取得するデータの範囲が限られている場合の
マルコフ連鎖モンテカルロ法を用いたベイズ推定についての検討

松香研究室 4 年　19L1085M　川口 結輝

1. はじめに

大学など限られた環境で実験を行う場合，同じ大学の学生に実験を協力してもらうことが多い。また，心理学における実験や調査では人が被験者になるため，多くのデータを収集することが困難になることがある。

このような時，限られた集団からデータを得ていることになるため，母分布に比べて偏りのあるデータになっている可能性がある。

実験の内容により，様々なデータ分析の方法があるため，それらすべての分析方法においてデータに偏りにある場合について検討をするのは難しいため，本研究ではデータが限られた範囲内から得られたデータ，かつデータ数が少ない場合におけるマルコフ連鎖モンテカルロ法を用いたベイズ推定の事前分布の影響を調べた。

・ベイズ推定

事前分布と尤度を利用して事後分布を推定

$p(q|\mathbf{D}) \propto p(\mathbf{D}|q) \times p(q)$（事後分布は事前分布と尤度の積に比例する）が成り立つ。

・事前分布　$p(q)$

データが得られる前に想定する，母数qが得られる確率を示した分布

・事後分布 $p(q|\mathbf{D})$

データが得られた後で想定する，qが得られる確率を示した分布

・尤度 $p(\mathbf{D}|q)$

事前分布のもとでqが得られたときに，データが得られる確率

・マルコフ連鎖モンテカルロ法

・マルコフ連鎖

未来の状態は現在の状態のみから決まる

・モンテカルロ法

乱数を生成して近似計算を行う手法

・マルコフ連鎖モンテカルロ法

直前の状態だけを考慮しながら，乱数を用いて近似計算を行う手法

・マルコフ連鎖モンテカルロ法によるベイズ推定

1. データ($\mathbf{D}$)を得る。
2. 事前分布をもとに乱数を生成し，qの値を設定する。
3. 得られたqの値を尤度$p(\mathbf{D}|q)$に当てはめて，尤度と事前分布の積$p(\mathbf{D}|q)p(q)$を求める。

4. 尤度と事前分布の積$p(\mathbf{D}|q)p(q)$の値を評価する*。

5. 2~4 を数万回繰り返し，事後分布を生成する。

＊$p(\mathbf{D}|q_n)p(q_n)$と$p(\mathbf{D}|q_{n+1})p(q_{n+1})$の比を用いて， 採用する値を確率的に決定する。（$q_n$は$n$回目の試行で得られた$q$の値）

2. シミュレーション 1

2.1. 目的

データ数，データ範囲，尤度関数の t 分布の自由度による事後分布への影響を調査した。

2.2. 方法（シミュレーション 2 でも同様）

統計解析ソフト R の Stan パッケージを利用した。

2.3. シミュレーションの状況設定（シミュレーション 2 でも同様）

ある実験で IQ についてのデータを集め，そのデータからマルコフ連鎖モンテカルロ法を用いたベイズ推定を行い，母集団の分布を推定する。

母集団は平均 100，分散 15 の正規分布に従っているが，データとして得た IQ は平均 100，分散 15 の正規分布を標準化した標準正規分布のうちの，限られた範囲から生成された値になっている。

この時， 推定した事後分布が標準正規分布に近似すれば， 正しく分析ができていると評価する。

2.4. シミュレーションの条件

事前分布は正規分布，尤度関数は t 分布を用いた。

シミュレーション 1 は以下のような条件の組み合わせで，計 27 条件で行った。

・データ数（以下 N とする）： 10, 30, 100

・データ範囲（以下 limit とする）： 0~200. 90~110, 110~130

・t 分布の自由度（以下 df とする）: 1, 3, 100

データ数が少ない場合において， 事前分布の分散が大きすぎると， 事後分布のばらつきが非常に大きくなるため，データ数ごとに事前分布の分散を変更した。

・N=10 のとき，事前分布は平均 0，分散 3 の正規分布

・N=30 のとき，事前分布は平均 0，分散 5 の正規分布

・N=100 のとき，事前分布は平均 0，分散 100 の正規分布

2.5. 結果

自由度が大きくなるにつれて， 事後分布のばらつきが大きくなり， 95%区間が広くなった。

limit=90~110 の事後分布は， limit=0~200 の事後分布とあまり変わらなかった。一方，limit=110~130 の事後分布は limit=0~200 の事後分布と比較して 95%区間に偏りが生じた(図 1)。

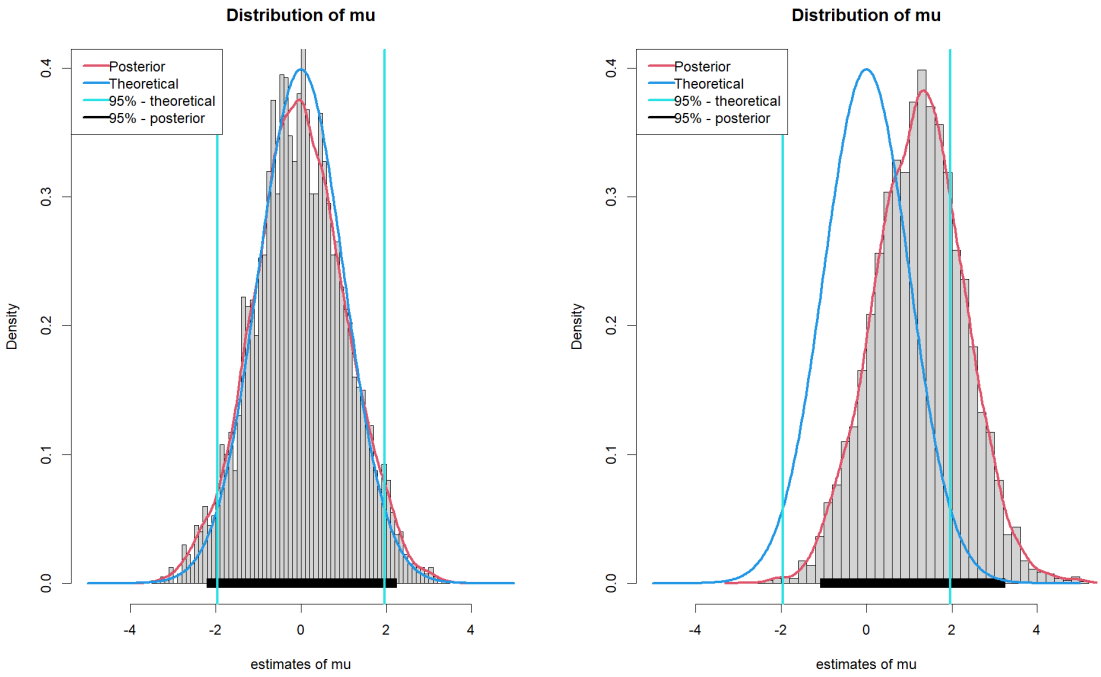

図 1

左: N=100, limit=0~200, df=1, 95%区間: -2.096~2.146, y 最大値: 0.375

右: N=100, limit=110~130, df=1, 95%区間: -0.969~3.143, y 最大値: 0.383

3. シミュレーション 2

3.1. 目的

データ範囲が母分布の中心と比べて偏っている場合における, データ数, 事前分布の分散, 尤度関数の分散, 尤度関数の t 分布の自由度による事後分布への影響を調査した。

3.2. シミュレーションの条件

事前分布は正規分布, 尤度関数は t 分布を用いた。

シミュレーション 2 は以下のような条件の組み合わせで, 計 36 条件で行った。

・データ数(以下 N とする): 10, 100

・事前分布の分散: 0.8, 1, 5

・尤度関数の分散: 10, 15, 30

・尤度関数の t 分布の自由度(以下 df とする): 1, 100

3.3. 結果

N=10 の場合, 事前分布の分散が母分布の分散と一致, もしくは母分布の分散よりも小さい時に偏りが小さくなり, 母分布に近づく。 N=100 の場合は, 事前分布の分散が母分布の分散と一致, もしくは母分布の分散よりも小さいことに加え, 尤度関数の分散の値を大きめに設定することで母分布に近づいた(図 2)。

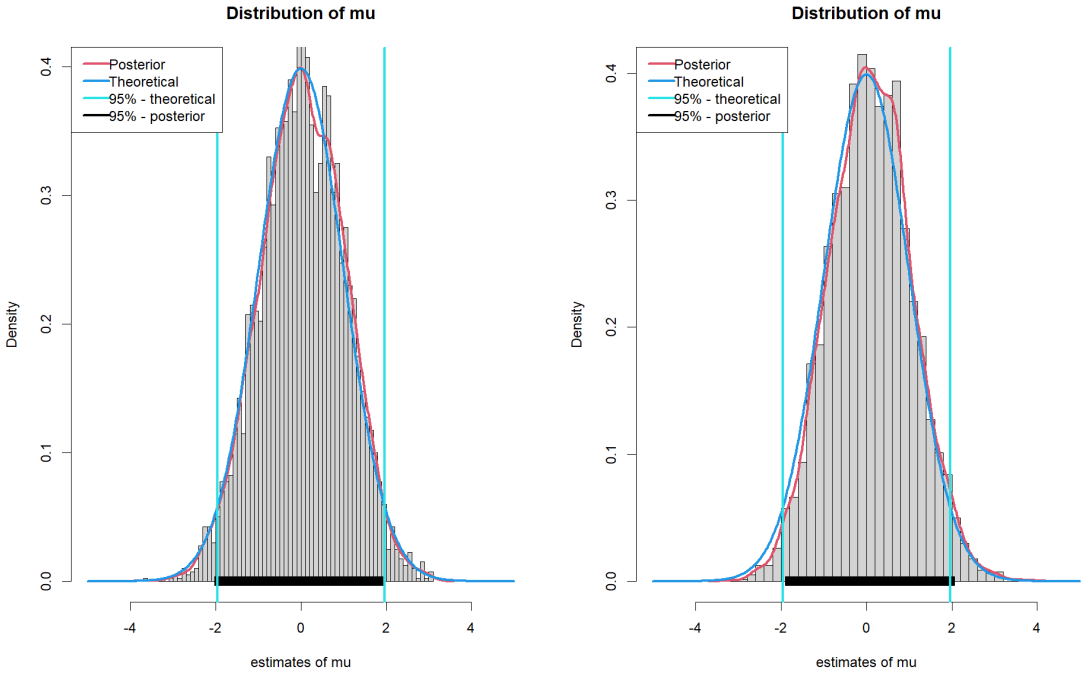

図 2

左：N=10，事前分布分散：1，尤度関数分散：30, df=100, 95%区間：-1.916~1.851，
y 最大値：0.400

右：N=100，事前分布分散：1，尤度関数分散：30, df=100, 95%区間：-1.797~1.968，
y 最大値：0.404

4. 総合考察

データ範囲が母分布の中心を捉えている場合は，データ範囲がないときと同じような事後分布となる。データ範囲が母分布の中心と比べて偏りがある場合は，事後分布が母分布に比べて偏りを持つが，事前分布の分散と尤度関数の分散を操作することで当てはまりを良くすることができる。

データ範囲に偏りがある場合に事後分布の偏りを緩和するためには，事前分布を母分布に近づける必要があり，事前分布とデータを利用して事後分布を形成するベイズ推定において，母分布に近い事前分布を手に入れている状態では，そこからデータを利用してベイズ推定を行う意義がなくなってしまうことが課題点として挙げられる。