ベイズ統計学に関する資料

統計学

頻度主義 (frequentism) 統計学

確率が結果 (データ) を決める

ベイズ主義 (Bayesianism) 統計学

データは定数、確率分布 (確率分布関数に含まれる定数) が分布をもつ

客観確率 (objective probability)

実験または理論的考察（思考実験）から求められ、客観的な観測結果と比較できるランダムな事象についての確率

主観確率 (subjective probability)

人間の主観的な信念あるいは信頼の度合

確率に関するベイズの定理

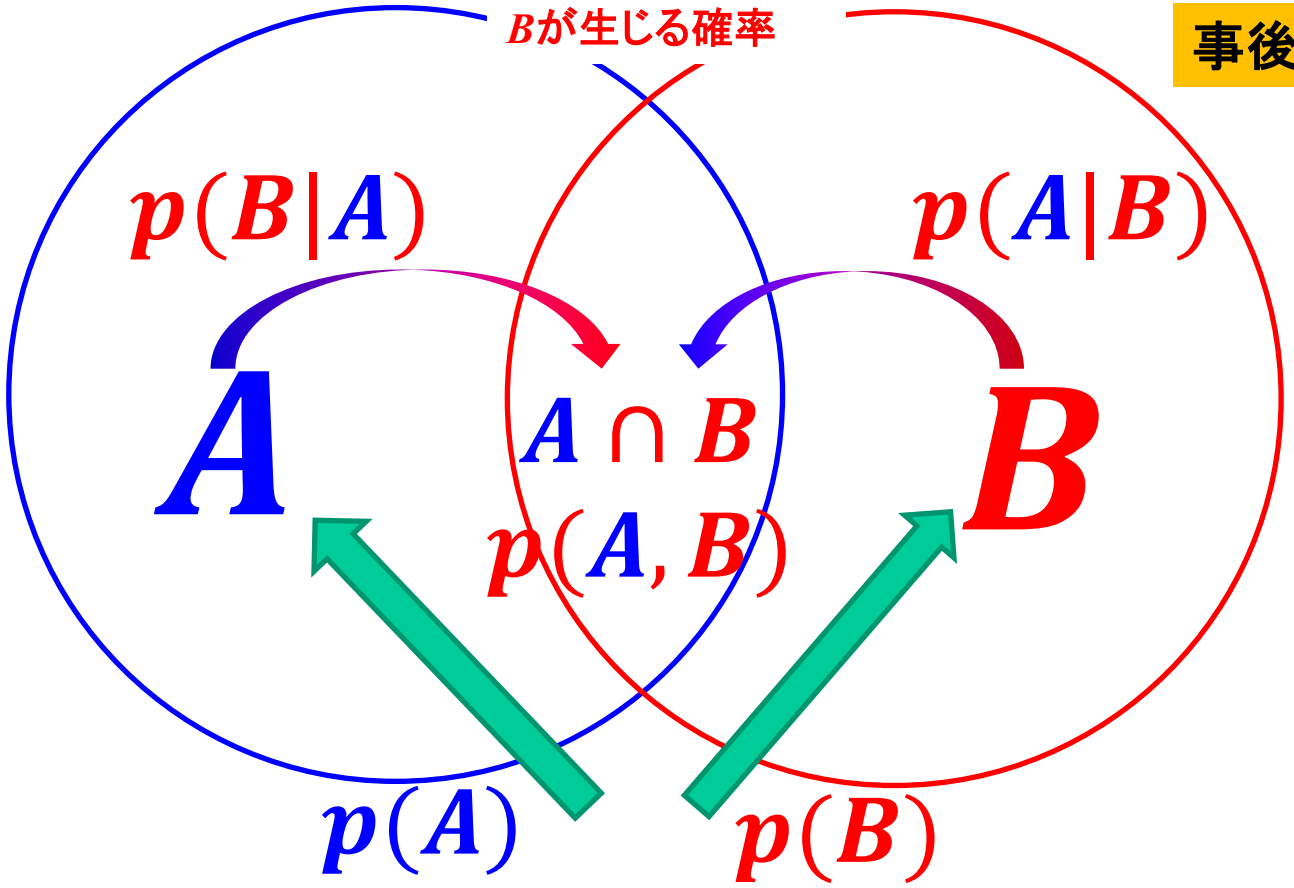
$$p(A,B) = p(B|A)p(A) = p(A|B)p(B)$$

同時確率
AとBが同時に生じる確率

事後確率
Aが生じた場合にBが生じる確率

事前確率
Aが生じる確率

事後確率 ＝ 条件付き確率



ベイズの定理: $p(A|B) = p(B|A)p(A)/p(B)$

事後確率　　事前確率

ベイズの定理と材料研究

$$p(A,B) = p(B|A)p(A) = p(A|B)p(B)$$

$p(A,B)$: AとBが同時に生じる

$p(B|A)$: Aが生じた場合にBが生じる確率

$p(A)$: Aが生じる確率

ベイズの定理: $p(A|B) = p(B|A)p(A)/p(B)$

$p(A|B)$: B が生じた場合にAが生じる確率

$p(A)$: 事前確率

A: 実験条件　B: 測定結果　と考えると、

$p(B|A)$: A の条件で実験を行った結果、B が得られる確率

ベイズの定理により、

B が得られる確率の高い条件 A を $p(A|B)$ から推定できる

問題: $p(B) = \sum_A p(B|A)$ を決めるためにはすべての実験をやらないといけない

↑今日はこの問題は忘れて説明を続ける

確率に関するベイズの定理

豊田秀樹著、基礎からのベイズ統計学、朝倉書店(2015)

$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$

A, B: 事象 (event)

$p(A)$: 事象Aが観察される確率

$p(A, B)$: 同時確率 (joint probability)。事象AとBが同時に観察される確率

$\sum_{i,j} p(A_i, B_j) = 1$

$\sum_i \ p(A_i, B_j) = p(B_j)$: 周辺確率 (marginal probability)

$p(B|A)$: 条件付き確率 (conditional probability)。

事象Aが観察されたという条件の下で、Bが観察される確率

$p(B_j|A_i) = \frac{p(A_i,B_j)}{p(A_i)}$

$\sum_j \ p(B_j|A_i) = 1$

乗法の定理 (multiplication theorem of probability): $p(A_i, B_j) = p(B_j|A_i)\, p(A_i) = p(A_i|A_j)\, p(B_j)$

全確率の公式 (law of total probability): $p(B_j) = \sum_i \ p(A_i, B_j)p(A_i)$

ベイズの定理 (Bayes' theorem): $p(A_i|B_j) = \frac{p(B_j|A_i)p(A_i)}{p(B_j)} = \frac{p(B_j|A_i)p(A_i)}{\sum_i \ p(B_j|A_i)p(A_i)}$

$p(A_i)$: 事前確率 (prior probability)

$p(A_i|B_j)$: 事後確率 (posterior probability)

$p(A_i|B_j) = p(B_j|A_i)$: AとBは互いに独立である(independent) => $p(A_i, B_j) = p(A_i)p(B_j)$

確率分布に関するベイズの定理

豊田秀樹著、基礎からのベイズ統計学、朝倉書店(2015)

母数 θ を確率変数として扱う

$f(\theta)$: 確率変数の確率分布。事前確率分布 (prior probability distribution)

$f(\theta, x) = f(x|\theta)f(\theta) = f(\theta|x)f(x)$: 確率変数が$\theta$をとり、データが x をとる同時確率

$f(\theta|x)$: データによる母数の条件付き分布。事後確率分布 (posterior probability distribution)

データが確定している場合のθの条件付き確率を

θの関数として尤度関数 (likelihood function) と呼ぶ

$f(x|\theta)$: 確率変数がθをとるとき、データが x を取る条件付き確率。

一般に 確率密度分布関数 と呼ばれる

$$f(x) = \int f(x|\theta)\, f(\theta) d\theta$$

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$: 確率分布に関するベイズの定理

$\Rightarrow$ 観測データ $\{x_i, y_i\}$ があるとき、分布関数の母数θ の分布がわかる

例えばフィッティング変数 w_i が正規分布

$$f(w_i|x_i, y_i) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

に従う場合、w_i の期待値 μ と標準偏差σ が求まる

主観確率 (subjective probability) の問題

豊田秀樹著、基礎からのベイズ統計学、朝倉書店(2015)

殺人事件が起こり、容疑者 X を捕まえた。Xの血液鑑定をしたところ、血液の特徴が一致した。
Xが犯人である確率 [事後確率 $p(犯人|一致)$] はいくらか。

条件付確率: 偶然血液の特徴が一致する確率は10^{-5}

$p(一致|犯人でない) = 10^{-5}$　　$p(一致|犯人) = 1$

(0) 単純に考えて、$p(犯人) = 1 - p(一致|犯人でない) = 0.99999$ としていいか？

日本人口 12,000万人のうち、血液型が一致するのは1,200人もいる　=> それじゃ、約1/1200？

世界人口 700,000万人のうち、血液型が一致するのは70,000人もいる　=> それじゃ、約1/70000？

事後確率 $p(犯人|一致) = p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{\sum_A p(B|A)p(A)}$

$$= \frac{p(一致|犯人)p(犯人)}{p(一致|犯人)p(犯人)+p(一致|犯人でない)\{1-p(犯人)\}} = \frac{p(犯人)}{p(犯人)+10^{-5}\{1-p(犯人)\}}$$

事前確率の仮定:

(1) 理由不十分の原則　: $p(犯人) = 1/2$

$p(B) = p(一致) = 0.50005$ => $p(犯人|一致) = 0.99999$: 犯人確定

(2) 一人当たりの殺人率: $p(犯人) = 10^{-5}$

$p(B) = p(一致) = 1.99999 \times 10^{-5}$ => $p(犯人|一致) = 0.5$

事前確率で「犯人をみつける確率は非常に低い」という思い込み (主観) があるため、事後確率が下がる

(3) 犯人を含んでいるとみられる地域には3700万人の人が住んでいる: $p(犯人) = 1/37000000$

$p(B) = p(一致) = 1.003 \times 10^{-5}$ => $p(犯人|一致) = 0.0027$

3700万人の対象者中、血液の特徴が一致する人は370人いる。$p(B) \sim 1/370$

COVID-19による罹病判定

前提:

COVID-19に感染している割合: 0.002

2020/5/16 YAHOO!ニュース歪んだ日本のPCR検査信仰、死者・感染者が少ないのには理由がある

https://headlines.yahoo.co.jp/article?a=20200516-00072596-gendaibiz-soci&p=3

感度　: 病気にかかっている人が　検診を受けると、70% (最大) の確率で陽性と判定される。

特異度: 病気にかかっていない人が検診を受けると、99% の確率で陰性と判定される。

10000人の集団を考える

- 感染者数　　20人: **入院する必要のある人数**
- 非感染者数　9980人

（1）10000人が無差別に検診をうける場合

- 感染者で　陽性と判定される　　20 * 0.7　=　14人: **正常判定で入院させられる人数**
- 非感染者で陰性と判定される　9980 * 0.99　=　9880人: 正、陰性判定
- 感染者で　陰性と偽判定される　20 – 14　=　6人: 誤、**野放しになっている感染者数**
- 非感染者で陽性と判定される　9980 – 9880 =　100人: 誤、**無駄に使われている病床**

結果: 陽性と判定された人のうち、正しく判定された人はたったの12.2%

20人しか入院の必要がないのに、100床のベッドが無駄遣いされる

感染しているのに野放しにされる人は6人

COVID-19による罹病判定

前提:

COVID-19に感染している割合: **0.002**

病気にかかっている人が　検診を受けると、**1 – 0.002**の確率で陽性と判定される。

病気にかかっていない人が検診を受けると、**1 – 0.002**の確率で陰性と判定される。

10000人の集団を考える

- 感染者数　　20人: **入院する必要のある人数**
- 非感染者数　9980人

(1') 10000人が無差別に検診をうける場合

- 感染者で　陽性と判定される　　20 * 0.998 =　20人: **正常判定で入院させられる人数**
- 非感染者で陰性と判定される　9980 * 0.998 = 9960人: 正、陰性判定
- 感染者で　陰性と偽判定される　20 –　20 =　0人: 誤、**野放しになっている感染者数**
- 非感染者で陽性と判定される　9980 – 9960 =　20人: 誤、**無駄に使われている病床**

結果: 陽性と判定された人のうち、正しく判定された人はやっと50%

検診の精度を0.998まで上げても、半数のベッドが無駄遣いされる

感染しているのに野放しにされる人はいない

COVID-19による罹病判定

前提:

COVID-19に感染している割合: 0.07 (東京都の病院での調査)

2020/5/16 YAHOO!ニュース歪んだ日本のPCR検査信仰、死者・感染者が少ないのには理由がある

https://headlines.yahoo.co.jp/article?a=20200516-00072596-gendaibiz-soci&p=3

感度　: 病気にかかっている人が　検診を受けると、70% (最大) の確率で陽性と判定される。

特異度: 病気にかかっていない人が検診を受けると、99% の確率で陰性と判定される。

10000人の集団を考える

- 感染者数　　700人: **入院する必要のある人数**
- 非感染者数　9300人

(1) 10000人が無差別に検診をうける場合

- 感染者で　陽性と判定される　　700 * 0.7　=　490人: **正常判定で入院させられる人数**
- 非感染者で陰性と判定される　9300 * 0.99　=　9207人: 正、陰性判定
- 感染者で　陰性と偽判定される　700 – 490　=　210人: 誤、**野放しになっている感染者数**
- 非感染者で陽性と判定される　9300 – 9207 =　93人: 誤、**無駄に使われている病床**

結果: 陽性と判定された人のうち、84% が正しく判定される

583床のベッドが必要

感染しているのに野放しにされる人は 210人

COVID-19による罹病判定

（1）10000人が無差別に検診をうける場合

- 感染者で　陽性と判定される　20 * 0.8 = 16人: 正、**正常判定で入院させられる人数**
- 非感染者で陰性と判定される　9980 * 0.9 = 8982人: 正、陰性判定
- 感染者で　陰性と偽判定される　20 – 16 = 4人: 誤、**野放しになっている感染者数**
- 非感染者で陽性と判定される　9980 – 8982 = 998人: 誤、**無駄に使われている病床**

結果: 陽性と判定された人のうち、正しく判定された人はたったの1.6%

20人しか入院の必要がないのに、998床のベッドが無駄遣いされる

感染しているのに野放しにされる人は4人

（2）他の症状や診断（事前診断）で、感染可能性が高い人を100人選択して検診をする場合

前提: 事前診断で選択した集団のうち、感染者の割合　10%

- 選択した集団のうち感染者　10人
- 選択した集団にとらえられなかった感染者　10人　**野放しになっている感染者数**

選択した集団のうち:

- 感染者で　陽性と判定される　10 * 0.8 = 8人: 正、**正常判定で入院させられる人数**
- 非感染者で陰性と判定される　90 * 0.9 = 81人: 正、陰性判定
- 感染者で　陰性と偽判定される　10 – 8 = 2人: 誤、**野放しになっている感染者数**
- 非感染者で陽性と判定される　90 – 81 = 9人: 誤、**無駄に入院している健常者数**

結果: 陽性と判定された人のうち、正しく判定された人は47%に増加

無駄遣いされるベッドは9床に減る

感染しているのに野放しにされる人は12人に増える

ベイズ更新 (Bayesian updating)

豊田秀樹著、基礎からのベイズ統計学、朝倉書店(2015)

メール A が迷惑メール A_1 である確率を求める。メールの特徴をB としたとき:

迷惑メールA_1である確率 $p(A_1|B)$、迷惑メールでないA_2である確率: $p(A_2|B)$

=> $p(A|B)$ が得られているとする。

ここに、ほかの特徴 C による判定を加える。

$$p(A,B,C) = p(A|B,C)p(B,C) = p(B,C|A)p(A)$$

$$\Rightarrow p(A|B,C) = \frac{p(B,C|A)p(A)}{p(B,C)}$$

A, Bが独立であれば$p(B,C) = p(B)p(C)$、

A が与えられた条件で Bと Cが独立であれば $p(B,C|A) = p(B|A)p(C|A)$

$$\Rightarrow p(A|B,C) = \frac{p(B|A)p(C|A)p(A)}{p(B)p(C)} = \frac{p(C|A)}{p(C)}\frac{p(B|A)p(A)}{p(B)}$$

$p(A,B) = p(A|B)p(B) = p(B|A)p(A)$ から、

$$p(A|B,C) = \frac{p(C|A)}{p(C)}p(A|B)$$

※ B の情報から A が迷惑メールであるかどうかの確率 $p(A|B)$ がわかっている場合に、

C による情報 $\frac{p(C|A)}{p(C)}$ を加えて、Aが迷惑メールである確率を更新 (改善) できる