Stataで学ぶベイズ統計

LightStone Corp

2019年12月
修正 2020年5月

Stataで学ぶベイズ統計

5回に分けてベイズ統計の基礎知識と, 関連するStataの操作方法を解説する

構 成

第1回 ベイズ統計とは
第2回 事前分布の選択と更新
第3回 マルコフチェーンモンテカルロ法(MCMC)
第4回 ベイズ統計によるモデル推定
第5回 ベイズプリフィックス

第1回ベイズ統計とは

- Thomas Bayes(1702-1761, 英国)
  - 長老派の牧師
  - 1742年　the Royal Society（王立協会）のフェローに選出
  - 数学的な研究論文は死後に発表された2本のみ
    - “An Essay Towards Solving a Problem in the Doctrine of Chances”
  - ベイズ統計は不確実性のあるイベントの，発生確率の計算に主観を組み込むことができる

2つの統計学

- ベイズ統計:
  - モデルのパラメータは何らかの確率分布に従うことを仮定する
  - この確率分布として既知の情報(事前分布)を利用する
  - 求めるものは事後分布としての「××分布」
- 頻度を用いた統計分析:
  - 頻度を用いる一般的な統計分析の手法ではパラメータの母数(点)を推定する
  - パラメータは正規分布(t分布)に従うと考える
  - 求めるものは「点推定値」と「標準誤差」

分析例(Stataマニュアルから)

- 小都市における感染症の感染率を推定する(Hoff 2009,3)
- 問題の感染症は極めてまれに観察される
- サンプルサイズは20人
- 目的の感染率はその都市における患者の割合でθとする

$$\theta \in [0, 1]$$

- サンプル中の感染者数をyとする

二項分布

- 感染者の人数は二項分布に従うと考えられる

$$y|\theta \sim \mathrm{Bin}(20, \theta)$$

- 他の都市における調査例から感染率は0.05から0.20程度であることが報告されている
- したがって, 平均的な感染率は0.10と考える(主観)
- この0.10を期待値として, ベイズ推定で利用することを考える
- 確率の期待値が0.10となるような確率分布は?

ベータ分布

- 区間[0,1]上の分布：Beta(α, β)
- ベータ分布の密度関数

$$f(x|\alpha,\beta) = B(\alpha,\beta)^{-1}x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$$

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

- 仮にBeta(2,20)という分布を考えると, 期待値は

$$E(X) = \frac{2}{2+20} = 0.09$$

事前分布

- 感染率θの事前分布としてBeta(2,20)を選択する
- 20人を対象にした調査の結果, 感染者の数は0であった.つまり, y=0.
- サンプルサイズが小さく, 稀に発生する感染症なので, 0であることは特に不合理なことではない
- 仮に真の感染率が0.05であるとき, 20人の調査結果がゼロになる確率は, 二項分布から

$$ {}_{20}C_0 \cdot 0.05^0 \cdot 0.95^{20} \simeq 36\% $$

ベイズ統計のモデル

- アウトカムとパラメータについて

$$y|\theta \sim \mathrm{Bin}(20, \theta)$$
$$\theta \sim \mathrm{Beta}(2, 20)$$

- 20人の調査結果で感染者がゼロというデータを使って事後分布を計算すると,

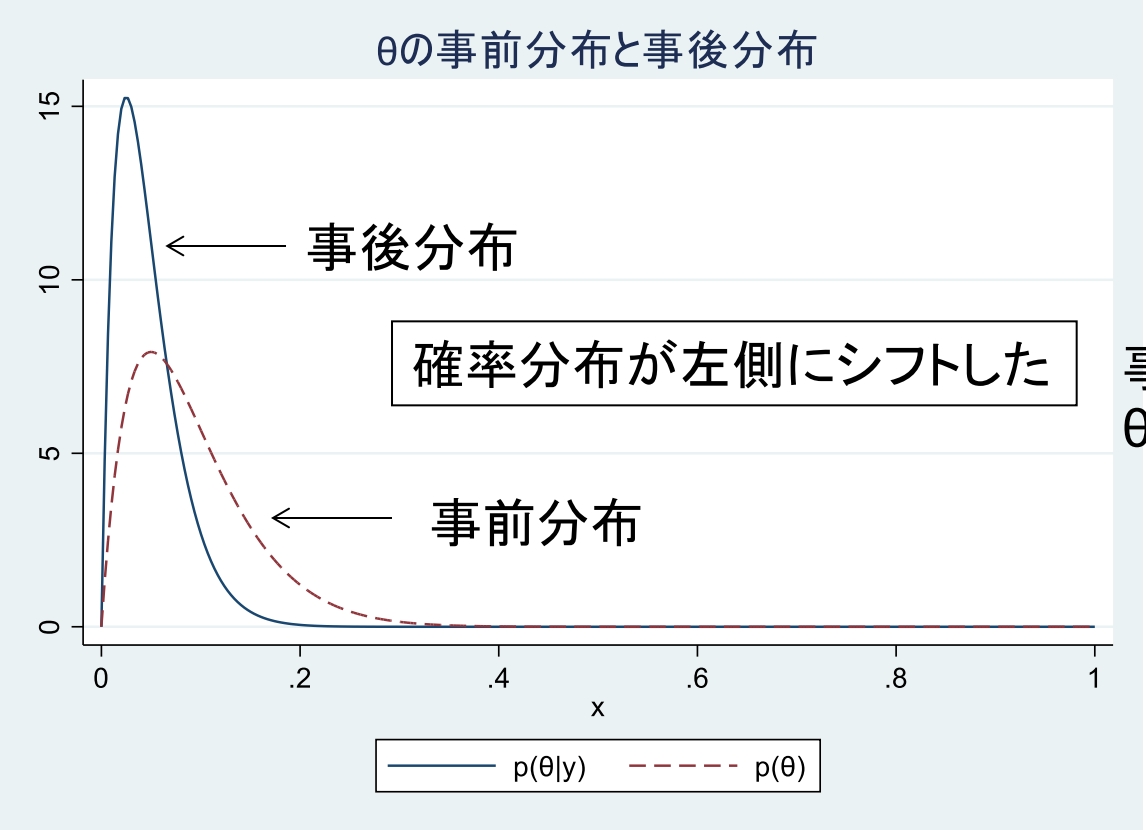
$$\theta|y \sim \mathrm{Beta}(\underline{2+0, 20+20-0}) = \mathrm{Beta}(2, 40)$$

*カッコ内がこのようになる理由は第二回で説明する

事前分布と事後分布

*ベータ分布のプロット

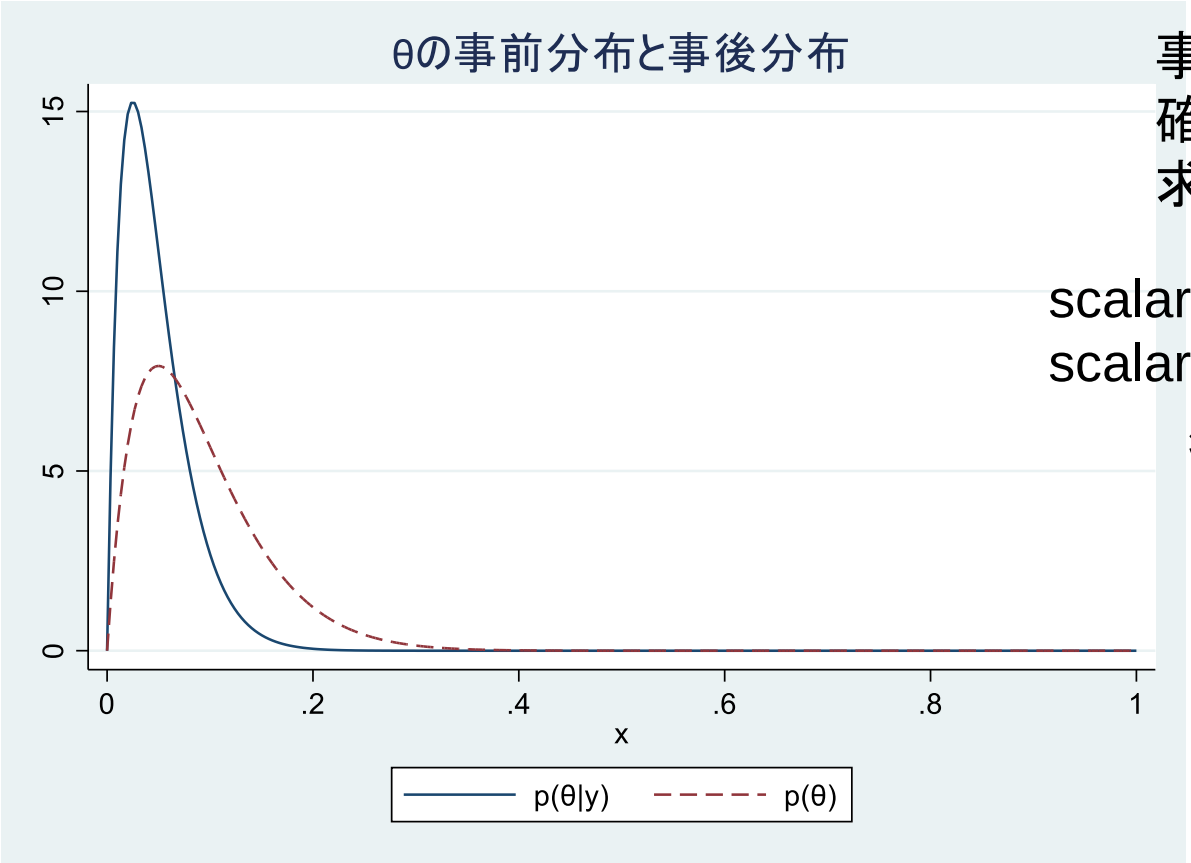
```
twoway function y=betaden(2,40,x), range(0 1)||function
                                y=betaden(2,20,x), range(0 1)
```



事前分布はp(θ)

事後分布の平均は
θ=2/(2+40)=0.048

事後分布



事後分布でθ<0.1となる確率は,次のコマンドで求めることができる

```
scalar test= ibeta(2,40,0.1)
scalar list
```

答えは約93%

ベイズ統計による感染率の期待値は4.8%

頻度で考える

- 母比率θを推定する方法で感染率を考える
- 標本平均から

$$\bar{y} = \frac{y}{n}$$

- 二項分布の期待値と分散の公式は,

$$E(X) = np$$
$$V(X) = np(1-p)$$

- pとして $\bar{y}$ を利用すると,

$$V(\bar{y}) = V(\frac{y}{n}) = \frac{n\bar{y}(1-\bar{y})}{n^2} = \frac{\bar{y}(1-\bar{y})}{n}$$

分析例の結論

- $\bar{y}$ の95%CIは

$$\left(\bar{y} - 1.96\sqrt{\bar{y}(1-\bar{y})/n},\quad \bar{y} + 1.96\sqrt{\bar{y}(1-\bar{y})/n}\right)$$

- 20人を調査した結果, 感染者がゼロの場合, 感染率の95%CIはゼロになってしまう
- 予測される感染者の数はゼロとしてよいのか?
- 頻度による統計学で考えると, 感染率はゼロになってしまう
- ベイズ統計では4.8%

事前分布の利用

- ここでは例として, 既知の情報を活かすために事前分布としてベータ分布を利用した
- 実際の分析では, 利用可能ないくつかの確率分布を事前分布に採用し, 事後分布の感度分析を行うことが望ましい

ベイズ統計学の基礎

- 次にベイズ統計学の基礎知識を確認する
- AとBを確率変数とする時,
  - 条件付き確率

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

  - ベイズの定理

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)} \qquad (1)$$

尤度

- 統計モデルの尤度

$$L(\theta; \mathrm{y}) = f(\mathrm{y}; \theta) = \Pi_{i=1}^{n} f(y_i; \theta)$$

- ベイズ統計学の目的:データが与えられた時のθの特徴を知ること

事前分布

- 事前分布を特定のπとする

$$p(\theta) = \pi(\theta)$$

- yとθは両方とも確率変数なのでベイズの定理(1式)に代入すると

$$p(\theta|\mathrm{y}) = \frac{p(\mathrm{y}|\theta)p(\theta)}{p(\mathrm{y})} = \frac{f(\mathrm{y};\theta)\pi(\theta)}{m(\mathrm{y})} \qquad (2)$$

データの分布については $m(\mathrm{y}) \equiv p(\mathrm{y})$

- yの周辺分布は定義から,

$$m(\mathrm{y}) = \int f(\mathrm{y},\theta)d\theta = \int f(\mathrm{y};\theta)\pi(\theta)d\theta \qquad (3)$$

事後分布

$$m(\mathrm{y}) = \int f(\mathrm{y},\theta)d\theta = \int f(\mathrm{y};\theta)\pi(\theta)d\theta \quad (3)$$

- 3式のθは積分計算されてしまうので, 結局, m(y)はθには依存しない一定の数である事が分かる
- 2式でm(y)はシータに無関係なので取り去ってしまうと,

$$p(\theta|y) \propto L(\theta;y)\pi(\theta) \quad (4)$$

- モデルパラメータの事後分布は尤度と事前分布の積に比例することを示している
- (4)式の対数をとって

$$\ln\{p(\theta|y)\} = l(\theta;y) + \ln\{\pi(\theta)\} - c \quad (5)$$

ただし, $c = \ln\{m(\mathrm{y})\}$

確率分布の確認

- 今回利用した確率分布はベータ分布と二項分布
- ベータ分布(確率変数Xは0から1の間に存在する)について確認する

$$f(x|\alpha,\beta) = B(\alpha,\beta)^{-1} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1$$

ベータ関数

$$0 < \alpha, \beta < \infty$$

$$B(\alpha,\beta) = \int_0^1 x^{\alpha}(1-x)^{\beta-1} dx$$

- 2つのパラメータが同時にゼロの時, 一様分布に等しい
- 密度関数の形状は次のコマンドで確認できる

twoway function y=betaden(a,b,x),

第1回のまとめ

- ベイズ統計の概略
- ベイズ統計学の基礎
- Stataによる密度関数のプロット
- ベータ分布と一様分布

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。

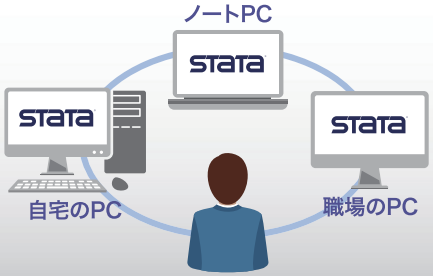
ライセンスはサブスクリプションがおすすめ！

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

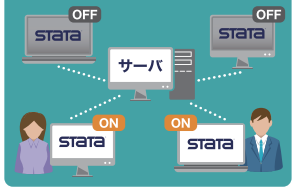
シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。

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



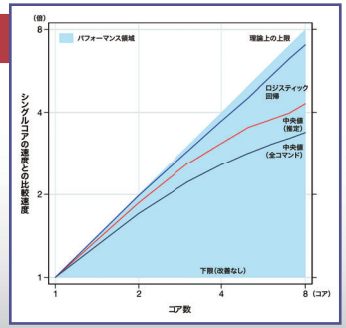
1台の仮想マシンにユーザがアクセスする運用
例:2ライセンスの場合

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/