統計基礎

古谷知之

本授業の目的と方法

- 統計学の基礎的な方法論や技法を習得することを目的
- 高校卒業程度の数学の知識とコンピュータの基礎知識を有することが前提（授業では数式を多用します）
- 履修者によって高校数学の習得状況にばらつきがあるので、必要に応じて適宜授業内で知識を補うよう心がける。ただし、授業だけでは補えない場合もあるので、その場合は各自自習すること

成績評価

- 授業第１回目の際に示します

利用するソフトウェア

- この授業では、Rという統計ソフトを使います。

- 他の統計ソフトが使える場合は、そのソフトウェアを使ってもらっても構いません。履修者の状況に応じて、利用するソフトウェアを柔軟に変更する場合があります。

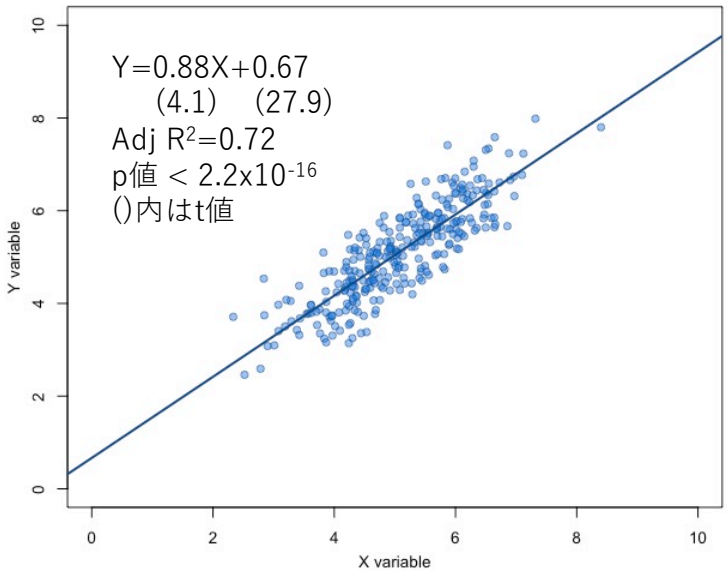
授業概要

＊履修者の状況に応じて変更される場合がありますが、概ね以下のような流れで授業を進めます。

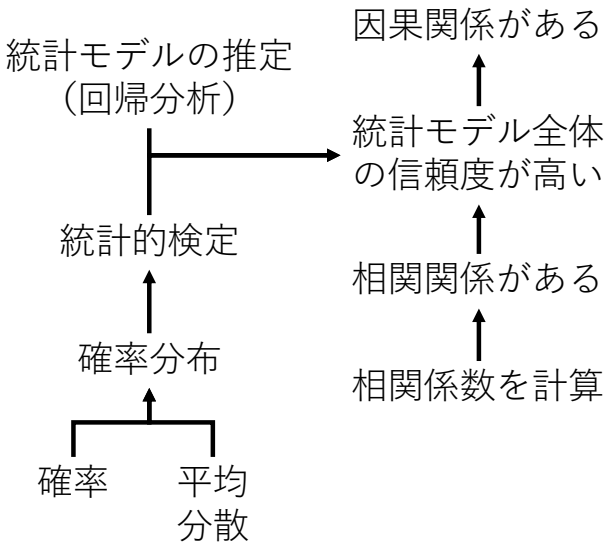
第1回　ガイダンス・確率
第2回　確率変数と確率分布(1)
第3回　確率変数と確率分布(2)
第4回　母集団と標本
第5回　単回帰分析(1)
第6回　単回帰分析(2)
第7回　仮説検定(1)
第8回　仮説検定(2)
第9回　重回帰分析
第10回　R演習(1)
第11回　R演習(2)
第12回　R演習(3)
第13回　R演習(4)
第14回　最終試験

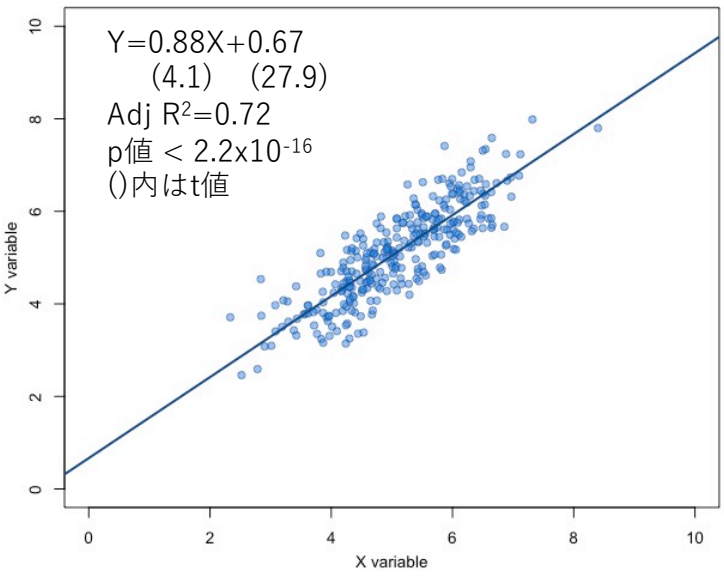
この授業の主な目的

- ある事象Xと別の事象Yとの間に、Xを原因としてYを結果とする因果関係があるということを、定量的かつ実証的に示す方法について理解する
- この方法を理解するのに必要な確率や統計に関する知識や技法を習得する



授業の全体像





確率

- 確率の定義
- 集合と確率
- 条件付き確率・同時確率・周辺確率
- ベイズの定理
- 逆確率
- 尤度
- 事前確率・事後確率

- 前提となる知識
- 場合の数、順列と組み合わせ
- 加法定理、余事象
- ベルヌーイ試行
- 反復試行、乗法定理

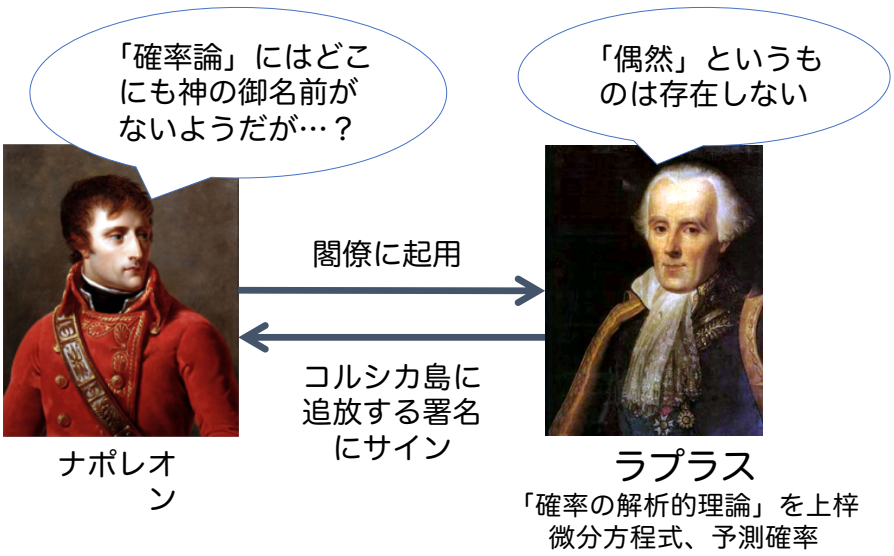
確率の「定義」（ラプラスの定義）

事象Aの起こる確率$P(A)$
＝事象Aが起こる場合の数r /全ての場合の数N

$$P(A) = \frac{r}{N}$$

「全ての場合の数」なんて、定義できるのか？
発生回数が少なくても確率は安定するのか？
偶然的な現象にも適用できるのか？

ラプラスとナポレオン



写真はWikipediaより

標本空間と事象

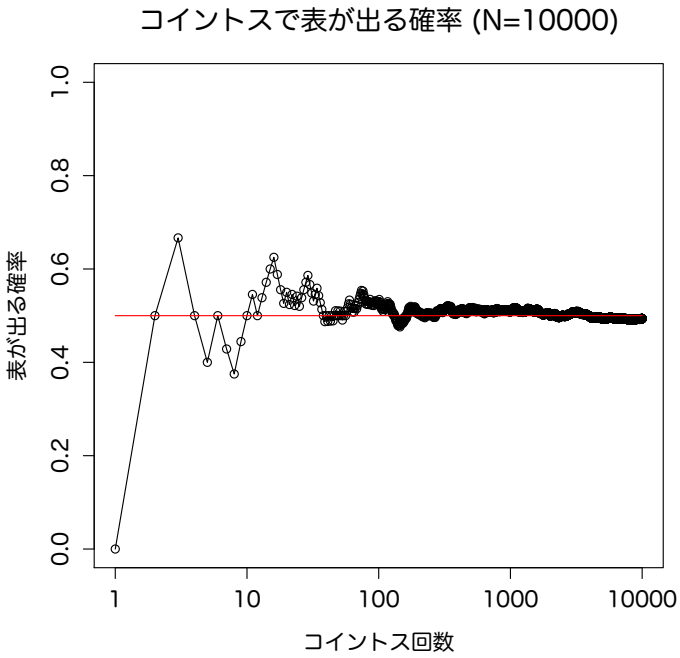
- 起こりうる結果を標本、標本の集合を標本空間という
- 事象とは標本空間の部分集合である。標本がまったくない集合を空集合という
- ただ一つの標本からなる事象を根元事象といい、これ以上分解できない事象を意味する。複数の標本を含み複数の根元事象に分解できる事象を複合事象という。

数学的確率と統計的確率

- 数学の教科書で出てくる確率
  - P(表)=[表の出る回数]/[コイントスの回数]
  - ラプラスの数学的確率と呼ぶことがある
- 多数回の試行で安定した相対頻度
  - 統計的確率と呼ぶことがある
- 結局は同じことを言っている
  - 現象の発生が偶然的であり、基本となる事象が全て等しい確率で生じる場合には、ある事象の確率は場合の数の比率で表される

統計的確率（頻度主義による確率）

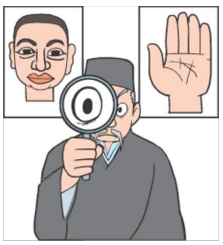
- 事象Aが起こりうる試行回数をn回繰り返すとする
- この時事象Aが起こった回数をnとする
- 試行回数nを無限大までのばすと、事象Aが生じる確率をαと定義できる



経験的確率（主観的確率）

- 過去の経験に則ってある事象が生じる確率を考えるとき、それを「経験的確率」という

- 経験的確率の例
  - 星占い、手相
  - 恋愛
  - 昔の犯罪捜査

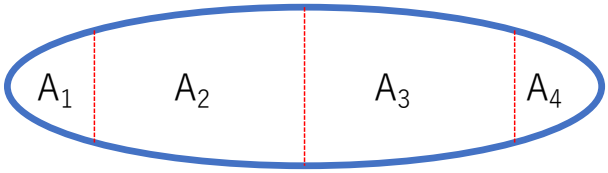


確率（ラプラスの定義）

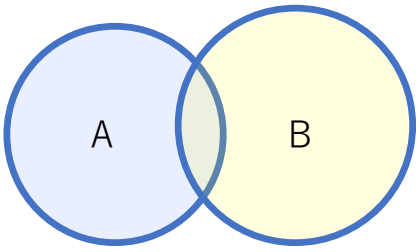
- イベント発生件数 / 生じうる全てのイベント件数
- $P(A)$：イベント（事象） Aの発生確率
- 疾病による死亡確率
  - （疾病による死亡者数）/（全人口）
  - （疾病による死亡者数）/（疾病罹患者総数）

集合の重なり方

- 集合内の部分集合が互いに背反のとき

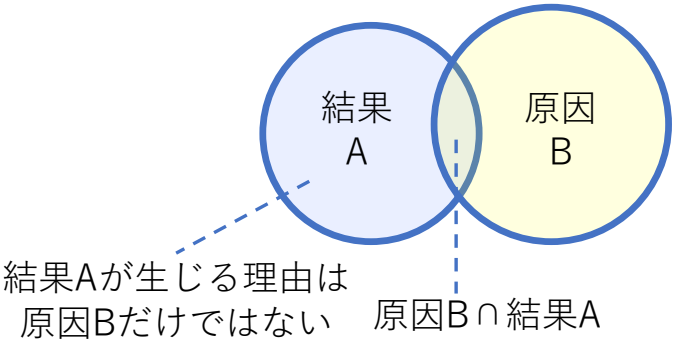


- 複数の集合が重なるとき



因果関係とベン図

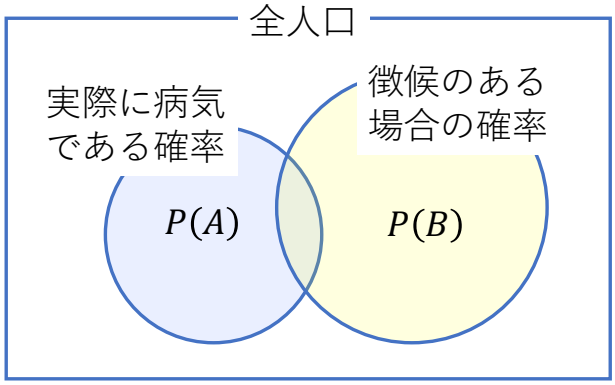
- 原因Bが理由となり結果Aが生じる場合
  =原因Bの割合に対する(原因B∩結果A)が生じる割合



原因と結果の発生確率

- 病気の徴候がある場合に実際に病気となる確率

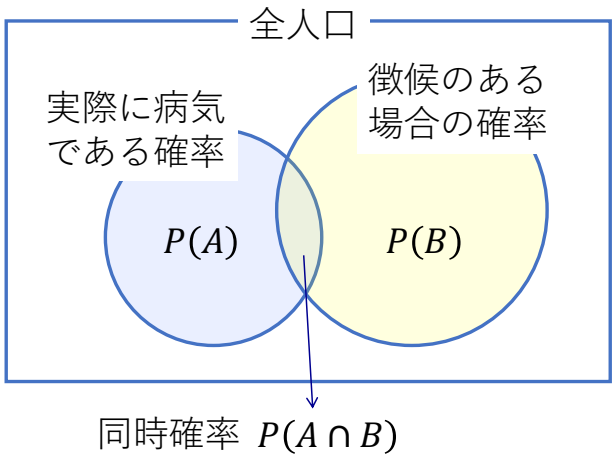
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$



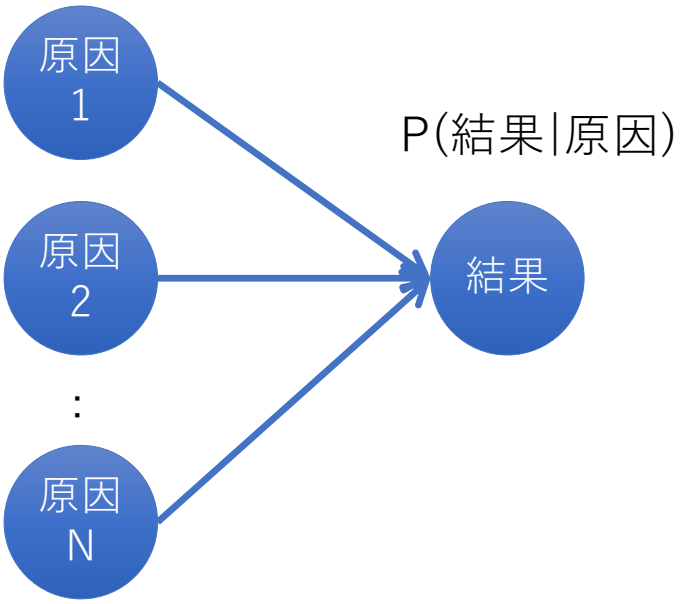
条件付き確率と同時確率

条件付き確率

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



因果関係



スクリーニングテストの例

- 兆候の有無：事象B
  - 兆候がある：B_+、兆候がない：B_-
- 健康状態：事象A
  - 健康である： $A_{健康}$、病気である： $A_{病気}$

健康状態	兆候（B）		
（A）	ー	＋	合計
健康	800	100	900
病気	25	75	100
合計	825	175	1000

スクリーニングテストの例

- 無作為に一人選んだ人が、兆候が＋でかつ病気である確率（同時確率）　$P(A_{病気} \cap B_+)$
- 無作為に選んだ人が病気である確率（周辺確率）　$P(A_{病気})$
- 無作為に選んだ人が兆候が+のとき病気であるである確率（条件付き確率）　$P(A_{病気}|B_+)$

健康状態	兆候（B）		
（A）	ー	＋	合計
健康	800	100	900
病気	25	75	100
合計	825	175	1000

S.セン（2005）、p.76

条件付き確率、同時確率、周辺確率

- 同時確率　$P(A_{病気} \cap B_{+})$ =75/1000
- 周辺確率　$P(B_{+})$ =175/1000
- 条件付き確率　$P(A_{病気}|B_{+})$=75/175
- 同時確率＝条件付き確率×周辺確率

健康状態	兆候（B）		
（A）	ー	＋	合計
健康	800	100	900
病気	25	75	100
合計	825	175	1000

S.セン（2005）、p.76

ベイズの定理の導き方

- より一般的に、二つの事象AとBがあるとする
- このとき、同時確率＝条件付き確率×周辺確率

$$P(A \cap B) = P(A|B)P(B)$$

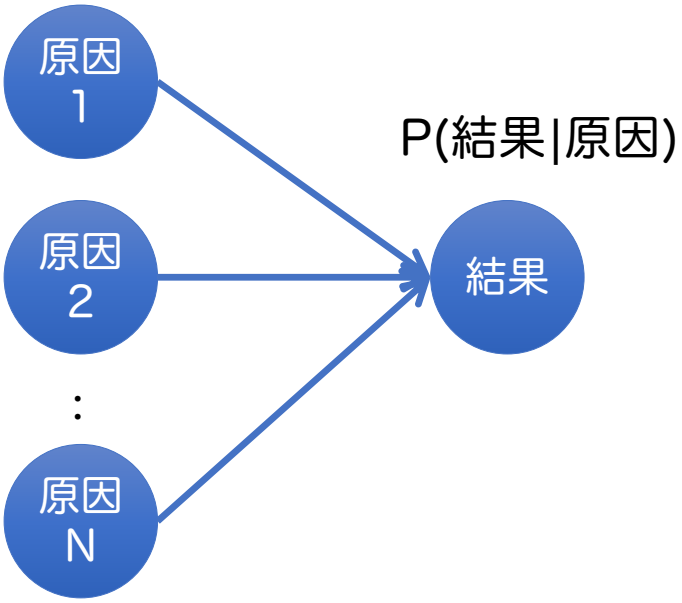
が成立する。

- 同様に、 $P(B \cap A) = P(B|A)P(A)$が成立する

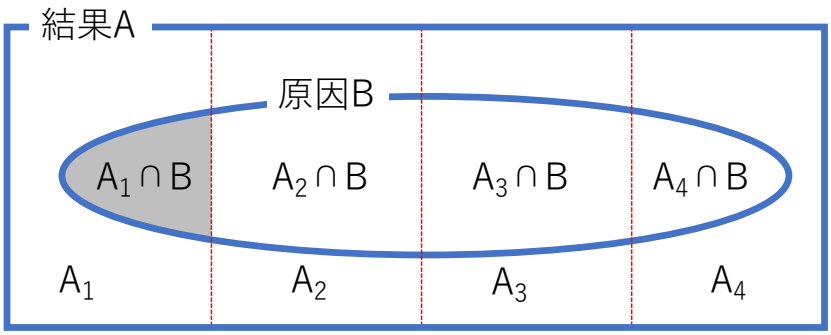
$$P(A \cap B) = P(B \cap A) \Leftrightarrow P(A|B)P(B) = P(B|A)P(A)$$
$$\Leftrightarrow P(A|B) = P(B|A)P(A)/P(B)$$

- これが「ベイズの定理」

因果関係と条件付き確率



原因と結果の関係



ベイズの定理

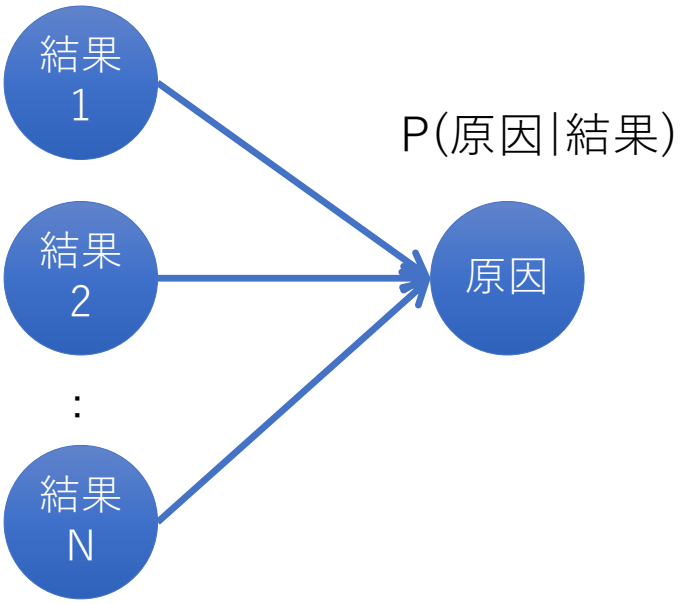
- 原因Bから結果A_1が生じうる確率

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{n} P(B \cap A_i)}$$

$$= \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

$$= \frac{P(B|A_1)P(A_1)}{P(B)}$$

逆確率による原因の推定



逆確率

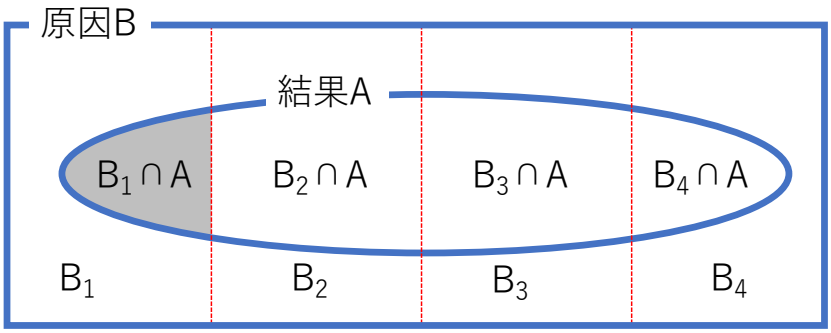
- 結果Aの原因BがB_1でありうる確率

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{n} P(A \cap B_i)}$$

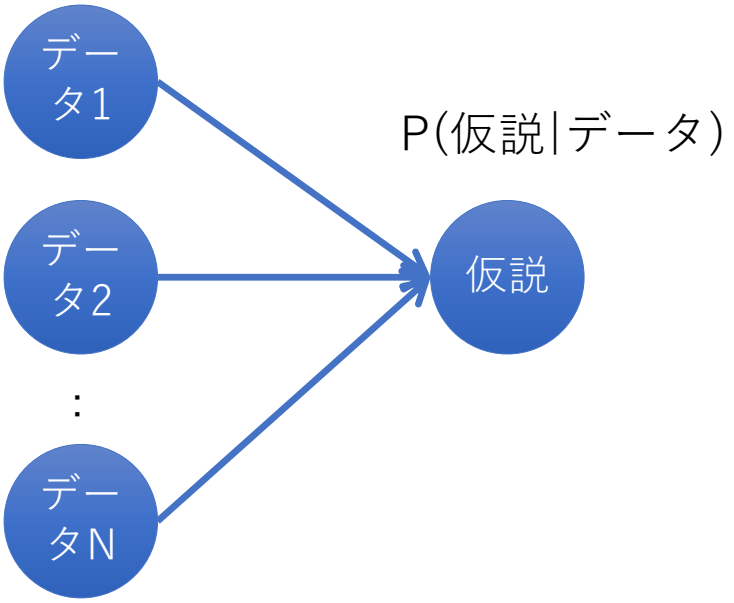
$$= \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

$$= \frac{P(A|B_1)P(B_1)}{P(A)}$$

逆確率における原因と結果の関係



データから仮説（モデル）を推定する



データから仮説（モデル）を推定する

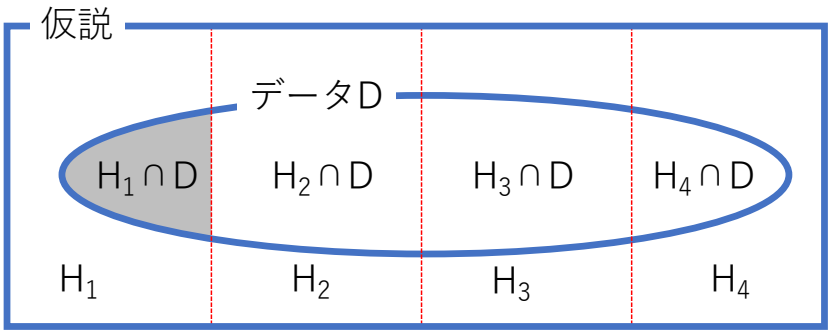
- データDから仮説H_1が成立する確率は以下のようになる

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{\sum_{i=1}^{n} P(D \cap H_i)}$$

$$= \frac{P(D|H_1)P(H_1)}{\sum_{i=1}^{n} P(D|H_i)P(H_i)}$$

$$= \frac{P(D|H_1)P(H_1)}{P(D)}$$

データから仮説（モデル）を推定する



ベイズの定理とベイズ統計

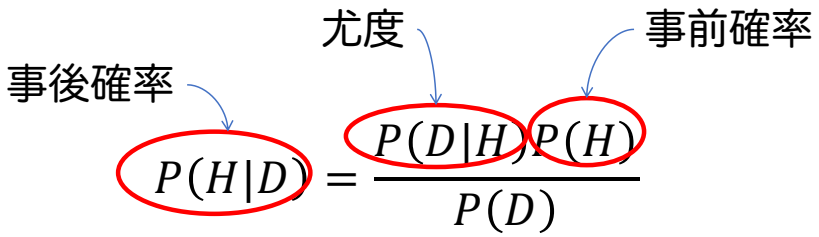
- ベイズ統計でのモデル推定は、データが与えられた条件の下で仮定されたモデルの成立する確率P(仮説|データ)を求めていることに他ならない
- 確率的には与えられたデータのもとで様々なモデル（仮説）が成立する可能性がある

尤度（ゆうど）

- 原因から結果が生じる確率P(結果|原因)を尤度という
- 原因と結果を、データと仮説（モデル）と読み替えると、
- ある仮説（モデル）を所与としてデータが得られる確率P(データ|仮説)といえる
- 仮説がデータに当てはまる当てはまりやすさ（尤もらしさ）を尤度という

事前確率・事後確率・尤度

- ベイズの定理にもとづいて、$P(H)$を事前確率、$P(D|H)$を尤度、$P(H|D)$を事後確率という



事前確率・事後確率・尤度

- データや仮説が複数（たくさん）あるとき、事前確率、事後確率、尤度はデータや仮説の値に対応した確率値をとる