ベイズ統計

2018 年 前期講義資料 （作成者：大塚 淳）

1　ベイズの定理

H, E を事象とする。ここで、H はある仮説、E は観察された証拠を念頭においている。なお以下では、H の余事象（対立仮説）を表すために H^c でなく $\neg H$ を用いる。条件付確率の定義から、次の**ベイズ定理**が導かれる

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

このとき

1. $P(E|H)$ は**尤度**（likelihood; 仮説のもとでどれだけ証拠が得やすいか）
2. $P(H)$ は**事前確率**（prior probability; 証拠が得られる前の段階で、仮説はどれだけ確からしかったか）
3. $P(H|E)$ は**事後確率**（posterior probability; 証拠が与えられたもとでの仮説の確からしさ）

と呼ばれる。ベイズ定理は 1,2 をもとに 3 を計算する、言い換えれば、仮説の説明力とそれが前々から持つ確からしさをもとに、証拠が与えられた後の仮説の確率をアップデートするためのルールを与える。

ちなみに、全確率の定理を分母の $P(E)$ に用いれば、ベイズ定理は次にように書き換えられる

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\neg H)P(\neg H)}$$

なぜこんなことをするのかというと、一般に証拠が得られる確率 $P(E)$ というのは評価し難いからである。それ比べ仮説の事前確率と尤度は見積もりやすいので、この式の方が計算上役に立つことが多い。

■**例**　くじ玉の入った二つの壺 A, B があって、A には 10 個中 1 個の割合でしかあたりが入っておらず、B は 10 個中 5 個があたりである。いま目の前に壺があるが、A,B どちらの壺かはわからない。一個ひいたらはずれであった。これを証拠 E としたとき、 この証拠によって「壺は A である」という仮説 H の確率はどのようにアップデートされるか。

- くじを引く前はどちらの壺かわからなかったので、事前確率は半々としよう。つまり $P(H) = P(\neg H) = 0.5$

- 壺 A は 10% の確率であたりなので $P(E|H) = 0.9$、一方壺 B は半分があたりなので $P(E|\neg H) = 0.5$
- これをベイズ定理に当てはめると、事後確率は

$$P(H|E) = \frac{0.9 \times 0.5}{0.9 \times 0.5 + 0.5 \times 0.5} = \frac{0.9}{1.4} \sim 0.64$$

つまりはずれが出たという証拠 E によって、目の前の壺が A であるという仮説の確からしさは 50% から約 64% まで高まったということが言える。

2　ベイズ定理を用いた帰納推論

ベイズ主義によれば、我々が行う帰納推論は、すべてベイズの定理を用いた確率計算にほかならない。以下では、アブダクション、蓋然的仮説の反証、枚挙的帰納法という 3 つの帰納推論について、それがベイズ主義の枠内でどのように扱われるのかを見ていく。

2.1　アブダクション

アブダクションとは、次のような推論である。

前提 1　細胞核内の核酸 A,T,G,C のうち、A/T, G/C の量比は等しい（シャルガフの比率）
前提 2　DNA が二重らせん構造をとっているなら、A/T, G/C の量比は等しくなる
結論　よって DNA は二重らせん構造をとっているだろう

ベイズの定理を用いて、この推論を評価するために必要なのは：

1. 尤度：
   - 二重らせん仮説のもとで A/T, G/C が同量になる確率 $P(E|H)$ 、および
   - 対立仮説のもとで A/T, G/C が同量になる確率 $P(E|\neg H)$
2. 事前確率：二重らせん仮説のもともとの確からしさ $P(H)$

これらが定まれば、二重らせん仮説の事後確率が求まる。

■練習

- $P(E|H) = 0.9, P(E|\neg H) = 0.1, P(H) = 0.5$ のとき、$P(H|E)$ を計算せよ。また $P(E|\neg H) = 0.7$ ではどうか。

2.2　反証

同様の原理で、蓋然的な仮説の予測が外れたとき、その仮説がどの程度反証されるかも評価することができる。

前提 1　細胞核内の核酸 A,T,G,C のうち、A/T, G/C の量比は等しい
前提 2'　ポーリングの三重らせん仮説 (H') からは、A/T $\neq$ G/C と予測される
結論'　よってポーリングの仮説は誤っているだろう

これも同様に尤度、事前確率を定めれば求まる。

2.3　枚挙的帰納法

枚挙的帰納法とは、次のような推論である

前提 1　カラス 1 は黒い
前提 2　カラス 2 は黒い
...
前提 n　カラス n は黒い
結論　よって全てのカラスは黒い

これは、H を「全てのカラスは黒い」という仮説、$E_1, E_2, \ldots, E_n$ をそれぞれのカラスの観測だとして、事後確率 $P(H|E_i)$ をアップデートしていくことに対応する。事前確率 $P(H) = 0.1$、仮説の尤度 $P(E_i|H) = 1$、対立仮説の尤度 $P(E_i|H) = 0.9$ とし、観測を繰り返したときの事後確率の変化を図 1 に示す。

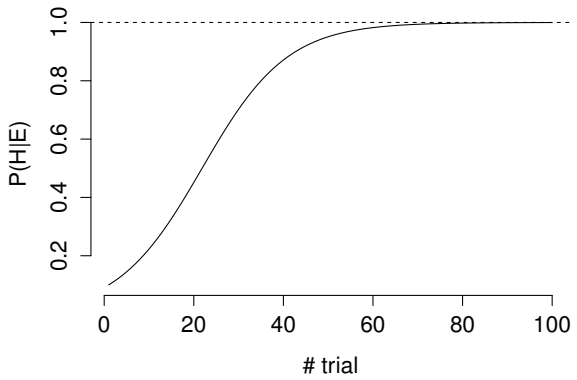

図 1　$P(H) = 0.1, P(E_i|H) = 1, P(E_i|H) = 0.9$ としたときの H の事後確率の変化

ここから分かるように、最初はありえないと思われていたような仮説でも、確証事例を重ねるにつれほとんど確実になっていく。このように観測・実験を重ねるにつれ事前確率の影響が払拭され、事後確率が尤度へと近づいていくことを、swamping of priors といい、これが枚挙的帰納法の根拠となっている。

3　事前確率

前述のように、ベイズの定理を用いて仮説の事後確率を求めるには、尤度と事前確率が必要であり、どちらも欠かすことができない。このうち特に事前確率を求めるところが、ベイズ主義者と他の立場を分ける点であり、論争の火種となってきた。この論争を考える前に、まず、ベイズ的推論において事前確率がいかに重要かということを、例から見てみよう。

3.1　基準率の誤謬

例 40 歳以上の女性のうち 1000 人に 4 人が乳がんになるといわれている。しかし今日のマンモグラフィーはかなり精確であり、9 割の確率で乳がんを発見でき、偽陽性率（健康なのにがんと誤診断する確率）も 1 割にすぎない。それを考えれば、40 歳以上の女性はぜひ定期的な検診を受け、乳がんの早期発見に努めるべきだ。

この主張にはどれほど説得力があるだろうか。乳がんであることを H、陽性を E で表すとすると、$P(H) = 0.004, P(E|H) = 0.9, P(E|\neg H) = 0.1$。したがってマンモグラフィーで陽性が出たときに実際に乳がんである確率は

$$P(H|E) = \frac{0.9 \times 0.004}{0.9 \times 0.004 + 0.1 \times 0.996} \sim 0.035$$

つまりたとえ試験結果が陽性だとしても、実際に乳がんである確率は 4% にも満たない！上の例同様、ベイズ推定は尤度に加えて事前確率によって決まるが、この場合は事前確率が極めて低い (0.004) ため、たとえ陽性だとしても事後確率はそれほど上がらないのである。でも我々はしばしば事前確率を忘れ、尤度が高ければ事後確率も高いだろうと早合点してしまう。これを**基準率の誤謬** (base rate fallacy) という。この事実ゆえ、稀な病気の医療診断は極めて難しい。試験が非常に高精度であっても、事前確率が低いために、どうしても誤診断の確率が増えてしまうのである。

3.2　事前確率の設定

上の例で明らかなように、証拠のもとでのある仮説の確からしさ（事後分布）を知るためには、事前分布が必要不可欠である。では、この事前分布はどこからやってくるのか？ベイズ統計では、主に次の 3 つの方法によって事前分布が決定される。

1. **無情報事前分布**
   考慮している仮説について何も事前に情報がない場合、そのどれもが同程度ありそうであると考え、全仮説に同じ確率を割り当てるような事前分布を考える。これを無情報事前分布と呼ぶ。例えば前述の壷 A, B にそれぞれ 0.5 を割り当てる分布は無情報事前分布である。またカラスの例では、「h% のカラスは黒い」という仮説を 0 から 1 までの値をとる確率変数 H で表せば、無情報事前分布は $0 \leq h \leq 1$ の間の一様分布となる。

2. **それ以前の事前分布**
   証拠が続けて得られる場合、以前の証拠によってアップデートされた仮説の事後確率を、新たな証拠を加味する際の事前分布として用いることが考えられる。
3. **経験ベイズ**
   割合などに関する仮説の場合は、事前に調査や観測を行い、その結果をそのまま事前分布とすることが考えられる。例えば上のマンモグラフィーの例では、1000 人に 4 人が乳がんであるという事前調査が事前分布として用いられた。

しかしどれも問題なしとは言えず、この点がベイズ統計学を巡る論戦の的となってきた。授業ではこれらについて見ていく。

練習問題

1. 節 1 で見た壺の例の続きで、一回ひいたはずれくじを壺に戻した上、さらにもう一回ひいてみたら、またはずれであった。このとき、仮説 H はどの程度確からしくなるか。
2. 「今年のインフルエンザ (H) の一般的な症状は、激しい頭痛と倦怠感 (E) です。これらの症状を認めたら、インフルエンザを疑って下さい。」
   a) $P(H) = 0.1, P(E|H) = 0.8, P(E) = 0.2$ としたとき、インフルエンザの事後確率を求めよ。
   b) $P(H) = 0.1, P(E|H) = 0.8, P(E|\neg H) = 0.1$ としたとき、インフルエンザの事後確率を求めよ。
3. 五色牛教授の授業では、期末試験で不合格だった生徒に変わった救済策がとられる。それは一種の運試しで、学生は A, B, C の三つの箱から一つを選ぶ。そのうち一つには牛の写真が入っていて、それを当てれば単位がもらえる。残りの二つは空で、その場合は落第である。さて、あなたはその哀れな学生で、A の箱を選んだとしよう。すると五色牛教授は C の箱を開けて、中身が空っぽであることを示す。そしてあなたに聞く。「今だったら B の箱に変えてもいいけど、変える？」— さて、あなたは変えるべきだろうか。その理由も合わせてのべよ。