SFC 第３週

ビジネスのデータサイエンス
データマイニング

情報量分析

慶應義塾大学総合政策学部
桑原 武夫

Keio University

授業のすすめ方： 授業システム

- SOL/Canvas （LMS）
  - https://sol.sfc.keio.ac.jp/courses/4115
- Zoom
  - https://keio-univ.zoom.us/j/86357890542?pwd=aFJOc3VkN3g0VGI1MGE2cktLNHdYZz09
  - ミーティングID: 863 5789 0542　パスコード: 421546

コースのねらいと概要

第１週
R言語とプログラミング第１歩
レポート

第２週
情報量分析
レポート

第３週
クラスター分析
レポート

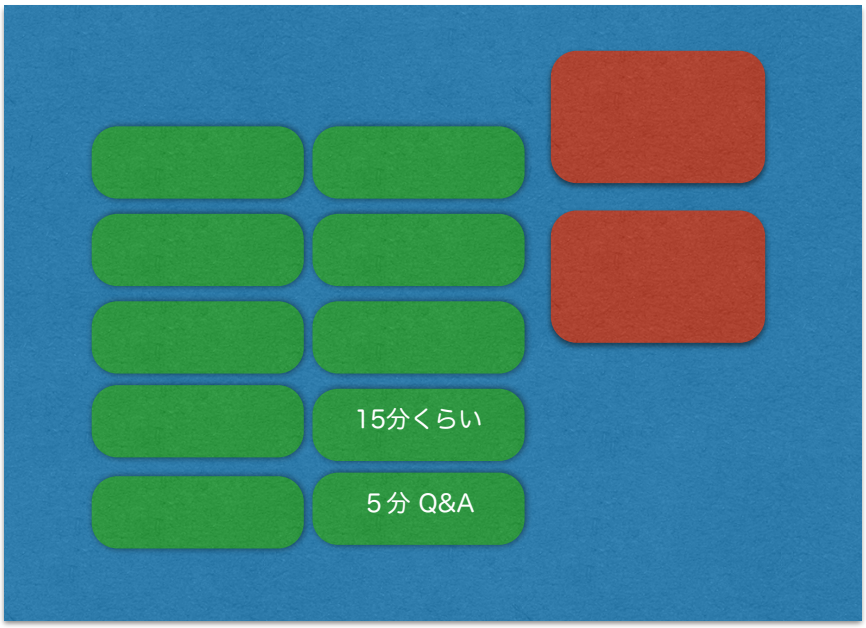
第４週
コレスポンデンス分析
レポート

第５週
決定木

第６週
最終報告会
総合レポート

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

第６週 グループ・プレゼン
データを使って、わかった！経験を発表する



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

第２週 レポート課題１

すてきな関数を作ろう！

Canvas / SOL を見てください。
締切は、学期末（忘れないようにしよう！）

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

準備1: データについて

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

出発点：知りたいこと

必要性

幸福度

身長

興味・関心

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

測定する

目に見えるもの

- 例えば、身長
  - 具体的、物理的

目に見えないもの

- 例えば、幸福度
  - 抽象的
  - 構成概念

Q. 幸せですか？

1. はい
2. どちらともいえない
3. いいえ

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

データ

- 例えば、身長
  - 具体的、物理的

175.7	174.2	169.7	170.4	171.7	171.5
171.5	156.4	177.0	171.2	176.6	164.5
158.5	172.9	182.7	166.3	159.3	158.9
162.0	160.8	171.9	162.5	155.6	159.2
145.0	186.3	180.1	159.8	173.0	155.7
168.0	167.1	162.8	162.8	152.1	161.0
172.3	159.6	158.0	166.2	167.4	169.4
153.4	168.1	162.1	162.4	157.2	158.7
148.3	160.1	170.6	165.0	159.0	155.4
169.5	167.5	163.6	166.6	163.1	167.5
159.7	165.4	171.6	157.3	143.1	168.8
176.1	153.1	153.7	176.0	158.2	145.3
144.5	155.9	159.1	164.8	158.2	167.3
157.5	170.8	154.4	163.7	177.8	161.0

程度
数量

- 例えば、幸福度
  - 抽象的
  - 構成概念

Q. 幸せですか？
1. はい
2. どちらともいえない
3. いいえ

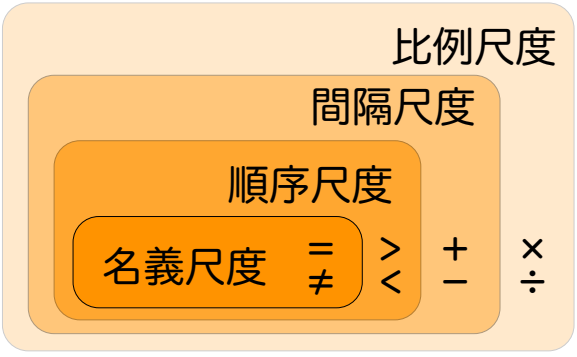
はい　はい　いいえ　どちらとも　いいえ
いいえ　どちらとも　はい　いいえ　はい
どちらとも　いいえ　はい　はい　いいえ
いいえ　どちらとも　はい　いいえ　はい
はい　はい　いいえ　どちらとも　いいえ
いいえ　どちらとも　はい　いいえ　はい
どちらと　1 1 2 3 2 3 1 2 2 1 3 3 2 3 1 2 3 2 1 1 1 2 3
いいえ　2 3 1 2 2 1 3 3 2 3 1 2 3 2 1 1 1 2 3 2 3 1 2
2 1 3 3 2 3 1 2 3 2 1 1 1 2 3 2 3 1 2 2 1 3 3
2 3 1 2 3 2 1 1 1 2 3 2 3 1 2 2 1 3 3 2 3 1 2
3 2 1 1 1 2 3 2 3 1 2 2 1 3 3 2 3 1 2 3 2 1 2

分類
カテゴリー

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

データの種類

- データ：測定項目＝尺度＝変数
- Stevens の４尺度
  - 情報量によってデータ（変数、尺度）を４段階に分類



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

大量のデータ

175.7	174.2	169.7	170.4	171.7	171.5	167.5	146.1	175.4	165.4	170.8	162.9	178.0	171.5	156.					76.6
164.5	158.5	172.9	182.7	166.3	159.3	158.9	162.0	160.8	171.9	162.5	155.6	159.2	145.0	186.					73.0
155.7	168.0	167.1	162.8	162.8	152.1	161.0	172.3	159.6	158.0	166.2	167.4	169.4	153.4	169					57.2
158.7	148.3	160.1	170.6	165.0	159.0	155.4	169.5	167.5	163.6	166.6	163.1	167.5	159.7	165.4					143.1
168.8	176.1	153.1	153.7	176.0	158.2	145.3	144.5	155.9	159.1	164.8	158.2	167.3	157.5	170.8	154.4	163.7	177.8		
161.0	159.7	178.0	161.9	164.0	166.3	175.0	162.8	172.9	159.2	155.6	176.0	154.5	160.6	151.1	162.6	156.3	162.7		
169.6	161.7	167.9	153.8	151.0	174.3	161.4	151.5	170.8	173.8	162.0	164.4	141.1	158.8	161.6	163.1	168.5	175.8		
172.8	179.0	162.1	146.1	164.2	173.5	164.3	150.0	160.7	161.7	161.2	179.4	163.8	164.9	173.8	163.0	165.6	159.6		
166.6	152.2	160.6	172.5	160.1	169.0	169.9	167.4	167.8	170.2	160.6	156.1	154.3	168.9	171.1	161.8	175.5	166.5		
166.5	156.1	168.4	153.9	153.4	181.1	158.2	178.6	163.7	142.2	161.8	166.0	162.0	180.1	167.5	169.9	156.7	164.5		
159.9	191.8	165.2	146.3	167.2	168.3	151.5	139.7	159.5	185.4	153.1	163.7	154.9	164.0	155.9	173.5	157.0	167.9		
162.7	172.7	158.7	162.8	164.0	152.1	166.3	163.6	176.2	161.3	149.0	164.0	153.7	160.3	161.3	164.2	161.7	165.9		
174.0	153.3	170.0	158.3	170.0	156.9	175.6	176.5	184.6	161.7	164.5	169.3	167.3	169.3	173.8	173.5	180.1	160.0		
160.2	180.7	171.5	170.3	はい	はい	いいえ	どちらとも	いいえ	いいえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	はい	いいえ	いいえ	どちら
183.7	153.7	155.6	165.3	とも	はい	いいえ	はい	はい	はい	いいえ	どちらとも	いいえ	いいえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	は
168.7	158.6	156.3	157.2	い	いいえ	いいえ	どちらとも	はい	いいえ	はい	はい	はい	いいえ	どちらとも	いいえ	いいえ	どちらとも	はい	いいえ	はい	ど
148.0	152.3	159.0	156.2	ちらとも	いいえ	はい	はい	いいえ	いいえ	どちらとも	はい	いいえ	はい	はい	はい	いいえ	どちらとも	いいえ	いいえ	どちらと	
166.5	181.5	163.7	164.4	も	はい	いいえ	はい	どちらとも	いいえ	はい	はい	いいえ	いいえ	どちらとも	はい	いいえ	はい	はい	はい	いいえ	どちらとも
158.9	159.4	171.0	152.8	いいえ	いいえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	はい	いいえ	いいえ	どちらとも	はい	いいえ	はい	はい	
158.9	157.2	178.7		はい	いいえ	どちらとも	いいえ	いいえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	はい	いいえ	いいえ	どちらとも		
				いいえ	はい	はい	いいえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	はい	いいえ	いいえ	どちらとも			
				い	はい	いいえ	どちらとも	はい	いいえ	いいえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	はい	い		
					いいえ	どちらとも	はい	いいえ	はい	はい	はい	いいえ	どちらとも	いいえ	いいえ	どちらとも	はい	いいえ	はい	どちらと	
					いえ	はい	はい	いいえ	いいえ	どちらとも	はい	いいえ	はい	はい	はい	いいえ	どちらとも	いいえ	いいえ	どちらとも	は
					いえ	はい	どちらとも	いいえ	はい	はい	いいえ	いいえ	どちらとも	はい	いいえ	はい	はい	はい	いいえ	どちらとも	いい
					いえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	はい	いいえ	いいえ	どちらとも	はい	いいえ	はい	はい	はい
					え	どちらとも	いいえ	いいえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	はい	いいえ	いいえ	どちらとも	はい	
					え	はい	はい	はい	いいえ	どちらとも	いいえ	いいえ	どちらとも	はい	いいえ	はい	どちらとも	いいえ	はい	はい	いいえ

Q. 幸せですか？

1. はい
2. どちらともいえない
3. いいえ

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

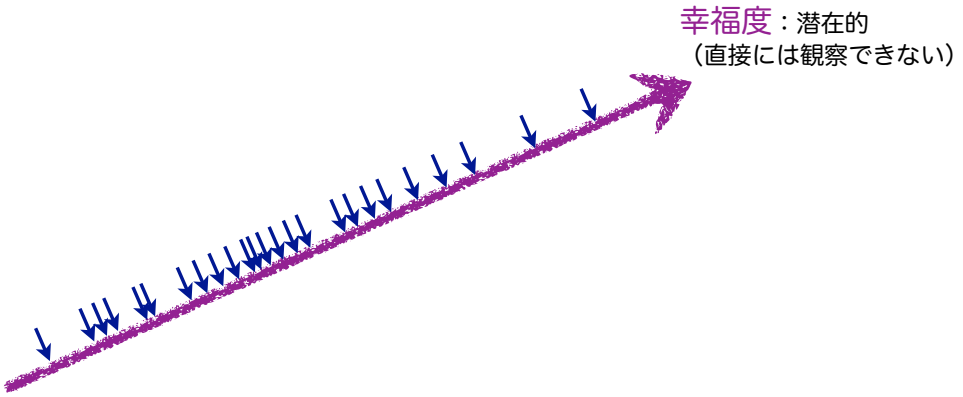
準備2: 良い測定、良い変数

どれだけの情報を顕在化できるか

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

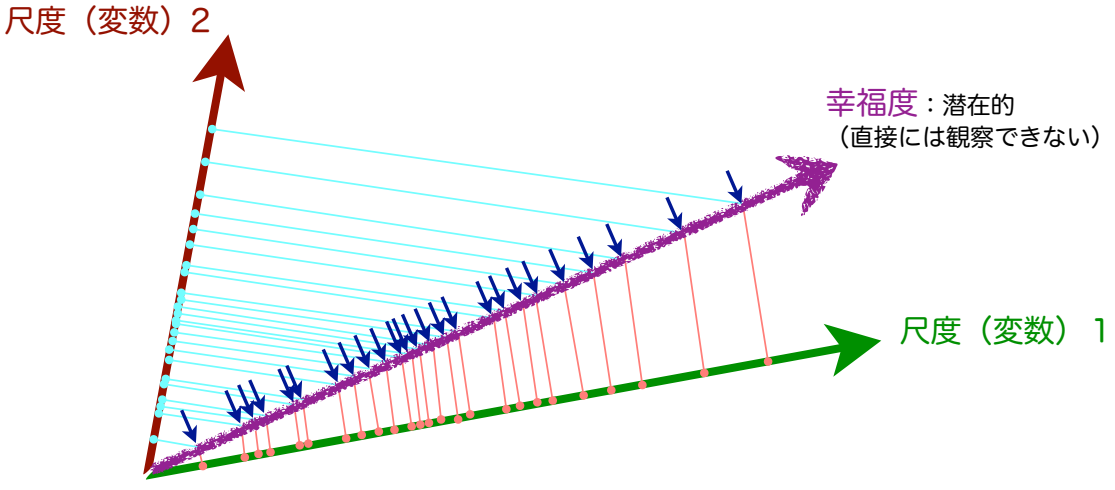
幸福度を測定する



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

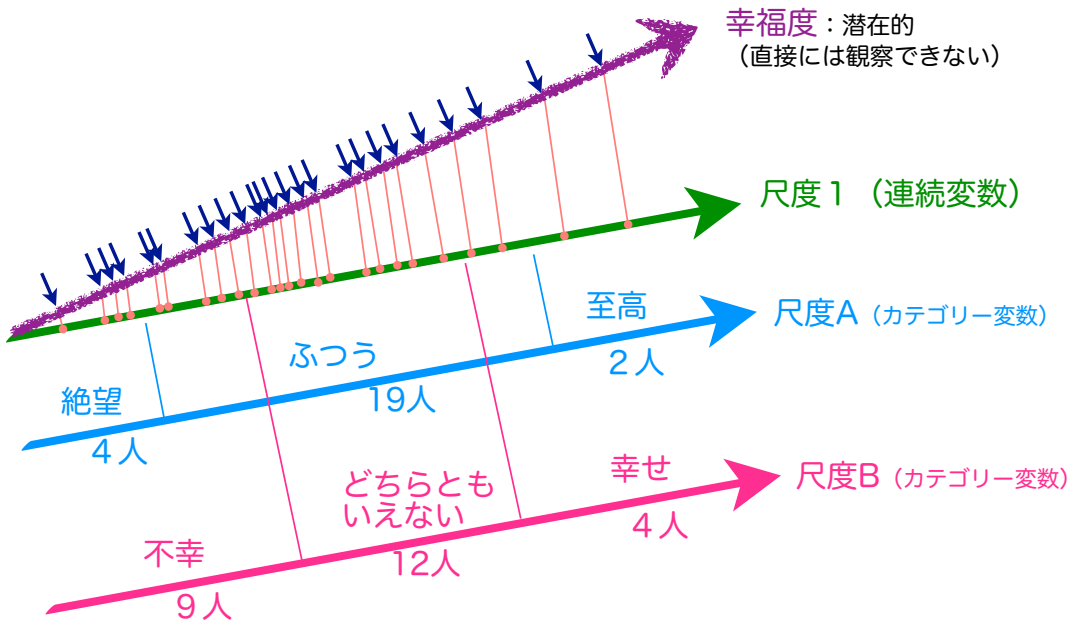
Keio University

良い尺度



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

カテゴリー（選択肢）で測定する場合



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

準備３：情報量

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

いろいろな状態（幸福度）

(; _ ;)　(つД｀)　(´・ω・｀)　(･3･)　(*ﾟーﾟ)　(´∀｀)　(*ﾟ▽ﾟ)ﾉ　＼(＾▽＾)／

表現するためのコード化

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

状態が２つだったら....

(´・ω・｀)　＼(＾▽＾)／

0　　**1**

状態区分	コード
(´・ω・｀)	**0**
＼(＾▽＾)／	**1**

状態数＝2
コード桁数＝1

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

状態が４つだったら....

(; _ ;)

↓

00

(´・ω・｀)

↓

01

(´∀｀)

↓

10

＼(＾▽＾)／

↓

11

状態区分	コード
(; _ ;)	0 0
(´・ω・｀)	0 1
(´∀｀)	1 0
＼(＾▽＾)／	1 1

状態数＝4
コード桁数＝2

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

状態が８つだったら....

(; _ ;)

↓

000

(つД｀)

↓

001

(´・ω・｀)

↓

010

(･3･)

↓

011

(*°ー°)

↓

100

(´∀｀)

↓

101

(*°▽°)ノ

↓

110

＼(＾▽＾)／

↓

111

状態数＝8
コード桁数＝3

状態区分	コード
(; _ ;)	0 0 0
(つД｀)	0 0 1
(´・ω・｀)	0 1 0
(･3･)	0 1 1
(*°ー°)	1 0 0
(´∀｀)	1 0 1
(*°▽°)ノ	1 1 0
＼(＾▽＾)／	1 1 1

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

つまり、状態が k 通りだったら、必要桁数＝情報量（bit）は、

状態区分	コード
(´・ω・`)	1
＼(ˆ▽ˆ)／	0

状態区分	コード
(；_；)	0 0
(´・ω・`)	0 1
(´∀`)	1 1
＼(ˆ▽ˆ)／	0 1

状態区分	コード
(；_；)	0 0 0
(つД`)	0 0 1
(´・ω・`)	0 1 0
(・з・)	0 1 1
(*°ー°)	1 0 0
(´∀`)	1 0 1
(*°▽°)ノ	1 1 0
＼(ˆ▽ˆ)／	1 1 1

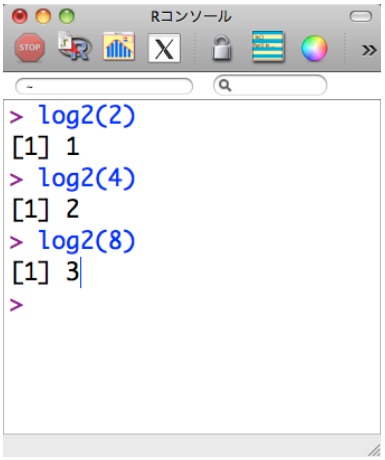
$$H = \log_2 k$$

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

情報量（bit）

状態数 k	情報量 $H=\log_2 k$ (bit)
1	
2	$\log_2 2 = 1$
3	
4	$\log_2 4 = 2$
5	
6	
7	
8	$\log_2 8 = 3$
9	
10	
11	



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

情報量（bit）

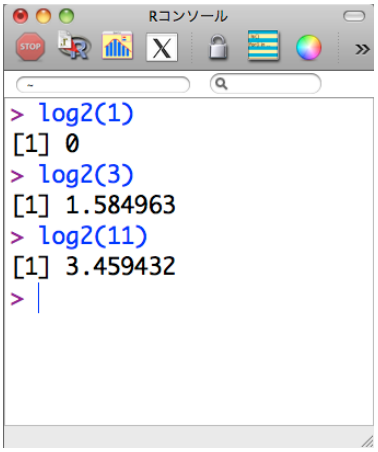
状態数 k	情報量 $H=\log_2 k$ (bit)
1	$\log_2 1$ =
2	$\log_2 2$ = 1
3	$\log_2 3$ =
4	$\log_2 4$ = 2
5	$\log_2 5$ =
6	$\log_2 6$ =
7	$\log_2 7$ =
8	$\log_2 8$ = 3
9	$\log_2 9$ =
10	$\log_2 10$ =
11	$\log_2 11$ =

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

情報量（bit）

状態数 k	情報量 $H=\log_2 k$ (bit)
1	$\log_2 1$ = 0
2	$\log_2 2$ = 1
3	$\log_2 3$ = 1.58
4	$\log_2 4$ = 2
5	$\log_2 5$ = 2.32
6	$\log_2 6$ = 2.58
7	$\log_2 7$ = 2.81
8	$\log_2 8$ = 3
9	$\log_2 9$ = 3.17
10	$\log_2 10$ = 3.32
11	$\log_2 11$ = 3.46



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

複数人いたとすると、

- ひとりについて、k 通りの状態があって、
- n人いる時、全体をコード化するのに必要な桁数は、

$$H = n\log_2 k$$

	1. (´・ω・`)	2. (^▽^)	...	k. (;_;)	
1					log2k
	1. (#ﾟДﾟ)	2. (*ﾟ―ﾟ)	...	k. (つДﾟ)	
2					log2k
			:		:
	1. (*ﾟ▽ﾟ)ﾉ	2. (*ﾟ―ﾟ)	...	k. (･з･)	
n					log2k

nlog2k

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

n人いたとすると、状態の数 k は

- 状態の数（k） は、データに基づいて区分できるかどうかを考えると、最大で、k = n まで

	1. (´・ω・`)	2. (^▽^)	3. (;_;)	4. (･з･)	5. (*ﾟ―ﾟ)	6. (つДﾟ)	...	n. (*ﾟ▽ﾟ)ﾉ	n+1. (#ﾟДﾟ)
1	1	0	0	0	0	0	…	0	0
2	0	1	0	0	0	0	…	0	0
3	0	0	1	0	0	0	…	0	0
4	0	0	0	1	0	0	…	0	0
5	0	0	0	0	1	0	…	0	0
6						1	…		
:	:	:	:	:	:	:	:	:	:
n	0	0	0	0	0	0	…	1	0

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

潜在情報量（bit）

- つまり、n人について、何らかの方法で識別できる最大限を考えると、必要桁数は、

$$H = n\log_2 n$$

人数 n	情報量 $H=n\log_2 n$ (bit)
1	$1\log_2 1 = 0$
2	$2\log_2 2 = 2$
3	$3\log_2 3 = 4.75$
4	$4\log_2 4 = 8$
5	$5\log_2 5 = 11.60$
6	$6\log_2 6 = 15.51$
7	$7\log_2 7 = 19.65$
8	$8\log_2 8 = 24$
9	$9\log_2 9 = 28.52$
10	$10\log_2 10 = 33.22$

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

潜在情報量（nit）

- bit のままでもいいのだが、後々便利なことが多いので、bitを約1.386倍し（自然対数にし）た nit を使う

$$H = 2n\log_e n$$

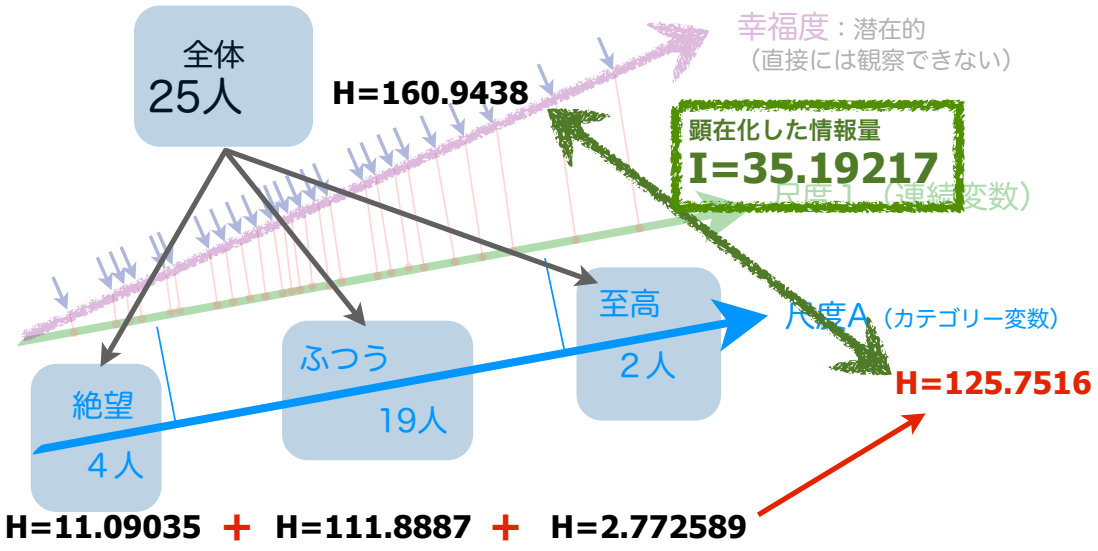
人数 n	情報量 $H=n\log_2 n$ (bit)
1	$1\log_2 1 = 0$
2	$2\log_2 2 = 2$
3	$3\log_2 3 = 4.75$
4	$4\log_2 4 = 8$
5	$5\log_2 5 = 11.60$
6	$6\log_2 6 = 15.51$
7	$7\log_2 7 = 19.65$
8	$8\log_2 8 = 24$
9	$9\log_2 9 = 28.52$
10	$10\log_2 10 = 33.22$

× 1.386

人数 n	情報量 $H=2n\log_e n$ (nit)
1	$2\times 1\log_e 1 = 0$
2	$2\times 2\log_e 2 = 2.77$
3	$2\times 3\log_e 3 = 6.59$
4	$2\times 4\log_e 4 = 11.09$
5	$2\times 5\log_e 5 = 16.09$
6	$2\times 6\log_e 6 = 21.50$
7	$2\times 7\log_e 7 = 27.24$
8	$2\times 8\log_e 8 = 33.27$
9	$2\times 9\log_e 9 = 39.55$
10	$2\times 10\log_e 10 = 46.05$

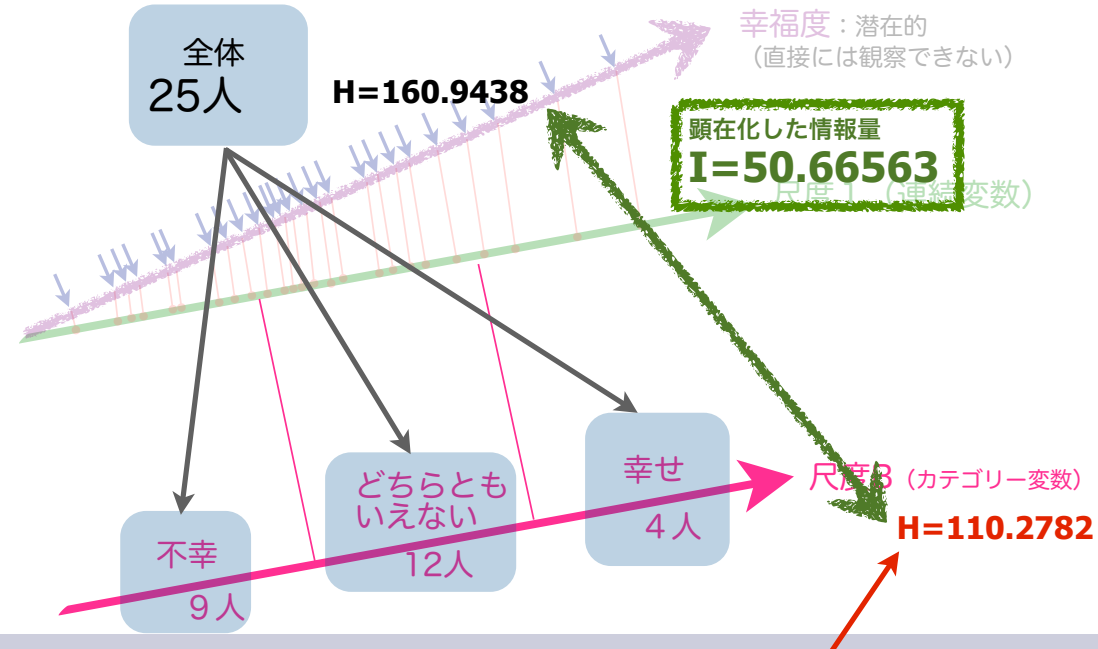
Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

分割すると情報は、顕在化する



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

尺度Bの方が顕在情報量が多い



Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

潜在情報量の計算

```
# 潜在情報量を計算する関数
# nit : function - the amount of information (if n=0 then nit=0 )
nit <- function(n) {
 v <- 2*n*log(n); if (is.nan(v))  v <- 0
 return(as.numeric(v))
 }
```

```
> nit(25)
[1] 160.9438

> nit(9)+nit(12)+ nit(4)
[1] 110.2782

> nit(25) - (nit(9)+nit(12)+ nit(4))
[1] 50.66563
```


Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

Keio University

第２週 レポート課題２にむけて

「ビジネスのためのデータサイエンス / データマイニング」履修者について、
説明する（＝情報を顕在化させる）と面白いと思われる変数は何でしょう？

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

第２週 レポート課題２（個人レポート）

「ビジネスのためのデータサイエンス / データマイニング」履修者を
もっともよく説明する（＝情報を顕在化させる）変数は何でしょう。その
検討過程含めてレポートを作成してください。

提出はSOL。
レポートはなるべくpdfで。
提出期限：厳守

提出遅れは、保証外

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp

ご清聴ありがとうございました

Takeo Kuwahara
kuwahara@sfc.keio.ac.jp