ベース機能を使っての表作成

Table of contents

オブジェクトからのデータの抽出

ベクトルからのデータ抽出

- インデックスによる抽出
  - ベクトル内の各要素には、その位置を示すインデックス（添え字、位置番号）が割り当てられている
  - R では最初の要素のインデックスは 0 ではなく 1[1]
  - 角括弧 [　] の中に取り出したい要素のインデックスを指定
- 論理型ベクトルによる抽出
  - TRUE に対応する要素のみを抽出

```
x <- c(5, 10, 15, 20)
x[2] # 2番目の要素
x[c(1,4)] # 1番目と4番目
x[-c(2,3)] # 2番目と3番目以外
x_flg <- c(TRUE,FALSE,FALSE,TRUE)
x[x_flg] # TRUEの要素のみ
```

[1]他の多くの言語（C, Python, Java など）とは異なることに注意

マトリックスからのデータ抽出

- 指定の仕方が 2 次元になる ([行番号, 列番号])
- 片方のインデックスのみ指定することで
  その次元の全要素（行全体または列全体）を抽出

```
m <- matrix(1:9, nrow = 3)
m[1, 2] # 1行2列目
m[ ,1] # 1列目（全行）
m[2 ,] # 2行目（全列）
```

リストからのデータ抽出

リスト: 異なる型（数値、文字列、ベクトル、他のリストなど）のオブジェクトを複数格納できる構造

- インデックスによる抽出
  - [[ ]] を使うと、リスト内の要素そのものを取り出す
  - [ ] を使うと、指定した要素を含むサブリストが返される
- 要素名による抽出 ($" 名前" または [[" 名前"]])
  - $ 記法がよく使われる

```
l <- list(name = "Tanaka", age = 50, scores = c(80,
    90))
l[[2]] # インデックスによる抽出(要素そのもの)
l[2] # インデックスによる抽出(サブリスト)
l[["name"]] # 要素名による抽出
l$name # $による抽出
```

data.frame からのデータ抽出

- リスト型とマトリックス型の双方の抽出方法が可能
- 特に以下は多用
  - 列名による抽出
  - 論理型のベクトルによる条件に合致した要素の抽出

```
df <- data.frame(id = 1:3, sex = c("M", "F", "M"))
df$sex
df[["sex"]]
df[, "sex"]
df[, 2]
df[1, ]
male_flg <- df$sex=="M" # 男性ならばTRUE
df[male_flg,] # 男性の行のみ抽出
```

演習問題 (復習込み)

1. `set.seed(123)` を使い、標準正規分布（平均 0, 標準偏差 1 の正規分布）に従う長さ 100 の乱数を生成してください
2. 10 番目の要素を抽出してください

```
set.seed(123)
x <- rnorm(100)
x[10] # -0.445662
```

一変数の集計

使用データセット：smoking_tbl_df

英国の健康調査に基づく、喫煙に関するデータ

- 対象：イギリスにおける成人 1691 名の調査データ
- 変数一覧：
  - gender：性別（male, female）
  - age：年齢（整数）
  - marital_status：婚姻状況（single, married など）
  - highest_qualification：最終学歴
  - nationality：国籍
  - ethnicity：民族的背景
  - gross_income：所得水準
  - region：地域
  - smoke：喫煙の有無（yes, no）
  - amt_weekends：週末の喫煙本数（整数）
  - amt_weekdays：平日の喫煙本数（整数）
  - type：喫煙者における喫煙スタイル（Packets,Hand-Rolled など）
- 出典：英国の喫煙行動に関する健康調査データ

データセットの読み込みと確認

```
# 初回のみパッケージインストール
if(!any(search() %in% "package:MedDataSets")){
  install.packages("MedDataSets",dep=TRUE)
}
# パッケージ読み込み
library(MedDataSets)
# データ読み込み
data("smoking_tbl_df")

# データの形式を確認
class(smoking_tbl_df) # tbl: パッケージ固有の属性
# tibble形式を通常のdata.frameに変換
smoking_tbl_df <- as.data.frame(smoking_tbl_df)

# データの内容を確認
head(smoking_tbl_df)
summary(smoking_tbl_df)
```

カテゴリ変数の度数と相対度数

- table()：各カテゴリの出現回数（度数）を集計
- prop.table()：度数表に基づき、全体に対する割合を計算
  相対度数は合計が 1 になるように計算される

```
freq_tab <- table(smoking_tbl_df$gender)
freq_tab
prop.table(freq_tab)
prop.table(freq_tab)*100 # パーセント表示（割合 × 100）
```

table() における NA の扱い

useNA 引数で NA の表示有無を制御

- "no"：デフォルト（NA 除外）
- "ifany"：NA がある場合のみ表示
- "always"：常に表示（0 も出力）

```
x <- c("A", "B", NA, "A", "C", NA)
table(x)
table(x, useNA = "ifany")
table(x, useNA = "always")
# NAを考慮した相対度数
freq_tab = table(x, useNA = "always")
prop.table(freq_tab)
# NA がないオブジェクトに対してもNAを集計
freq_tab <- table(smoking_tbl_df$gender, useNA = "
    always")
prop.table(freq_tab)
```

連続変数の階級分けと頻度集計

cut( ) 関数: 連続値を区間に分け、factor 型に変換

- breaks 引数で区間の区切り位置を指定する[2]

```
# 年齢を区間(0,20], (20,40], (40,60], (60,∞)に分割
age_category <- cut(smoking_tbl_df$age, breaks = c(0,
    20, 40, 60, Inf))
# 分類結果を確認
head(age_category)
# 度数を集計
table(age_category)
# 相対度数を計算
prop.table(table(age_category))
```

[2] Inf は正の無限大を表す数値型の特別な値

二変数：離散変数のクロス集計

二変数のクロス集計と割合

- table( ) の引数に 2 つのベクトルを指定
- prop.table の margin 引数を指定
  行方向・列方向の割合を計算

```
tab2 <- table(smoking_tbl_df$gender, smoking_tbl_df$
    smoke)
tab2
prop.table(tab2) # 全体の割合
prop.table(tab2, margin=1) # 行ごとの割合
prop.table(tab2, margin=2) # 列ごとの割合
```

演習問題

1. amt_weekends を「20 本以下」と「20 本より多い」に階級分けしてください
2. 階級分けした amt_weekends の頻度と相対度数を NA を考慮して求めてください
3. 階級分けした amt_weekends と smoke のクロス集計表を作成し、amt_weekends の欠損の理由を確認してください

演習問題 R コード

```
# データ確認
summary(smoking_tbl_df$amt_weekends) # NAのあるデータ
# 1. amt_weekends を 0〜20本未満・20本以上に階級分け
amt_weekends_cat <- cut(smoking_tbl_df$amt_weekends,
    breaks = c(-Inf, 20, Inf))
# 2. 度数と相対度数を計算
tab_amt_weekends = table(amt_weekends_cat, useNA="
    always")
prop.table(tab_amt_weekends)
# 3. amt_weekends と smoke のクロス集計
cross_tab <- table(amt_weekends_cat, smoking_tbl_df$
    smoke, useNA="always")
cross_tab # 喫煙していない人が欠損している
```

要約統計量の算出

要約統計量の算出（連続変数）

- mean( ): 平均
- median( ): 中央値
- sd( ): 標準偏差
- var( ): 分散
- 欠損値（NA）がある場合は、na.rm = TRUE を指定

```
mean(smoking_tbl_df$age)
median(smoking_tbl_df$age)
sd(smoking_tbl_df$age)
var(smoking_tbl_df$age)
# 欠損値がある場合、na.rmの指定が必要
mean(smoking_tbl_df$amt_weekends) #NA
mean(smoking_tbl_df$amt_weekends,na.rm = TRUE)
```

層別の集計と要約統計量の計算

層別のクロス集計

層（stratum）とは、あるカテゴリ変数に基づいて、互いに重ならないグループに分割したもの

層ごとにデータを分け、各グループで集計

```
# 男性だけのデータ
male_df <- smoking_tbl_df[smoking_tbl_df$gender == "
    male", ]
# 女性だけのデータ
female_df <- smoking_tbl_df[smoking_tbl_df$gender ==
    "female", ]

# 男性の喫煙有無の集計
table(male_df$smoke)
prop.table(table(male_df$smoke))
# 女性の喫煙有無の集計
table(female_df$smoke)
prop.table(table(female_df$smoke))
```

層別の要約統計量

層ごとに、連続変数（例：年齢）の要約統計量を計算

```
# 男性の年齢の平均と標準偏差
mean(male_df$age, na.rm = TRUE)
sd(male_df$age, na.rm = TRUE)
# 女性の年齢の平均と標準偏差
mean(female_df$age, na.rm = TRUE)
sd(female_df$age, na.rm = TRUE)
```

演習問題

喫煙に関する変数について、以下を層別に集計・要約してください

1. gender ごとに喫煙スタイル（type）の度数を集計、gender ごとの割合を算出
2. gender ごとに週末の喫煙本数（amt_weekends）の平均と標準偏差を計算
3. 年齢を 40 歳未満, 40--64 歳, 65 歳以上に階級分けし、各層ごとに喫煙有無（smoke）の度数を集計、各年齢層ごとに割合を算出
4. 各年齢層ごとに平日の喫煙本数（amt_weekdays）の中央値を計算

☕演習：2 変数以上による層別 💻

演習問題：R コード例

```r
# 1. 性別ごとの喫煙スタイルの度数, 割合
male_df <- smoking_tbl_df[smoking_tbl_df$gender == "
    Male", ]
female_df <- smoking_tbl_df[smoking_tbl_df$gender ==
    "Female", ]
table(male_df$type)
prop.table(table(male_df$type))
table(female_df$type)
prop.table(table(female_df$type))

# 2. 性別ごとの週末の喫煙本数の平均・標準偏差
mean(male_df$amt_weekends, na.rm = TRUE)
sd(male_df$amt_weekends, na.rm = TRUE)
mean(female_df$amt_weekends, na.rm = TRUE)
sd(female_df$amt_weekends, na.rm = TRUE)
```

演習問題：R コード例（続き）

```r
# 3. 年齢層ごとにデータを分割
young_df <- smoking_tbl_df[smoking_tbl_df$age < 40, ]
middle_df <- smoking_tbl_df[smoking_tbl_df$age >= 40
    & smoking_tbl_df$age <= 64, ]
elderly_df <- smoking_tbl_df[smoking_tbl_df$age >=
    65, ]
# 各年齢層の喫煙有無の度数と割合
table(young_df$smoke)
prop.table(table(young_df$smoke))
table(middle_df$smoke)
prop.table(table(middle_df$smoke))
table(elderly_df$smoke)
prop.table(table(elderly_df$smoke))

# 4. 年齢層ごとの平日喫煙本数の中央値
median(young_df$amt_weekdays, na.rm = TRUE)
median(middle_df$amt_weekdays, na.rm = TRUE)
median(elderly_df$amt_weekdays, na.rm = TRUE)
```