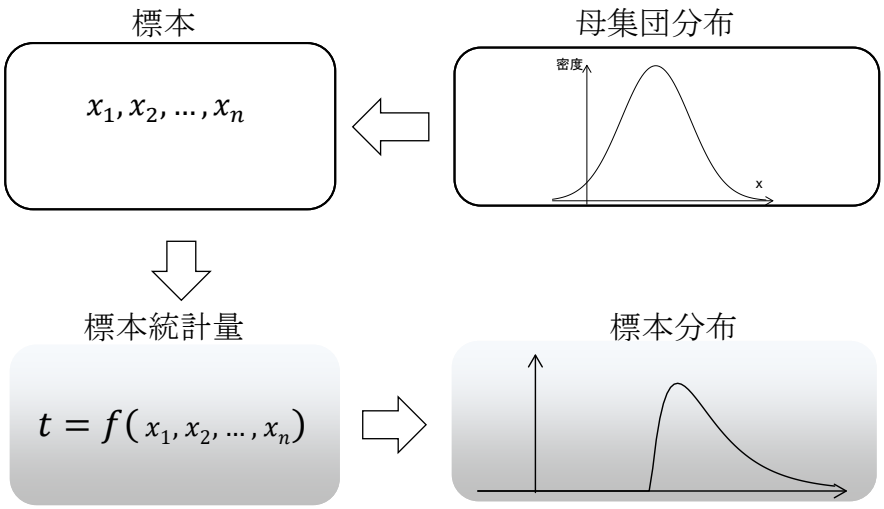
第1講　推測統計学の四つの重要概念（名義尺度編）

学びのポイント

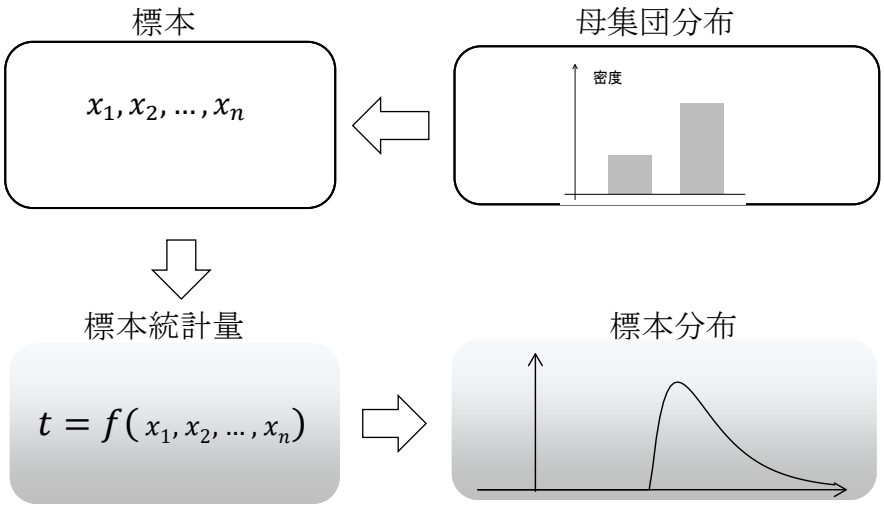
- 比率尺度データだけでなく、名義尺度データでも**母集団**、**標本**、**統計量**、**標本分布**という四つの概念が想定されることが分かる。
- 名義尺度の従属変数が従う分布には、**ベルヌーイ分布**、**二項分布**、**カテゴリカル分布**、**多項分布**が存在することが理解でき、それらの相互の関係が説明できる。
- 二項分布のサイズ n の値を大きくした極限に、**正規分布**、**ポワソン分布**が出現することが理解でき、どのようなときにどちらになるのか説明ができる。
- 標準正規分布に従う確率変数の和が従う分布としてχ^2分布という分布が提案されていることが分かる。
- 名義尺度データの主な関心は頻度を扱うことであることが分かり、頻度には、**粗頻度**と**相対頻度**という二つの区別があることが理解できる。
- 相対頻度（表）に基づく統計量として**情報量**があり、情報量の表に対してエントロピーが定義されることが分かる。
- **エントロピー**は、確率分布のデコボコさを測る指標であることが説明できる。
- 二つ以上のカテゴリーが存在する相対頻度表に対して、**結合エントロピー**や**条件付エントロピー**が定義されることが分かる。
- 確率分布の距離を測る統計量に、**KL 情報量**、**JS 情報量**、**相互情報量**、**χ^2値**などが利用されることが分かり、その違いを説明できる。
- 統計量が従う標本分布が理論的に明白ではないときに、**ブートストラップ法**を用いて標本分布を推定するアプローチがあることが理解できる。

見取り図

【前期】


標本
母集団分布
$x_1, x_2, \ldots, x_n$
密度
x
標本統計量
標本分布
$t = f(x_1, x_2, \ldots, x_n)$

【後期】


標本
母集団分布
$x_1, x_2, \ldots, x_n$
密度
標本統計量
標本分布
$t = f(x_1, x_2, \ldots, x_n)$

■ 目標

前期の「言語統計学 A」では、t 検定、重回帰分析などの具体的な統計モデルを扱う前に、これらのモデルで利用される基本的な統計量の話をしていました。

この「言語統計学 B」でも最終的な目標は第 2 講以降で扱う統計モデルを熟知して、使いこなすための知識を身に着けてもらうことですが、まずは前期同様この第 1 講では、名義尺度変数に対して計算される基本的な統計量を学びます。前期の比率尺度データについては、次の四つの概念が明確に区別されていました。

（1）標本 Sample

得られたデータのことを標本（サンプル）と呼ぶ。多くの場合全てを報告するのではなく何らかの指標で標本全体を代表させる。

（2）母集団 Population

研究者が想定する「考えられる対象をすべて含む集団」。

（3）（標本）統計量 Sample Statistic

標本から計算される指標を（標本）統計量と呼ぶ。

（4）標本分布 Sampling distribution

母集団から標本がサンプリングされる際の確率的な揺らぎのせいで、統計量が見せる確率的挙動を表したもの。

これらの四つの概念は、後期で扱う名義尺度データについてもそのまま当てはまります。しかし、後期の主眼は、あるカテゴリーが何回登場したのかという「頻度」（あるいは「頻度表」）です。連続値（実数）を取る比率尺度データとは異なり、標本が離散値（自然数）しかとらないという制約を持つので、前期とは異なる扱いが必要です。

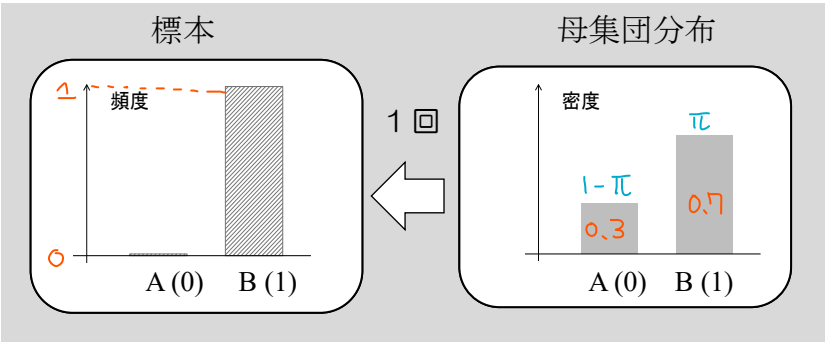
そこで、第 1 講の前半では、この頻度をもたらす確率分布たちを新しく学んでいきます。これらの分布は相互に関連しており、それぱかりか、前期に習った正規分布さえ今回習う分布から派生したものだという点が明らかになり、確率分布一般に対する理解も深まることでしょう。

第 1 講の後半では、頻度（表）にまつわる統計量とそれらが従う標本分布の想定について学んでいきます。名前にはなじみがないものが多いかもしれません。しかし、一つ一つの概念はつながっていて、丁寧に追いかけていくと、それらは非常に納得のいく理由で提唱されてきた概念であることが分かるでしょう。

📖 ノート１　母集団：代表的な確率分布

（１）ベルヌーイ分布 Bernoulli Distribution

これは 1 か 0 という二値の値を取る変数が従う確率分布。（名義尺度）
確率πで 1、確率 $1-\pi$ で 0 を取るように設計されている。



① 密度関数　$p(y|\pi) = \pi^y(1-\pi)^{1-y}$

② 期待値　$E[y] = \pi$

③ 分散　$Var[y] = \pi(1-\pi)$

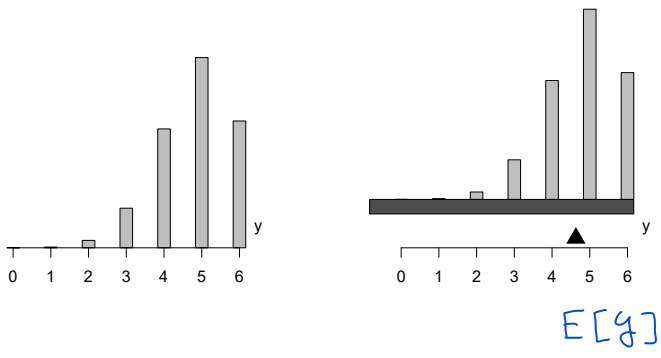
ベルヌーイ試行の具体例

（例１）□ NP の空欄が不定冠詞か定冠詞か。

（例２）「ろっかく」の変換が「六角」か「六画」か。

（例３）アンケート調査の解答が「はい」か「いいえ」か。

確率分布の期待値



Ⓟ記法：π と φ

これはローマ字 p に対応するギリシア文字

π（パイ pi）

φ（フィー phi）ファイ

Ⓟ記法：密度関数

① 定値な発想

π（y=1のとき）

1-π（y=0のとき）

② 巧妙な対応

場合わけを回避

(i) y=1 のとき

$\pi^y(1-\pi)^{1-y}$

$= \pi^1(1-\pi)^0$

$= \pi$

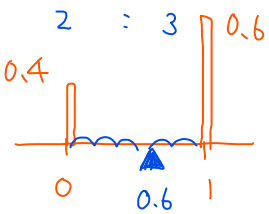
(ii) y=0 のとき

$\pi^y(1-\pi)^{1-y}$

$= \pi^0(1-\pi)^1$

$= 1-\pi$

Ⓟベルヌーイ分布の期待値



Ⓟ 変数変換

変数を何らかの関数で変換すること.

$f(x) = 2x+1$

x を $2x+1$ に変換

$g(y) = y^2$

y を y^2 に変換

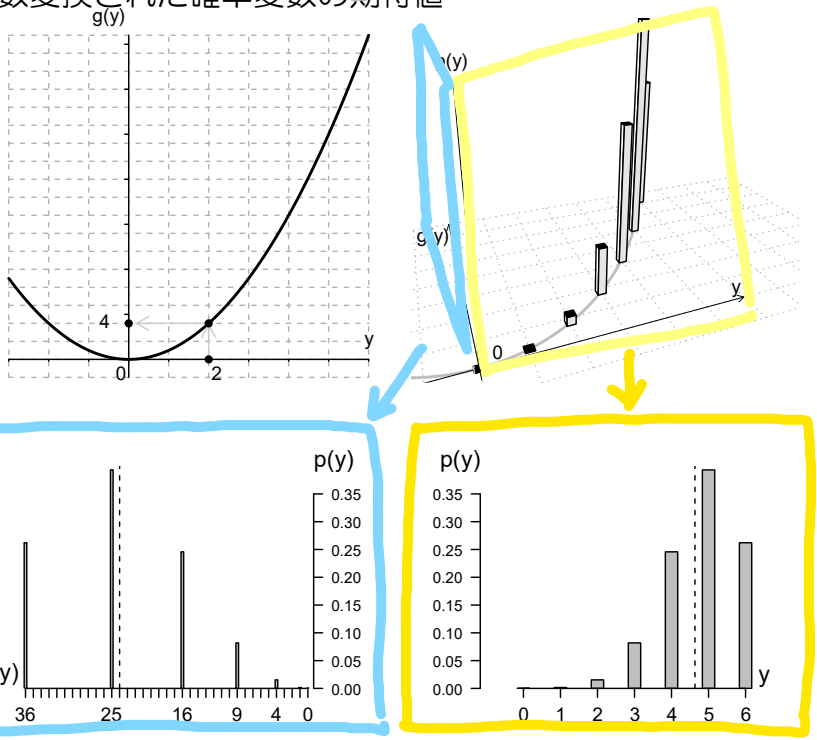
$g(y) = (y - E[y])^2$

↑

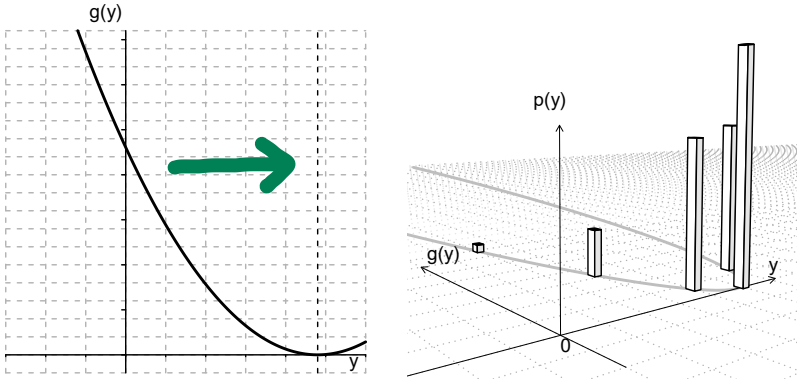
これが分散を計算するときに用いる変換

確率変数の変数変換と分散

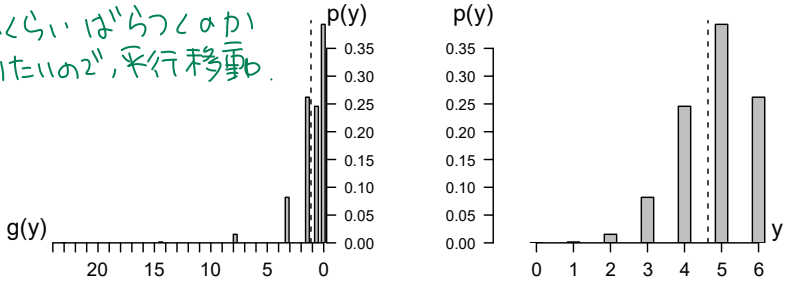
① 変数変換された確率変数の期待値



② 確率変数の分散

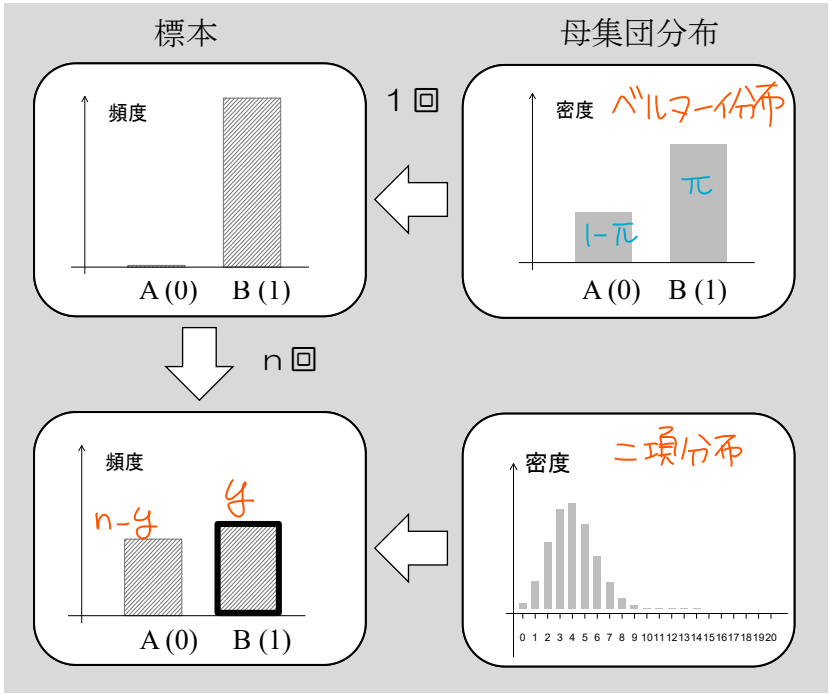


期待値のまわりにどのくらいばらつくのか知りたいので，平行移動.



（2）二項分布 Binomial Distribution

これは、確率 π で 1 を取るベルヌーイ試行を n 回行ったときに 1 が出る回数が従う確率分布。



① 密度関数　$p(y|n,\pi) = {}_nC_y\pi^y(1-\pi)^{n-y}$

② 期待値　$E[y] = n\pi$

③ 分散　$Var[y] = n\pi(1-\pi)$

二項分布に従う確率変数の具体例

（例 1）□ NP を 10 回観測したとき、そのうち何回が、空欄に不定冠詞を用いていたか。

（例 2）floor を 20 人のアメリカ人に発音してもらい、そのうち何人が最後の[r]を発音するか。

（例 3）「私は友達に贈り物をあげた」を 15 人のネイティブに翻訳してもらった時、そのうち何人が SVOO で表現したか。

（例 4）「走りません」「走らないです」を合計を 500 例集めた。そのうち何文が「ます」を使った表現か。

Ⓟ 二項分布とベルヌーイ分布

ベルヌーイ分布は，二項分布において くり返しの数 n を 1 に設定した場合。

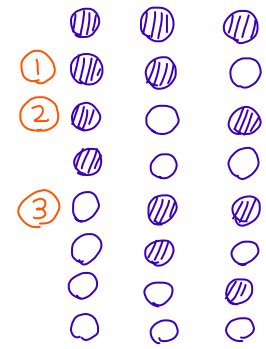
Ⓟ 記法 : $p(y|\cdots)$

① $p(\cdots)$ … の確率

② $y \mid n, \pi$ ↑ y given n and π　n と π が与えられたときの y.

Ⓟ 組み合わせ ${}_nC_y$

例えば，n = 3 のとき



の 8 つの可能性がある（一般に，n のとき，2^n）.

さらに，y = 2

① の確率　$\pi \times \pi \times (1-\pi)$

② の確率　$\pi \times (1-\pi) \times \pi$

③ の確率　$(1-\pi) \times \pi \times \pi$

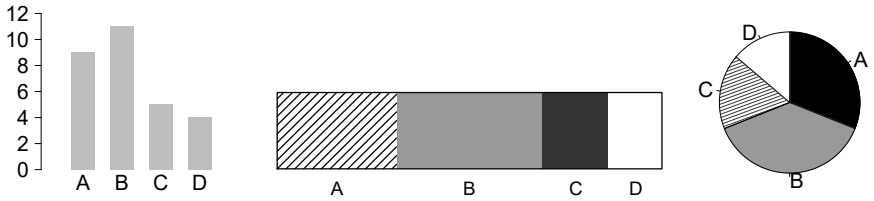
このため．

$3 \times \pi^2 (1-\pi)$

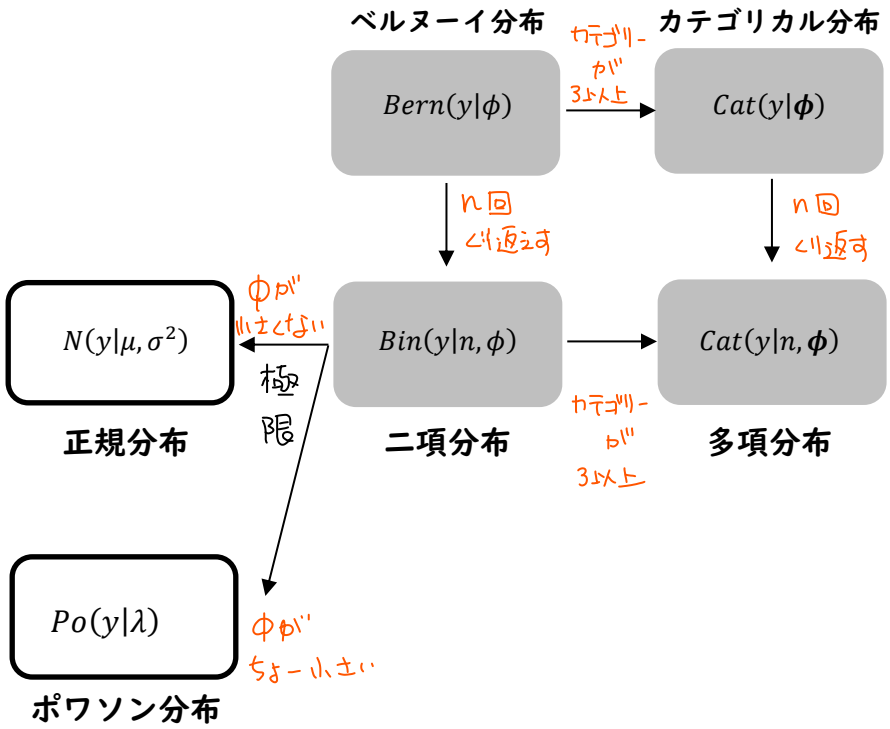
$= {}_3C_2$

視覚化

① 棒グラフ　② 帯グラフ　③ 円グラフ

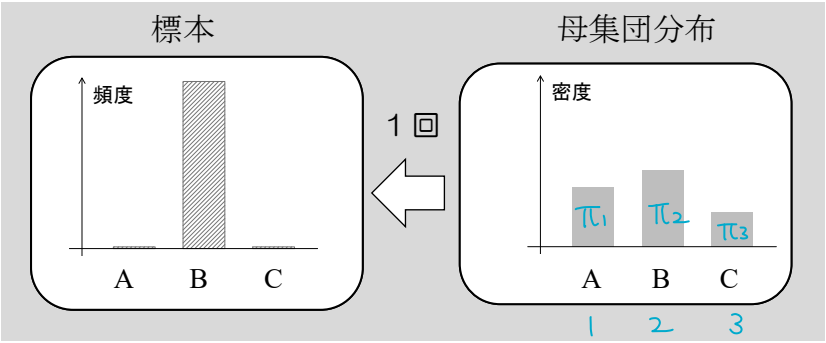


確率分布の関係（１）



（3）カテゴリカル分布 Categorical Distribution

これは、値がvになる確率がπ_vであるV個の離散値から一つ取り出す試行を1回行った時それぞれの値が出る回数が従う分布。



① 密度関数　$p(\boldsymbol{y}|\boldsymbol{\pi}) = \pi_1^{y_1}\pi_2^{y_2}\cdots\pi_V^{y_V} = \prod_{v=1}^{V}\pi_v^{y_v}$

② 期待値　$E[y_v] = \pi_v$

③ 分散　$Var[y_v] = \pi_v(1-\pi_v)$

ベクトル表記

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_V \end{pmatrix}$$

カテゴリカル分布に従う確率変数の具体例

（例1）drive 代名詞 crazy 構文をコーパスで1例採取したとき、その代名詞が me であるか。

（例2）尊敬語を含む韓国語の例文を一つ翻訳するとき、「お…になる」、「…なさる」、「お…なさる」という三つのタイプの中で「…なさる」構文が使われるかどうか。

（例3）日本語の学習者が rapidity という単語を訳すとき、「速さ」「速度」「スピード」の中で「速さ」が選択されるか。

Ⓟ カテゴリカル分布とベルヌーイ分布

ベルヌーイ分布は、カテゴリカル分布のカテゴリー数を2に限定した特殊な場合。

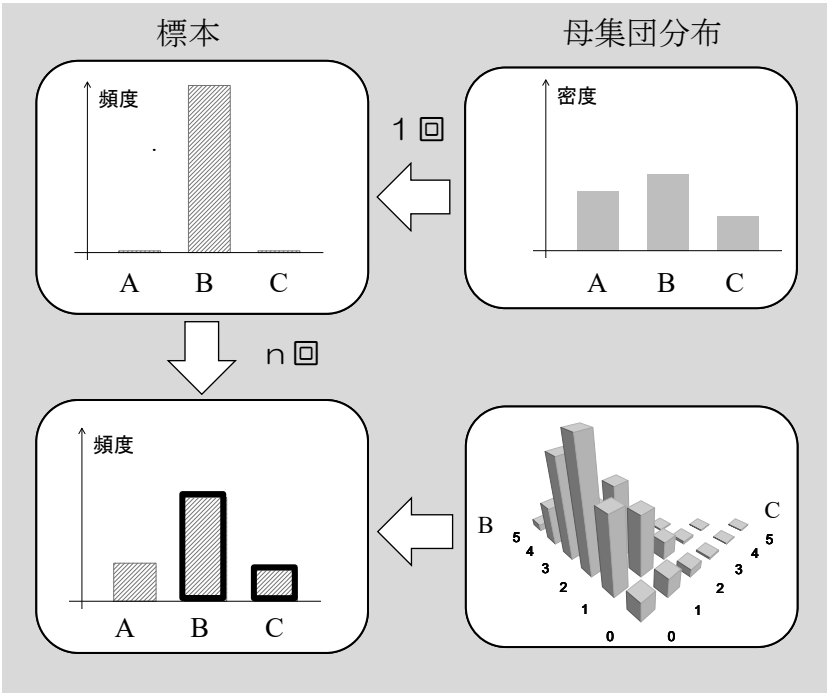
Ⓟ 確率密度分布

$\pi_1^{y_1}$ $\pi_2^{y_2}$ $\pi_3^{y_3}$

$y_1 = 1$ のとき
$(\Leftrightarrow y_2 = 0 \wedge y_3 = 0)$
$\pi_1^1\ \pi_2^0\ \pi_3^0$
$= \pi_1 \times 1 \times 1$
$= \pi_1$

（4）多項分布 Multinomial Distribution

これは、値がvになるπ_vであるV個の離散値から一つ取り出す試行を n 回行った時それぞれの値が出る回数が従う分布



① 密度関数　$Multinom(y|n,\boldsymbol{\phi})$

$$= \frac{n!}{y_1!\,y_2!\cdots y_V!}\pi_1^{y_1}\pi_2^{y_2}\cdots\pi_V^{y_V}$$

$$= \frac{n!}{y_1!\,y_2!\cdots y_V!}\prod_{v=1}^{V}\pi_v^{y_v}$$

② 期待値　$E[y_v] = n\pi_v$

③ 分散　$Var[y_v] = n\pi_v(1-\pi_v)$

多項分布に従う確率変数の具体例

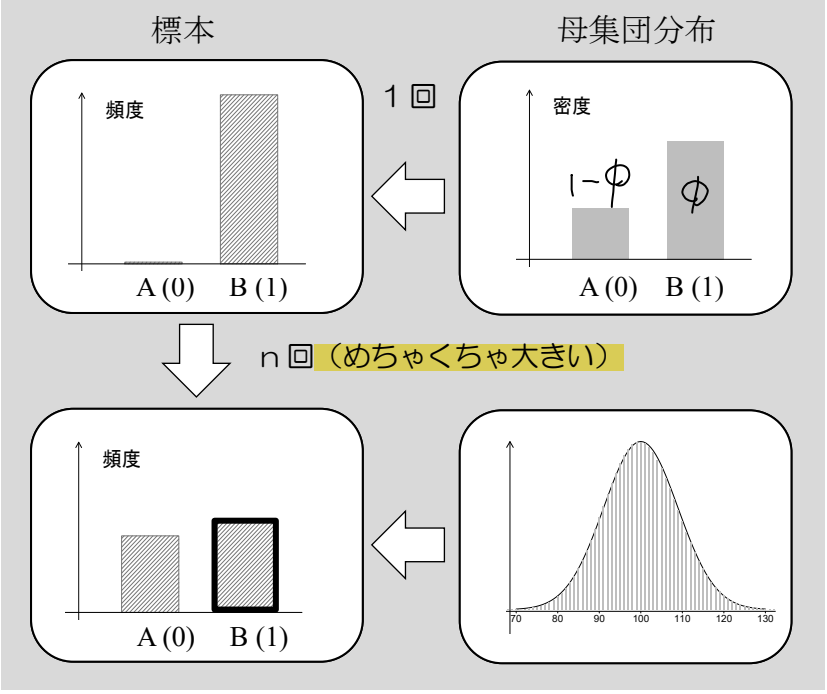
（例１）drive 代名詞 crazy 構文をコーパスで 100 例採取したとき、そのうち何例が me という代名詞を取っているか。

（例２）「お…になる」、「…なさる」、「お…なさる」という三つのタイプの尊敬語を使った構文を 532 例集めたとき、「…なさる」構文を使うのは何例か。

（５）正規分布 Normal Distribution

これは、「稀ではない現象」を「大量に観測した」際に二項分布の極限として登場する確率分布。

（＝ランダムな誤差が積みあがると出現する分布）



① 密度関数　$p(y|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$

② 期待値　$E[y]=\mu$

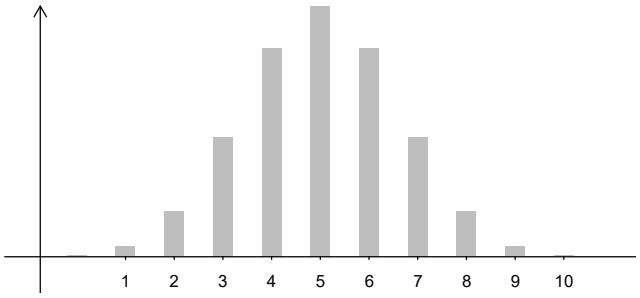
③ 分散　$Var[y]=\sigma^2$

二項分布が正規分布に近づいていくということ

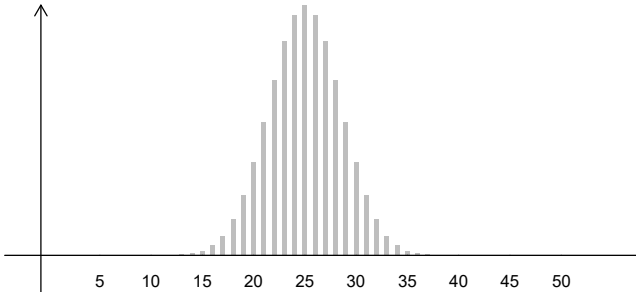
二項分布$B(y|n,\phi)$に従う確率変数Xはϕがそこそこ大きく、かつ、n が大きいとき、近似的に$N\left(n\phi, n\phi(1-\phi)\right)$に従う。

$\phi = 0.5$のとき

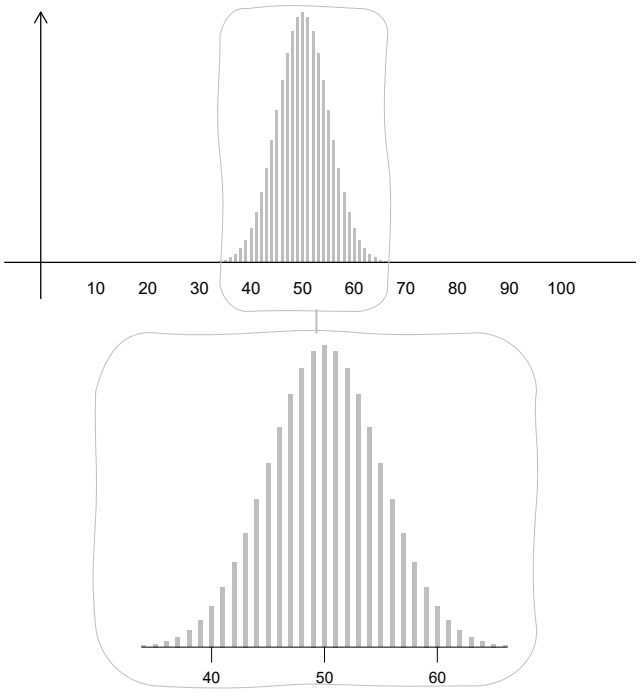
$n = 10$



$n = 50$

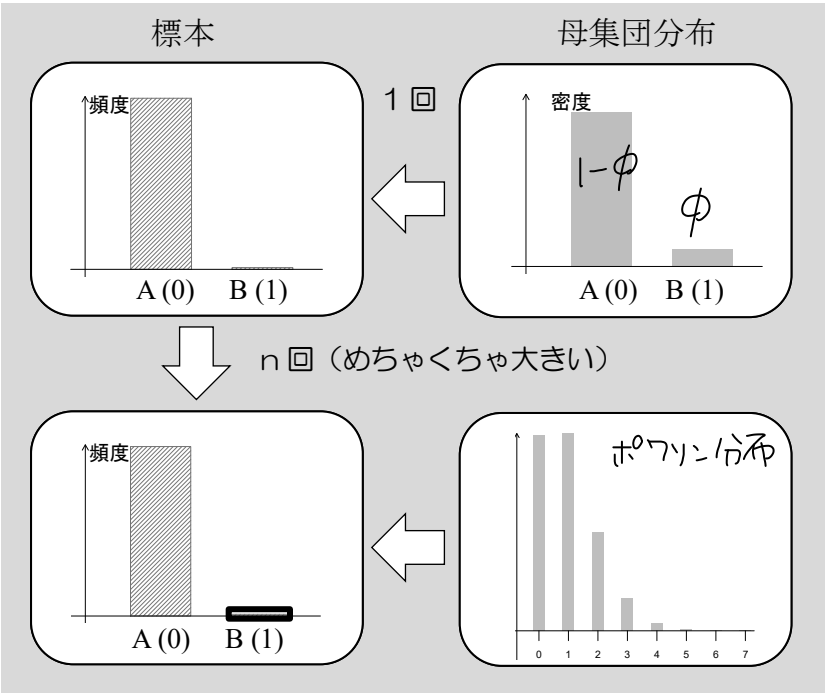


$n = 100$



（6）ポワソン分布 Poisson Distribution

これは、「稀な現象」を「大量に観測した」際に、得られる発生回数が従う二項分布の極限として登場する確率分布。



① 密度関数　　$Po(y|\lambda) = \frac{\lambda^y}{y!}\exp[-\lambda]$

② 期待値　　$E[y] = \lambda$

③ 分散　　$Var[y] = \lambda$

期待値がλなので「単位時間あたりに平均λ回起こる現象が、単位時間にy 回起きる確率現象」のモデルとして使われる。

ポワソン分布に従う確率変数の具体例

（例１）〇×新聞記事の一面で「言語」という単語は何回登場するか。

（例２）ある釣り場で１時間釣りをしたときに、何匹の魚を釣ることができるか。

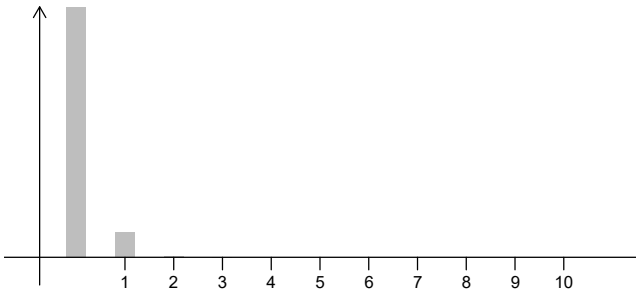
（例３）一年間で何回兵士が馬に蹴られて怪我をしてしまうか。

二項分布がポワソン分布に近づいていくということ

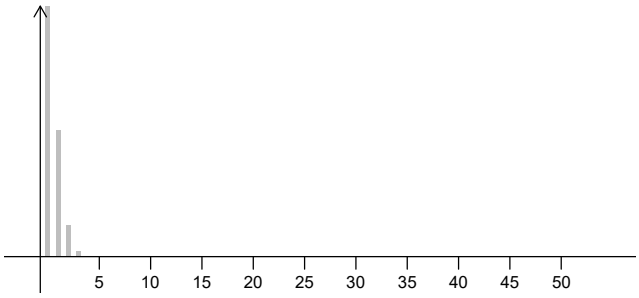
二項分布$B(y|n,\phi)$に従う確率変数 X はϕがとても小さく、かつ、n が大きいとき、近似的に$Po(n\phi)$に従う。

$\phi = 0.01$のとき

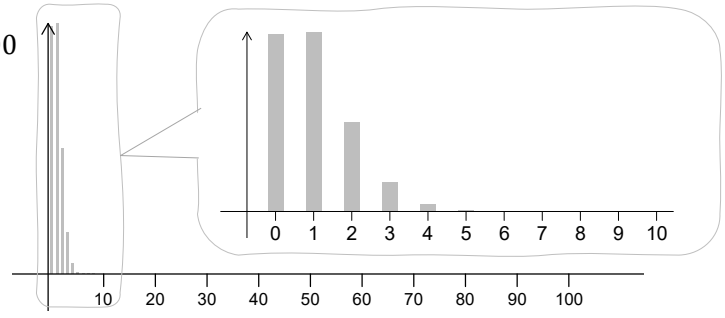
$n = 10$



$n = 50$

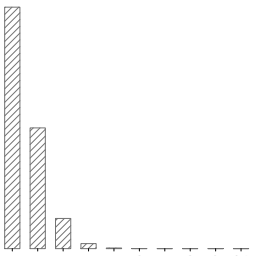


$n = 100$

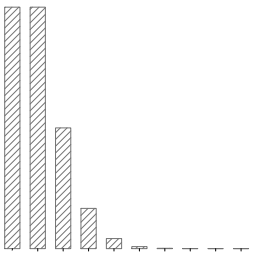


ポワソン分布の例

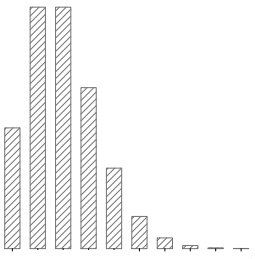
$\lambda = 0.5$



$\lambda = 1$



$\lambda = 2$

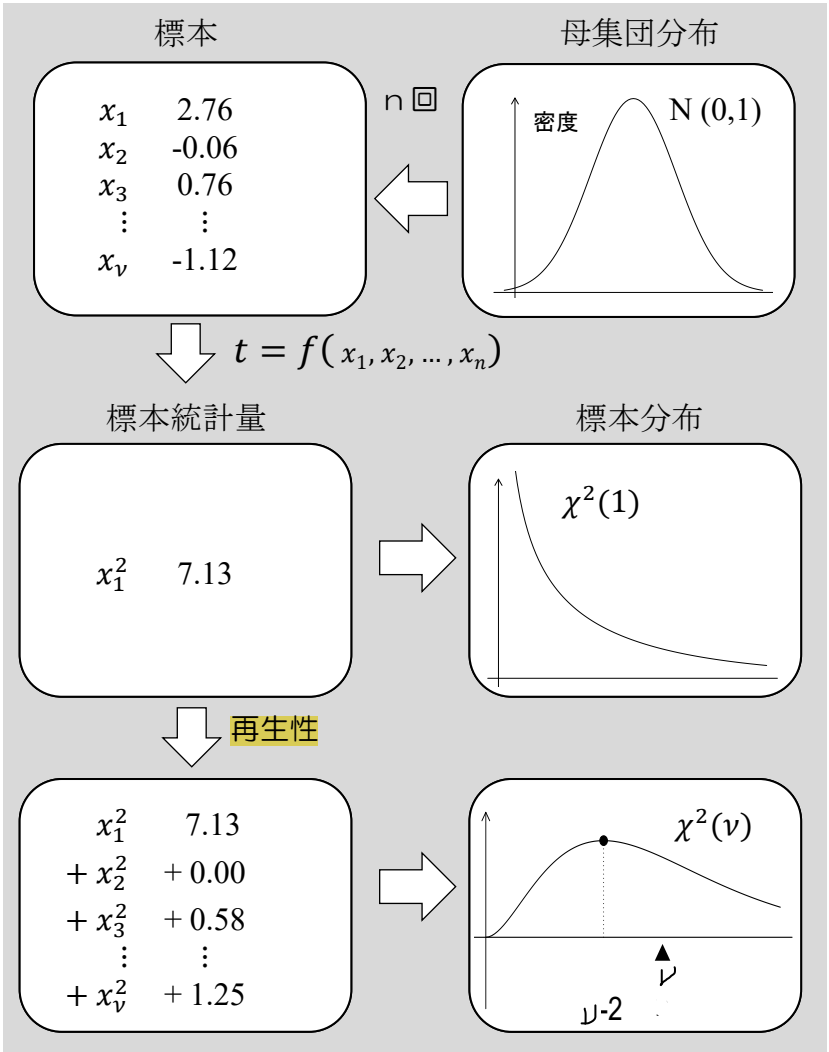


Ⓟ ギリシア文字 ニュー ν

これは，νの
ギリシア文字。

（7）χ^2分布 Chi-square Distribution

これは、N(0,1)に従うν個の確率変数の二乗和が従う分布。



① 密度関数 $\chi^2(y|\nu) = \dfrac{2^{-\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} y^{\frac{\nu}{2}-1} \exp\left[-\dfrac{y}{2}\right]$

② 期待値 $E[y] = \nu$

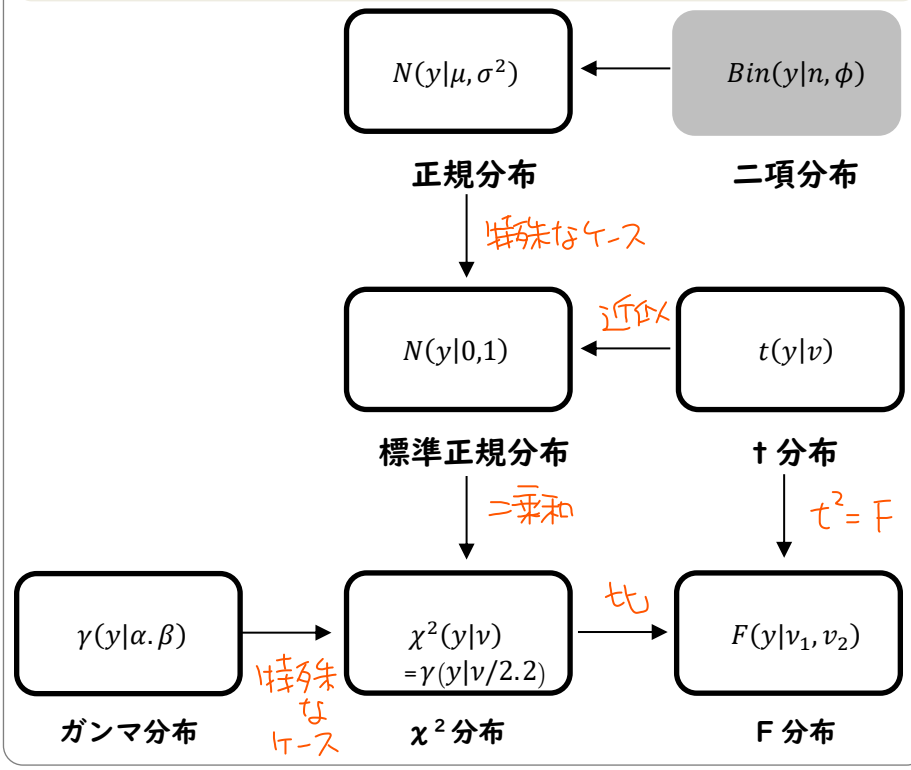
③ 分散 $Var[y] = 2\nu$

標準正規分布

これは、平均が０，分散が１の正規分布。

$z_i \sim N(0,1)$

確率分布の関係（２）



分布の再生性　Reproductive Property

同一分布に従う複数の独立な確率変数の和が元の分布に従うとき、その分布には再生性があると言います。

① 正規分布

$$\frac{\begin{array}{c}X_1 \sim N(\mu_1, \sigma_1^2) \\ X_2 \sim N(\mu_2, \sigma_2^2)\end{array}}{X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)}$$

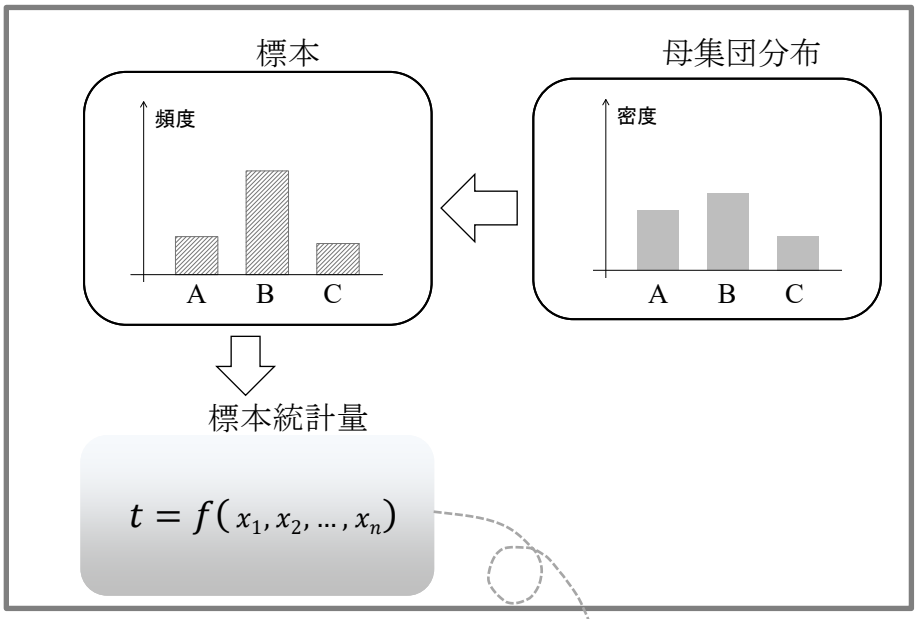
② 二項分布

$$\frac{\begin{array}{c}X_1 \sim Binom(n_1, \phi) \\ X_2 \sim Binom(n_2, \phi)\end{array}}{X_1 + X_2 \sim Binom(n_1 + n_2, \phi)}$$

③ χ^2分布

$$\frac{\begin{array}{c}X_1 \sim \chi^2(n_1) \\ X_2 \sim \chi^2(n_2)\end{array}}{X_1 + X_2 \sim \chi^2(n_1 + n_2)}$$

🕮 ノート２　基本的な統計量



■ 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
		↓
		KL 情報量 JS 情報量
自己相互情報量	→	相互情報量
χ^2値	→	χ^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。

crazy
25

$Freq(X = x) = x$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。

crazy
0.658

$RelFreq(X = x) = x$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。

crazy	*nuts*	*mad*	*out of one's mind*
25	*7*	*4*	*2*

合計 38回

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。

crazy	*nuts*	*mad*	*out of one's mind*
0.66	*0.18*	*0.11*	*0.05*

合計 1

（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。

	crazy	*nuts*	*mad*	*out of one's mind*	計
PRS	*25*	*7*	*4*	*2*	38（回）
PST	*17*	*2*	*2*	*0*	21（回）
					59（回）

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）。

	crazy	*nuts*	*mad*	*out of one's mind*
PRS	*0.42*	*0.12*	*0.07*	*0.03*
PST	*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）。

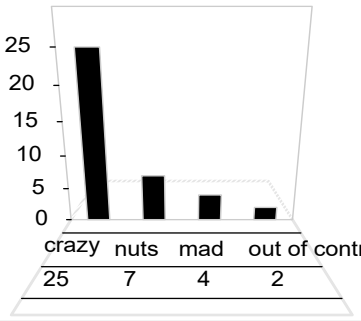
	crazy	*nuts*	*mad*	*out of one's mind*
PRS	*0.66*	*0.18*	*0.11*	*0.05*
PST	*0.45*	*0.05*	*0.05*	*0.00*

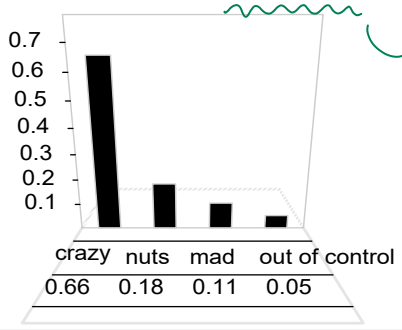
経験分布としての相対頻度表

相対頻度表は離散的な経験分布として解釈される。

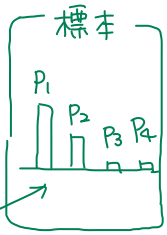
母集団における確率分布：出る確率

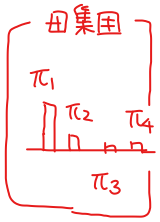
標本における確率分布：　出た確率（経験分布）



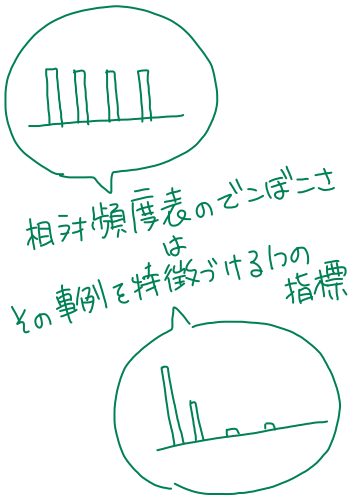


Ⓟ 母集団分布と経験分布

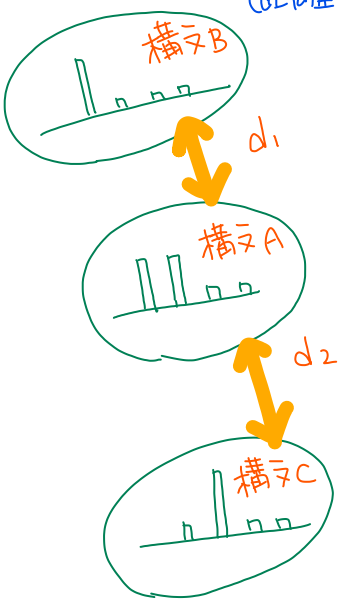




Ⓟ デコボコさを考える



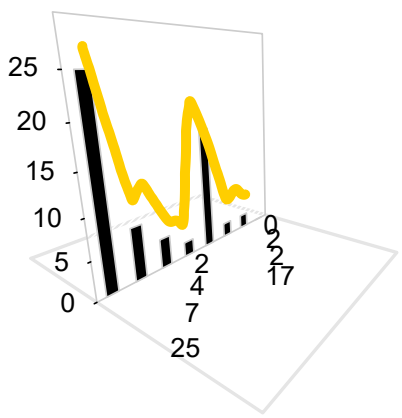
Ⓟ 相対頻度表の類似度
(距離)

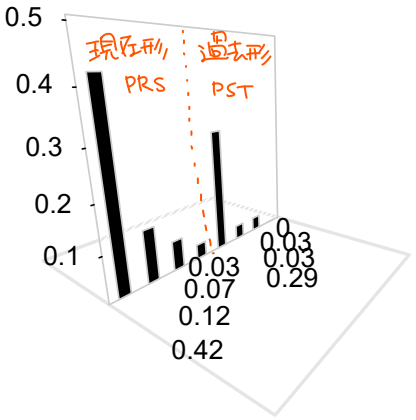


経験分布としての分割表（相対頻度）

（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。



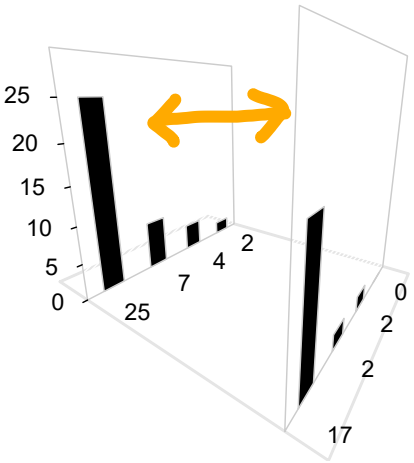


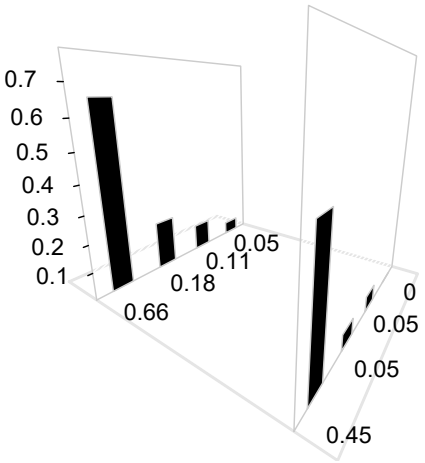
⇒　どのくらい「デコボコするのか」という点に関心が向く。

（☞　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）

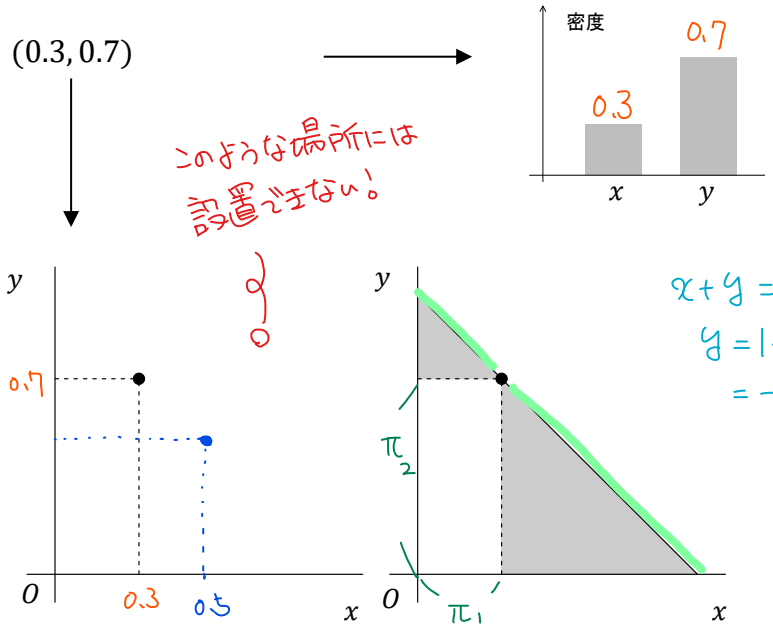
⇒　わけられた二つの確率分布はどのくらい違うのかという点に関心が向く。

（☞　KL／JS 情報量など）

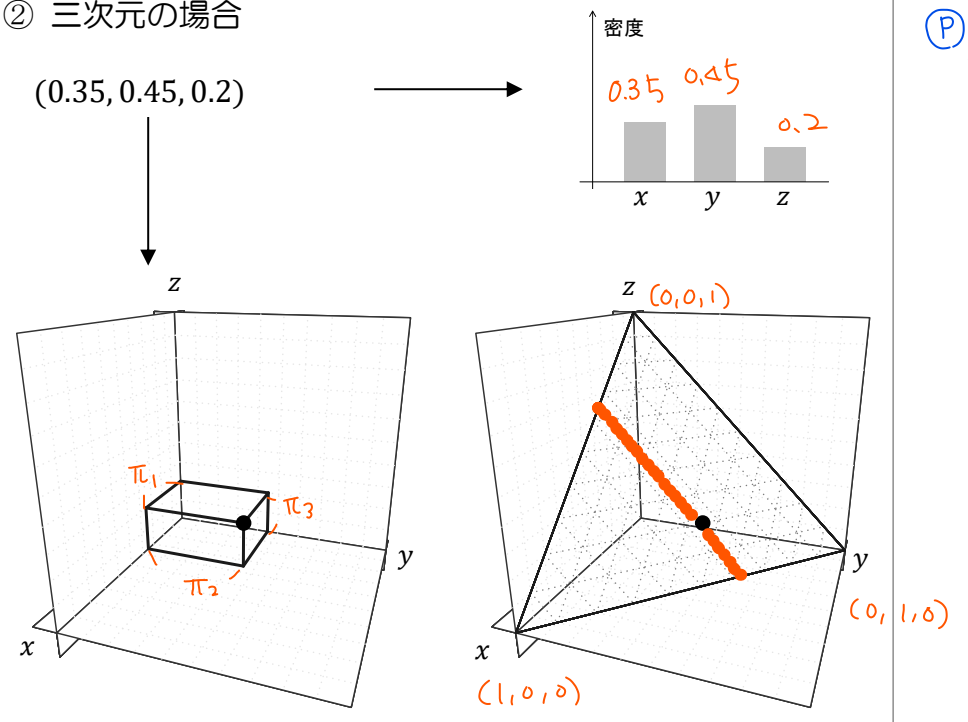
確率分布（相対頻度表）の集合の幾何的な解釈①

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

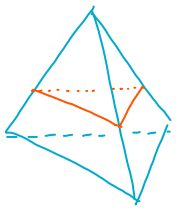


② 三次元の場合

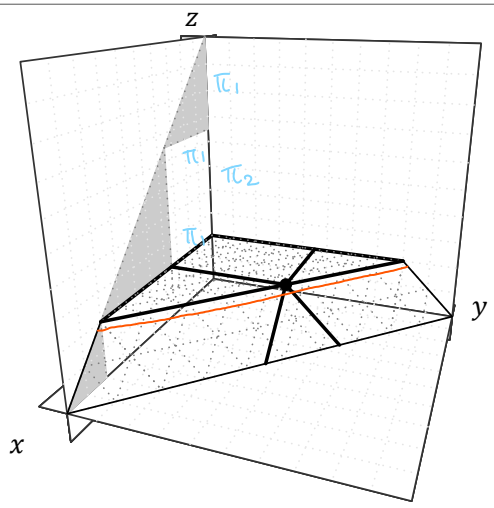


ⓟ 四次元の場合

正四面体

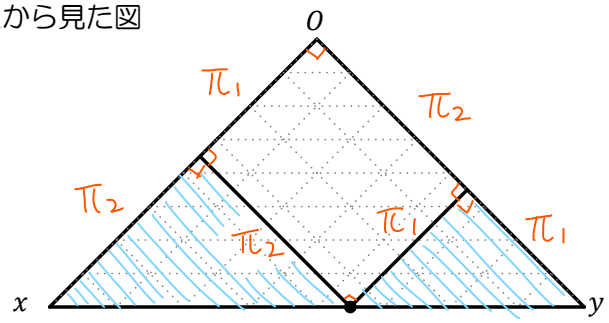


このようなものを確率単体 Simplex

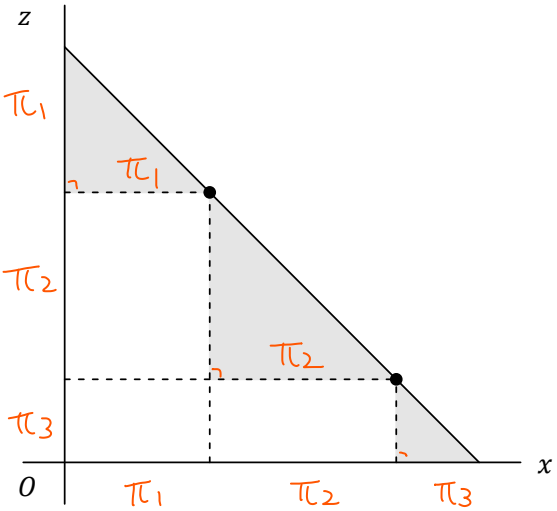


(?, ?, 0.2) 固定

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

Ⓟ 読み下し方

$I(X=x)$
X（出身地）が
x（大阪）をとる
ときの I（情報量）

（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。

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

$$= -\log p(X=x)$$

$= -\log p(x)$

※ 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。

出身地　大阪

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

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。

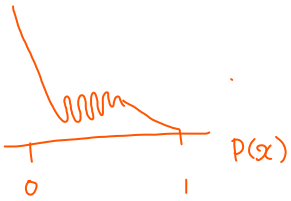
$$y = \log_3 x$$

① 掛け算（積）	② 指数	③ 対数
$1 \div 3 \div 3 \cdots \div 3$	$= 3^{-n}$	$-n = \log_3 3^{-n}$
⋮	⋮	⋮
$1 \div 3 \div 3$	$= 3^{-2}$	$-2 = \log_3 3^{-2}$
$1 \div 3$	$= 3^{-1}$	$-1 = \log_3 3^{-1}$
→ 1	$= 3^{0}$	$0 = \log_3 1$
1×3	$= 3^{1}$	$1 = \log_3 3^{1}$
$1 \times 3 \times 3$	$= 3^{2}$	$2 = \log_3 3^{2}$
⋮	⋮	⋮
$1 \times \underbrace{3 \times 3 \times \cdots \times 3}_{n\text{個}}$	$= 3^{n}$	$n = \log_3 3^{n}$

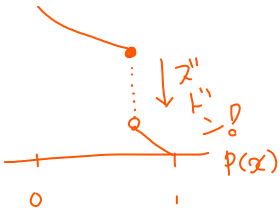
÷3　÷3　÷3　÷3

Ⓟ 単調に減少

次のようなものは
やめてほしい



Ⓟ 連続

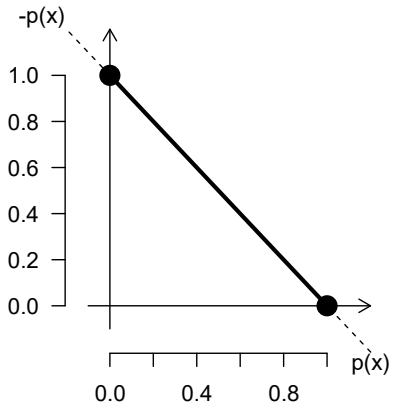


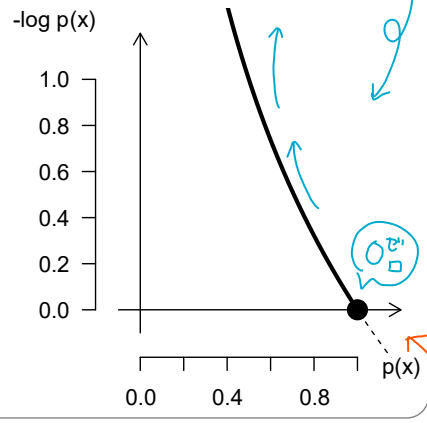
情報量の直感的な意味合い

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）

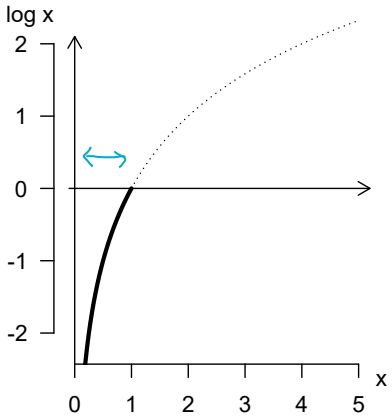
この三つの条件を満たす関数$I(x)$は、$I(x) = -\log_2 p(x)$という形しかありえないことが知られています（証明略）。



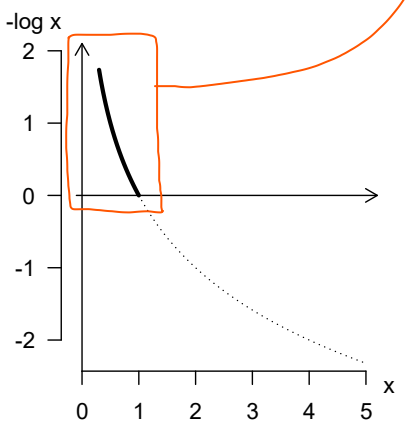


対数関数 log のグラフ

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。



上下反転



（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

crazy	*nuts*	*mad*	*out of one's mind*
0.658	*0.184*	*0.105*	*0.052*

② 情報量の表

crazy	*nuts*	*mad*	*out of one's mind*
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）

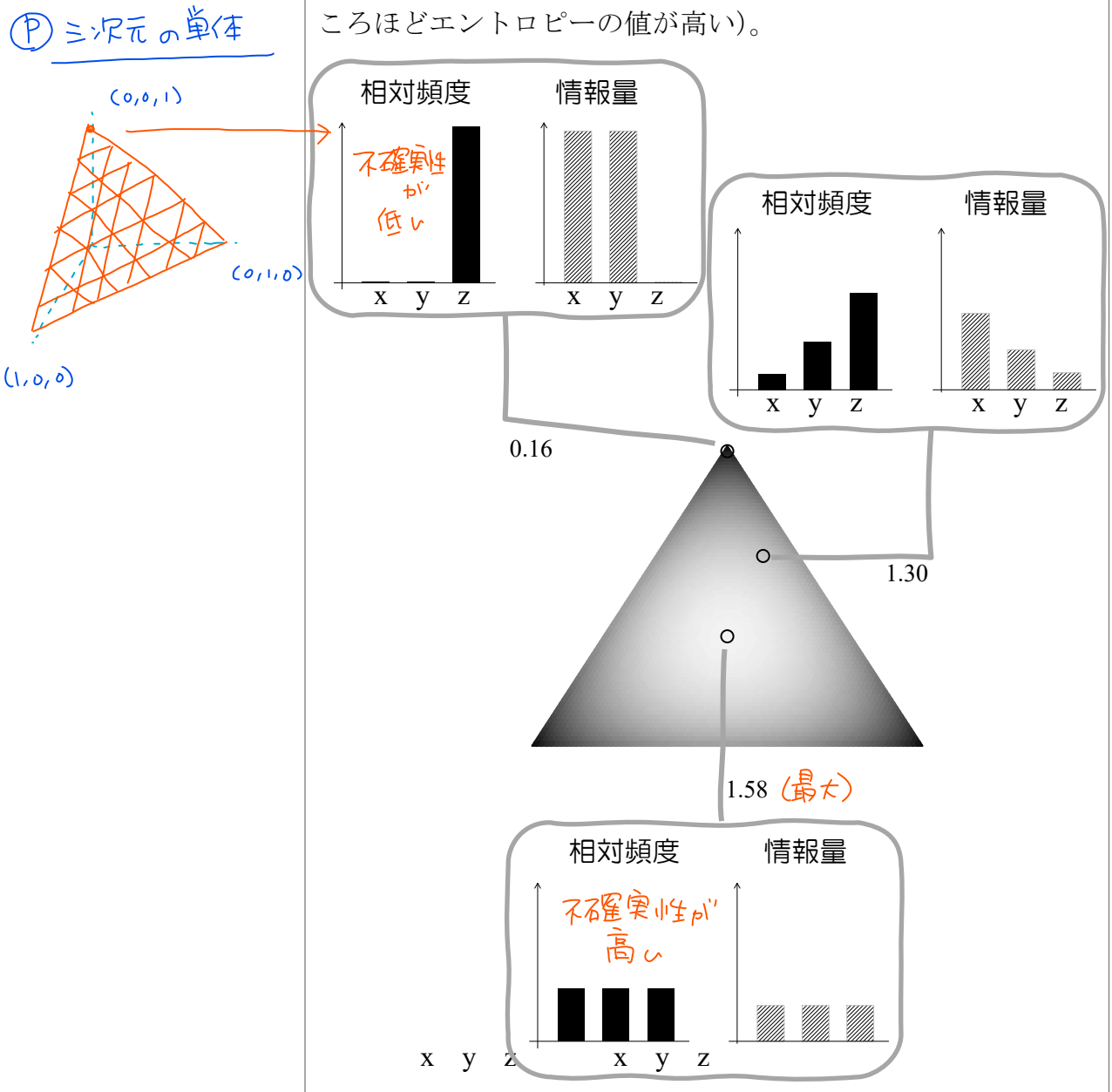
↓経験的な確率分布↓　　↓

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。

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

※ 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

$H(x) = p(x_1) \times (-\log p(x_1)) + p(x_2) \times (-\log p(x_2)) + \cdots + p(x_n) \times (-\log p(x_n))$

Ⓟ 記法：ギリシア文字H

H：ギリシア文字エータの大文字

Ⓟ 重みをつける意義

× $(0 + 20 + 50) \div 3 = \frac{70}{3} = 23.3$

○ $(90 \times 0 + 5 \times 20 + 5 \times 50) \div 100 = 35$

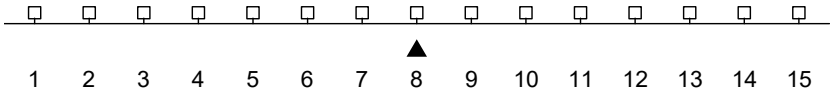
$0.9 \times 0 + 0.05 \times 20 + 0.05 \times 50$

確率という重みをつけている

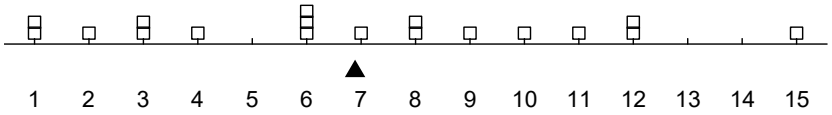
エントロピーの直感的な意味合い：重心

① 線分の重心

（重みがすべて同じとき）



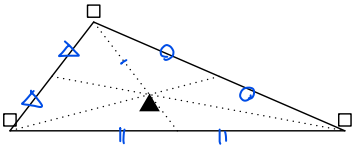
（重みがまちまちのとき）



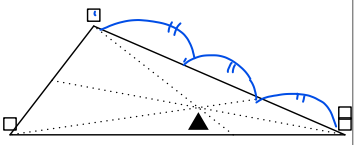
② 三角形の重心

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交点である。

（重みがすべて同じとき）

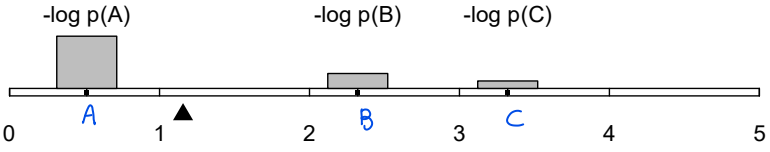


（重みがまちまちの時）



③ 情報量の重心

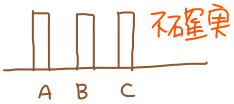
	単語 A	単語 B	単語 C
情報量	-log 0.7	-log 0.2	-log 0.1
確率（重み）	0.7	0.2	0.1



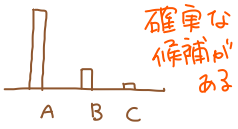
$$
\begin{aligned}
\mathrm{H}(P) &= \sum_{x\in X} -p(X=x)\log p(X=x) \\
&= p(X=A) - \log p(X=A) \\
&\quad + p(X=B) - \log p(X=B) \\
&\quad + p(X=C) - \log p(X=C)
\end{aligned}
$$

Ⓟ エントロピーの性質

(ケース1) 最大！



(ケース2) 小さい



（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。

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

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（エントロピー）。

＝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」を知ることで得られるエントロピーをそれぞれ$H(\mathrm{X})$、$H(\mathrm{Y})$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（エントロピー）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

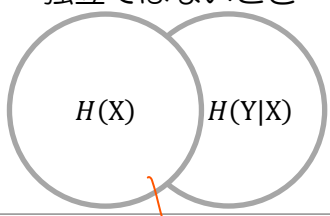
$$0\le H(X,Y)\le H(X)+H(Y)$$

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

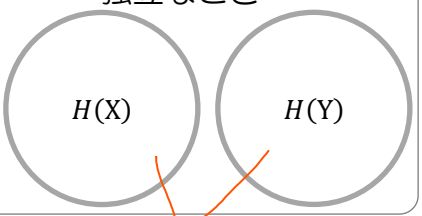
（性質２）条件付エントロピーとの関係

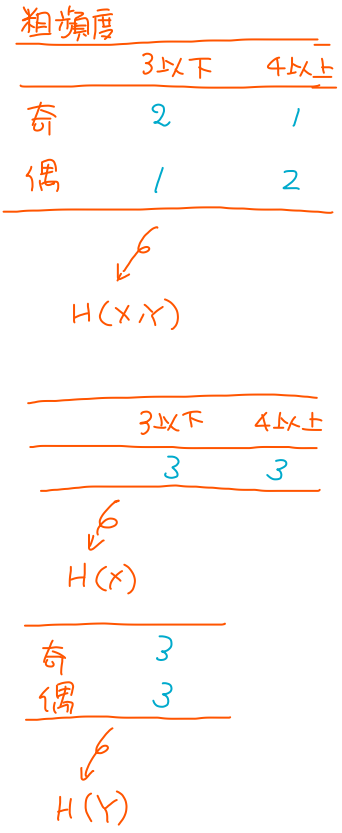
$$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)$$

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

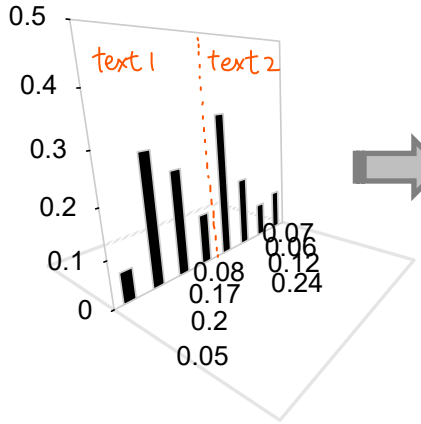
① 分割表（粗頻度）

	make	*take*	*do*	*have*
text 1	*4*	*17*	*14*	*7*
text 2	*20*	*10*	*5*	*6*

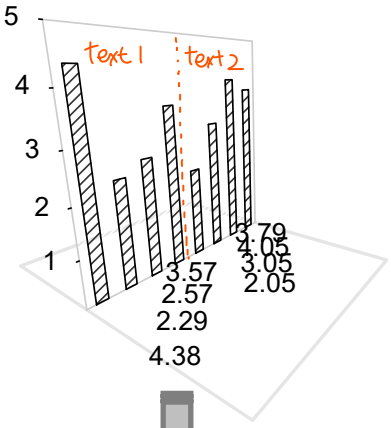
② 分割表（相対頻度）

	make	*take*	*do*	*have*	合計
text 1	*0.05*	*0.20*	*0.17*	*0.08*	*0.51*
text 2	*0.24*	*0.12*	*0.06*	*0.07*	*0.49*

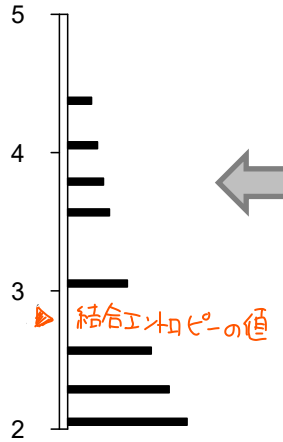
③ 同時確率密度分布
（相対頻度の分割表）

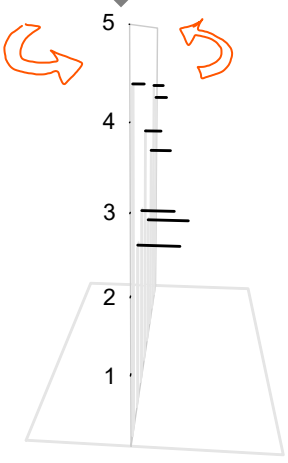


④ 情報量を算出



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。

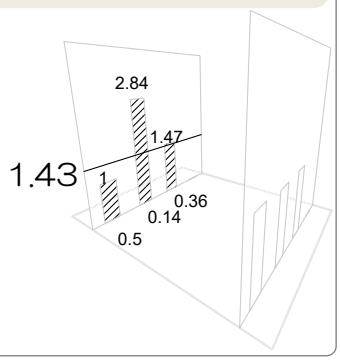
$$\mathrm{H}(X|Y=y) = \sum_{i=1}^{n} -p(x|y)\log p(x|y)$$

晴|曇=夏

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



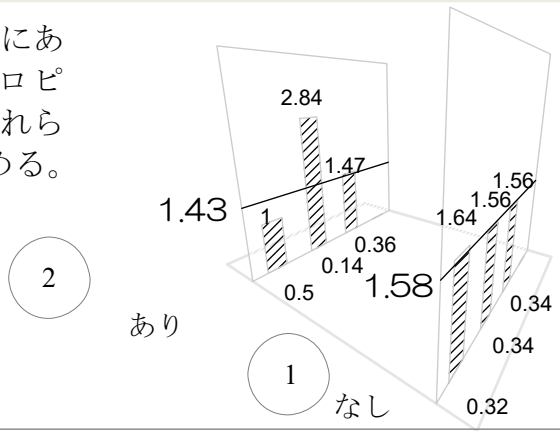
② Yが与えられた時のXの条件付きエントロピー

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。

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

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。



Ⓟ条件つき確率

	晴れ	雨	計
夏	40	10	50
冬	20	30	50
計	60	40	100

P(晴) = 60/100 = 0.6

P(晴|夏) = 40/50 = 0.8

P(夏|晴) = 40/60 = 0.66

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

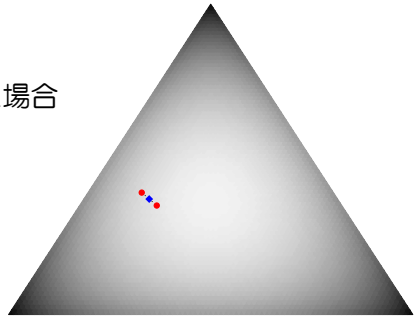
ケース１では、「教科書以外」の数が少ない分、「教科書以外」という条件の下もたらされたエントロピーはその分少なめに貢献させられている。

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。

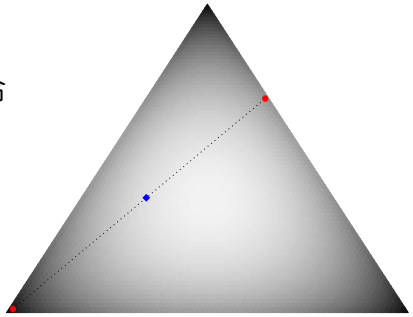
条件を考えてもあまり意味がなかった場合

	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)

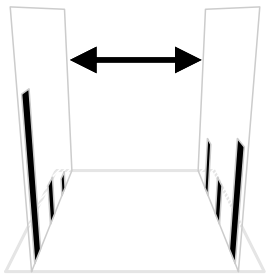
これは、二つの確率分布間の距離を表す統計量の一つ。相対エントロピー（Relative Entropy）とも呼ばれる。

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

確率分布間の距離

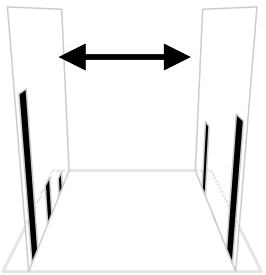
たくさん相対頻度表（経験的な確率分布）を手にした時、それらの中でどれが似ているのかという類似度（または距離）を測りたいという場面が、この後たくさん登場する。

ケース１



分布 p　　分布 q

ケース２



分布 p　　分布 r

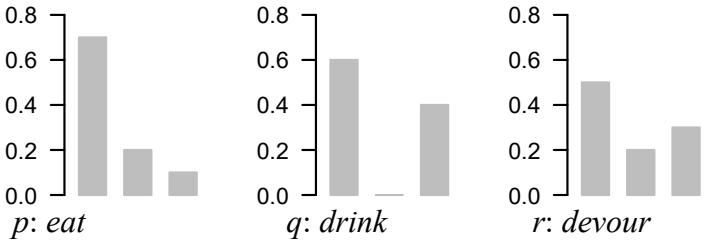
コーパス言語学における利用例

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

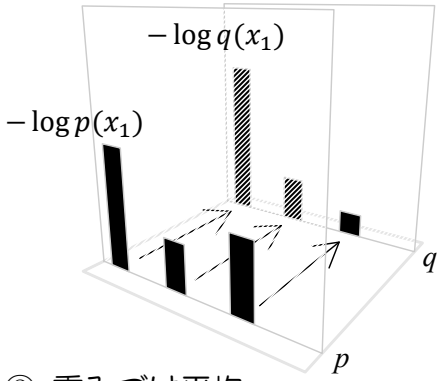
S V O の位置に来る相対頻度表（確率分布）を見る。p の分布を作った動詞と近い「距離」の動詞は意味が似てるはず。



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

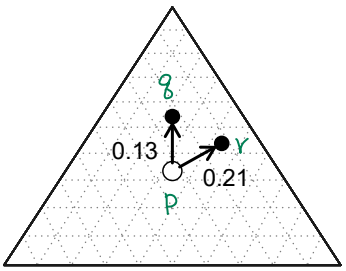


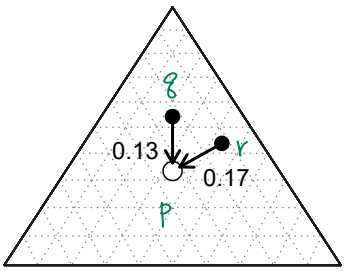
$$
\begin{aligned}
&\underset{q}{-\log q(x_1)} - \{\underset{p}{-\log p(x_1)}\} \\
&= -\log q(x_1) - \{-\log p(x_1)\} \\
&= -\log q(x_1) + \log p(x_1) \\
&= \log p(x_1) - \log q(x_1) \\
&= \log \frac{p(x_1)}{q(x_1)}
\end{aligned}
$$

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。

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

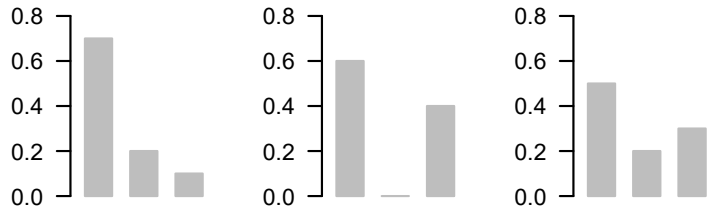
計量文体論における利用例

① モチベーション

文体上類似したテクストを見つけたい！

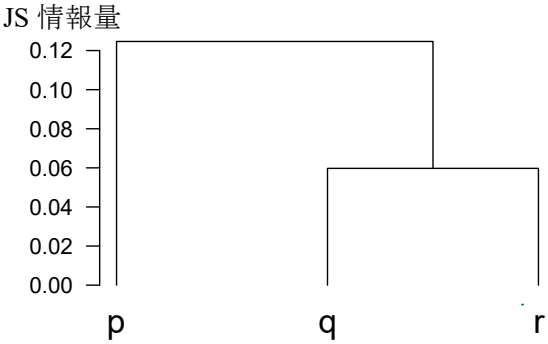
② 例

それぞれのテク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。



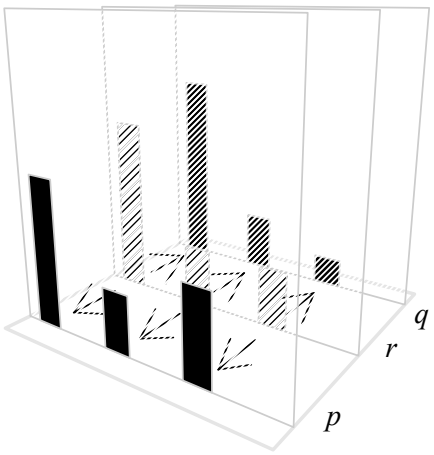
p: 吾輩は猫である q: こころ　　r: 高瀬舟

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。

「実際にコーパスから計算された状況」が

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

「理想的な独立な状況」からどのくらい遠いのかを測る

$$= \sum_{i}^{n}\sum_{k}^{K} p(x_i,y_k)\log\frac{p(x_i,y_k)}{p(x_i)p(y_k)}$$

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。

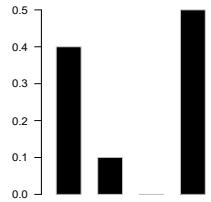
周辺確率の積

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

xとyが同時に出る確率

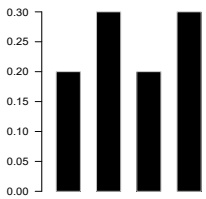
（A）経験分布の相対頻度表（同時確率）

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



P(Y=crazy) 周辺確率

同時確率 P(X=drive かつ Y=crazy)

P(X=drive) 周辺確率

ⓟ 変数の独立性

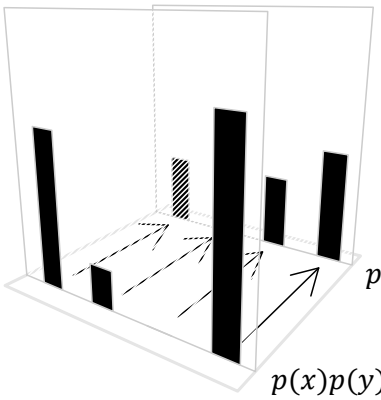
① 同時確率
これは，xとyが同時に得られる確率
P(x,y)

② 周辺確率
これは，二つの変数x,yが話題に挙がっているときに，一方を無視し，他方のみを対象にした確率
P(x)
P(y)

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。

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



$p(x,y)$：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。

相対頻度表（確率分布）全体が対象

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

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

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

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \quad \leftarrow \text{加点} \quad \leftarrow \text{減点}$$

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

↑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

Ⓟスムージング(平滑化)

	B	Not B
A	0	8
Not A	6	10

Add-one smoothing (+1)

	B	Not B
A	1	9
Not A	7	11

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

Ⓟ同時確率の分解

① 一般的なケース

$P(x,y)$ 周辺確率 条件つき確率 $= P(x) \times P(y|x)$

A/B	B	Not B	計
A			☆
Not A			
計	□		1

$P(x)$

$P(x) = \square$

$P(y|x) = \frac{▨}{\square}$

$P(x,y) = ▨ = \square \times \frac{▨}{\square}$

$= P(x)P(y|x)$

② x と y が独立なケース

$P(y|x) = P(y)$

x が与えられていようが x の値を知らなかろうが 確率は変わらない

$P(x,y) = P(x)P(y)$

自己相互情報量の別の捉え方

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$ ← 実際に観測した値 / ← 理想（独立）状態の値

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

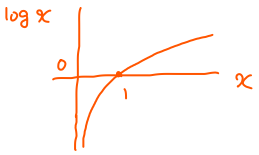
$$= \log \frac{p(y|x)}{p(y)}$$

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

↑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、$PMI = -\infty$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）。

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

- O_i　観測度数（粗頻度）
- $\hat{E}_i$　理論度数（粗頻度）：np_i

※ 両者に差がないという H0 の下でχ^2分布に従うため GOF の代わりにχ^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

	say	*nod*	*change*	*arrive*	合計
PST	*35*	*10*	*5*	*10*	**60**
PRF	*10*	*0*	*20*	*10*	**40**
合計	**45**	**10**	**25**	**20**	**100**

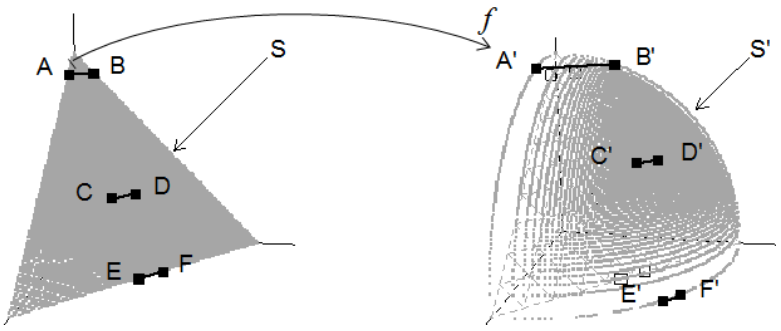
② 理論度数：理論上こうなっているだろうという理想データ　$\hat{E}_i$

	say	*nod*	*change*	*arrive*	合計
PST	*27*	*6*	*15*	*12*	**60**
PRF	*18*	*4*	*10*	*8*	**40**
合計	**45**	**10**	**25**	**20**	**100**

その他の分布間の距離を測る指標

① ヘリンジャー距離 Hellinger Distance

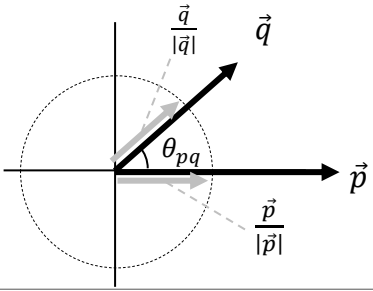
$$d_H(\boldsymbol{p},\boldsymbol{q}) = \sqrt{\sum_{i=1}^{n}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2/2}$$
$$= \frac{1}{\sqrt{2}}\left\|\sqrt{P}-\sqrt{Q}\right\|_{L2}$$



② コサイン類似度

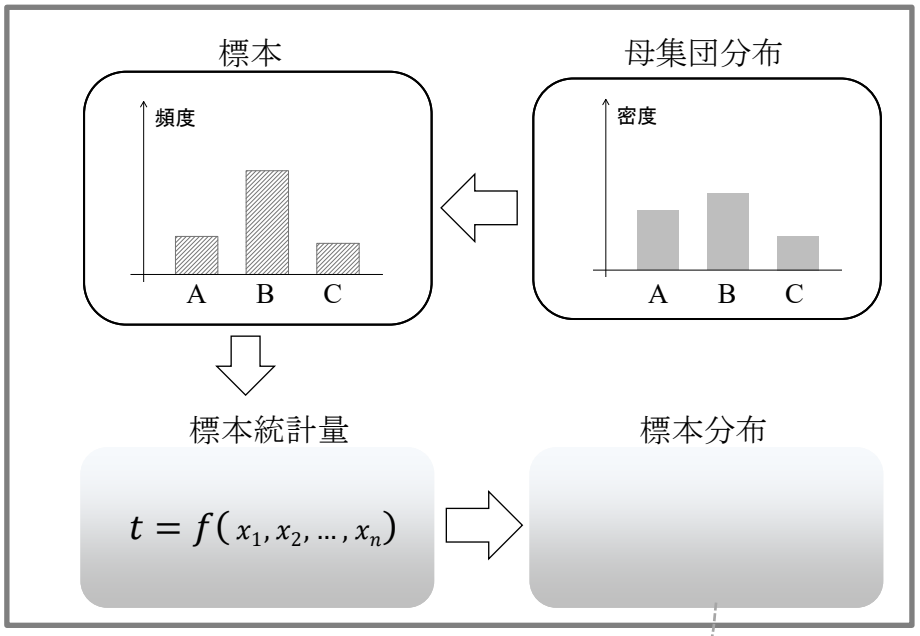
相対頻度表（経験確率分布）をベクトルと見なし、$\vec{p}$と$\vec{q}$のなす角θ_{pq}のコサインを類似度と考える。

$$d_{cos}(\boldsymbol{p},\boldsymbol{q}) = \cos\theta_{pq}$$
$$= \left(\frac{\vec{p}}{|\vec{p}|}\right)\cdot\left(\frac{\vec{q}}{|\vec{q}|}\right)$$
$$= \frac{\vec{p}\cdot\vec{q}}{|\vec{p}||\vec{q}|}$$
$$= \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$$



※　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。

📖 ノート３　標本分布



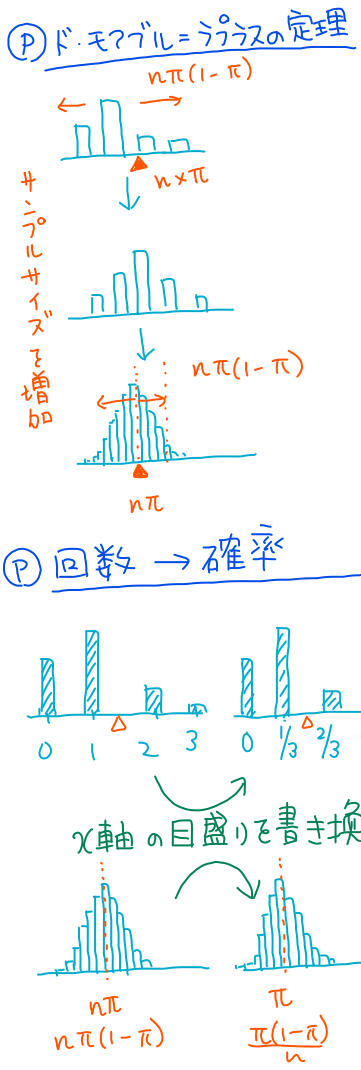
■ このノートで習う標本分布

（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πが特定の値π_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 H0： $\pi = \pi_0$
対立仮説 H1: $\pi \neq \pi_0$

② ド・モアブル＝ラプラス（De Moivre-Laplace）の定理

サイズが n で、確率π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。

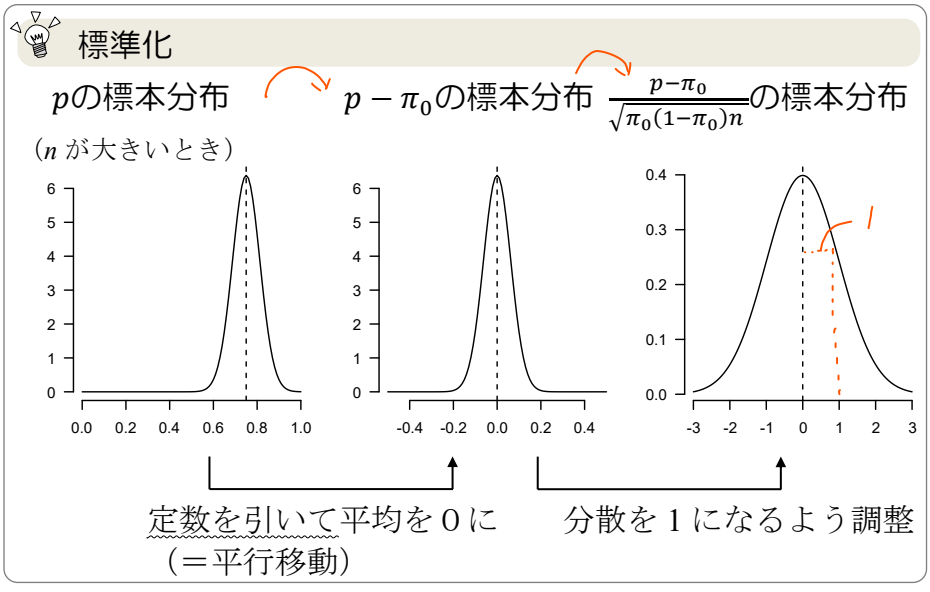
③ 標準化

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 H0 :「二つの母集団の母比率は同じ」 $\pi_1 = \pi_2 (= \pi)$

対立仮説 H1:「　　〃　　」は偽 $\pi_1 \neq \pi_2$

② ド・モアブル＝ラプラス（De Moivre-Laplace）の定理

サイズが n で、確率π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。

Ⓟ適合度検定を用いるとき（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：$O_i \sim Bin\,(n, \pi_i)$
対立仮説：$O_i \sim Bin\,(n, \pi_i)$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π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

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。

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

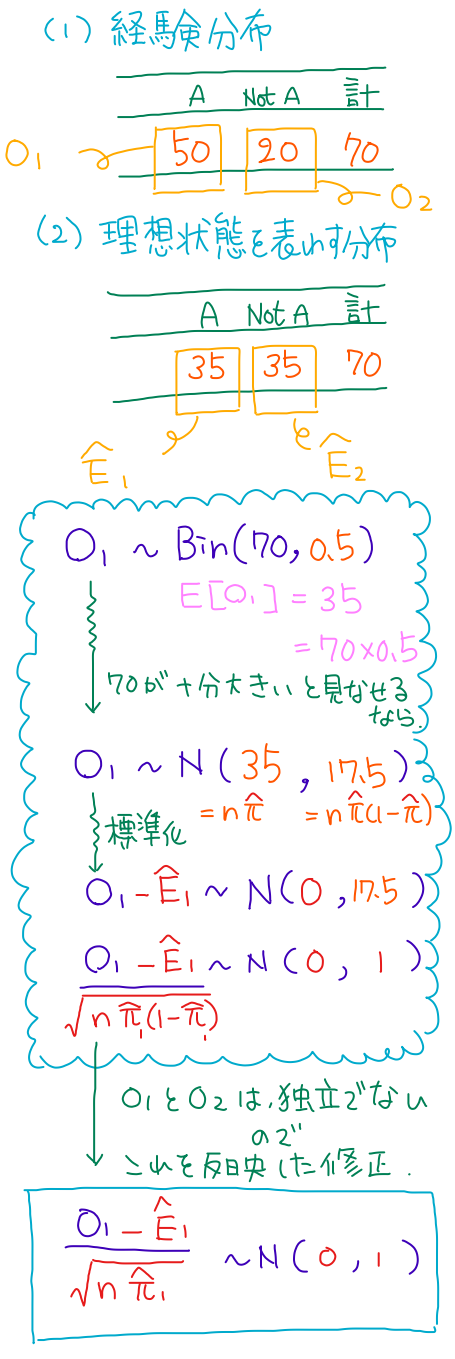
④ 標本分布

標準正規分布に従う確率変数の二乗の和が従う分布がχ^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

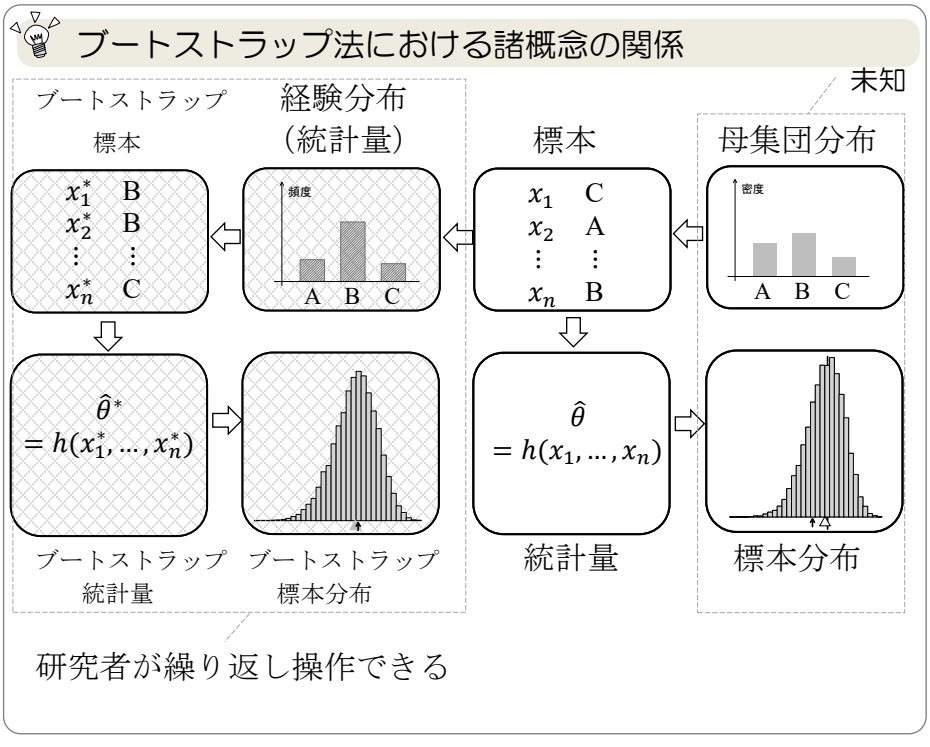
母集団分布の推定値と見なされた**経験分布**から**復元抽出**で得た標本で推定値の信頼性を評価する手法。

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。

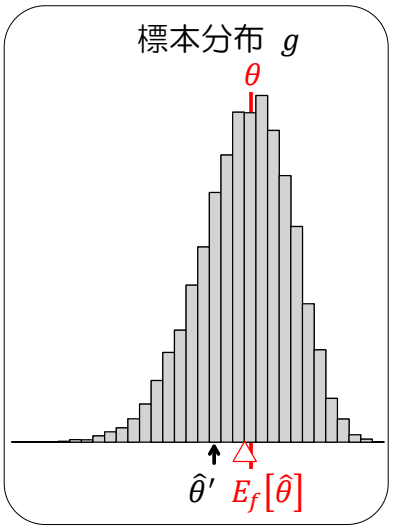


（2）ブートストラップ法によるバイアスと分散（標準誤差）

① 母集団分布 f に基づく量

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

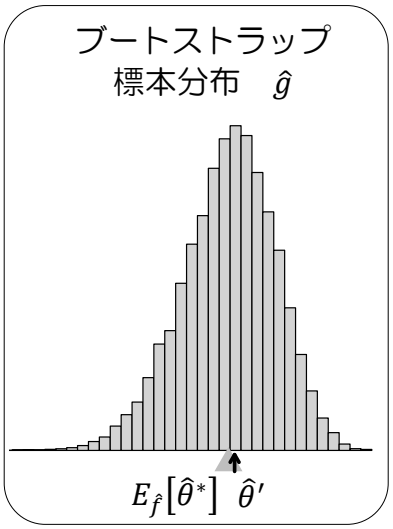
$$\sigma^2_{\hat{\theta}} = E_f\left[\left(\hat{\theta} - E_f[\hat{\theta}]\right)^2\right]$$



② 経験分布（母集団分布の推定量 $\hat{f}$ ）に基づく量

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

$$\sigma^2_{\hat{\theta}^*} = E_{\hat{f}}\left[\left(\hat{\theta}^* - E_{\hat{f}}[\hat{\theta}^*]\right)^2\right]$$



※ 復元抽出 Sampling with Replacement

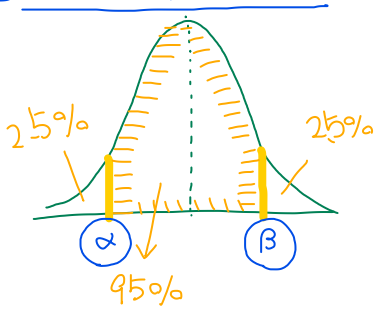
一度抽出したデータも次の抽出の対象になる抽出。

Ⓟ信頼区間と棄却域

（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。

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



ブートストラップ標本分布では，
仮に1000個のデータがあれば，下から25番目と，975番目の点の位置を拾ってくればいい