        design = nhms_surdsg,
        na.rm = T)
```
```

```
               mean     SE
known_cholN 0.13479 0.0051
```

5.2.2.5 Estimating Confidence Interval for Prevalence

1. To calculate the confidence interval for prevalence, function `svyciprop(_)` from package `survey` `::` will be used.
    - Generally, a generic function `confint(_)` can be used to calculate the confident interval for model parameter.
    - In R however, the function will treat proportion as mean of binary outcomes. While treating proportion as mean of binary outcomes is reasonable accepted to calculate the prevalence, however, when calculate the CI, it is preferable to treat apply logit transformation and transformed back to the original scale
    - the default method used in `svyciprop(_)` function is "logit"
    - however, to replicate result from SPSS and SUDAAN, the method parameter need to change to "xlogit" :::

```
```{r}
svyciprop(formula = ~known_cholN,
          design = nhms_surdsg,
          method = "xl") %>%
  attr(., "ci")
```
```

```
     2.5%     97.5% 
0.1251425 0.1450549 
```

> **Note**
>
> - function `attr(_)` is used to pull the attribute from the object (i.e., the output of the `svyciprop(_)` function), while the parameter `"ci"` in the `attr(_, "ci")` function is to pull the CI from the `svyciprop(_)`

5.2.2.6 Estimating the Unweighted Sample Proportion

> **Tip**
>
> Can you calculated the sample proportion using the same function?.
> Hint:
>
> 1. Sample Proportion = Unweighted Proportion.
> 2. Unweighted design vs. Weighted design.

5.2.2.7 Estimating by Subpopulation

5.2.2.8 Total Sample and Estimated Population

Can you try calculate the total sample? Using the example from calculating the total number sample with the outcome.

The tutorial on estimated total population will be cover in Bonus II: Population Pyramid part

5.3 Bonus I:

5.4 Bonus II: Population Pyramid

References