Stataで学ぶベイズ統計

LightStone Corp

2020年1月

修正2020年2月

第2回 事前分布の選択と更新

- 今回の内容はベイズ統計の基礎知識に関すること
  - 再び「ベイズの定理」
  - 原因の確率
  - 主観確率の役割
  - 確率の更新(ベイズ更新)
  - 多数の原因
  - 事後分布
  - 事前分布

ベイズの定理

- 話を簡単にするため、最初は「確率」を使って話を進める
- 最初にベイズの定理について、もう一度考えてみる

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

- 事象A: 初期不良の原因
- 事象B:スマートフォンの初期不良の発生
- ここでは初期不良(B)の原因(A)は3つ(液晶画面、通信機能、充電池)しかないものとする

原因の確率

- 3つの不具合AはBの分割事象と考えることができる

$$P(B) = \sum_{i=1}^{3} P(A_i \cap B) = \sum_{i=1}^{3} P(A_i)P(B|A_i)$$

- よって、スマホの初期不良が発生したとき、原因がAiである確率は次のように表現できる

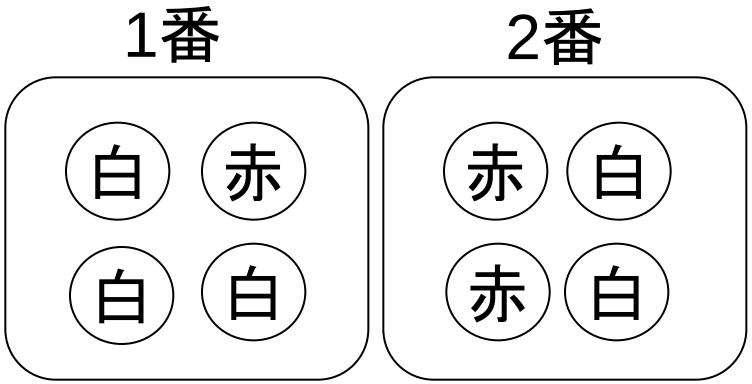
$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{3} P(A_i)P(B|A_i)}$$

- 一般的に書くと、原因の確率(ベイズの定理)は

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum P(A_i)P(B|A_i)}$$

主観確率

- 2つの箱の中に4つのボールが入っているものとする



- 事象A:1番または2番の箱の一方を選ぶ
- 事象B:白いボールを引く

主観確率

- 次のような設定を考える

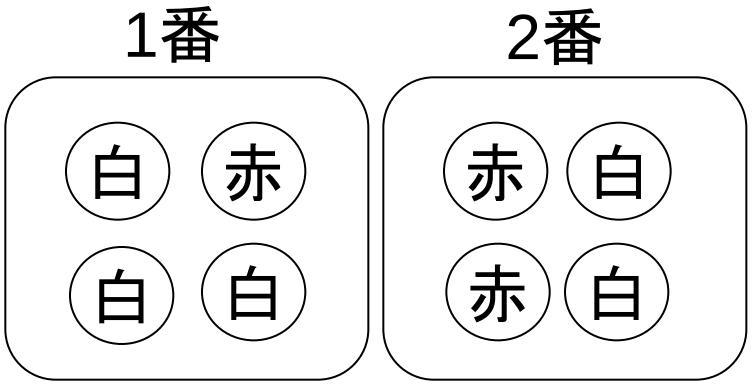
$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum P(A_i)P(B|A_i)}$$

- 目的:白いボールを引いた時(B)、選択した箱(A)が1番である確率と2番である確率を求める
- 最初にどちらの箱を選ぶのか、まったく予想はできないのでP(A)の事前確率は0.5とする

$$P(A_1) = P(A_2) = 0.5$$

条件付き確率

- それぞれの箱で白を引く確率は



$$P(B|A_1) = 3/4, \quad P(B|A_2) = 2/4$$

- したがって、

$$P(A_1|B) = \frac{P(A_1)\cdot P(B|A_1)}{P(A_1)\cdot P(B|A_1)+P(A_2)\cdot P(B|A_2)} = 0.6$$

$$P(A_2|B) = \frac{P(A_2)\cdot P(B|A_2)}{P(A_1)\cdot P(B|A_1)+P(A_2)\cdot P(B|A_2)} = 0.4$$

事後確率

- ベイズの定理による計算の結果、箱の選択確率が変化した

試行前の事前確率

$$P(A_1) = 0.5 \rightarrow 0.6$$
$$P(A_2) = 0.5 \rightarrow 0.4$$

1回目の試行により求めた事後確率

- 箱にボールを戻し、再度、選択を行ったら、やはり白いボールだった。それぞれの箱を選択する確率はどのように変化するか?

事後確率

- 1回目の結果(情報:データ)を使って、再び選択確率を計算する

$$P(A_1|B) = \frac{P(A_1)\cdot P(B|A_1)}{P(A_1)\cdot P(B|A_1) + P(A_2)\cdot P(B|A_2)}$$

$$= \frac{0.6\times(3/4)}{0.6\times(3/4) + 0.4\times(2/4)} = 0.69$$

$$P(A_2|B) = \frac{P(A_2)\cdot P(B|A_2)}{P(A_1)\cdot P(B|A_1) + P(A_2)\cdot P(B|A_2)}$$

$$= \frac{0.4\times(2/4)}{0.6\times(3/4) + 0.4\times(2/4)} = 0.31$$

確率の変化

- 2回の試行の結果、事前確率は次のように変化した

$$P(A_1) = 0.5 \rightarrow 0.6 \rightarrow 0.69$$
$$P(A_2) = 0.5 \rightarrow 0.4 \rightarrow 0.31$$

練習

- ベイズの定理の考え方を整理するために次の問題を解いてみよう
- 3つの箱があって、黒と白のカードがそれぞれ異なる割合で入っている
  - 箱1(黒5枚、白2枚)
  - 箱2(黒1枚、白1枚)
  - 箱3(黒2枚、白3枚)
- カードの1枚引いたとき、黒だった。どの箱を選択した可能性が高いか?ベイズの定理を使って箱ごとの事後確率$P(A_i|B)$を求めてみよう

練習

- 事象A:箱の選択
- 事象B:黒いカードを引く事
- 3つの箱からどれを選択するか、特別な理由はないものとして

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$$

- 各箱で黒いカードを選択する確率は

$$P(B|A_1) = 5/7, \quad P(B|A_2) = 1/2, \quad P(B|A_3) = 2/5$$

練習

- 1枚引いたカードが黒(B)である時、それが箱1、2、または、3である確率は

$$P(A_1|B) = \frac{P(A_1)\cdot P(B|A_1)}{P(A_1)\cdot P(B|A_1)+P(A_2)\cdot P(B|A_2)+P(A_3)\cdot P(B|A_3)} = 44.2\%$$

$$P(A_2|B) = \frac{P(A_2)\cdot P(B|A_2)}{P(A_1)\cdot P(B|A_1)+P(A_2)\cdot P(B|A_2)+P(A_3)\cdot P(B|A_3)} = 31.0\%$$

$$P(A_3|B) = \frac{P(A_3)\cdot P(B|A_3)}{P(A_1)\cdot P(B|A_1)+P(A_2)\cdot P(B|A_2)+P(A_3)\cdot P(B|A_3)} = 24.8\%$$

練習

- 1回目に引いたカードを元に戻して、2回目のカードを引いた所、また、黒いカードだった。事後確率を先と同じ要領で計算すると、次のように変化することが分かる

$$
\begin{aligned}
&\text{箱1: } 33.3\% \rightarrow 44.2\% \rightarrow 55.5\% \\
&\text{箱2: } 33.3\% \rightarrow 31.0\% \rightarrow 27.2\% \\
&\text{箱3: } 33.3\% \rightarrow 24.8\% \rightarrow 17.3\%
\end{aligned}
$$

ベイズの定理

$$P(A_i|B) = \frac{P(A_i)\cdot P(B|A_i)}{\sum P(A_i)\cdot P(B|A_i)}$$

- ここで原因Aをθ、結果BをZ、さらに事前確率P(Ai)をw(θi)、事後確率P(Ai|B)をw'(θ|z)と表記すると、

$$w'(\theta_i|z) = \frac{w(\theta_i)p(z|\theta_i)}{\sum w(\theta_i)p(z|\theta_i)}$$

ベイズの定理

- 先のカードの例で言えば、zはカードの色(z_1:黒、z_2:白)、θは箱の選択
- 箱の選択確率は

$$w(\theta_1) = w(\theta_2) = w(\theta_3) = 1/3$$

- 尤度は

$$p(z_1|\theta_1) = 5/7,\ p(z_1|\theta_2) = 1/2,\ p(z_1|\theta_3) = 2/5,$$

$$p(z_2|\theta_1) = 2/7,\ p(z_2|\theta_2) = 1/2,\ p(z_2|\theta_3) = 3/5$$

- 事後確率は

$$w'(\theta_1|z_1) = 44.2\%,\ w'(\theta_2|z_1) = 31.0\%,\ w'(\theta_3|z_1) = 24.8\%$$

ベイズの定理

$$w'(\theta_i|z) = \frac{w(\theta_i)p(z|\theta_i)}{\sum w(\theta_i)p(z|\theta_i)}$$

- ここでθ(箱)が連続的に選択できるとしたら、

$$w'(\theta|z) = \frac{w(\theta)p(z|\theta)}{\int w(\theta)p(z|\theta)d\theta}$$

- w:事前分布、w':事後分布と呼ぶ
- θ;母集団の母数(パラメータ)、zは標本、p(z|θ)は尤度
- 尤度はf(z|θ)と表現することも多い

ベイズの定理

$$w'(\theta|z) = \frac{w(\theta)p(z|\theta)}{\int w(\theta)p(z|\theta)d\theta}$$

- 分母はθについて積分するので、一定の数になる
- 比例の表現を利用すれば、事後分布は

$$w'(\theta|z) \propto w(\theta)p(z|\theta)$$

事後に有する情報 = 事前に有する情報 × 標本の情報 × 定数

事前分布

- ベイズ統計学は事前分布を選択することから始まる
- 選択した事前分布から比較的簡単に事後分布を求めることのできる「共役事前分布」から話を始める
- 運動会における白組の勝敗について考えてみる
- 白組が勝ったらx=1(確率θ)、負けたらx=0

$$x = \begin{cases} 1, & \theta \\ 0, & 1-\theta \end{cases}$$

- 過去の情報を考慮しなければ、勝つ確率は1/2
- ここでは過去のn回の結果がzとして与えられているとする

$$z = (x_1, x_2, \ldots, x_n)$$

事前分布

- データを利用した尤度は

$$p(z|\theta) = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

- 尤度の形に合わせて事前分布w(θ)を、母数(α, β)のベータ分布Beta (α, β)とする

$$w(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

MEMO

- ベータ分布の比例定数はベータ関数の逆数

$$\frac{1}{\mathrm{B}(\alpha,\beta)} \longleftarrow \text{比例定数}$$

$$\mathrm{B}(\alpha,\beta) = \int_0^1 u^{\alpha}(1-u)^{\beta-1}du$$

事前分布

$$w'(\theta|z) \propto w(\theta)p(z|\theta)$$

- 上式の右辺を計算すると、

$$\begin{aligned} w(\theta)p(z|\theta) &= \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} \\ &= \theta^{\alpha+\Sigma x_i-1}(1-\theta)^{\beta+n-\Sigma x_i} \end{aligned}$$

ここで、$\underline{\alpha' = \alpha + \sum x_i}$, $\underline{\beta' = \beta + n - \sum x_i}$ とすれば

事後分布はパラメータを更新したベータ分布になる

$$w'(\theta) \propto \theta^{\alpha'-1}(1-\theta)^{\beta'-1}$$

パラメータの更新

- 第1回の資料(スライド9)の計算は次の要領で計算

$$\alpha = 2, \beta = 20, \sum x_i = 0, n = 20$$

$$\theta|y \sim \mathrm{Beta}(2+0, 20+20-0) = \mathrm{Beta}(2, 40)$$

- ベータ分布の期待値は2つのパラメータで決まり、

$$E(\theta) = \frac{\alpha}{\alpha+\beta}$$

- よって, 感染率の期待値は

$$E(\theta) = \frac{2}{2+40} \simeq 4.8\%$$

勝率の変化

- 運動会の例…白組の勝率を予測は毎年、どのように更新されるか?同じように二項分布とベータ分布を利用する
- 第1回目
  - 事前の情報は何もなかった
  - ベータ分布でα=β=1の時、一様分布になる
  - そして第1回目は白組が勝った
  - 事後確率はどのように変化するか?

$$n = 1, x_1 = 1, \sum x_i = 1$$

$$\alpha' = 1 + 1 = 2$$

$$\beta' = 1 + 1 - 1 = 1$$

$$E(\theta) = \frac{2}{2+1} = 66.7\%$$

勝率の変化

- 第2回目も白が勝って2戦2勝となった。これで、次に白組が勝つ確率は、つぎのように更新される

$$n = 1, \sum x_i = 1$$

$$\alpha' = 2 + 1 = 3$$

$$\beta' = 1 + 1 - 1 = 1$$

$$E(\theta) = \frac{3}{3+1} = 75.0\%$$

- 第3回目も白が勝った。同じ要領で更新すると

$$n = 1, \sum x_i = 1$$

$$\alpha' = 3 + 1 = 4$$

$$\beta' = 1 + 1 - 1 = 1$$

$$E(\theta) = \frac{4}{4+1} = 80.0\%$$

ベイズ統計による来年の運動会の勝率は80%.頻度だと100%

共役事前分布

- 特定の事前分布(共役事前分布:conjugate prior distribution)と尤度の組合せにより簡単に事後分布のパラメータが分かる

- 代表的な共役事前分布

尤度	事前分布	事後分布
二項分布	ベータ分布	ベータ分布
正規分布	正規分布	正規分布
ポワソン分布	ガンマ分布	ガンマ分布

ポアソン分布とガンマ分布

- 尤度としてポアソン分布、事前分布としてガンマ分布を利用する例について考える
- ポアソン分布

$$p(X = x) = \frac{e^{-\theta}\theta^{x}}{x!}$$

- パラメータθの下で、あるイベントがx回発生する確率を意味する
- パラメータθは発生回数の平均を示す。例えば、一日に発生する交通事故の回数を標本とすれば、観察期間中の平均回数
- 集めた情報(標本)を利用して尤度を計算する際に利用する

尤度

- n日に観察したデータセットをzとすると

$$z = (x_1, x_2, \ldots, x_n)$$

- ポアソン分布を用いた尤度は

$$p(z|\theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \propto e^{-n\theta} \theta^{\Sigma x_i}$$

- θは平均発生回数を示すパラメータ

ガンマ分布

- ポアソン分布のパラメータθの分布をガンマ分布と考える
- ガンマ分布の表現方法(パラメータの表記方法)にはいくつかの方法がある
- ここではベイズ定理での利用に適している次の表現方法を利用する

$$w(x) = \frac{\alpha^{\lambda}}{\Gamma(\lambda)} e^{-\alpha x} x^{\lambda-1}$$

- λ:shapeパラメータ(正値)。この値が1の時, 指数分布の密度関数となる
- α:rateパラメータ(正値)
- x:ゼロ以上の値をとる連続変数。ポアソン分布のxとは無関係

ベイズの定理

- ポワソン分布のパラメータがガンマ分布に従うとすれば、

$$w(\theta) = \frac{\alpha^\lambda}{\Gamma(\lambda)} e^{-\alpha\theta}\theta^{\lambda-1} \propto e^{-\alpha\theta}\theta^{\lambda-1}$$

- よってベイズの定理から

$$\begin{aligned} w'(\theta) &\propto w(\theta) \cdot p(z|\theta) \\ &\propto e^{-\alpha\theta}\theta^{\lambda-1} \cdot e^{-n\theta}\theta^{\Sigma x_i} \\ &= e^{-(n+\alpha)\theta}\theta^{(\Sigma x_i+\lambda)-1} \\ &= e^{-\alpha'\theta}\theta^{\lambda'-1} \end{aligned}$$

$$\alpha' = n + \alpha, \quad \lambda' = \Sigma x_i + \lambda$$

練習

- ガンマ分布の期待値と分散の公式

$$E(\theta) = \frac{\lambda}{\alpha}, \quad V(\theta) = \frac{\lambda}{\alpha^2}$$

- 過去5年間の冬季に熊が町内で観測された回数は(6,5,7,3,7)の28回である。
- 今年の冬季は8回観測された。ガンマ分布のパラメータはどのように変化したか、考えてみよう。

練習

- 最初に期待値と分散の公式と、過去5年間の情報から

$$\lambda = 5.6, \alpha = 1$$

- 今シーズンの情報を反映させると

$$\lambda' = 8 + 5.6 = 13.6$$
$$\alpha' = 1 + 1 = 2$$

- 頻度で考えると、過去6年間の平均は6回ですが、新しいガンマ分布のパラメータを利用すると、予想される回数は6.8回になります

共役事前分布

- 尤度と共役事前分布から簡単に事後分布のパラメータを計算できる
- 常に便利な共役事前分布がある訳ではない
- 共役事前分布が利用できない場合は、やや複雑な計算を行う必要がある(第3回で解説)

事前分布の選択

- ベイズ分析の難しさは事前分布(研究者の経験や知識に依存)と実際に観測したデータ(尤度)のバランスを取るところにある
- 事前分布の影響がデータの示す情報を、掻き消してしまうほど大きいような状態は避ける
- 逆に、ビッグデータと呼ばれるような情報が入手可能な状況では、事後分布は事前分布と独立となる
- この時、ベイズ分析の結果と、頻度による統計分析の結果はほぼ同じものになる(Bernstein-von Misesの定理)
- 事前分布として幾つか異なるものを利用して、事後分布のパラメータへ変化を考察する(感度分析)

第2回のまとめ

- 事前分布の役割
- 共役事前分布
- 観測したデータを使って事前分布のパラメータを更新する仕組み

参考文献

- 松原 望「入門　ベイズ統計」 東京図書

研究者向けの統計解析ソフトウェア Stata® 評価版のお申込み

機能制限なし！30日間無料で使える評価版でまずはお試しください！ https://www.lightstone.co.jp/stata/evaluate.html

※学生の方への評価版の提供はございません。大学の講義で使用するなど単位取得に関わる形でのご利用となる場合は学生版のご購入をご検討ください。

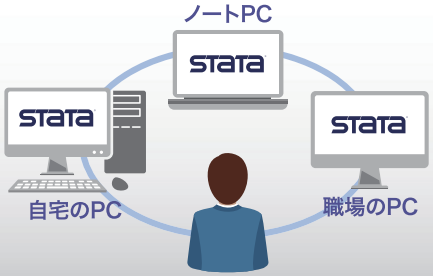
ライセンスはサブスクリプションがおすすめ！

サブスクリプションライセンスを選ぶメリット

- 常にStataの最新バージョンが利用できます
- 技術サポートの対象となります
- 毎年の経費としてStataを導入できます
- 利用者の増減に柔軟に対応可能
- 初期導入費を抑えたい場合にもおすすめ

シングルライセンス（サブスクリプション）

職場のPC・ノートPC・自宅のPC等、個人所有のPCにインストール可能です。
同時に使用できるのは1台のみとなります。



マルチユーザライセンス（サブスクリプション） 2ユーザ以上

ボリュームライセンス

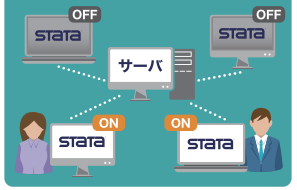
シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。

同時起動ライセンス

ご利用環境により下記3通りの運用方法からご選択いただけます。

PC毎にスタンドアロンインストールをして運用
例:2ライセンスの場合

クライアント＆サーバ環境によるネットワーク運用
例:2ライセンスの場合



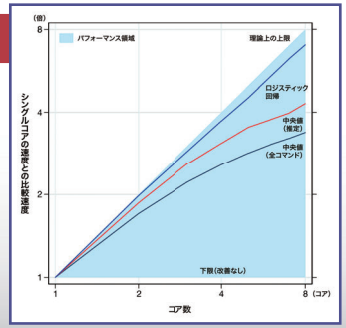
1台の仮想マシンにユーザがアクセスする運用
例:2ライセンスの場合

Stata／MPの演算能力

データ処理時間の削減ならMP

Stata MPはPCの持つマルチコアの特性を活かして、処理を分散・並列する機能を備えます。およそ85%以上のコマンドで処理速度が向上し、コア数に応じた計算時間の短縮が期待できます。
コマンドごとの処理速度の向上の度合いについては以下のページや資料をご覧ください。

コア数	全てのコマンド	推定コマンド	ロジスティック回帰
2	1.7倍	1.8倍	1.9倍
4	2.6倍	3.1倍	3.8倍
8	3.3倍	4.2倍	6.8倍



弊社Webページ内の『Stata／MP』
https://www.lightstone.co.jp/stata/statamp.html

株式会社 ライトストーン
〒101-0031
東京都千代田区東神田 2-5-12 龍角散ビル 7F
TEL：03-3864-5211
FAX：03-3865-0050
e-mail： sales@lightstone.co.jp
WEB： https://www.lightstone.co.jp/