Q. & A. 続

Q.16　条件つき予測分布などというものを使っていいのですか？

x^* について推論するのであれば，x^* 以外の変数は周辺化（積分消去）する事後予測分布がベイズ統計学の基本です。EAP などの推定値を代入するだけの条件つき予測分布などというものを使っていいのでしょうか。

x^* について推論するのであれば，x^* 以外の変数は周辺化（積分消去）する事後予測分布がベイズ統計学の主流の方法の 1 つです。でもそれだけ、ではありません。p.39 にも書いた通り、条件つき予測分布には情報損失があります。しかし条件つき予測分布は著者の作った用語ではなく、古くから使用されてきた伝統ある用語です。昔は、計算機の能力の限界で条件つき予測分布しか利用できないことも珍しくありませんでした。計算機がこれだけ発達し、これからも発達するであろう今日において、条件つき予測分布という用語を本書で使用したのには、教育面・実践面の 2 つの理由があります。

教育面の理由

予測分布は将来のデータの分布ですから、初心者が是非習得すべき重要な概念です。条件付き予測分布は、正規分布などの分かり易い単純な式の形式になるので、初心者にとって理解が容易です。「未知だった母数にデータを利用した推定値を代入し、明示的なモデル式を使用して将来を予測する」という流れが明快です。それに対して事後予測分布は、正規分布のような名前の付いた分かり易い分布や、モデル式にはなりません。事後予測分布は、名も無き分布です。これは初心者にとっては、掴みどころがなく、とても難しい概念となります。事後予測分布だけで予測を理解させるのは初心者教育になじみません。

実践面の理由

実践というからには著者が実際に実践的分析をしている領域の実例を挙げます。それは、たとえば教育測定分野の項目反応理論においてです。この分野では EAP 推定値を代入しただけのモデル式でテストの予測確率を求めることが普通です。条件付き予測分布を利用するのがメインのやり方です。

まず事後予測分布から説明します。時期 1 でのテスト実施データ T_1 から、受験者の能力に関する事後分布、$f(\theta|T_1)$ を求めます。この受験者が時期 2（次期）のテストを受けるとどうなるかは、T_2 の予測分布で得られ、データ生成モデル $g(T_2|\theta)$ を用いて、

$$f(T_2|T_1) = \int g(T_2|\theta) f(\theta|T_1) d\theta$$

と求めます。しかし我々の分野ではこの事後予測分布はめったに使いません。CBT（Computer-Based Testing）では、受験者が回答したら、一瞬のうちに次の問題を提示しなければなりません。プール内の多数の問題候補ごとに事後予測分布を作っていたのでは、レスポンスが遅くなり、受験者が回答した直後に問題提示できません。事後予測分布の僅かばかりの情報効率の良さなど、現実的要請の前には何の役にも立ちません。その代りに、条件付き予測分布

$$g(T_2|\theta_{eap})$$

を利用します。条件付き予測分布なら、プール内の多数の問題候補ごとに予測分布を作って比較しても、瞬時の対応が可能です。

また即時性を要求されない場合でも、事後予測分布ではなく条件付き予測分布を多用します。仮に何か月か先の行動予測でも、常に前回の試験結果 T_1 を持ち歩かなければならないのは実際的でなく、現実的ではありません。テスト事業は、現実的に困ったことが起きると直ぐにマスコミが騒ぐし、場合によっては訴訟に発展します。現実的対応がシビアに要求される分野です。そこで問題なく長年使用されていることは、条件付き予測分布が現実的な対応を与えてくれることの証拠です。事後予測分布を使用して項目選択するシステムは、論文で理論的に提案されています。しかし IRT の教科書を書くために 2011 年に調べた際には、事後予測分布を使用して項目選択している大規模 CBT テストシステムを、1 つも見つけられませんでした。

いくら計算機が発展しても、人間はもっと big な、更に big なデータを扱おうとするでしょう。したがって式の評価だけで将来を予測できる簡便でスピーディーな条件つき予測分布は、今後も使われ続けます。条件付き予測分布は、big データと格闘するデータ分析者にとって必須の、なくてはならないツールであり続けるでしょう。

Q.17　「確率の確率分布」などというものを使っていいのですか？

本書には２５％点の事後分布や、優越率の事後分布など、たくさんの「確率の確率分布」が登場します。データから知見を引き出す際に困ったことは生じませんか？　妥当な分析知見が得られますか？

理論面からの説明

t 値が何故 t 分布するかは、解析学の観点からは変数変換によって説明されます。解析学による変数変換は変換式の個別の事情に影響されるので、伝統的な統計学における分布論は数学的に大変複雑になります。対して「統計科学フロンティア 12　計算統計 II　マルコフ連鎖モンテカルロ法とその周辺　岩波書店」の p.189 には、「θ の実数値関数 $g(\theta)$ の周辺事後密度関数を求めるには $g(\theta^{(t)})(t=1,2,\cdots,n)$ を用いた密度関数の推定を行えばよい。」とあります。要するに生成量は、変換式の個別の事情によらずに、プリミティブに確実に変数変換を実行した結果であり、変換された指標の事後分布を与えます。

正規分布モデルにおいて、MCMC 法により $\mu^{(t)}, \sigma^{(t)}$ を妥当に抽出するということは、$\mu^{(t)} + (-0.675) \times \sigma^{(t)}$ を妥当に抽出することと同義であり、即ちそれは $(25\%点)^{(t)}$ を妥当に抽出しているということです。「μ や σ は確率変数だけれども、25%点は絶対に確率変数ではない」と言い切ったら、それは誤りです。

優越率は、抽出された母数に対して一意に値が定まります。値域が $[0,1]$ という安定的な変換 $g(\quad)$ を施しているだけですから、妥当に母数を抽出することと妥当に $\pi_d^{(t)}$ を抽出することは同義です。母数が確率分布するなら、優越率も確率分布します。

実践面からの要請

「確率の確率分布」は何より分析現場の要請として避けて通れません。本書に登場する A 君は、10 杯の牛丼を調べ、第 2 章において「4 回に 1 回は 76.7 g 未満を覚悟しなければならない、また 95%の確信で、それは [71.5g,80.3] の間だ」と主張しています。しかし B 君という新しい登場人物が、「僕は 1000 杯の牛丼を調べたよ。4 回に 1 回に覚悟しなければいけないのは、76.0 g だよ。また 95%の確信でそれは [75.8g,76.2] の間だ。76.7 g だなんて、A 君はなんてお人よしなんだ！　笑っちゃうよ。」と言ったとします。人々はどちらを信じるでしょう。妥当にデータを取っている限りにおいて、もちろん B 君を信じます。優越率も同じです。20 名のデータを取った実験より 2000 名のデータから計算された優越率の揺れが小さいことは直観的に自然で、分析現場の要請に答えるために、理論体系としてそれを示す必要があります。

教程としての必要性

分析現場の要請ばかりでなく、教育的側面からも「確率の確率分布」は必須です。伝統的な頻度論統計学では、２５％点の信頼区間の計算方法や、優越率の信頼区間の計算方法が提案されています。本書の一番の主旨は、「伝統的な頻度論による入門教程をベイズ的アプローチに置き代えること」です。だから「なんだ、頻度論で簡単にできることが、ベイズでは出来ないのか！」と言われては教程として成り立ちません。２５％点の確信区間や優越率の確信区間の計算方法を提示することは、オルタナティブの教程として外せません。

妥当性の確認

「牛丼問題」は、わずか $n=10$ です。にも係わらず、事後予測分布の 25%点と、25%点の事後分布による点推定値は、ほとんど一致します。「数学教授法問題」は、各群わずか 20 人です。にも係わらず、事後予測分布の優越率と、優越率の事後分布による点推定値は、ほとんど一致しています。もし条件付き予測分布を使った確率の事後分布が理論的に過っているなら、こうも数値が似通ることは考えられません。パラメータ数が少ないので、公刊前にシミュレーションもしています。T が大きければ、n が相当に小さい段階から両者のかい離は実践的には無視できる程度です。理論的には、事後予測分布より情報効率は落ちるのですが、その損失は少ないと判断できます。本書で提案している範囲の「確率の確率分布」を使った分析は、安心してご利用ください。