第4回 DSRT
ベイズ統計学の医薬品の臨床開発での活用について

議題1-1：ベイズ統計学入門

ベイズ流アプローチとは

- ベイズ流アプローチの主要な考え方

個人の事前知識*（事前確率）と尤度（データ）を結合し、確率を更新すること（事後確率）

  – 人間の学習プロセスを模倣した考え方である

出典：Emmanuel Lesaffre, Andrew B. Lawson　宮岡悦良　監訳『医薬データ解析のためのベイズ統計学』より抜粋し、一部表現を変更

*個人の事前知識は、個人ごとに異なり得るものであり、主観性を持っている可能性がある。

ベイズ流アプローチの原点

- ベイズの定理(1763年)

条件付き確率が逆になっている

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

- Thomas Bayesが初めて逆確率を‘発案’
- ベイズの他界後、Richard Priceにより1763年にベイズの定理として定式化、論文投稿

ベイズ流アプローチの原点

- Pierre-Simon Laplaceにより、ベイズの定理をパラメータθ、データYで表現

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

- 注目は**パラメータθが確率変数***として使われている点

⇒ ベイズ流アプローチの原点

  – これにより、多くの頻度論者（特に影響力のあった、NeymanやFisher ）の批判に遭い、長い間無視されることになった。

*パラメータθは頻度論同様、θoとしておくこともできるが、真のパラメータθoはわからないため、確率変数としておくことが一般的

ベイズの定理の整理

θ:パラメータの確率変数
y:データの確率変数

$$P(\theta|y) = \frac{P(y|\theta) \times P(\theta)}{P(y)}$$

$$\Longleftrightarrow \quad P(\theta|y) \propto P(y|\theta) \times P(\theta)$$

比例

事後分布(確率) ∝ 尤度 × 事前分布(確率)

posterior　　　　　　　　　　prior

事前分布(事前確率)に新たなデータ(尤度)をかけ合わせることにより、更新した事後分布(事後確率)を得ることができる

事前分布(prior)

$$P(\theta|y) \propto P(y|\theta) \times \boldsymbol{P(\theta)}$$

これ

- どのような確率分布（二項分布、正規分布、ポアソン分布etc.）を想定することも可能
- 事前情報が無いor使いたくないとき
  - 事後分布の計算に、事前分布の影響を極力なくしたい

→ 無情報事前分布*(noninformative(NI) prior)の利用

例：Jeffreyの事前分布

  - 実用的には、無情報分布に近似できる、弱情報事前分布の利用

例：正規事前分布N(0,10^6)

*ベイジアンは過去に無情報であることを表現することに多大な努力をしてきたが、現在では厳密に無情報であることを示すことは難しく、実際には無情報ではないことを認めている。

事前分布(prior)

$$P(\theta|y) \propto P(y|\theta) \times \boldsymbol{P(\theta)}$$

これ

- <u>**適切な**</u>事前情報がある場合、情報を盛り込んだ事前分布を設定
  例1：過去の試験データから事前分布を構成する
  例2：専門家の意見を元に事前分布を構成する

※事前分布に依存して事後分布が変化する。すなわち、解析結果が変化する。事前情報を盛り込むことはベイズ流アプローチの一番特徴であるが、事前情報の設定により客観性を損ねる可能性があるため、慎重な議論が必要である。

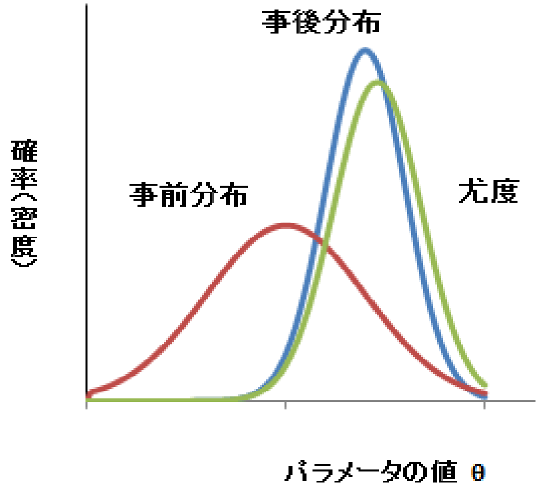
- 求めた事後分布を次試験の<u>事前情報</u>*として利用することも可能。（ベイズ更新）
  例：PhaseⅡ試験で得られたデータと設定した事前分布$P(\theta)$から事後分布$P(\theta|y_2)$を算出する。PhaseⅢ試験では、PhaseⅢ試験で得られたデータとPhaseⅡ試験で求めた事後分布$P(\theta|y_2)$を事前分布$P(\theta)$として利用し、事後分布$P(\theta|y_3)$を算出する。

*そのまま事前分布として用いても良いが、その他の情報とあわせて事前分布とすることも可能。

事後分布(posterior)

これ

$$\boldsymbol{P(\theta|y)} \propto P(y|\theta) \times P(\theta)$$



- 事前分布と尤度から算出するため、

分布形は事前分布と尤度により決定する

（パラメータμ、分散既知に対して）

事前分布：正規分布　尤度：正規分布

⇒ **事後分布：正規分布**

- 大標本のとき、事前分布・尤度の分布にかかわらず、事後分布は漸近的に<u>正規分布</u>に近似する*（ベイズ流の中心極限定理）→事前分布の選択が重要ではない
  – 1785年にラプラスが証明。古典的な中心極限定理よりも前であった

*例数が大きいときの事後分布は<u>尤度への依存が大きくなる。</u>

事後分布算出の例

μ_0, σ^2:既知　データ:$\mathrm{y} = (y_1, \dots, y_n)$, $y_i \sim N(\mu, \sigma^2)$ $(i = 1,2, \dots, n)$

平均値:$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

- 尤度:$L(\mu|y) \propto L(\mu|\bar{y}) = P(\bar{y}|\mu)$　$\bar{y}|\mu \sim N(\mu, \sigma^2/n)$
- 事前分布:$P(\mu)$　$\mu \sim N(\mu_0, \sigma^2/n_0)$ (n_0は定数)
- 事後分布:$P(\mu|\bar{y}) \propto P(\bar{y}|\mu) \times P(\mu)$

$$\mu|\bar{y} \sim N\left(\boxed{\frac{n_0\mu_0 + n\bar{y}}{n_0 + n}}, \frac{\sigma^2}{n_0 + n}\right)$$

  – 事後分布の平均値が事前分布の平均値μ_0と尤度の平均値$\bar{y}$の重み(事前分布、尤度の例数)付き平均値となっている。

事後分布の積分計算の問題

- 求めた事後分布から確率計算や要約を行う際に、解析的に積分計算を行うことが難しいことが多い

事前分布:コーシー分布 $P(\theta)=\frac{1}{\pi\gamma[1+\left(\frac{\theta-\theta_0}{\gamma}\right)^2]}$

尤度:正規分布 $L(\theta|y)=\frac{1}{(2\pi)^{n/2}\sigma^2}\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2]$ **解析的に計算が難しい**

事後分布の平均値:$\frac{1}{\pi\gamma(2\pi)^{n/2}\sigma^2}\frac{1}{P(y)}\boxed{\int\frac{\theta\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2]}{1+\left(\frac{\theta-\theta_0}{\gamma}\right)^2}d\theta}$

  - さらにパラメータが複数あると、より計算量が増える

難しい事後分布の積分計算の対処法①
～数値積分～

- 近似的に積分計算の解を求める

例

- 長方形による区分求積法
- 台形法　など

難しい事後分布の積分計算の対処法②
～事後分布からのサンプリング～

- モンテカルロ積分法
  - 事後分布からシュミレーションにより生成した確率変数の標本平均を用いて事後分布の平均を近似する

より難しい積分を要する場合（多変量分布等）・・・

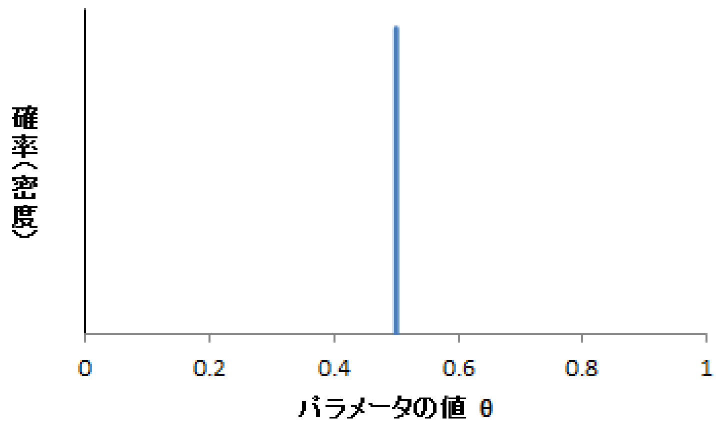
- **マルコフ連鎖モンテカルロ法（MCMC法）**

例：ギブス・サンプラー、メトロポリスアルゴリズム

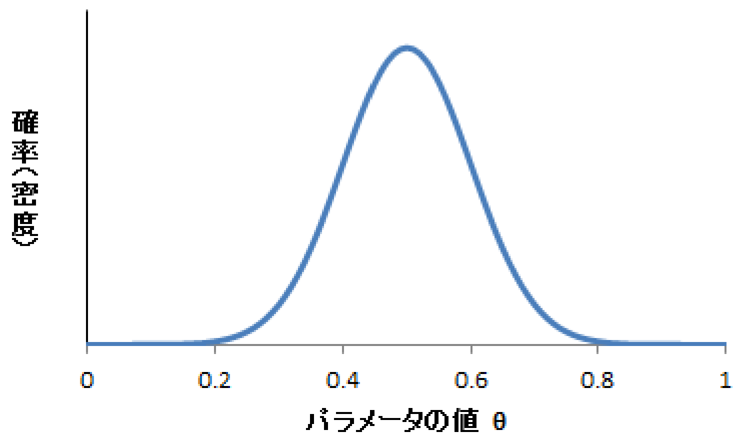
解析的な計算が難しかったベイズ統計学だが、1980年代後半になると、ギブス・サンプラーを用いたWinBUGSが実装され、実務的に利用されるようになっていった

頻度論とベイジアン（パラメータの扱い）

- 頻度論
  - パラメータθは固定値



- ベイジアン
  - パラメータθは確率変数とし、確率分布を持つ



頻度論とベイジアン（要約統計量による推測）

- 頻度論
  - データからパラメータθの推定値（平均値、標準偏差等）を計算する

- ベイジアン
  - 確率変数であるパラメータθの事前分布とデータ（尤度）から事後分布を計算し、事後分布の要約量を算出する

例. 事後平均値　$\bar{\theta} = \int \theta P(\theta|y) d\theta$

事後分散　$\bar{\sigma}^2 = \int (\theta - \bar{\theta})^2 P\,(\theta|y) d\theta$

頻度論とベイジアン（区間推定）

- 頻度論
  - 信頼区間（Confidence Interval）を用いる
  - 95%信頼区間：信頼区間を繰り返し構成した場合、95%は真のパラメータθを含む（**確率的に変動するのは区間**）
- ベイジアン
  - 信用区間（Credible Interval）を用いる
  - 100(1-α)%信用区間：$P(a \leq \theta \leq b) = 1 - \alpha$を満たす区間$[\boldsymbol{a}, \boldsymbol{b}]^*$
  - 95%信用区間：信用区間に95%の確率でパラメータθが存在する区間

（**確率的に変動するのはθ**）

*どこの95%区間を取るのかどうかでa,bの値は変動する。等裾信用区間、最高事後密度区間などがよく用いられる。

参考文献

- Emmanuel Lesaffre and Andrew B.Lawson著　宮岡悦良　監訳
『医薬データ解析のためのベイズ統計学』
（共立出版株式会社、2016）