

Rで学ぶ確率統計

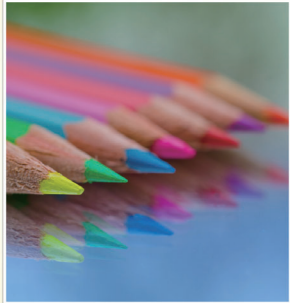
第2回 データの要約とレポート作成

1

Rで学ぶ確率統計

第2回
データの要約と
レポート作成

講師 秋光 淳生



2

第2回 データの要約とレポートの作成

- ① 表の作成
- ② グラフの作成
- ③ レポートの作成



3

①-1 表の作成

	勉強	結果
1	勉強した	合格
2	勉強した	不合格
3	勉強しなかった	合格
4	勉強した	不合格
5	勉強しなかった	不合格
	⋮	⋮

クロス集計表

	不合格	合格	
勉強した	50	100	150
勉強しなかった	25	25	50
	75	125	200

行の周辺度数 (row marginal totals) and 列の周辺度数 (column marginal totals) are indicated.

4

①-2 相対度数

	不合格	合格	計
勉強した	50	100	150
勉強しなかった	25	25	50
計	75	125	200

▼列の周辺度数による相対度数

	不合格	合格	計
勉強した	0.667	0.8	0.75
勉強しなかった	0.333	0.2	0.25
計	1	1	1

▼行の周辺度数による相対度数

	不合格	合格	計
勉強した	0.333	0.667	1
勉強しなかった	0.5	0.5	1
計	0.375	0.625	1

▼全体度数による相対度数

	不合格	合格	計
勉強した	0.25	0.5	0.75
勉強しなかった	0.125	0.125	0.25
計	0.375	0.625	1

単に割合と言っても何を知りたいかの目的に依存する

第2回 データの要約とレポート作成

5

①-3 多重クロス表

	勉強	結果	性別
1	勉強した	合格	男性
2	勉強した	不合格	男性
3	勉強しなかった	合格	女性
4	勉強した	不合格	女性
5	勉強しなかった	不合格	男性
	⋮	⋮	
	⋮	⋮	

性別	勉強	不合格	合格	計
女性	勉強した	3	2	5
	勉強しなかった	23	20	43
	計	26	22	48
男性	勉強した	47	98	145
	勉強しなかった	2	5	7
	計	49	103	152

6

①-4 シンプソンのパラドックス

性別	勉強	不合格	合格	計
女性	勉強した	0.6	0.4	1
	勉強しなかった	0.535	0.465	1
	計	0.542	0.458	1
男性	勉強した	0.324	0.676	1
	勉強しなかった	0.286	0.714	1
	計	0.322	0.678	1

	不合格	合格	計
勉強した	0.333	0.667	1
勉強しなかった	0.5	0.5	1
計	0.375	0.625	1

勉強したほうが合格率が高い

男性、女性ともに「勉強しなかった」ほうが合格する割合が大きい

7

①-5 表の作成

```
> b01 <- read_csv("A02L01.csv")
> b02 <- table( b01$study, b01$result)
または table( b01[,1:2])
```

ファイルの読み込み
表の作成
1:2 は c(1,2) でもよい

```
> prop.table( b02 )
> prop.table( b02, 1)
> prop.table( b02, 2)
```

なし 全体度数
1 行の周辺度数による
2 列の周辺度数による

```
> addmargins(b02)
> addmargins(b02, 1)
> addmargins(b02, 2)
```

なし 全体度数
1 列の周辺度数のみ計算
2 行の周辺度数のみ計算

```
table(b01[,1:2]) %>% prop.table () %>% addmargins()
```

8

②-1 グラフの作成(1)

グラフについて

1. 情報を伝える有効な手法

2. 用途にあった適切なグラフ

3. 必須項目(凡例等)

4. 作弄的なグラフ

5. 言葉でも表現

日付	値
3月21日	22
3月22日	22
3月23日	22
3月24日	24
3月25日	24

Rで学ぶ確率統計

第2回 データの要約とレポート作成

9

②-1 グラフの作成(2)

ggplot The Grammar of Graphics <https://r4ds.had.co.nz/data-visualisation.html>

```
> ggplot() + # +で続く
              # +のあとに改行
  geom_point() # 散布図
  geom_line()  # 折れ線グラフ
  geom_bar()   # 棒グラフ
              (円グラフ)
  geom_histogram() # ヒストグラム
  geom_box() +    # 箱ひげ図

  stat_smooth() + # 回帰曲線

  theme()        # 背景など
```

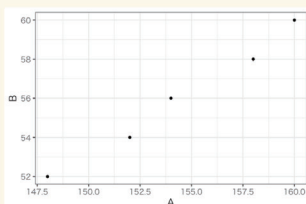
ggplotでのグラフ作成

1. ggplot (DATA) でデータを指定
2. geom 関数でグラフの種類を選ぶ
3. aes 関数で項目 (列) を指定する
4. レイヤーの重ね書きができる
5. オプションが多い
6. スクリプトを残しておく

10

②-2 散布図、折れ線グラフ(1)

```
> a01 <- read_csv("A02L02.csv")
> ggplot(a01, aes(x=A, y=B)) +
  geom_point()
```



```
> a01
# A tibble: 5 x 3
  name1 A B
<chr> <dbl> <dbl>
1 A1 148 52
2 A2 152 54
3 A3 154 56
4 A4 158 58
5 A5 160 60
```

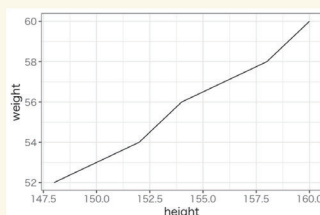
```
# a01 というデータ
# aes(要素) x座標が A列、y座標が B列

# ggplot(a01) + geom_point( aes(x=A, y=B) )
# スクリプトは + のあとに改行
```

11

②-2 散布図、折れ線グラフ(2)

```
> ggplot(a01, aes(x=A, y=B)) +
  geom_line() +
  labs(x="height", y="weight") +
  theme(text = element_text(size=16))
```



```
# geom_line で折れ線グラフ
# labs() で座標のラベルを設定
# データから自動で設定
# 列の名前を工夫しておく

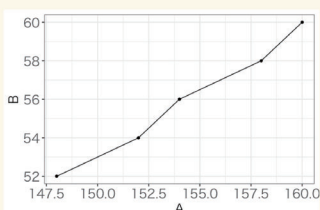
# theme で枠組みの色や軸の設定
#
# geom_line(color=2) は赤の線

# グラフはRStudio のExport で保存できる
# ggsave("ファイル名") でも保存可
# Rmarkdown でレポートにする
```

12

②-2 散布図、折れ線グラフ(3)

```
> ggplot(a01, aes(x=A, y=B)) +
  geom_point() + geom_line() +
  theme(text = element_text(size=16))
```



```
ggplot(データ) +
  geom_point(mapping = aes(要素)) +
  geom_line(mapping = aes(要素)) +
  theme(フォントなど)
```

レイヤーの重ね合わせ

```
theme()
+
geom_line()
+
geom_point()
+
ggplot(データ)
```

折れ線グラフ

散布図

座標系

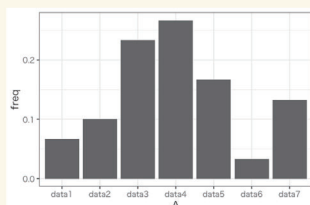
R で学ぶ確率統計

第2回 データの要約とレポート作成

13

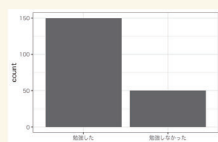
②-3 棒グラフ、積み上げ棒グラフ、円グラフ(1)

```
> c01 <- read_csv("A02L03.csv")
> ggplot(c01, aes(x=A, y=freq)) +
  geom_bar(stat="identity")
```



```
# geom_bar() のみ
# カテゴリーデータを集計する
```

```
> ggplot(b01, aes(x=study)) + geom_bar()
```

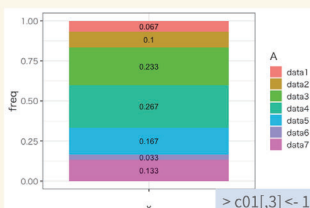


```
# 列の値をそのままグラフにする場合
# stat="identity"
```

14

②-3 棒グラフ、積み上げ棒グラフ、円グラフ(2)

```
> p1 <- ggplot(c01, aes(x="", y=freq, fill=A)) +
  geom_bar(stat="identity", position="fill") +
  geom_text(aes(label=freq, y=pos1))
> p1
```



```
> c01[,3] <- 1-(cumsum(c01$freq) / (0.5*c01$freq))
```

```
# 積み上げ棒グラフ
# stat="identity"
# position="fill"
```

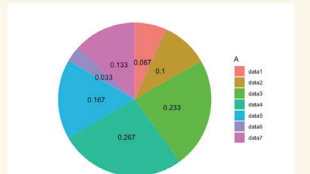
```
# A tibble: 7 x 3
#   A   freq pos1
#   <chr> <dbl> <dbl>
1 data1 0.067 0.966
2 data2 0.1   0.883
3 data3 0.233 0.716
4 data4 0.267 0.466
5 data5 0.167 0.249
6 data6 0.033 0.149
7 data7 0.133 0.0665
```

ラベルのy座標

15

②-3 棒グラフ、積み上げ棒グラフ、円グラフ(3)

```
> p2 <- p1 +
  coord_polar("y", direction=-1) +
  theme_void()
> p2
```



円グラフ：扇形の面積が大きさ

```
# 1要素の積み上げ棒グラフ
# p1 という積み上げ棒グラフを
# coord_polar で極座標に
```

```
# 色によって膨張してみえる
# 値の比較には不向き
# 内訳を見る
```

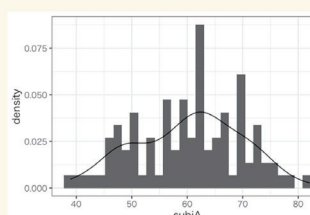
direction = 1 が反時計周り

theme_void() は枠を消す

16

②-4 ヒストグラム、密度グラフ

```
> c02 <- read_csv("A02L04.csv")
> ggplot(c02, aes(x=subjA, y=..density..)) +
  geom_histogram() + geom_density()
```



```
# A tibble: 100 x 2
#   subjA subjB
#   <dbl> <dbl>
1 49 56
2 60 52
3 59 69
4 65 63
5 70 39
```

ヒストグラム

y = ..density.. は密度関数用

何も無ければ y 軸は count

bins = 30 など階級数を変更できる

binwidth で階級幅

第2回 データの要約とレポート作成

② 今回はWord形式を選ぶ

③-2 レポートの作成(1)

HTML マークアップ
(Hyper Text Markup Language)
R Markdown

Rのコード、結果、注釈本文
文章制御

```

1 # Report.Rmd
2 ---
3 title: "学生履修簿統計"
4 author: "学生番号 1234567890", "氏名 松光淳生"
5 date: "2021/04/05"
6 output:
7   word_document:
8     reference_docx: sample21.docx
9 ---
10
11 この中身はコメントとして処理では無視される
12 下の行は R の chunk のメモ
13
14 ```{r setup, include=FALSE}
15 knitr::opts_chunk$set(echo = TRUE)
16 ```
17
18 # レポート
19
20 レポートはRmdで作成し、__Word__ 形式で変換する
21
22 <--
23 ## は変出し
24
25
26 # もう一つの例は変出し
  
```

3-2 レポートの作成(2)

保存すると
タプルファイル名

Knit ボタンを押すと変換

Knit ボタンを押すと変換

HTML マークアップ
(Hyper Text Markup Language)
R Markdown

この部分はヘッダ(タイトル等)

R のコード、結果、注釈本文
文章制御

` ` で囲んだ部分：チャック
チャック内：#行：コメント
チャックに名前をつけられる

[r] (` ` は Shift + @)
で囲まれた部分が R のコード

R のコード横の
実行を確認できる

は段落タイトル
まとまりごとに1行空ける

③ー2 レポートの作成(3)

✎で学ぶ線形統計

レポート

レポートは Rmd で作成し、Word 形式で提出する

```
summary(cars)
```

```
##
##      speed          dist
##  Mean:   15.4      Min:    2.00
##  1st Qu: 12.00     1st Qu: 26.00
##  Median: 15.0     Median: 36.00
##  Mean:   15.4      Mean:   42.98
##  3rd Qu: 17.0     3rd Qu: 52.00
##  Max.:   25.0      Max.: 128.00
```

もう一つ小さい見出し

このように小さい見出しも作成できる

```
## 文字の大きさなどレイアウトは今ひとつ
##
## R Markdown を編集し直すことがあり
## 変換後の Word は編集できない
##
## 完成後にコピーを別名で保存すれば
## レイアウトなどを編集できる

## スクリプトを作成しておいてコピーする
```

学生番号 1234567890

氏名 秋田洋生

2021/04/05

Rで学ぶ確率統計

第2回 データの要約とレポート作成

21

③-2 レポートの作成(4)

タイトル等で設定した項目

Rで学ぶ確率統計

学生番号 1234567890

氏名 秋元淳生

2021/04/05

段落タイトル

レポート

レポートは Rmd で作成し、Word 形式で、その下に実行結果が表示されている

summary(cars)

speed dist
Min.: 7.4 Min.: 2.89
1st Qu.:12.0 1st Qu.: 26.86
Median: 15.4 Median: 35.89
Mean: 15.4 Mean: 42.98
3rd Qu.:19.0 3rd Qu.: 55.89
Max.: 25.0 Max.: 128.09

もう一つ小さい見出し

このように小さい見出しも作成できる

summary(cars) という R のコードがあり、データを要約する summary() 関数を実行

R にもともと cars というデータがあり、データを要約する summary() 関数を実行

文字の大きさなどレイアウトは今ひとつ
R Markdown を編集し直すことがあり
変換後の Word は編集できない
完成後にコピーを別名で保存すれば
レイアウトなどを編集できる
スクリプトを作成しておいてコピーする