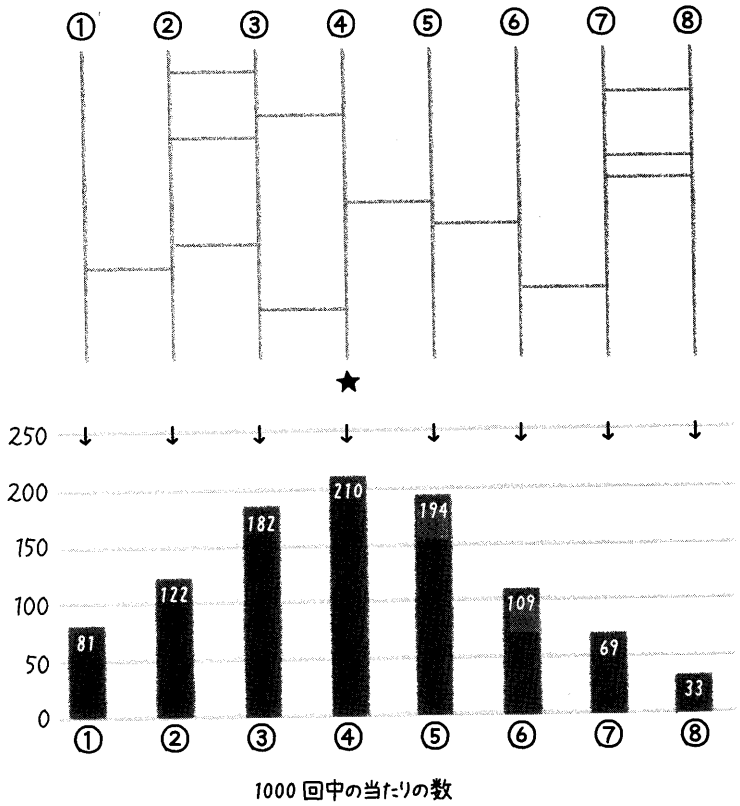
統計学が最強の学問である　西内啓　にしうち・ひろむ

あみだくじ
参加者 4 名。8 本の縦棒。左から 4 番目に星印をつけ、3 人に 4 ほんずつ横の線を引いてもらうというルール。



試しにこのルールで 1000 回繰り返したとして、縦棒ごとの当たる回数をシミュレーションしてみると図表 2 のような結果になる。一番当たる確率が高いのは当たりの真上である④で 1000 回申 210 回、つまり 21・0%の確率で当たる。次いでその右隣では 19・4%の確率で当たることになる。一方、最も低い右端では 3・3%の確率でしかない。

原因不明の疫病を防止するための学問を「疫学」と呼ぶが、世界で最初の疫学研究は 19 世紀のロンドンで、コレラという疫病に対して行なわれた。この疫学の中でも統計学は大きな役割を果たす。

・・・

当時のロンドンは産業革命の真っ最中で、農業で食べていけなくなつた人々が都会へ押し寄せ、工場で労働者として働くようになりはじめた時期であった。急激な人口の増加に都市の発達が追い付かず、狭く不潔な地域に粗末な家がひしめき、その家の中に貧しい人が押し込められ、下水も整備されないためにゴミや排泄物が庭や地下室、道端といったそこらじゆうに貯めこまれていた。当然その悪臭たるや悲惨なものだっただろうが、そうした「臭い地域」に住む臭い労働者たちの多くがコレラで死亡していたため、悪臭を取り除きさえすればコレラもなくなるのではないかと考えたのだ。

さらには、もっと果敢にこの汚物を取り除こうとした役人もいた。**彼は街中の汚物を片っ端から清掃し、下水を整備し汚物を川へ流せるようにする、という政策を取った**。この役人が活躍したのは主にコレラの一度目と二度目の大流行時の間の期間だが、彼らの努力にかかわらず、二度目の大流行時（死亡者約 7 万人）は、むしろ最初の大流行時（死亡者約 2 万人）よりも大量の死亡者を出している。

図表 3　スノウのレポート

	家屋の数	コレラによる死亡者	1 万軒あたりの死亡者数
水道会社 A を利用	40046	1263	315
水道会社 B を利用	26107	98	37

出所：ジョン・スノウ『コレラの伝染様式について』

だからスノウの提案したコレラ流行の解決策もごくシンプルなものになる。

「とりあえずしばらく水道会社 A の水を使うのを止める。以上！」

なおスノウがこの考えを発表してからおよそ 30 年後、ドイツの細菌学者ロベルト・コツホはコレラの病原体である「コレラ菌」を発見し、その結果、コレラが水中に生息することや、コレラ患者の排泄物に含まれること、そしてコレラ菌の存在する水を飲むことでコレラに感染することも証明された。

じつは水道会社 A と水道会社 B の違いは、前者がロンドンの中心を流れるテムズ川の下流から、後者はテムズ川の上流からそれぞれ採水しているというものだった。そして当時のテムズ川には、前述の勇敢な役人の努力によって大量にコレラ患者の排泄物が流し込まれていたのである。つまり**彼は、効率的にコレラ患者を拡大再生産させる社会システムを意図せず作りあげてしまった**のだ。

図表 7　標準誤差を算出する式

$$\text{標準誤差} = \sqrt{\frac{\text{全体の人数} - \text{サンプルの人数}}{\text{全体の人数} - 1} \cdot \frac{\text{真の割合}\left(1 - \text{真の割合}\right)}{\text{サンプルの人数}}}$$

この標準誤差というのがどういったものかというと、サンプルから得られた割合（たとえば失業率）に対して標準誤差の 2 倍を引いた値から標準誤差の 2 倍を足した値までの範囲に真の値が含まれている信頼性が約 95%、という値である。

たとえばサンプリング失業率が 25%という調査結果が得られ、その標準誤差が 0・5%だったとすれば、全数調査をした結果得

られるであろう真の失業率も 24%～26%の間にあると考えてほぼ
間違いない、ということを統計学者たちは 80 年以上前に証明して
いるのだ。

参考までに「**A／B テスト**」とは、デザインにせよ機能にせよ、A パターンと B パターンを両方試してみて比較する、という意味である。

・・・

比較されるのはたいていバナークリック率や商品の売上、有料会員への入会率といった利益に直結する数字についてであり、A パターンと B パターンのどちらが優れていたかという判断のもと、その後優れていたパターンがサイトに正式採用されるのだ。

なお、同時に 3 パターン以上試す場合についても「A／B／C テスト」などとは呼ばれず、「A／B テスト」と表現する。さらに余談だが統計学においてはこうしたデータの取り方のことを「A／B テスト」とは言わず**ランダム化比較実験**と呼ぶ（なお A パターンと B パターンの条件の変え方にランダムさが含まれていない実験は準実験と呼ぶ)。

スタッツ（統計値）

～～～～～～～～～～
スタッツ【stats】 の意味
《statistics（統計）の略》スポーツで、チームや個人のプレーの成績をまとめたもの。
～～～～～～～～～～

近年ウェブ界隈で「A／B テスト」と呼ばれ、統計家が長
年「ランダム化比較実験」と呼ぶものがどれだけ強力か、という話が中心的なテーマとなるだろう。

こうしたランダム化比較実験がどれだけ強力か、本節で説明するその最も大きな理由は、「人間の制御しうる何物についても、その因果関係を分析できるから」である。

統計学が「最強の学問」となったのはその汎用性の高さ、すなわち、政治だろうが教育だろうが経営だろうがスポーツだろうが、最速で最善の答えを導けるところにある、という話は先に書いたところである。そうした統計学の汎用性は、前節で紹介したようにどんなことの因果関係も科学的に検証可能な「ランダム化比較実験」によって大きく支えられている。

もう少し大げさに言い換えるならば、フィッシャーが打ち立てたランダム化比較実験という方法論は、科学の領域そのものを変えたと言っても過言ではないのである。

・・・
だが、フィッシャーのランダム化比較実験がなけ
れば、人類は**「誤差のある現象」を科学的に扱うことはできなかった**のである。

だが残念なことにこの武器はいつでも使えるわけではない。

世の中にはランダム化を行なうこと自体が不可能な場合、行なうことが許されない場合、そして行なうこと自体は本来何の問題もないはずだが、やると明らかに大損をする場合、という**3つの壁**がある。1つめの壁は「**現実**」、2つ目の壁は「**倫理**」、そして3つ目の壁のことを「**感情**」と呼ぶこともできるだろう。以下、それぞれの壁について説明していきたい。

「1回こっきりのチャンス」あるいは、あったとしてもせいぜ
い数回程度しかチャンスの与えられないもの自体を取り扱うことに対して、ランダム化しようがしまいが統計学は無力である。

ゴルトンが思いついたのは、この進化論の人間への応用である。彼は1883年に著した『人間の知性とその発達』において、「より環境に適した人種や血統を優先して、より多くの機会を与える」という、優生学の考え方を提示した。

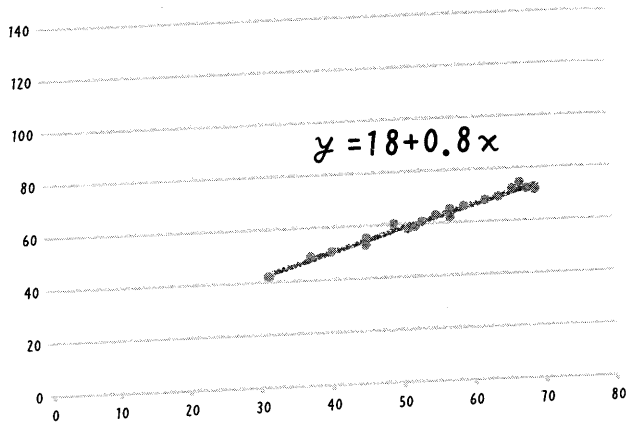
優生学とはすなわち、人間の知性は遺伝によって決定されるのだから、積極的に知性の低い人間は淘汰し、知性の高い人間ができるだけ多くの子孫を残すようにすれば人類の知性はどんどん向上していくから、これが人類の目指すべき正義ではないかというのだ。

こうした彼の思想はその後しばらく欧米で大流行することになる。その背景には、19世紀のヨーロッパに依然として存在する貴族と労働者の間の大きな階級格差があったのだろう。既得権益を守りたかった貴族たちにとって、ゴルトンの優生学というアイディアはとても都合がよかった。優生学に基づけば、貴族が恵まれているのは「環境に適した知能の高い血統」であるためであり、彼らが優遇され子々孫々まで栄えることは人類全体のためになるというのだ。さらに言えばこのゴルトンのアイディアは、富裕層から高い税金を徴収し貧しく弱いものを救済するような社会のやり方こそが人類の進化を止める諸悪の根源なのだ、と貴族が政府の税制と社会保障政策を批判する根拠にだってできる。

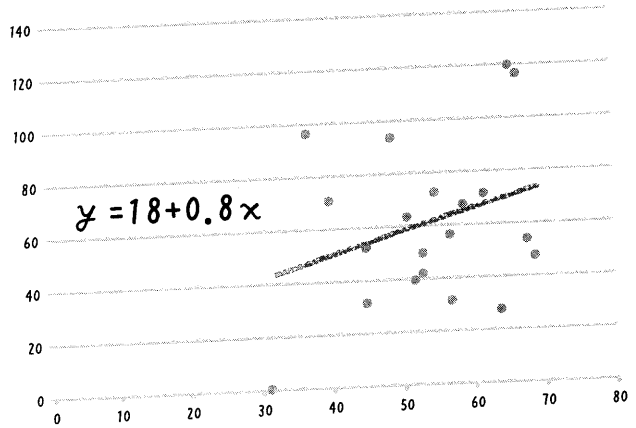
優生学の考え方が現在タブーとされているのはナチスが本気でこの思想を信じ、あるいは利用し、「劣等人種」の虐殺に繋がったからだと言われている。ただし、ほんの50年ほど前まではアメリカにおいてすら知的障害者や性犯罪者の遺伝子を残さないようにする「断種」が法的に認められていたのだから、ナチスだけを悪者にはできないだろう。このような法律はかつて世界中に存在しており、その背後には多かれ少なかれ優生学の考え方があった。

そのため現代的な統計学においては、実際に得られたデータ自体に「比較的大きな値を持つものもいれば小さな値を持つものもいる」というバラつきが存在しているだけでなく、**得られた回帰係数自体にバラつきが存在している**と考える。すなわち、仮に今後100回「たまたま得られたデータ」から回帰係数を計算したとしたら、「比較的大きな値となることもあれば小さな値となることもある」というバラつきを考慮しなければいけないのだ。

図表20 バラツキが小さい回帰分析の例



図表21 バラツキが大きい回帰分析の例



図表23 図表20のグラフの回帰分析の結果

変数	回帰係数の推定値	標準誤差	95%信頼区間	P値
切片	18	1.5	14.9 ～ 21.2	＜ 0.001
x	0.8	0.03	0.7 ～ 0.9	＜ 0.001

図表24 図表21のグラフの回帰分析の結果

変数	回帰係数の推定値	標準誤差	95%信頼区間	P値
切片	18	35	-55.5 ～ 91.5	0.61
x	0.8	0.6	-0.5 ～ 2.1	0.23

・P値：仮に回帰係数が0だった場合にデータのバラつきのせいだけでこれぐらいの回帰係数が推定されてしまう確率

やはり**慣例的には 5%を下回ると「さすがに回帰係数 0 と考えるのはキビシイ」と判断**される。
→だから回帰係数は有意。

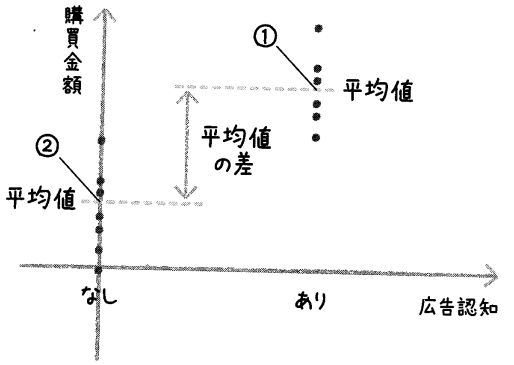
図表 24）p 値を見てみても 5%以下ということはなく、**仮に回帰係数の真値が 0 であっても切片については 61%、x の傾きについては 23%程度の確率で、この程度の結果はデータのバラツキによって生じてしまう**だろうという結果だ。

図表 25 一般化線形モデルをまとめた 1 枚の表

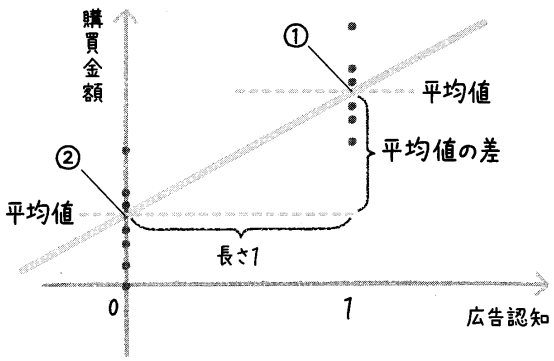
比較したいもの（結果変数）	分析軸（説明変数） 2グループ間の比較	多グループ間の比較	連続値の多寡で比較	複数の要因で同時に比較
連続値	平均値の違いをt検定	平均値の違いを分散分析	回帰分析	重回帰分析
あり/なしなどの二値	集計表の記述とカイ二乗検定		ロジスティック回帰	

さらに言えば、一番右の「複数の説明変数で同時に比較」したいときに使う手法を、1 つの説明変数しかない場合に使ってもかまわないし、その場合説明変数がグループ間の比較だろうが、連続値の多寡だろうがまったく問題がない。つまり t 検定をすべき場面で重回帰分析を行なおうとしても（ただしこの場合分析軸が 1 つしかないため重回帰分析とは呼ばれず回帰分析と呼ばれることになるが）、**カイ二乗検定を行なうべき場面でロジスティック回帰をしてもまったく同じ p 値が得られる**のである。だから関連性を分析する手法のほとんどが広義の回帰分析であると言えるのだ。

図表 26 t 検定の考え方



図表 27 回帰分析の考え方



このようにグループ間の違いを 0 か 1 かで表現しさえすれば「平均値の差」と「回帰係数」はまったく同じ備になるのである。

このように本来数値というわけではない「2 つのグループ」あるいは「二倍の変数」を、0 か 1 かで表現するやり方のことを**ダミー変数**と呼び、多くの論文でよく使われている。回帰分析の表に「男性ダミー」とか「高齢者ダミー」と書いてあれば、それはすなわち「男性なら 1・女性なら 0」とか「高齢者なら 1・未満なら 0」といった変数を回帰分析に使ったということである。

この節だけで、基礎統計学の教科書 1 冊分の手法を説明することができたのも、一般化線形モデルという素晴らしい枠組みがあってのことである。

このように学習者にとってわかりやすいコンセプトがネルダーとウェダーバーンによって提唱されてからもう 40 年ほど経つが、未だに一般的な教科書の記述に活かされていない、というのは個人的に残念である。

以前、ハーバードの大学院生と私的な統計学の勉強会を催した際にこのような枠組みの話をすると、「なんで今まで誰もこんな風に教えてくれなかったの？」というリアクションをいただいたこともあるので、我が国の統計教育だけの問題というわけでもなさそうだ。

・・・

なお余談だが、**一般化線形モデル**（Generalized linear model）の一部である t 検定・分散分析・回帰分析・重回帰分析といったもの（カイ二乗検定とロジスティック回帰は含まれない）を、一般線形モデル（General linear model）としてまとめるアイディアは、ネルダーとウェダーバーンに先んじて 1968 年にコーエンという名の統計学者によって提唱されている。

数学的な部分については専門書などを参照してもらうことにしてここでは説明を省くが、**もともと 0 か 1 かという二値の結果変数を変換し、連続的な変数として扱うことで重回帰分析を行なえるようにした、というのがロジスティック回帰の大まかな考え方**である。

ロジスティック回帰では、回帰係数をオッズ比つまり「約何倍そうなりやすいか」で示すということさえ知っていれば、結果の理解に問題はないだろう（なお正確には、**推定された対数オッズ比を変換しオッズ比の形で結果を示す**）。重回帰分析と同様に、回帰係数の推定値、標準誤差と信頼区間、それに p 値というのが読み取るべき値であり、その回帰係数の結果の読み方だけが少し異なるのだ。

図表 33「家庭や塾での勉強時間がゼロ」の生徒の特徴

	オッズ比の推定値	p 値
男子（ダミー）	0.77	0.05～0.10
読み聞かせ（ダミー）	1.11	0.05 以上
宿題（ダミー）	0.55	0.001 未満
文化階層下位（ダミー）	1.78	0.01 未満
文化階層上位（ダミー）	0.69	0.05-0.10
父・大卒（ダミー）	0.60	0.01 未満

また家庭の階層が下位グループであれば「勉強時間がゼロ」になる割合が 1.78 倍、父親か大卒であれば 0・60 倍と塾通いの有無以外にも家庭環境による学習習慣の影響は大きいのではないかという結果が示唆されている（図表 33）。

このような問題に対して有効な解決策が 1983 年に提案されている。

それはローゼンバウムとルービンという統計学者によって発表された**傾向スコア**あるいは**プロペンシティスコア**と呼ばれる手法である。この方法は主に疫学分野でランダム化が不可能あるいは困難な因果関係の特定に重宝されてきた。

傾向スコアとは、興味のある二値の説明変数について「どちらに該当するか」という確率のことを言う。**「どちらに該当するか」という傾向を示す値だから傾向スコア**というわけである。なお傾向スコア自体は、すでに紹介したロジスティック回帰によって簡単に得ることができる。

～～～～～～～～～～～～～

傾向スコア例

https://www.jstage.jst.go.jp/article/tenrikiyo/19/2/19_19-008/_pdf

～～～～～～～～～～～～～

ちなみに形態素解析とは異なるアプローチとして、辞書を使わない N-Grm と呼ばれるやり方もある。これはすなわち機械的に重複を許した N 文字ずつの文字列を切り出し、そこから求める単語を探すというやり方だ。もし N が 5 だとすれば「統計学が最強の学問である。」という文章からは、「統計学が最」（1－5 文字目）、「計学が最強」（2－6 文字目）「学が最強の」（3－7 文字目）…「間である。」（最後の 5 文字）といった 5 文字ずつの Gram が生成される。

・・・

Google で、あまり一般的でない単語を検索しても該当するページにヒットするのは、その背後に膨大な量の N-Gram データが存在しているからだ。

多くの統計家が二値の結果変数に対
してロジスティック回帰を用いる一方で、**計量経済学者はプロビット回帰という手法を好んで使う**。プロビット回帰のほうがロジスティック回帰よりも数理的にキレイではあるのだが、推定された回帰係数がロジスティック回帰のオッズ比のように「約X倍になる」という形にはならず、直感的な解釈がしにくいという難点を持っている。

ここまでに社会調査や心理統計学、データマイニングや計量経済学といったさまざまな分野における統計学に対する考え方の違いを紹介したが、最後に分野をまたいで存在する「確率自体の考え方」についての対立を紹漉しょぅ。

それが**頻度論者かベイズ論者か、という対立軸**である。両者の違いを一言で表すとするならば、「事前に何らかの確率を想定するか」「しないか」と言い換えてもいいかもしれない。

両者の違いを理解するために、たとえばここに 2 種類のコインがあったとしよう。一方は表が出る確率も裏が出る確率も 5 分 5 分の「本物のコイン」で、もう一方は表が出る確率が 8 割、裏が出る確率が 2 割の「イカサマのコイン」だ。両者は見た目や重さなどからはまったく区別がつかないものとするが、何回か投げた回数を集計・分析し、どちらのコインであるかをそれぞれの立場から判断してみよう。

◆頻度論者

まず**このコインが本物だと仮定**する。

そしてその仮定のもとで 10 回中全部が表になる確率を計算するだろう。

すなわち、「2 分の 1 の確率で出る表が偶然 10 回全部出る確率は 2 の 10 乗分の 1、つまり 0・10%しかない」ということだ。この 0・10%という確率はいわゆる p 値と呼ぶものである。

つまり、このような確率の奇跡が起こったと考えるよりは、そもそもの「このコインは本物」という**仮定を「考えにくい」と捨て去ったほうが理にかなっている**という判断を行なうのだ。

次に**「このコインはイカサマのコイン」だと仮定**したらどうなるだろう？先ほどと同様の計算を行なえば、「80%の確率で出る表が偶然 10 回全部出る確率は 10・74%程度」ということになる。p 値が 10・74%程度なら別に奇跡的と言うほどではない。

だからこの仮定を捨てきることはできない。

一方の「本物のコイン」という仮説が捨て去られ、他方の「イカサマのコイン」という仮説は捨て去ることができないのであれば、すなわちこれはイカサマのコインだと考えたほうが妥当だということになる。

◆ベイズ論者

コインが本物だった場合に 10 回全部が表になる条件付き確率も、イカサマのコインだった場合に 10 回全部表になる条件付き確率も、頻度論と計算方法は変わらずそれぞれ 0・10%

あるいは10・74%であるが、その次の計算方法がベイズでは少し異なる。

すなわち、本物だった場合、イカサマだった場合のそれぞれの状況において、事前確率と条件付き確率の掛け算を行なうのだ。この場合、以下のよう計算が行なわれる。

① 本物である事前確率×本物である場合に10回全部が表になる条件付き確率
＝50%×0.10%
＝0.05%

② イカサマである事前確率×イカサマである場合に10回全部が表になる条件付き確率
＝50%×10.74%
＝5.37%

また、コインが本物かイカサマかという選択肢しかない以上、どんなときでもコインが本物である確率とイカサマである確率を合計すれば必ず1になるはずである。それは「10回全部が表になった」という結果が得られた時点においても例外ではない。つまり、①と②の合計値は「1」になるはずなのだ。

そこで、①と②の合計値である5・42%という値で、①、②それぞれの値を割ってやる。
そうすると「10回全部が表になる場合にこのコインが本物である確率」は、

① ÷（①+②） ＝ 0.05÷5.42 ＝ 0.90%

であり、一方「10回全部が表になる場合にこのコインがイカサマである確率」は、

② ÷（①+②） ＝ 5.37÷5.42 ＝ 99.10%

と計算することができる。

すなわち、10回全部が表になったデータから、これは**99・1%の確率でイカサマのコインであるとベイズ論者は判断する**のである。

このように事前確率とデータに基づいて算出された確率のことを事後確率と呼ぶ。
このような計算結果をまとめると図表53のようになる。

図表53 ベイズ的な確率の計算①

	本物	イカサマ	合計
事前確率	50.00%	50.00%	100.00%
条件付き確率	0.10%	10.74%	
事前確率×条件付き確率	0.05%	5.37%	5.42%
事後確率	0.90%	99.10%	100.00%

これまでに紹介した分野別の傾向で言えば、**社会調査、疫学、生物統計学、心理統計学分野には頻度論者が多く**、計量経済学者にはベイズ論者が増えており、データマイニング屋は特にどちらとも意識していないがベイズ論者寄り、といったところだろうか。

ベイズ論者からすれば、もし仮に最初の事前確率として90%これは本物であると考えていても、「10回全部表」というデータから導かれる事後確率は「92・43%イカサマ」であり、事前確率による影響は小さ

いと主張したいだろう（図表 54）。

図表 54 ベイズ的な確率の計算②

	本物	イカサマ	合計
事前確率	90.00%	10.00%	100.00%
条件付き確率	0.10%	10.74%	
事前確率 × 条件付き確率	0.09%	1.07%	1.16%
事後確率	7.57%	92.43%	100.00%

一方で、たとえば**3回しかコインを投げられなかった状態で頻度論は「どちらかはわからない」としか判断できないが、ベイズ的な考え方であれば、少なくともどちらである可能性が高いかは判断できる**のだ。

そのため「間違いが許されない」という保守的な判断が求められる分野ほど頻度論に基づく傾向にある。

図表 57 代表的な論文データベース

名前	URL	分野
ERIC	http://www.eric.ed.gov/	教育学
PsycINFO	http://www.apa.org/psycinfo/	心理学
Econlit	http://www.aeaweb.org/econlit/	経済学
Pubmed	http://www.ncbi.nlm.nih.gov/pubmed	医学
JSTOR	http://www.jstor.org/	綜合

また、Google の提供する Google Scholar（http:// scholar. Google.co.jp/）も便利な文献検索サービスである。英語だけでなく日本語の文献を探すこともできるし、Google 的な「重要度の高い文献ほど上位に」といった機能も実装されているようだ。

p-value （p 値）
significant (有意つまり誤差の範囲でないこと)