ベイズのはなし

ゆっきん

（すうがく徒のつどい　第 5 回）

前置き

- **ゆっきんについて**
  - 理学部数学科卒→高校数学教師→SE→高校プログラミング講師（NOW）
  - 数学と情報の学習コミュニティとボードゲームコミュニティを運営
  - QRコードからX アカウントをフォローしていただけると嬉しいです
    （仲良くしていただけると、もっと嬉しいです）

- **本講座について**
  - 本講座はベイズ統計学の入門講演です
  - 入門講演として、初学者の方でも楽しめるように心がけます
    ぜひ楽しい時間にしましょう！！

目次

1. **条件付き確率とベイズの定理**
   条件付き確率の復習と、本講座の根幹であるベイズの定理について説明します。

2. **確率分布**
   ベイズ推論のための準備として、確率分布について説明します。

3. **ベイズ推論**
   コイン投げの例を通して、ベイズ推論の基本的な考え方を説明します。

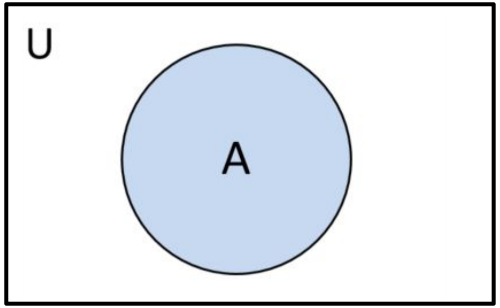
4. **ベイズ線形回帰**
   ベイズ推論の機械学習への応用として、ベイズ線形回帰を紹介します。

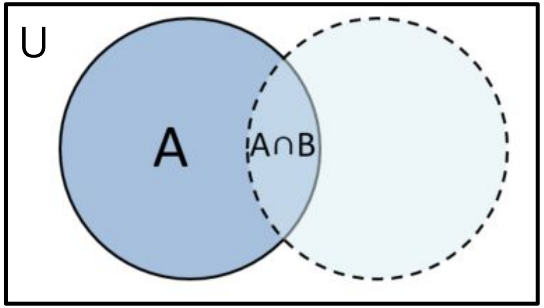
1. 条件付き確率とベイズの定理

根元事象がすべて同様に確からしい試行において、

事象 A の起こる確率は $\boxed{P(A) = \dfrac{n(A)}{n(U)}}$

事象 A が起こった時に事象 B が起こる条件付き確率は $\boxed{P(B|A) = \dfrac{n(A \cap B)}{n(A)} = \dfrac{P(A \cap B)}{P(A)}}$



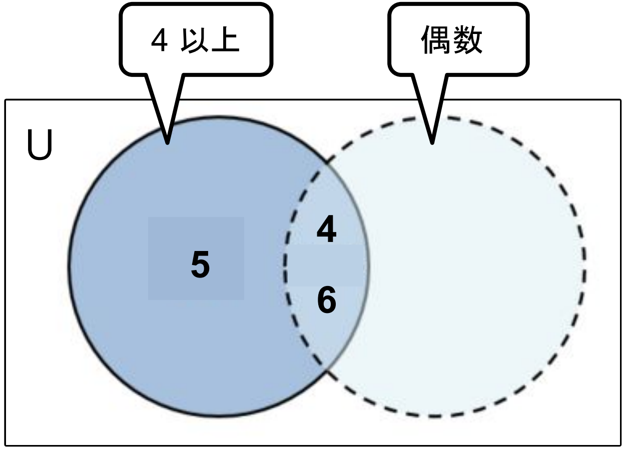


条件付き確率の例

さいころを１回振ります。
出た目が４以上のとき、その目が偶数である確率を求めなさい。

条件付き確率の例

さいころを 1 回振ります。
出た目が 4 以上のとき、その目が偶数である確率を求めなさい。



確率は $\dfrac{2}{3}$

問題「2人の子ども①」

スミスさんには子どもが2人います。2人のうち、年上の子は女の子です。
では、2人とも女の子である確率を求めなさい。
ただし、男の子と女の子は等確率で生まれるものと仮定します。

答えはどれでしょうか？

① $\frac{1}{2}$

② $\frac{1}{3}$

③ $\frac{1}{4}$

問題「２人の子ども①」

スミスさんには子どもが2人います。２人のうち、年上の子は女の子です。
では、２人とも女の子である確率を求めなさい。
ただし、男の子と女の子は等確率で生まれるものと仮定します。

	年上の子が女の子	年上の子が男の子
年下の子が女の子		
年下の子が男の子		

問題「２人の子ども①」

スミスさんには子どもが2人います。２人のうち、年上の子は女の子です。
では、２人とも女の子である確率を求めなさい。
ただし、男の子と女の子は等確率で生まれるものと仮定します。

	年上の子が女の子	年上の子が男の子
年下の子が女の子		
年下の子が男の子		

正解は

① $\frac{1}{2}$

ゆっきんと申します。よろしくお願いします！　#Techmath Project　#ボードゲーム会 Lilac

問題「２人の子ども②」

スミスさんには子どもが2人います。２人のうち、少なくとも1人は女の子です。
では、２人とも女の子である確率を求めなさい。
ただし、男の子と女の子は等確率で生まれるものと仮定します。

答えはどれでしょうか？

① $\frac{1}{2}$

② $\frac{1}{3}$

③ $\frac{1}{4}$

問題「２人の子ども②」

スミスさんには子どもが2人います。２人のうち、少なくとも1人は女の子です。
では、２人とも女の子である確率を求めなさい。
ただし、男の子と女の子は等確率で生まれるものと仮定します。

	年上の子が女の子	年上の子が男の子
年下の子が女の子		
年下の子が男の子		

問題「2人の子ども②」

スミスさんには子どもが2人います。2人のうち、少なくとも1人は女の子です。
では、2人とも女の子である確率を求めなさい。
ただし、男の子と女の子は等確率で生まれるものと仮定します。

	年上の子が女の子	年上の子が男の子
年下の子が女の子		
年下の子が男の子		

正解は ② $\frac{1}{3}$

問題「２人の子ども③　ー火曜日に生まれた少女ー」

スミスさんには子どもが2人います。２人のうち、少なくとも1人は火曜日に生まれた女の子です。では、２人とも女の子である確率を求めなさい。
ただし、男の子と女の子は等確率で生まれるものと仮定します。

火曜日に生まれたという一見関係なさそうな条件が、確率に影響を与えるのでしょうか？　ぜひ考えてみてください！

答えは、「火曜日に生まれた少女」と検索すると出てきます。

$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$ より、

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

この式をベイズの定理と呼びます。

B が何らかの観測、A をその原因としたとき、

$P(A)$: 原因Aの発生確率

$P(B)$: 観測結果Bの発生確率

$P(B|A)$: Aが発生した際に観測結果Bが発生する確率（時間順行）

$\boxed{P(A|B)}$: Bが発生した際に原因Aが起こっていた確率（時間逆行）

問題「病気に罹患している確率」

ある病気の罹患率は1%です。この病気に罹患しているか検査する方法があり、罹患している人は99%の確率で陽性と診断され、健康な人は97%の確率で陰性と診断されます。この検査で陽性と診断されたとき、実際に罹患している確率を求めなさい。

答えはどれでしょうか？

①
25%

②
50%

③
70%

④
99%

問題「病気に罹患している確率」

ある病気の罹患率は1%です。この病気に罹患しているか検査する方法があり、罹患している人は99%の確率で陽性と診断され、健康な人は97%の確率で陰性と診断されます。この検査で陽性と診断されたとき、実際に罹患している確率を求めなさい。

$$P(\text{罹患}|\text{陽性}) = \frac{P(\text{罹患})\cdot P(\text{陽性}|\text{罹患})}{P(\text{陽性})}$$

問題「病気に罹患している確率」

ある病気の罹患率は1%です。この病気に罹患しているか検査する方法があり、罹患している人は99%の確率で陽性と診断され、健康な人は97%の確率で陰性と診断されます。この検査で陽性と診断されたとき、実際に罹患している確率を求めなさい。

$$P(\text{罹患}|\text{陽性}) = \frac{P(\text{罹患})\cdot P(\text{陽性}|\text{罹患})}{P(\text{陽性})}$$

$$= \frac{\frac{1}{100}\cdot\frac{99}{100}}{\underset{\text{罹患かつ陽性}}{\frac{1}{100}\cdot\frac{99}{100}} + \underset{\text{健康かつ陽性}}{\frac{99}{100}\cdot\frac{3}{100}}}$$

問題「病気に罹患している確率」

ある病気の罹患率は1%です。この病気に罹患しているか検査する方法があり、罹患している人は99%の確率で陽性と診断され、健康な人は97%の確率で陰性と診断されます。この検査で陽性と診断されたとき、実際に罹患している確率を求めなさい。

$$P(\text{罹患}|\text{陽性}) = \frac{P(\text{罹患})\cdot P(\text{陽性}|\text{罹患})}{P(\text{陽性})}$$

$$= \frac{\frac{1}{100}\cdot\frac{99}{100}}{\underbrace{\frac{1}{100}\cdot\frac{99}{100}}_{\text{罹患かつ陽性}}+\underbrace{\frac{99}{100}\cdot\frac{3}{100}}_{\text{健康かつ陽性}}}$$

$$= \frac{1}{4}$$

正解は ① 25%

意外と低い！

問題「病気に罹患している確率」

ある病気の罹患率は1%です。この病気に罹患しているか検査する方法があり、罹患している人は99%の確率で陽性と診断され、健康な人は97%の確率で陰性と診断されます。この検査で陽性と診断されたとき、実際に罹患している確率を求めなさい。

＜別解＞人口を 10000 人と仮定します。

	陽性	陰性
罹患している人（100人）	99人	1人
罹患していない人（9900人）	297人	9603人

$$\frac{99}{99+297}=\frac{1}{4}$$

問題「病気に罹患している確率」

ある病気の罹患率は1%です。この病気に罹患しているか検査する方法があり、罹患している人は99%の確率で陽性と診断され、健康な人は97%の確率で陰性と診断されます。この検査で陽性と診断されたとき、実際に罹患している確率を求めなさい。

もともとの病気の罹患率

1%

事前確率

情報追加

検査結果：陽性

情報追加後の病気の罹患率

25%

事後確率

得た情報によって、確率が更新されました。これを**ベイズ更新**といいます。

ベイズの定理の活用例「迷惑メールフィルター」

届いたメールはどのようにして、迷惑メールかどうか判断されるのでしょうか？

ベイズの定理の活用例「迷惑メールフィルター」

届いたメールはどのようにして、迷惑メールかどうか判断されるのでしょうか？

届いたメールが迷惑メールの確率

10% → 情報追加 → **50%** → 情報追加 → **80%**

事前確率

URLが含まれている

「出会い」という
言葉が含まれている

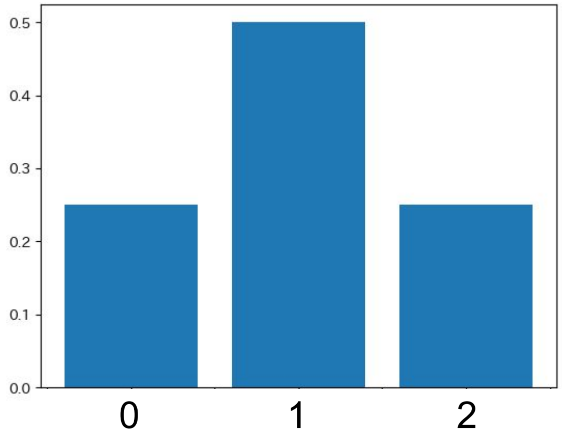
少ない情報からでも確率を出すことができ、情報が増えるごとに事後確率の精度は高くなっていきます。

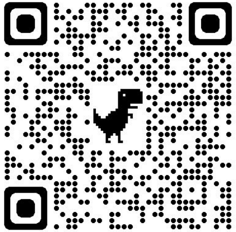
2. 確率分布

どの値を取るかが確率的に決まる変数のことを確率変数と呼びます。
また、確率を表す関数のことを確率分布と呼びます。

(例)コインを 2 回投げたときの表が出た回数X

$$P(X)=\begin{cases}\frac{1}{4} & (X=0)\\ \frac{1}{2} & (X=1)\\ \frac{1}{4} & (X=2)\end{cases}$$

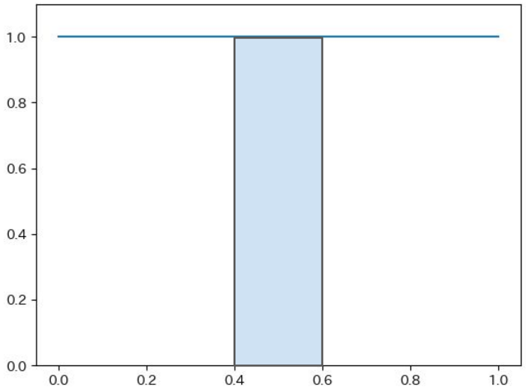




連続型の確率変数、確率分布を考えることもできます。

（例）0 から 1 までのランダムな実数 X

- $P(X = 0.5) = 0$ （一点の確率は０）
- $P(0 \leqq X \leqq 1) = 1$ （全体の確率は１）
- $P(0.4 \leqq X \leqq 0.6) = 0.2$ （区間の確率は面積）



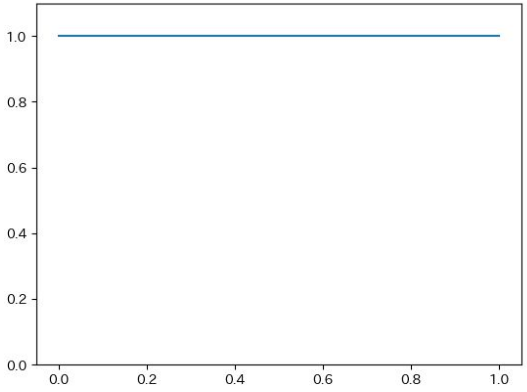
$f(x) = 1 \ \ (0 \leqq x \leqq 1)$ を X の確率密度関数と呼びます。

確率分布の例① 連続一様分布

確率変数がどのような値でも、確率密度関数が一定の値をとる分布

$$f(x) = \frac{1}{b-a} \quad (a \leqq x < b)$$

（例）0 から 1 までの実数をランダムで決定するときの値が従う確率分布



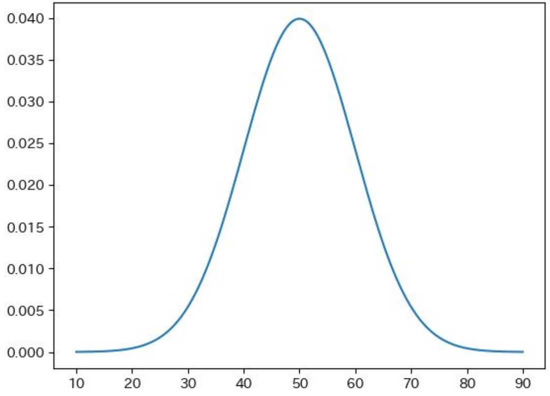
確率分布の例② 正規分布

統計における最重要分布(中心極限定理)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

μ:平均
σ:標準偏差

(例)平均50、標準偏差10の正規分布(偏差値)



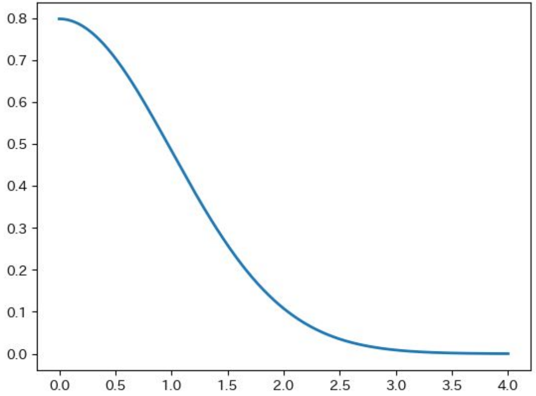
確率分布の例③ 半正規分布

正の値のみをとる正規分布を考えたい時に使用する分布

σ:標準偏差

$$f(x) = \sqrt{\frac{2}{\pi\sigma^2}} exp(-\frac{x^2}{2\sigma^2}) \qquad (x > 0)$$

（例）標準偏差1の半正規分布

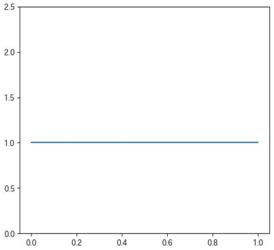


確率分布の例④ ベータ分布

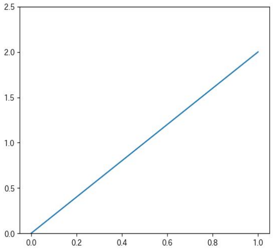
α と β の2つのパラメータによって特徴づけられる分布

面積を1にする
ための調整係数

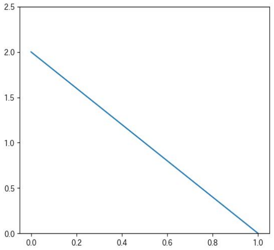
$$f(x) = C \cdot x^{\alpha-1}(1-x)^{\beta-1} \qquad C = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}$$



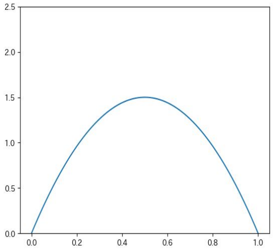
α=1, β=1



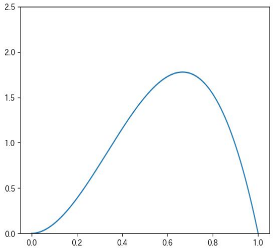
α=2, β=1



α=1, β=2



α=2, β=2



α=3, β=2

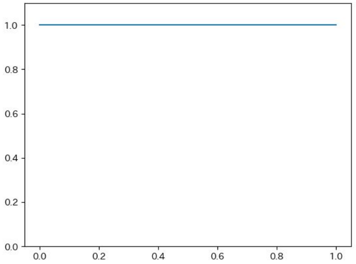
連続一様分布

3. ベイズ推論

ベイズ推論とは、パラメータｐの確率分布を推論することです。

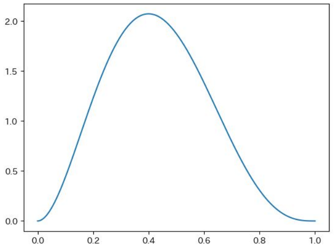
ベイズ推論の流れは以下となります。

1. パラメータｐの事前の確率分布を設定する（事前分布）
2. 情報が得られる
3. 情報によって、パラメータｐの確率分布が更新される（事後分布）



事前分布

情報追加



事後分布

問題「くじ引きで当たりを引く確率」

当たりの確率が一定のくじを5 回引いたとき、結果は「当たり・外れ・外れ・当たり・外れ」でした。このくじが当たる確率p はどのくらいでしょうか？

この問題を最尤推定という方法と、ベイズ推論という方法の2通りで考えましょう。

問題「くじ引きで当たりを引く確率」

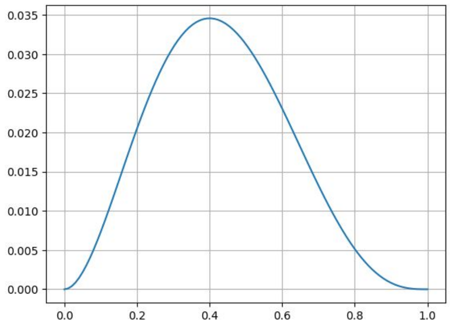
当たりの確率が一定のくじを5 回引いたとき、結果は「当たり・外れ・外れ・当たり・外れ」でした。このくじが当たる確率p はどのくらいでしょうか？

＜最尤推定＞

結果が「当たり・当たり・外れ・当たり・外れ」となる確率は、

$$p \cdot (1-p) \cdot (1-p) \cdot p \cdot (1-p) = p^2(1-p)^3$$

尤度関数



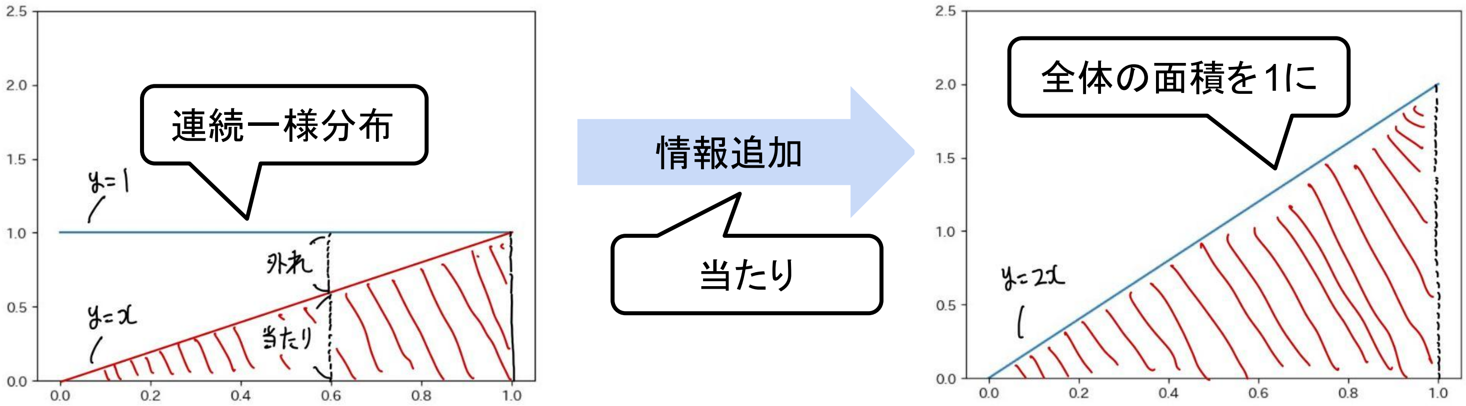
尤度関数が最大となるのは　$p = 0.4$ のとき。

問題「くじ引きで当たりを引く確率」

当たりの確率が一定のくじを5 回引いたとき、結果は「当たり・外れ・外れ・当たり・外れ」でした。このくじが当たる確率p はどのくらいでしょうか？

＜ベイズ推論＞

p の事前分布として、連続一様分布を考えます。当たりの情報で確率分布が更新されます。

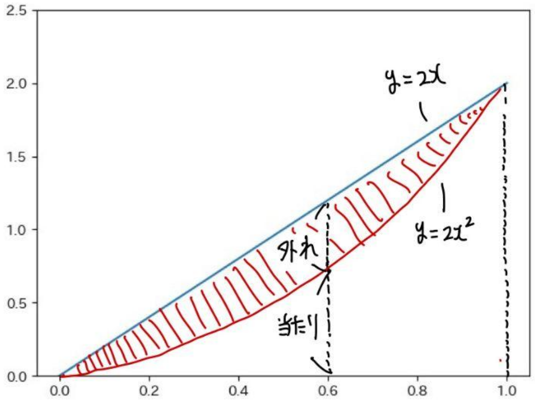


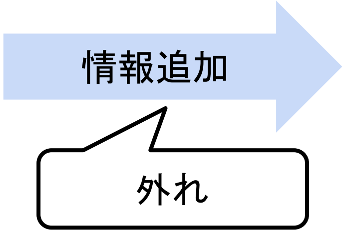
問題「くじ引きで当たりを引く確率」

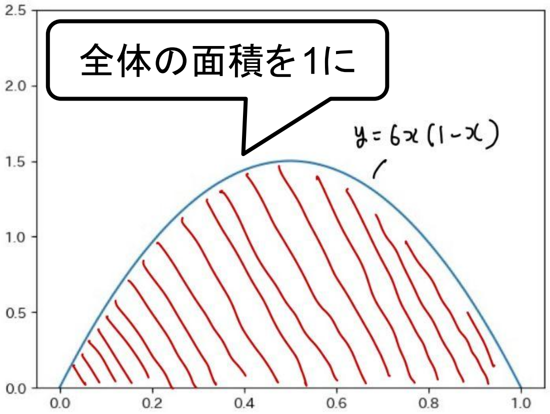
当たりの確率が一定のくじを5 回引いたとき、結果は「当たり・外れ・外れ・当たり・外れ」でした。このくじが当たる確率p はどのくらいでしょうか？

＜ベイズ推論＞

次に外れの情報で、また確率分布が更新されます。





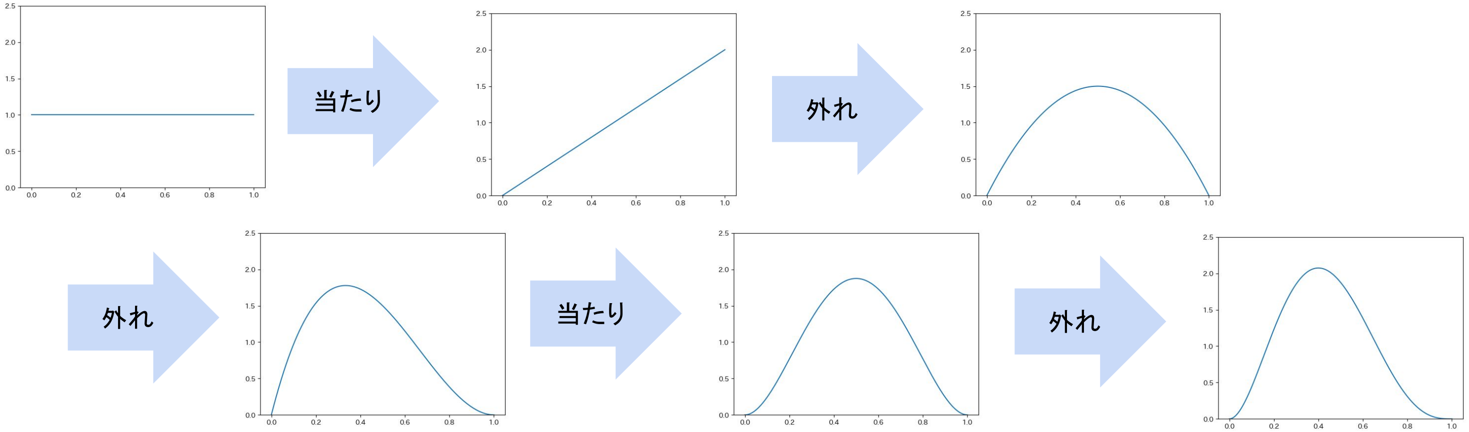


問題「くじ引きで当たりを引く確率」

当たりの確率が一定のくじを5 回引いたとき、結果は「当たり・外れ・外れ・当たり・外れ」でした。このくじが当たる確率 p はどのくらいでしょうか？

＜ベイズ推論＞

同じ流れで、確率分布は以下のように更新されます。



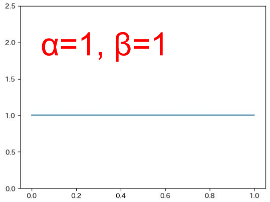
問題「くじ引きで当たりを引く確率」

当たりの確率が一定のくじを5 回引いたとき、結果は「当たり・外れ・外れ・当たり・外れ」でした。このくじが当たる確率 p はどのくらいでしょうか？

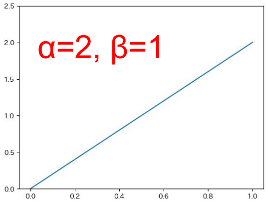
＜ベイズ推論＞

同じ流れで、確率分布は以下のように更新されます。

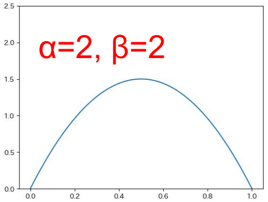
実は全て
ベータ分布！



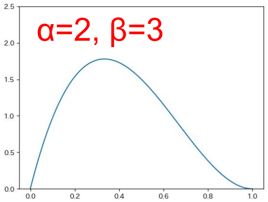
当たり



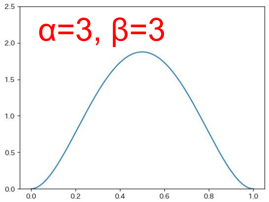
外れ



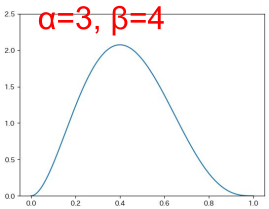
外れ



当たり



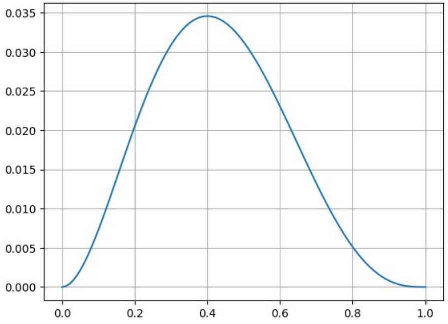
外れ



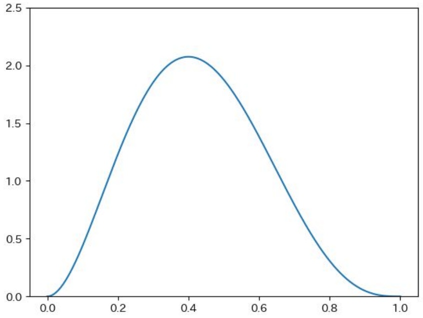
問題「くじ引きで当たりを引く確率」

当たりの確率が一定のくじを5 回引いたとき、結果は「当たり・外れ・外れ・当たり・外れ」でした。このくじが当たる確率 p はどのくらいでしょうか？

今回の場合、最尤推定における尤度関数のグラフと、ベイズ推論で得られる事後分布は同じ形状となりました。



尤度関数のグラフ



ベイズ推論の事後分布

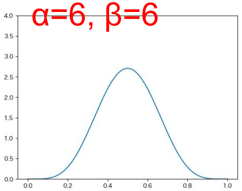
問題「くじ引きで当たりを引く確率」

当たりの確率が一定のくじを5 回引いたとき、結果は「当たり・外れ・外れ・当たり・外れ」でした。このくじが当たる確率p はどのくらいでしょうか？

＜ベイズ推論＞

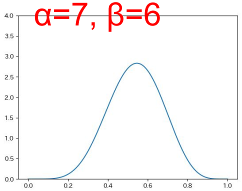
p の事前分布として、連続一様分布以外を考えることもできます。

α=6, β=6



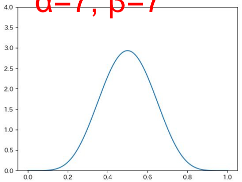
当たり

α=7, β=6



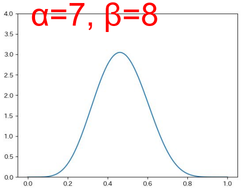
外れ

α=7, β=7



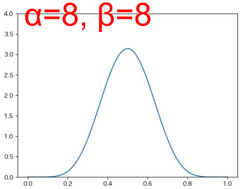
外れ

α=7, β=8



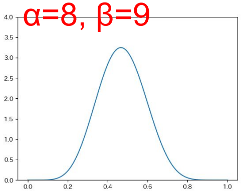
当たり

α=8, β=8



外れ

α=8, β=9



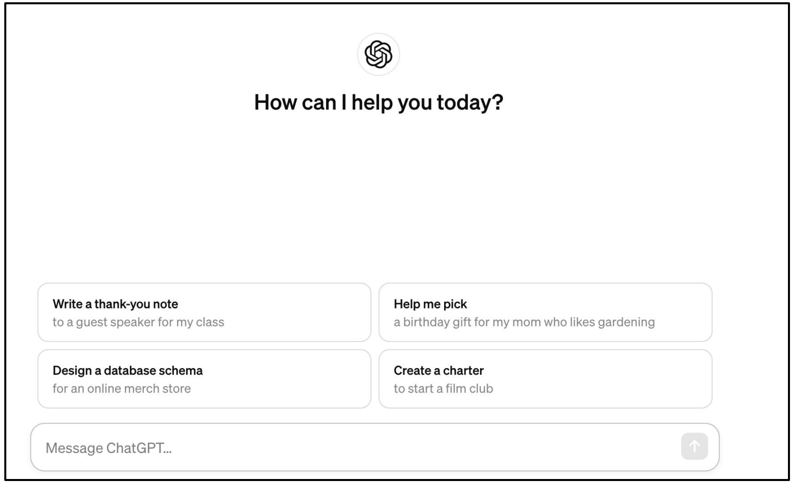
α=6, β=6 の事前分布は、「当たり×5、外れ×5」の事前データを持っていることと同じ

ゆっきんと申します。よろしくお願いします！　#Techmath Project　#ボードゲーム会 Lilac

4. ベイズ線形回帰

AI（人工知能）は、ルールベースの手法と機械学習の手法に分かれます。

- ルールベースの手法：人がルールを定める
- 機械学習の手法：機械（コンピュータ）がデータをもとにルールを学習する



機械学習はさらに教師あり学習、教師なし学習、強化学習に分かれます。

このうち教師あり学習とは、正解つきのデータをもとにルールを学習し、未知のデータの正解を予測することを言います。

予測の中でも数値を予測することを、回帰と呼びます。

回帰の例　カリフォルニアの住宅価格

- 世帯所得
- 住宅の築年数
- 住宅の部屋数
- 居住人数

説明変数（入力）

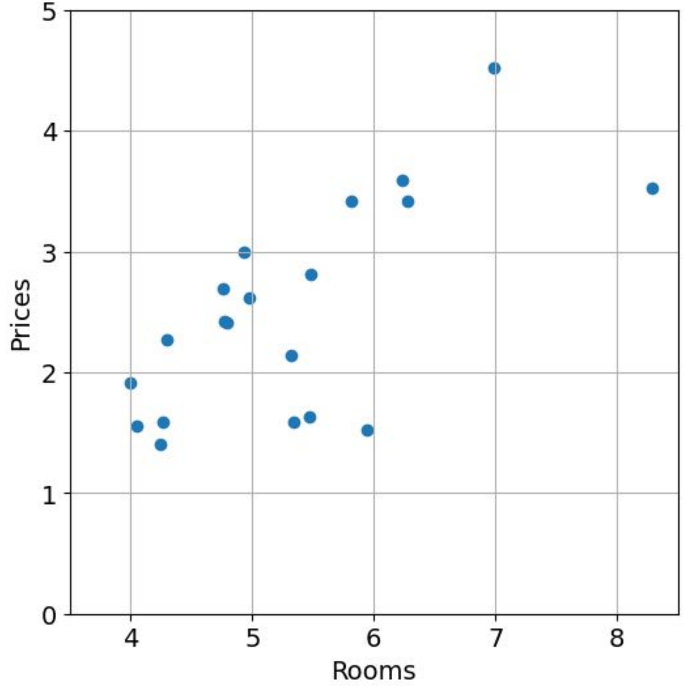
→

- 住宅価格

目的変数（出力）

回帰の例「カリフォルニアの住宅価格」

住宅の部屋数と住宅価格の関係性を調べてみましょう。

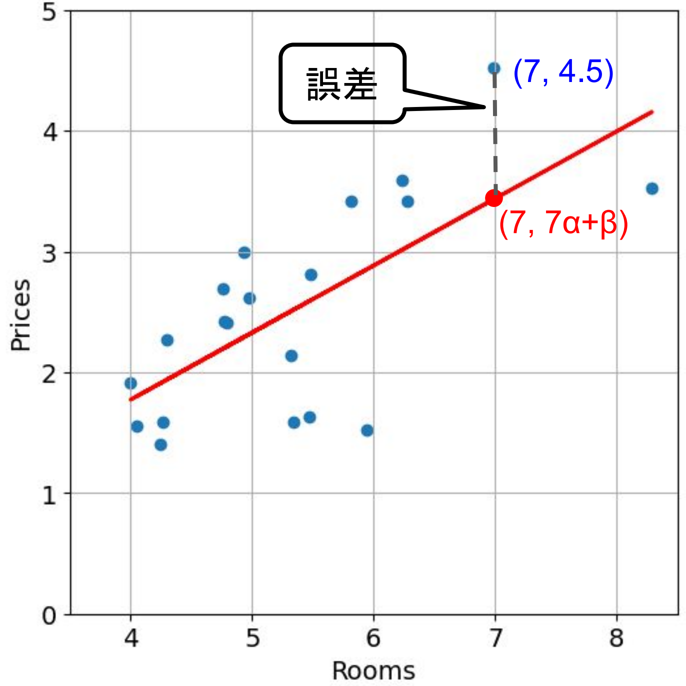


1 次式 $y = \alpha x + \beta$ で 2 変数の関係を近似しましょう。これを線形回帰といいます。

最小 2 乗法を用いた一般的な線形回帰と、
ベイズ線形回帰の 2 通りの方法で考えます。

回帰の例「カリフォルニアの住宅価格」

住宅の部屋数と住宅価格の関係性を調べてみましょう。



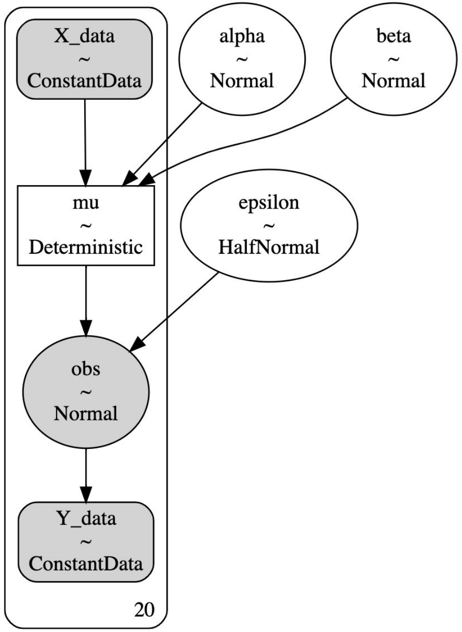
最小 2 乗法では、直線と各データの y 座標の誤差を調べます。

誤差の 2 乗和は α と β の 2 次関数になります。この 2 次関数が最小となる α と β を求めれば OK！

最小 2 乗法では、α や β が 1 つの値に定まります。

回帰の例「カリフォルニアの住宅価格」

住宅の部屋数と住宅価格の関係性を調べてみましょう。



ベイズ推論とは、パラメータの確率分布を推論することでした。

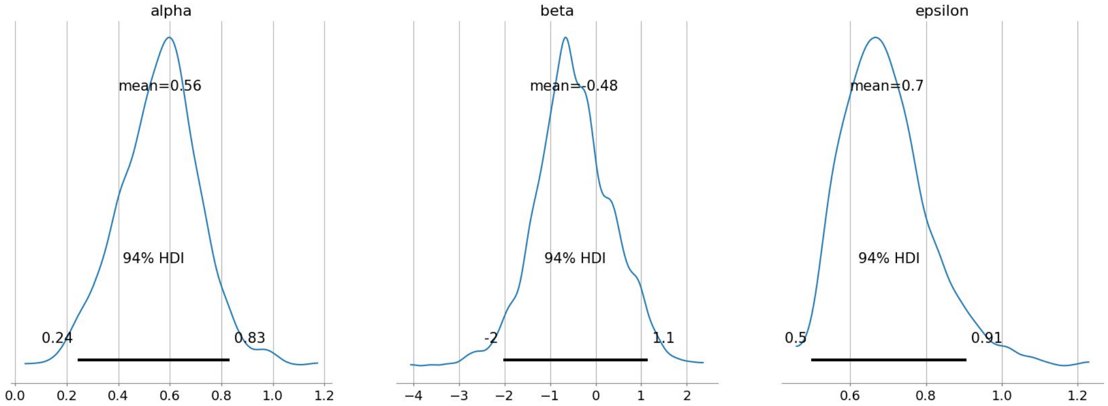
今回パラメータとして考えるのは次の3 種類です。

- 傾き α
  事前分布は平均値0、標準偏差10の正規分布
- 切片 β
  事前分布は平均値0、標準偏差10の正規分布
- 誤差 ε
  事前分布は標準偏差1の半正規分布

回帰の例「カリフォルニアの住宅価格」

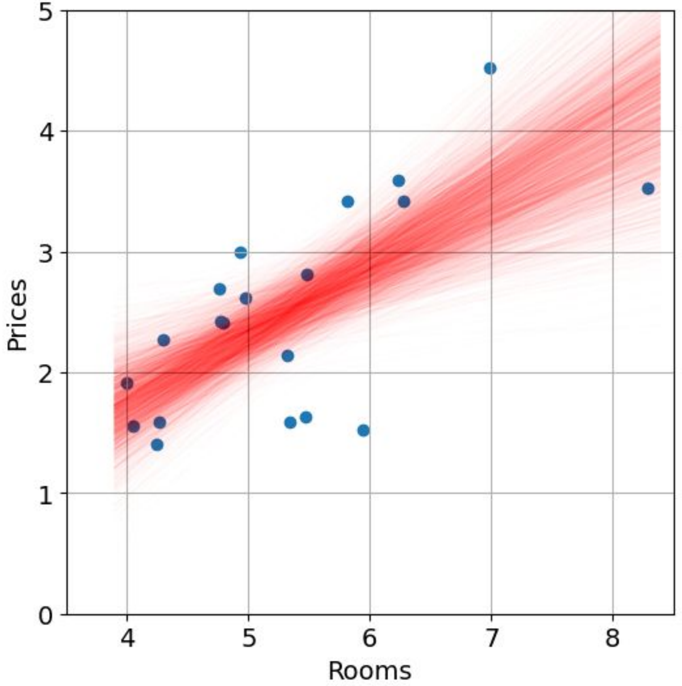
住宅の部屋数と住宅価格の関係性を調べてみましょう。

情報によって更新された後のα, β, ε の事後分布は次のようになります。
事後分布は通常は解析的に解けない複雑な形をしているため、サンプリングによって近似的に事後分布を求めます。(マルコフチェインモンテカルロ法(MCMC))



回帰の例「カリフォルニアの住宅価格」

住宅の部屋数と住宅価格の関係性を調べてみましょう。



ベイズ線形回帰では、α や β の確率分布が得られますので、それをもとに回帰直線を描くと幅を持った状態で示されます。

これにより、不確実性がどの程度であるかを表現できています。

まとめ

- 確率は情報を得ることで更新されます。（ベイズの定理）
- ベイズ推論は、ベイズの定理を土台とした推論の方法です。パラメータの事前分布と得られたデータをもとに、パラメータの事後分布を推論します。
- ベイズ推論には次のようなメリットがあります
    - 推論の結果が確率分布であることから、不確実性が表現されている
    - データが不十分な場合にも使うことができる
    - 事前知識や経験を事後推定に組み込むことができる

発表は以上となります。

ご清聴ありがとうございました！