CEDEC2010 (C10_I0037) 2010-08-31

はじめての確率統計学

— おいしいとこ取りをめざして —

平岡和幸

1. 本格「風」確率論 …… 確率は面積だ！
2. 七つの御利益 …… 面積と思えばいろいろすっきり
3. 確率・統計的な情報処理 …… 簡単な例の試食

0. イントロ

ねらいなど

自己紹介（表）

平岡 和幸

- 和歌山高専で数学を担当
- 専門は数理工学（機械学習・パターン認識）
- 確率・統計の専門家ではないが，道具として常用

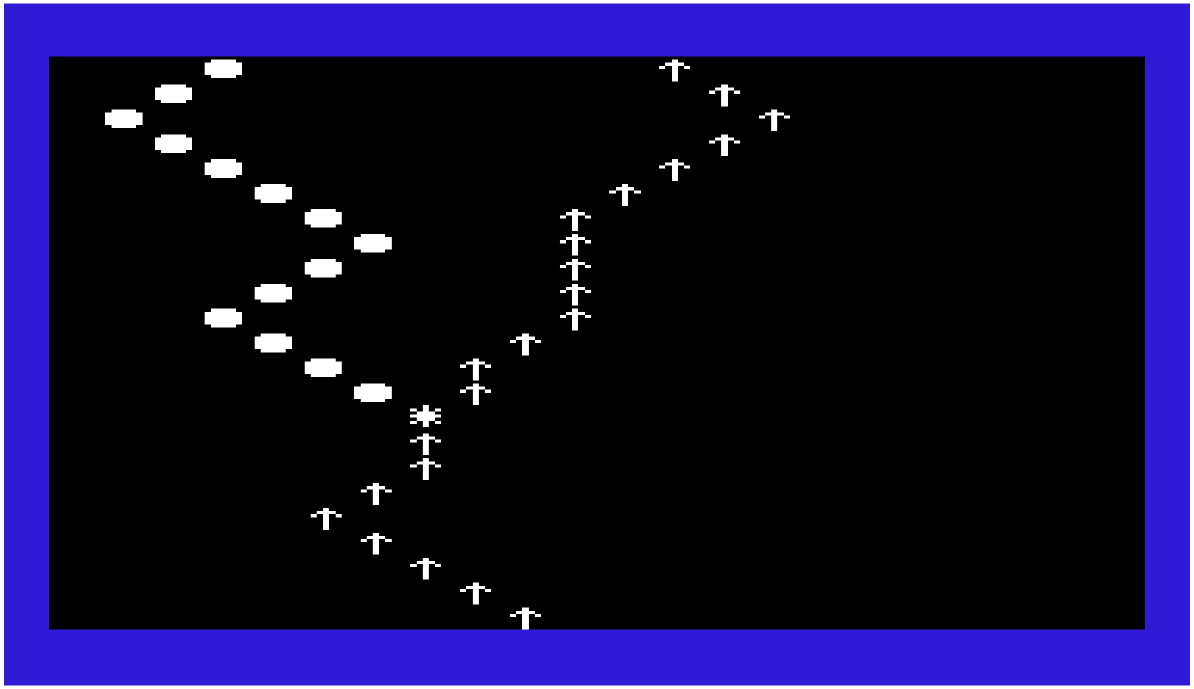
自己紹介（表）

- 『プログラミングのための線形代数』　[LACS]
- 『プログラミングのための確率統計』　[PSCS]

いずれも平岡・堀（オーム社）

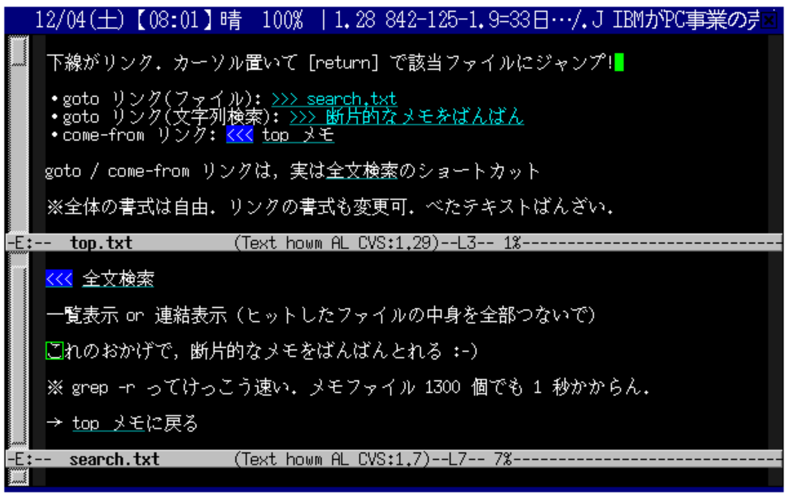
自己紹介（裏）

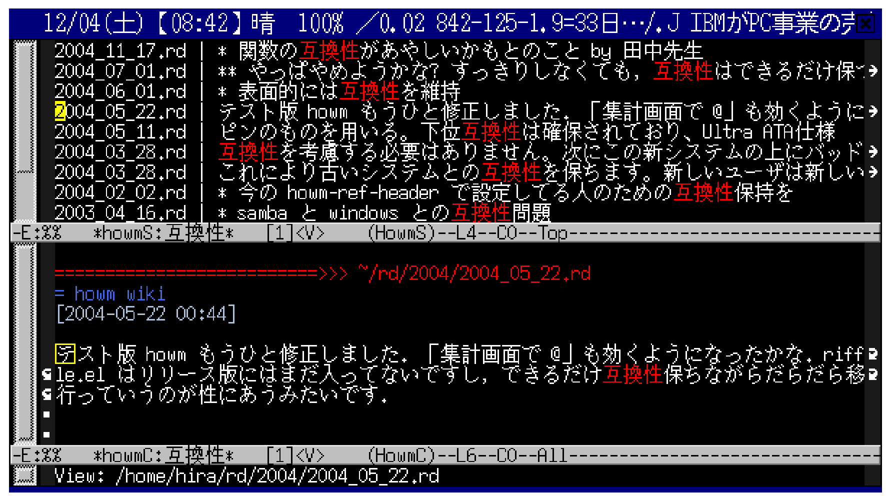
はじめて作ったアクションゲーム



自己紹介（裏）

メモとり環境 howm （一人お手軽Wikiもどき）





→ 2ch UNIX板 howmスレ

本講演のねらい

× 度数分布表，独立性の検定，クラメールの連関係数，……
→ 『マンガでわかる統計学』（オーム社）

○ **確率・統計的な情報処理**に使う数学の勘所

- 現実では「A なら 100% 必ず B」とはなかなかならず，どうしてもゆらぎが生じる
- そのゆらぎを確率・統計的視点で取り扱うことによって，推定なり判定なりをうまくやりたい

本講演のねらい

受講スキル：

パターン認識や機械学習やデータマイニングといった「不確定なゆらぎを取り扱う情報処理手法」に興味があるけれど，数学が原因で手を出しづらく感じている方.

得られるであろう知見：

確率・統計的手法の教科書が読めるようになる.

たとえば『パターン認識と機械学習 ― ベイズ理論による統計的予測』（シュプリンガー）など.

位置づけ（私見）

何の入門書を読むべきか：

不確定なゆらぎ？

- 確率でとらえる
  - → 他人を説得したい
    - → 統計
  - → 自分がうまくやりたい
    - → 統計的○○
- その他
  - → カオス, ファジー など

途中でも質問してください

1. 本格「風」確率論

確率は面積だ！

趣旨

数学者ずるい！俺らにもそれ使わせろ！

本格的な確率論を厳密にやるのはしんどい

$$\Rightarrow \begin{cases} \times \ \text{本格はあきらめよう} \\ \bigcirc \ \text{厳密じゃなくやろう} \end{cases}$$

「本格風」だけでも十分おいしい.
これを理学部に独占させるものか.

趣旨

※「ずるい」はもちろん冗談です.
いじわるしているわけではないはず.

確率とは？

- サイコロで3が出る確率
- こんどの台風が上陸する確率
- この絵画が贋作である確率

……哲学をやっていると進まないので，パス

確率は面積だ！

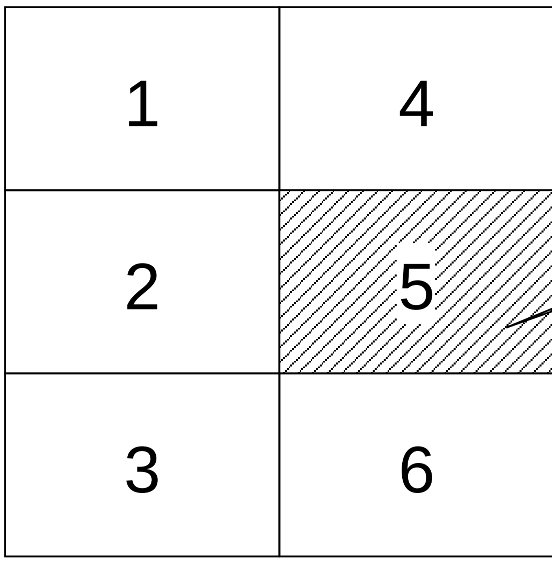
岩波数学辞典いわく……

- 確率変数，独立など<u>すべての確率論の概念</u>は，$(\Omega, \mathscr{F}, P)$の上で，<u>測度</u>論の言葉で定義される．(p.158)
- 確率測度Φとは，ある可測空間$(S, \mathscr{S})$上の<u>測度</u>であって，$\underline{\Phi(S) = 1}$を満たすもののことである．(p.146)

$\Rightarrow$ **全体が1** な **面積・体積など** が「確率」だ！

確率の図解から神様視点へ

サイコロの目



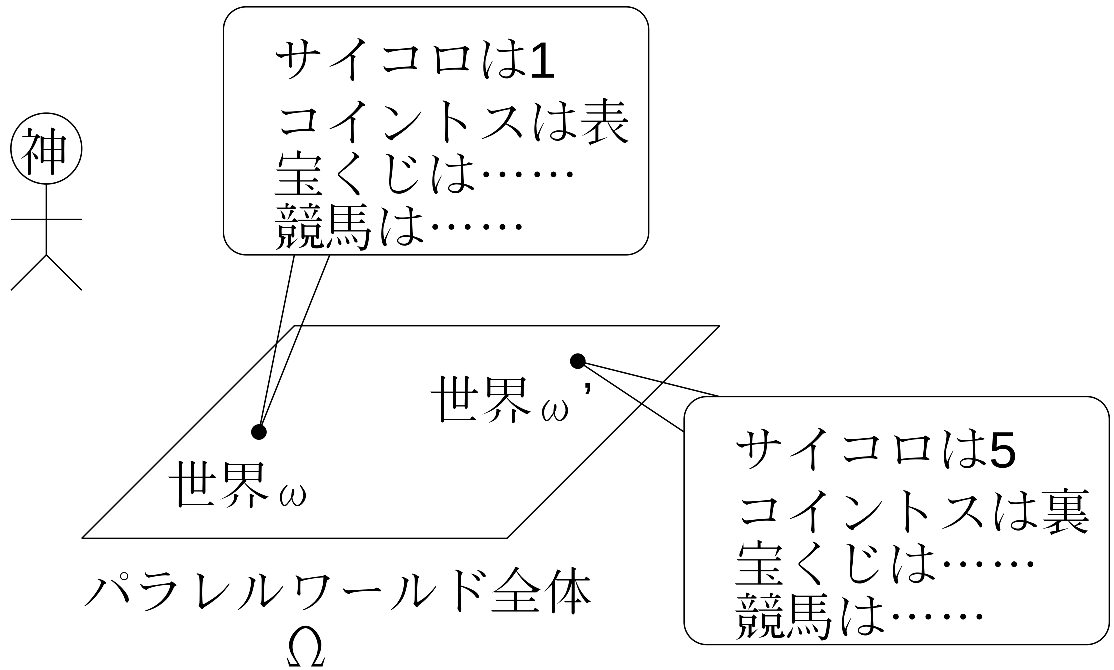
面積1/6 → 確率1/6

P(5が出る) = 1/6

全体の面積 ＝ 1

これをあえて パラレルワールド と解釈（ 神様視点 ）

神様視点

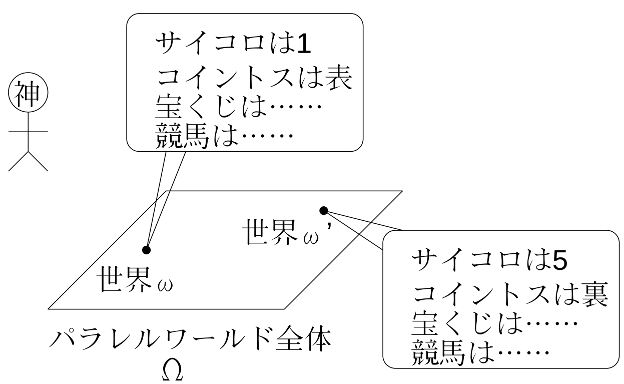


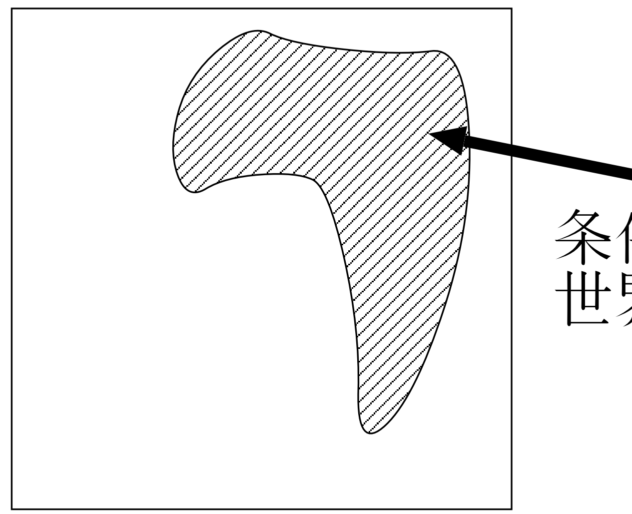
[PSCS 下書きより]

各世界ではすべてが 確定的 ！

神様視点

P(○○) ＝ ○○の確率 ＝ ○○な世界たちの面積





条件○○を満たす
世界$_{\omega}$たち

全体の面積=1

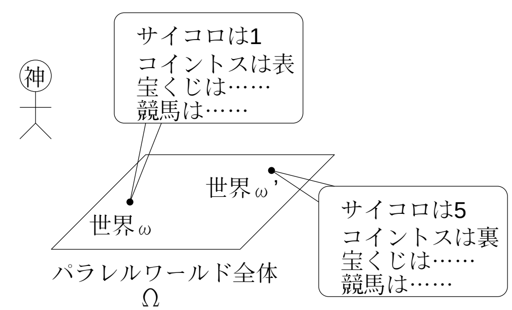
[PSCS 下書きより]

念押し

- パラレルワールドを本気で信じろとは言わない
- 確率の問題を解く際は，こう考えると便利
- 現実とどう対応するかは，「哲学」なのでパス
  （各自で折り合いをつけてください）

神様視点のうれしさ

- 不確定な話と確定した話とを分離し，
  ゆらぎのない「静止画」に
- あとはただの面積計算に帰着



$\Rightarrow$
- 哲学論争を回避
- 込み入った概念も視覚的にイメージ

2. 七つの御利益

面積と思えばいろいろすっきり

趣旨

本格「風」確率論 ＝ **神様視点** ＝ **確率は面積だ！**

確率論の基礎をひと巡りしながら，その御利益を見ていこう

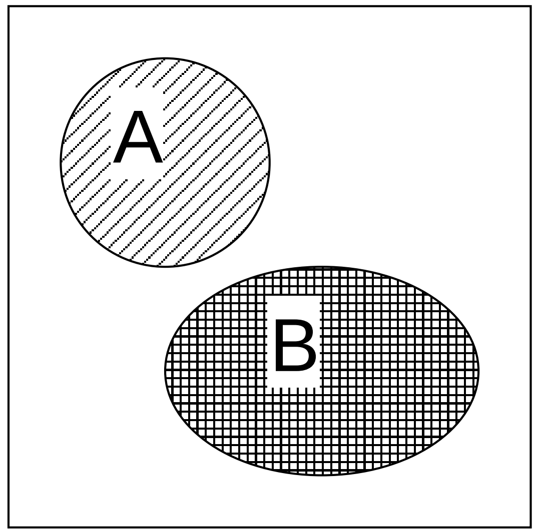
1. 確率の諸性質
2. 確率変数と確率分布の区別
3. ベイズの絵書き歌
4. 独立性のイメージ
5. 期待値の性質
6. 期待値と平均値の区別
7. 大数の法則の真意

(1) 確率の諸性質

面積の性質としてあたりまえ.

（例）AとBが同時には起きないなら,

P(AまたはB) = P(A) + P(B)

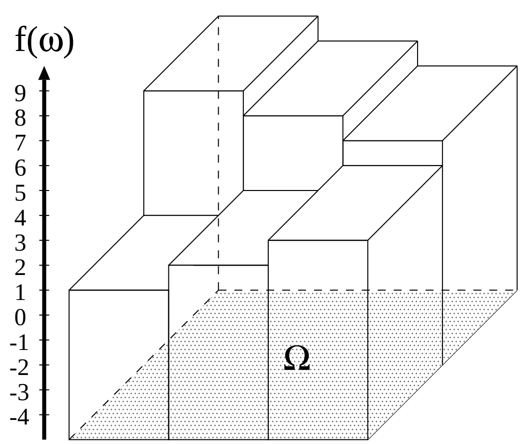


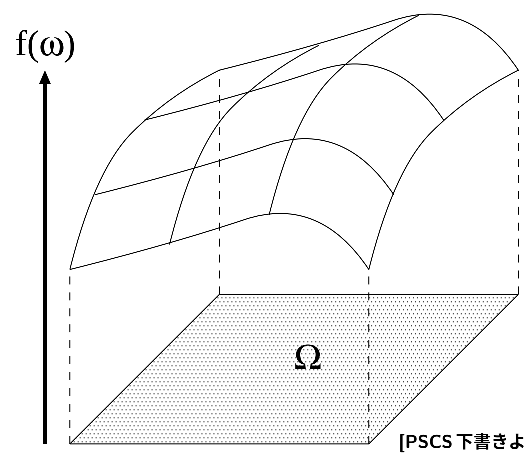
全体の面積は1

(2) 確率変数と確率分布の区別

確率変数：「確率的にゆらぐ値」のこと

→ 神様視点では **ただの関数**





[PSCS 下書きより]

(2) 確率変数と確率分布の区別

確率分布：

「どんな値が出る確率がいくらか」の一覧

値	その値が出る確率
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

(2) 確率変数と確率分布の区別

サイコロを二回ふる．一回目 X．二回目 Y．

- **確率分布** はどちらも同じ.
  （$1 \sim 6$がどれも確率$1/6$で出る）
- **確率変数** X, Y は同じではない.
  （一回目は5が出ても二回目は3が出たり）

(3) ベイズの絵書き歌

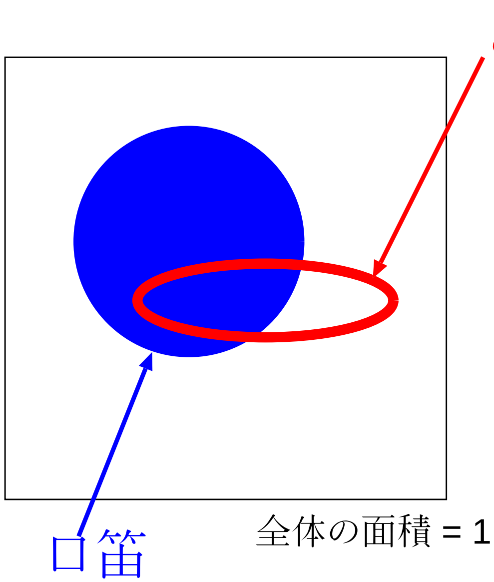
いろいろな応用で使われる **ベイズの公式** も，
神様視点なら絵書き歌で楽々．

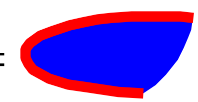
※（私見）
ベイズ法の哲学や位置づけについて，ネットでは一面的なコメントが目につく．でも初心者はそこは軽く流しておけばよいのでは.

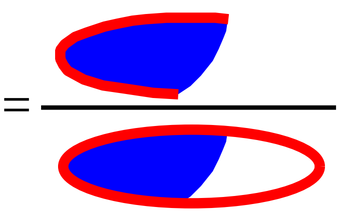
- 「哲学」にこだわりすぎると進まない
- ベイズ派も一枚岩ではなく，いろいろな立場がある

(3) ベイズの絵書き歌

準備：同時確率と条件つき確率



同時確率 P(ご機嫌, 口笛) = 

条件つき確率 P(口笛 | ご機嫌) = 

※ P(ご機嫌, 口笛) = P(口笛 | ご機嫌) P(ご機嫌)

(3) ベイズの絵書き歌

条件つき確率こそ応用の鍵

- P(不正 | カード使用状況)
- P(顔 | 画素値)
- P(中キック | バックステップ)

条件つき確率を逆算するのが**ベイズの公式**

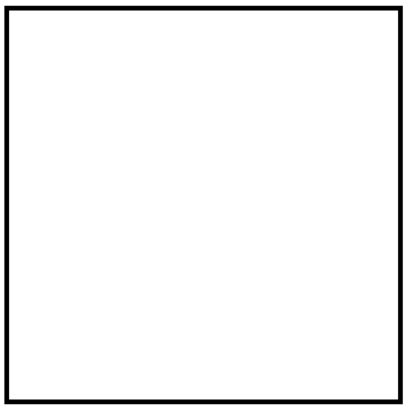
(3) ベイズの絵書き歌

(例)

$$\begin{cases} \mathrm{P}(\textbf{ご機嫌}) = 0.4 \\ \mathrm{P}(\textbf{口笛}|\textbf{ご機嫌}) = 0.7 \\ \mathrm{P}(\textbf{口笛}|\textbf{ご機嫌でない}) = 0.5 \end{cases}$$

$$\mathrm{P}(\textbf{ご機嫌}|\textbf{口笛}) = ?$$

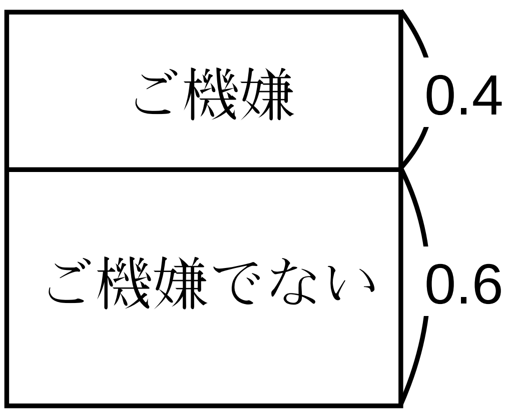
(3) ベイズの絵書き歌



全体の面積 = 1

P(ご機嫌) = 0.4
P(口笛 | ご機嫌) = 0.7
P(口笛 | ご機嫌でない) = 0.5

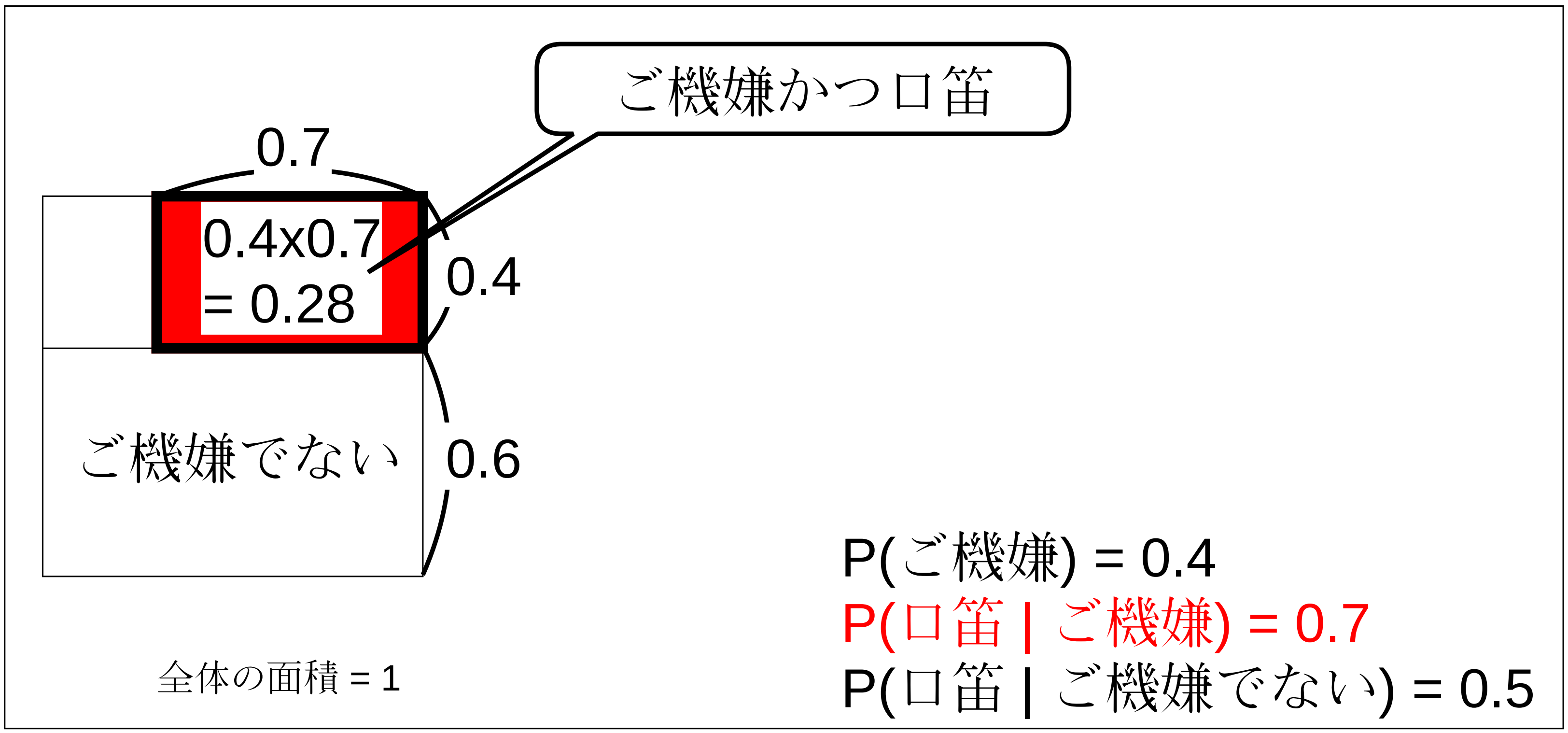
(3) ベイズの絵書き歌



全体の面積 = 1

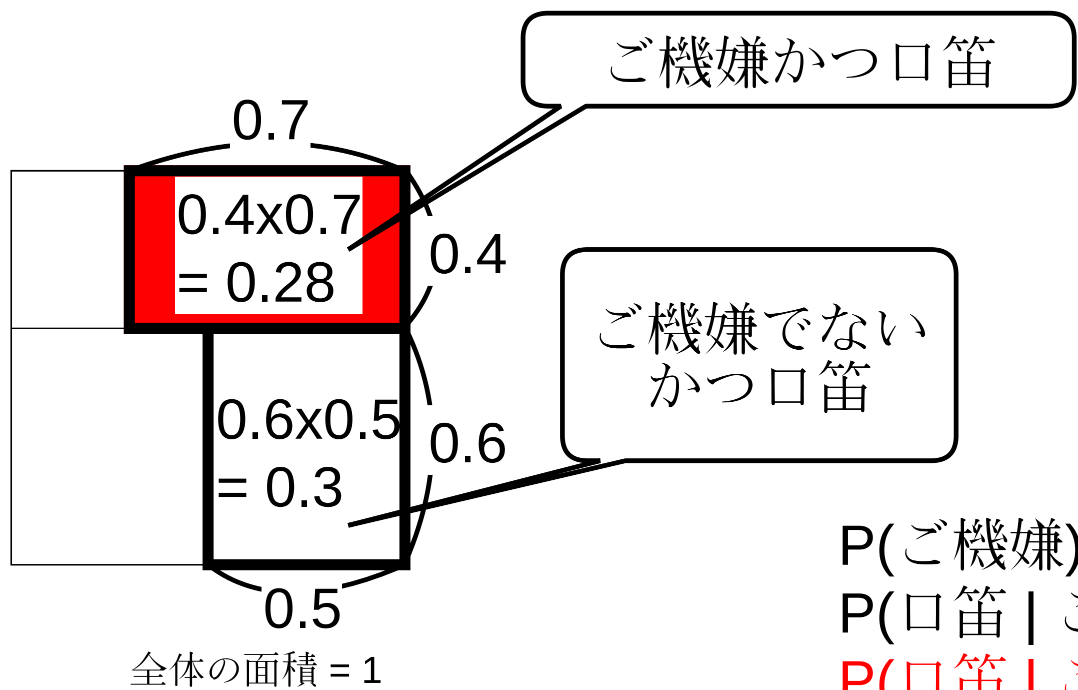
P(ご機嫌) = 0.4
P(口笛 | ご機嫌) = 0.7
P(口笛 | ご機嫌でない) = 0.5

(3) ベイズの絵書き歌



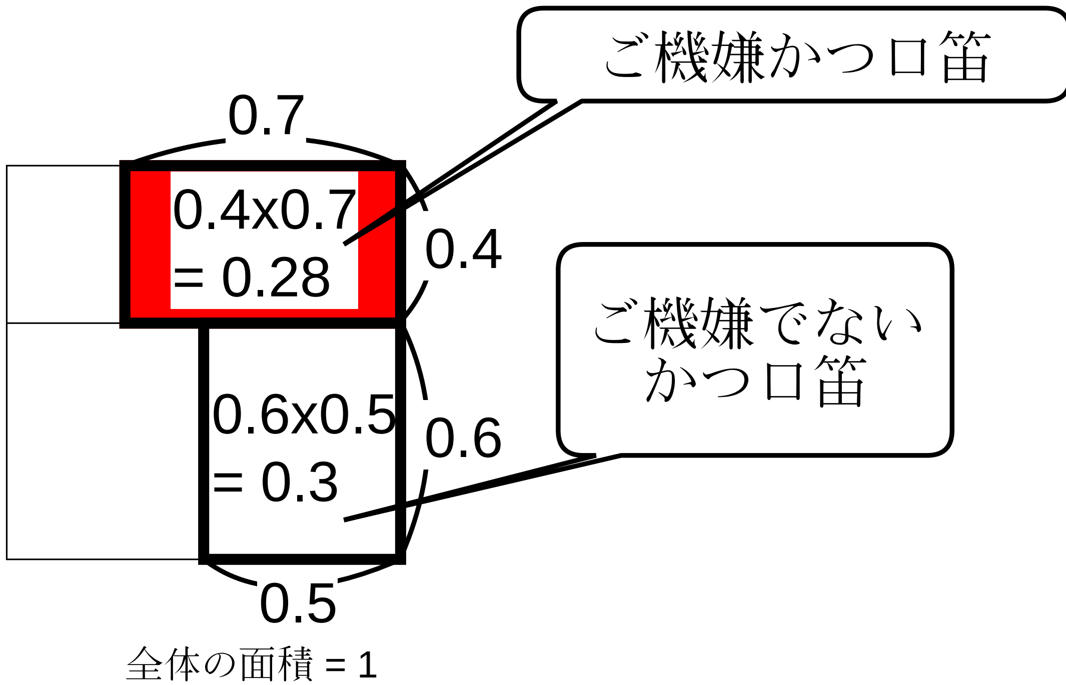
ご機嫌かつ口笛
0.7
0.4x0.7
= 0.28
0.4
ご機嫌でない
0.6
全体の面積 = 1
P(ご機嫌) = 0.4
P(口笛 | ご機嫌) = 0.7
P(口笛 | ご機嫌でない) = 0.5

(3) ベイズの絵書き歌



P(ご機嫌) = 0.4
P(口笛 | ご機嫌) = 0.7
P(口笛 | ご機嫌でない) = 0.5

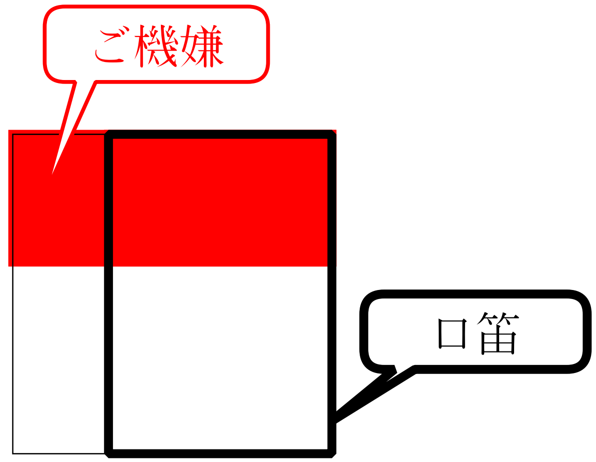
(3) ベイズの絵書き歌



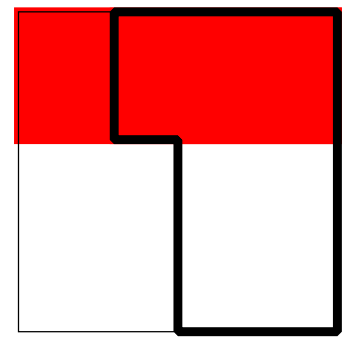
P(口笛)
= 0.28 + 0.3 = 0.58

P(ご機嫌 | 口笛)
= 0.28 / 0.58

(4) 独立性のイメージ



独立 …… 口笛とご機嫌が無関係

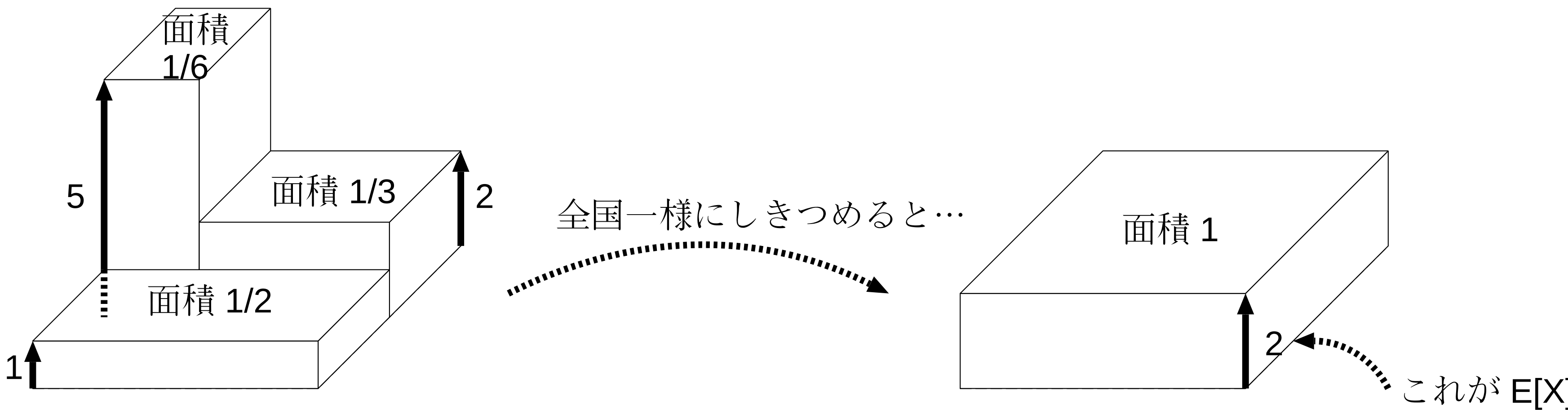


独立でない

$$
\begin{aligned}
\mathrm{P}(\textbf{口笛}\,|\,\textbf{ご機嫌}) &= \mathrm{P}(\textbf{口笛}\,|\,\textbf{ご機嫌でない}) \\
\mathrm{P}(\textbf{口笛}\,|\,\textbf{ご機嫌}) &= \mathrm{P}(\textbf{口笛}) \\
\mathrm{P}(\textbf{口笛}, \textbf{ご機嫌}) &= \mathrm{P}(\textbf{口笛})\,\mathrm{P}(\textbf{ご機嫌})
\end{aligned}
$$

(5) 期待値の性質

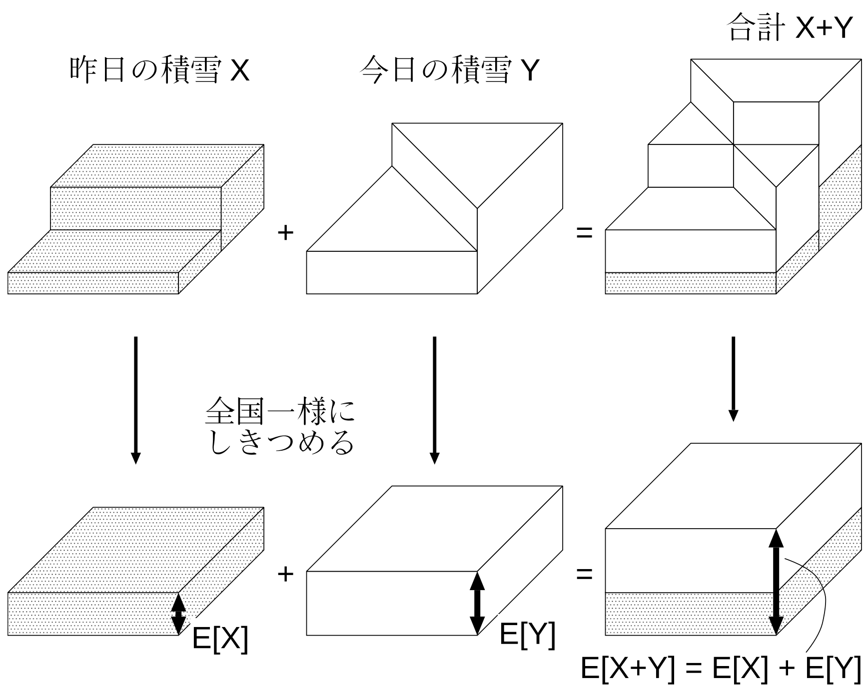
- 応用では **期待値** の計算が頻出
  ∵運しだいでゆらぐ「うれしさ」の期待値を最大化
- X の期待値 $\mathrm{E}[X]$ とは，単に X の **グラフの体積**



[PSCS 下書きより]

(5) 期待値の性質

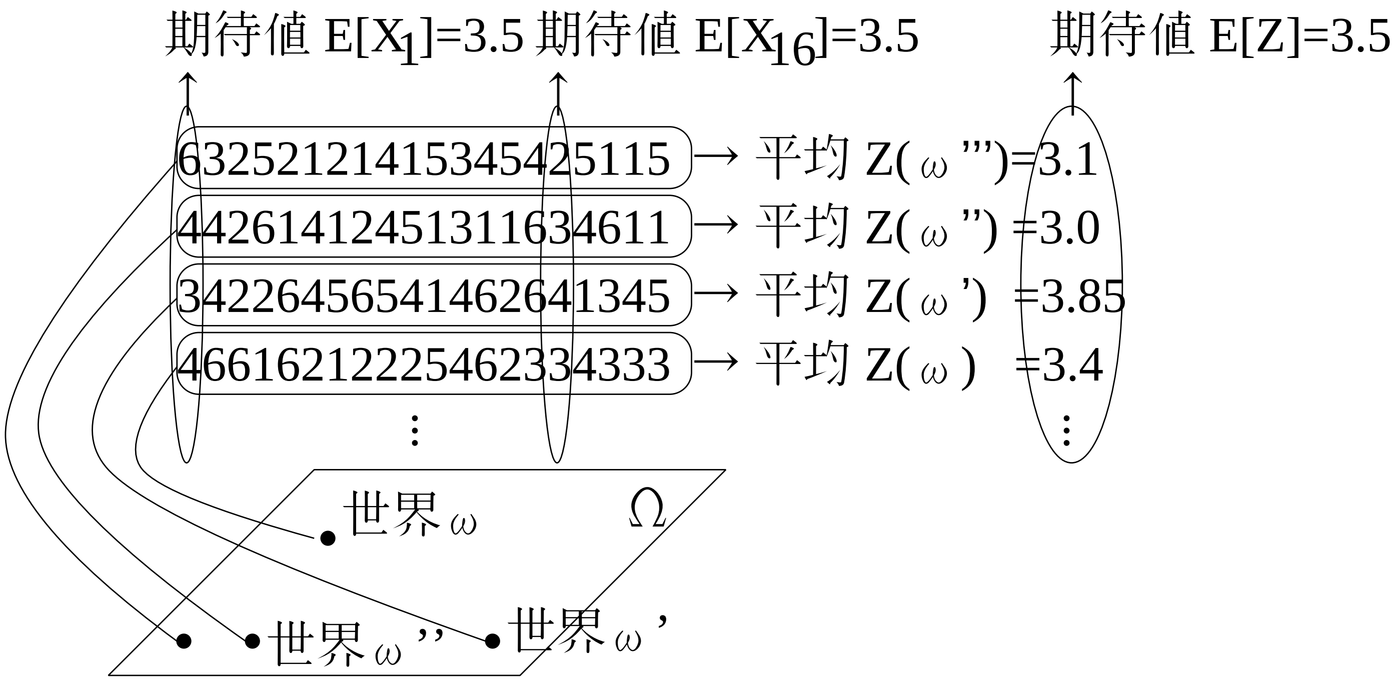
期待値の諸性質は，体積と思えばあたりまえ



[PSCS 下書きより]

(6) 期待値と平均値の区別

サイコロを何度もふる. n 回目：X_n



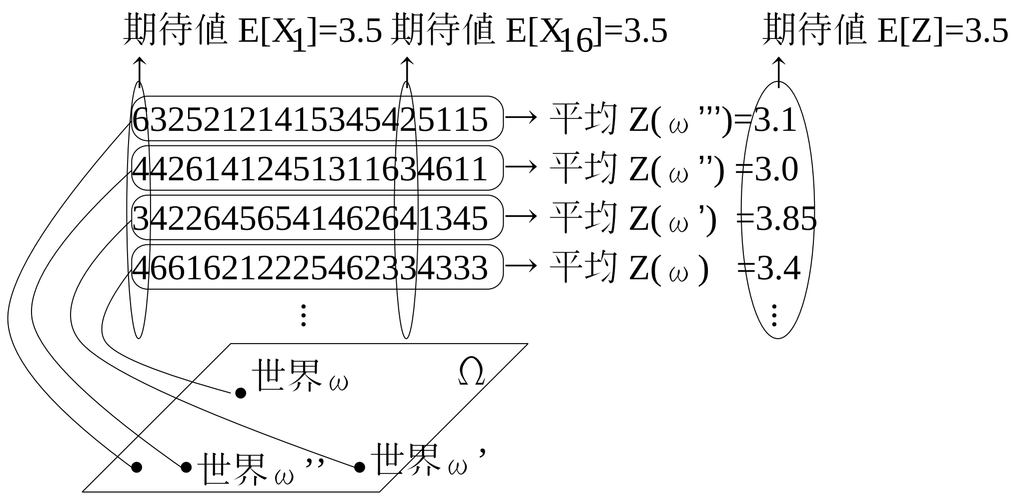
[PSCS 下書きより]

(7) 大数の法則の真意

サイコロのn回目までの平均：Z_n

大数の強法則

「$n \to \infty$で$Z_n \to 3.5$となる確率」$= 1$



[PSCS 下書きより]

3. 確率・統計的な情報処理

簡単な例の試食

趣旨

- 確率・統計的な情報処理のバリエーションは様々.
- 簡単な例を題材に挙げつつ，その例に限らない
  - 独特の考え方
  - ひっかかりどころ

を述べたい.

問題設定

spam filterを題材に：

$$\textbf{メール内容}\ X = (X_1, X_2, \ldots, X_{10000}), \quad \textbf{種類}\ Y$$

$$X_w = \begin{cases} \bigcirc & \textbf{単語}\,w\,\textbf{を含む} \\ \times & \textbf{単語}\,w\,\textbf{を含まない} \end{cases}, \quad Y = \begin{cases} \textbf{spam} \\ \textbf{ham} \end{cases}$$

XからYを当てたい.

w	単語
1	愛
2	合印
⋮	⋮

問題設定

用意するもの：訓練データ

……(X, Y)の実現値をたくさん

やりたいこと：

見たことのないXに対して，Yを当てる

メール内容 $X = (X_1, X_2, \ldots, X_{10000})$,　種類 Y

$$X_w = \begin{cases} \bigcirc & \text{単語}\,w\,\text{を含む} \\ \times & \text{単語}\,w\,\text{を含まない} \end{cases}, \quad Y = \begin{cases} \text{spam} \\ \text{ham} \end{cases}$$

Step.0 前提：確率でとらえる

「データはある確率分布に従って生成される」
と信じる．（**真の分布**と呼ぶ）

メール内容 $X = (X_1, X_2, \ldots, X_{10000})$,　種類 Y

$$X_w = \begin{cases} \bigcirc & \text{単語}\,w\,\text{を含む} \\ \times & \text{単語}\,w\,\text{を含まない} \end{cases}, \quad Y = \begin{cases} \text{spam} \\ \text{ham} \end{cases}$$

Step.1 準備：モデルをたてる

真の分布がどんな格好か仮定（例：ナイーブベイズ）

$$
\begin{aligned}
&\mathrm{P}(X_1 = \bigcirc, \ldots, X_{10000} = \times, Y = \textbf{spam}) \\
&= s_1(\bigcirc | \textbf{spam}) \cdots s_{10000}(\times | \textbf{spam})\, r(\textbf{spam})
\end{aligned}
$$

- **未知パラメータ**を含む …… 上の s や r
- 近似でもよい
- 複雑すぎ・広すぎはデメリットあり（後述）

Step.2 訓練：モデルのパラメータを推定

訓練データに合うよう, 未知パラメータを推定する.

$$\begin{aligned} &\mathrm{P}(X_1 = \bigcirc, \ldots, X_{10000} = \times, Y = \textbf{spam}) \\ &= \; s_1(\bigcirc|\textbf{spam}) \cdots s_{10000}(\times|\textbf{spam})\, r(\textbf{spam}) \end{aligned}$$

- この例なら, 単に訓練データ中の割合でよい
- ここがしんどすぎるモデルは大規模化に不向き

Step.3 本番：そのパラメータで運用

分布$\mathrm{P}(X = x, Y = y)$の推定がStep.2で得られた

→ 新たなメールに対して

$$\begin{cases} \mathrm{P}(Y = \textbf{spam} \mid X = \text{そのメール}) \\ \mathrm{P}(Y = \textbf{ham} \mid X = \text{そのメール}) \end{cases}$$

の大小で spam/ham を判定

実験時の注意（経験損失と期待損失）

結局は「いろいろ試してうまくいった手法を採用」.
その際に落し穴が……

- 訓練データに対する成績（**経験損失**，学習誤差）
- 本番の成績（**期待損失**，汎化誤差，予測誤差）

両者は別物！

実験時の注意（経験損失と期待損失）

授業と同じ問題で試験するなら丸暗記最強.

でもそれは実力じゃない.

はじめて見る問題での成績が実力.

- 訓練データとは別に テスト用データ を確保
- 各手法を訓練データでチューニング
- テスト用データで成績を測り, 採用手法を決定

※ データがもったいないならCross Validationなども