

第 2 章 中央値

2 中央値

2.1 準備

2.1.1 必要なライブラリをインストールする

下の行は console で実行すると良い。

```
# code
install.packages("tidyverse")
```

2.1.2 必要なライブラリを読み込む

```
# code
library(tidyverse)
library(systemfonts)
library(ragg)
```

2.2 中央値を求める

2.2.1 中央値とは何か

「中央値とは文字どおり、データの「真ん中」にある値という意味で、データを小さい順（もしくは大きい順）に並べたときに、ちょうど真ん中にくる値です。」（テキスト 13）

2.2.2 データを入力する

ここでは、CSV ファイルを作成してアップロードして使用する。

CSV ファイル名の例: zenki_chap_02_01_utf8.csv

ファイルの中身

```
car_id,year_nen
1,11
2,1
```

3,10

(以下省略)

2.3 データフレームの作成

2.3.1 read_(csv) の使用

```
# code
df_eigyousha <- read_csv("zenki_chap_02_01_utf8.csv")

## Rows: 16 Columns: 2
## -- Column specification -----
## Delimiter: ","
## dbl (2): car_id, year_nen
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

df_eigyousha

## # A tibble: 16 x 2
##   car_id year_nen
##   <dbl>   <dbl>
## 1     1       11
## 2     2        1
## 3     3       10
## 4     4       12
## 5     5        2
## 6     6        3
## 7     7       10
## 8     8       11
## 9     9        2
## 10    10        1
## 11    11       12
## 12    12        2
## 13    13       11
## 14    14       12
## 15    15        8
## 16    16       11
```

2.3.2 データの確認

```
# code
glimpse(df_eigyousha)

## Rows: 16
## Columns: 2
## $ car_id    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
## $ year_nen  <dbl> 11, 1, 10, 12, 2, 3, 10, 11, 2, 1, 12, 2, 11, 12, 8, 11
```

2.3.3 変数の型の確認

```
# code
class(df_eigyousha)

## [1] "spec_tbl_df" "tbl_df"      "tbl"        "data.frame"
```

tibble であることがわかります。

2.3.4 平均値の確認

```
# code
mean(df_eigyousha$year_nen)
```

```
## [1] 7.4375
```

2.4 中央値を求める

2.4.1 median()

中央値の関数は `mdeian()` です。Excel でも `MEDIAN()` を使いますので同じ名前です。

`median()` を使う場合。

```
# code
median(df_eigyousha$year_nen)
```

```
## [1] 10
```

2.4.2 summarize()

`summarize()` を使うこともできます。これは中央値を出すための関数ではありません。

```
# code
df_eigyousha |>
  summarise(
    median_value = median(year_nen)
  )
```

```
## # A tibble: 1 x 1
##   median_value
##         <dbl>
## 1           10
```

平均値と中央値を並べて表示してみましょう。

```
# code
df_eigyousha |>
  summarise(
    mean_value = mean(year_nen),
    median_value = median(year_nen)
  )
```

```
## # A tibble: 1 x 2
##   mean_value median_value
##         <dbl>         <dbl>
## 1      7.44          10
```

2.5 章末問題

2.5.1 問題

以下のデータは、ある部品の単価（円）を示したものです。中央値を求めてください。

520、480、720、890、490、980、1500

2.5.2 解答例

2.5.3 ベクトルを作成する場合

```
# code
data_buhin_tanka <- c(520, 480, 720, 890, 490, 980, 1500)
```

中央値を計算する。

```
# code
median(data_buhin_tanka)
```

```
## [1] 720
```

2.5.4 データフレームを作成する場合

```
# code
df_buhin_tanka <- tibble(
  tanka_yen = c(520, 480, 720, 890, 490, 980, 1500)
)
df_buhin_tanka
```

```
## # A tibble: 7 x 1
##   tanka_yen
##   <dbl>
## 1      520
## 2      480
## 3      720
## 4      890
## 5      490
## 6      980
## 7     1500
```

中央値を計算する。

```
# code
median(df_buhin_tanka$tanka_yen)
```

```
## [1] 720
```

2.5.5 CSV ファイルからデータフレームを作成する場合

CSV ファイル名の例: zenki_chap_02_shoumatu_utf8.csv

CSV ファイルの中身

```
tanka_yen
520
480
720
（以下省略）
```

データフレームの作成

```
# code
df_buhin_tanka2 <- read_csv("zenki_chap_02_shoumatu_utf8.csv")

## Rows: 7 Columns: 1
## -- Column specification -----
## Delimiter: ","
## dbl (1): tanka_yen
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
df_buhin_tanka2
```

```
## # A tibble: 7 x 1
##   tanka_yen
##   <dbl>
## 1      520
## 2      480
## 3      720
## 4      890
## 5      490
## 6      980
## 7     1500
```

中央値の計算

```
# code
median(df_buhin_tanka2$tanka_yen)
```

```
## [1] 720
```

2.6 テキストと参考文献

2.6.1 テキスト

- 玄場公規, 湊宣明, 豊田裕貴, 2016, 『Excel で学ぶビジネスデータ分析の基礎ビジネス統計スペシャリスト・エクセル分析ベーシック対応』, オデッセイコミュニケーションズ.

2.6.2 参考文献

- ウィットカム, 2024, 『R ではじめるデータサイエンス第2版』, オライリー・ジャパン.