SPPH 604 001 Lab Exercise: Data wrangling

Contents

Problem

Use the functions we learned in Lab 1 to complete Lab 1 Exercise. We will use Right Heart Catheterization Dataset saved in the folder named 'Data/wrangling/'. The variable list and description can be accessed from this website (https://biostat.app.vumc.org/wiki/pub/Main/DataSets/rhc.html).

A paper you can access the original table from this paper (doi: 10.1001/jama.1996.03540110043030). We have modified the table and corrected some issues. Please knit your file once you finished and submit the knitted file **ONLY**.

```
# install LaTex
tinytex::install_tinytex()
```

```
# Load required packages
library(dplyr)
library(tableone)
library(tidyverse)
# Data import: name it rhc
rhc <- read.csv("C:/Users/ccyal/Downloads/rhc.csv", header = TRUE)
#head(rhc)
```

Part (a) Basic Manipulation [60%]

(I) Continuous to Categories: Change the Age variable into categories below 50, 50 to below 60, 60 to below 70, 70 to below 80, 80 and above [Hint: the `cut` function could be helpful]

```
#summary(rhc$age)
rhc$age <- cut(rhc$age, c(0, 50, 60, 70, 80, Inf), right = TRUE,
                   labels = c("<50", "50-59", "60-69", "70-79", "80+"))

table(rhc$age)
```

```
##
##   <50 50-59 60-69 70-79   80+
##  1424   917  1389  1338   667
```

(II) Re-order: Re-order the levels of race to white, black and other

```
#table(rhc$race)
rhc$race <- factor(rhc$race, levels = c("white", "black","other"))
levels(rhc$race)
```

```
## [1] "white" "black" "other"
```

(III) Set reference: Change the reference category for gender to Male

```
#table(rhc$sex)
rhc$sex <- factor(rhc$sex, levels = c("Male", "Female"))
levels(rhc$sex)
```

```
## [1] "Male"   "Female"
```

(IV) Count levels: Check how many levels does the variable "cat1" (Primary disease category) have? Re-group the levels for disease categories to "ARF", "CHF","MOSF","Other". [Hint: the `nlevels` and `list` functions could be helpful]

```
num_levels <- nlevels(rhc$cat1)

cat("The variable 'cat1' has", num_levels, "levels\n")
```

```
## The variable 'cat1' has 0 levels
```

```
levels(rhc$cat1) <- list(ARF = c("ARF(ARF)"),
                         CHF = c("CHF(CHF)"),
                         Other = c("Other(Cirrhosis)", "Other(Colon Cancer)",
                                   "Other(Coma)", "Other(COPD)",
                                   "Other(Lung Cancer)"),
                         MOSF = c("MOSF(MOSF w/Malignancy)",
                                  "MOSF(MOSF w/Sepsis)"))
num_levels_new <- nlevels(rhc$cat1)

cat("The variable 'cat1' now has", num_levels_new, "levels\n")
```

```
## The variable 'cat1' now has 4 levels
```

(V) Rename levels: Rename the levels of "ca" (Cancer) to "Metastatic","None" and "Localized (Yes)", then re-order the levels to "None","Localized (Yes)" and "Metastatic"

```r
# Original freq
# table(rhc$ca)
# rename categories
rhc$ca_new <- factor(
  x = rhc$ca,
  levels = c("No", "Yes", "Metastatic"),
  labels = c("None", "Localized (Yes)", "Metastatic")
)
# table(rhc$ca_new)
# Reorder the levels of "ca"
rhc$ca_new <- factor(
  x = rhc$ca_new,
  levels = c("None", "Localized (Yes)", "Metastatic")
)
# Display levels
levels(rhc$ca_new)
```

```
## [1] "None"            "Localized (Yes)" "Metastatic"
```

(VI) comorbidities:

- create a new variable called "numcom" to count number of comorbidities illness for each person (12 categories) [Hint: the `rowSums` command could be helpful],
- report maximim and minimum values of numcom:

```r
rhc <- rhc %>%
  rowwise() %>%
  mutate(numcom=sum(c(cardiohx,chfhx,dementhx,psychhx,chrpulhx,renalhx,liverhx,
                      gibledhx,malighx,immunhx,transhx,amihx)))

max_numcom <- max(rhc$numcom)

cat("The maximum number of a person have for the comorbidities illness is",
    max_numcom,"\n")
```

```
## The maximum number of a person have for the comorbidities illness is 6
```

```r
min_numcom <- min(rhc$numcom)

cat("The minimum number of a person have for the comorbidities illness is",
    min_numcom)
```

```
## The minimum number of a person have for the comorbidities illness is 0
```

(VII) Anlaytic data: Create a dataset that has only the following variables

- "age", "sex", "race","cat1", "ca", "dnr1", "aps1", "surv2md1", "numcom", "adld3p", "das2d3pc", "temp1", "hrt1", "meanbp1", "resp1", "wblc1", "pafi1", "paco21", "ph1", "crea1", "alb1", "scoma1", "swang1", and
- name it rhc2.

```
rhc2 <- rhc %>%
  select(age, sex, race,cat1, ca, dnr1, aps1, surv2md1, numcom, adld3p,
         das2d3pc, temp1, hrt1, meanbp1,
         resp1, wblc1, pafi1, paco21, ph1, crea1, alb1, scoma1,swang1)
#dim(rhc2)
```

Part (b) Table 1 [20%]

(i) Re-produce the sample table from the rhc2 data (see the Table that was provided with this assignment). In your table, the variables should be ordered as the same as the sample. Please re-level or re-order the levels if needed. [Hint: the `tableone` package might be useful]

```
table_title <- "Table 1: Characteristics of Critically Ill Patients"

# Create a summary table using CreateTableOne
table1 <- CreateTableOne(vars = c("age", "sex", "race","cat1","ca","dnr1",
                                  "aps1", "surv2md1", "numcom", "adld3p",
                                  "das2d3pc", "temp1", "hrt1", "meanbp1",
                                  "resp1", "wblc1", "pafi1", "paco21", "ph1",
                                  "crea1", "alb1", "scoma1"),
                         strata = "swang1",
                         data = rhc2,
                         test=FALSE)

# Print the summary table
cat("##", table_title, "\n\n")
```

```
## ## Table 1: Characteristics of Critically Ill Patients
```

```
print(table1)
```

```
##                       Stratified by swang1
##                        No RHC          RHC
##   n                      3551            2184
##   age (%)
##      <50                  884 (24.9)      540 (24.7)
##      50-59                546 (15.4)      371 (17.0)
##      60-69                812 (22.9)      577 (26.4)
##      70-79                809 (22.8)      529 (24.2)
##      80+                  500 (14.1)      167 ( 7.6)
##   sex = Female (%)       1637 (46.1)      906 (41.5)
##   race (%)
##      white                2753 (77.5)     1707 (78.2)
##      black                 585 (16.5)      335 (15.3)
##      other                 213 ( 6.0)      142 ( 6.5)
##   cat1 (%)
##      ARF                  1581 (44.5)      909 (41.6)
##      CHF                   247 ( 7.0)      209 ( 9.6)
##      Cirrhosis             175 ( 4.9)       49 ( 2.2)
##      Colon Cancer            6 ( 0.2)        1 ( 0.0)
##      Coma                  341 ( 9.6)       95 ( 4.3)
```

```
##      COPD                  399 (11.2)       58 ( 2.7)
##      Lung Cancer            34 ( 1.0)        5 ( 0.2)
##      MOSF w/Malignancy     241 ( 6.8)      158 ( 7.2)
##      MOSF w/Sepsis         527 (14.8)      700 (32.1)
##   ca (%)
##      Metastatic            261 ( 7.4)      123 ( 5.6)
##      No                   2652 (74.7)     1727 (79.1)
##      Yes                   638 (18.0)      334 (15.3)
##   dnr1 = Yes (%)           499 (14.1)      155 ( 7.1)
##   aps1 (mean (SD))       50.93 (18.81)   60.74 (20.27)
##   surv2md1 (mean (SD))    0.61 (0.19)     0.57 (0.20)
##   numcom (mean (SD))      1.52 (1.17)     1.48 (1.13)
##   adld3p (mean (SD))      1.24 (1.86)     1.02 (1.69)
##   das2d3pc (mean (SD))   20.37 (5.48)    20.70 (5.03)
##   temp1 (mean (SD))      37.63 (1.74)    37.59 (1.83)
##   hrt1 (mean (SD))      112.87 (40.94)  118.93 (41.47)
##   meanbp1 (mean (SD))    84.87 (38.87)   68.20 (34.24)
##   resp1 (mean (SD))      28.98 (13.95)   26.65 (14.17)
##   wblc1 (mean (SD))      15.26 (11.41)   16.27 (12.55)
##   pafi1 (mean (SD))     240.63 (116.66) 192.43 (105.54)
##   paco21 (mean (SD))     39.95 (14.24)   36.79 (10.97)
##   ph1 (mean (SD))         7.39 (0.11)     7.38 (0.11)
##   crea1 (mean (SD))       1.92 (2.03)     2.47 (2.05)
##   alb1 (mean (SD))        3.16 (0.67)     2.98 (0.93)
##   scoma1 (mean (SD))     22.25 (31.37)   18.97 (28.26)
```

(ii) Table 1 for subset

Produce a similar table as part (b) but with only male sex and ARF primary disease category (cat1). Add the overall column in the same table. [Hint: `filter` command could be useful]

```r
table_title <- "Table 1 Subset: Characteristics of Male Patients with ARF"
# Create a summary table for males with ARF primary disease category (cat1)
table1_filtered <- rhc2 %>%
  filter(sex == "Male", cat1 == "ARF") %>%
  CreateTableOne(
    vars = c("age", "numcom", "adld3p", "das2d3pc", "temp1", "hrt1", "meanbp1",
             "resp1", "wblc1", "pafi1", "paco21", "ph1", "crea1", "alb1",
             "scoma1"),
    strata = "swang1",
    addOverall = TRUE,  # Add an overall column
    test = FALSE
  )

# Print the summary tables
cat("##", table_title, "\n\n")
```

```
## ## Table 1 Subset: Characteristics of Male Patients with ARF
```

```r
print(table1_filtered)
```

```
##                       Stratified by swang1
```

```
##                          Overall          No RHC           RHC
##   n                        1382             888             494
##   age (%)
##      <50                    382 (27.6)       267 (30.1)      115 (23.3)
##      50-59                  198 (14.3)       127 (14.3)       71 (14.4)
##      60-69                  299 (21.6)       174 (19.6)      125 (25.3)
##      70-79                  340 (24.6)       201 (22.6)      139 (28.1)
##      80+                    163 (11.8)       119 (13.4)       44 ( 8.9)
##   numcom (mean (SD))       1.34 (1.16)      1.32 (1.16)     1.38 (1.15)
##   adld3p (mean (SD))       1.00 (1.79)      1.00 (1.78)     1.01 (1.80)
##   das2d3pc (mean (SD))    21.74 (5.62)     21.67 (5.72)    21.87 (5.44)
##   temp1 (mean (SD))       37.96 (1.71)     38.02 (1.69)    37.84 (1.76)
##   hrt1 (mean (SD))       115.96 (39.26)   115.52 (39.39)  116.76 (39.06)
##   meanbp1 (mean (SD))     79.08 (36.38)    83.69 (36.81)   70.80 (34.11)
##   resp1 (mean (SD))       29.01 (14.35)    30.27 (14.21)   26.73 (14.33)
##   wblc1 (mean (SD))       15.80 (12.03)    15.92 (11.50)   15.58 (12.93)
##   pafi1 (mean (SD))      188.09 (100.74)  208.05 (102.50) 152.20 (86.70)
##   paco21 (mean (SD))      37.45 (10.03)    38.08 (10.56)   36.32 (8.89)
##   ph1 (mean (SD))          7.40 (0.10)      7.40 (0.10)     7.39 (0.10)
##   crea1 (mean (SD))        2.22 (2.25)      2.10 (2.33)     2.45 (2.09)
##   alb1 (mean (SD))         3.07 (0.68)      3.12 (0.66)     2.98 (0.70)
##   scoma1 (mean (SD))      18.42 (27.05)    19.48 (28.07)   16.51 (25.02)
```

Part (c) Considering eligibility criteria [20%]

Produce a similar table as part (b.i) but only for the subjects who meet all of the following eligibility criteria: (i) age is equal to or above 50, (ii) age is below 80 (iii) Glasgow Coma Score is below 61 and (iv) Primary disease categories are either ARF or MOSF. [Hint: `droplevels.data.frame` can be a useful function]

```r
rhc2_eligible <- rhc2 %>%
  filter(age %in% c("50-59", "60-69", "70-79")) %>%
  filter(scoma1 < 61)

#dim(rhc2_eligible)
#table(rhc2_eligible$age)

# First age
rhc2_eligible$age <- factor(rhc2_eligible$age, levels = c("50-59", "60-69", "70-79"))
# Then coma score
# Next to include only subjects with primary disease categories "ARF" or "MOSF"
rhc2_cat1Exclusive <- rhc2_eligible %>%
  filter(cat1 %in% c("ARF", "MOSF w/Malignancy","MOSF w/Sepsis"))
# Remove unused levels from categorical variables
rhc2_cat1Exclusive <- droplevels(rhc2_cat1Exclusive)

#recode cat1
rhc2_cat1Exclusive$cat1 <- gsub("MOSF w/Malignancy", "MOSF",
                                rhc2_cat1Exclusive$cat1)
rhc2_cat1Exclusive$cat1 <- gsub("MOSF w/Sepsis", "MOSF",
                                rhc2_cat1Exclusive$cat1)
# rename categories
rhc2_cat1Exclusive$ca_new <- factor(
  x = rhc2_cat1Exclusive$ca,
```

```
  levels = c("No", "Yes", "Metastatic"),
  labels = c("None", "Localized (Yes)", "Metastatic")
)
# table(rhc$ca_new)
# Reorder the levels of "ca"
rhc2_cat1Exclusive$ca_new <- factor(
  x = rhc2_cat1Exclusive$ca_new,
  levels = c("None", "Localized (Yes)", "Metastatic")
)
# Display levels
#levels(rhc2_cat1Exclusive$ca_new)

table_title <- "Table 1: Characteristics of Eligible Patients"

# Create a summary table using CreateTableOne
table_eligible <- CreateTableOne(
  vars = c("age", "sex", "race", "cat1", "ca_new", "dnr1", "aps1", "surv2md1",
           "numcom", "adld3p", "das2d3pc", "temp1", "hrt1", "meanbp1", "resp1",
           "wblc1", "pafi1", "paco21", "ph1", "crea1", "alb1", "scoma1"),
  strata = "swang1",
  data = rhc2_cat1Exclusive,
  test = FALSE
)

# Print the summary table
cat("##", table_title, "\n\n")
```

```
## ## Table 1: Characteristics of Eligible Patients
```

```
print(table_eligible)
```

```
##                         Stratified by swang1
##                          No RHC          RHC
##   n                        1226            1102
##   age (%)
##      50-59                  321 (26.2)      263 (23.9)
##      60-69                  429 (35.0)      416 (37.7)
##      70-79                  476 (38.8)      423 (38.4)
##   sex = Female (%)          550 (44.9)      469 (42.6)
##   race (%)
##      white                  977 (79.7)      908 (82.4)
##      black                  187 (15.3)      135 (12.3)
##      other                   62 ( 5.1)       59 ( 5.4)
##   cat1 = MOSF (%)           417 (34.0)      545 (49.5)
##   ca_new (%)
##      None                   812 (66.2)      816 (74.0)
##      Localized (Yes)        283 (23.1)      212 (19.2)
##      Metastatic             131 (10.7)       74 ( 6.7)
##   dnr1 = Yes (%)            150 (12.2)       67 ( 6.1)
##   aps1 (mean (SD))        54.57 (18.13)   61.82 (18.77)
##   surv2md1 (mean (SD))     0.59 (0.16)     0.55 (0.16)
##   numcom (mean (SD))       1.51 (1.14)     1.50 (1.12)
```

```
##   adld3p (mean (SD))     1.18 (1.82)     1.22 (1.90)
##   das2d3pc (mean (SD))  20.22 (5.22)    20.57 (4.82)
##   temp1 (mean (SD))     37.85 (1.72)    37.67 (1.82)
##   hrt1 (mean (SD))     116.68 (39.28)  120.02 (40.26)
##   meanbp1 (mean (SD))   79.99 (37.36)   67.57 (33.53)
##   resp1 (mean (SD))     30.57 (12.90)   26.34 (13.90)
##   wblc1 (mean (SD))     16.35 (13.38)   16.91 (13.06)
##   pafi1 (mean (SD))    226.30 (110.48) 178.37 (94.37)
##   paco21 (mean (SD))    37.94 (11.25)   36.38 (10.44)
##   ph1 (mean (SD))        7.40 (0.10)     7.38 (0.10)
##   crea1 (mean (SD))      2.20 (2.34)     2.53 (2.06)
##   alb1 (mean (SD))       3.09 (0.66)     2.94 (1.08)
##   scoma1 (mean (SD))    14.22 (18.48)   13.65 (17.91)
```

Optional 1: Missing values

(I) Any variables included in rhc2 data had missing values? Name that variable. [Hint: `apply` function could be helpful]

```
# Check for missing values in the rhc2 dataset
missing <- apply(rhc2, 2, function(x) anyNA(x))

# Identify variables with missing values
var_name <- names(missing[missing])

# Print the names of variables with missing values
cat("The variable name is", var_name)
```

```
## The variable name is adld3p
```

(II) Count how many NAs does that variable have?

```
na_counts <- colSums(is.na(rhc2[, var_name]))

cat("This variable",var_name,"has",na_counts ,"missing")
```

```
## This variable adld3p has 4296 missing
```

(III) Produce a table 1 for a complete case data (no missing observations) stratified by `swang1`.

```
# Create a subset of rhc2 with only non-missing data
rhc2_nonMissing <- rhc2[complete.cases(rhc2), ]
table_title <- "Table 1: Characteristics of Critically Ill Patients with No Missing Data"

# Create a summary table using CreateTableOne
table_nonMissing <- CreateTableOne(
  vars = c("age", "sex", "race", "cat1", "ca", "dnr1", "aps1", "surv2md1",
           "numcom", "adld3p", "das2d3pc", "temp1", "hrt1", "meanbp1", "resp1",
           "wblc1", "pafi1", "paco21", "ph1", "crea1", "alb1", "scoma1"),
  strata = "swang1",
  data = rhc2_nonMissing,
```

```
  test = FALSE
)

# Print the summary table
cat("##", table_title, "\n\n")
```

```
## ## Table 1: Characteristics of Critically Ill Patients with No Missing Data
```

```
print(table_nonMissing)
```

```
##                         Stratified by swang1
##                          No RHC          RHC
##   n                        1049            390
##   age (%)
##      <50                    264 (25.2)      113 (29.0)
##      50-59                  160 (15.3)       85 (21.8)
##      60-69                  261 (24.9)       99 (25.4)
##      70-79                  238 (22.7)       70 (17.9)
##      80+                    126 (12.0)       23 ( 5.9)
##   sex = Female (%)          480 (45.8)      137 (35.1)
##   race (%)
##      white                  813 (77.5)      297 (76.2)
##      black                  176 (16.8)       67 (17.2)
##      other                   60 ( 5.7)       26 ( 6.7)
##   cat1 (%)
##      ARF                    429 (40.9)      127 (32.6)
##      CHF                    174 (16.6)      129 (33.1)
##      Cirrhosis               71 ( 6.8)        5 ( 1.3)
##      Colon Cancer             2 ( 0.2)        0 ( 0.0)
##      Coma                     2 ( 0.2)        2 ( 0.5)
##      COPD                   179 (17.1)       15 ( 3.8)
##      Lung Cancer             12 ( 1.1)        2 ( 0.5)
##      MOSF w/Malignancy       68 ( 6.5)       25 ( 6.4)
##      MOSF w/Sepsis          112 (10.7)       85 (21.8)
##   ca (%)
##      Metastatic              81 ( 7.7)       20 ( 5.1)
##      No                     797 (76.0)      324 (83.1)
##      Yes                    171 (16.3)       46 (11.8)
##   dnr1 = Yes (%)             87 ( 8.3)       11 ( 2.8)
##   aps1 (mean (SD))       48.36 (16.34)   49.38 (19.71)
##   surv2md1 (mean (SD))    0.70 (0.15)     0.69 (0.17)
##   numcom (mean (SD))      1.74 (1.22)     1.76 (1.23)
##   adld3p (mean (SD))      1.24 (1.86)     1.02 (1.69)
##   das2d3pc (mean (SD))   20.36 (7.28)    20.36 (6.96)
##   temp1 (mean (SD))      37.35 (1.66)    37.24 (1.61)
##   hrt1 (mean (SD))      112.23 (38.20)  108.66 (39.22)
##   meanbp1 (mean (SD))    87.35 (37.97)   70.91 (33.38)
##   resp1 (mean (SD))      30.43 (11.65)   25.25 (12.73)
##   wblc1 (mean (SD))      14.45 (11.16)   14.75 (13.09)
##   pafi1 (mean (SD))     250.90 (112.53) 238.90 (104.11)
##   paco21 (mean (SD))     41.77 (14.86)   37.16 (8.57)
##   ph1 (mean (SD))         7.39 (0.10)     7.40 (0.09)
##   crea1 (mean (SD))       2.03 (2.27)     2.22 (2.05)
```

```
##   alb1 (mean (SD))       3.26 (0.65)     3.19 (0.64)
##   scoma1 (mean (SD))     5.25 (15.83)    6.54 (17.20)
```

Optional 2: Calculating variance of a sample

Write a `function` for Bessel's correction to calculate an unbiased estimate of the population variance from a finite sample (a vector of 100 observations, consisting of numbers from 1 to 100).

```
Vector <- 1:100

#variance.est <- function(?){?}
variance_est <- function(rhc2) {
  n <- length(rhc2)  # Number of observations
  mean_val <- mean(rhc2)  # Mean of the data

  # Initialize the sum of squared differences
  sum_squared_diff <- 0

  # Calculate the sum of squared differences from the mean
  for (i in 1:n) {
    sum_squared_diff <- sum_squared_diff + (rhc2[i] - mean_val)^2
  }

  # Calculate the population variance with Bessel's correction
  population_variance <- sum_squared_diff / (n - 1)

  return(population_variance)
}

#variance.est(Vector)
variance_result <- variance_est(Vector)
print(variance_result)
```

```
## [1] 841.6667
```

```
# Calculate the standard deviation
estimated_std_dev <- sqrt(variance_result)

# Print the standard deviation
print(estimated_std_dev)
```

```
## [1] 29.01149
```

Hint: Take a closer look at the functions, loops and algorithms shown in lab materials. Use a `for loop`, utilizing the following pseudocode of the algorithm:

Naïve algorithm [edit]

A formula for calculating the variance of an entire population of size N is:

$$\sigma^2 = \overline{(x^2)} - \bar{x}^2 = \frac{\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2/N}{N}.$$

Using Bessel's correction to calculate an unbiased estimate of the population variance from a finite sample of n observations, the formula is:

$$s^2 = \left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right) \cdot \frac{n}{n-1}.$$

Therefore, a naive algorithm to calculate the estimated variance is given by the following:

- Let $n \leftarrow 0$, Sum $\leftarrow 0$, SumSq $\leftarrow 0$
- For each datum x:
  - $n \leftarrow n + 1$
  - Sum $\leftarrow$ Sum $+ x$
  - SumSq $\leftarrow$ SumSq $+ x \times x$
- Var = (SumSq − (Sum × Sum) / n) / (n − 1)

Verify that estimated variance with the following variance function output in R:

```
var(Vector)
```

```
## [1] 841.6667
```

Knit your file

Please knit your file once you finished and submit the knitted PDF or doc file. Please also fill-up the following table:

Group name: ** Antique Ruby **

Student initial	% contribution
YL	48%
JE	26%
MJ	26%