確率と統計

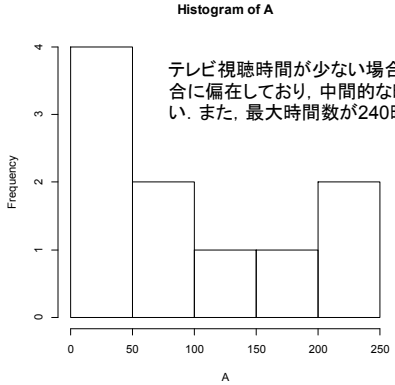
中山クラス
第4週

0

（1）大学ごとのヒストグラム

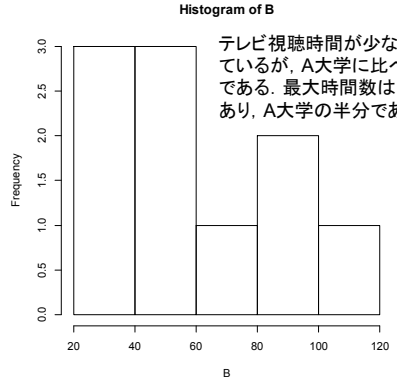
```
> A <-c(60,100,50,40,50,230,120,240,200,30)
> A
 [1]  60 100  50  40  50 230 120 240 200  30
> hist(A)

> B <- c(50,60,40,50,100,80,30,20,100,120)
> B
 [1]  50  60  40  50 100  80  30  20 100 120
> hist(B)
```

2



テレビ視聴時間が少ない場合と多い場合に偏在しており，中間的な時間が少ない．また，最大時間数が240時間と長い．

3



テレビ視聴時間が少ない方に偏っているが，A大学に比べると平坦である．最大時間数は120時間であり，A大学の半分である．

4

（2）大学ごとの平均，標準偏差

```
> A_mean <- mean(A)
> A_mean
[1] 112  ･･･答え
> B_mean <- mean(B)
> B_mean
[1] 65  ･･･答え
> A_var <- var(A)*(length(A)-1)/length(A)
> A_var
[1] 6056
> B_var <- var(B)*(length(B)-1)/length(B)
> B_var
[1] 1005
> A_std <- sqrt(A_var)
> A_std
[1] 77.82031  ･･･答え
> B_std <- sqrt(B_var)
> B_std
[1] 31.70173  ･･･答え
```

5

A大学の平均＝112
B大学の平均＝ 65

A大学の分散＝6056
B大学の分散＝1005

A大学の標準偏差＝77.82
B大学の標準偏差＝31.70

ヒストグラムからも分かるように，A大学では視聴時間が長い学生がおり，かつ，幅広く分布している．これに対してB大学では全体的に視聴時間が短く，かつ，分布範囲も狭い．これらが，平均と標準偏差に表れている．

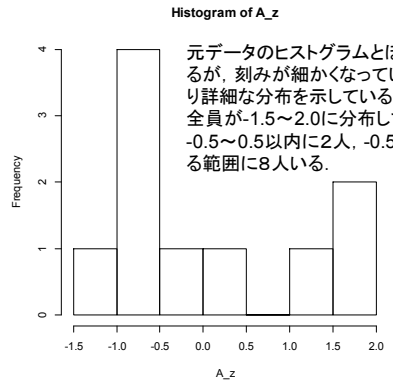
6

（3）大学ごとの標準化

標準化の一つとしてz得点を求める.

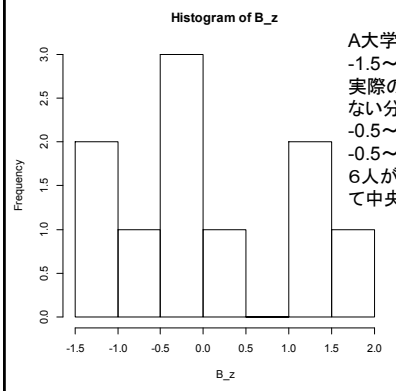
```
> A_z <- (A-A_mean)/A_std
> A_z
 [1] -0.6682061 -0.1542014 -0.7967072 -0.9252084
 -0.7967072  1.5163138
 [7]  0.1028009  1.6448149  1.1308103 -1.0537096

> B_z <- (B-B_mean)/B_std
> B_z
 [1] -0.4731602 -0.1577201 -0.7886004 -0.4731602
1.1040405  0.4731602
 [7] -1.1040405 -1.4194807  1.1040405  1.7349208
```

7



元データのヒストグラムとほぼ同じであるが，刻みが細かくなっているので，より詳細な分布を示している.
全員が-1.5～2.0に分布している.
-0.5～0.5以内に2人，-0.5～0.5を超える範囲に8人いる.

8



A大学と同様に全員が-1.5～2.0に分布しており，実際の時間数に左右されない分布状況が分かる.
-0.5～0.5以内に4人，
-0.5～0.5を超える範囲に6人がおり，A大学に比べて中央に分布している.

9

ヒストグラムの範囲と刻みの指定方法

```
hist(aaa, breaks=seq(10,100,5))
```

10～100の範囲で5刻みのヒストグラムを作図
10～100は全てのデータを含むように指定
（例）データ分布：　5～125
　　範囲指定：　5～125，0～130など

10

第9章　データフレーム

Rで利用されるデータの保存形式
数値や文字など異なるタイプのデータを扱うことができる

◆データフレームの作成法
Excelで表を作成（数値，文字混在）
csvファイルとして保存→aaa.csv
txtファイルとして保存 →bbb.txt

read.csv(“aaa.csv”)　表題（ヘッダー）があることが前提
read.csv(“aaa.csv”, header=FALSE)　ヘッダーがない場合

read.table(“aaa.csv”)　ヘッダーがないことが前提
read.table(“aaa.csv”, header=TRUE)　ヘッダーがある場合

read.table(bbb.txt”)　txtファイルはread.table( )を使用

14

テキストファイル（table.txt）

No	名前	性別	数学	統計	心理テスト	統計テスト1	統計テスト2	指導法
1	大村	男	嫌い	好き	13	6	10	C
2	本多	男	嫌い	好き	14	10	13	B
3	杉内	女	好き	嫌い	6	6	14	A

Excelで編集して，txtファイルとして保存する→作業ディレクトリ（推奨する）

適当なエディタで編集して，txtファイルとして保存する→作業ディレクトリ
　半角スペースで区切ること．全角スペースは文字と見なされる．
　改行は認識される．

15

```
> tbl <- read.table("table.txt")    ヘッダー無しが標準
> tbl
  V1  V2  V3  V4  V5       V6        V7        V8    V9
1 No. 名前 性別 数学 統計 心理テスト 統計テスト1 統計テスト2 指導法
2  1   大村   男  嫌い 好き       13         6         10         C
3  2   本多   男  嫌い 好き       14        10         13         B
4  3   杉内  女  好き  嫌い        6         6         14         A

> tbl_ht <- read.table("table.txt", header=TRUE)  ヘッダーがある場合
> tbl_ht
  No. 名前 性別 数学 統計 心理テスト 統計テスト1 統計テスト2 指導法
1  1 大村   男 嫌い 好き       13         6         10         C
2  2 本多   男 嫌い 好き       14        10         13         B
3  3 杉内   女 好き 嫌い        6         6         14         A
```

16

この頁以降の内容はcsvファイルを読み込んだ場合でも同じである.

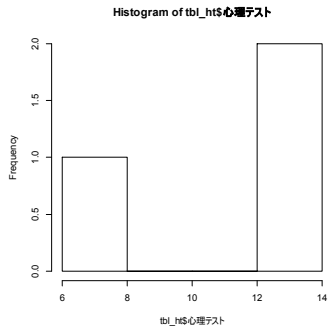
```
> table(tbl_ht$指導法)

A B C
1 1 1

> table(tbl_ht[,9])

A B C
1 1 1
```

17

```
> tbl_ht[2,]
  No. 名前 性別 数学 統計 心理テスト 統計テスト1 統計テスト2 指導法
2  2 本多   男   嫌い  好き       14        10         13         B

> tbl_ht[,4]
[1] 嫌い 嫌い 好き
Levels: 嫌い 好き
```

18

```
> hist(tbl_ht$心理テスト)
```



19

```
> mean(tbl_ht$統計テスト1)
[1] 7.333333

> var(tbl_ht$統計テスト2)
[1] 4.333333

> sd(tbl_ht$統計テスト1)
[1] 2.309401

> min(tbl_ht$統計テスト1)
[1] 6

> max(tbl_ht$統計テスト1)
[1] 10
```

20

```
> for (i in 6:8){print(mean(tbl_ht[,i]))}
[1] 11
[1] 7.333333
[1] 12.33333

> print("Rによるやさしい統計学")
[1] "Rによるやさしい統計学"
```

21

```
> Sta_z <- scale(tbl_ht$心理テスト)
> Sta_z
          [,1]
[1,]  0.4588315
[2,]  0.6882472
[3,] -1.1470787
attr(,"scaled:center")     平均
[1] 11
attr(,"scaled:scale")      標準偏差
[1] 4.358899
```

22

```
> tbl_ht$数学
[1] 嫌い 嫌い 好き
Levels: 嫌い 好き

> as.integer(tbl_ht$数学)
[1] 1 1 2

> cor(as.integer(tbl_ht$数学),as.integer(tbl_ht$統計) )
[1] -1
```

23

今日の自習＆演習

◆これまでの復習
◆第2回レポートの作成，質問
◆小テストの予想問題1，2の検討，質問
◆第9章：9.1～9.3の例題

24