広島大学

データサイエンス基礎

Fundamental Data Science

第9回　確率

広島大学　数理統計学グループ

講義の予定

第　1回　「ガイダンスと導入」
第　2回　「データの取得とオープンデータ, データサイエンスの倫理」
第　3回　「データの種類とデータの要約」
第　4回　「データの要約」
第　5回　「R によるデータの視覚化」

記述統計

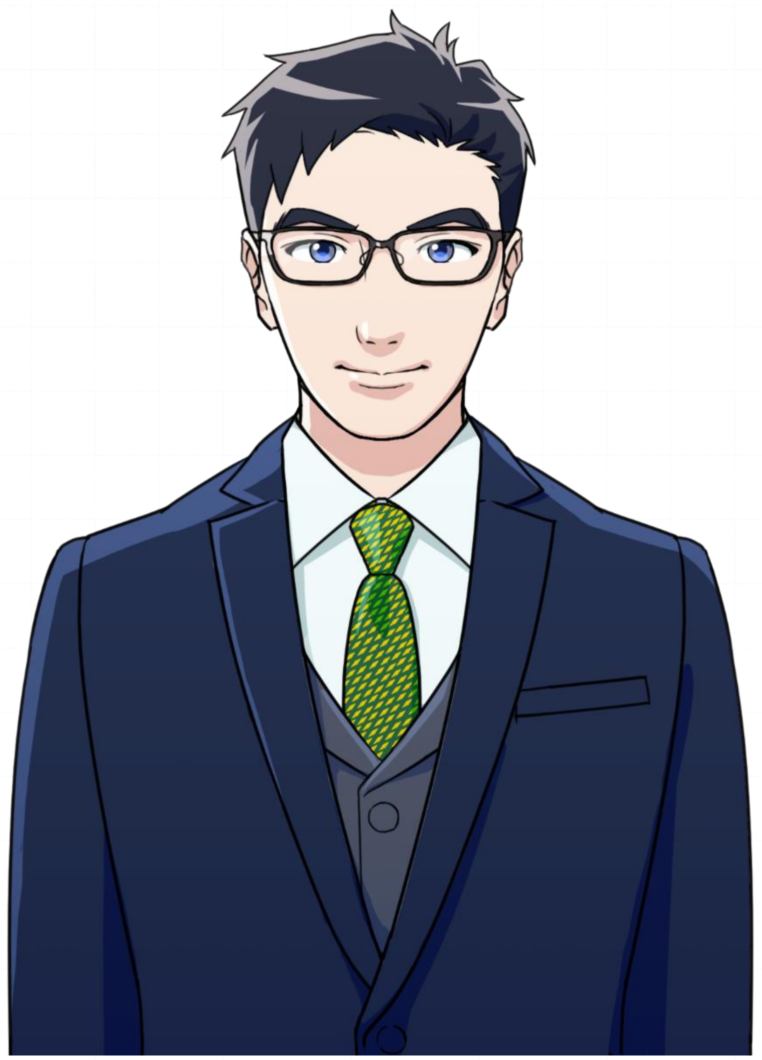
第　6回　「相関と回帰」
第　7回　「Excel による単回帰分析」
第　8回　「R による主成分分析・クラスター分析」

データ分析手法

第　9回　「確率」
第10回　「確率変数と確率分布」
第11回　「主要な確率分布」
第12回　「二変量確率分布」
第13回　「データ収集法」
第14回　「点推定と区間推定」
第15回　「区間推定」

推測統計

これからの講義では, まず
区間推定を理解するための
準備をしていきます



推測統計の例

推測統計

全体の傾向や特徴を収集した**一部のデータ**から**推測**すること

まずは
確率について
知っておかなきゃですね

区間推定

一部の受験者の点数から受験者**全員**の平均点は
確率95%（信頼度95%）で**55点**以上**60点**以下
と推測!!

確率的に!!

推測統計では確率がキーワード!!

データサイエンス基礎 第9回

私たちがナビゲートします！

今日の内容

- ▶ **確率**
- ▶ **条件付き確率**
- ▶ **ベイズの定理**

確率とは

確率

偶然起こる現象の，現象全てに対する割合．
起こりやすさを数値で表したもの．（参照：Wikipedia）

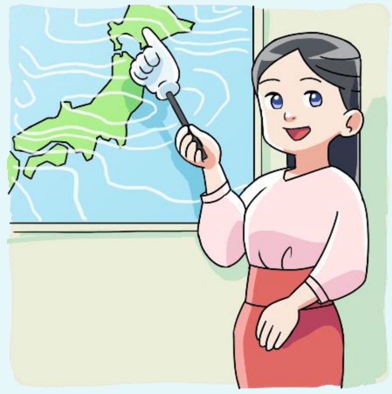
「～の確率は0.5（50%）」のように表現 ※ 確率は0以上1以下の数値

身近な確率

サイコロを1回投げたとき，1の目が出る確率

ある野球選手がヒットを打つ確率

降水確率

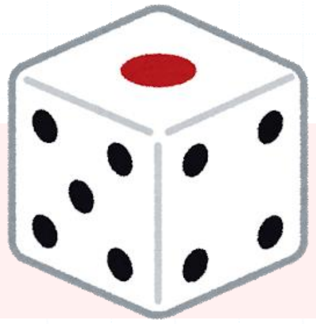


確率は現象の「妥当性」や「信用度」を数値で与える

「意思決定」を行うための道具

確率の計算

Q サイコロを1回投げたとき，1の目が出る確率は?



A $\frac{1}{6}$ 　 $\dfrac{\text{1の目が出る場合の数: }\mathbf{1通り}}{\text{1の目が出る,...,6の目が出る: }\mathbf{6通り}}$

確率の用語で言うと…

$$\frac{\text{事象の起こりうる場合の数}}{\text{起こりうるすべての場合の数}}$$

では，確率の計算方法から
話していきますよ

試行，標本空間

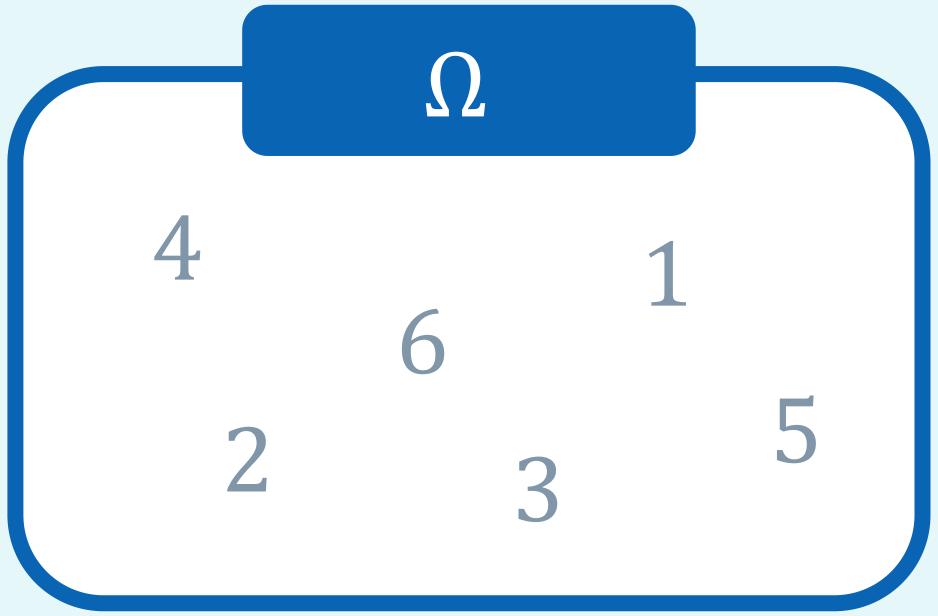
試行
偶然性を伴う実験や調査

例 「サイコロを1回投げる」


標本空間（全事象）Ω
試行の際に起こりうる結果の全体

例 サイコロの出る目全体

イメージ図



数式でかくと

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

※ **数字**だけでも**言葉**でかいても OK

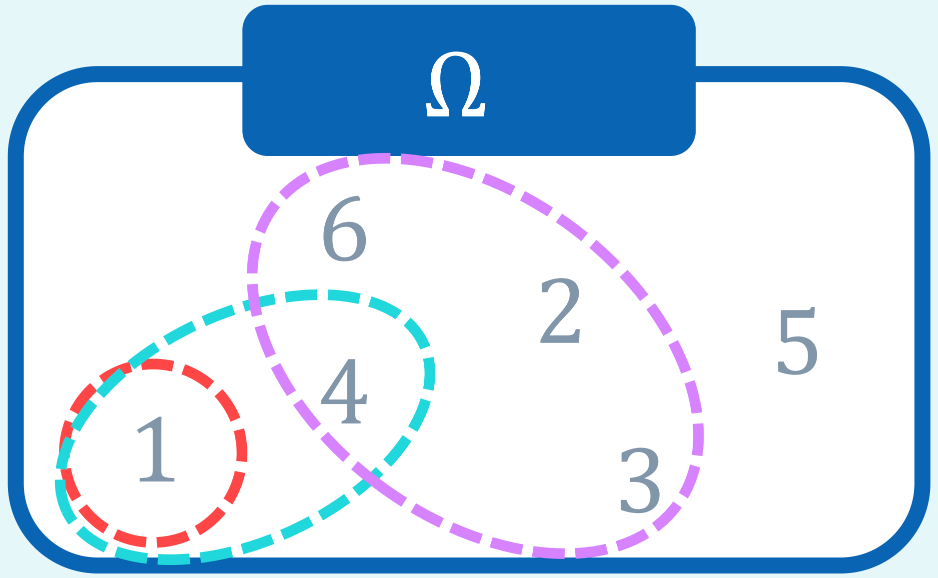
$\Omega = \{$1の目が出る, 2の目が出る, 3の目が出る, 4の目が出る, 5の目が出る, 6の目が出る$\}$

事象，場合の数

事象

試行の結果としてありうる事柄で，Ω の一部分

例



1の目が出る事象：$A = \{1\}$

1または4が出る事象：$B = \{1, 4\}$

1と5以外が出る事象：$C = \{2, 3, 4, 6\}$

$\Omega = \{1,2,3,4,5,6\}$

※ Ω を言葉でかいたら事象も言葉でかく：$A = \{$**1の目が出る**$\}$

場合の数

事象が起こりうるすべての場合を重複なく数え上げた数．

事象 A が起こりうる場合の数を $\#(A)$ とかく

例

$\#(A) = 1$

$\#(B) = 2$

$\#(C) = 4$

$\#(\Omega) = 6$

(古典的) 確率

事象 A が起こりうる確率 (Probability):

$$\color{red}{P(A)} = \frac{\text{事象 } A \text{ の起こりうる場合の数}}{\text{起こりうるすべての場合の数}} = \frac{\#(A)}{\#(\Omega)}$$

古典的確率ともいう

これは
小・中学校などで学んだ
確率ではないでしょうか?

例 **サイコロを1回投げたとき，1の目が出る確率**

標本空間: $\Omega = \{1,2,3,4,5,6\}$

事象: $A = \{1\}$

➡ $P(A) = 1/6$

例 **コインを1回投げたとき，表が出る確率**

標本空間: $\Omega = \{\text{表}, \text{裏}\}$

事象: $A = \{\text{表}\}$

➡ $P(A) = 1/2$

確率の使用例

あるテレビショッピング

このサプリメントはダイエットに抜群に効きます．
実際に**5人中5人**
全員の体重が減少しています．

本当に効くなら買いたい

Aさん

サプリの効果と関係なく，
たまたま全員の体重が
減っただけかも？

サプリの効果関係なしに偶然
5人全員の体重が減少する確率 **< 0.01 (1%)** なら買う
> 0.01 (1%) なら買わない

確率の使用例

5人全員の体重が偶然減少する確率を計算

起こりうるすべての場合の数

1人目		2人目		3人目		4人目		5人目	
2	×	2	×	2	×	2	×	2	＝ 32通り

減るか減らないの2通りを5人分

5人全員の体重が減る場合の数 **＝ 1通り**

確率 ＝ 1/32 ＝ 0.03125 ＝ 約3%

買わない!!

1%より大きい ⇒ 偶然の確率が大きい (効果があるとはいえない)

確率の使用例

あるテレビショッピング

このサプリメントはダイエットに抜群に効きます．
実際に**10人中10人全員の体重が減少**しています．

10人中10人の体重が偶然減少する確率

$= 1/1024$ ＝約0.00097 ＝ **ほぼ0%** → **1%を境目で判断すると買う!!**

確率を計算する前に情報を正しく把握しよう!!

(古典的) 確率の注意点

例 **いびつなサイコロを1回投げたとき, 1の目が出る確率**

いびつなサイコロ:
1,6 が出る本当の割合: 1/6 より**大きい**

標本空間: $\Omega = \{1, 2, 3, 4, 5, 6\}$ 事象: $A = \{1\}$

古典的確率:
いびつなサイコロでも,
$P(A) = 1/6$ となる

計算結果と現実のギャップ!!

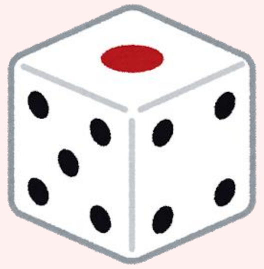
でも実際の割合は
1/6 より大きい

(古典的) 確率ではそれぞれの事象が起こる確率が等しいことが前提!!

条件付き確率

条件付き確率

Q サイコロを1回投げた.
どの目が出たかは確認できていないとする.
このとき, 3以下の目が出た確率は?



A $\frac{1}{2}$ (通常の確率)

情報の追加

Q 投げるのを見ていた人が「**奇数が出ていた**」と教えてくれた.
このとき, 3以下の目が出た確率は?

A $\frac{2}{3}$ (条件付き確率)

情報の追加で確率が変化!!

奇数が出ていたなら1,3,5の
目だけ考えればいいですよね.
3以下は1と3だけです

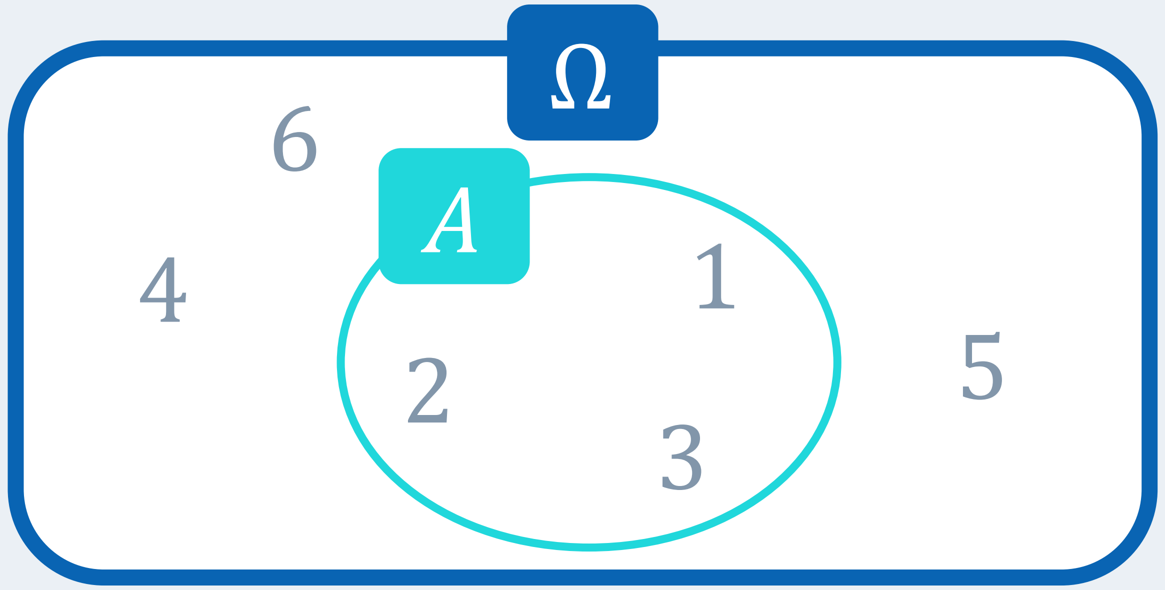
条件付き確率の仕組み

例 サイコロを1回投げる

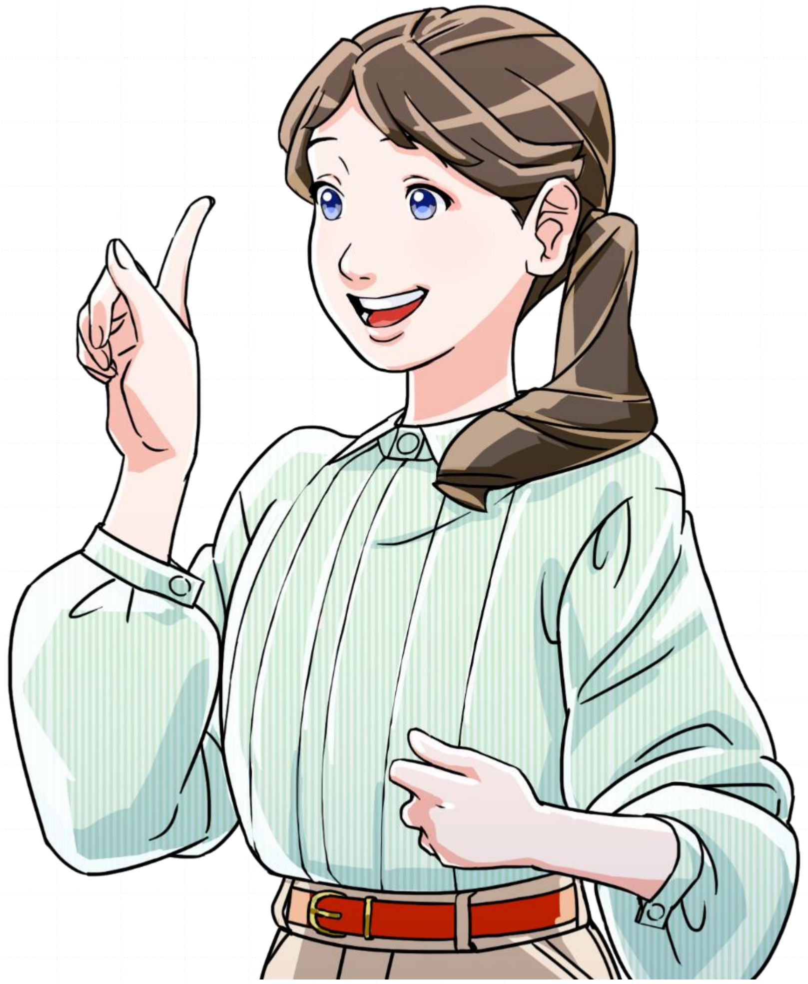
標本空間：$\Omega = \{1,2,3,4,5,6\}$

事象：$A = \{1,2,3\}$：出目が3以下

情報なしの場合



条件付き確率を
図でイメージしてみましょう!

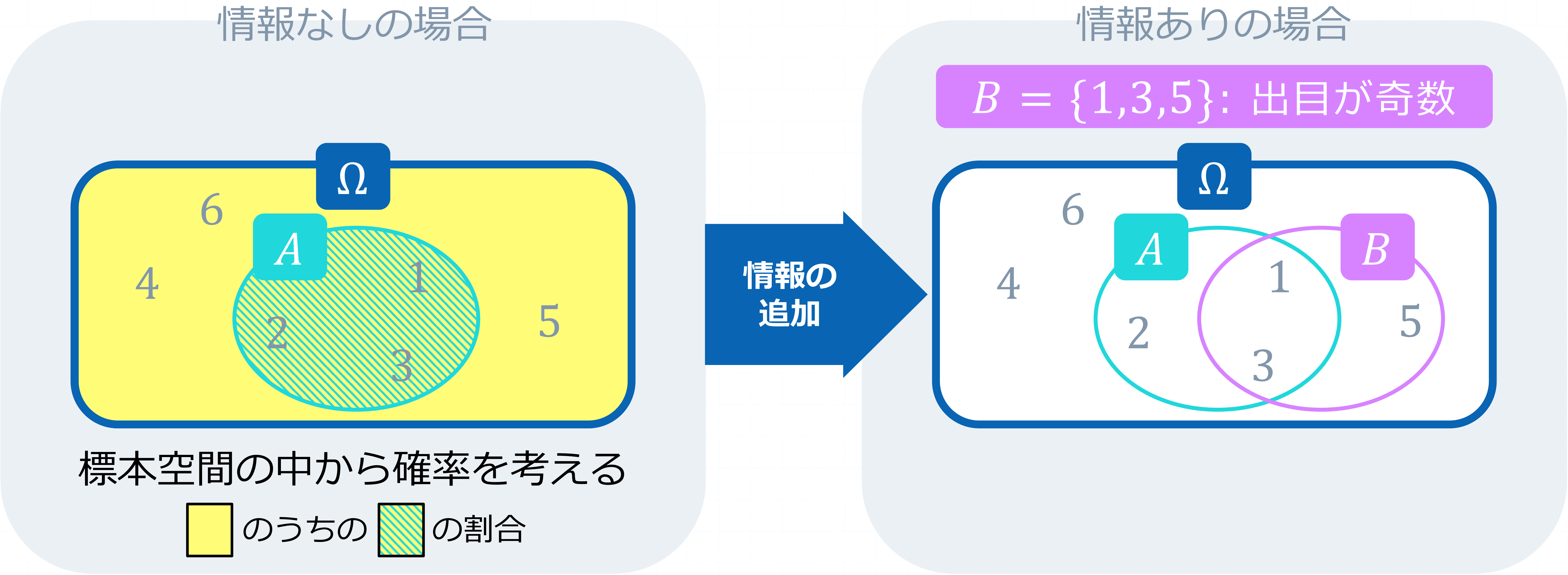


条件付き確率の仕組み

例 **サイコロを1回投げる**

標本空間: $\Omega = \{1,2,3,4,5,6\}$

事象: $A = \{1,2,3\}$: 出目が3以下

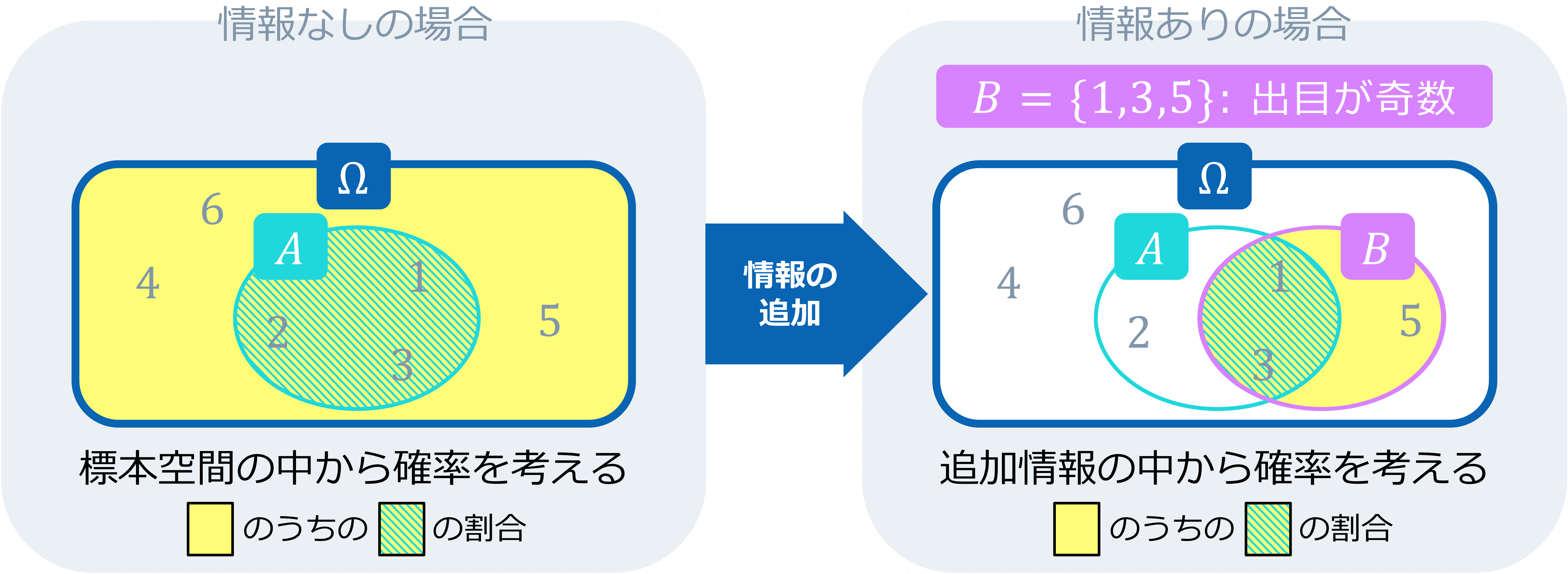


条件付き確率の仕組み

例 **サイコロを1回投げる**

標本空間: $\Omega = \{1,2,3,4,5,6\}$

事象: $A = \{1,2,3\}$: 出目が3以下



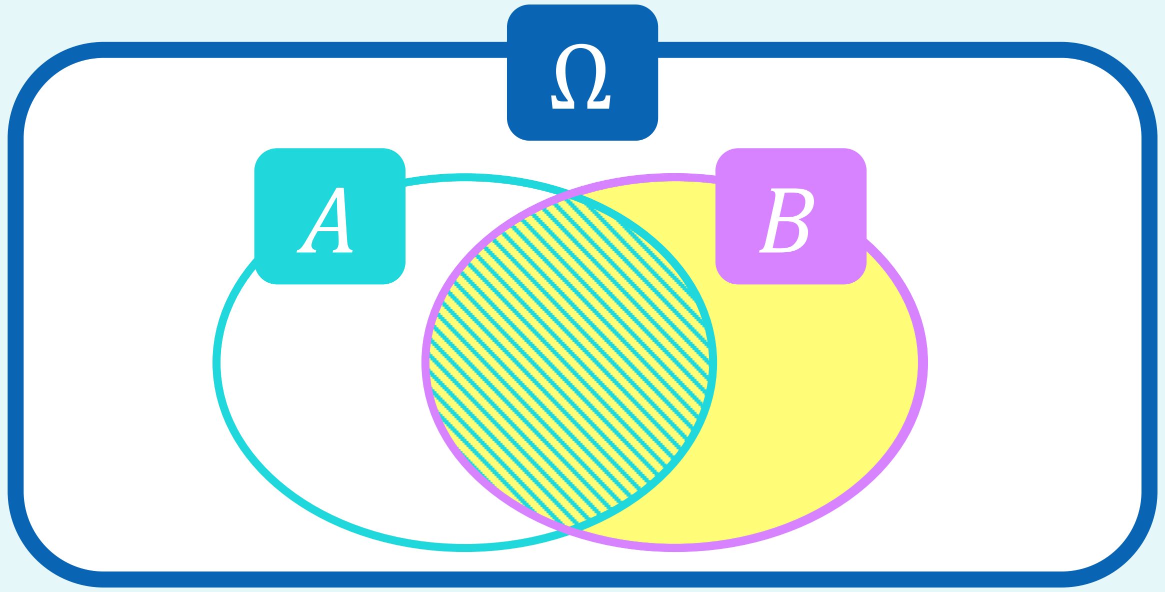
条件付き確率

2つの事象 A,B (ただし, $P(B)>0$) に対して,
事象 B が起こったときに事象 A が起こる**条件付き確率**:

$A \cap B$: A と B が重なる部分

$$\boldsymbol{P(A|B)} = \frac{\#(A\cap B)}{\#(B)}$$

$$= \frac{P(A\cap B)}{P(B)}$$



場合の数で表しても
確率で表しても, 全く同じ
計算結果になります

条件付き確率の例1

知り合い夫婦に子供が2人いるらしい.
このとき, 2人とも男の子である確率は?

情報1 「少なくとも1人は男の子」

情報2 「年上は男の子」

Q 情報なし, 情報1あり, 情報2ありのときの
それぞれの確率は?

条件付き確率の例1

情報なしの場合

事象 A: {2人とも男の子}

$\#(A) = 1$ → $P(A) = \frac{1}{4}$

情報1「少なくとも1人は男の子」の場合

事象 B: {少なくとも1人は男の子}

$\#(B) = 3$　$\#(A \cap B) = 1$

→ $P(A|B) = \frac{\#(A \cap B)}{\#(B)} = \frac{1}{3}$

情報2「年上は男の子」の場合

事象 C: {年上は男の子}

$\#(C) = 2$　$\#(A \cap C) = 1$

→ $P(A|C) = \frac{1}{2}$

全部で4通り

年上 子1	年下 子2
男	男
男	女
女	男
女	女

どんな情報かによって確率が変化

条件付き確率の例2

Q Aさんはある病気にかかっているかどうかの検査を受け，陽性と言われた．

このとき，本当に病気にかかっている確率は?

病気に関する直接的な情報　Aさんの地域では，**0.01%**の人がその病気にかかっている

受けた検査の精度は**99%**

病気の人を病気であると正しく判定する確率 (感度)：99%

病気でない人を病気でないと正しく判定する確率 (特異度)：99%

検査の精度も
確率の計算に取り入れましょう

条件付き確率の例2

陽性と判定されたとき病気にかかっている確率　$P(A|B)$

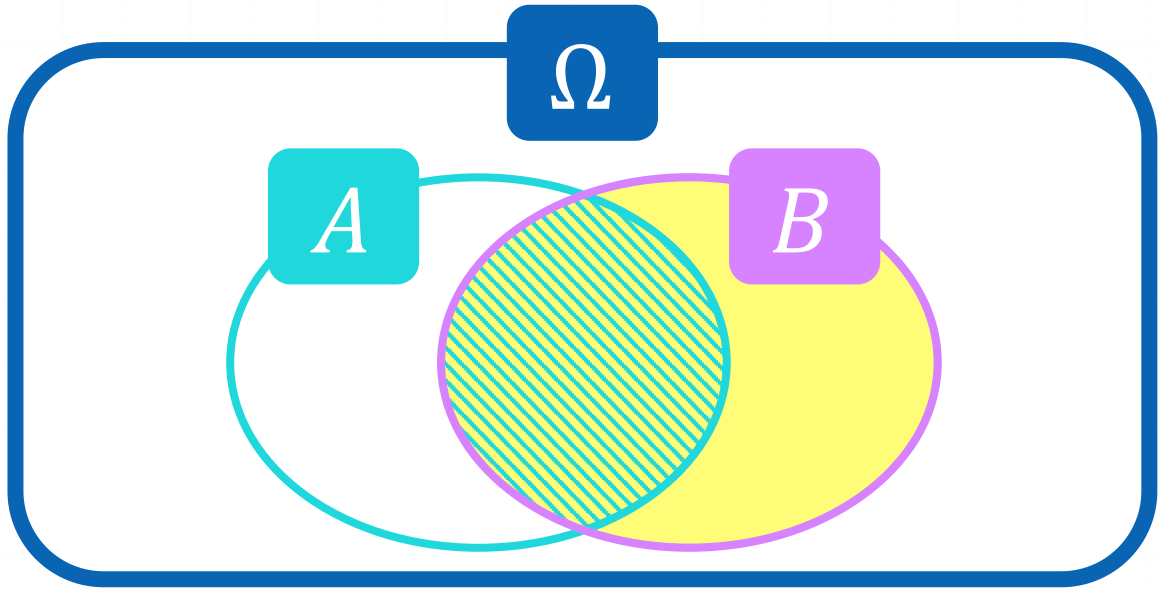
事象 A: {病気にかかっている}

事象 B: {陽性と判定される}

に対応

100万人のうちの割合

	陽性 B	陰性	合計
病気 A	99 (99%)	1 (1%)	100
病気でない	9,999 (1%)	989,901 (99%)	999,900
合計	10,098	989,902	1,000,000



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\#(A \cap B)}{\#(B)} = \frac{99}{10098} = 0.009804 = \text{約}1\%$$

ベイズの定理
（条件付き確率の応用）

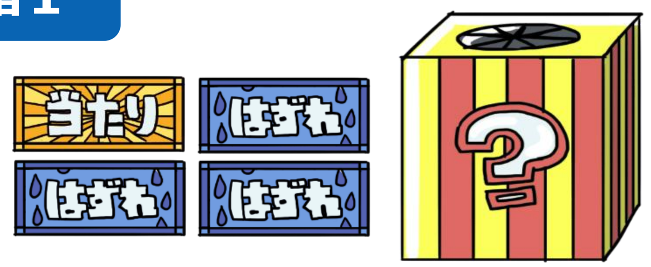
どっちの箱?

Q 外見は同じだけど，当たりの枚数が異なる
2つのくじ引きの箱がある．

お店の人にどちらか1つの箱を目の前に出され，
そこからくじを引くと当たりだった．

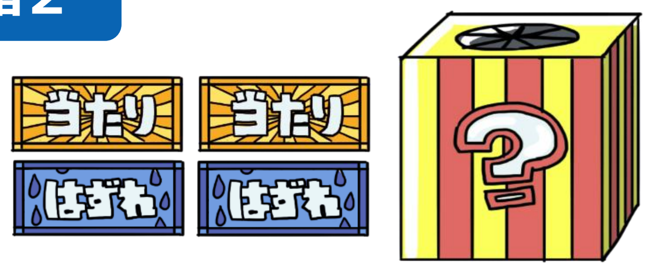
このとき，どっちの箱からくじを引いた確率が高いか?

箱1



4枚中1枚当たり

箱2



4枚中2枚当たり

結果

原因

でも当たりを
引いたんだから
当たりが多い
箱2の気がする…

箱1か箱2かの2択だから
確率は 1/2 ??
(通常の確率)

結果の情報を利用したい!!

ベイズの定理

ベイズの定理

$$P(B|A) = P(B) \times \frac{P(A|B)}{P(A)}$$

原因 B が起きたもとで 結果 A が起きた確率

結果 A が起きたもとで 原因 B が起きた確率

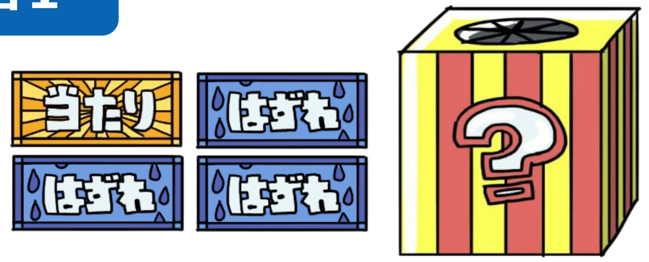
結果 をもとに その原因 の確率 (逆の確率) を計算できる!!

事象 A: 結果 , 事象 B: その原因 と思ってみる!!

例 A: {当たりを引く}, B: {箱1からくじを引く}

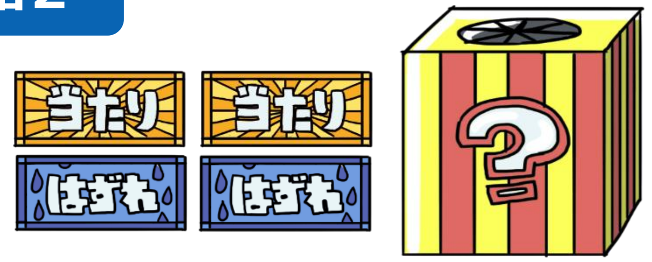
どっちの箱?

箱1



4枚中1枚当たり

箱2



4枚中2枚当たり

結果 引いたら当たりだった

原因 箱1からくじを引いた

当たりだったときに箱1からくじを引いた確率

事象（結果）A：{当たりを引く}

事象（原因）B：{箱1からくじを引く}

$$P(B|A) = P(B) \times \frac{P(A|B)}{P(A)}$$

箱1からくじを引く確率: $P(B) = \dfrac{1}{2}$

情報がないから等確率としておく

箱1から当たりを引く確率: $P(A|B) = \dfrac{1}{4}$

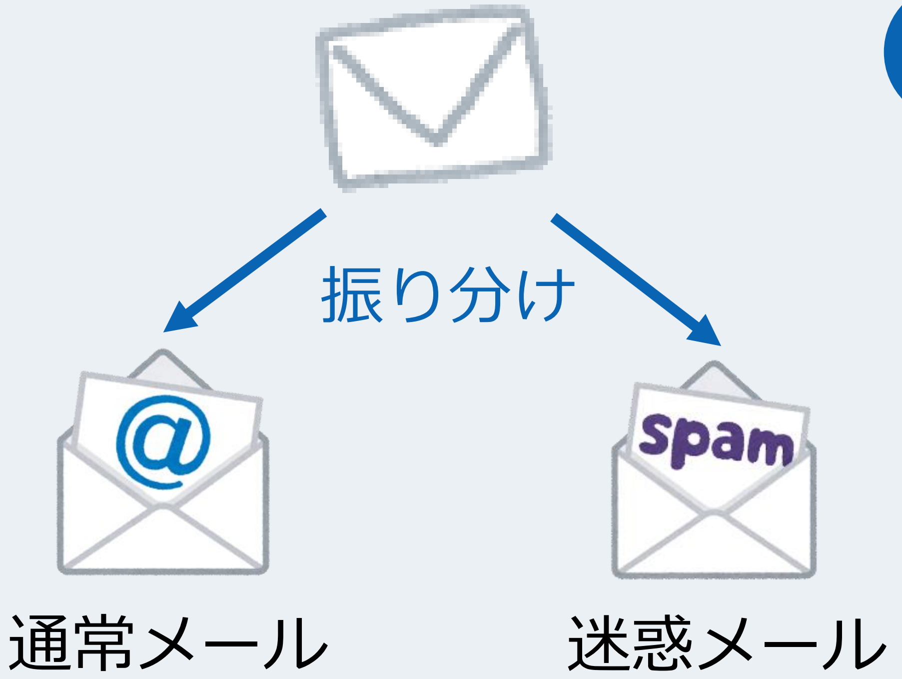
2つの箱から当たりを引く確率: $P(A) = \dfrac{3}{8}$

$$P(B|A) = \frac{1}{3}$$

箱2からくじを引いた確率のほうが大きい

ベイズの定理の実用例

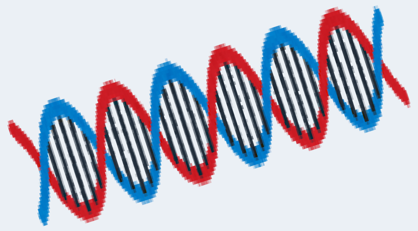
迷惑メールの自動振り分け機能



例

「完全無料」を含む メールが
迷惑メールである 確率

DNA 親子鑑定



例

子の遺伝子が〇〇 のとき，
自分が父親である 確率

ベイズの定理は
実際に身近なところに
使われているんです

今日のまとめ

▶ **確率**

統計学では**確率**がキーワード!!
確率を求めることで**意思決定**につながる

▶ **条件付き確率**

情報の追加で確率が変わる!!

▶ **ベイズの定理**

結果からその**原因**の確率がわかる!!

確率は統計学の中でも
大事なものでしたね