第一回ベイズ統計学・機械学習研究会

～ベイズってなに？～

目次

- ベイズ統計学の歴史
- ベイズの定理
- ベイズ統計学の用途
- 最尤法とベイズ統計学
  - コインを使った例

ベイズ統計学の歴史

- 1740年ごろトーマス・ベイズがベイズの定理を発見
- フィッシャーとネイマン、ピアソンがベイズ統計学を否定
  - 事前分布の決め方に主観が入ることがあり数学的に厳密ではないから
- 1950年ドイツ軍の暗号をイギリスがベイズ統計学を用いて解読
- コンピューターの計算能力が向上し今まで手計算では太刀打ちできなかった問題が解けるように
  - マルコフ・チェーン・モンテカルロ（ＭＣＭＣ）の登場

ベイズの定理

- 使う道具はこれだけ

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

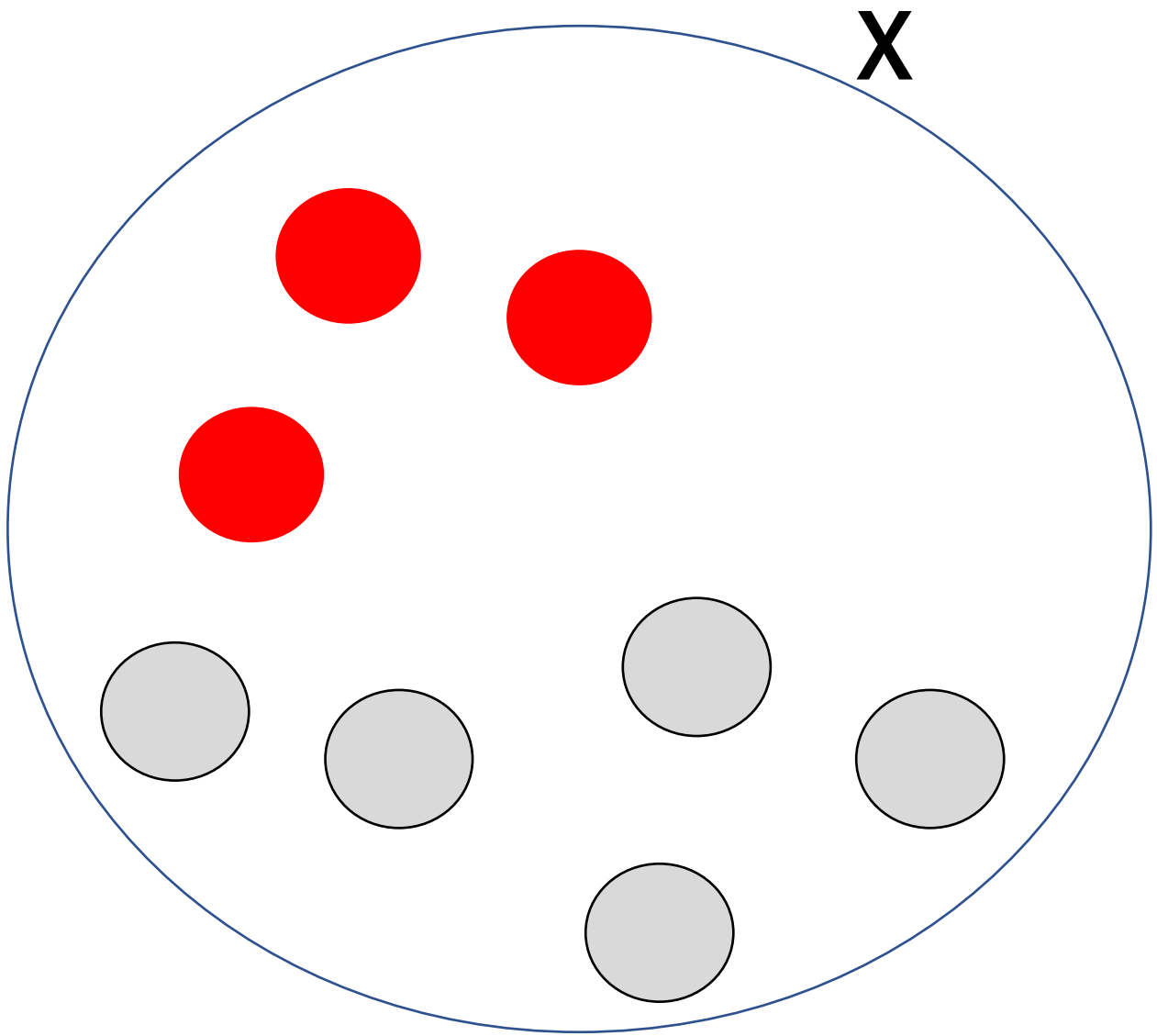
例題

- XとYという二つの袋を用意する。Xには赤玉が3つ白玉が5つ、Yには赤玉が1つ白玉が3つ入っている。二分の一の確率でXかYを選び、その中の玉を一つとる。

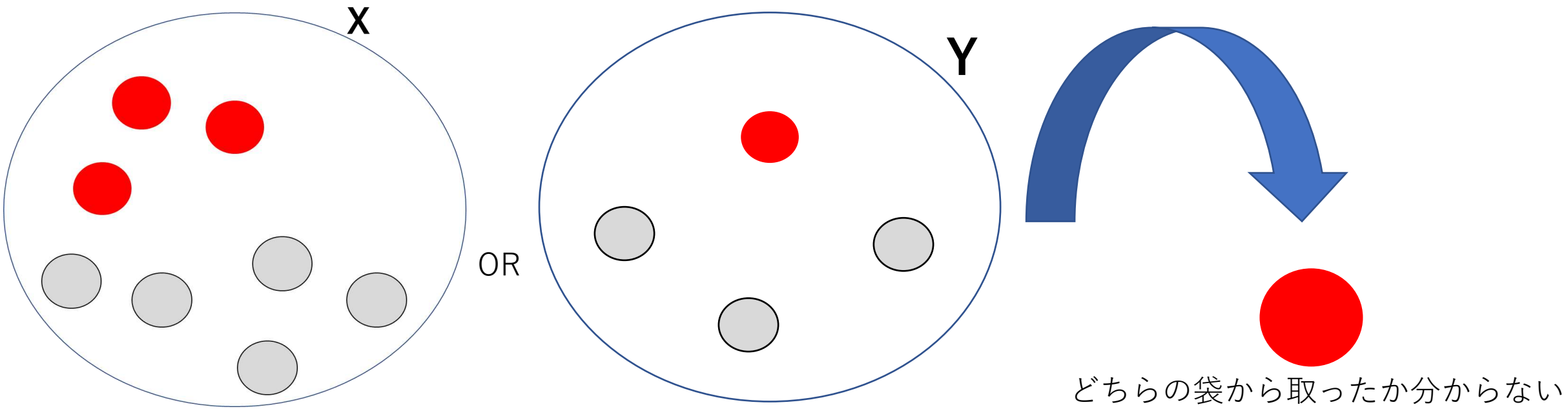
- Xを選んだ時赤玉を取る確率は？
- 赤玉を取った時、それがXの袋に入っていたものである確率は？

- Xを選んだ時赤玉を取る確率は？

条件付確率
P(赤玉|X)=3/8



- 赤玉を取った時、それがXの袋に入っていたものである確率は？



$$P(X|赤玉)=\frac{P(X)*P(赤玉|X)}{P(赤)}$$

P(X)=1/2 (Xの袋を選ぶ確率)

P(赤)=(1/2*3/8)+(1/2*1/4)

=5/16 (赤い球を引く確率)

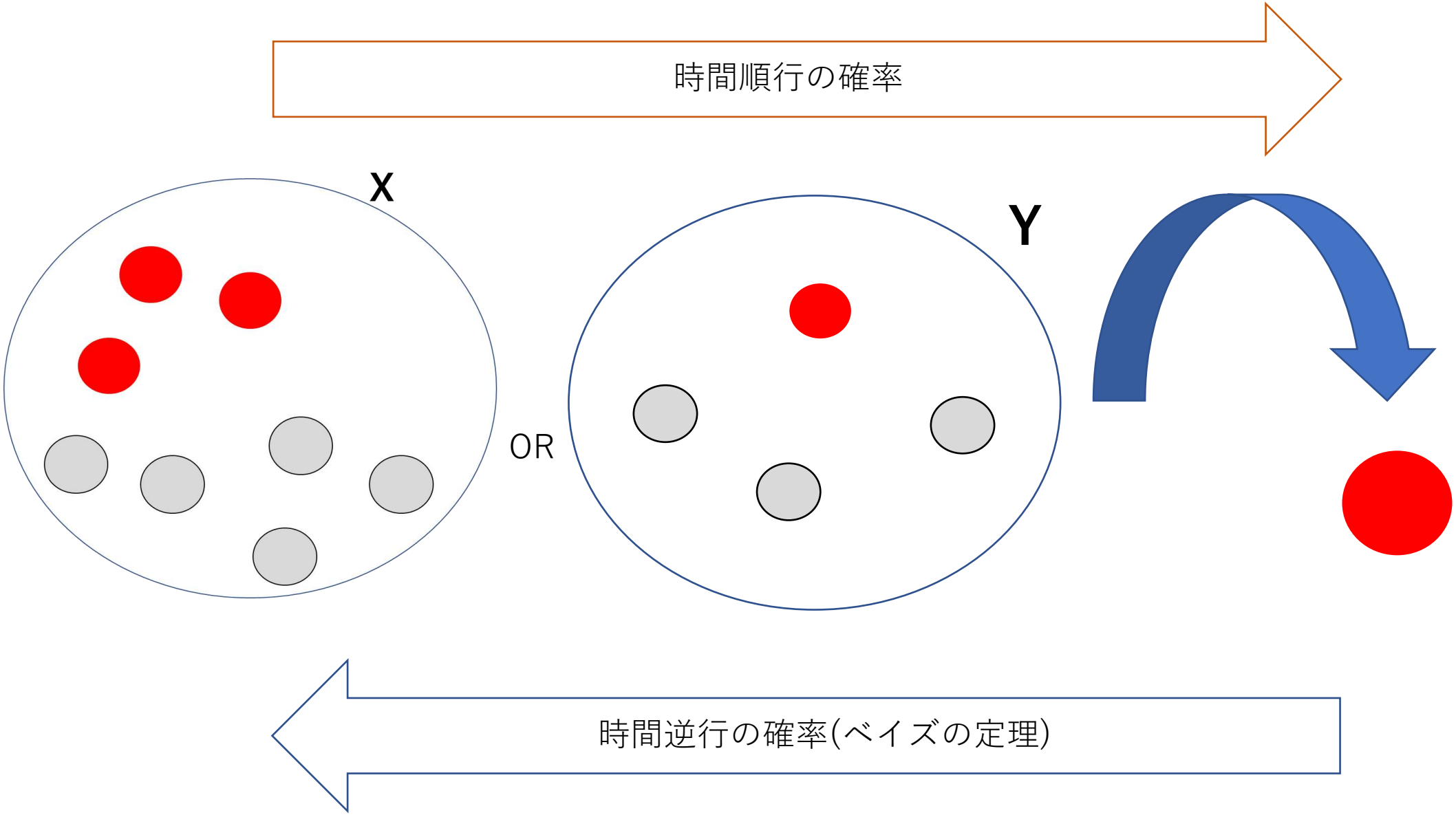
P(赤|X)=3/8　(Xを選んだ時赤玉を取る確率)

$$P(X|赤玉)=\frac{P(X)*P(赤玉|X)}{P(赤)}$$

=3/5

時間の流れ通りの情報から

時間の流れに逆らう分析ができる


時間順行の確率
X
OR
Y
時間逆行の確率(ベイズの定理)

ベイズ統計の用途

- 結果から原因を推測
- データから母集団のパラメーターを予測
- 迷惑メールのフィルタリング
  - ユーザーに送られてきた迷惑メールからどんなメールが迷惑か推測
- 少ないデータからそれっぽい分析をする

最尤推定とベイズ統計学

- 最尤推定（さいゆうすいてい）
- 与えられた観測値から、尤度を最大にするような母数や理論値を推定する方法

→最もそれっぽい値を予測する方法（みなさんも使ってます）

例：n回投げてk回表が出たコインの表が出る確率P0は

P0=k/n　と予測できる

Q. n回投げてk回表が出たコインの表が出る確率 p_0 を予測せよ。

コインの表が出る確率は、

$${}_nC_k\, p^k (1-p)^{n-k}$$ と表せる。

pの値によって確率は変わる
→これが最大になるpを求める

$$L(p) = {}_nC_k \cdot p^k \cdot (1-p)^{n-k}$$

↑これを微分して最大になるpを求める

$L(p)$ を最大にするpも $\log L(p)$ を最大にするpも同じなので、対数をとって

$$\log L(p) = \log {}_nC_k \cdot p^k \cdot (1-p)^{n-k}$$

$$= \log {}_nC_k + k\log p + (n-k)\log(1-p)$$

$$\frac{d}{dp}\log L(p) = \frac{k}{p} - \frac{(n-k)}{(1-p)}$$

これが0になれば良いので

$$0 = \frac{k}{p} - \frac{(n-k)}{(1-p)}$$

$$p = \frac{k}{n}$$

おまけ

$(n-k)\log(1-p)$ の微分、

$\log(1-p)$ を合成関数だと考えて、

$y = \log x$ $\quad x = 1-p$ をそれぞれ微分して、

$$\frac{dy}{dx} = y' = \frac{1}{x}$$ xを代入して、

$$= \frac{1}{1-p} \cdots ①$$

$$\frac{dx}{dp} = x' = -1 \cdots ②$$

①、②より

$$\frac{dy}{dx}\cdot\frac{dx}{dp} = \frac{dy}{dp} = \frac{1}{1-p} \times -1$$

$$= -\frac{1}{(1-p)}$$

最尤推定の限界

例題：AとBの二つのコインを用意する。Aは3回投げて2回表、
Bは100回投げて60回表が出た。どちらのコインのほうが表が出やすいか答えよ。

最尤推定を用いて予測すると…
A:P(A)=2/3
　　=66.7%
B:P(B)=60/100
　　=60%
答え:Aのほうが表が出る確率は高い。

でも本当にそうだろうか…

ここでベイズ統計の力

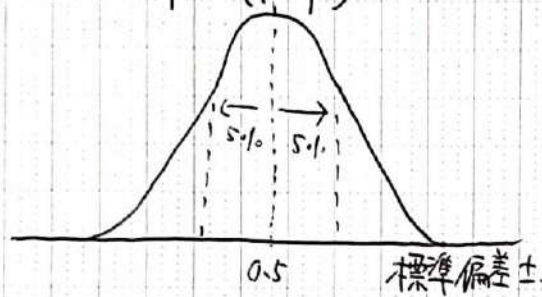
- ベイズ統計では事前分布を使って確率を求める
- 事前分布とは、経験や一般常識、今まで集めたデータなどを確率分布の形であらわしたもの。すでに経験的に知っていることをデータ分析に盛り込むことができる。
  - 事前分布の求め方は今後やります。

尤度関数　　事前分布

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

・今回つかう事前分布

$$C\,p^{50}(1-p)^{50}$$



100回投げたら、だいたい50%表が出るだろう

コインA

事前分布 $C\,p^{50}\cdot(1-p)^{50}$ を仮定

$$p^{post}(p) = C\times p^{50}\times(1-p)^{50}\times D\cdot p^{2}\times(1-p)^{1}$$

$$= E\times p^{52}\times(1-p)^{51}$$

$$\Big\downarrow$$

$$P_A = \frac{52}{52+51} = \frac{52}{103} = 50.49\%$$

コインBも同様に.

$$p^{post}(p) = C\,p^{50}(1-p)^{50}\times D\cdot p^{60}\times(1-p)^{40}$$

$$= E\times p^{110}\times(1-p)^{90}$$

$$\Big\downarrow$$

$$P_B = \frac{110}{110+90} = \frac{110}{200} = 55\%$$

ベイズ統計の力をつかうと、、

コインAが表になる確率 50.49%

コインB ＜ 55 %

よりそれっぽい分析ができる!!

- 今日は例を示すため手計算でやりましたが、大体の計算はコンピューターがしてくれます
- コンピューターがどんな計算をしているのかを理解してベイズ統計学を使いこなしましょう
- 計算が理解できなくても、コンピューターに計算させるうちに理解できてくるようです
- **この講座が終わるころには皆さんを立派なベイジアンにします**