

Rによるクラスター分析

- Rによるクラスター分析

- 手順①(データセットを読み込む)
- 手順②(分類)

- 階層的クラスター分析

- (1) `dist(<データセット>, method="<データ間距離の定義>")`

- でデータ間距離を計算

- (2) `hclust(<(1)の結果*>, method="<クラスターとの距離の定義>")`

- で近いもの同士の結合を繰り返す

- ※デンドログラムを描画するには、`plot(<(2)の結果>, hang=-1>`

- ※k個のクラスターへの分類結果を見るには、`cutree(<(2)の結果>, k=<クラスター数>)`

- 非階層的クラスター分析(kmeans法)

- `kmeans(<データセット>, <クラスター数>, nstart=<試行回数>)`

- ※分類結果を見るには、上の結果に\$clusterを付けて見る

- 手順③(各クラスターの特徴の解釈)

仮想データセット

	A	B	C	D	E	F
1	id	eigo	sugaku	kokugo	rika	syakai
2	1	77	73	87	58	71
3	2	78	80	91	64	72
4	3	65	74	78	46	64
5	4	71	98	76	65	58
6	5	63	82	81	49	67
7	6	72	85	80	55	66
8	7	65	77	85	53	62
9	8	73	58	77	40	60
10	9	74	79	82	45	63
11	10	69	71	93	47	75
12	11	70	68	74	50	65
13	12	79	100	89	61	74
14	13	52	48	68	43	53
15	14	67	53	71	39	57
16	15	65	76	84	52	61
17	16	76	60	86	41	73
18	17	62	56	83	44	68
19	18	75	72	73	51	69
20	19	69	81	94	48	77
21	20	69	64	70	42	59

Rによる階層的クラスタ分析

*ワード法の場合は注意(後述)

- 事前準備
とくになし
- 手順
 - ① データセットを読み込む
 - ② 分類
 - (1) `dist(<データセット>, method="<データ間距離の定義>")`
でデータ間距離を計算
 - (2) `hclust(<(1)の結果*>, method="<クラスターとの距離の定義>")`
で近いもの同士の結合を繰り返す

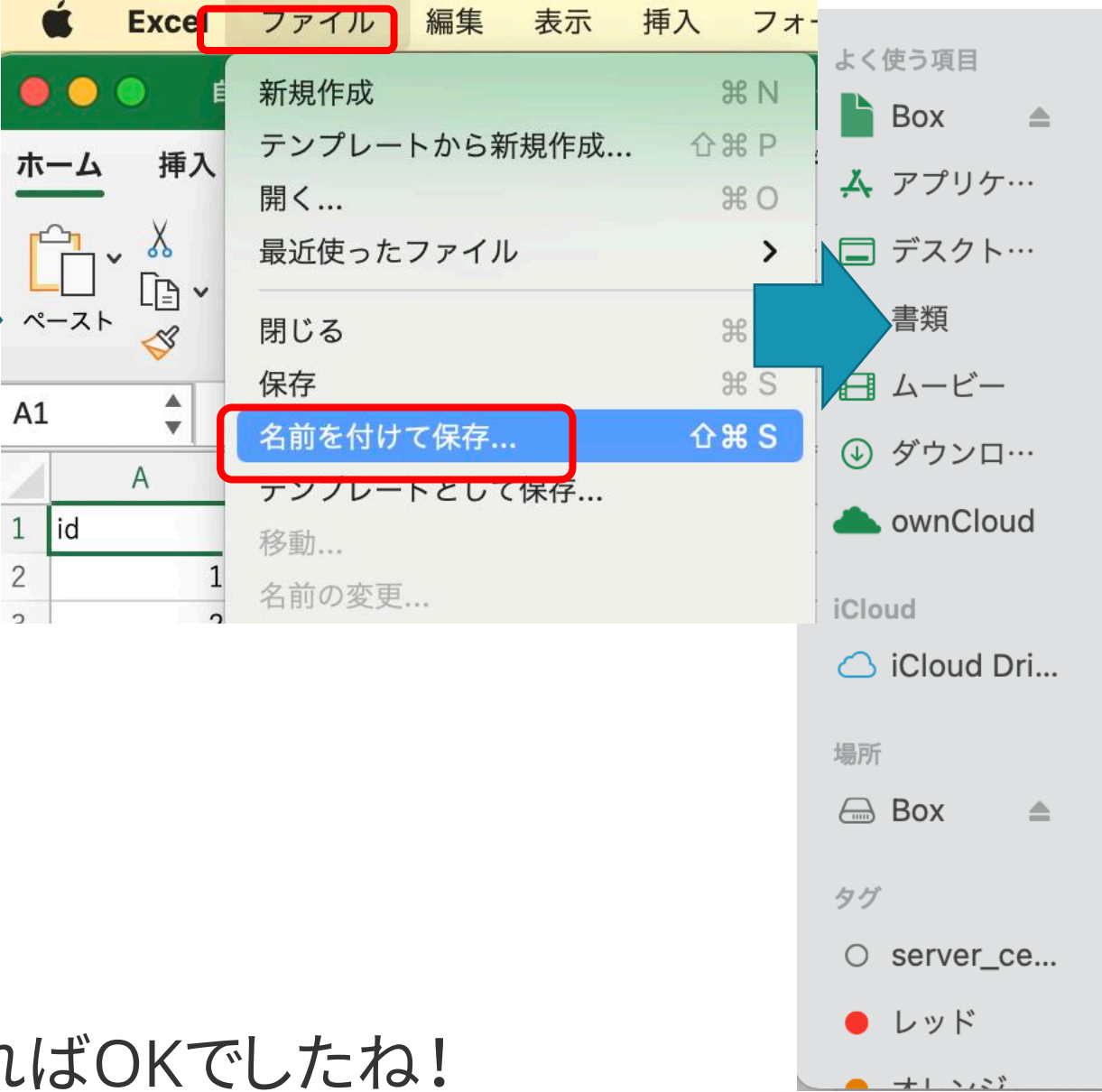
※デンドログラムを描画するには、`plot(<(2)の結果>, hang=-1>`


※k個のクラスターへの分類結果を見るには、`cutree(<(2)の結果>, k=<クラスター数>)`
 - ③ 各クラスターの特徴の解釈

手順① (データセットを読み込む)

EXCELで先ほどの仮想データセットをつかって、
csv形式で、ワーキングディレクトリ(作業用フォルダ)に保存しましょう

	A	B	C	D	E	F
1	id	eigo	sugaku	kokugo	rika	syakai
2	1	77	73	87	58	71
3	2	78	80	91	64	72
4	3	65	74	78	46	64
5	4	71	98	76	65	58
6	5	63	82	81	49	67
7	6	72	85	80	55	66
8	7	65	77	85	53	62
9	8	73	58	77	40	60
10	9	74	79	82	45	63
11	10	69	71	93	47	75
12	11	70	68	74	50	65
13	12	79	100	89	61	74
14	13	52	48	68	43	53
15	14	67	53	71	39	57
16	15	65	76	84	52	61
17	16	76	60	86	41	73
18	17	62	56	83	44	68
19	18	75	72	73	51	69
20	19	69	81	94	48	77
21	20	69	64	70	42	59





名前: **testScore.csv**

タグ:

場所: **DL3**

ファイル形式: **CSV (コンマ区切り) (.csv)**

ボタン: 新規フォルダ, キャンセル, **保存**

ワーキングディレクトリに保存する

※注意: 環境によっては、日本語は文字化けします

あとは、下記を実行すればOKでしたね!

```
dat <- read.csv("testScore.csv", header=T, row.names=1, na.strings=".")
```

1行目は列ラベルとして読み込むよう指定

1列目は行ラベルとして読み込むよう指定

半角ピリオドを欠損値として読み込むよう指定

今回も欠損値はありませんが、クセにしておくといはいます
row.names=1は指定しなくても今回は動きます。

手順②(分類)

分類に使用する変数だけが含まれている必要あり

(1) `dist(<データセット>, method="<データ間距離の定義>")`でデータ間距離を計算

*通常、`method="euclidean"`として、ユークリッド距離を指定します。p個の変数があるとき、i番目とj番目のデータのユークリッド距離は、 $d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2}$ で求めます。
その他、マンハッタン距離 (manhattan)、最長距離 (maximum)、ミンコフスキー距離 (minkowski)、キャンベラ距離 (canberra) などが指定できます (青木、2009)。

```
> ddd <- dist(dat, method="euclidean")
> round(ddd, digits=2)
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2	10.15																		
3	20.47	27.60																	
4	31.62	28.20	31.83																
5	20.25	24.06	9.75	26.12															
6	15.87	17.29	16.09	18.71	11.31														
7	16.43	20.86	10.54	26.80	9.27	12.57													
8	28.04	37.75	19.31	47.27	28.67	31.62	25.73												
9	17.41	23.24	11.14	28.83	12.77	12.37	12.61	22.38											
10	15.52	21.54	19.29	40.44	19.21	22.78	17.92	26.74	18.92										
11	18.52	27.24	9.70	34.34	17.29	18.84	15.65	15.59	15.17	21.91									
12	27.48	20.45	36.30	22.56	28.93	21.33	30.81	50.80	29.87	33.97	39.14								
13	46.48	55.05	32.79	58.60	40.96	47.18	38.47	26.08	41.77	44.08	30.87	68.11							
14	36.24	45.83	24.33	52.37	34.01	38.30	31.58	10.34	30.30	34.64	20.69	58.74	17.06						
15	17.26	22.34	9.22	27.60	9.70	13.42	2.00	24.12	12.12	18.52	14.46	32.36	36.80	29.78					
16	21.54	30.97	22.07	48.68	27.89	30.36	26.00	16.25	22.29	16.09	19.72	44.93	38.05	24.80	25.42				
17	27.11	36.22	19.44	49.35	26.59	32.79	23.90	15.52	26.46	20.78	18.28	50.85	24.80	18.00	22.87	15.97			
18	15.94	23.98	13.38	31.91	17.78	15.87	17.94	20.45	14.25	22.11	7.68	34.37	38.08	26.78	17.38	20.74	23.98		
19	17.69	19.29	22.23	35.59	17.52	19.77	19.05	34.45	19.44	10.30	26.80	25.77	51.53	42.40	20.32	24.88	29.87	25.12	
20	28.88	37.83	14.87	41.55	24.37	27.71	23.24	10.30	20.47	29.31	11.53	48.40	24.21	11.79	21.45	22.76	19.16	17.03	35.00

例えばこれは、id=13とid=19
の人のユークリッド距離

(2) `hclust(<(1)の結果>, method="<クラスターとの距離の定義>")` で近いもの同士の結合を繰り返す。

*通常、`method="average"`とする群平均法、または`method="ward.D"`とするウォード法がよく使われます。

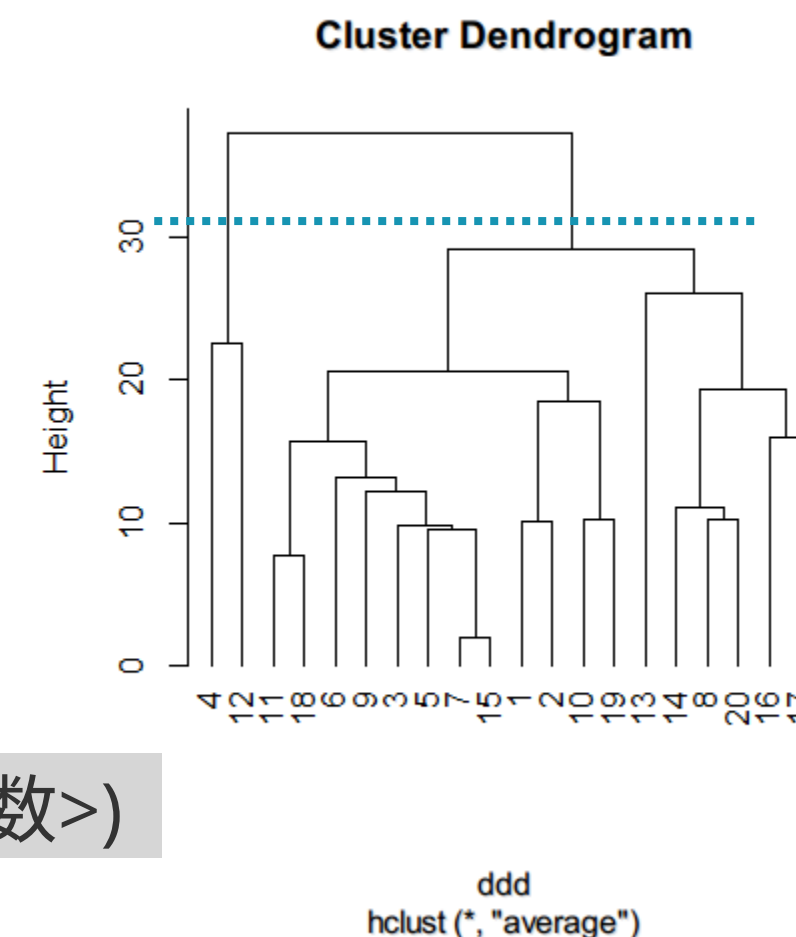
ただし、ウォード法は、距離の入力の仕方に注意が必要です(後述)。その他、最短距離法(single)、最長距離法(complete)なども指定できます(青木、2009)。

※デンドログラムを描画するには、`plot(<(2)の結果>, hang=-1>`

```
> ddd <- dist(dat, method="euclidean")
> result <- hclust(ddd, method="average")
```

hang=-1を付けないとidラベル
がガタガタになって見にくくなる

上記を実行すると、右図が表示される



※k個のクラスターへの分類結果を見るには、`cutree(<(2)の結果>, k=<クラスター数>)`

```
> cutree(result, k=2)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 1  1  1  2  1  1  1  1  1  1  1  2  1  1  1  1  1  1  1  1
```

例えば、id=1の人は、
クラスター1に割り振られたと読む

手順③ (各クラスターの特徴の解釈)

手順②で分類されたクラスター間で、分類に使用した変数の平均値を比較すれば良いのでしたね！
anovakunを使えば、群ごとの平均値をみることができますね！

分類結果のクラスターは「対応のない要因」で、分類に使用した変数は「対応のある要因」になるので、
2要因混合計画 (AsB) を指定すればOKです。

```
> dat$cluster <- cutree(result, k=2)
> source("anovakun_485.txt")
> anovakun(dat[c(6, 1:5)], "AsB", 2, 5)
```

clusterという新しい列(変数)を作成して分類結果を代入

anovakunは、対応のない要因の列が最前列になるようにしておく必要があったので、6列目(cluster)が先頭になるよう並べ替えています。

[AsB-Type Design]

This output was generated by anovakun 4.8.5 under R version 3.6.3.
It was executed on Thu Jun 24 15:59:04 2021.

<< DESCRIPTIVE STATISTICS >>

A	B	n	Mean	S.D.
a1	b1	18	68.9444	6.4486
a1	b2	18	69.8333	10.9504
a1	b3	18	80.9444	7.7950
a1	b4	18	48.1667	6.6177
a1	b5	18	65.6667	6.4807
a2	b1	2	75.0000	5.6569
a2	b2	2	99.0000	1.4142
a2	b3	2	82.5000	9.1924
a2	b4	2	63.0000	2.8284
a2	b5	2	66.0000	11.3137

a1: クラスター1
a2: クラスター2

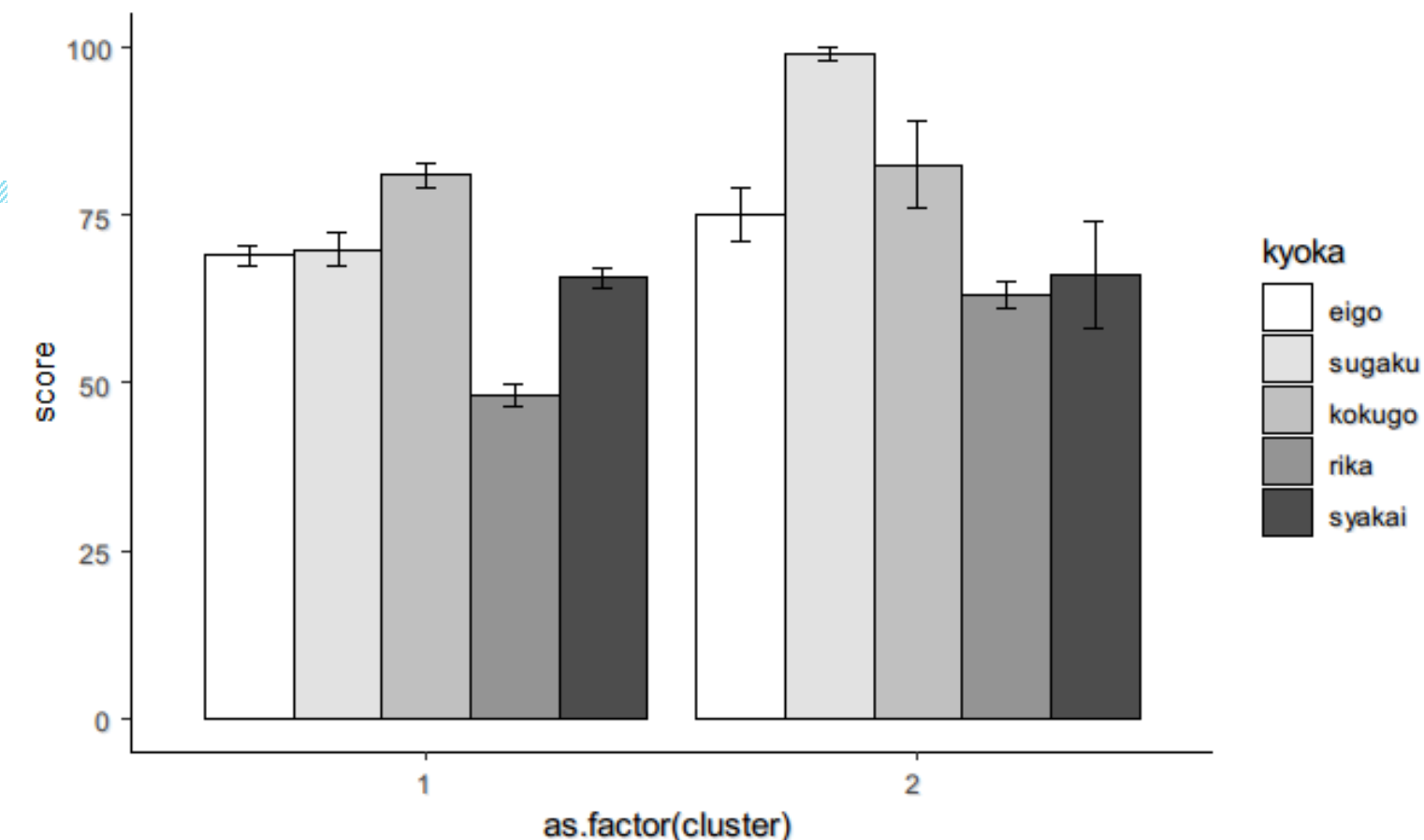
b1:eigo
b2:sugaku
b3:kokugo
b4:rika
b5:syakai

(復習) 2要因混合計画 (AsB) のデータセットの並べ方

		イカアン		ボスビッグ		ビビッテル	
		A	B	C	D	E	
1	id	A		b1	b2	b3	
2		1	a1	冷蔵庫	6	10	11
3		2	a1		4	8	12
4		3	a1		5	10	12
5		4	a1		3	8	10
6		5	a1		2	9	10
7		6	a2	常温	5	7	12
8		7	a2		4	6	8
9		8	a2		2	5	5
10		9	a2		2	4	6
11		10	a2		2	3	4

棒グラフにできれば、わかりやすいですね。
分散分析のときの棒グラフの作図用スクリプトが使えます。
余力のある人はぜひやってみてください。

anovakunの検定結果とあわせてみると、
数学と理科に群間差がみられるので、
1:理系科目不得意群、2:理系科目得意群と解釈できそうです。



```
#作図
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('Rmisc')) install.packages('Rmisc'); library('Rmisc')
if (!require('reshape2')) install.packages('reshape2'); library('reshape2')

dat.long <- melt(dat, id.vars="cluster")
colnames(dat.long) <- c("cluster", "kyoka", "score")
dat.long.summary <- summarySE(dat.long, measurevar=c("score"), groupvars=c("cluster", "kyoka"))
# Rmiscで基本統計量を計算

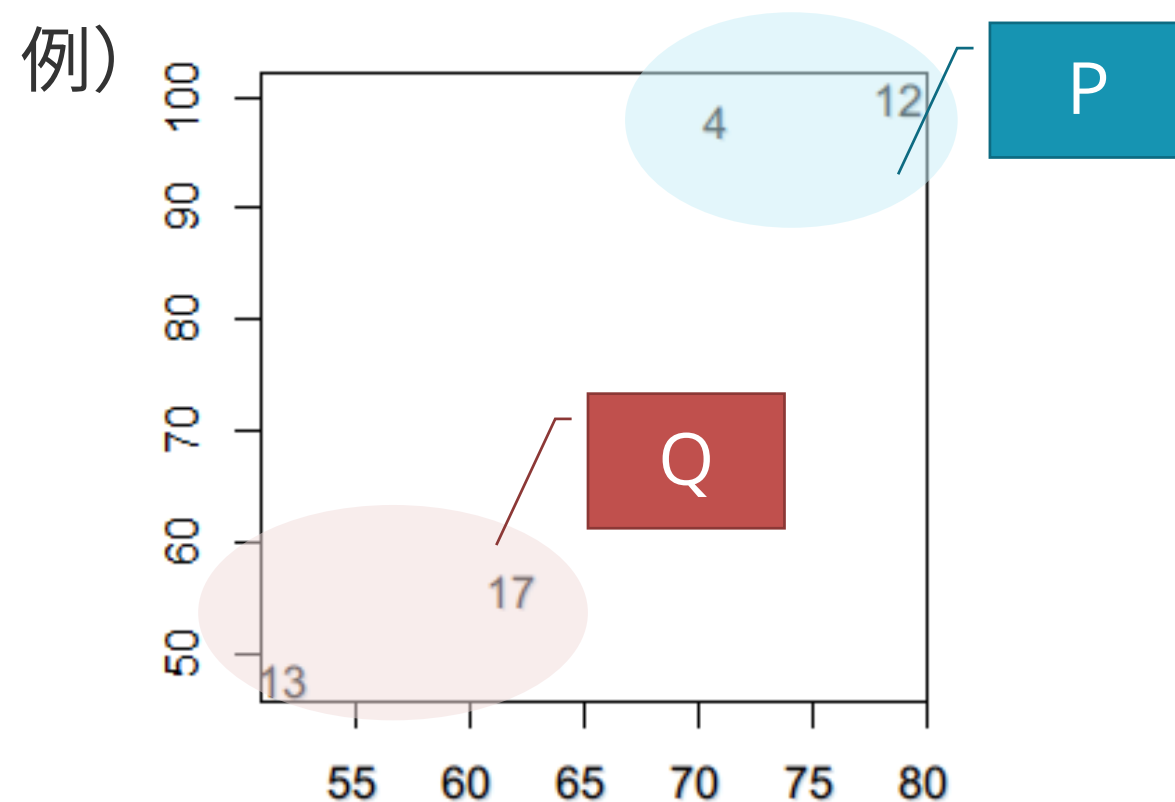
g <- ggplot(dat.long.summary, aes(x=as.factor(cluster), y=score, fill=kyoka)) #X軸にクラスター, Y軸に平均点, 凡例に教科を指定
g <- g + geom_bar(position=position_dodge(0.9), stat='identity', colour="black") #棒を埋め込む
g <- g + geom_errorbar(aes(ymin=score-se, ymax=score+se), colour="black",
  position=position_dodge(0.9), size=0.5, width=.2) #エラーバーを埋め込む
g <- g + scale_fill_grey(start = 1.0, end = 0.3)
g <- g + theme_classic()
plot(g)
```

clusterの値が1, 2のため、そのままだと、
量的変数として処理されてしまうため、
質的変数として処理するように変換しています。

ワード法

各クラスターについて、重心と所属するデータの距離の二乗和を求め、それを全クラスター分足したもの

結合を検討しているクラスターPとQがあるとき、PとQを一つにすることで増加するクラスター内平方和を、PとQの距離として定義する方法



	A	B	C
1	id	eigo	sugaku
2	4	71	98
3	12	79	100
4	13	52	48
5	17	62	56

●結合前のクラスター内平方和

$$\text{重心}_P = \left(\frac{71 + 79}{2}, \frac{98 + 100}{2} \right) = (75, 99)$$

$$\begin{aligned} \text{クラスター内平方和}_P &= (\text{距離}_{P,4})^2 + (\text{距離}_{P,12})^2 \\ &= \{(75 - 71)^2 + (99 - 98)^2\} + \{(75 - 79)^2 + (99 - 100)^2\} \\ &= 34 \end{aligned}$$

$$\text{重心}_Q = \left(\frac{52 + 62}{2}, \frac{48 + 56}{2} \right) = (57, 52)$$

$$\begin{aligned} \text{クラスター内平方和}_Q &= (\text{距離}_{Q,13})^2 + (\text{距離}_{Q,17})^2 \\ &= \{(57 - 52)^2 + (52 - 48)^2\} + \{(57 - 62)^2 + (52 - 56)^2\} \\ &= 82 \end{aligned}$$

$$\text{結合前クラスター内平方和} = 34 + 82 = 116$$

●結合後のクラスター内平方和

$$\text{重心}_R = \left(\frac{71 + 79 + 52 + 62}{4}, \frac{98 + 100 + 48 + 56}{4} \right) = (66, 75.5)$$

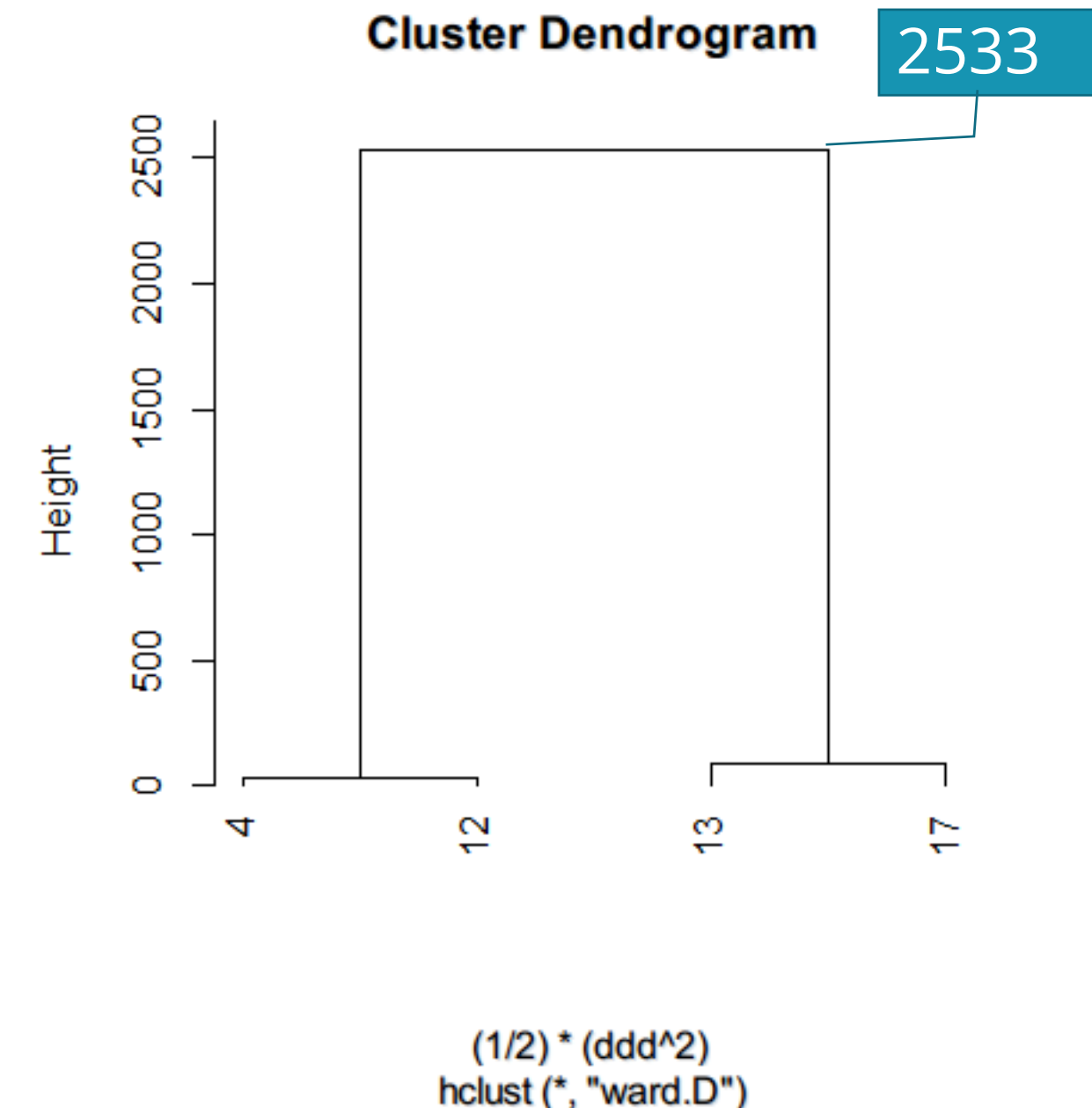
$$\begin{aligned} \text{クラスター内平方和}_R &= (\text{距離}_{R,4})^2 + (\text{距離}_{R,12})^2 + (\text{距離}_{R,13})^2 + (\text{距離}_{R,17})^2 \\ &= \{(66 - 71)^2 + (75.5 - 98)^2\} + \{(66 - 79)^2 + (75.5 - 100)^2\} \\ &\quad + \{(66 - 52)^2 + (75.5 - 48)^2\} + \{(66 - 62)^2 + (75.5 - 56)^2\} \\ &= 2649 \end{aligned}$$

$$\therefore \text{距離}_{P,Q} = \text{増加量} \Delta = 2649 - 116 = 2533$$

前ページで定義された距離(クラスター内平方和の増分)をそのまま分析に用いるには、次のようにして、**ユークリッド距離の二乗の1/2倍**を、`hclust()`の引数として渡す必要があります*。
(1/2倍にしなくても結合のされ方には影響ありませんが、その場合は、デンドログラムで描かれる要素間の距離は、クラスター内平方和の増分の2倍になるので注意が必要です(川端ほか, 2018)。

```
> result <- hclust((1/2)*(ddd^2), method="ward.D")  
> plot(result, hang=-1)
```

ウォード法は、分類結果を解釈しやすい場合が多いといわれていて、階層的クラスター分析では、とくによく使われます。



Rによる非階層的クラスター分析

- 事前準備
とくになし

- 手順

① データセットを読み込む

② 分類

```
kmeans(<データセット>, <クラスター数>, nstart=<試行回数>)
```

※分類結果を見るには、上の結果に\$clusterを付けて見る

③ 各クラスターの特徴の解釈

手順①(データセットを読み込む)

(略)

※さっきのdatが残っている人は、\$clusterを挿入してしまっているので、リセットするために読み込み直しておきましょう。

手順②(分類)

分類に使用する変数だけが含まれている必要あり

ここで指定された試行回数だけkmeans法を繰り返して、最もクラスター内平方和を小さくできた分類結果を結果として返します。あまり大きい数を指定すると時間がかかります。

```
kmeans(<データセット>, <クラスター数>, nstart=<試行回数>)
```

※分類結果を見るには、上の結果に\$clusterを付けて見る

```
> result <- kmeans(dat, 2, nstart=1000)
> result$cluster
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
1  1  1  1  1  1  1  2  1  1  2  1  2  2  1  2  2  1  1  2
```

手順③ (各クラスターの特徴の解釈)

先ほどと同様に行えばOKですね！

```
> dat$cluster <- result$cluster
> anovakun(dat[c(6, 1:5)], "AsB", 2, 5)

[ AsB-Type Design ]
```

clusterという新しい列(変数)を作成して分類結果を代入

anovakunは、対応のない要因の列が最前列になるようにしておく必要があったので、6列目(cluster)が先頭になるよう並べ替えています。

This output was generated by anovakun 4.8.5 under R version 3.6.3.
It was executed on Thu Jun 24 16:34:32 2021.

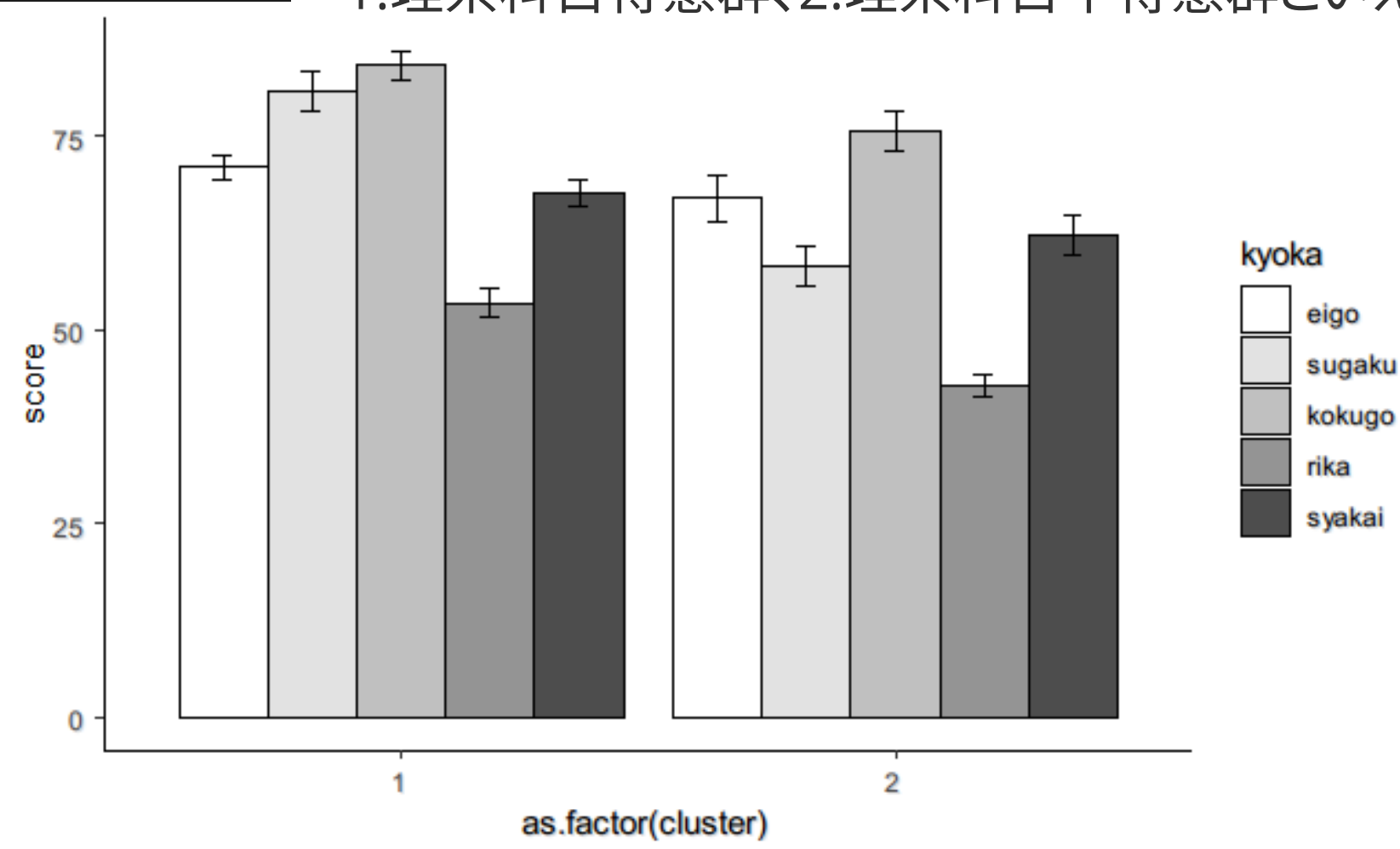
<< DESCRIPTIVE STATISTICS >>

A	B	n	Mean	S.D.
a1	b1	13	70.9231	5.4385
a1	b2	13	80.6154	9.1519
a1	b3	13	84.0769	6.5506
a1	b4	13	53.3846	6.7767
a1	b5	13	67.6154	5.9237
a2	b1	7	67.0000	7.9582
a2	b2	7	58.1429	6.6940
a2	b3	7	75.5714	6.8034
a2	b4	7	42.7143	3.6384
a2	b5	7	62.1429	6.8903

a1: クラスター1
a2: クラスター2

b1:eigo
b2:sugaku
b3:kokugo
b4:rika
b5:syakai

先ほどより差異が小さくなったので解釈が難しいですが、依然として、とくに数学と理科に群間差がみられるので、1:理系科目得意群、2:理系科目不得意群といえそうです。



課題

- 前回のアンケート調査データのうち、コンビニでの様々な商品の購入頻度やサービスの利用頻度に関する5項目に対して階層的クラスター分析(ワード法^{*})を適用し、分類されたクラスターの解釈結果を下記の文章にそって報告してください。分類するクラスターの数、デンドログラムをみて判断しましょう(デンドログラムは、図1として報告してください)。

***スライド11枚目の説明の通り、ユークリッド距離の2乗の1/2倍をhclust()にわたしてください**

- また、ふだんコンビニに行った際にレジで支払う金額について、クラスター×コンビニでの支払い方法(現金or 現金以外)の交互作用があるのか否か分散分析で検討してください。結論付けるのに必要な検定結果については、 F 値(括弧内に自由度)、 p 値等の各種統計量を、論文と同じように報告してください。

本研究では、コンビニの利用の仕方によって、平均的にレジで支払う金額が異なるのか否かを検討する。Web調査の結果、大学生〇〇名が回答した。まず、食べ物や飲み物・生活用品の購入頻度、ならびにATM・コピー機の利用頻度を問う5項目を用いて、階層的クラスター分析(ワード法、対象間の非類似度は平方ユークリッド距離の1/2倍)を行った。デンドログラム(図1参照)を参考に解釈可能性について検討した結果、〇つのクラスターへの分類が妥当と考えられた。各クラスターの平均値をみたところ、第1クラスター($n=〇〇$)はとくに〇〇な傾向が強かったため「〇〇群」、第2クラスター($n=〇〇$)はとくに〇〇な傾向が強かったため「〇〇群」、・・・と命名した。続いて、ふだんコンビニに行った際にレジで支払う、およその金額について、分類されたクラスターおよび支払い方法を対応の〇〇要因とする分散分析を行った結果、・・・

***クラスターの特徴を解釈する際の分散分析については、検定結果の報告は不要です。**

課題の補足

- dist関数には、距離を計算するのに用いたい変数のみが含まれているデータセットを読み込ませる必要があります(今回の場合は、コンビニでの様々な商品の購入頻度やサービスの利用頻度に関する5項目のみ)。データセットから特定の列だけ抽出するには、以下のようにします。

1～4列目だけ抽出する場合

```
> dat
  eigo sugaku kokugo rika syakai
1   77    73    87   58    71
2   78    80    91   64    72
3   65    74    78   46    64
4   71    98    76   65    58
5   63    82    81   49    67
```

```
> dat[1:4]
  eigo sugaku kokugo rika
1   77    73    87   58
2   78    80    91   64
3   65    74    78   46
4   71    98    76   65
5   63    82    81   49
```

1列目と5列目だけ抽出する場合

```
> dat[c(1, 5)]
  eigo syakai
1   77    71
2   78    72
3   65    64
4   71    58
5   63    67
```

3列目だけ取り除いて抽出する場合

```
> dat[-3]
  eigo sugaku rika syakai
1   77    73   58    71
2   78    80   64    72
3   65    74   46    64
4   71    98   65    58
5   63    82   49    67
```

- クラスターの解釈や分類後の主分析に必要な分散分析は、これまでの復習をしながらチャレンジしよう!

チャレンジ問題(任意)

先の課題とは別の方法でクラスター分析を行って、分類されたクラスターの特徴を報告してください(どの方法で分類したかはわかるように付記しておいてください)。