Bayes Analysis Maniax

フリーソフトで始める ベイズ統計解析

後藤和智（後藤和智事務所 OffLine）

Bayes Analysis Maniax

フリーソフトで始める ベイズ統計解析

著：後藤和智（後藤和智事務所 OffLine）
発行：2017 年 8 月 13 日
（コミックマーケット 92）

0.1 まえがき

66 冊目の同人誌となります、後藤和智です。今回は、ベイズ統計学の解説書を書いてみました。

弊サークルは評論や東方 Project 分野でいろいろな統計学の解説書や統計学を用いた同人誌を書いてきましたが、ベイズ統計学を取り扱うことはあまりありませんでした。せいぜい 2013 年に出した『改定増補版 紅魔館の統計学なティータイム――市民のための統計学 Special2』（コミックマーケット 85）で少しだけ触れた程度ですが、統計学を取り扱っている以上、ベイズ統計学を避けて通ることはできないのではないかと思い、書いた次第であります。

ベイズ統計学は、意思決定によく用いられると言われます。ベイズ統計学では、事前に想定した分布を、情報を用いてどんどん更新していくものですが、それが近年ではいろいろな意思決定に応用されているようです。また、最近はとんと聞かれなくなりましたが、ビッグデータの活用にもベイズ統計学、ベイズ推計が使われているようです。

またベイズ統計学を用いると、データ間の因果関係を描くことも不可能ではありません。そしてデータの間の関係を調べるためには、ベイズ統計の知識はこれからどんどん必要になっていくでしょう。

ベイズ統計学を用いた解説書はここ数年でかなり増えてきており、統計学を学びたい人には必ず読んで欲しい本である『完全独習 統計学入門』（ダイヤモンド社、2006 年）の著者・小島寛之氏も、2015 年にはベイズ統計学の解説書を刊行されました（『完全独習 ベイズ統計学入門』ダイヤモンド社、2015 年）。ベイズ統計学の需要は、ビジネスでも学術でも高まってきています。

本書は、今までの統計学の解説書と同様に、自分の手で動かして理解することを目的にしています。そしてそれに最適なソフトが、フリーの統計ソフト「R」だと思っています。R は最近では、最早 SPSS に取って代わるような統計ソフトとして認知されており、ほとんど無料で手に入ることから、市民が分析したり学習したりするのにも最適です。R にはベイズ統計解析を学んだり行ったりするパッケージもいろいろありますので、是非とも学んでいただければと思います。

2017 年 7 月 仙台市内にて 後藤和智

目次

第 1 章

理論編：ベイズ統計学の基礎

1.1 はじめに

本書では、フリーの統計ソフト「R」を中心としたフリーソフトを用いたベイズ統計解析の手法について解説します。そもそも統計学の考え方として大きく 2 つわけて「頻度主義」と「ベイズ主義」と呼ばれるものがあり、一般的な統計学で使われる「頻度主義」とは、母集団のパラメータが決まっているものであり、それを標本を用いて推定するというものです。

これに対して、ベイズ主義に基づく統計学とは、母集団のパラメータも含め、あらゆるパラメータが確率的であるという考え方に立ちます。このような考え方は、頻度主義に基づく統計学に比べて、人間の思考法に近く、統計学の実用性をさらに高めるものと評価されています。

ベイズ主義に基づく統計学においては、母集団のパラメータの確率分布について考えます（**事前分布**）。そして最尤法などによる推定によりその事前分布の確からしさを調べて、より母集団の確率分布に近い分布に更新していくという方法をとることができます。頻度主義に基づく統計学は、母集団のパラメータが確率分布ではなく、あらかじめ決まっているものなのでこういう操作をすることはできません。

このような手法は、例えば迷惑メールの対策に役に立っています。迷惑メール対策では、特定の単語などを含むものを迷惑メールとして処理しますが、たまにそれが誤っている場合が、あります。ベイズ統計学に基づく判別では、その誤りを学習させることにより、より制度を高めることができます。メールソフトでは、例えば「Mozilla Thunderbird」や「Shuriken」（ジャストシステム）などがベイズ統計学に基づく迷惑メールフィルタ（ベイジアンフィルタ）を用いているようです。

また、ベイズ統計学に基づくと、回帰分析の制度を高めることができます。ベイズ主義に基づく回帰分析では、主観的な確率分布を分析に入れることができます。頻度主義に基づいた統計学にはできないような分析を、ベイズ主義の統計解析は実現してくれるのです。

1.2 ベイズの定理

そもそもベイズ統計学で前提とする**ベイズの定理**とは、次のようなものです（なおこの定理は頻度主義でもベイズ主義でも成り立つ）。

ベイズの定理

事象 A と B の起こる確率について、次の関係が成り立つ。

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1.1}$$

$P(A|B)$ のような表記法について、慣れない方もいるかもしれません。そもそも $P(A|B)$ というのは、事象 B が起こっている前提で事象 A が起こる確率のことを言います（場合によっては $P_B(A)$ のような表記をすることもあります）。条件付き確率は、事象 A と B が同時に起こる確率 $P(A \cap B)$ を用いて、次のように表されます。

条件付き確率

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.2}$$

例を示してみます。

例

赤い玉が 1 個、白い玉が 3 個入っている箱 A と、赤い玉が 2 個、白い玉が 2 個入っている箱 B があるとする。このとき、箱 A を選んだという前提で赤い玉を引く確率と、箱 B を選んだという前提で赤い玉を引く確率は次のように示される。

$$P(\text{赤}\,|A) = \frac{1}{1+3} = \frac{1}{4} \tag{1.3}$$

$$P(\text{赤}\,|B) = \frac{2}{2+2} = \frac{2}{2} \tag{1.4}$$

ベイズ統計学は、次のような問題の解決に役立ちます。

例題

赤い玉が 1 個、白い玉が 3 個入っている箱 A と、赤い玉が 2 個、白い玉が 2 個入っている箱 B があるとする。このとき、赤い玉を引いたとき、その玉が箱 A のものである確率を求めよ。ただし、どちらの箱を選ぶかは確率的に等しいものとする。

この例題で求める確率は $P(A|\text{赤})$ ということになります。また、文中の「どちらの箱を選ぶかは確率的に等しいものとする」というのが、ベイズ統計解析における事前確率ということになります。この確率を求めるには、先のベイズの定理を使って次のようなものを使います。

$$P(A|\text{赤}) = \frac{P(\text{赤}\,|A)P(A)}{P(\text{赤})} \tag{1.5}$$

それぞれの確率は次のように表されます。

$$P(\text{赤}\,|A) = \frac{1}{4} \qquad P(A) = \frac{1}{2} \qquad P(\text{赤}) = \frac{1+2}{4+4} = \frac{3}{8} \tag{1.6}$$

この数値を当てはめると、$P(A|\text{赤})$ は次のように導き出すことができます。

$$P(A|\text{赤}) = \frac{P(\text{赤}\,|A)P(A)}{P(\text{赤})} = \frac{1}{4} \times \frac{1}{2} \div \frac{3}{8} = \frac{1 \times 1 \times 8}{4 \times 2 \times 3} = \frac{1}{3} \tag{1.7}$$

従って、赤い玉を引いた場合、それが箱 A からのものである確率は $\frac{1}{3}$ ということになります。また、B からのものでる確率は $\frac{2}{3}$ です。このように求めた確率を事後確率といいます。

なお、ベイズの定理は、事象 B が互いに排反な事象 $B_1, B_2, \ldots, B_k$ で表されるとき、次のように拡張することができます。

ベイズの定理

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)} \tag{1.8}$$

$$\text{ただし、} \sum_{j=1}^{k} P(B_j) = P(B_1) + P(B_2) + \cdots + P(B_k) \tag{1.9}$$

1.3 事前確率の設定と理由不十分の原則

ところで、先ほどの例題における、二つの箱の内ひとつが選ばれる確率は等しいというのは、どういう根拠に基づいているのでしょうか？　実は、これは特に一定の根拠に基づいているわけではなく、問題のために適当に考えたものに過ぎません。

ベイズ統計においては、このように事前確率を理由や根拠が不十分のまま適当に設定するというものも許容されます(**理由不十分の原則**)。頻度主義に基づく統計学においては、どちらの箱が選ばれるかについてはあらかじめパラメータとして与えられますが、ベイズ主義の場合はこれも確率変数になるばかりではなく、主観的な予測を入れることもできます。このような考え方は、客観性を重視する頻度主義の考え方からは受け入れることは難しいかもしれませんが、主観を入れることにより、思考のプロセスを重視することができるのです。

例えば先ほどの例題の場合、玉を引く人が直前に「俺は A の箱から引いてやる！」とか言っていたのだとしたら、常識的に考えれば A が選ばれる確率は飛躍的に高まります。仮に A が選ばれる確率 $P(A)$ を 0.8 と考えたとすると ($P(B) = 0.2$)、赤い玉を引いた場合、その玉が A から来た確率は次のようになります。

$$P(A|\text{赤}) = \frac{1}{4} \times \frac{4}{5} \div \frac{3}{8} = \frac{1 \times 4 \times 8}{4 \times 5 \times 3} = \frac{8}{15} \tag{1.10}$$

と、A である確率は飛躍的に高くなります。このような、頻度主義から考えれば素っ頓狂な発想も、ベイズ統計解析では入れることができるのです。

1.4 確率の更新

ベイズ主義に基づく解析では、計算によって求められた事前分布を、次の計算の事後分布として使うことができます。これを**確率の更新**と言います。

例えば、前々節の例題の試行のあと、玉を箱に戻して、同じ箱からもう一度玉を引くと、また赤だったとします。こうなると、赤の出る確率の低い A である可能性は低くなります。そこで、先の例題で求めた、目の前の赤い玉が A からきたものである確率は $\frac{1}{3}$ であるというものを新たな事前分布として設定すると、赤い玉を引いたときにそれが箱 A から来たものである確率は次のようになります。

$$P(A|\text{赤}) = \frac{P(\text{赤}|A)P(A)}{P(\text{赤})} = \frac{1}{4} \times \frac{1}{3} \div \frac{3}{8} = \frac{1 \times 1 \times 8}{4 \times 3 \times 3} = \frac{2}{9} \tag{1.11}$$

この確率を別の観点から見てみましょう。赤を 2 回連続で引く確率を P(赤赤) とする場合、箱 A,B それぞれにおける確率は次に用になります。

$$P(\text{赤赤}\,|A) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16} \qquad P(\text{赤赤}\,|B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \qquad P(\text{赤赤}) = \frac{3}{8} \times \frac{3}{8} = \frac{9}{64} \tag{1.12}$$

これに、当初の事前分布である「A と B が選ばれる確率は等しい」を使うと、次のようになります。

$$P(A|\ \text{赤赤}) = \frac{P(\text{赤赤}\,|A)P(A)}{P(\text{赤赤})} = \frac{1}{16} \times \frac{1}{2} \div \frac{9}{64} = \frac{1 \times 1 \times 64}{16 \times 2 \times 9} = \frac{2}{9} \tag{1.13}$$

と、先ほどの値と一致することが分かります。ベイズ統計は、情報を取得することによってより正確さを増していく解析手法であると言うことができます。

第 2 章

R によるベイズ統計解析の基本

2.1 はじめに

ここでは、ベイズ統計学の考え方を、R の簡単な操作によって学んでいきます。前章でも示したとおり、ベイズ統計学は、最初に事前分布を設定し、それを観測された実態に応じて更新していくというものです。そのような考え方を R で追体験できる方法を、ここで実際に操作しながら学んでいきましょう。

2.2 離散的な事前分布を用いる方法

次のような問題で考えます。

例題

あるベルヌーイ試行（成功するか失敗するかの試行）において、その成功率の確度が次の割合であるとする。

成功率	0	0.25	0.5	0.75	1
成功率が上記のものである確度	$\frac{1}{15}$	$\frac{5}{15}$	$\frac{7}{15}$	$\frac{2}{15}$	$\frac{1}{15}$

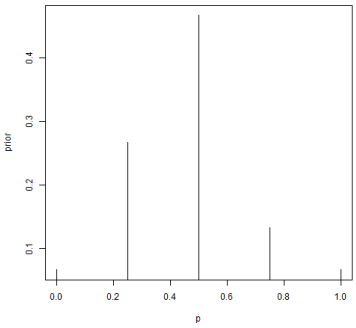
真の成功率を確かめるためこの試行を 30 回行ったところ、12 回成功して 18 回失敗した。このとき、先の確度の分布を事前分布として設定し、事後分布を求めよ。

まず、事前分布として用いる成功率の確度の分布を R に入力します。成功率を p に、確度の分布を prior に代入します。

```
> p <- c(0,0.25,0.5,0.75,1)
> prior <- c(1/15,4/15,7/15,2/15,1/15)
```

まず確度の分布を図で見るために、plot コマンドを使って中身を見てみましょう。

```
> plot(p,prior,type="h")
```



これを見ると、成功率がやや低い方に偏っているかもしれない試行であると考えることができます。さて、R におけるベイズ統計の学習用パッケージに、LearnBayes があります。この中に収録されているプログラムを使って、ベイズ統計を学んでみましょう。まずはパッケージを読み込ませます。

```
> library(LearnBayes)
```

次に、成功した回数と失敗した回数をベクトル形式で読み込ませ、これを事後分布を求めるコマンド pdisc に入れます。

```
> data <- c(12,18)
> post <- pdisc(p, prior, data)
```

これで事後分布の予測が終わります。早速、post の中身を見てみましょう。

```
> round(post,5)
[1] 0.00000 0.17092 0.82897 0.00012 0.00000
```

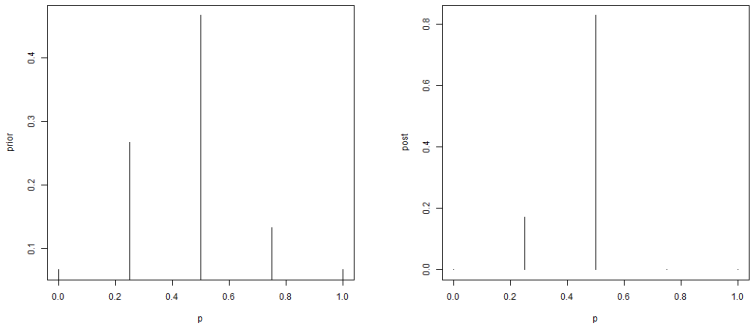
事後分布は次のような感じになりました。実験での成功率が $12/(12+18) = 0.4000$ だったので、0.5 以下にさらに偏った形になっています。

成功率	0	0.25	0.5	0.75	1
成功率が上記のものである確度	0.0000	0.1709	0.8289	0.0001	0.0000

事前分布と事後分布の違いを見てみるために、事後分布を出力してみます。

```
> plot(p,post,type="h")
```

2 つの分布を並べてみると、次のようになります（左：事前分布、右：事後分布）。やはり、事後分布は 0.4 の近くにかなり偏ることが分かるはずです。

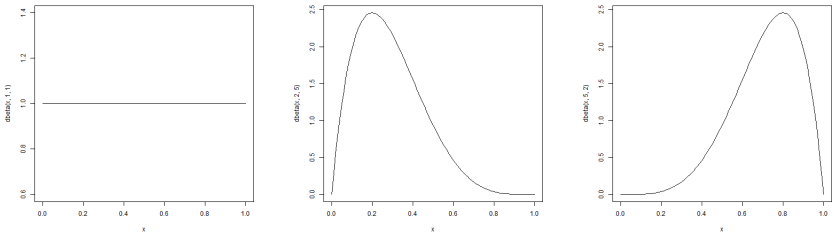


2.3 ベータ分布を用いる方法

先ほどの作業では事前分布が離散的な場合を使用しましたが、続いて連続的な場合を使います。この例題のような問題を解く場合、よく使われる分布がベータ分布です。ベータ分布は、次の確率密度関数で示される分布です。

$$f(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \qquad \text{ただし、} B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \tag{2.1}$$

このベータ分布を用いると、ベルヌーイ試行の成功率が p 以下である確率が q、というふうに表すことができます。下記のグラフは、左から $(\alpha, \beta) = (1,1), (2,5), (5,2)$ のベータ分布の確率密度関数です。



ここで、成功率の事前分布を、次のような設定に従うベータ分布にしてみましょう。

1. 成功率が 0.3 以下となる確率は 0.5
2. 成功率が 0.5 以下となる確率は 0.9

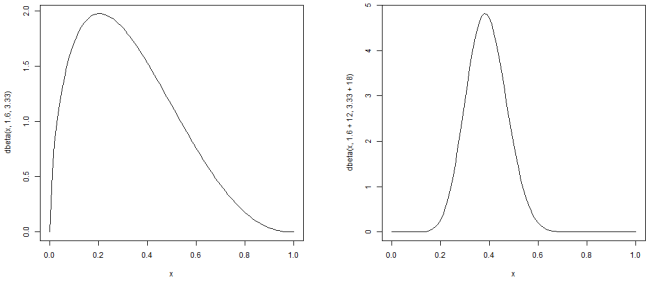
これを満たすベータ分布を求めるためには、コマンド `beta.select` を使います。次のように入力すると、ベータ分布のパラメータ α, β を求めることができます。

```
> quantile1 <- list(p=0.5,x=0.3) #x=0.3 以下となる確率が 0.5
> quantile2 <- list(p=0.9,x=0.6) #x=0.6 以下となる確率が 0.9
> beta.select(quantile1,quantile2)
[1] 1.60 3.33
```

これによると、条件を満たすベータ分布は $\alpha = 1.60, \beta = 3.33$ であることがわかります。これを用いて推計を行います。証明は省略しますが、事後分布は、成功した回数を s、失敗した回数を f で表すと、次のようになります。この性質がベータ分布をベイズ推定で使う強みです。

$$f(p|\alpha + s, \beta + f) = \frac{p^{\alpha+s-1}(1-p)^{\beta+f-1}}{B(\alpha + s, \beta + f)} \tag{2.2}$$

これを使うと、事後分布は $\alpha = 1.60 + 12, \beta = 3.33 + 18$ のベータ分布になることがわかります。これをプロットすると次のようになります（左：事前分布、右：事後分布）。



なお、$\alpha = 1, \beta = 1$ のベータ分布は、事前情報のない事前分布（無情報事前分布）によく使われます。またベータ分