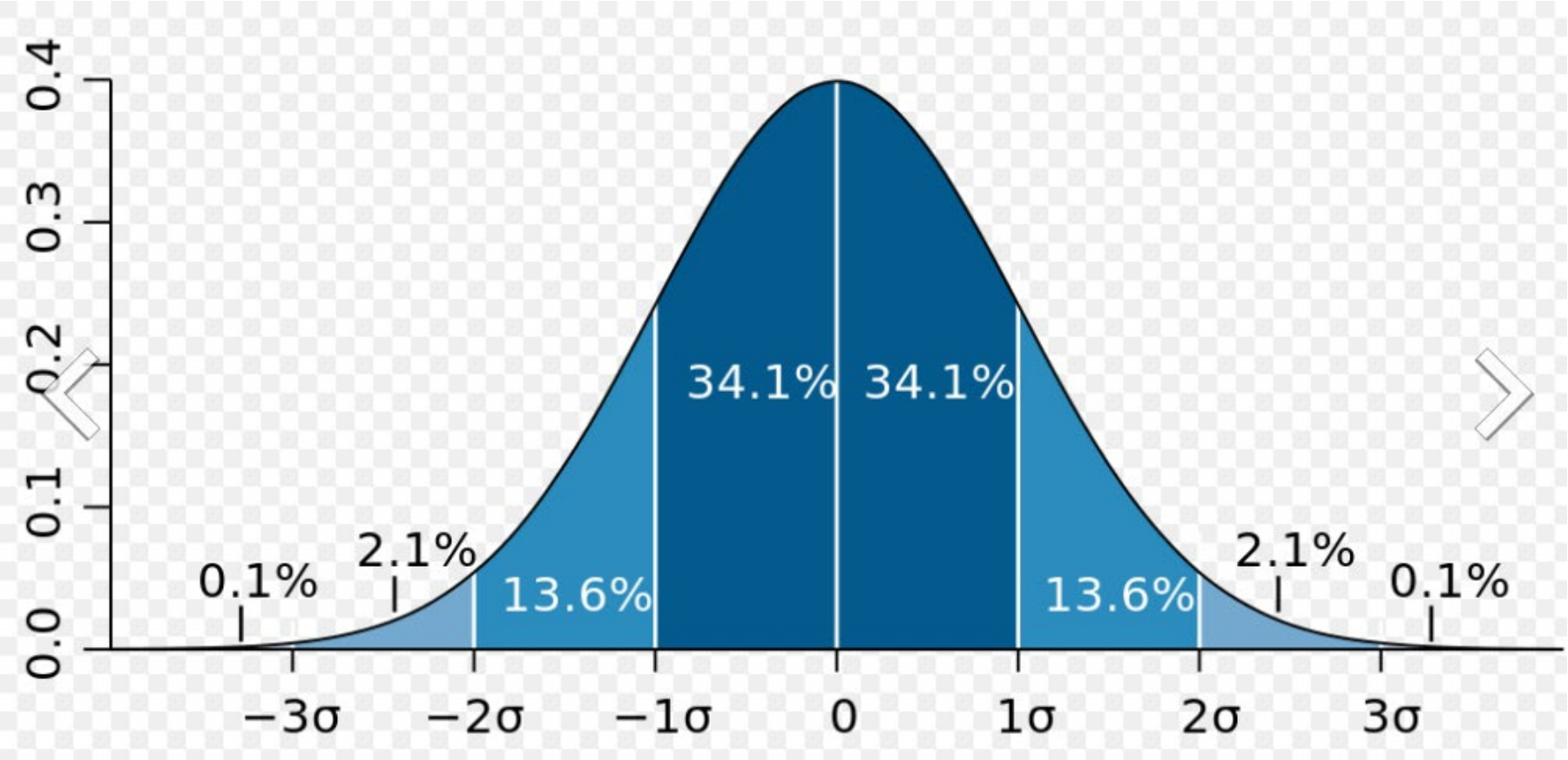
推定・検定とベイズ推定

推定



標準正規分布がもつ確率密度関数のグラフ

出典 Wikipedia

統計的推論

- A君の夏休みの自由研究発表:「尾久の原公園で蝶を10頭捕まえたら、青いものが3頭だった。この公園にいる蝶全体のうち、青いものは30%だ。」
- B先生「10頭じゃあ、ばらつきが大きそうだね」
- A君「困った.. 何頭調べようか。100頭? 1000頭?」

このような、標本(母集団の一部分) の観察にもとづいて、公園全体の蝶の色など(母集団の性質)に関するなんらかの推測的議論を行うことを統計的推論という

推定と検定

- 推定：母集団の特徴を表すなんらかの特性値（パラメータ、母数）の未知の値を標本の観察にもとづいて推測すること
- 検定：母集団の性質についてわれわれが想定すること（仮説）が標本の観察結果によって支持されるかどうかを調べること
- 推定(統計)量：推定の目的のため用いられる統計量
- 検定統計量：検定のために用いられる統計量

区間推定-比率

- n回の観察でx回ある事柄が起こった。この標本比率$\frac{x}{n}$を真の比率pの推定値として用いる。次のzは標準正規分布$N(0,1)$に従う。

$$z = \frac{\frac{x}{n} - p}{\sqrt{\frac{pq}{n}}}$$

zが-1.96と1.96の間にある確率は95%である。この式の解をp_1,p_2とすると、**母集団における比率pはp_1とp_2の間にあることが95%確かである。**

母集団におけるあるパラメータの値をある区間の間にあると推定すること:
区間推定
その区間:**信頼区間**
そのことの信頼度:**信頼係数**
zは、標準偏差を単位として、平均が基準点pからどれくらい離れているかの指標:
効果量(effect size)

区間推定-近似計算法

- 前ページのpの区間決定の式は
- $Pr\left\{\frac{x}{n}-1.96\sqrt{\frac{pq}{n}}<p<\frac{x}{n}+1.96\sqrt{\frac{pq}{n}}\right\}$=0.95
- これを書き直す。不等式の両端の辺で、

  $-\ p=\frac{x}{n}, q=1-\frac{x}{n}$

- $\Pr\left\{\frac{x}{n}-1.96\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}<p<\frac{x}{n}+1.96\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}\right\}=0.95$

- この式は見た目とは異なり簡単に計算でき、多くの場合実際的に十分である。モンテカルロ計算の結果や放射線計測の結果もこの式にもとづいて区間を推定することが多い

平均値の区間推定-大標本の場合

- $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$が標準正規分布$N(0,1)$であることから

$$Pr\{-1.96 < z < 1.96\} = 0.95$$

を用いてμの区間を推定する。

σが未知の場合には、標本標準偏差$\hat{\sigma}$を用いる。

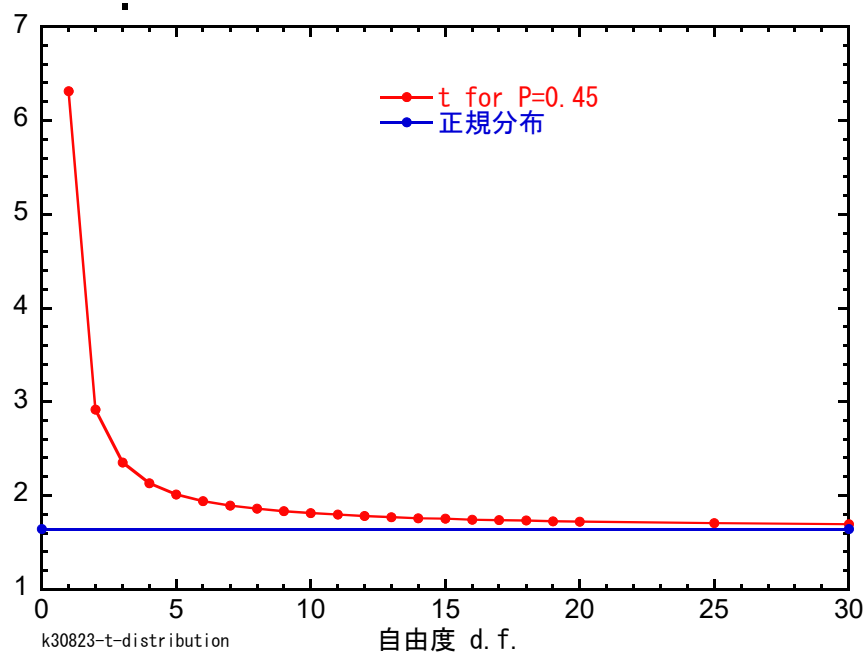
1.96は95%の信頼区間に対応。99%の場合は2.58を用いる。

平均値の区間推定-正規母集団で小標本の場合

- $t = \frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{n}}$は標本数が小さい場合、標準正規分布ではなく、自由度$n-1$のt分布に従う。
- $Pr\{-t_{0.05}(n-1) < t < t_{0.05}(n-1)\} = 0.90$
- 正規分布：0～1.645

の間に45%が含まれる。

t分布：自由度が小さいほどその範囲が広がる。



標本の大きさの決定-比率の推定

- 標本の大きさnが大であれば標本比率 $\frac{x}{n}$ は近似的に正規分布 $N\left(p, \frac{pq}{n}\right)$ に従うから、

$$\text{信頼係数95\%で} \frac{\left|\frac{x}{n} - p\right|}{\sqrt{\frac{p(1-p)}{n}}} < 1.96$$

これを*n*について解くと、$n = \left(\frac{1.96}{E}\right)^2 P(1-P)$

ただし、許容誤差 $E = |{}^{x}/_{n} - p|$

P=0.5で*n*は最大。*p*について情報があれば標本数を節約可能

*E*を小さくしようとすると標本数が多くなる。

【例】先ほどの高尾山の青い蝶の比率を許容誤差3%、信頼係数95%で求める。*p*=0.3を用いる。

$n = \left(\frac{1.96}{0.03}\right)^2 0.3\ 0.7 = 896$

平均値の推定と標本の大きさ

- 標本の大きさが大であれば、標本平均値$\bar{x}$は正規分布$N(\mu, \sigma^2/n)$に従うので、信頼係数95%で
- $\frac{|\bar{x}-\mu|}{\sigma/\sqrt{n}} < 1.96$
- これを、nについて解くと

$$\mathrm{n} = \left(\frac{1.96\sigma}{E}\right)^2$$

ただし許容誤差$E = |\bar{x} - \mu|$

標準偏差σが既知であるかあるいは近似値が必要

区間推定-分散

- $C^2 = \frac{\hat{\sigma}^2}{\sigma^2}$ が修正カイ二乗分布に従うことを用いる。
- 修正χ^2分布の表により
  - $P_r\{C^2 < {C_1}^2\}$ =0.025
  - $P_r\{C^2 < {C_2}^2\}$ =0.975

  となる${C_1}^2$と${C_2}^2$を選ぶ。（自由度に注意）

  $P_r\{{C_1}^2 < C^2 < {C_2}^2\}$ =0.95

  信頼係数95%のσ^2の信頼区間は次式で得られる。

  $P_r\left\{\frac{\hat{\sigma}^2}{{C_2}^2} < \sigma^2 < \frac{\hat{\sigma}^2}{{C_1}^2}\right\}$ =0.95

推定の練習問題

- 宮川公男著「基本統計学」第8章「推定」の文中および章末の練習問題1-27を解け。

検定

仮説の検定の手順

1. 仮説H_0を設ける。
2. 仮説を検定するための適当な統計量を選ぶ。
3. 統計量の値についてある境界値を設定し、仮説の成立にとってその値よりも不利な値の領域（仮説が正しいとするとその統計量が得られる確率が非常に小さくなるような領域）では仮説を否定する。これを「棄却域R」、逆を「採択域A」という。
4. 標本観察を行い、統計量がRに落ちれば仮説を否定し、Aに落ちれば肯定する。
5. RとAの境界：通常5%または1% 「有意水準」

はじめから否定することを狙って仮説を設けているので、検定のための仮説を帰無仮説と呼ぶことがある。

偶発的変動で標本観察結果が大きく変動することはありうる。
しかし、その確率が5%以下、1%以下のように小さいときには、H_0が誤りであるという意味のある原因によるものと考えようということ。
これが「有意」という言葉の意味。

比率の検定

- 片側検定
  - 棄却域を片側だけに設けるもの
- 両側検定
  - 棄却域を両側に設けるもの
- 母集団の比率pがある特定の値p_0に等しいといえるかどうかを調べる
- 次のz_0が棄却されるかどうかで判定する
  - z_0と正規分布関数を比較する。

$$z_0 = \frac{x/n - p_0}{\sigma}$$

分子：標本比率とあるパラメータの差、
分母：標本平均のばらつき

$$\sigma = \sqrt{p(1-p)/n}$$

比率の差の検定

- 2つの母集団の間で、ある特性を持つものの割合に差があるか？
- 2つの母集団がパラメータp_1, p_2を持つとする。

それぞれから大きさn_1, n_2の標本をとり、観察した結果、標本比率x_1/n_1とx_2/n_2を得たとする。このとき$x_1/n_1 - x_2/n_2$はn_1, n_2がともに大きければ近似的に正規分布に従う。

- 次のz_0を求める。これが0であるという仮説は、パラメータに差が無い、ということ。z_0がある範囲よりも外ならば、「差がない」という仮説が棄却され、差があることとなる。例えば有意水準5%の場合$z_0 < -1.96$または$z_0 > 1.96$の場合に「差がある」こととなる。

$$z_0 = \frac{x_1/n_1 - x_2/n_2}{\sigma}$$

分子：標本比率の差、分母：標本平均のばらつき

$$\sigma = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

P：両方の標本を一緒にして計算したものを用いる。

平均値の検定-正規分布

- 平均値の検定は、「母集団の平均値μがある特定の値μ_0 に等しい」と言えるかどうかを調べる
- 母集団の分布は正規分布、その分散の値がわかっているとする。
- N個の標本をとりその平均を$\bar{x}$とする。zは標準正規分布$N(0,1)$に従う

$$-z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

- ここで検定仮説 $H_0: \mu = \mu_0$に対してz_0を計算し、$z_0 > 1.64$

などでH_0を棄却する。(有意水準5%の片側検定の場合)

- $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

ここで右辺の分子は平均とある値の差、分母は平均の標準偏差

平均値の検定-t分布

- 母集団の分布は正規分布$N(\mu, \sigma^2)$と仮定し、母集団の分散σ^2が未知であって、その推定値として$\hat{\sigma}^2$を用いるとき統計量tが自由度n-1のt分布に従うという性質を用いて検定を行う。
- $t = \frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{n}}$
- t分布と正規分布の相違点
  - 1.64 (片側検定、$\alpha = 5\%$)などの代わりにt分布の$t_{0.05}$ (自由度n-1)などを用いること
  - 例 $t_{0.05} = 1.833$ ($n - 1 = 9$の場合)

平均値の差の検定

- 2つの異なる母集団の間で平均値がことなっているかどうかを標本観察によって検定する
- それぞれの母集団からのそれぞれ独立に大きさn_1 n_2の標本をとりその平均を$\bar{x}_1$および$\bar{x}_2$とする。また、標本分散を$\hat{\sigma}_1$および$\hat{\sigma}_2$とする
- 大標本（z_0を正規分布と比較）

$$-\ z_0 = \frac{\overline{x_1} - \overline{x_2}}{\sigma}$$

$$-\ \sigma = \sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}$$

- 小標本（t_0を自由度$n_1 + n_2 - 2$のt分布と比較）

$$-\ t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sigma}$$

$$-\ \sigma = \sqrt{\frac{(n_1-1)\hat{\sigma}_1^{\,2} + (n_2-1)\overline{\sigma}_2^{\,2}}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

検定の練習問題

- 宮川公男著「基本統計学」第9章「検定」の文中および章末の練習問題1-13、30-33を解け。
- *t*分布を学習した後に、同14-24を解け。
- χ^2分布および*F*分布を応用した検定を学習した後に、同25-29を解け。

付録　ベイズ推定

- 参考図書
  - 涌井良幸「道具としてのベイズ統計」

ベイズ統計の導入

- 「事前確率」・「主観確率」を用いる。
  - 多くの場合「あいまい」: 短所
  - 「柔軟性」: 長所
- 「ベイズ更新」
  - 新たにデータを得るごとに確率を変更すること

	データ	母数(パラメータ)*
従来の統計学 (推定と検定の部分)	確率変数	母集団固有の唯一の値が存在すると仮定
ベイズ統計学	情報の源	確率変数であり、その分布を調べようとする

*データが従う分布を決定する定数. 正規分布の場合、平均値と分散が母数

ベイズ統計と伝統的統計学の相克

南へ8.5km from 荒川C

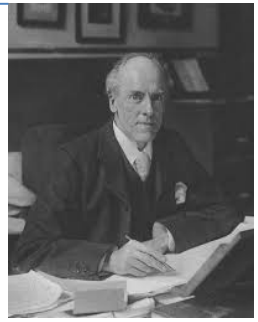
- ベイズ統計
  - トーマス・ベイズ(1702赤穂浪士討ち入り-1761)

排斥

- 伝統的な統計学（推論・検定）
  - ネイマン(1894-1891)
  - フィッシャー（1890－1962）

論敵

  - カール・ピアソン(1857-1936)
  - エゴン・ピアソン(1895-1980)



- 「21世紀のマイクロソフトの基本戦略はベイズテクノロジー」ビルゲイツ

アラン・チューリング

画像出典:Wikipedia

条件付き確率

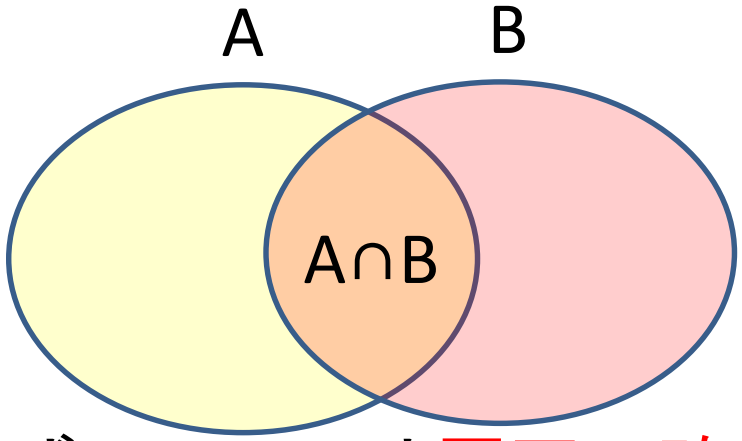
- 条件付き確率
  - ある事象Aが起こったという条件で事象Bの起こる確率
  - AのもとでBの起こる条件付き確率
  - $P(B|A), PA(B)$などという記号で表す
  - $P(B|A) = \frac{P(A \cap B)}{P(A)}$
    - $P(A \cap B)$:AとBの同時確率

- 例題　ある授業の受講者のうち、80%が日本人、72%が日本人男性である。日本人の中から一人を選んだ時、男性である確率は？
- $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{72}{100}}{\frac{80}{100}} = \frac{72}{80} = 0.9$

乗法定理

- 乗法定理
  - $P(A \cap B) = P(B|A)P(A)$
  - $P(A \cap B) = P(A|B)P(B)$
    - 「|」はギブン given と読む
- 例題1　100本中10本が当たりのくじをA氏、Bさんの順にひく。A氏もBさんも当たる確率は? くじは戻さない。
- $P(A \cap B) = P(B|A)P(A) = \frac{9}{99} * \frac{10}{100} = \frac{1}{110}$

ベイズの定理



- ベイズの定理
  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
  - $P(A|B)$と$P(B|A)$は逆確率
  - Aを原因、Bを結果と考えれば、$P(A|B)$は原因の確率
  - $P(A)$は事前確率、$P(A|B)$は事後確率
- 例題2 3枚のカードe,f,gが箱に入っている. eは両面白、fは白黒、gは両面黒である. これらの3枚の内1枚を取り出して、置いたら、上面が白であった。そのカードがfである確率は?
- 伝統解法 : 白はe表、e裏、f表の3通りなので、fの確率は$\frac{1}{3}$

例題のベイズの定理による解法

- $P(F|White) = \frac{P(White|F)P(F)}{P(White)}$
  - $P(White) = \frac{3}{6} = 0.5$
  - $P(F) = \frac{1}{3}$
  - $P(White|F) = \frac{1}{2} = 0.5$

  これらを上式に代入して

- $P(F|White) = \frac{P(White|F)P(F)}{P(White)} = \frac{0.5 \times \frac{1}{3}}{0.5} = \frac{1}{3}$
- カードは色の原因、色はカードの結果である.
- 結果の「色」から原因のカードを選ぶ確率を計算しているので、答えは「原因の確率」である.

ベイズの定理の変形（排反型）

- ベイズの定理
  - $P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)}$
  - AをA_1と書き換えた
- Bの原因としてA_1のほかにA_2, A_3も考え、互いに共通部分は無いとする。（排反）
  - $P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3)$　（排反）
  - $P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$　（乗法定理）
  - $P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+P(B|A_3)P(A_3)}$　← 代入

排反型ベイズの例題1

- 二つのツボa,bがあって、ツボaには赤玉3個、白玉2個、ツボbには赤玉8個、白玉4個入っている。ツボa,bが選ばれる確率は1:2. ツボを見ずに玉一個を取り出したら赤だった。これがツボaから選ばれた確率は？
- 事象定義: A:ツボaから玉を取り出す, B:ツボbから玉を取り出す, Red: 取り出した玉が赤玉.
- $P(A|Red) = \frac{P(Red|A)P(A)}{P(Red|A)P(A)+P(Red|B)P(B)}$

$$= \frac{\frac{3}{5}\times\frac{1}{3}}{\frac{3}{5}\times\frac{1}{3}+\frac{8}{12}\times\frac{2}{3}} = \frac{9}{29}$$

A : B

1 : 2

排反型ベイズの例題2

- 区別のつかないツボa,bがあって、ツボaには白玉1個と赤玉3個、ツボbには白玉2個と赤玉2個が入っている。どちらかのツボから玉を1個取り出したら赤玉だった。この赤玉がツボaから取り出された確率は?
- ツボa,bの選ばれる確率が与えられていないので、「両者が選ばれる確率は等しい」とする。
- 理由不十分の原則 （融通性あり vs 恣意的）

- $P(A|Red) = \frac{P(Red|A)P(A)}{P(Red|A)P(A)+P(Red|B)P(B)} = \frac{\frac{3}{4}\times\frac{1}{2}}{\frac{3}{4}\times\frac{1}{2}+\frac{2}{4}\times\frac{1}{2}} = \frac{3}{5} = 0.6$

A　　B

1　:　2

例題3

- 5回に1回帽子を忘れる人が、飲食店A,B,Cに寄って家に帰ったら帽子が無かった。Bに忘れた確率は?
- 事象定義: A:Aに入るとき帽子あり, B,C同様, F:家に帰った時帽子が無い.
- $P(B|F) = \frac{P(F|B)P(B)}{P(F|A)P(A)+P(F|B)P(B)+P(F|C)P(C)}$
  - $P(F|A) = P(F|B) = P(F|C) = \frac{1}{5}$
  - $P(A) = 1, P(B) = 1 - \frac{1}{5} = \frac{4}{5}, P(C) = (1 - \frac{1}{5})^2 = (\frac{4}{5})^2$
- $P(B|F) = \frac{\frac{1}{5}\times\frac{4}{5}}{\frac{1}{5}\times 1+\frac{1}{5}\times\frac{4}{5}+\frac{1}{5}\times(\frac{4}{5})^2} = \frac{20}{61}$

例題5

- ある病気にかかっている人に、検査法Tを適用すると98%の確率で病気であると正しく診断される
- ある病気にかかっていない人に、Tを適用すると5%の確率で誤って病気にかかっていると診断される
- 人全体からなる集団である病気にかかっている人と、かかっていない人の割合は3%と97%である
- 母集団から無作為に抽出された一人にTを適用して病気にかかっていると診断されたとき、本当に病気にかかっている確率は？

例題5の通常解

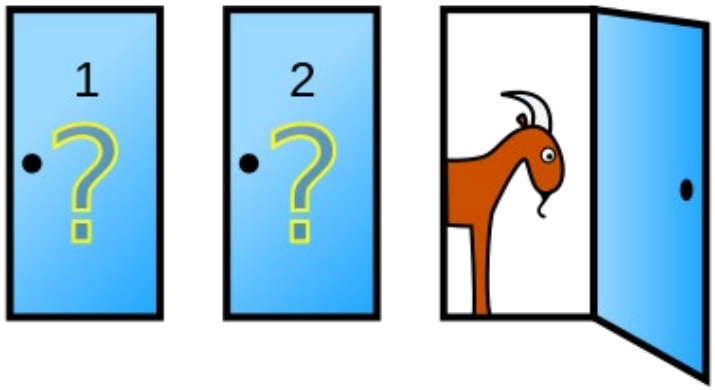
- 10000人を仮定
- 病気ではない人の数$= 10000 \times 0.97 = 9700$人
  - うち病気と診断される人数$= 9700 \times 0.05 = 485$人
- 病気の人の数$= 10000 \times 0.03 = 300$人
  - そのうち病気と診断される人数$= 300 \times 0.98 = 294$人
- $\frac{\text{病気にかかり病気と診断される人数}}{\text{病気と診断される人数}} = \frac{294}{485+294} = \frac{294}{779} \cong 38\%$

例題5のベイズ解

- 事象定義 A:病気, $\bar{A}$: 病気ではない, B:病気と診断
- $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P\left(B\middle|\bar{A}\right)P(\bar{A})}$
  - $P(A) = 0.03, P(\bar{A}) = 0.97, P(B|A) = 0.98, P(B|\bar{A}) = 0.05$ を代入
- $P(A|B) = \frac{0.98\times0.03}{0.98\times0.03+0.05\times0.97} = \frac{294}{779} \cong 38\%$

例題6 モンティーホール問題

- 設定：3枚のドアのどれか1枚に賞金
- 手順
  - 回答者が1枚のドアを選ぶ
  - 正解を知っている番組の司会者が、「残りのドアのうち、これは外れ」と言って一枚のドアを開ける
  - 回答者は、その段階でドアを変えても変えなくてもいい
- 問題：ドアを変える, 変えない, どちらが得か？



画像：Wikipediaから引用

数え上げ評価

ドアはA,B,C. 当たりはAとする

		2 司会者		
		A	B	C
1 回答者	A	レ	○	○
	B			
	C			

ドアを変えない

回答者が最初にAを選ぶ確率:$\frac{1}{3}$（レ）

そのあとで、司会者がBまたはCを選ぶ確率:$\frac{2}{2}$

賞金獲得確率=$\frac{1}{3}\frac{2}{2}=\frac{1}{3}$

赤:回答者が選んだので、司会者は選べない
緑:当たりなので、司会者は選べない
○:賞金獲得

ドアを変える

		2 司会者		
		A	B	C
1 回答者	A		はずれ	はずれ
	B	γ ○	α	β
	C	ハ○	ロ	イ

回答者が最初にBを選ぶ確率:$\frac{1}{3}$（α）

そのあとで、司会者がCを選ぶ確率:1(Aは当たりなので選べない)(β)

そのあとで、回答者がAに変える確率:1(Aしか残っていない)（γ）

回答者が最初にCを選ぶ確率:$\frac{1}{3}$（イ）

そのあとで、司会者がBを選ぶ確率:1(Aは当たりなので選べない)ロ

そのあとで、回答者がAに変える確率:1(Aしか残っていない)ハ

賞金獲得確率=$\frac{1}{3}+\frac{1}{3}=\frac{2}{3}$

ドアを変えるほうが2倍お得!

ベイズの定理によるモンティホール問題

回答者が最初にAを選ぶ

- 定義 A:ドアAが当たり, B:ドアBが当たり
  - D: ドアCが開く
- $P(A|D) = \frac{P(D|A)P(A)}{P(D|A)P(A)+P(D|B)P(B)} = \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}+1\times\frac{1}{3}} = \frac{1}{3}$ (* , #)
- $P(B|D) = \frac{P(D|B)P(B)}{P(D|A)P(A)+P(D|B)P(B)} = \frac{1\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}+1\times\frac{1}{3}} = \frac{2}{3}$

ドアを変えるほうが2倍お得!

*Aが当たりの場合、ドアBまたはドアCが等確率で開く

#Bが当たりの場合、ドアCが必ず開く(∵Aは回答者が最初に選んでいたので、開くことが無い)

例題7 迷惑メールのベイズフィルタ

- 100通のメールのうち迷惑メールと通常メールはそれぞれ70通と30通.「グラビア」という文字が迷惑メールの40通と通常メールの10通に含まれていた。グラビアという文字が含まれているメールの内、迷惑メールの割合は？
- 事象定義 M:迷惑メール, N:通常メール, G:グラビアという文字を含む
- $P(M|G) = \frac{P(G|M)P(M)}{P(G)}$
  - $P(G) = \frac{40+10}{100} = 0.5, P(M) = \frac{70}{100} = 0.7, P(G|M) = \frac{40}{70}$を代入して
- $P(M|G) = \frac{\frac{40}{70} \times \frac{70}{100}}{0.5} = \frac{40}{50} = 0.8$

ベイズ統計の基本

ベイズの定理の変形
-仮定･結果から原因･データへ-

- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
  - A（仮定）⇒H(原因), B(結果)⇒D(データ)と読み替え
- $P(H|D) = \frac{P(D|H)P(H)}{P(D)}$
- お互いに排反な原因がn個ある場合
- $P\left(H_i|D\right) = \frac{P(D|H_i)P(H_i)}{P(D|H_1)P(H_1)+P(D|H_2)P(H_2)+\cdots+P(D|H_n)P(H_n)}$

例題8(原因・結果版のベイズ定理)

- 1個のツボの中に白と赤の玉が合計3個入っている。その中から1個取り出したら赤玉がでた。ツボの中に入っている赤玉の個数の確率分布は？
- 事象定義：D:赤玉が出る,
  ツボの中の赤玉は、1個: H_1,2個: H_2 , 3個: H_3
- 解：$P(H_i|D)$をすべてのiについて求める.
- $$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D|H_1)P(H_1)+P(D|H_2)P(H_2)+\cdots+P(D|H_n)P(H_n)}$$

例題8の続き

- 代入量
  - $P(D|H_1)=\frac{1}{3}, P(D|H_2)=\frac{2}{3}, P(D|H_3)=\frac{3}{3},$
  - $P(H_1)=\frac{1}{3}, P(H_2)=\frac{1}{3}, \quad P(H_3)=\frac{1}{3}, \quad \because$理由不十分の原則
- 分母$=\frac{1}{3}\times\frac{1}{3}+\frac{2}{3}\times\frac{1}{3}+\frac{3}{3}\times\frac{1}{3}=\frac{2}{3}$
- $P(H_1|D)=\frac{\frac{1}{3}\times\frac{1}{3}}{\frac{2}{3}}=\frac{1}{6}, P(H_2|D)=\frac{\frac{2}{3}\times\frac{1}{3}}{\frac{2}{3}}=\frac{2}{6}, P(H_3|D)=\frac{\frac{3}{3}\times\frac{1}{3}}{\frac{2}{3}}=\frac{3}{6},$

ツボの中の赤玉の数	1	2	3	合計
確率	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	1

$P(H_i)$：事前確率, $P(H_i|D)$：事後確率, $P(D|H_i)$：仮定H_iの尤度(ゆうど)

ベイズの公式のまとめ

事後確率
データ*D*が得られたときに、その原因が*H*である確率

尤度
原因が*H*である時にデータ*D*が得られる確率

事前確率
原因*H*が発生する確率

- $P(H|D) = \frac{P(D|H)P(H)}{P(D)}$

得られたデータから、原因の分布を決める

ベイズの定理をさらに変形(母数化)

- ベイズの定理の仮定Hを母数θに置き換える
- $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$
- 母数を離散値から連続的な確率変数に変更
  - (事前確率) $P(\theta)$ → (事前分布) $\pi(\theta)$
  - (尤度) $P(D|\theta)$ →(尤度) $f(D|\theta)$
  - (事後確率) $P(\theta|D)$ → (事後分布) $\pi(\theta|D)$

事後分布
データDを得たときの、
母数θの確率密度分布

尤度
母数がθである時に
データDを得る確率

事前分布
母数θの確率密度関数

$$\pi(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{P(D)}$$

ベイズの定理をさらに変形（規格化定数をいったんはずす）

- 前ページの式で、$P(D)$はデータDの得られる確率である。通常データDが得られた後のことを考えるのでこの部分は定数でありθを含まないので、定数kと書き直す.
- $\pi(\theta|D) = \frac{1}{k} f(D|\theta)\pi(\theta)$
- kは総和、積分が1であることを担っており、規格化の条件である. kを省略し、等号=を比例∝に変える.
- $\pi(\theta|D) \propto f(D|\theta)\pi(\theta)$

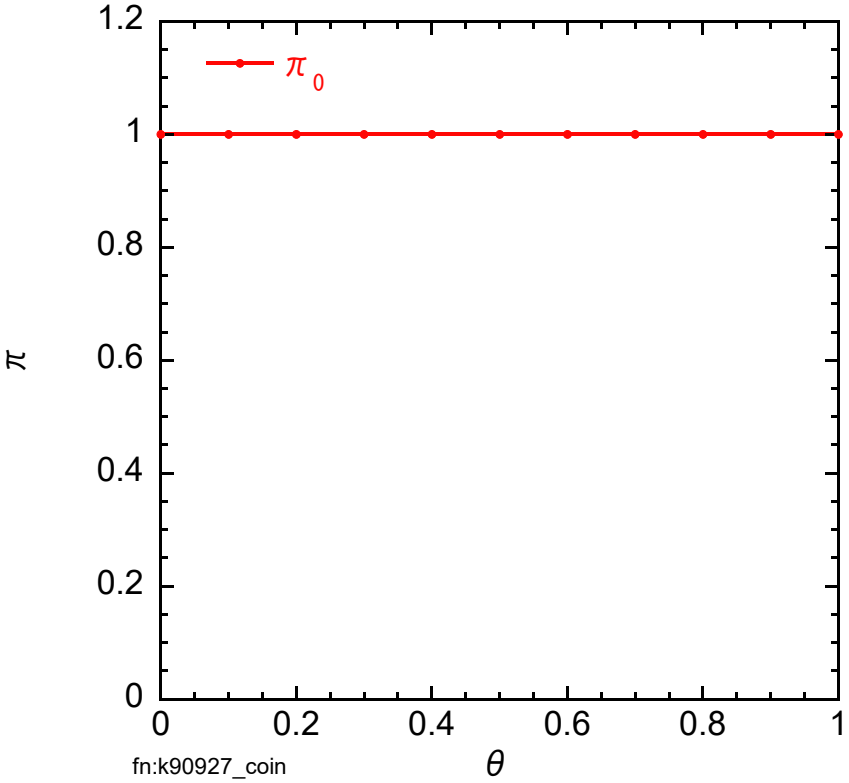
事後分布　尤度　事前分布

規格化定数をいったんはずす…例題1:お菓子

- お菓子の重さの平均値を調べる
  - 3つ取り出したところ、99g,100g,101gだった
  - 製品の重さの分散は3
  - 昨年の経験から平均値μは平均値100, 分散1の正規分布と予想.
  - このお菓子の重さの平均値μの事後分布は?
- 正規分布: $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- 尤度= $\frac{1}{\sqrt{2\pi\times 3}} e^{-\frac{(99-\mu)^2}{2*3}} \frac{1}{\sqrt{2\pi\times 3}} e^{-\frac{(100-\mu)^2}{2*3}} \frac{1}{\sqrt{2\pi\times 3}} e^{-\frac{(101-\mu)^2}{2*3}}$
- 事前分布=$\frac{1}{\sqrt{2\pi}} e^{-\frac{(100-\mu)^2}{2\times 1}}$
- 事後分布$\propto \frac{1}{\sqrt{2\pi}} e^{-\frac{(100-\mu)^2}{2\times 0.5}}$ …分散が半分になり、精度があがった

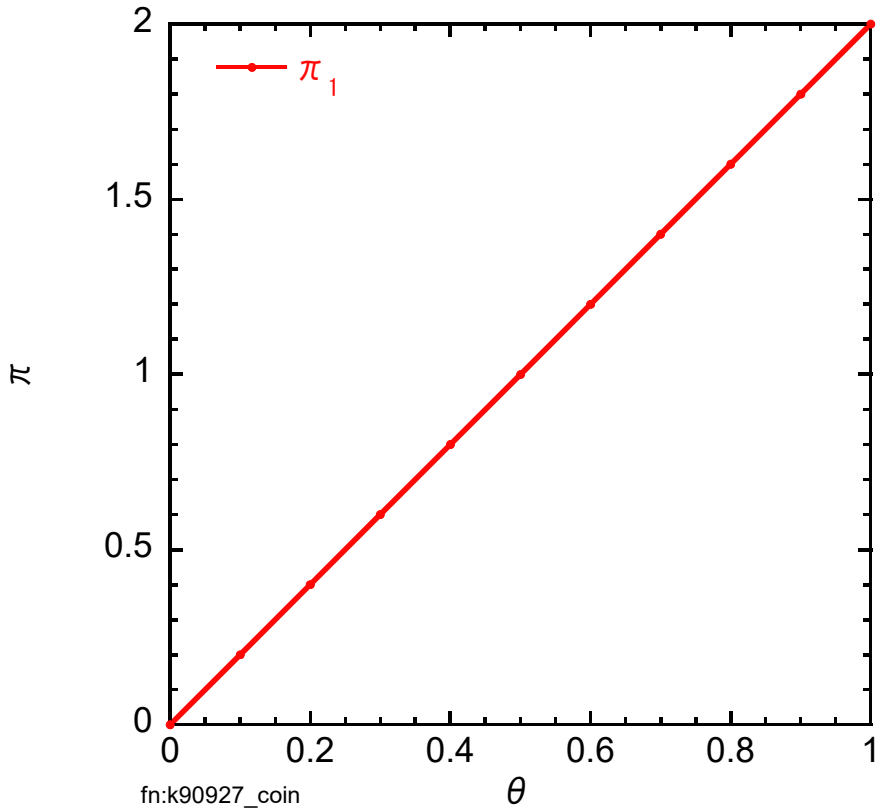
規格化定数をいったんはずす…例題2:コイン

- あるコインを4回投げたら、表、表、裏、裏と出た。この時の表の出る確率θは?
- 尤度: $f(\text{表}|\theta) = \theta, f(\text{裏}|\theta) = 1 - \theta$
- 事前分布:$\pi(\theta) = 1$
- ∵理由不十分の原則から一様分布とする



コイン例題の続き1

- 「1回目に表」のデータを取り込む
  - 1回目の事後分布$\pi(\theta|D_1) \propto \theta \times 1 = \theta$
  - 規格化条件($0 \leq \theta \leq 1$で確率の総和が1)を考慮
  - $\pi_1(\theta) = 2\theta$

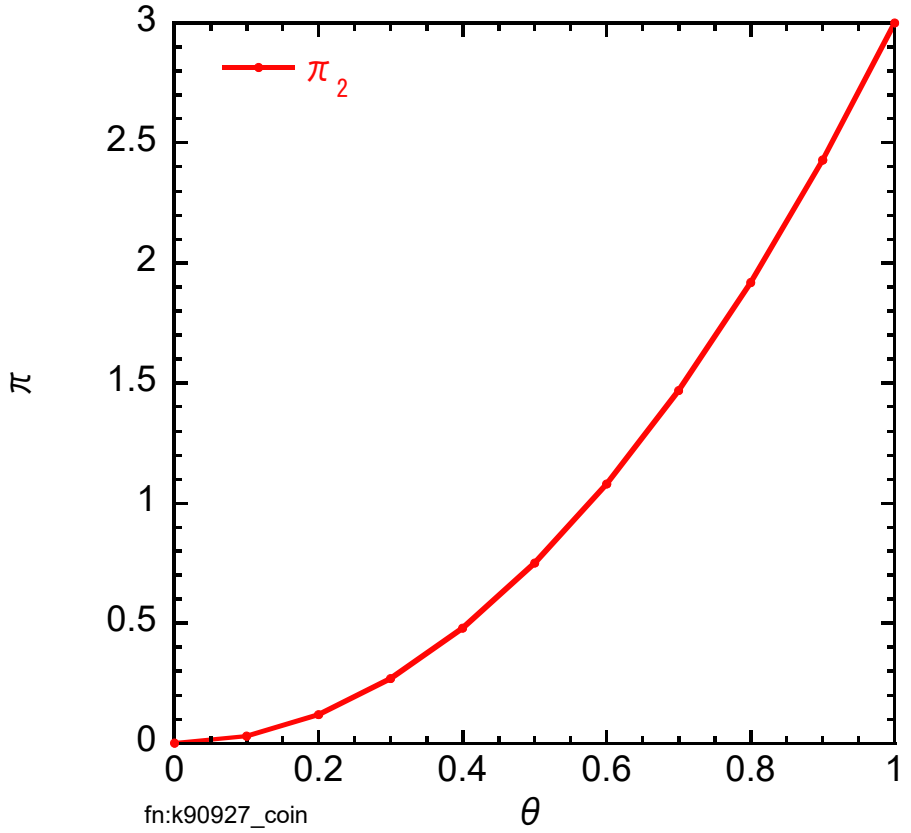


コイン例題の続き2

- 「2回目に表」のデータを取り込む
  - 2回目の事後分布$\pi(\theta|D_2) \propto \theta \times 2\theta = 2\theta^2$
  - 規格化条件($0 \leq \theta \leq 1$で確率の総和が1)を考慮
  - $\pi_2(\theta) = 3\theta^2$

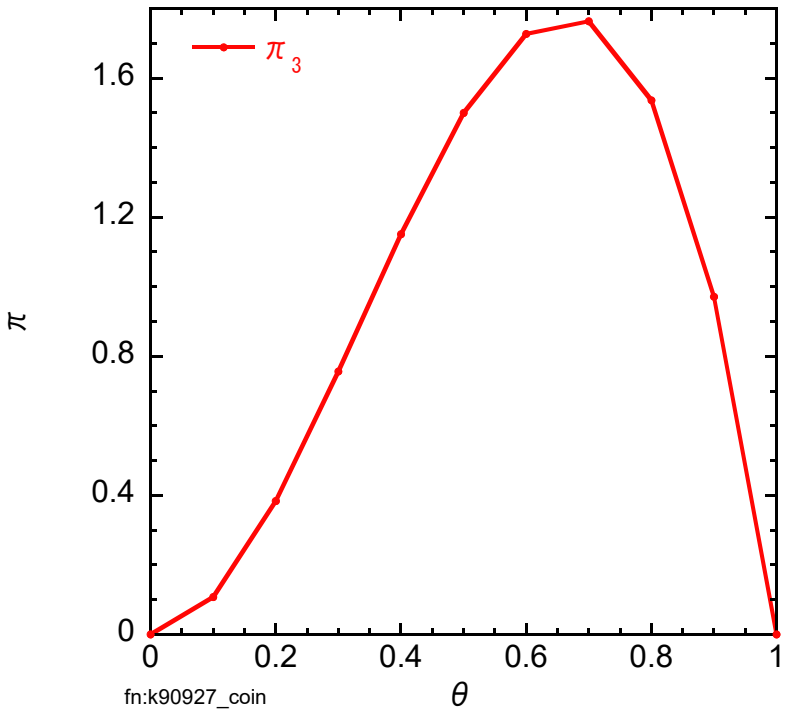
ベイズ更新

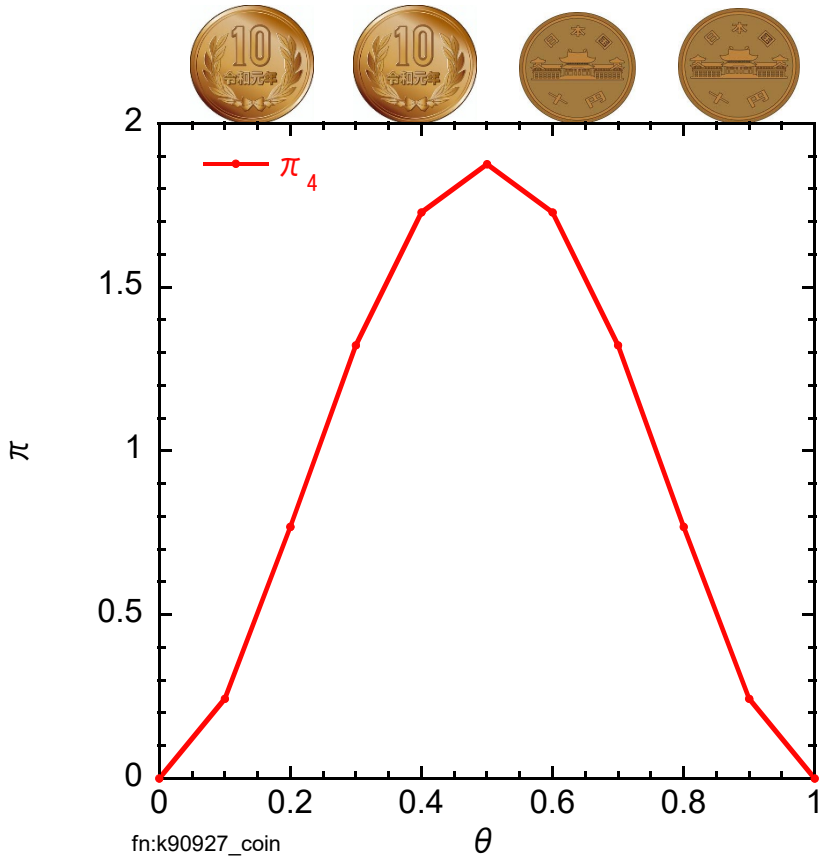
データを追加するごとに、母数θの確率分布が更新.
「母数一定」の伝統的統計学と対照的.



コイン例題の続き3

- 「3回目に裏」、「4回目に裏」のデータを取り込む
  - 3回目の事後分布$\pi(\theta|D_3) \propto (1-\theta) \times 3\theta^2$
  - 規格化条件を考慮して$\pi_3(\theta) = 12(1-\theta)\theta^2$
  - $\pi_4(\theta) = 30(1-\theta)^2\theta^2$





薬の効用問題

- 新薬の効果を調べるため5人の治験者に投薬したところ4人に効き、1人に効かなかった。この新薬の効き具合の分布は?
- 尤度 $f(D|\theta) = {}_5C_4\theta^4(1-\theta)$　二項分布の考え方から
- 事前分布$\pi(\theta) = 1$
- 事後分布$\pi(\theta|D)$

  $\propto {}_5C_4\theta^4(1-\theta) \times 1$
- 事後分布$\pi(\theta|D)$

  $= 30\theta^4(1-\theta)$

分析：θ>0.5が89%あり、有効性が期待できる

尤度に掛け合わせて、事前分布が同じ種類の事後分布に変換されるとき、その分布を尤度の自然な共役分布という. 本例題では事前分布、事後分布ともベータ分布.