

■ このノートで習う統計量の一覧

(要素1つ)      (カテゴリー1つ)      (カテゴリー2つ以上)

相対頻度    →    相対頻度表    →    分割表 (相対頻度表)

↓経験的な確率分布↓

情報量    →    情報量の表    →    情報量の表の集合

情報量

確率分布の  
デコボコさ

エントロピー →

結合エントロピー

条件付エントロピー

自己相互情報量

→

相互情報量

$\chi^2$ 値

→

$\chi^2$ 値

確率分布の距離

## ■ 頻度から経験分布へ

### (1) 頻度

これは、標本において各カテゴリーが何回出現したかを表す統計量。

#### ① 粗頻度 Raw Frequency

これは、その確率変数の実現値が、何回生じたのかを表す統計量。

<hr/> <i>crazy</i> <hr/> <hr/> 25 <hr/>	$Freq(X = x) = x$ が観測された数
---	---------------------------

#### ② 相対頻度

これは、その確率変数の実現値が、どのくらいの割合で生じたのかを表す統計量。

<hr/> <i>crazy</i> <hr/> <hr/> 0.658 <hr/>	$RelFreq(X = x) = x$ が観測された割合
--	-------------------------------

### (2) 頻度表

#### ① 粗頻度表 Raw Frequency Table

これは、確率変数の実現値に対し粗頻度がいくつかを表示したもの。すべての値を足すと総観測数になる。

<hr/> <i>crazy</i> <hr/>	<hr/> <i>nuts</i> <hr/>	<hr/> <i>mad</i> <hr/>	<hr/> <i>out of one's mind</i> <hr/>
25	7	4	2

#### ② 相対頻度表

これは、確率変数の実現値に対し相対頻度がいくつかを表示したもの。すべての値を足すと1になる。

<hr/> <i>crazy</i> <hr/>	<hr/> <i>nuts</i> <hr/>	<hr/> <i>mad</i> <hr/>	<hr/> <i>out of one's mind</i> <hr/>
0.66	0.18	0.11	0.05

### (3) 分割表 Contingency Table

#### ① 分割表 (粗頻度)

二つ (以上) のカテゴリーの組み合わせで作られた条件に  
適う事例の頻度を表した頻度表。

	<i>crazy</i>	<i>nuts</i>	<i>mad</i>	<i>out of one's mind</i>
<i>PRS</i>	25	7	4	2
<i>PST</i>	17	2	2	0

#### ② 分割表 (相対頻度)

(ケース1) 二つ (以上) のカテゴリーの組み合わせの  
相対頻度表 (行列内の和が 1)。

	<i>crazy</i>	<i>nuts</i>	<i>mad</i>	<i>out of one's mind</i>
<i>PRS</i>	0.42	0.12	0.07	0.03
<i>PST</i>	0.29	0.03	0.03	0.00

(ケース2) 二つ (以上) のカテゴリーの組み合わせの  
相対頻度表 (行の和が 1)。

	<i>crazy</i>	<i>nuts</i>	<i>mad</i>	<i>out of one's mind</i>
<i>PRS</i>	0.66	0.18	0.11	0.05
<i>PST</i>	0.45	0.05	0.05	0.00

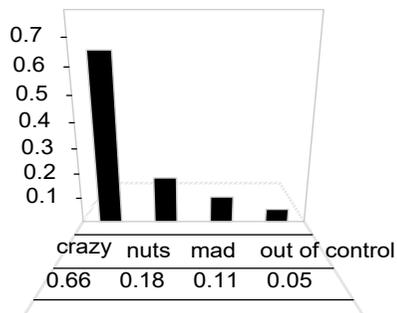
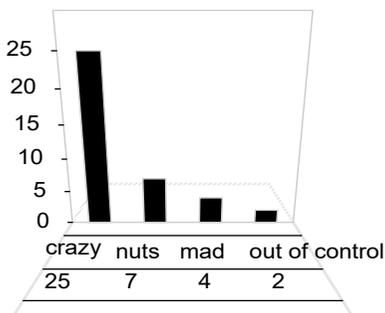


#### 経験分布としての相対頻度表

相対頻度表は離散的な経験分布として解釈される。

母集団における確率分布： 出る確率

標本における確率分布： 出た確率 (経験分布)

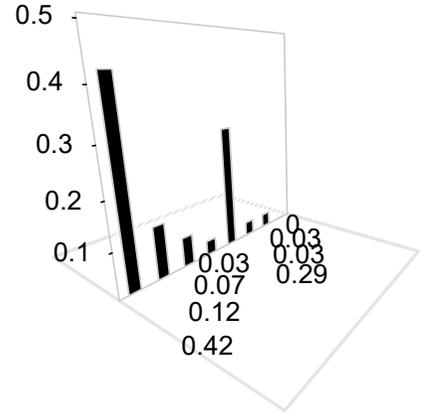
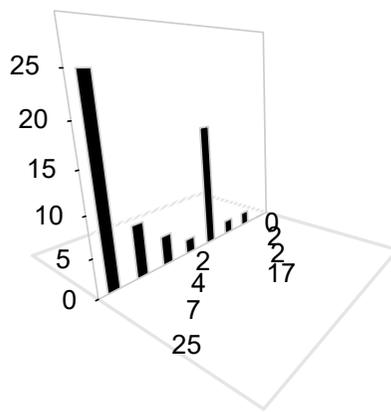




## 経験分布としての分割表（相対頻度）

### （ケース1）行列内の和が1の場合

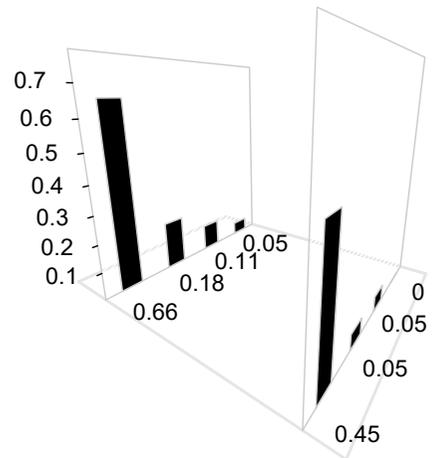
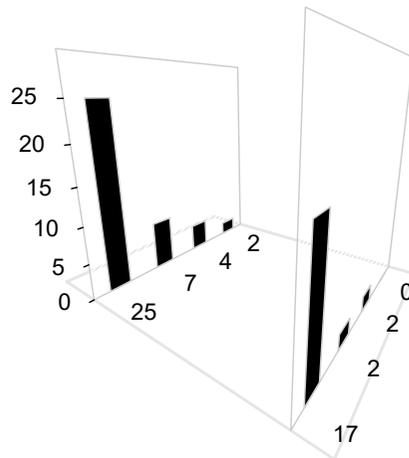
長い離散的な経験分布として解釈できる。



⇒ どのくらい「デコボコするのか」という点に関心が向く。  
(☞ 結合エントロピー)

### （ケース2）行の和が1の場合

複数の離散的な経験分布の集合として解釈できる。



⇒ わけたことで、平均してどのくらいデコボコするようになったのかという点に関心が向く。

(☞ 条件つきエントロピー)

⇒ わけられた二つの確率分布はどのくらい違うのかという点に関心が向く。

(☞ KL/JS 情報量など)

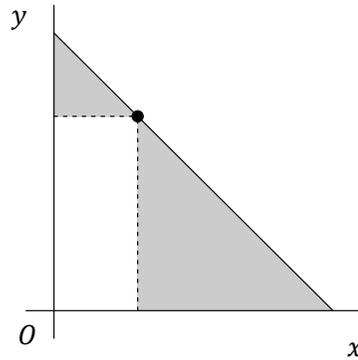
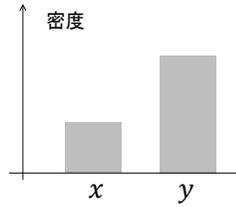
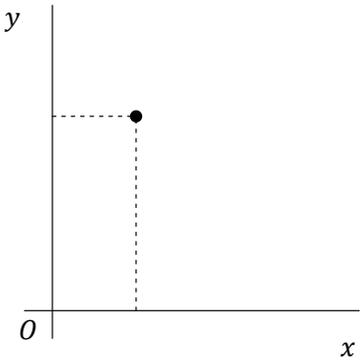
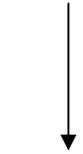


## 確率分布（相対頻度表）の集合の幾何的な解釈①

相対頻度表の集合を考えるとそこに多様体が生まれる。

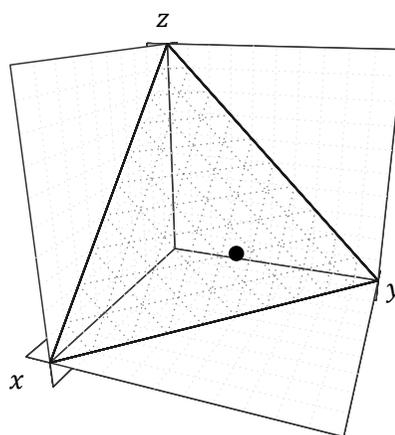
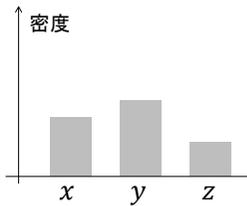
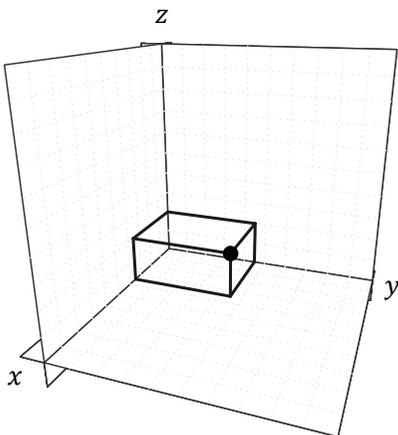
### ① 二次元の場合

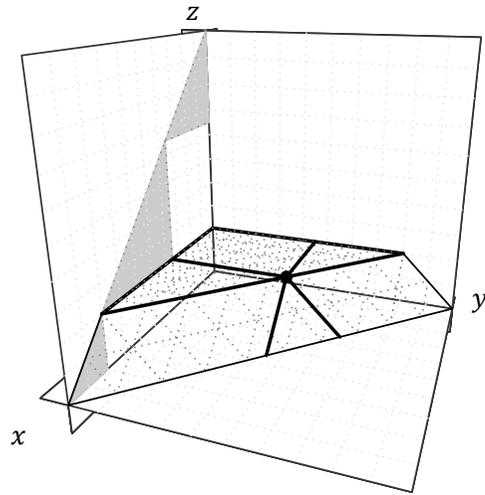
$(0.3, 0.7)$



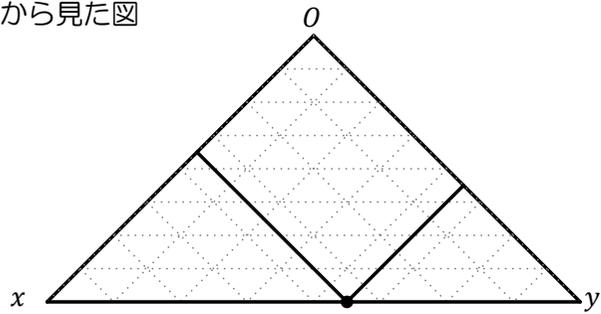
### ② 三次元の場合

$(0.35, 0.45, 0.2)$

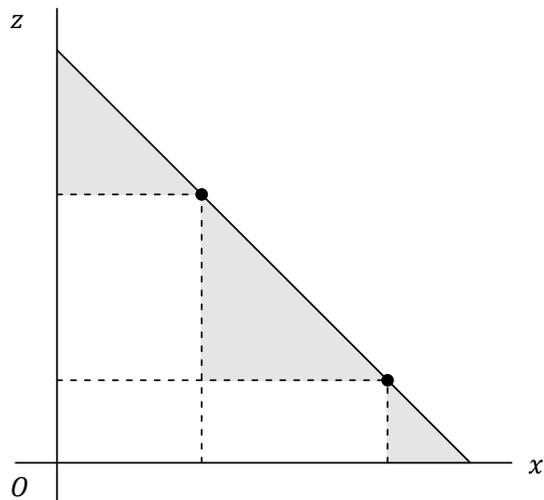




上から見た図



左横から見た図



## ■ 確率分布に対して定められる統計量 1 : 驚きの度合い

### (1) 情報量 Amount of Information

これは、ある事象の持つ情報の大きさ（その事象が観察されることがどれだけ驚くべきことか）を表した統計量。

$$\begin{aligned} I(X = x) &= -\log_2 p(X = x) \\ &= -\log p(X = x) \end{aligned}$$

※ 底が2のときの情報量の単位をビット(bit)という。



#### 確率と情報量の記法

確率変数  $X$  が  $x$  という値を取る確率を次のように表す。

$$p(X = x) = p(x)$$

そこで、情報量を次のようにも表記する。

$$I(x) = -\log_2 p(x)$$



#### 指数と対数

「3を何乗したら  $x$  になるのか」の答えを次のように書く。

$$y = \log_3 x$$

##### ① 掛け算 (積)

$$1 \div 3 \div 3 \cdots \div 3$$

⋮

$$1 \div 3 \div 3$$

$$1 \div 3$$

$$1$$

$$1 \times 3$$

$$1 \times 3 \times 3$$

⋮

$$1 \times 3 \times 3 \times \cdots \times 3$$

##### ② 指数

$$= 3^{-n}$$

⋮

$$= 3^{-2}$$

$$= 3^{-1}$$

$$= 3^0$$

$$= 3^1$$

$$= 3^2$$

⋮

$$= 3^n$$

##### ③ 対数

$$-n = \log_3 3^{-n}$$

⋮

$$-2 = \log_3 3^{-2}$$

$$-1 = \log_3 3^{-1}$$

$$0 = \log_3 1$$

$$1 = \log_3 3^1$$

$$2 = \log_3 3^2$$

⋮

$$n = \log_3 3^n$$

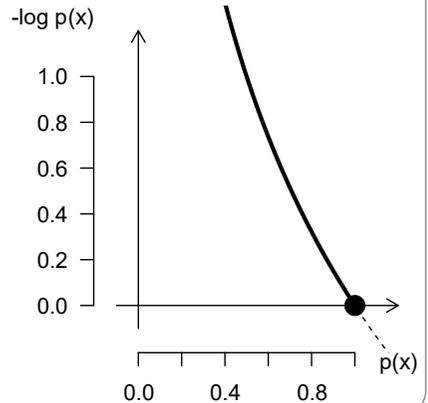
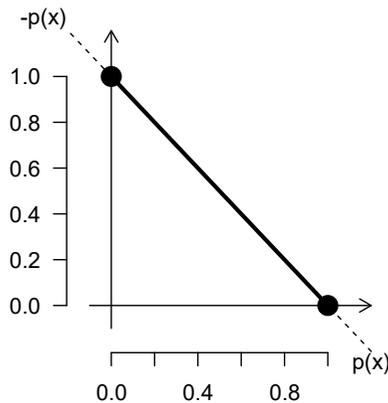


### 情報量の直感的な意味合い

情報量の根底にある動機は「珍しいことが起こったときに高い値を出す統計量を作りたい」というもので、次の性質が満たされるように設計されています。

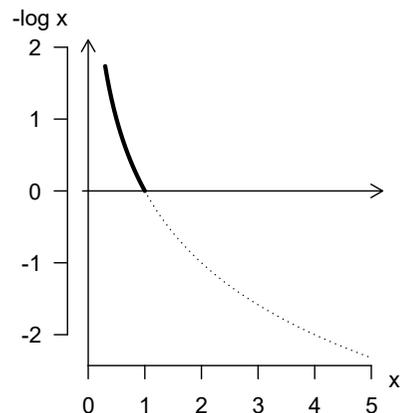
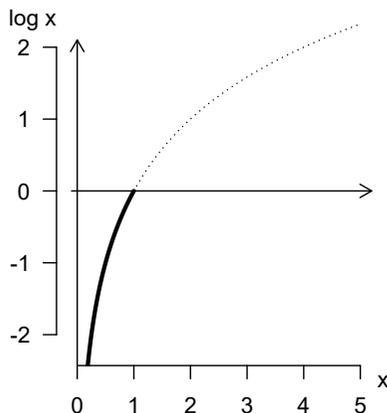
- ①  $I(x)$ は、 $0 \leq p(x) \leq 1$ の区間で、単調に減少する。
- ②  $I(x)$ は、 $0 \leq p(x) \leq 1$ の区間で、連続。
- ③  $x_1$ が生じることと $x_2$ が生じることが完全に独立なとき、両者が同時に起こるときの情報量が和で計算できる。  
(これは計算機にやさしい設計となっている)

この三つの条件を満たす関数 $I(x)$ は、 $I(x) = -\log_2 p(x)$ という形しかありえないことが知られています (証明略)。



### 対数関数 log のグラフ

真数の部分には $0 \leq p(x) \leq 1$ の確率が来る。返り値を扱いやすい正の値にするために、情報量では負の対数を用いる。



## (2) 情報量の表 (ベクトル)

分析においては、単一の情報量を議論することはまれで、複数の情報量の表・ベクトルを扱う。

### ① 相対頻度表

<i>crazy</i>	<i>nuts</i>	<i>mad</i>	<i>out of one's mind</i>
0.658	0.184	0.105	0.052



### ② 情報量の表

<i>crazy</i>	<i>nuts</i>	<i>mad</i>	<i>out of one's mind</i>
0.60	2.44	3.25	4.25



### 💡 ここまで登場した統計量のまとめ

(要素1つ)      (カテゴリー1つ)      (カテゴリー2つ以上)

相対頻度 → 相対頻度表 → 分割表 (相対頻度表)

↓ 経験的な確率分布 ↓

↓

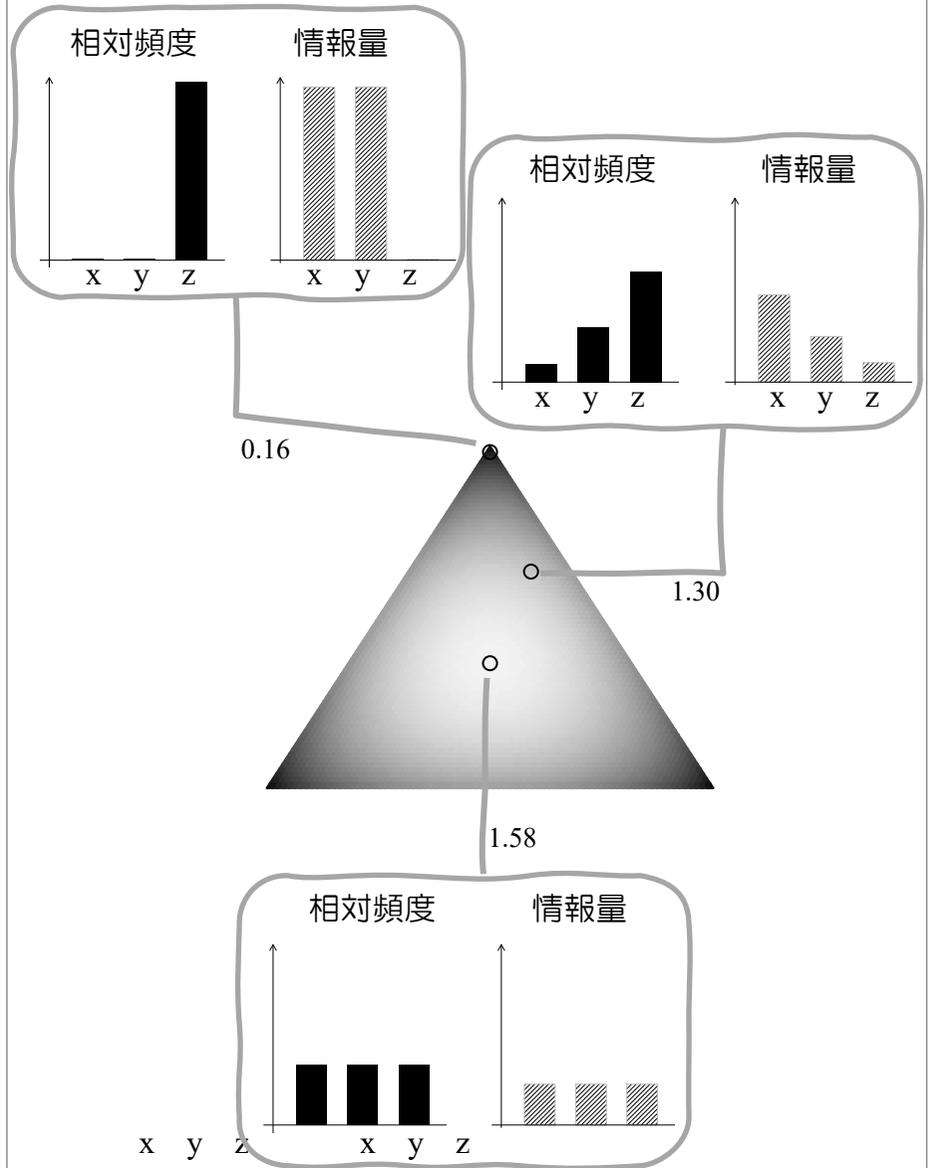
情報量 → 情報量の表 → 情報量の表の集合

情報量



## エントロピーの幾何的な理解

エントロピーは、確率分布（相対頻度表）に対して定義できる。そして、相対頻度表は確率単体として表現できる。そこで、この単体上の各点（相対頻度表）に対し対応するエントロピーがどのくらいなのかを色の濃さで示したのが下図である（色が白いところほどエントロピーの値が高い）。



## ■ 確率分布に対して定められる統計量2：デコボコさ

### (1) エントロピー Entropy (シャノンの情報量)

これは、離散的な経験確率分布に対して定義される情報量の平均を意味する統計量である。 $0 \leq H(P) \leq \log(n)$ となる。

$$H(X) = \sum_{i=1}^n -p(x_i) \log p(x_i)$$

※ 情報の乱雑さを測るために使われる。



### 重みづけ平均

#### ① 重みづけ平均 (加重平均)

例1：天気と効用 (損失)

地域 A	晴れ	雨	雪
損失	0	20	50
確率 (重み)	0.9	0.05	0.05
地域 B	晴れ	雨	雪
損失	0	20	100
確率 (重み)	0.9	0.05	0.05
地域 C	晴れ	雨	雪
損失	0	20	50
確率 (重み)	0.05	0.9	0.05

例2：お店の商品の値段

	リンゴ	オレンジ	バナナ
値段	500	200	100
確率 (重み)	0.1	0.3	0.6

#### ② 情報量の重みづけ平均 (加重平均)

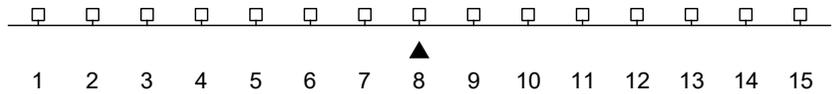
	単語 1	単語 2	単語 3
情報量	$-\log 0.7$	$-\log 0.2$	$-\log 0.1$
確率 (重み)	0.7	0.2	0.1



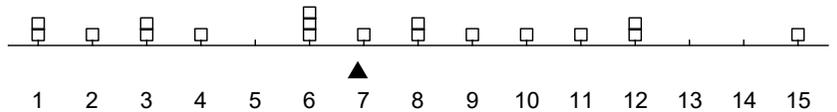
## エントロピーの直感的な意味合い：重心

### ① 線分の重心

(重みがすべて同じとき)



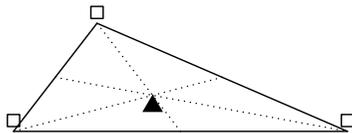
(重みがまちまちのとき)



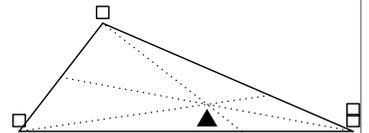
### ② 三角形の重心

重みが全て等しい時、三角形の重心は各頂点から対応する斜辺の中点へ引いた線分が交わる点である。

(重みがすべて同じとき)

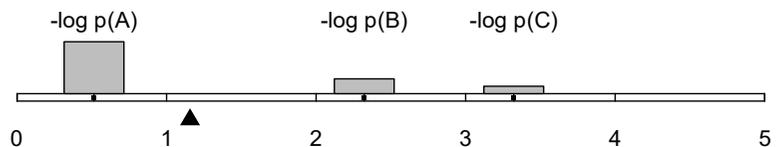


(重みがまちまちの時)



### ③ 情報量の重心

	単語 A	単語 B	単語 C
情報量	$-\log 0.7$	$-\log 0.2$	$-\log 0.1$
確率 (重み)	0.7	0.2	0.1



$$\begin{aligned}
 H(P) &= \sum_{x \in X} -p(X=x) \log p(X=x) \\
 &= p(X=A) - \log p(X=A) \\
 &\quad + p(X=B) - \log p(X=B) \\
 &\quad + p(X=C) - \log p(X=C)
 \end{aligned}$$

## (2) 結合エントロピー Joint Entropy

これは、離散的な経験的同時確率分布に対して定義される情報量の平均を意味する統計量である。

$$H(X, Y) = \sum_{i=1}^n \sum_{k=1}^K -p(x_i, y_k) \log p(x_i, y_k)$$

二つの情報を両方同時に知ることで得られる平均情報量。

=複数のカテゴリーで分割表を作ったときに、その分割表全体でどのくらいデコボコしているのかを表す指標。



### 結合エントロピーの例

サイコロを振って出た目を考える。「3以下か4以上か」「奇数か偶数か」を知ることによって得られるエントロピーをそれぞれ  $H(X)$ 、 $H(Y)$  と置く。 $H(X, Y)$  は、この二つの情報を両方手にしたときに得られる平均情報量である。



### 結合エントロピーの性質

それぞれのエントロピーと結合エントロピーの間には次のような性質が成り立つことが知られている。

(性質1) それぞれのカテゴリーのエントロピーとの関係

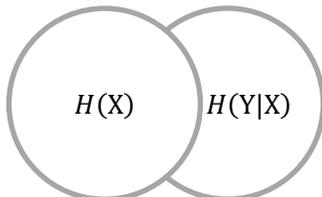
$$0 \leq H(X, Y) \leq H(X) + H(Y)$$

※  $H(X, Y) = H(X) + H(Y)$  という統合成立条件は、 $X$  と  $Y$  が独立の時だけである。

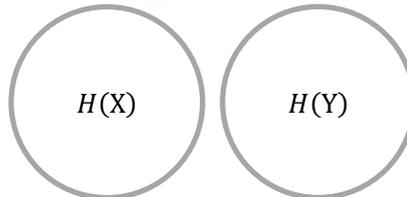
(性質2) 条件付エントロピーとの関係

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

独立ではないとき



独立なとき





## 結合エントロピーの直感的な意味合い

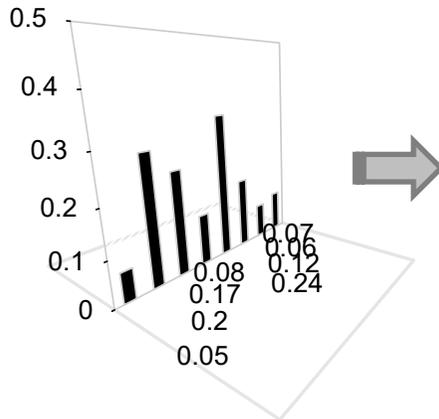
### ① 分割表（粗頻度）

	<i>make</i>	<i>take</i>	<i>do</i>	<i>have</i>
<i>text 1</i>	4	17	14	7
<i>text 2</i>	20	10	5	6

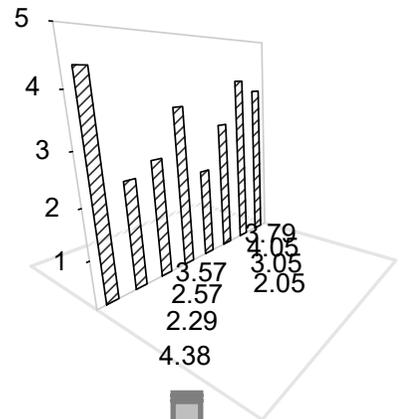
### ② 分割表（相対頻度）

	<i>make</i>	<i>take</i>	<i>do</i>	<i>have</i>	合計
<i>text 1</i>	0.05	0.20	0.17	0.08	0.51
<i>text 2</i>	0.24	0.12	0.06	0.07	0.49

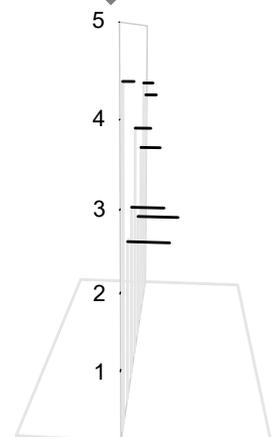
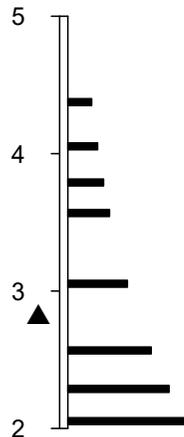
### ③ 同時確率密度分布 （相対頻度の分割表）



### ④ 情報量を算出



### ⑤ エントロピーを計算



### (3) 条件付きエントロピー Conditional Entropy

ある事象が与えられたときのエントロピーのこと。

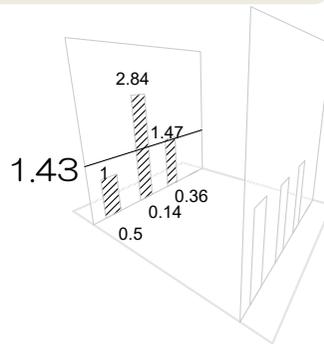
- ①  $Y = y$  が与えられた時の  $X$  の条件付きエントロピー  
 これは、 $Y = y$  という特定の値が与えられた時の確率変数  $X$  の乱雑さを表した統計量。

$$H(X|Y = y) = \sum_{i=1}^n -p(x|y) \log p(x|y)$$

#### 💡 条件つきエントロピーの計算過程 (1)

例：「教科書かどうか」という情報でエントロピーに変化があるか？

	速度	速さ	スピード	合計
教科書	50	14	36	100
教科書以外	16	17	17	50
合計	66	31	53	150



- ②  $Y$  が与えられた時の  $X$  の条件付きエントロピー  
 これは、確率変数  $Y$  が与えられた時の確率変数  $X$  の乱雑さを表した統計量。

$$H(X|Y) = \sum_{i=1}^n \sum_{k=1}^K -p(x, y) \log p(x|y)$$

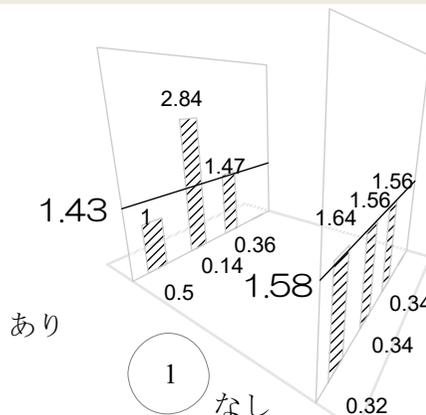
#### 💡 条件つきエントロピーの計算過程 (2)

それぞれの「板」の上にある確率分布のエントロピーを計算したのち、それらの重みづけ平均を求める。

2

あり

1  
なし





### なぜ2で割らず、加重平均を用いるのか？

次の二つの二つのようなケースの違い、つまり、各カテゴリーに属する確率を反映させたいため、加重平均が用いられている。

#### ケース1

	速度	速さ	スピード	合計	Ent.
教科書	50	14	36	100	1.43
教科書以外	16	17	17	50	1.58
合計	66	31	53	150	1.52

1.48

#### ケース2

	速度	速さ	スピード	合計	Ent.
教科書	50	14	36	100	1.43
教科書以外	32	34	34	100	1.58
合計	82	48	70	200	1.52

1.51

ケース1では、「教科書以外」の数が少ない分、「教科書以外」という条件の下もたらされたエントロピーはその分少なめに貢献させられている。

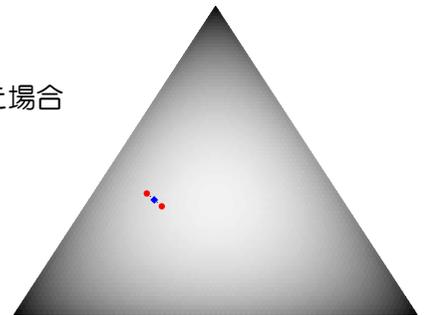


### 条件つきエントロピーの幾何的な理解

条件付エントロピーは、ある情報（条件）を与えたとき、平均してどのくらいのエントロピーになるのかを教えてくれる。

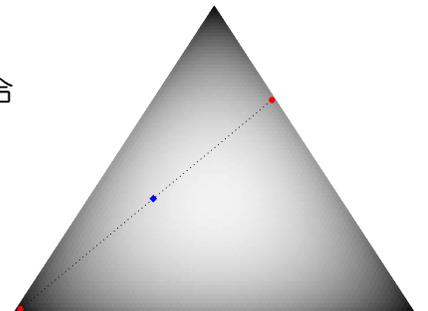
#### 条件を考えてもあまり意味がなかった場合

	x	y	z	合計	Ent.
A	50	14	36	100	1.43
not A	48	20	32	100	1.58
合計	98	34	68	200	1.52



#### 条件を考えることで効果があった場合

	x	y	z	合計	Ent.
B	97	2	1	100	0.16
not B	1	33	67	100	0.98
合計	98	34	68	200	1.52



## ■ 確率分布に対して定められる統計量3：距離／類似度

### (1) カルバック・ライブラー・ダイバージェンス (KL 情報量) Kullback-Leibler divergence (KL divergence)

これは、二つの確率分布間の距離を表す統計量の一つ。相対エントロピー (Relative Entropy) とも呼ばれる。

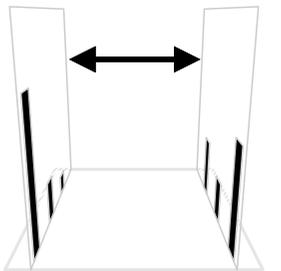
$$D_{KL}(p||q) = \sum_{i=1}^n p(x_i) \log \frac{p(x_i)}{q(x_i)}$$



#### 確率分布間の距離

たくさん相対頻度表 (経験的な確率分布) を手にした時、それらの中でどれが似ているのかという類似度 (または距離) を測りたいという場面が、この後たくさん登場する。

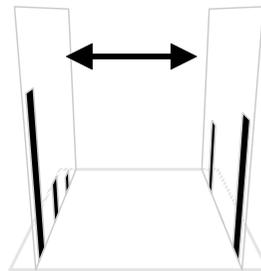
ケース1



分布  $p$

分布  $q$

ケース2



分布  $p$

分布  $r$



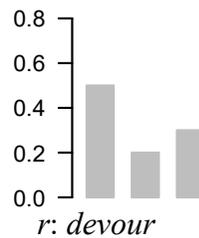
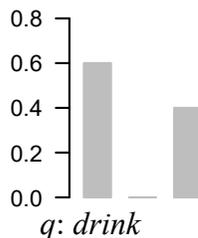
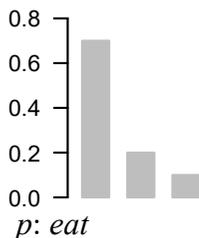
#### コーパス言語学における利用例

##### ① モチベーション

コーパスをもとに、「意味の類似性」を測りたい!

##### ② 例

SV O の位置に来る相対頻度表 (確率分布) を見る。  $p$  の分布を作った動詞と近い「距離」の動詞は意味が似てるはず。

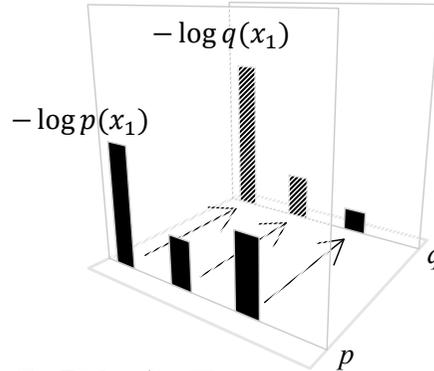




## KL 情報量の直感的な意味合い

確率分布  $p$  から確率分布  $q$  へ変化することによって生じる情報量の変化に対し確率分布  $p$  で加重平均を計算したもの。

### ① 各カテゴリーの情報量の変化



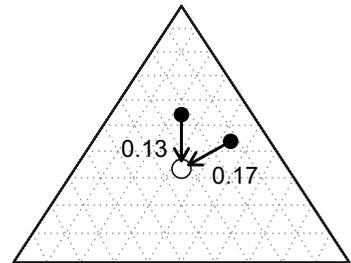
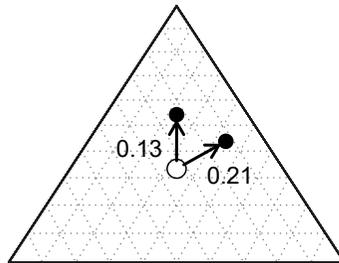
$$\begin{aligned} & -\log q(x_1) - \{-\log p(x_1)\} \\ &= -\log q(x_1) - \{-\log p(x_1)\} \\ &= -\log q(x_1) + \log p(x_1) \\ &= \log p(x_1) - \log q(x_1) \\ &= \log \frac{p(x_1)}{q(x_1)} \end{aligned}$$

### ② 重みづけ平均

$$D_{KL}(p||q) = \sum_{i=1}^n p(x_i) \log \frac{p(x_i)}{q(x_i)}$$



## KL 情報量と単体



## KL 情報量の性質

(性質 1) 非負性

$$D_{KL}(p||q) \geq 0$$

(性質 2) 非退化性

$$D_{KL}(p||q) = 0 \Leftrightarrow p = q$$

(性質 3) 非対称性

$$D_{KL}(p||q) \neq D_{KL}(q||p)$$

(性質 4) 三角不等式を満たさない!

$$D_{KL}(p||q) \neq D_{KL}(r||p) + D_{KL}(q||r)$$

## (2) ジェンセン・シャノン・ダイバージェンス Jensen-Shannon divergence

これは、二つの確率分布間の距離を表す統計量の一つ。

$$D_{JS}(p||q) = \frac{D_{KL}(p||r) + D_{KL}(q||r)}{2}$$

※ ただし、 $r$  という確率分布は、次式で定義される  $p$  と  $q$  の「平均」とも言えるような確率分布。

$$r(x) = \frac{p(x) + q(x)}{2}$$



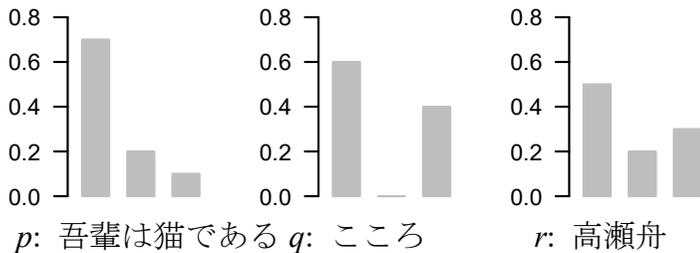
### 計量文体論における利用例

#### ① モチベーション

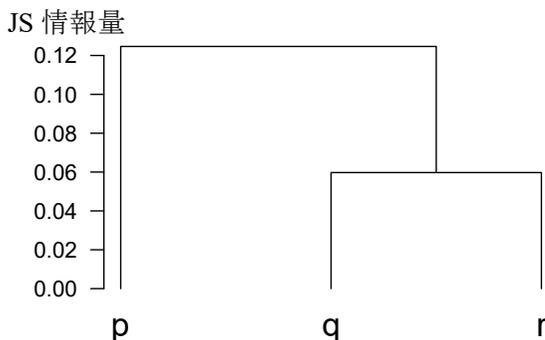
文体上類似したテキストを見つけたい！

#### ② 例

それぞれのテキストから相対頻度表を作り、JS 情報量の点で近いものからグルーピングをしていく。



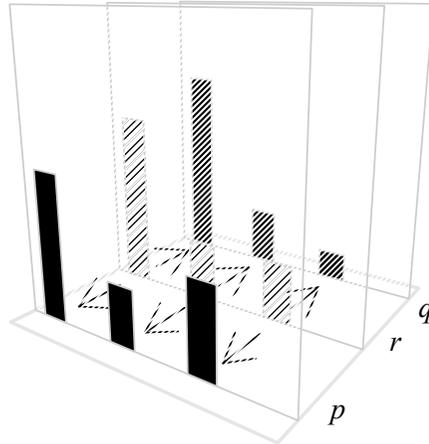
#### ③ 凝集型階層的クラスタリング





### JS 情報量の直感的な意味合い

KL 情報量のような非対称性を持つ統計量ではなく、対照性を持つ確率分布間の距離として考案された。



距離を計算する前に、 $p, q$  の二つの確率分布の「平均」に相当する確率分布  $r$  を作り、そこから  $p, q$  への距離を KL 情報量で測り、その平均を取ったもの。



### JS 情報量の性質

(性質 1) 非負性

$$D_{JS}(p||q) \geq 0$$

(性質 2) 非退化性

$$D_{JS}(p||q) = 0 \Leftrightarrow p = q$$

(性質 3) 対称性

$$D_{JS}(p||q) = JS(q||p)$$

(性質 4) 三角不等式を満たさない！

$$D_{JS}(p||q) \neq D_{JS}(r||p) + D_{JS}(q||r)$$

### (3) 相互情報量 Mutual Information

これは、 $X$  と  $Y$  が独立な場合の同時分布と実際の同時分布の距離を示す統計量。確率変数間の依存度合いを表す。

$$MI(X, Y) = D_{KL}[ p(x, y) || p(x)p(y) ]$$

$$= \sum_i^n \sum_k^K p(x_i, y_k) \log \frac{p(x_i, y_k)}{p(x_i)p(y_k)}$$



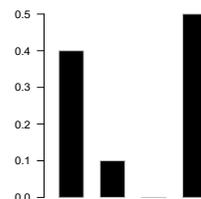
#### 同時確率と変数の独立

二つの変数が独立している場合にのみ、同時確率が周辺確率の積で表せる。すなわち、次の関係が成り立つ。

$$p(x, y) = p(x)p(y)$$

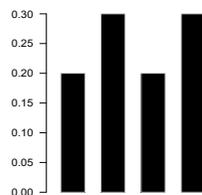
(A) 経験分布の相対頻度表 (同時確率)

X/Y	crazy あり	crazy なし	合計
drive 構文	0.4	0.1	0.5
drive 構文以外	0	0.5	0.5
合計	0.4	0.6	1



(B) 二つの変数が独立した場合の相対頻度表 (同時確率)

X/Y	crazy あり	crazy なし	合計
drive 構文	0.2	0.3	0.5
drive 構文以外	0.2	0.3	0.5
合計	0.4	0.6	1

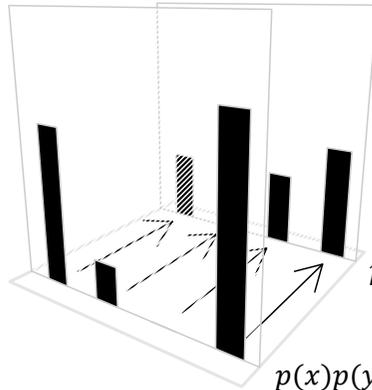




### 自己相互情報量の直感的な意味合い

「X と Y が独立だという理想的な場合」から近いのか遠いのか、距離を出すことで、X と Y の依存度（=どれだけ独立ではないのか）を算出している。

$$MI(X, Y) = D_{KL}[ p(x, y) \parallel p(x)p(y) ]$$



$p(x, y)$  : 今回のデータ (経験分布)

$p(x)p(y)$  : 完全に独立なとき



### 計量文体論における利用例

#### ① モチベーション

どの二つの単語の組み合わせで共起の強さが大きいのか知りたい！

#### ② 例

10,000 万語のテキストにおける単語の使用頻度と単語の共起頻度から相互情報量を計算する。

単語 A	(頻度)	単語 B	(頻度)	共起頻度	MI
生成	100	文法	200	90	5.49
認知	200	文法	200	50	3.64
生成	100	言語学	100	2	1.00
認知	200	言語学	100	50	4.64

- (4) 自己相互情報量 Pointwise mutual information (PMI)  
 これは、二つの確率変数の「特定の実現値の組  $x, y$  に対して」の関連の度合いを測るときに用いられる統計量。

$$MI(X, Y) = \sum_i^n \sum_k^K p(x_i, y_k) \log \frac{p(x_i, y_k)}{p(x_i)p(y_k)}$$

↑  
↓

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

↑  
↓

相対頻度表 (確率分布) 全体が対象  
 相対頻度表の一つのセルが対象

二つの単語が関連している度合いを測る統計量であり、相互情報量はこれに  $p(x, y)$  で重みを付けた平均である。



#### 自己相互情報量の直感的な意味合い

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad \begin{array}{l} \leftarrow \text{加点} \\ \leftarrow \text{減点} \end{array}$$

$$= \log p(x, y) - \log p(x)p(y)$$

↑ 加点            ↑ 減点

基本的には、同時確率が大きければ、加点をしていく (分子)。ただし、ただ同時確率にのみ注目してしまうと、(a) teacher と school というコンビも、(b) the と school というコンビも、ともに高い値を持ってしまう。しかし、(b) のような事例は排除したい。そこで、そもそも出現頻度が多い単語である場合、ペナルティとして減点し、調整を試みている。

- ①  $p(x, y) = p(x)p(y)$  のとき  
すなわち、 $x$  と  $y$  が独立の時は、0 になる。
- ②  $p(x, y) > p(x)p(y)$  のとき  
すなわち、 $x$  と  $y$  に強い共起関係があるとき、 $PMI > 0$
- ③  $p(x, y) < p(x)p(y)$  のとき  
すなわち、 $x$  と  $y$  に強い相補分布関係があるとき、 $PMI < 0$



## 自己相互情報量の見ている部分

### (A) 独立なとき

X/Y	teacher あり	teacher なし	合計
school あり	0.2	0.3	0.5
school なし	0.2	0.3	0.5
合計	0.4	0.6	1

### (B) 経験分布 1: teacher と school の共起関係

X/Y	teacher あり	teacher なし	合計
school あり	0.4	0.1	0.5
school なし	0	0.5	0.5
合計	0.4	0.6	1

### (C) 経験分布 2: school と the の共起関係

X/Y	the あり	the なし	合計
school あり	0.02	0.01	0.03
school なし	0.58	0.39	0.97
合計	0.6	0.4	1



## 自己相互情報量の別の捉え方

$$\begin{aligned}
 \text{PMI}(x, y) &= \log \frac{p(x, y)}{p(x)p(y)} \\
 &= \log \frac{p(x)p(y|x)}{p(x)p(y)} \\
 &= \log \frac{p(y|x)}{p(y)} \\
 &= \log p(y|x) - \log p(y)
 \end{aligned}$$

↑ 加点
↑ 減点



## 自己相互情報量の限界

(1) 出現頻度が低い単語同士の場合  
非常に大きくなってしまう

(2) 0 が出てくるとき

$p(x, y) = 0$  のとき、 $\text{PMI} = -\infty$  となり計算できない。このため、スムージング処理を行うことが多い。

(5) 適合度 (Goodness-of-fit) 統計量

観測値の頻度表と期待値の頻度表の離れ具合を測る統計量 (📖ノート 3)。

$$GOF = \sum_{i=1}^k \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$$

- $\left\{ \begin{array}{l} O_i \text{ 観測度数 (粗頻度)} \\ \hat{E}_i \text{ 理論度数 (粗頻度) : } np_i \end{array} \right.$

※ 両者に差がないという  $H_0$  の下で  $\chi^2$  分布に従うため GOF の代わりに  $\chi^2$  という値で書かれることもある。



観測度数と期待度数

① 観測度数：実際の観測データ  $O_i$

	<i>say</i>	<i>nod</i>	<i>change</i>	<i>arrive</i>	合計
<i>PST</i>	35	10	5	10	<b>60</b>
<i>PRF</i>	10	0	20	10	<b>40</b>
合計	<b>45</b>	<b>10</b>	<b>25</b>	<b>20</b>	<b>100</b>

② 理論度数：理論上こうなっているだろうという理想データ  $\hat{E}_i$

	<i>say</i>	<i>nod</i>	<i>change</i>	<i>arrive</i>	合計
<i>PST</i>	27	6	15	12	<b>60</b>
<i>PRF</i>	18	4	10	8	<b>40</b>
合計	<b>45</b>	<b>10</b>	<b>25</b>	<b>20</b>	<b>100</b>

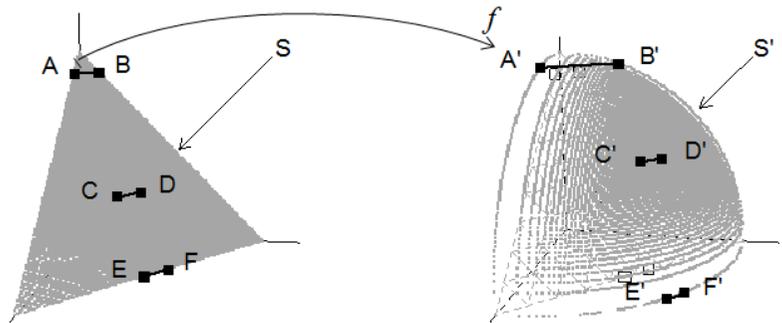


## その他の分布間の距離を測る指標

### ① ヘルンジャー距離 Hellinger Distance

$$d_H(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2 / 2}$$

$$= \frac{1}{\sqrt{2}} \|\sqrt{\mathbf{P}} - \sqrt{\mathbf{Q}}\|_{L_2}$$



### ② コサイン類似度

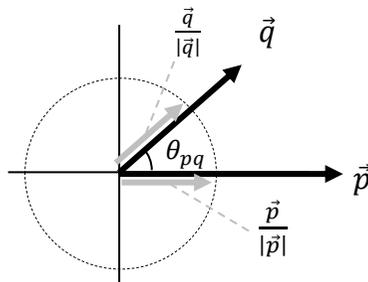
相対頻度表（経験確率分布）をベクトルと見なし、 $\vec{p}$ と $\vec{q}$ のなす角 $\theta_{pq}$ のコサインを類似度と考える。

$$d_{\cos}(\mathbf{p}, \mathbf{q}) = \cos \theta_{pq}$$

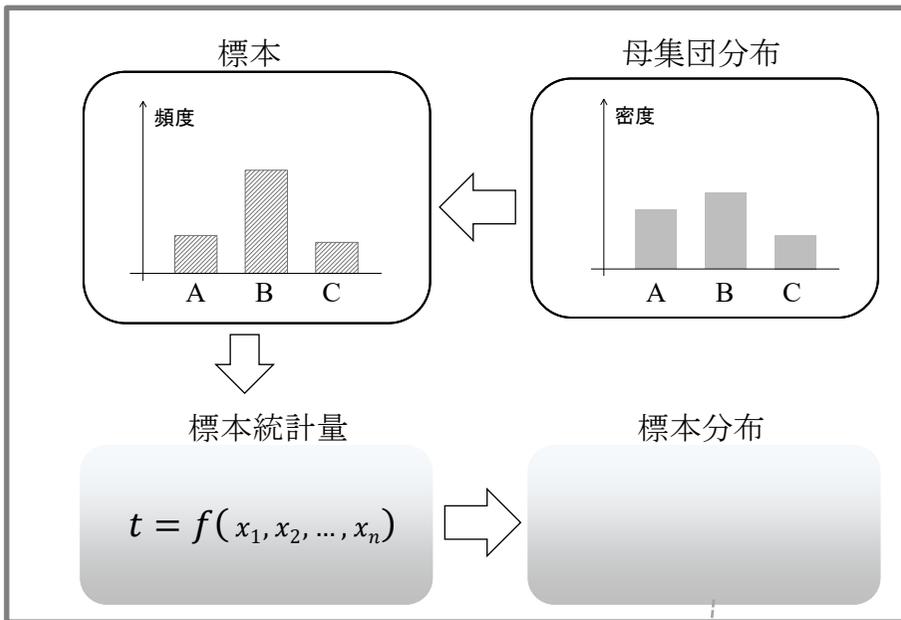
$$= \left( \frac{\vec{p}}{|\vec{p}|} \right) \cdot \left( \frac{\vec{q}}{|\vec{q}|} \right)$$

$$= \frac{\vec{p} \cdot \vec{q}}{|\vec{p}| |\vec{q}|}$$

$$= \frac{\sum_{i=1}^n p_i q_i}{\sqrt{\sum_{i=1}^n p_i^2} \sqrt{\sum_{i=1}^n q_i^2}}$$



※ そもそも内積という概念自体二つのベクトルの類似度を表す概念。ベクトルの頂点を比較できるように、単位円周上に載せてから内積を取っている。



■ このノートで習う標本分布

- (ケース1) 標本分布の形がよく分かっているもの
  - ① 母比率の検定
  - ② 比率の差の検定
  - ③ 独立性の検定
- (ケース2) 標本分布の形がよく分かっていないもの  
ブートストラップ法

■ 分布の形がよく分かっているもの

(1) 母比率の検定

これは、母集団比率 $\pi$ が特定の値 $\pi_0$ と等しいかどうかを検定する方法。

① 帰無仮説

帰無仮説  $H_0 : \pi = \pi_0$

対立仮説  $H_1 : \pi \neq \pi_0$

② ド・モアブル＝ラプラス (De Moivre-Laplace) の定理

サイズが  $n$  で、確率 $\pi$ の  $Bin(n, \pi)$  に従う  $X$  があるとき、

(i)  $X$  は  $n$  が大きい時、 $N(n\pi, n\pi(1-\pi))$  に従う。

(ii)  $X/n (=p)$  は  $n$  が大きい時、 $N(\pi, \pi(1-\pi)/n)$  に従う。

③ 標準化

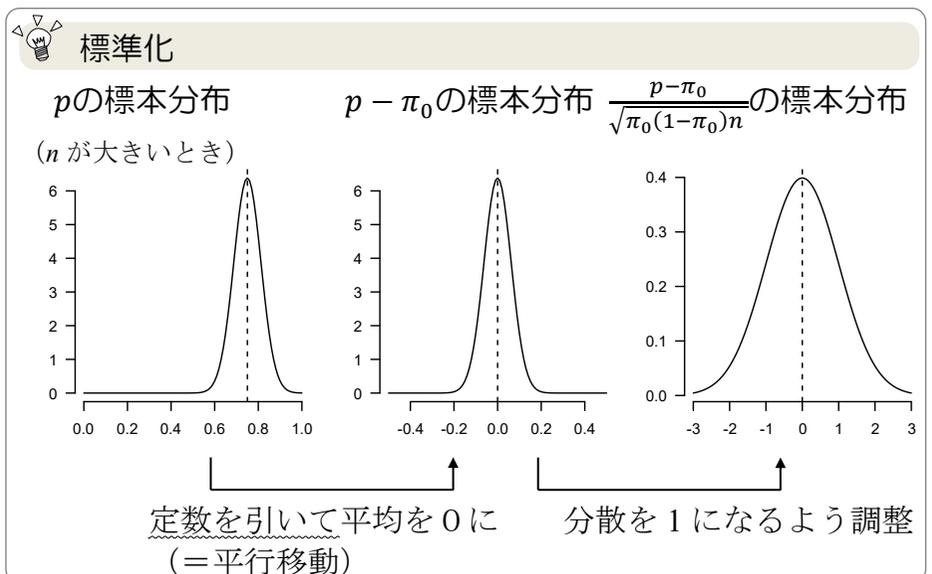
正規分布に従うのなら、標準化した統計量を作成する。

帰無仮説が真のとき、 $z$  は標本分布  $N(0,1)$  に従う。

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$$

④ 仮説の判定

$z$  の値が標本分布のどこに位置しているかで帰無仮説との整合性を判断する。



(2) 比率の差の検定 (独立した二群の比率の差の検定)

これは、集団 1 における標本比率  $p_1$  と集団 2 における標本比率  $p_2$  から母集団比率に差がないのかを検定する方法。

① 帰無仮説

帰無仮説  $H_0$  :  $\pi_1 = \pi_2 (= \pi)$

対立仮説  $H_1$  :  $\pi_1 \neq \pi_2$

② ド・モアブル＝ラプラス (De Moivre-Laplace) の定理

サイズが  $n$  で、確率  $\pi$  の  $Bin(n, \pi)$  に従う  $X$  があるとき、

(i)  $X$  は  $n$  が大きい時、 $N(n\pi, n\pi(1 - \pi))$  に従う。

(ii)  $X/n (=p)$  は  $n$  が大きい時、 $N(\pi, \pi(1 - \pi)/n)$  に従う。

③ 標準化

正規分布に従うのなら、標準化した統計量を作成し、標本分布として  $N(0,1)$  を使えばいい。

$$z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1 - \hat{\pi}) / \left(\frac{1}{m} + \frac{1}{n}\right)}}$$

$$\hat{\pi} = \frac{m}{m+n} p_1 + \frac{n}{m+n} p_2$$

④ 仮説の判定

$z$  の値が標本分布のどこに位置しているかで帰無仮説との整合性を判断する。

### (3) 適合度の検定

観測値から作られた経験分布が、理論的に導出される期待される確率分布と同一だとみなしてよいか検定する方法。

#### ① 帰無仮説

帰無仮説： $O_i \sim \text{Bin}(n, \pi_i)$

対立仮説： $O_i \sim \text{Bin}(n, \pi_i)$ とは考えられない。

#### ② ド・モアブル＝ラプラス (De Moivre-Laplace) の定理

サイズが  $n$  で、確率  $\pi$  の  $\text{Bin}(n, \pi)$  に従う  $X$  があるとき、

(i)  $X$  は  $n$  が大きい時、 $N(n\pi, n\pi(1-\pi))$  に従う。

(ii)  $X/n (=p)$  は  $n$  が大きい時、 $N(\pi, \pi(1-\pi)/n)$  に従う。

#### ③ 標準化

$O_i$  が正規分布に従うことから、平均を引き標準偏差で割ることで標準化する。

$$\frac{O_i - n\pi_i}{\sqrt{n\pi_i(1-\pi_i)}} \sim N(0,1)$$

ただし、相対頻度表では、すべてのセルが独立に動くわけではないので、調整が必要。

$$GOF = \sum_{i=1}^k \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$$

#### ④ 標本分布

標準正規分布に従う確率変数の二乗の和が従う分布が  $\chi^2$  分布である。

$$GOF \sim \chi^2(k-1)$$

#### ⑤ 仮説の判定

GOF の値が標本分布のどこに位置しているかで帰無仮説との整合性を判断する。

■ 分布の形がよく分かっていないもの：ブートストラップ法



問題点はここ！

標本分布の性質（輪郭や密度式）が分かっていないと、  
標準誤差、バイアス、信頼区間 が分からない！

(1) ブートストラップ法 Bootstrapping

母集団分布の推定値と見なされた経験分布から復元抽出で得た標本で推定値の信頼性を評価する手法。

① 経験分布 Empirical Distribution

母集団分布  $f$  から得られた標本をもとに構築した相対頻度表（確率分布）  $\hat{f}$ 。母集団分布の推定値と見なす。

② ブートストラップ標本 Bootstrap Sample

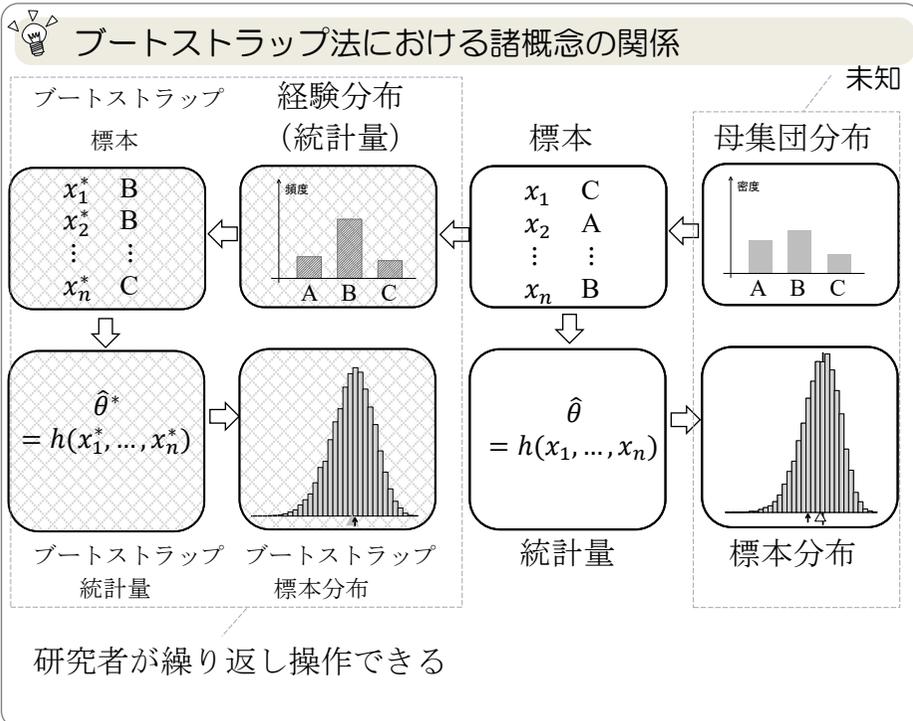
経験分布から抽出された標本のこと。

③ ブートストラップ統計量 Bootstrap Statistic

ブートストラップ標本から計算された量。

④ ブートストラップ標本分布 Bootstrap Sampling Distribution

ブートストラップ統計量が従う標本分布のこと。

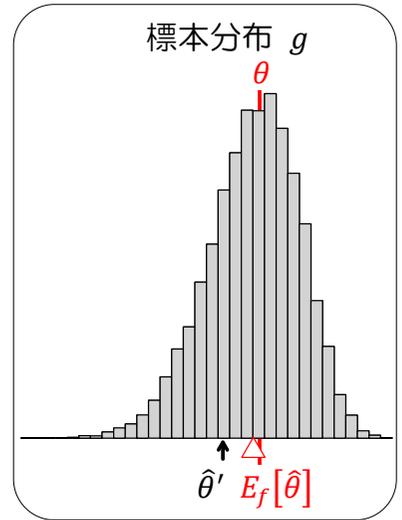


(2) ブートストラップ法によるバイアスと分散 (標準誤差)

① 母集団分布  $f$  に基づく量

$$bias_f(\hat{\theta}) = E_f[\hat{\theta}] - \theta$$

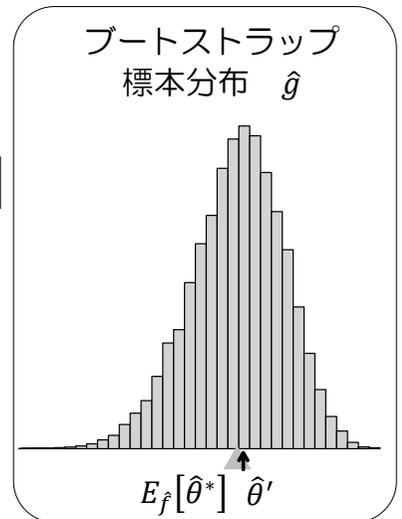
$$\sigma_{\hat{\theta}}^2 = E_f[(\hat{\theta} - E_f[\hat{\theta}])^2]$$



② 経験分布 (母集団分布の推定量  $\hat{f}$ ) に基づく量

$$bias_{\hat{f}}(\hat{\theta}^*) = E_{\hat{f}}[\hat{\theta}^*] - \hat{\theta}'$$

$$\sigma_{\hat{\theta}^*}^2 = E_{\hat{f}}[(\hat{\theta}^* - E_{\hat{f}}[\hat{\theta}^*])^2]$$



※ 復元抽出 Sampling with Replacement

一度抽出したデータも次の抽出の対象になる抽出。

(3) ブートストラップ信頼区間

複数の方法があるが、最も簡便なものにパーセンタイル法 (percentile method) がある。

$$CI_{bs} = (\hat{\theta}' - \hat{g} \text{ の } 0.975 \text{ の点}, \hat{\theta}' - \hat{g} \text{ の } 0.025 \text{ の点})$$