
データの分析と知識発見
Introduction to Data Analysis

今回の構成

尺度水準について学ぶ

クロス集計表について学ぶ

Rを用いてクロス集計表を作る

尺度

名義尺度と順序尺度

名義尺度の例(選択肢)

次の手法の中で、分散共分散行列
または相関行列をもとに
固有値・固有ベクトルを計算し、
新しくより少ない成分でデータを
表すための手法はどれか？

1. 主成分分析
2. 回帰分析
3. クラスター分析
4. 判別分析

異なるものを区別し、
分類するために用いる

順序尺度の例

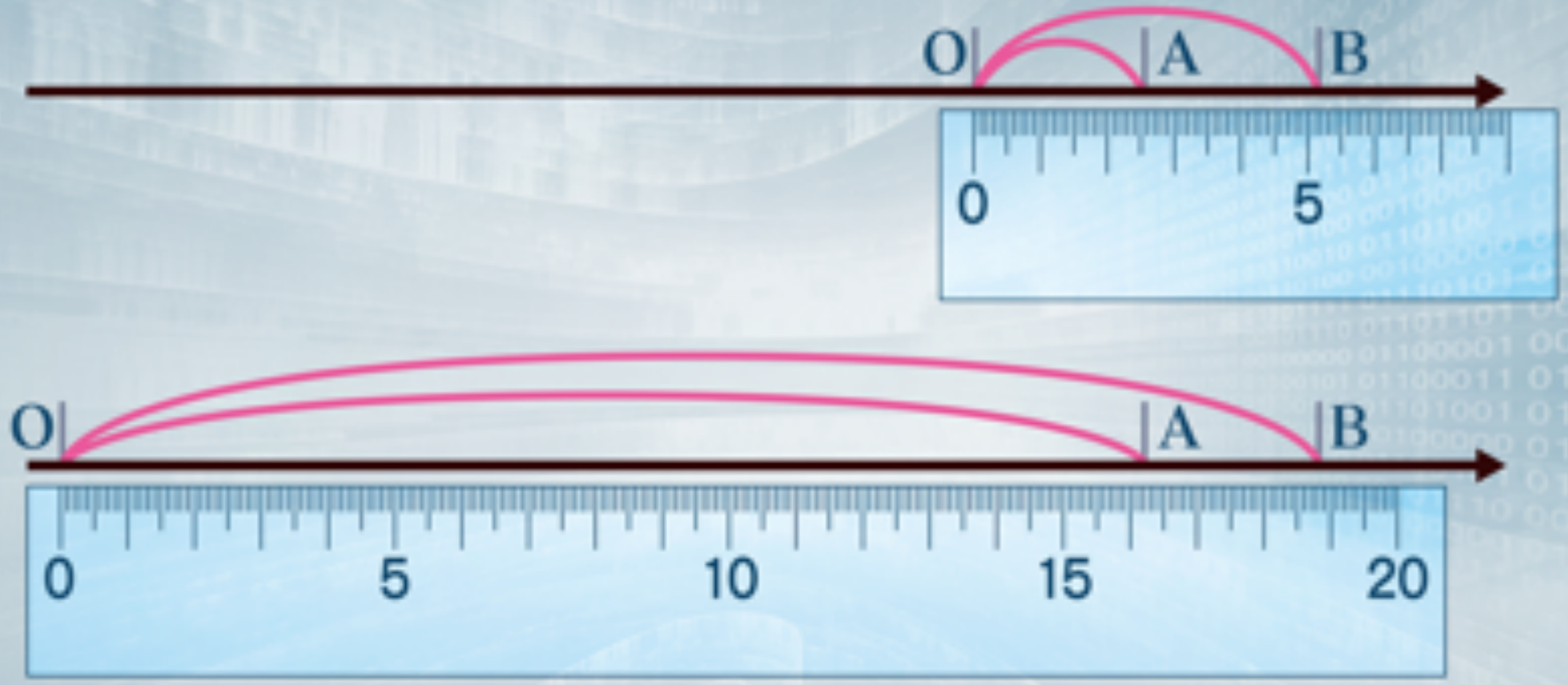
成績一覧	1位	A	100点	50点
	2位	B	50点	10点
	3位	C	40点	30点
	4位	D	10点	

コンピューターを操作することは

嫌い	嫌い	普通	好き	好き
	どちらかというと		どちらかというと	
1	2	3	4	5

間隔が同じとは限らない

間隔尺度と比例尺度



クロス集計表(1)

	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	合格
	⋮	⋮

単純集計

勉強した	勉強しなかった
150	50

合格	不合格
125	75

クロス集計表(2)

	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	合格
	⋮	⋮

クロス集計

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

クロス集計表(2)

	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	合格
	⋮	⋮

クロス集計

(行の)周辺度数

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

(列の)周辺度数

全体度数

クロス集計表(3)

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

(行の)周辺度数による相対度数

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

(列の)周辺度数による相対度数

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

全体度数による相対度数

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

多重クロス表

年齢	勉強	合格	不合格	合計
40歳未満	勉強した	0.676	0.324	1
	勉強しなかった	0.714	0.286	1
	合計	0.678	0.322	1
40歳以上	勉強した	0.400	0.600	1
	勉強しなかった	0.465	0.535	1
	合計	0.458	0.542	1

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

シンプソンのパラドックス

年齢	勉強	合格	不合格	合計
40歳未満	勉強した	0.676	0.324	1
	勉強しなかった	0.714	0.286	1
	合計	0.678	0.322	1
40歳以上	勉強した	0.400	0.600	1
	勉強しなかった	0.465	0.535	1
	合計	0.458	0.542	1

40歳未満も40歳以上もどちらも
勉強しなかった方が合格している割合が高い

シンプソンのパラドックス

年齢	勉強	合格	不合格	合計
40歳未満	勉強した	0.676	0.324	1
	勉強しなかった	0.714	0.286	1
	合計	0.678	0.322	1
40歳以上	勉強した	0.400	0.600	1
	勉強しなかった	0.465	0.535	1
	合計	0.458	0.542	1

40歳未満も40歳以上もどちらも
勉強しなかった方が合格している割合が高い

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

勉強した方が
合格している割合が高い

クロス集計表における指標(1)

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

ファイ係数

$$\phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

x \ y		項目2		
		y_1	y_2	合計
項目1	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
	合計	$a+c$	$b+d$	$a+b+c+d$

ユール係数

$$Q = \frac{ad-bc}{ad+bc}$$

クロス集計表における指標(2)

x \ y		項目2						
		y_1	y_2	$\cdots$	y_j	$\cdots$	y_n	合計
項目1	x_1	a_{11}	a_{12}	$\cdots$	a_{1j}	$\cdots$	a_{1n}	$a_{1\cdot}$
	x_2	a_{21}	a_{22}	$\cdots$	a_{2j}	$\cdots$	a_{2n}	$a_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\ddots$		$\vdots$	$\vdots$
	x_i	a_{i1}	a_{i2}	$\cdots$	a_{ij}	$\cdots$	a_{in}	$a_{i\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\ddots$		$\vdots$	$\vdots$
	x_m	a_{m1}	a_{mn2}	$\cdots$	a_{mj}	$\cdots$	a_{mn}	$a_{n\cdot}$
	合計	$a_{\cdot 1}$	$a_{\cdot 2}$	$\cdots$	$a_{\cdot j}$	$\cdots$	$a_{\cdot n}$	$a_{\cdot\cdot}$

期待度数

$$\hat{a}_{ij}=\frac{a_{\cdot j}}{a_{\cdot\cdot}}\times a_{i\cdot}=\frac{a_{\cdot j}\cdot a_{i\cdot}}{a_{\cdot\cdot}}$$

カイ2乗値

$$\chi^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(a_{ij}-\hat{a}_{ij})^2}{\hat{a}_{ij}}$$

クロス集計表における指標(3)

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

期待度数

$$\hat{a}_{ij}=\frac{a_{\cdot j}}{a_{\cdot\cdot}}\times a_{i\cdot}=\frac{a_{\cdot j}\cdot a_{i\cdot}}{a_{\cdot\cdot}}$$

勉強	合格	不合格	合計
勉強した	93.75	56.25	150
勉強しなかった	31.25	18.75	50
合計	125	75	200

カイ2乗値

$$\chi^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(a_{ij}-\hat{a}_{ij})^2}{\hat{a}_{ij}}$$

$$(100-93.75)^2/100$$
$$+(50-56.25)^2/56.25$$
$$+(25-31.25)^2/31.25$$
$$+(25-18.75)^2/18.75$$
$$=4.44\cdots$$

シンプソンのパラドックス

	勉強	結果	年齢
1	勉強した	合格	40歳未満
2	勉強 しなかった	不合格	40歳未満
3	勉強 しなかった	合格	40歳以上
4	勉強した	不合格	40歳以上
5	勉強 しなかった	合格	40歳未満

年齢	勉強	合格	不合格	合計
40歳 未満	勉強した	98	47	145
	勉強 しなかった	5	2	7
	合計	103	49	152
40歳 以上	勉強した	2	3	5
	勉強 しなかった	20	23	43
	合計	22	26	48

Rによる表の作成

テーブルの作成

```
> test1 <- read.csv("ch041.csv",header=T,row.names=1,fileEncoding="UTF-8")
> table1 <- table (test1$study,test1$result)
> table2 <- addmargins (table1)
```

相対度数の計算

```
> prop.table (table1, 1)
> addmargins (prop.table (table1, 1), 2)
```

Rによる表の作成

テーブルの作成

```
> test1 <- read.csv("ch041.csv",header=T,row.names=1,fileEncoding="UTF-8")
> table1 <- table (test1$study,test1$result)
> table2 <- addmargins (table1)
```

1：行の相対度数
2：列の相対度数
何も指定しない：全体度数

相対度数の計算

```
> prop.table (table1,1)
> addmargins (prop.table (table1,1), 2)
```

1：列の合計
2：行の合計
何も指定しない：両方の合計