Calculate Descriptive Statistics of Participants

Mingcheng Hu

Table of contents

```
library(gtsummary)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
source("/work/users/y/u/yuukias/BIOS-Material/BIOS992/utils/csv_utils.r")
```

```
train_data <-
 ↪  read.csv("/work/users/y/u/yuukias/BIOS-Material/BIOS992/data/train_data_unimputed.csv")

test_data <-
 ↪  read.csv("/work/users/y/u/yuukias/BIOS-Material/BIOS992/data/test_data_unimputed.csv")

print(dim(train_data))
```

```
[1] 28127   101
```

```r
print(dim(test_data))
```

```
[1] 7032  101
```

```r
total_data <- bind_rows(train_data, test_data) %>% select(-c(X))

print(dim(total_data))
```

```
[1] 35159   100
```

```r
# We only care about the covariates
columns_all <- colnames(total_data)
columns_covariate_idx <- sapply(columns_all, function(x) {
    if (determine_category(x) == "covariate") {
        return(TRUE)
    } else {
        return(FALSE)
    }
})
columns_covariate <- columns_all[columns_covariate_idx]
data_covariate <- total_data %>% select(c(all_of(columns_covariate), event))
```

```r
# data <- data_covariate %>%
# filter(ethnicity != "e")
data <- data_covariate

# * We also need to manually relevel the categorical variables
data <- data %>%
    mutate(
        event = factor(event,
            levels = c("0", "1"),
            labels = c("No CVD", "CVD")
        ),

        # Set "Never" (0) as baseline for smoking
        smoking = factor(smoking,
            levels = c("0", "1", "2", "-3"),
            labels = c("Never", "Previous", "Current", "Prefer not to
            ↪  answer")
```

```
        ),

        # Set "No" (0) as baseline for diabetes
        diabetes = factor(diabetes,
            levels = c("0", "1", "-1", "-3"),
            labels = c("No", "Yes", "Do not know", "Prefer not to answer")
        ),

        # Ensure other categorical variables are properly factored
        ethnicity = factor(ethnicity,
            levels = c("1", "2", "3", "4", "5", "6"),
            labels = c("White", "Mixed", "Asian/Asian British", "Black/Black
             ↪  British", "Chinese", "Other")
        ),
        education = factor(education,
            levels = c("1", "2", "3", "4", "5", "6", "-7", "-3"),
            labels = c(
                "College/University degree", "A levels/AS levels",
                "O levels/GCSEs", "CSEs", "NVQ/HND/HNC",
                "Other professional", "None of the above",
                "Prefer not to answer"
            )
        ),
        activity = factor(activity,
            levels = c("0", "1", "2"),
            labels = c("Low", "Moderate", "High")
        ),
        sex = factor(sex,
            levels = c("0", "1"),
            labels = c("Female", "Male")
        ),
        hypertension_treatment = factor(hypertension_treatment,
            levels = c("0", "1"),
            labels = c("No", "Yes")
        )
    )
```

```
head(data)
```

```
  age    sex           ethnicity     BMI  smoking diabetes systolic_bp
1  62 Female Asian/Asian British 26.5089    Never       No          144
```

```
2  52   Male               White 27.9123 Previous       No         129
3  42   Male               White 28.2933    Never       No         136
4  67 Female               White 28.5573  Current       No         166
5  64 Female               White 23.2140    Never       No         119
6  45   Male               White 27.7743    Never       No         138
  hypertension_treatment total_chol hdl_chol                education activity
1                     No      5.337    2.013 College/University degree Moderate
2                     No      5.485    1.473                     CSEs      Low
3                     No      5.773    0.924 College/University degree Moderate
4                    Yes      6.086    1.975        A levels/AS levels     <NA>
5                     No      5.871    1.387      Prefer not to answer     High
6                     No      6.812    1.354 College/University degree Moderate
  max_workload max_heart_rate  event
1           60            123 No CVD
2          120            115 No CVD
3          120            144    CVD
4           40            125    CVD
5           70             96 No CVD
6          120            123 No CVD
```

```
# digits for missing rate
list("tbl_summary-fn:percent_fun" = label_style_number(scale = 100, digits =
 ↪  2)) |>
  set_gtsummary_theme()
data %>%
    tbl_summary(
        by = "event",
        digits = list(
            all_continuous() ~ 1,
            all_categorical() ~ c(0, 2)  # 0 for count, 2 for percentage
        ),
        statistic = list(
            all_continuous() ~ c("{mean}±{sd}"),
            all_categorical() ~ c("{n} ({p}%)")
        ),
        missing_stat = "{N_miss} ({p_miss}%)",  # show missing rate
         ↪  percentage instead of count
        missing_text = "Missing"
    ) %>%
    add_p() %>%
    as_kable_extra(
        booktabs = TRUE,
```

```
        longtable = TRUE,
        linesep = ""
    )
```

```
Warning: 'xfun::attr()' is deprecated.
Use 'xfun::attr2()' instead.
See help("Deprecated")

Warning: 'xfun::attr()' is deprecated.
Use 'xfun::attr2()' instead.
See help("Deprecated")
```

Characteristic	**No CVD** N = 30,719	**CVD** N = 4,440	**p-value**
age	54.8±8.1	59.9±7.1	<0.001
sex			<0.001
Female	17,664 (57.50%)	1,829 (41.19%)	
Male	13,055 (42.50%)	2,611 (58.81%)	
ethnicity			<0.001
White	26,607 (86.76%)	3,935 (88.95%)	
Mixed	1,494 (4.87%)	202 (4.57%)	
Asian/Asian British	1,751 (5.71%)	216 (4.88%)	
Black/Black British	249 (0.81%)	24 (0.54%)	
Chinese	164 (0.53%)	10 (0.23%)	
Other	402 (1.31%)	37 (0.84%)	
Missing	52 (0.17%)	16 (0.36%)	
BMI	26.7±4.3	27.8±4.5	<0.001
Missing	11 (0.04%)	7 (0.16%)	
smoking			<0.001
Never	17,738 (57.84%)	2,283 (51.60%)	
Previous	10,111 (32.97%)	1,669 (37.73%)	
Current	2,734 (8.92%)	457 (10.33%)	
Prefer not to answer	84 (0.27%)	15 (0.34%)	
Missing	52 (0.17%)	16 (0.36%)	
diabetes			<0.001
No	30,122 (98.22%)	4,242 (95.89%)	
Yes	482 (1.57%)	166 (3.75%)	
Do not know	47 (0.15%)	13 (0.29%)	
Prefer not to answer	16 (0.05%)	3 (0.07%)	
Missing	52 (0.17%)	16 (0.36%)	

systolic_bp	136.9±18.3	144.3±18.4	<0.001
Missing	139 (0.45%)	18 (0.41%)	
hypertension_treatment	5,083 (16.63%)	1,350 (30.60%)	<0.001
Missing	155 (0.50%)	28 (0.63%)	
total_chol	5.9±1.1	5.9±1.1	<0.001
Missing	2,305 (7.50%)	356 (8.02%)	
hdl_chol	1.5±0.4	1.4±0.4	<0.001
Missing	3,712 (12.08%)	550 (12.39%)	
education			<0.001
College/University degree	11,957 (38.99%)	1,382 (31.24%)	
A levels/AS levels	3,855 (12.57%)	494 (11.17%)	
O levels/GCSEs	6,509 (21.22%)	922 (20.84%)	
CSEs	1,896 (6.18%)	211 (4.77%)	
NVQ/HND/HNC	1,775 (5.79%)	348 (7.87%)	
Other professional	1,403 (4.57%)	259 (5.85%)	
None of the above	3,043 (9.92%)	763 (17.25%)	
Prefer not to answer	229 (0.75%)	45 (1.02%)	
Missing	52 (0.17%)	16 (0.36%)	
activity			0.6
Low	4,013 (15.57%)	585 (16.17%)	
Moderate	10,498 (40.72%)	1,464 (40.48%)	
High	11,267 (43.71%)	1,568 (43.35%)	
Missing	4,941 (16.08%)	823 (18.54%)	
max_workload	85.2±24.1	82.0±24.9	<0.001
max_heart_rate	115.1±13.8	111.5±14.6	<0.001
Missing	1 (0.00%)	0 (0.00%)	

[1] Mean±SD; n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test