Phân tích số liệu sử dụng R

Phân tích mô tả thật đơn giản

Ba Khuong Cao - caobakhuongytcc@gmail.com

! Ưu điểm

- Chỉ cần một câu lệnh là có cả bảng kết quả sẵn sàng đưa vào bài báo/luận văn
- Không cần phải phân tích cho từng biến số và chuyển kết quả vào bảng một cách thủ công.
- Có thể xuất kết quả ra file word và copy/paste trực tiếp. Việc này giúp hạn chế sai sót trong quá trình phân tích do không cần phải đánh tay kết quả vào bảng.

i Các bước thực hiện

1. Làm quen với dataset (Mô tả ngắn gọn)
2. Loading packages
3. Loading dataset
4. Thực hiện một số bước cần thiết để mã hóa dữ liệu
5. Phân tích kết quả và xuất ra file Word (.docX)

1. Mô tả dataset

Trong phần phân tích ngày hôm nay, chúng ta sẽ sử dụng bộ số liệu được public trên tạp chí. Bài báo khoa học này nói về việc nhiễm HIV ở phụ nữ hoạt động mại dâm.

Link: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0197251

Phần phân tích này sử dụng dataset có tên `data_file_Table_1.csv`. Phần mô tả cụ thể của dataset này như sau (Theo công bố trong bài báo):

Variable name (Tên biến)	Variable description (Mô tả)	Codes (Mã hóa)
unique	Subject identification number	Numeric
q103	Nationality	1 = Benin 2 = Ghana 3 = Nigeria 4 = Togo 5 = Others
age	Age in years	Numeric
age_3cat	Age in years (categories)	1 = 18–24 2 = 25–34 3 = 35
q105	Education	0 = Not educated 1 = Primary 2 = Secondary 3 = Postsecondary, university
q114	Marital status	1 = Married 2 = Single 3 = Divorced 4 = Widowed
q107	Place of work	1 = Brothel 2 = Home 3 = Bars 4 = Hotels 5 = Street 6 = Nightclub 7 = Others
duree	Duration in sex work (months)	Numeric
q205	Number of clients (last 7 days of work)	Numeric
q219	Number of sexual intercourses (last 3 days)	Numeric
constante	Consistent condom use with clients (last 7 days of work)	1 = Yes 2 = No
q211	Condom use with last client	1 = Yes 2 = No
groupe	HIV/treatment status	1 = HIV positive, treated 2 = HIV positive, not treated 3 = HIV negative
anal	Ever had anal sex (any partner)	1 = No 2 = Yes

Variable name (Tên biến)	Variable description (Mô tả)	Codes (Mã hóa)
hla9	Ever had oral sex (any partner)	1 = Yes 2 = No
gono	NG prevalence	1 = Positive 2 = Negative 9 = No answer
chlam	CT prevalence	1 = Positive 2 = Negative 9 = No answer
gonochlam	NG/CT prevalence	1 = Positive 2 = Negative 9 = No answer
cd4_nb	CD4 cell count/mm^3	Numeric
charge_viral_copies	Viral load	Numeric

2. Loading packages

Chúng ta sẽ sử dụng một vài packages: `gmodels`, `dplyr`, `psych`, `gtsummary`, `readxl`, `readr`, `finalfit`, `flextable`.

```
# Check whether pacman is available and install if needed
options(repos = c(CRAN = "https://cloud.r-project.org"))
if (!requireNamespace("pacman", quietly = TRUE)) install.packages("pacman")

# Use pacman to install (if needed) and load the required packages
pacman::p_load(gmodels, dplyr, psych, gtsummary, readxl,readr,finalfit,flextable)
```

3. Loading dataset

Các datasets liên quan đến bài báo này có thể tải xuống theo link:

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0197251

Chúng ta có thể load dataset này vào R để phân tích theo lệnh sau:

```r
# Read CSV
hiv_tab_1_data <- read_csv(file.path(data_folder, "data_file_Table_1.csv"))

# Inspect the structure of the dataset
str(hiv_tab_1_data)
```

```
spc_tbl_ [319 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ unique              : chr [1:319] "D001T" "D002T" "D003T" "D004T" ...
 $ q103                : num [1:319] 2 1 4 2 4 2 2 3 1 4 ...
 $ age                 : num [1:319] 47 51 42 44 49 45 51 25 44 44 ...
 $ age_3cat            : num [1:319] 3 3 3 3 3 3 3 2 3 3 ...
 $ q105                : num [1:319] 1 1 0 1 0 1 0 1 0 1 ...
 $ q114                : num [1:319] 3 3 2 4 3 2 2 2 2 2 ...
 $ q107                : num [1:319] 5 5 1 1 5 1 1 1 2 1 ...
 $ duree               : num [1:319] 48 180 60 36 24 60 180 72 48 120 ...
 $ q205                : num [1:319] 19 32 72 35 54 11 27 13 1 19 ...
 $ q219                : num [1:319] 12 14 38 14 28 9 13 9 0 13 ...
 $ constante           : num [1:319] 1 2 1 1 1 2 1 1 1 1 ...
 $ q211                : num [1:319] 1 1 1 1 1 1 1 1 1 1 ...
 $ groupe              : num [1:319] 1 1 1 1 1 1 1 1 1 1 ...
 $ anal                : num [1:319] 1 1 1 1 1 1 1 1 1 1 ...
 $ hla9                : num [1:319] 2 2 2 2 2 2 2 2 2 2 ...
 $ gono                : num [1:319] 1 2 2 1 2 2 2 2 2 2 ...
 $ chlam               : num [1:319] 2 2 2 2 2 2 2 2 2 2 ...
 $ gonochlam           : num [1:319] 1 2 2 1 2 2 2 2 2 2 ...
 $ cd4_nb              : num [1:319] 219 297 191 464 525 145 263 710 734 280 ...
 $ charge_virale_copies: num [1:319] 50 50 1683 848 50 ...
 - attr(*, "spec")=
  .. cols(
  ..   unique = col_character(),
  ..   q103 = col_double(),
  ..   age = col_double(),
  ..   age_3cat = col_double(),
  ..   q105 = col_double(),
  ..   q114 = col_double(),
  ..   q107 = col_double(),
  ..   duree = col_double(),
  ..   q205 = col_double(),
  ..   q219 = col_double(),
  ..   constante = col_double(),
  ..   q211 = col_double(),
  ..   groupe = col_double(),
```

```
  ..   anal = col_double(),
  ..   hla9 = col_double(),
  ..   gono = col_double(),
  ..   chlam = col_double(),
  ..   gonochlam = col_double(),
  ..   cd4_nb = col_double(),
  ..   charge_virale_copies = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

> **Notes**
>
> Hầu hết tất cả các biến trong dataset này đều là dạng `num`(`numeric`), trừ biến `unique`. Như vậy để phù hợp với dạng dữ liệu được mô tả theo bảng ở trên, một số biến sẽ cần biến đổi thành `factor`(Biến phân loại/nhị phân)

4. Biến đổi/Mã hóa dữ liệu

Chuyển một số biến sang dạng `factor`

```
# Convert to factors ----
factor_vars <- c(
  "q103", "age_3cat", "q105", "q114", "q107",
  "constante", "q211", "groupe", "anal", "hla9",
  "gono", "chlam", "gonochlam")

hiv_tab_1_data[factor_vars] <- lapply(hiv_tab_1_data[factor_vars], as.factor)
```

Gán nhãn giá trị

Các nhãn giá trị của các biến có thể được gán (theo bảng mô tả các biến ở trên)

```
#Labeling for values of variables ----
levels(hiv_tab_1_data$q103)=c("Benin","Ghana","Nigeria", "Togo", "Others")
levels(hiv_tab_1_data$age_3cat)=c("18 - 24","25 - 34",">=35")
levels(hiv_tab_1_data$q105)=c("Not educated","Primary","Secondary",
                              "Postsecondary, university")
levels(hiv_tab_1_data$q114)=c("Married","Single","Divorced", "Widowed")
```

```
levels(hiv_tab_1_data$q107)=c("Brothel","Home","Bars", "Hotels","Street",
                              "Nightclub","Others")
levels(hiv_tab_1_data$constante)=c("Yes","No")
levels(hiv_tab_1_data$q211)=c("Yes","No")
levels(hiv_tab_1_data$groupe)=c("HIV positive, treated",
                                "HIV positive, not treated", "HIV negative")
levels(hiv_tab_1_data$anal)=c("No","Yes")
levels(hiv_tab_1_data$hla9)=c("Yes","No")
levels(hiv_tab_1_data$gono)=c("Positive","Negative", "No answer")
levels(hiv_tab_1_data$chlam)=c("Positive","Negative", "No answer")
levels(hiv_tab_1_data$gonochlam)=c("Positive","Negative", "No answer")
```

5. Phân tích kết quả và xuất file

Để có bảng kết quả (Bảng 1) như trong bài báo đã công bố, chúng ta thực hiện như sau:

```
#Table 1. Baseline characteristics  ----
des_table1 <- hiv_tab_1_data %>%
  tbl_summary(
    include = c(q103, age, age_3cat, q105, q114, q107, duree, q205, q219,
                constante, q211, groupe, anal, hla9, gono, chlam, gonochlam,
                cd4_nb, charge_virale_copies),
    type = list(
      all_continuous() ~ "continuous2"
    ),
    statistic = all_continuous() ~ c("{median} ({p25}, {p75})"),
    percent = "column",
    label = list(
      q103 ~ "Nationality",
      age ~ "Age in years",
      age_3cat ~ "Age in years (categories)",
      q105 ~ "Education",
      q114 ~ "Marital status",
      q107 ~ "Place of work",
      duree ~ "Duration in sex work, months",
      q205 ~ "Number of clients, last 7 days of work",
      q219 ~ "Number of sexual intercourses, last 3 days",
      constante ~ "Consistent condom use with clients (last 7 days of work)",
      q211 ~ "Condom use with last client",
      groupe ~ "HIV/treatment status",
```

```
      anal ~ "Ever had anal sex (any partner)",
      hla9 ~ "Ever had oral sex (any partner)",
      gono ~ "NG prevalence",
      chlam ~ "CT prevalence",
      gonochlam ~ "NG/CT prevalence",
      cd4_nb ~ "CD4 cell count/mm³",
      charge_virale_copies ~ "Viral load"
    ),
    digits = list(
      all_continuous() ~ 2,
      all_categorical() ~ 1
    ),
    missing = "no"
  ) %>%
  modify_header(label = "**Variable**") %>%         # change label column header
  modify_caption("**Table 1. Baseline characteristics of 319 female sex workers,
                 Cotonou, Benin, 2008-2012**")   # add table name/title

des_table1 %>%
  as_kable(
    format = "latex",
    longtable = TRUE,
    booktabs = TRUE
  ) %>%
  kableExtra::kable_styling(latex_options = c("repeat_header"))
```

Table 2: **Table 1. Baseline characteristics of 319 female sex workers, Cotonou, Benin, 2008–2012**

Variable	**N = 319**
Nationality	
Benin	123.0 (38.6%)
Ghana	49.0 (15.4%)
Nigeria	66.0 (20.7%)
Togo	74.0 (23.2%)
Others	7.0 (2.2%)
Age in years	
Median (Q1, Q3)	34.00 (27.00, 41.00)
Age in years (categories)	
18 - 24	48.0 (15.0%)

Table 2: **Table 1. Baseline characteristics of 319 female sex worke *(continued)*

Variable	**N = 319**
25 - 34	112.0 (35.1%)
>=35	159.0 (49.8%)
Education	
Not educated	107.0 (33.5%)
Primary	126.0 (39.5%)
Secondary	79.0 (24.8%)
Postsecondary, university	7.0 (2.2%)
Marital status	
Married	21.0 (6.6%)
Single	102.0 (32.0%)
Divorced	149.0 (46.7%)
Widowed	47.0 (14.7%)
Place of work	
Brothel	126.0 (39.6%)
Home	24.0 (7.5%)
Bars	37.0 (11.6%)
Hotels	40.0 (12.6%)
Street	76.0 (23.9%)
Nightclub	6.0 (1.9%)
Others	9.0 (2.8%)
Duration in sex work, months	
Median (Q1, Q3)	36.00 (16.00, 60.00)
Number of clients, last 7 days of work	
Median (Q1, Q3)	13.00 (5.00, 26.00)
Number of sexual intercourses, last 3 days	
Median (Q1, Q3)	5.00 (1.00, 11.00)
Consistent condom use with clients (last 7 days of work)	246.0 (77.1%)
Condom use with last client	279.0 (87.5%)
HIV/treatment status	
HIV positive, treated	49.0 (15.4%)
HIV positive, not treated	82.0 (25.7%)
HIV negative	188.0 (58.9%)
Ever had anal sex (any partner)	16.0 (5.0%)
Ever had oral sex (any partner)	62.0 (19.4%)
NG prevalence	
Positive	13.0 (4.1%)

Table 2: **Table 1. Baseline characteristics of 319 female sex worke *(continued)*

Variable	**N = 319**
Negative	301.0 (94.4%)
No answer	5.0 (1.6%)
CT prevalence	
Positive	9.0 (2.8%)
Negative	305.0 (95.6%)
No answer	5.0 (1.6%)
NG/CT prevalence	
Positive	21.0 (6.6%)
Negative	293.0 (91.8%)
No answer	5.0 (1.6%)
CD4 cell count/mm³	
Median (Q1, Q3)	389.00 (214.50, 583.00)
Viral load	
Median (Q1, Q3)	8,153.00 (154.00, 31,321.00)

```
des_table1
```

Bảng kết quả có thể xuất ra file word

```
des_table1 %>%
 as_flex_table() %>%
 save_as_docx(path = "Table 1. Baseline characteristics.docx")
```

Variable	**N = 319**[1]
Nationality	
Benin	123.0 (38.6%)
Ghana	49.0 (15.4%)
Nigeria	66.0 (20.7%)
Togo	74.0 (23.2%)
Others	7.0 (2.2%)
Age in years	
Median (Q1, Q3)	34.00 (27.00, 41.00)
Age in years (categories)	
18 - 24	48.0 (15.0%)
25 - 34	112.0 (35.1%)
>=35	159.0 (49.8%)
Education	
Not educated	107.0 (33.5%)
Primary	126.0 (39.5%)
Secondary	79.0 (24.8%)
Postsecondary, university	7.0 (2.2%)
Marital status	
Married	21.0 (6.6%)
Single	102.0 (32.0%)
Divorced	149.0 (46.7%)
Widowed	47.0 (14.7%)
Place of work	
Brothel	126.0 (39.6%)
Home	24.0 (7.5%)
Bars	37.0 (11.6%)
Hotels	40.0 (12.6%)
Street	76.0 (23.9%)
Nightclub	6.0 (1.9%)
Others	9.0 (2.8%)
Duration in sex work, months	
Median (Q1, Q3)	36.00 (16.00, 60.00)
Number of clients, last 7 days of work	
Median (Q1, Q3)	13.00 (5.00, 26.00)
Number of sexual intercourses, last 3 days	
Median (Q1, Q3)	5.00 (1.00, 11.00)
Consistent condom use with clients (last 7 days of work)	246.0 (77.1%)
Condom use with last client	279.0 (87.5%)
HIV/treatment status	
HIV positive, treated	49.0 (15.4%)
HIV positive, not treated	82.0 (25.7%)
HIV negative	188.0 (58.9%)
Ever had anal sex (any partner)	16.0 (5.0%)
Ever had oral sex (any partner)	62.0 (19.4%)
NG prevalence	
Positive	13.0 (4.1%)
Negative	301.0 (94.4%)
No answer	5.0 (1.6%)
CT prevalence	
Positive	9.0 (2.8%)