第 3 章: 測定 (3.2. R で欠損データを扱う)

今井耕介 著

『社会科学のためのデータ分析入門 (QSS)』

2026-03-09

3.2 Rで欠損データを扱う

欠損値 (Missing Data) とは

- 調査において、回答者が質問に答えなかったり、回答を拒否したりした場合に生じる。
- R では、欠損値は `NA` (Not Available) という特別な記号で表される。
- **注意**: `NA` が含まれるデータに対してそのまま計算（平均など）を行うと、結果も `NA` になってしまう。

3.2.1 欠損値の確認

afghan データの読み込み (1)

- アフガニスタン調査データの読み込みを行います。

```
# 1. ローカルに保存した afghan データの読み込み（推奨）
afghan <- read.csv("afghan.csv")

#（参考）URL から直接読み込むことも可能
# afghan <- read.csv("https://ayumu-tanaka.github.io/QSS/QSS_Data/afghan.csv")
```

is.na() 関数 (2)

- データが NA かどうかを判定します。

```
# 1. 最初の 10 人分の収入データを確認
head(afghan$income, n = 10)
```

```
##  [1] "2,001-10,000"  "2,001-10,000"  "2,001-10,000"  "2,001-10,000"
##  [5] "2,001-10,000"  NA              "10,001-20,000" "2,001-10,000"
##  [9] "2,001-10,000"  NA
```

```
# 2. 欠損値かどうかを判定 (TRUE が欠損)
head(is.na(afghan$income), n = 10)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE
```

欠損値の数と割合の計算

- is.na() の結果（TRUE/FALSE）を sum() や mean() に渡すことで集計できます。

```
# 1. 欠損値の総数をカウント
sum(is.na(afghan$income))
```

```
## [1] 154
```

```
# 2. 全体に占める欠損値の割合を計算
mean(is.na(afghan$income))
```

```
## [1] 0.05591866
```

3.2.2 欠損値がある場合の計算

na.rm オプション

- 多くの統計関数には、NA を除外して計算するための na.rm = TRUE という引数があります。

```
# サンプルベクトルの作成
x <- c(1, 2, 3, NA)

# そのまま平均を計算すると NA になる
mean(x)
```

```
## [1] NA
```

```
# NA を除外して平均を計算
mean(x, na.rm = TRUE)
```

```
## [1] 2
```

集計表における欠損値の表示

- table() 関数で exclude = NULL を指定すると、NA も一つのカテゴリとして表示されます。

```
# 被害経験のクロス集計に NA も含める
# exclude = NULL を入れないと NA は無視される
table(ISAF = afghan$violent.exp.ISAF,
      Taliban = afghan$violent.exp.taliban,
      exclude = NULL)
```

```
##        Taliban
## ISAF       0    1 <NA>
##   0     1330  354   22
##   1      475  526   22
##   <NA>     7    8   10
```

3.2.3 欠損値の除外

na.omit() による一括除外

- **リストワイズ削除**: 1 つでも NA を含む行（回答者）をデータセット全体から削除します。

```
# 1. 1 つでも欠損値がある行をすべて削除した新しいデータセットを作成
afghan.sub <- na.omit(afghan)

# 2. 元の行数と、削除後の行数を比較
nrow(afghan)      # 元のデータ
```

```
## [1] 2754
```

```
nrow(afghan.sub) # 削除後のデータ
```

```
## [1] 2554
```

```
# 3. 特定の変数だけから NA を除く
length(na.omit(afghan$income))
```

```
## [1] 2600
```

3.2.4 まとめ

このセクションのまとめ

- **NA の理解**: R において欠損値は `NA` で表され、計算に影響を及ぼす。
- **確認と集計**:
  - `is.na()` で確認し、`sum()` や `mean()` で量を確認する。
  - `table(..., exclude = NULL)` で回答拒否なども含めた全体像を把握する。
- **処理方法**:
  - `mean(..., na.rm = TRUE)` のように、その都度除外する。
  - `na.omit()` でデータセットから欠損のある行を一括削除する（分析対象が減ることに注意）。
- **R の操作**: `is.na()`, `na.rm`, `na.omit()`, `nrow()`。