2-2 結果から原因を探っていく

事前確率は恣意的か？

ベイズ統計が長い間、日の目を見なかった原因は、計算の煩雑さだけではありません。「はじめに」でも述べたように、「事前確率に曖昧性を含んでいる場合があることも、厳密性を求める数学では嫌われたからだ」という点です。実際、事前確率を決めるには、こうしなければならないという決まりはありません。事前確率が分からなければ、その確率を1にしてしまうことも、0.5にしてしまうことも可能です。恣意的に決められるのです。

例えば、飛行機事故で墜落した機体を捜索する場合、「どこに墜落したか」を知りたくても、最初は確定的なことは言えません。あくまでも、さまざまなデータから「取りあえず確率の高そうな場所」を設定します。これを**事前確率**と考えるのです。これこそ、ベイズ理論の柔軟性であり、特徴でもあるのです。

そして現在では、この柔軟性がさまざまな場面で応用が利く[1]として捉えられているのも皮肉なものです。

多様性が求められる現代では、ガチガチに固められた従来の統計学（**頻度論**とも言う）よりも、感性や知見などを取り込めるという長所は、むしろ歓迎されたと言えます。ただ、計算はかなり複雑になることが多く[2]、辟易していたところに、コンピュータが一般化し普及したこともこの理論の再登場を後押ししたことは間違いありません。

いまでは、一昔前のスーパーコンピュータの性能をパソコンで利用できますし、ベイズ統計の処理をこなすソフトウェア[3]も充実しています。個人で利用し、検証することも可能になっています。これから先どのように変貌していくかは分かりませんが、ベイズの手法は統計分野で一つのジャンルを作り出したことは確実です。

結果から原因を探る

ベイズの定理はご覧の通り、非常に単純な構造をしています（再掲）。

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

しかし単純だからこそ、解釈の幅が拡がるとも言えます。例えば、Aを原因（cause）、Bを結果（effect[4]）とし、ベイズの定理を

$$P(\text{原因}|\text{結果})=\frac{P(\text{結果}|\text{原因})P(\text{原因})}{P(\text{結果})}$$

と見立てると、「ある得られた結果から、その原因の確率」を知ることができる、ということが分かります。

例えば、飛行機事故は、さまざまな要因から起き、調査によれば「操縦

1) 例えばスパムメールを排除したい場合、スパムメールは最初はスパムかどうか判定しにくいですが、ある程度、「この言葉はスパムに関係するだろう」という「単語」のデータを集め、そこから「その単語が含まれるメールがスパムであるだろう」という確率を決め、これを事前確率にします。これは主観的確率ですね。以後、その単語が含まれるメールを利用者に「スパムか、そうでないか」を判定させるなどして、精度を上げていくことができます。

2) ベイズ統計ではデータ一つひとつが重要な役割を果たすため、多くのデータを扱う場合には複雑になりやすいのです。従来の統計（頻度論）でデータが大事でないというわけではありませんが、データの扱いは、母数を引き出すための要素としての性質が強いので、データが主役のベイズ統計とはデータの扱い方が異なります。

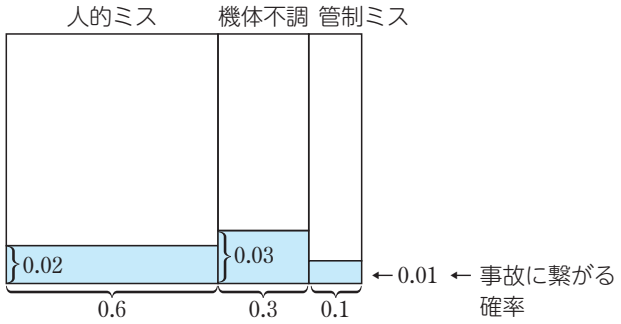
3) 例えば、R（アール）というソフトウェア。無料で提供されている統計処理ソフトです。統計処理全般を扱いますが、ベイズ統計ももちろん利用できます。後の章で述べるMCMC法を扱うためのパッケージも用意されています。

4) 結果と言うとresultが思い浮かぶかもしれません。これは一般的語彙で結果そのものに注目した語です。原因(cause)に対してはeffectが対応します。ペアと思ってください。

ミス（人的ミス）」「機体不調」「管制ミス」などがあります。ここではシンプルに考えるため、原因はこの３つだけであるとしましょう。そこで今回（あってはならないことですが）事故が起こったとして、その事故発生の原因が「管制ミス」であった確率を見てみます。データとして、次の表の通りだとします。

	発生確率	事故に繋がる確率
人的ミス	0.6	0.02
機体不調	0.3	0.03
管制ミス	0.1	0.01

これを図で表わすと、



となりますから、

$$P(\text{管制ミス}|\text{結果}) = \frac{P(\text{結果}|\text{管制ミス})P(\text{管制ミス})}{P(\text{結果})}$$

したがって、

$$P(\text{管制ミス}|\text{結果}) = \frac{0.01\times 0.1}{0.6\times 0.02+0.3\times 0.03+0.1\times 0.01} \fallingdotseq 0.045$$

という結果が出ます。この場合の「結果」は「事故発生」と読み替えてください。

「ある結果が得られ、それがある原因である確率」を知ることができるというのは、大きな成果です。シンプルな定理から、いやシンプルであるからこそ、さまざまな見方を与えてくれるのかもしれません。

データの更新で「原因の推測」も変わる――ベイズ更新

ベイズの定理は、得られた結果に対して、それがある原因によるものかどうかの確率を与えますが、その後、さらに新たな結果（データ）が得られた時、今度は、先に得た結論（事後確率）を使って更新することができます。

例えば、コインを３回投げて「表表表」と、３回連続して「表」が出たとします。さて、４回目に表が出る確率はどうでしょうか。

陰の声：そんなの簡単。コインの表・裏が出る事象は独立だから、たとえ直前に「表表表」３回出たとしても、次は$\frac{1}{2}$に決まっているでしょう。確率の常識だよ。

そうかなぁ？　では、10 回投げて、「表表表表裏表表表裏表」と出たとします。表が 8 回、裏が 2 回です。では、11 回目に表が出る確率はいかに？

陰の声：ううっ。独立試行だから、誰が何と言おうと、$\frac{1}{2}$だね。

それって、確信を持って言えますか？　もしかしたら、イカサマ・コインかもしれませんよ。ここまでくると、このコインの表が出る確率が$\frac{1}{2}$と考えることの方に不安が出てきます。コインには表と裏しかないので、試行回数が少ない間は表が連続して出ても不思議ではないけれど、試行回数が増えれば、表の出る確率は$\frac{1}{2}$に近づいていくのではないかと期待してしまいます。

そう確率は「期待」を表わしているのです。しかし、従来の考え方では、イカサマかどうかはさておいて、表の出る確率は$\frac{1}{2}$にしていました。そこで、従来の統計学（頻度論）では、この確率がどの程度の信頼度があるかをデータを元に検証する、という姿勢をとるのです。

ベイズの考え方は違います。「得られたデータが与えられたすべての情報」で、これを使って、逆にコインの表の出る確率を導き出すのです。データが得られる度にその確率は変化します。イカサマ・コインでなければ$\frac{1}{2}$に近づくでしょう。そうでなければそのコインの持っている特性の確率に落ち着くはずです。

このように得られるデータで確率を更新していくことを**ベイズ更新**（baysian update）と言います。

この構造が優れているのは、ある時点での結論を利用して、さらに最新の知見をも加えていくらでも更新ができるような柔軟な構造を持っていることです。この恩恵は大きいものです。従来型の統計処理ではなかなかそうはいきません。すべてのデータ（過去のものも最新のものも）を再度、計算させなければなりません。

数年であっても生データは散逸しやすいものですから、ベイズの方法論がいかに優れているかが分かるでしょう。

2-3 ベイズの展開公式

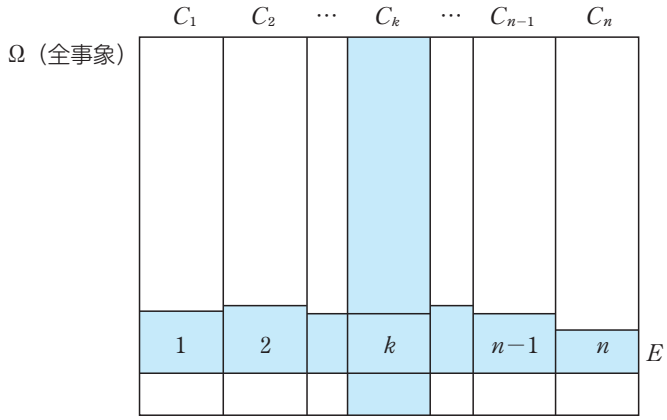
ここでは、ベイズの定理を変形し、さきほど保留にしておいた尤度（ゆうど）についても述べておきます。

まず、Ωを全事象とし、Ωに含まれる事象をC_1, C_2, …, C_nとします（考え得る「原因」としましょう）が、これは重なり合う部分がないので（排反）、$\Omega = C_1 \cup C_2 \cup \cdots \cup C_n = \bigcup_{k=1}^{n} C_k$ですが、$C_i \cap C_j = \phi (i \neq j)$なので、

$$\Omega = C_1 + C_2 + \cdots + C_n = \sum_{k=1}^{n} C_k$$

と書くことにします。

このとき事象E（「結果」と考えましょう）が得られた場合を想定します。言葉だけでは少々分かりづらいので、図を見てください。



ここで、事象Eは、図中で数字を振った部分、1、2、…、k、…、n（この部分はϕでもよい）の和になりますから、その確率は、$P(E) = P(\Omega \cap E)$