

医療におけるエビデンスとP値

医学統計解析グループ（代表：前谷俊三）

天理よろづ相談所 医学研究所

証拠に基づく医療においてはランダム化比較試験（RCT）は最も確かなエビデンスを提供するとみなされている。P値はRCTの結果を要約し、新治療を行政機関が認可したり、臨床医がどの治療を患者に適用すべきかを定める基準として広く使用されている。しかし統計学者の中にはP値が論理的妥当性をもち使用に値するものかに疑念を抱くものもいる。本論文では臨床医と患者の視点からP値の意味を再検証する。

P値の弱点は以下の通りである。1) P値とは、真偽の定かでない仮定の下で求めた条件付確率であるので、非専門家にわかりにくく、かつ臨床的に適切な尺度とはいえない；2) たとえP値が同じでも、症例数に応じて証拠としての重みが変わる；3) 臨床医や患者が知りたいのは治療効果の差が有意かどうかよりは、差の範囲がどの程度かであるので、P値よりは信頼区間のほうがその問いに答えられる。以上の問題点はP値の新しい解釈によりある程度は克服できる。すなわち、P値とは、平均すれば他の治療よりよく効く治療を施行すると、反って悪い結果をきたすリスクと解釈できる。しかし通常P値は並行比較デザインによるRCTに基づき、群の平均を比較して得られるので、それは個々の患者の得失よりも集団の利害を反映する。理想からいえば、同一患者で治療効果を比較することにより、個人にとってのリスクを推定すべきである。

キーワード：P値，ランダム化臨床試験，証拠に基づく医療，治療効果の大きさ，悪い治療を受けるリスク

はじめに

現今の医療改革において主要な三つ柱を挙げるとすれば、第一は証拠に基づく医療

（EBM）である¹。過去に証拠と信じられたものが学問的にみて証拠に値するものかを洗い直し、より確かな証拠に支えられた質の高い医療を実現する。第二は医療の透明性である。これは社会にとっては評価可能な医療であり、医療提供者にとっては説明

【別刷請求先】

〒632-8552天理市三島町200
天理よろづ相談所 医学研究所
前谷俊三

責任を果たす医療である。² これによって医療の効果、効率、安全性、経済性が客観的に評価され改善されなければならない。第三は患者中心の医療である。医療情報が患者のわかる言葉で伝えられ、患者が医療提供者と情報を共有して、自分の利害に合う医療が行われることを納得しなければならない。

さてEBMにおいて強いエビデンスとは、疾患の原因や治療の機序を解明して得られる証拠ではない。適切な結果(アウトカム)に支えられた証拠である。一例を挙げれば、ある疾患の原因が特定の細菌によるものであり、ある抗生剤がその細菌の細胞壁合成を阻害することが実験的に確かめられたとする。この事実はその疾患に薬剤を使用するための強いエビデンスとはいえない。実際にその薬を患者に使ってみて、患者が本当によくなることが強いエビデンスなのである。これは専門知識がなくても患者自身が実感でき、かつ評価できる。しかしEBMでは結果がよければ全てよしとはしない。有利な条件を備えた患者が治療群側に偏っていたために有効と見誤ることがあるからである。EBMでは個人的臨床経験を強いエビデンスとみなさないのはこのためである。

この誤りを統計的手法によって最小限に抑えようとするのがランダム化臨床試験(RCT)である。RCTでは医療の評価基準をアウトカムにおいている。その点では病因や治療の作用機序など医学的専門知識がない者でも医療の評価ができる。しかし他方ではその統計的手法が非専門家にとって極めて難解であり、これが評価を妨げる壁として立ちはだかっている。

有意性の検定によく使用されるP値もそ

の例に漏れない。これはRCTにおいてある治療が有効である証拠の強さを示す尺度として使われてきた。また実際にもP値は行政機関が治療法を許認可したり、臨床医がどの治療を患者に施行すべきかを決める際に重要な基準ともなっている。しかしP値は統計学者の間でも論議の多い値である。さらにまた医療改革の時代においては専門家だけでなく、臨床医や患者や社会がその意味を理解し、それぞれの立場からその妥当性を検証する必要がある。しかし現状ではどれだけの関係者がそれを正しく理解しているだろうか。本稿では、まずP値に関する我々の考え方を卑近な例を挙げて説明し、ついで臨臨床的視点からその妥当性を再検討する。なお数式は極力避け、必要な場合のみ括弧内に記載した。

毒物混入事件の犯人とP値

ある会場で研究発表会に続いて立食パーティーが行われた。宴がたけなわの頃、多数の出席者が重篤な中毒症状で倒れ、会場は大混乱に陥った。調査の結果、出された料理から毒物が検出された。それは、料理が副室に一時的に置かれている間に何物かによって混入されたことがわかった。幸い副室の前には防犯カメラが設置されていたので、副室に入った者を全て確認することができた。その結果、副室に入った3人が容疑者として拘束された。この事件では参加者が副室に入ったかどうかをカメラで確認することが、重要な手がかりとなる。もし「犯人がどうかをイエスかノーで答えよ」と言われれば、目下のところ最善の基準は、「カメラに写れば犯人、写らなければ無実」である。これを仮に判定基準とよぶことにする。

しかしこの判定基準だけではどれだけの証拠になるだろうか。その証拠の強さを測る一つの尺度がP値である。その考え方は以下の通りである。すなわち「もし容疑者が無実だとしても副室に入る可能性はある。その可能性はどれだけか」と考える。それがP値に相当する。尋問の結果、容疑者の一人は「便所に行こうとして間違っただけで副室に入った」と答えた。仮に便所に行くためには必ず副室を通り抜けねばならないとすれば、この判定基準だけで犯人とするのは不公平である。逆に副室が一番奥にあり、犯行以外の目的で其処に入ることは殆ど考えられないならば、副室に入った者全ては犯人の可能性が高い。つまりP値が小さい程容疑者が犯人である確率が高くなる。

実際にPがどの程度かを知るためには、会場に来た人の中でどれだけが副室に迷い込むか、日を改めて(犯人のいない時に)調べればよい。もしPが例えば5%以下、すなわち、無実の人が副室に迷い込む割合が20人に1人かそれより少なければ、カメラに写った人を犯人とする。これが有意性の検定の原理であり、 $P=0.05$ ($1/20$)を有意水準という。

もしこの検定法を聞けば、「医療ではこの程度のことをエビデンスとするのか。司法ではこれだけでは犯人とはしない」という者がいるかもしれない。しかし逆に「犯人を野放しにして犯行が繰り返されるぐらいなら、P値はもっと大きくてもやむを得ない。それが社会全体の利益だ」という意見が出るかもしれない。いずれにせよP値は統計学者だけで論議すべき問題ではなく、すべての関係者が関わり合うべき問題である。

しかしその前にP値に関して非専門家が陥りやすい誤解を知っておかねばならない。もしPが0ならば容疑者が犯人である可能性は100%であるが、Pが1だからといって、容疑者が100%無実だとはいえない。というのはPが1とは無実の人でも必ず副室に入るというだけに過ぎず、一方、犯人も必ず副室に入るからである。P値とは容疑者が無実かまたは犯人であるかの可能性を測る尺度ではない。カメラによる判定基準が、どれだけ頼りになるかを測る一つの尺度なのである。裏を返せば、P値とはその判定基準に頼ればどれだけ誤りを犯す危険があるかを表す一つの尺度である。

ここで一つの尺度と断ったのは、Neyman-Pearson統計学では尺度は二つ、誤りも二つあるからである。このあたりから話が込み入ってくる。それは二つの尺度のためだけではない。「何々とすれば」という仮定での話であるからである。既に述べたように、第一の誤りとは、もし容疑者が無実とすれば(この仮定を帰無仮説という)、これを犯人と見誤ることである(この誤りを第1種の過誤またはエラーという)。Pが小さいということは、この判定基準を使えば無実の人に濡れ衣を着せる危険が少ないことを表す。これに対して第2種の過誤(エラー)とは、容疑者を犯人と仮定した場合(対立仮説)これを見逃すことであり、この誤りをしても重大な結果を招くことになる。本事件で使った判定基準では犯人を見落とすことはない。しかし仮にカメラが副室の入り口全体を写し出していなかったならば、見落としの危険は0ではない。カメラに写らなかったからといって100%無実とはいえない。そうなればP値という尺度

だけで判定基準を評価するのは片手落ちとなる。見落とし（エラー）の確率を P 値に対して例えば Q 値と命名して、 P, Q の 2 本立てで評価する必要がある。実際には Q という名前はないが、 $1 - Q$ には検出力という名前が付けられ、 P 値と検出力という二つの尺度で評価が行われている。

それではどちらのエラーの防止を重視すべきであろうか、現行の RCT では エラーがより小さくなるように設計されている（例えば $\alpha = 0.05$, $\beta = 0.2$ ）。いわば犯人を見逃すことよりも、無実な人に濡れ衣を着せることを避けようとする姿勢が窺われる。ここにも全ての関係者が関与して議論する余地が残っている。

治療の有効性の判定における P 値

上の例で容疑者を治療法に代え、有罪を有効に代えれば、治療効果の判定における P 値の意味がわかる。すなわち、 P が 0 ならばその治療は確実に有効といえるが、 P が 1 でも、絶対に無効（無治療と同等）とはいえない。 P 値は有効、無効の確率を測る尺度ではない。例えば $P = 0.05$ であれば、無効である確率が 5% であり、95% は有効という意味ではない。 P 値が同じ 0.05 でも、無効である確率は種々の値になる。これからいえることは、判定基準に基づけば、無効の治療を有効と見誤る条件付確率が 0.05 であるということだけである。

それでは判定基準として何をを使うかという、ある治療が無治療（または偽薬）と比べて治療効果をどれだけ高めたかという成績である。この平均的治療効果の増加分（差）を治療効果の大きさ（effect size）という。effect size という判定基準がどれだけ確

実かを測る一つの尺度が P 値である。例えば、ある治療が無治療に比べて 5 年生存率を 10% 増加させ、 $P = 0.01$ で有意であったとする。その意味するところは「effect size 10% を有効性の判定基準に使えば、効かぬ治療を効くと間違える危険（エラー）は 1% である^{注1}」ということである。ただしこの判定基準を使えば、効く治療を効かぬと見誤る危険（エラー）がどれだけかは P 値をみてもわからない。これを測るのは検出力である。ただ一般にいえることは α を小さくするような判定基準を使うほど、検出力は低下する（エラーを犯す危険は増加する）。これは無実な人を検挙しないようにするほど、犯人を見逃す危険が増えることと同様である。

^{注1} 実際に使用する判定基準は effect size ではなく、 α レベル（有意水準）であり、通常 5% に設定される。これによって エラーは 5% 以下に抑えられる。

検査の信頼性と疾患予測値

P 値と治療の有効性とはともすれば混同しやすい。この違いは検査の信頼性と診断的中率（予測値）を例にとれば理解しやすい。表 1 に示すように両者は分割表を縦に読むか、横に読むかの違いにたとえられる。検査の信頼性を評価するためには、表 1 を上から下に辿って感度と特異度を求める。感度とは疾患をもつ人を検査にかければ、どれだけの割合で疾患がみつけれられるかを表し、検出力に相当する。一方、特異度とは疾患をもたない人の中でどれだけの割合で検査が陰性に出るかを表す。これが $1 - \beta$ 値に相当することは図 1 と表 2 をみればわかる。感度と特異度が高いほど診断的中

表1. 検査の信頼度と診断的中率

		疾患		
		あり	なし	診断的中率
検査	陽性	真陽性数 TP	偽陽性数 FP	陽性予測値 TP/(TP+FP)
	陰性	偽陰性数 FN	真陰性数 TN	陰性予測値 TN/(TN+FN)
検査の信頼度		感度 TP/(TP+FN)	特異度 TN/(TN+FP)	

表2. P値と治療効果の差の確率

		治療効果の真の差		
		あり	なし	差の確率
判定 従 え 基 準 に ば に	差あり	真陽性数 TP	偽陽性数 FP	差のある確率 TP/(TP+FP)
	差なし	偽陰性数 FN	真陰性数 TN	差のない確率 TN/(TN+FN)
判定基準の信頼度		検出力(1 -) TP/(TP+FN)	(P値*) FP/(TN+FP)	

* effect size を判定基準とした場合 $\alpha = P$

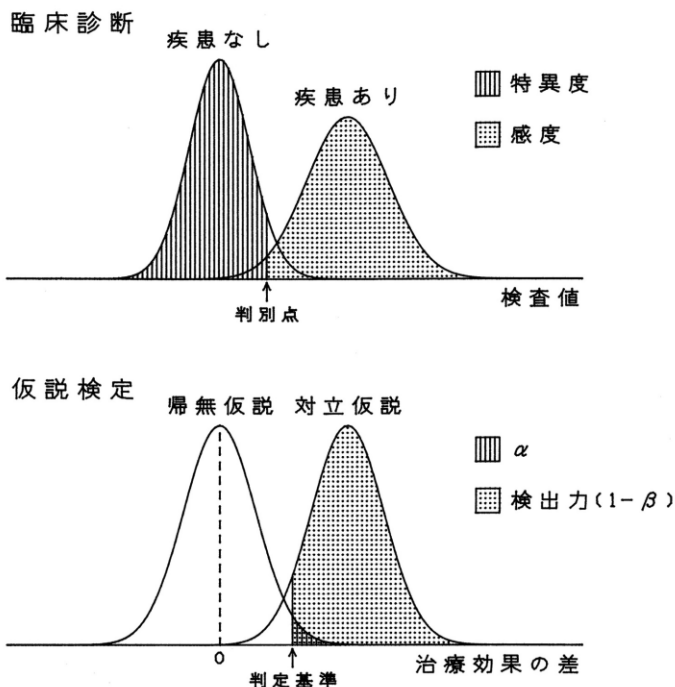


図1. 臨床診断と仮説検定の類似性

率は上がる傾向はあるが、それだけでは決まらない。例えばエイズの診断で感度99.9%、特異度99.9%という極めて正確な検査を被験者にしたところ陽性の結果が出たとする。だからといってその被験者がエイズに感染している確率は99.9%ではない。本当は約17%に過ぎず、むしろエイズでない可能性のほうが強い。これは日本でエイズ感染者の割合(有病率)が5千人当たり1人と仮定すれば、Bayesの定理から容易に導くことができる(陽性予測値 $=1+(1-\text{有病率})/\text{有病率} \times (1-\text{特異度})/\text{感度}$)¹⁾。患者を前にして臨床医が知りたいのは感度や特異度という検査の性能ではない。検査結果が陽性に出た場合、どれだけの確実さで「あなたはエイズです」と言ってよいか(陽性予測値)、また検査が陰性の場合どれだけ自信をもって「あなたは大丈夫です」と言い切れるか(陰性予測値)である。患者が知りたいのもこの診断的中率であり、これは表1を左から右に辿れば求められる。実際にこれを感度と特異度から求めようとすれば、その値は有病率(事前確率)に大きく影響される。

同様に治療についても我々が知りたいのは「どれだけの確率でその治療が他の治療よりよく効くといえるか」である。この疑問に対してP値で答えようとすれば、それは本来横に読むべき表2を、縦に読んで出した不適切な(irrelevant)答えかもしれない。これは「特異度が0.999(P値に換算して0.001)の検査が陽性に出たのだから、エイズに間違いない」というようなものである。感度と特異度は診断の確かさを測る尺度ではなく、検査法の確かさを測る尺度である。同様に、P値も仮説の真偽を判定する

尺度というよりは、真偽を判定する基準を評価するための尺度といえる。いわば尺度を評価するもう一つの尺度という解釈もできる。これはP値を帰無仮説に相反する証拠とみなすFisher本来の考え方とは異なる。

以上を数学的に表せば、P値とは「治療効果に差がないとすれば」という仮説Hの下でDという結果(データ)が観察される条件付確率Pr(D|H)である。しかし真偽の定かでない仮定の上での話は難解であり、治療の損得と結びつけにくい。患者や臨床医が本当に知りたいのはその逆の確率Pr(H|D)である。つまり「観察結果に基づけば仮説はどの程度確かか」である。後述するように、新しい解釈によればP値はPr(H|D)に読み変えることができる。

症例数(サンプルサイズ)とP値

1976年Petoら世界的に高名な統計学者10名が医師向けにRCTの解析法を解説した³⁾。そこで彼らは次のように述べている。「大きい臨床試験で得られたP値は、小さい試験で得られた同じP値と比べて、治療効果に差があることを示すより強い証拠となる」。これは「大きいことはいいことだ」と解釈され、大規模臨床試験を支持する一つの根拠ともなった。ところが1993年Freemanは全く反対の結論を述べた⁴⁾。すなわち、「P=0.05という値は、少数例から得たものならば強い証拠となるが、多数例から得たものであれば、その証拠は常に極めて弱く、極端な場合は差がないという逆の証拠になる」。RoyallはPeto等の説はP値を0.05より大きい小さいかの2値として扱った場合にのみ正しいと述べている⁵⁾。いずれにせよ証拠の強さが症例数に影響されることは

大きい問題である。多くの有識者はP値を使った有意性検定をとらえどころのない (elusive) 概念の顕れとみている。⁵

信頼区間の推定とP値

医学研究において治療効果を報告する場合、仮説検定に拠るとある治療が「効くか効かぬか」と二者択一的な決定となり、情報の喪失が起きる。これに対して「どの程度効くか」を信頼区間で表すべきであるという見解は1980年以前にもみられた。⁶ しかし1980年代の後半から、effect sizeをその信頼区間で表すべきであるという勧告がトップジャーナルに屢登場した。⁷⁻¹⁰ 中には有意性検定のもつ全ての情報は信頼区間に含まれるとして「P値の出番は終わったか」という論説まで現われた。¹¹ またEBMにおいても信頼区間を記載することが推奨された。¹ こうして信頼区間はP値よりも有用かにみえたが、信頼区間はP値のもつ問題を全て解消したわけではない。⁴ 信頼区間を表すためには2つの数値が必要である。同じ2個の数値を使うならば、effect sizeとP値を併記したほうがよいのではないだろうか。実はこの論議は以下に述べるP値の新しい解釈によってP値が復活するかどうかにかかっている。

P値の新しい解釈

1990年AnscombeはP値には全く別の解釈があることを示した。¹² 2つの治療の中で平均治療効果の大きい方を選ぶことは期待値基準と呼ばれ、決定分析で常用されてきた。^{13,14} しかしこの選択基準では常により結果が得られるとは限らない。彼はこれが裏目に出るリスク。つまりよいはずの治

療を選べば反って損をする確率がP値(片側)に一致することを示した。1999年我々もこの報告を知らず、同じ確率をエラー(後にP')という名で発表した。^{15,16} これを以下にグラフで説明する。

今ある病気が一人の患者で繰り返し起こると仮定する。これに対して2つの治療A, Bを順序をアランダムにして同一患者に施行し、どちらが早く病気を治せるかを比較した。図2上段左は治癒日数の差(B-A)を多数の患者で求めて、その分布を描いたものとする。これが母集団の分布に近似すると仮定し(第三の仮説)、かつそれが正規分布と仮定する。平均するとAはBよりも10日(矢印)治癒日数を短縮する。しかしその個人差は大きく(標準偏差=10)、差が0より小さい、つまりBをした方が早く治る患者が縞模様の面積だけある。これをP' とすると、約38%となる。すなわち、平均ではAのほうがよい治療であるにも拘わらず、Aをすると損をする患者がP' だけいる。

図2の上段右は個々の患者の治癒日数の差ではなく、n人の患者を一組にして差の平均を求め、これを何組にもわたって繰り返し、その分布を描いたものである。nを10人から20人、40人、80人と増やして行くと、当然のことながら分布の幅が狭くなる(平均の標準誤差=標準偏差/√n)。その結果P' 値も小さくなり、16%、8%、2% から0.2%となる(打点部)。つまりAの治療を受けて損をする危険が減少する。

一方、図2下段はA, Bの治療効果に差がないと仮定した場合(帰無仮説)の治癒日数の差の分布である。当然のことながら、これはnの大きさに拘わらず、0を中心とし

て分布する．その分布の広がりには図2上段と同様 n が大きいほど狭くなる．しかし同じ n 同士ならばその広がりも上段と同じ(等分散)と仮定している．さてこの帰無仮説の下での P 値(片側)とは effect size(矢印)より外側の分布曲線の面積(縞模様または打点部)となる．その理由は既に述べたように effect size を有効か無効かの判定基準(境目)としているためである．上段と下段を見比べれば, P 値と P' 値が一致することが視覚的に捉えられる．(最初の分布が正規分布でなくても n が大きくなれば $P = P' = (-$ 標準化 effect size)となる)．

P 値対 P' 値

新しい P 値の解釈についてはそれほど注目も払われず, まだ教科書にも記載されていない．しかし医療提供者や消費者にとっては画期的なことである．何故ならば, 帰無仮説を用いた従来の P 値の説明は難解である．たとえ理解しても P 値とはある種の誤りを犯す条件付確率である．その誤りがどれだけの損失をもたらすかはわからない．新しい解釈では P' 値は患者が損失を蒙るリスクそのものを表す．これならば, 医師, 患者, 社会も理解でき, また強い関心を寄せるはずである．Anscombe はこれを称して「 P

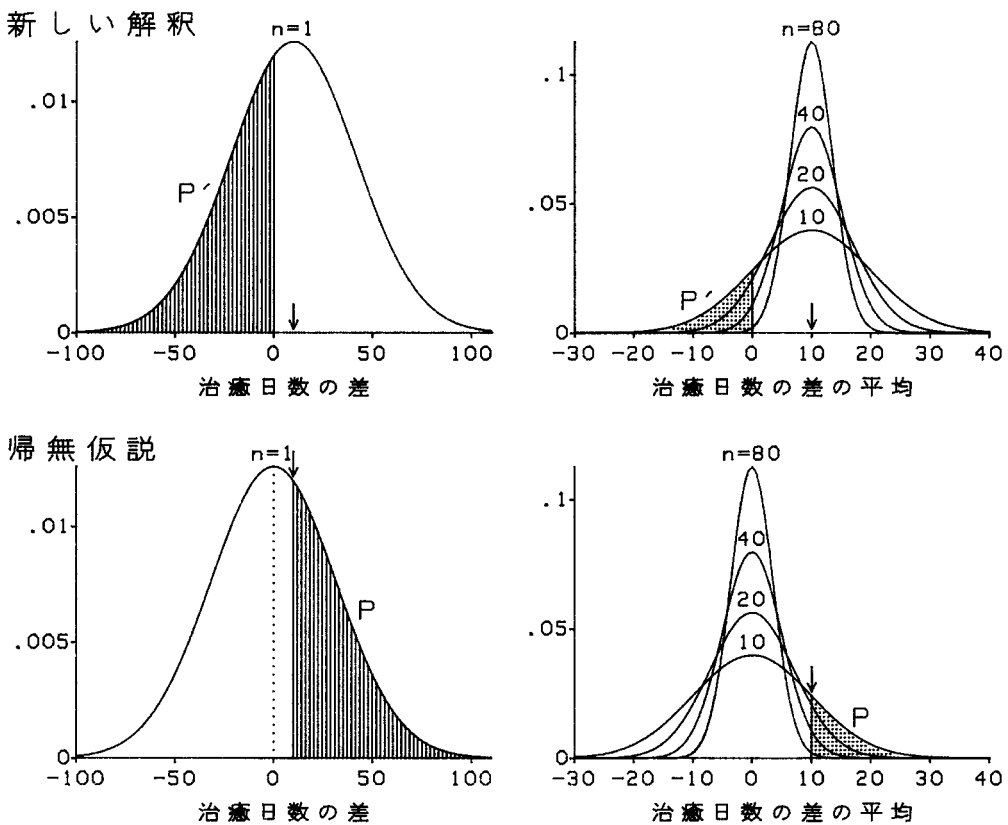


図2. 帰無仮説に基づく P 値とその新しい解釈(P' 値)
矢印は effect size を表す

値で臨床試験の結果を伝えること自体は悪かったとはいえないが、その理由が間違っていた(the right thing is done for the wrong reason)」と述べている。¹² いわば「嘘から出たまこと」がP値である。

従来のP値は「治療A,Bの効果は同じか、それとも差があるか」という設問に答えようとしている。もしこれが「DNAではチミンとアデニンのモル数は同じか」というように現実であり得る問いであれば納得できる。しかし、いやしくもRCTで比較しようとする二つの治療が全く等しい効果をもつことは考えにくい。もしそうならば最初から効果に差があることを前提として、さらに突っ込んだ設問が考えられる。すなわち、「平均的な差(effect size)はどれだけか」と、「どれだけの割合の患者ではAの方が効き、どれだけではBのほうが効くか」という問いである。P値は「差があるかないか」を測るには不適當な尺度であるが、後の問いに答えるには適切な尺度である。医師や患者が知りたいのも後者である。ある治療が他の治療の効果と同等であっても、もともとである。治療する者、される者が共に抱く懸念は「やろうとする治療は裏目に出ないか」である。またその懸念が現実のものとなることも稀ではない。ところがRCTの検定では、この懸念に対応する仮説もその答えも用意していない。¹⁷ これでは医師が患者に対して十分な説明責任を果たしているとはいえない。この意味でもP'の新しい解釈の方が臨床的には望ましい。

ここでもう一度図2の上段と下段を比べると、上段の分布曲線は実際に観察されたデータからありのままに描くことができる。一方下段の分布曲線は仮定に基づく曲線で

ある。すなわち、曲線を規定する母数の中で、平均(位置母数)を0とすることは帰無仮説からみて妥当であるが、分散(尺度母数)はそれだけでは決められない。そこで等分散の仮定に従い、上段の曲線と同じ値を当てはめているが、この仮定に問題がないとはいえない(特にn=1の場合に上下の曲線が同じとみなすことには無理がある)。¹⁶ 従ってその曲線から求めたP値が本来求めている値かどうか検証できない。臨床医が抱く疑問は「それでも帰無仮説を立て、仮説検定をするメリットがあるのか」である。

我々は最初P'をエラー(第3の過誤)と呼んだ。¹⁵ これはSchwartzとLellough¹⁸が命名したもので、「本当はAがBより効くが、両者を取り違えてBのほうがよく効くと見誤ること」と定義される。ところがデータをありのままにみれば、BがAよりよく効いたという事実をエラーといってしまうだろうか。ここには「治療対象は一つの高質な母集団からなる」という前提がある。しかし実際の治療対象にはAが効く集団やBが効く集団など、異質な集団が混じっていても不思議ではない。特に最近では薬物治療における効果の違いがpharmacogenomicsで説明されるようになり、また長期生存を達成するはずの根治手術で反って癌の増殖が加速される機序が分子生物学的に解明されようとしている。¹⁹ 治療法と患者の交互作用はそれほど稀とは思われない!データをありのままに」という理由から、我々はエラーをP'という名前に変更した。PとP'は同じ値でもそのコンセプトや背景は全く異なる。

集団にとっての P' 値と個人にとっての P' 値

P 値にせよ P' 値にせよeffect sizeが同じでも n が増えるとその値が小さくなることは図2で述べた。これは具体的に何を意味するのだろうか。例えばある会社の従業員に上記の病気が集団で発生し、会社が蒙る損失は病気欠勤の延べ日数に比例すると仮定する。もし1人1日の欠勤で1万円の損失とすると、治療Aを選べばBに比べて平均10日欠勤日数を短縮するので、10万円の利益となる。逆にAを選んで反って会社が損をするリスクはどれだけだろうか。これが図2上段右の n に対応する P' (打点部)となる。例えば80人の病欠者を出した大会社ではそのリスクはわずか0.2%である。しかし、病欠者が10人の小さい会社ではそのリスクは16%となり、大会社の P' 値は適用できない。いずれにせよ以上は会社という集団にとってのリスクである。一方、個人にとってのリスクは別であり、これは個人の属する会社の大小には関係ない(図2上段左)。もし従業員の給与が日当で1日1万円支払われるとする。病欠者個人にとって治療Aを選ぶ方がBを選ぶより平均10万円の得となることは、会社の場合と変わらない。しかし個人がAを選んで損をするリスクは、集団にとってのリスクよりも大きく38%となる。極端な例としては、平均でみればAの治療効果が有意に大きいにも拘わらず、Aを選んで損をする患者が50%を越えるというパラドックスも成立し得る。¹⁹

以上が P' 値の具体的な意味である。もし P 値を信頼区間に変えたとしても、 n の増加と共に信頼区間の幅が小さくなることに関しては P 値と同様である。問題は集団にとってのリスクと、個人にとってのリスク

を峻別することにある。例えば大規模臨床試験でeffect sizeの95%信頼区間が0を含まなければ($P < 0.05$ ならば)、大部分の患者ではその治療が他の治療よりよく効くと誤解する者は少なくない。逆にもし有意でなければ、 n を増やすことで問題が解決できると考える者は多い。対応のある t 検定において P 値を求める場合、もし例数が40であれば、 $n=40$ の差の平均の分布から P を計算するのが通例である。これは $n=40$ という集団にとってのリスクであり、個人にとってのリスクではない。

津谷によれば世界医師会のヘルシンキ宣言が32年間にわたり誤って和訳されていたという。²⁰ それは「わたしの患者の健康を先ず最初に配慮する(the health of my patient will be my first consideration)」の文中の「わたしの患者」が単なる「患者」と訳されていた。しかし患者個人の利益と患者一般の利益とを区別しなかったのはヘルシンキ宣言の翻訳者だけであろうか。個人のリスクを知りたければ図2の右ではなく、左上の分布から P' を求めるのが本筋である。なお P' 値を記載する場合には当然 n を併記すべきである。教科書には書かれていないが、これは P 値についても言えることである。

臨床医や患者にとっての理想の尺度とは

EBMを提唱したSackettらのグループはエビデンスを5つの順位に分けた。²¹ 上位のエビデンスはRCTから得られたもので、第一種と第二種の過誤が共に小さければ第一位、大きければ第二位とした。しかし他方ではSackettらはEBMの対象を集団ではなく、個々の患者のケア - (the care of individual patients)に置いている。²² これは臨

床医にとっては当然のことである。しかし現行のRCTでは個々の患者にとって治療AがBよりどれだけ得かはわからない。これを知るためには同一患者で2つの治療を比較しなければならないが、大部分のRCTでは、2治療の効果を別々の患者で比較している。つまり並行比較 (parallel comparison) である。⁹ これは治療群間で平均と平均の比較しかできず、これからわかるのは集団にとっての治療の得失である。ここに現行のRCTの限界がある。

集団にとってのP'値はnが変われば変動する。臨床医や患者にとっての理想は、個人にとってのP'を知ることである。これならば症例数を増減してもあまり変動はない。最終目標は $P < 0.05$ を達成することではなく、 $P'(n=1)$ をできるだけ小さくすることに置くべきである。後者を求めるためにはエンドポイントに代えて、繰り返し観察可能な代替エンドポイント¹⁹を使って、n of 1 trial²³ができないか今後検討すべき課題である。さらに、もし治療Aに反応する集団とBに反応する集団とが混在すれば、両者を分離することによりそれだけ $P'(n=1)$ は小さくなり、同時にeffect sizeは大きくなる。これこそは臨床医が追求すべき理想であるが、いずれも実現することは容易なことではない。Yusufら²⁴は「もし完全性を望むあまり大勢を知ることを敵に回すようなことになれば不幸なことである」と述べている。しかし現実の問題点から目をそらせば新しい道は拓けない。理想の追求と現実的アプローチは両立可能と考える。

参考文献

1. Sackett DL, Straus SE, Richardson WS, et al. Evidence-based medicine: How to practice and teach EBM. 2nd Ed. Churchill Livingstone, London, 2000.
2. Relman AS. Assessment and accountability: the third revolution of medical care. N Engl J Med 1988;319:1220-1222.
3. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976;34:585-612.
4. Freeman PR. The role of P-value in analyzing trial results. Stat Med 1993;12:1443-52.
5. Royall RM. The effect of sample size on the meaning of significance tests. American Statistical Association 1986;40:313-315.
6. Rothman KJ. A show of confidence. N Engl J Med 1978;299:1362-1363.
7. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986;105:429-435.
8. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986;292:746-750.
9. Ware JH, Mosteller F, Infelfinger JA. P values. In: Bailar JC, Mosteller F, eds. Medical uses of statistics. Boston, the Massachusetts Medical Society; 1986.
10. 津谷喜一郎, 折笠秀樹. 医学統計学の活用. 東京:サイエンティスト社; 1995:125-141.
11. Bulpit CJ. Confidence intervals. Lancet 1987; 494-497.
12. Evans SJ, Mills P, Dawson J. The end of p value? Br Heart J 1988;60:177-180.
13. Anscombe FJ. The summarizing of clinical experiments by significance levels. Stat Med 1990; 9:703-8.
14. Weinstein WC, Fineberg HV. Clinical decision analysis. WB Saunders, 1980.
15. Bunn DW. Analysis of optimal decision John Wiley & Sons, Chisester, 1982.

15. 天理よろづ相談所医学統計解析グループ . 情報の価値からみた医学統計学の再検討 . 臨床評価1999;27:393-407.
16. 天理よろづ相談所医学統計解析グループ . Neyman-Pearson 統計学から新しい臨床統計学へ : エビデンスよりは説明責任を . 天理医学紀要 2002;5:100-116(<http://www.tenriyorozu-hp.or.jp/01hospital/06kenkyu/igaku-kiyou/vol-5.html> でダウンロード可能)
17. Maetani S. Another approach to clinical trial numbers. Lancet 1990;335:114.
18. Schwarz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. J Chron Dis 1967;20:637-648.
19. 前谷俊三 , 小野寺 久 , 今村正行 . 外科におけるランダム化比較試験の問題点 . 外科 2003;65:993-999.
20. Revine RJ, 津谷喜一郎 , 坂上正道 , 他 . 医薬品開発のグローバリゼーションにおける臨床試験の倫理 臨床評価1999;26:341-390.
21. Cook DJ, Guyatt GH, Laupacis A, et al. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 1992;102 Suppl:305-311.
22. Sackett DL, Rosenberg WMC, Gray JAM, et al. Evidence-based medicine: what is it and what it is'nt: It's about integrating individual clinical expertise and the best external evidence. Br Med J 1996;312:70-72.
23. Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The n-of-1 randomized controlled trial: clinical usefulness. Ann Int Med 1990;112:293-9.
24. Yusuf S, Collins R, Peto R. Why do we need some large sample randomized trials. Stat Med 1984;3:404-420.

Evidence in medicine and *P* value

Statistical Analysis Group

Tenri Institute of Medical Research

In evidence-based medicine, randomized controlled trials (RCT) are considered to provide the most reliable evidence. *P* values summarizing the results of RCTs have been widely used by regulatory agencies in approving new treatments and by clinicians in making decisions as to which treatment should be applied to their patients. However, some statisticians question whether the *P* value is logically defensible, deserving of its widespread use. In this paper we re-examine the implications of the *P* value from the perspective of clinicians and patients.

The weaknesses of the *P* value are that: 1) it is hard to understand, and is not a clinically relevant measure since it is the conditional probability of **observing an outcomes given a hypothesis**; 2) an equal *P* value does not mean that the strength of evidence is equal, the latter varying with the sample size; 3) the interest of clinicians and patients could better be served by the confidence interval since they are interested in the reliable range of the effect size, rather than whether it is statistically significant. These problems are partly overcome by a new interpretation of the *P* value, as the risk that the expected better treatment, compared to the worse treatment, actually produces inferior outcomes. However, this risk serves the interest of a group of people rather than individual patients, since it is usually obtained by comparison of group means in parallel design RCTs. The risk for individuals should ideally be estimated by comparison of treatments in the same patients.

Keywords: *P* value, randomized controlled trial, evidence-based medicine, effect size, risk of receiving inferior treatment

依頼コメント

手良向 聡

京都大学医学部附属病院 探索医療センター検証部

はじめに

医学統計解析グループの「医療におけるエビデンスとP値」(以下,論文)では,経験豊富な臨床医の眼から,臨床試験に用いられている統計的検定という論理形式とP値について考察がなされている。「P値は統計学者の間でも議論の多い値である。医療改革の時代においては専門家だけでなく,臨床医や患者や社会がその意味を理解し,それぞれの立場からその妥当性を検証する必要がある。」という緒言には,全く同感である。1930年前後にJ. NeymanとE. Pearsonが「Hypothesis Test (仮説検定)」を, R. A. Fisherが「Significance Testing (有意性検定)」を提唱して以来,それらの論理形式とP値の解釈について多くの批判や考察がなされている(例えば 柳本 1991, Goodman 1999)。

P値の解釈について

簡単に言うと, Fisherは, P値をデータと帰無仮説との乖離を表す1つの尺度と考えていた。NeymanとPearsonのアプローチは, 仮説検定の結果は有意か有意でないかの2種類であってP値は解釈しないというものである。現在, 医学分野ではこれらの両者の考え方が都合よく組み合わせて用いられている。

論文でのP値の説明とその問題点の指摘はやや分かりにくい。その要因は, 臨床試

験の開始前に設定される「第 種の過誤の確率(検定の有意水準あるいは エラー)」と臨床試験の終了後に計算される「P値(有意確率)」が混同されていることである。特に, 表1と表2を用いた説明は私には難解である。臨床試験を科学的な技術評価の1つの方法と考えると, 臨床試験を計画する際に使用する道具(帰無仮説, 対立仮説, 第 種の過誤, 第 種の過誤, 期待治療効果など)と臨床試験の結果を解釈際に使用する道具(P値, 信頼区間, 推定治療効果など)との整合性が要求される。統計的推論の時間的な順序と整合性を考慮しないで議論すると問題点が正確に把握できないと考える。

統計的検定と意思決定について

論文で紹介されている方法(Schwartz and Lellouch, Anscombe, Maetani)は統計的検定と意思決定を論理的に結び付けようという試みであり, たいへん魅力的である。しかしながら, これらの手法は, 単一の治療の優劣を決定づける指標が存在するときにしか適用できず, そのような状況は臨床現場ではほとんどないと考えられる。臨床現場における意思決定で考慮される因子には, 少なくとも有効性, 安全性, 利便性, 経済性が含まれ, しかも単一の臨床試験の結果のみから, これらを同時に評価することは

困難である。そこで、可能な限りの情報を収集して、決定分析などの手法を用いてコンセンサスを形成していくというプロセスが重要となる。この意思決定のプロセスを明示的に示すことが、いわゆるEBMであると私は理解している。

これからの医療統計家

いずれにしても、臨床試験から得られる結果をどのように個人の意思決定に役立てるかという視点は重要である。それは、論文に述べられているように「集団にとってのリスクと個人にとってのリスクを峻別すること」である。しかしながら、最初から個人のベネフィットとリスクを推定することが可能であれば、集団のベネフィットとリスクを推定する必要はない。従って、個々の臨床試験において集団に対するベネフィットとリスクを推定した上で、それらの情報を総合して分析し、個人の治療に還元していくというプロセスが現時点で我々

が取り得る最良の戦略と考える。医療現場で働く医療統計家としては、個々の臨床試験のデータを分析するだけでなく、その診断、治療に関するあらゆる情報を総合したり、分析したり、評価したりしながら、個人の意思決定に役立つ情報を提供していくことが使命であり、その点で論文の主張には大きな共感を覚える。いわゆる「分析」だけを行う統計家は医療の分野ではあまり価値を認められないと感じる。医療統計家は、研究の計画、複数の研究の総合、集団レベルでの意思決定、個人レベルでの意思決定などにどこまで関与し、役に立つかが問われている。

参考文献

- 柳本武美. 統計的検定における帰無仮説の理解. 応用統計学 1991;20:97-107.
- Goodman SN. Toward evidence-based medical statistics. 1: the *P* value fallacy. *Annals of Internal Medicine* 1999;130:995-1004.

依頼コメント

折笠秀樹

富山医科薬科大学 臨床統計学

著者は、要旨の中で「 P 値とは、2つの統計学的誤りの中の1つを犯す条件付き確率である」と述べている。しかし、 P 値とは帰無仮説が正しいと仮定したとき、今得られたデータもしくはそれ以上極端（対立仮説寄り）なデータが出現する（条件付き）確率であるので、誤りを犯す確率という表現は正しくない。「 P 値とは・・・臨床的に適切な尺度といえない。」「 P 値は並行比較デザインによるRCTに基づき、群の平均を比較して得られるので、それは個々の患者の得失よりも集団の利害を反映する。」については、その通りだと思う。「個人にとってのリスクを推定すべきである。」も、尤もな話かと思う。

毒物混入事件の犯人と P 値のセクションにおいて「その証拠の強さを測る一つの尺度が P 値である。」と述べられている。証拠の強さは P 値というよりは、効果サイズあるいはNNTで測るとEBMでは言われている。例えばメタアナリシスの P 値は $P < 0.0001$ になるのはざらなので、メタアナリシスでは P 値は意味がないと言われている。

症例数と P 値のセクションでは「 $P = 0.05$ という値は、少数例から得たものならば強い証拠となるが、多数例から得たものであれば、その証拠は常に極めて弱く、極端な場合は差がないという逆の証拠になる」と書かれているが、尤もだと思う。

P 値対 P' 値のセクションにおいて著者の主張したエラーについては、幸い値が極小になることがほとんどなので、Neyman-Pearsonでは取り扱わなかったからといって、それほど影響が大きいとは思えない。ただ、エラーは大切な概念かと思う。「 P 値にせよ P' 値にせよeffect sizeが同じでも n が増えるとその値が小さくなる・・・」については、その通りかと思う。

集団にとっての P' 値と個人にとっての P' 値のセクションにおいて「個人のリスクを知りたければ図2の右ではなく、左上の分布から P' を求めめるのが本筋である。」と述べられている。 $n = 1$ の分布と言うのは、つまりは個人内変動を示す分布になる。これを知るためにはN-of-1デザイン試験、あるいは少なくとも個人内で複数個の観察値がないと作れない。Bayesian流に考えれば、データが蓄積するにつれ、個人に関するリスクというか予測確率が計算可能である。Neyman-Pearsonでは集団の平均的像についての判断になり、集団のリスク（確率）となる。

臨床医や患者にとっての理想の尺度とはこのセクションに書かれた「ここに現行のRCTの限界がある。」は賛成である。私も常々平均的結果が出てきているにすぎないのだと述べてきた。あとどう解釈するかは1人1人の患者を前にして考えることであり、そこではFraction、つまりこの患者は

RCTの平均的患者より重症か軽症かを考え、割り引いて勘案することをEBMでは推奨している。

全体を通じて、著者の述べる個人のリスクを確率として求めるのは臨床現場では重要な課題であろうが、RCTからそれは得られないのではないかと思う。個人内試験を行う（通常の外来診療を繰り返す、というのでいいと思う）ことが必須になると思う。ただ、ある個人の患者さんから「あと1か月は生きられるでしょうか」と尋ねられて、その確率を正しく伝えるのは難しいと思う。RCTなどでの平均的数値から、こ

の患者さんの様態や背景を勘案して、その確率を修飾するくらいではないだろうか。また、「確率でいわれても困る。結局生きられるかそうでないか、これについて先生の意見を聞きたい」と言われたらどうすべきだろうか。

まとめ

Neyman-Pearson流の P 値は、著者も述べるように限界がある。それは、Bayesian流の P 値に変更すれば、多少は現実に合った指標になるのではないか。また、現場では P 値よりもNNTのほうが重要な指標かと思う。

統計科学の足もと p 値と信頼区間

河合統介^a, 栗林和彦^b, 濱崎俊光^a, 後藤昌司^c

^a ファイザー(株) デベロップメント・オペレーション統括部・生物統計部

^b 日本イーライリリー(株) 臨床開発部 臨床統計室

^c 大阪大学 大学院基礎工学研究科 システム創成専攻・数理科学領域

1. 序に代えて

「p 値と信頼区間」は、統計科学において古くて新しい主題である。Day¹の「医薬統計学の変遷」によれば、製薬企業の内部では、統計的考え方については初期の問題に絡むだけで新しい試みは何もなされず、とくに1980年代に脚光を浴びたのがp値であると述べている。統計家は、統計的方法論の研究に主たる関心をおきがちであるが、実際のデータ解析の過程におけるデータの省察と結果の解釈を軽々と済ませることがある。前谷²は「証拠に基づく医療 (EBM: Evidence-Based Medicine)」よりも、統計家がデータ解析の過程で、「説明責任をもて」と鋭く指摘している。前谷³の「医療におけるエビデンスとp値」もその流れの中での提案である。本小文では、前谷³の提案に関する検討と、医学統計研究会の定例会を始めとする諸種の会合において、著者らが折に触れ検討してきた文献の省察に基づき、二三の見解と留意点を提示したい。

2. p 値と信頼区間：文献の省察

一般に、p値は、効果がないあるいは効果

に差がないという仮定（帰無仮説）のもとで、実際に観測された結果と同じかより極端な結果が得られる確率の実現値として定義される。^{4,5} Fisher⁶は、p値をデータと帰無仮説の間の不一致を測る指標として解釈することを提案している。前谷³の提案に関する議論の前に、最初に、不幸にも広範囲に蔓延しているp値の誤った解釈を指摘しておきたい。

まず、最も陥り易い誤解は、p値を仮説の確率として捉えること、すなわちp値が0.05であることを、帰無仮説が真である確率が5%しかないと解釈することである。⁵これをいいかえて、帰無仮説が偽である確率が95%であると解釈するものもいる。⁷これは完全に誤った解釈である。何故なら、p値は帰無仮説が真であるとの仮定のもとで算出される値であるからである。

もう一つの陥り易い誤解は、p値と有意水準（第I種の過誤率：水準とも呼ばれる）を混同することである。これは、臨床家だけでなく、統計家にも起こり得る。例えば、Hungら⁸は、p値を帰無仮説に対する証拠の測度としながらも、p値と第I種の過誤率

を混同する記述を続けている：「水準は実験前の第I種の過誤率であり、実際に H_0 が真のときに、 H_0 を棄却する誤りを犯す実験で観測されるp値を以下の確率に制御するために用いられる」⁹ また、「p値は有意な最小の水準、すなわち「境界線水準」と解釈できる」と述べている文献¹⁰もあるが、こういった記述もp値と有意水準の不用意な混同を招いている。p値(証拠)と有意水準(過誤)は異なる概念であり、明確に区別することが必要である。p値と有意水準に関する更なる興味深い議論については、Hubbard & Bayarri⁹を参照してほしい。

p値が提案されたとき、複数の統計家から、p値の論理的な基盤と実用上の有用性に対する批判があがった^{11,12}。おそらく、それらの批判の大半は、p値が証拠の測度として、観測された効果の大きさを反映していない点に向けられたと考えられる⁷。すなわち、大標本における小さい効果が、小標本における大きい効果と同じp値をとりえる点である。いいかえれば、前谷³によって指摘されているように、たとえp値が同じでも、標本サイズに応じて証拠の強さが変わる¹³。したがって、標本サイズが異なる場合に、あるいは異なる実験でp値の大きさ自体を比較することは、残念ながら、ほとんど意味をもたない。

この批判が、p値よりも信頼区間(仮説検定よりも推定)に重点をおく今日の流行の基盤となっている。上記の批判の他に、p値よりも信頼区間の有用性を主張する論文¹⁴⁻¹⁸で述べられている内容を要約すると、p値との対比での信頼区間の利点は、以下のとおりである。

- 研究者の疑問は「効果に差があるか」よりもむしろ「効果の差の大きさがどの程度であるか」にある。この疑問には、信頼区間を算出することにより答えられる。p値は、効果の差の大きさについて何の情報も与えない。
- 信頼区間は、効果が測定された単位のまま表示される。その結果、読み手に結果の臨床上の妥当性を批判的に考察することを可能にする。
- 信頼区間は、仮説検定がもつすべての情報を提供し、その情報の臨床上の妥当性も示唆する。つまり、信頼区間がゼロを含むとき、その結果は、有意でない結果を与える仮説検定と同値である。また、信頼区間の上限(あるいは下限)が、その効果の差が臨床的に有用であるかどうかを精査することも可能にする。
- 仮説検定の結果が、小標本の場合に誤って解釈されることがある。それは、統計的に有意でない差が、実際に差が存在しないことを意味すると捉えられることである。逆に、研究者が有意な差を見出したいくない場合には、不適切に少ない標本サイズを用いればよい。研究者が、結果の報告と彼らの知見の解釈を、適宜、信頼区間を用いて行うのであれば、これらの誤解を与える機会を減らすことができる。

しかしながら、p値から信頼区間への転換ですべての問題が解消される訳ではない⁵。p値は、帰無仮説に基づいて算出されるので、対立仮説のもとでの分布が明示的に得られない場合に有用であると考えられる。また、Weinberg¹⁹は、推定だけでは不十分で仮説検定が重要な補足的役割を演じる場面

を例示している。さらに, Freeman⁵ は, 頻度論流の観点だけではなく Bayes 流の観点からの接近の重要性を述べている。上述のように, p 値と信頼区間が仮説の確率を測っていると解釈することは誤りであり, もし仮説の確率を知りたいのであれば, Bayes 流接近法を用いるべきである。²⁰ Bayes 流の p 値の解釈については, 3 節および 4 節で議論する。

3. 前谷³の提案と解釈

EBM では, 適切に管理された無作為化臨床試験の結果が信頼のおける証拠であると見做されており, 無作為化臨床試験での主要な比較結果は p 値で要約されることが多い。このことから, 前谷³ は, p 値は医療における証拠の強さの測度になり得るかについて考察し, Anscombe²¹ をもとに, p 値を臨床医および患者の立場からわかりやすい「量」として解釈することを試みている。

先述のように, p 値は, 実際には, 帰無仮説のもとで, 得られた実験 (観察) 結果またはそれよりも極端な結果の得られる確率の実現値である。ここに, 「極端」は, 規定した尺度 (検定統計量) 上で測られる。仮説検定では, 事前に規定した有意水準と対比して, p 値が有意水準未満であれば, 帰無仮説を棄却する。因に p 値が「実現値」であることに留意して, 「 p 値の期待値 (期待 p 値)」を提案したものに Sakrowitz & Samuel-Cahn²² がある。さらに, その代替として Bhattacharya & Habtzgh²³ は, p 値の期待値の代わりに中央値を提案している。

Anscombe²¹ では, 簡単のために, 新治療と標準治療の比較において, その期待利益の差を t として, 統計量 t が実験 (観察) が

ら得られた t の推定値であり, 平均 μ , 既知の分散 σ^2 の正規分布に従う状況を想定している。ここに, t の大きな値が新治療の優位性を示すこととする。

この場合に, t の事前分布に一様分布を仮定すると, t が与えられたときの t の事後分布は平均 μ , 分散 σ^2 の正規分布になる。Anscombe²¹ はこのことを利用し, t 以上に極端な結果の得られる確率, すなわち両治療に差がないという帰無仮説の新治療が標準治療に優るという対立仮説に対する片側 p 値が, t が与えられたもとで t が負値になる確率に等しくなることから, p 値は, t が与えられたもとで新治療が標準治療よりも劣る確率を測ると解釈できるとしている。

前谷³ は, このことに着目し, t が与えられたもとで t が負値になる確率を p' 値として, これに頻度論流の解釈を与えている。すなわち, 新治療を施すと損をする患者が p' だけ存在するとしている。これは, 暗に次のように考えていることになる。クロスオーバー試験などで患者ごとの期待利益差が推定可能である場合に, t_i が患者によって異なり, 集団全体でみれば一様分布 (事前分布) に従っているとすると, ここに, 添字 i は患者個人を表す。実験 (観察) から t_i が得られたもとでの t_i の事後分布の期待値が t_i であるので, t_i の標本分布に基づいて t_i が負値の確率を p' 値として推定することができる。これは, 経験 Bayes 推定に他ならない。そうすると, p' 値は, 新治療よりも標準治療の利益の方が大きい患者の割合, すなわち新治療を施すと損をする患者の割合と解釈できるであろう。

しかしながら, この議論は Anscombe²¹ が想定している場面, すなわち t_i が分散既知

の正規分布に従い、事前分布が一様分布の場合に限定される。たとえば²が未知の場合、²の事前分布として一様分布を仮定することが不適切な場合、期待利益を割合で測る場合などは「 p 値」「 p' 値」となる。このような場合に、 p' 値だけを評価することは可能であるが、事後分布を陽に表すことができない場合には、 p' 値の計算にMarkov連鎖モンテカルロ(MCMC)法などのコンピュータ集約型接近法が必要になる。

4. Bayes流の解釈と実践的接近法

Anscombe²¹とは別の観点で、Goodman^{24,25}により提案された、「Bayes因数」を用いた p 値のBayes流の解釈について述べる。ここでは、事前(事後)確率として「帰無仮説が真である確率」を導入することにより、帰無仮説に対する証拠の強さとして p 値を解釈するうえで、示唆的な見解を与えている。

いま、Bayesの定理において、データが得られる前に設定された帰無仮説が真である確率のオッズは「事前オッズ」と呼ばれ、データが得られた後での帰無仮説が真である確率のオッズは「事後オッズ」と呼ばれる。Bayes因数は、この二つのオッズがどの程度まで乖離しているか、すなわち、試験から得られたデータが、事前に設定したオッズを最初の位置からどの程度まで移動させるかを示す指標である。このことから、Bayes因数の対数値は「証拠の重さ」と呼ばれることもある。上記のことは、次のような関係式で表される。

$$\begin{aligned} & \text{帰無仮説の事前オッズ} \times \text{Bayes 因数} \\ & = \text{帰無仮説の事後オッズ} \end{aligned}$$

上式からわかるとおり、Bayes 因数は事前オッズに対する事後オッズの比である

が、これらのオッズを用いずにデータから計算可能である。それは、二つの競合する仮説のもとでのデータの確率(あるいは尤度)の比で表される。例えば、ある対立仮説に対する帰無仮説のBayes因数が $1/10$ であるとき、この試験の結果が帰無仮説の相対オッズを $1/10$ に減少させることを意味する。つまり、帰無仮説の事前オッズが 1 (帰無仮説が真である事前確率は 0.5)であるとき、試験後のオッズは $1/10$ (帰無仮説が真である事後確率は 0.09)に減少する。

Goodman²⁴は、 p 値の算出に用いた数値で計算が可能なおとと解析的な導出が容易であることから、数あるBayes因数のなかで最小Bayes因数の利用を推奨している。最小Bayes因数は、データが最も良く支持する対立仮説(最大尤度)を分母にもつBayes因数である(すなわち、データに基づくBayes因数の最小値である)。とくに、検定が正規近似に基づく場合には、最小Bayes因数は $\exp(-z^2/2)$ と簡単に表される。ここで、 z は帰無仮説での規準化偏差を表す。例えば、 t 検定の場合には検定統計量 t で z をおきかえればよい。ここで、標本サイズは固定されているとする。

この p 値と最小Bayes因数の関係式を用いて、諸種の z スコアにおける p 値とそれに対応する最小Bayes因数の値を表1に示す。例えば、 z スコアが 1.96 であるとき(すなわち、 p 値が 0.05 であるとき)、最小Bayes因数は 0.15 であり、このことは、今回の試験の結果が最も良く支持する仮説を 100% とすると、帰無仮説がその 15% の支持(信頼度)を受けていることを意味する。これは、 p 値の値 0.05 より 3 倍大きく、 p 値が 0.05 であることが提示しているほど、帰無仮説

が偽である証拠が強くはないことを示唆している。²⁴

表1の右側は、p値が、帰無仮説の事前確率から事後確率への変化に与える影響の大きさを示している。効果に差がないとする帰無仮説の事前確率を0.05としたとき、試験後に最小Bayes因数が0.15(対応するp値は0.05)との結果が得られれば、帰無仮説に対する信頼度が13%まで減少することがわかる。各p値に対応する最後の行は、帰無仮説について、どの程度の事前の信頼度が、データが得られた後に5%の信頼度(すなわち、帰無仮説が真ではない信頼度が95%)になるかを示している。p値が0.05のとき(Bayes因数は0.15以上)、帰無仮説が偽であることを95%の信頼度で結論づけるためには帰無仮説の事前確率が26%以下でなければならないことがわかる。

要約すると、Bayes流の枠組みでは、ほとんどの場合に、p値が0.05であるとき、帰無仮説が真である可能性がなお残る。このことは、ちょうど0.05水準で有意ということが、帰無仮説を偽と仮定することの強力な正当化になっていないことを示唆している。

Goodman²⁴は、p値が提示している大きさほど帰無仮説に対する証拠の重さは強くはなく、p値が帰無仮説に対する証拠を誇張していると指摘している。この見解については異論もあるが(Goodman²⁶とSenn²⁷による議論を参照)、Bayes流の枠組みでp値を捉えることは、p値の解釈の誤解(p値を帰無仮説が真である確率として解釈すること)を防ぎ、また帰無仮説に対する証拠の強さとしてp値を解釈するうえで、有用であると思われる。

表1. p値のBayes流の解釈⁷(文献引用)

p値 (zスコア)	最小Bayes因数	帰無仮説が真である確率		証拠の強さ
		事前確率(%)	事後確率(%)	
0.1 (1.64)	0.26 (1/3.8)	75	44	弱い
		50	21	
		17	5	
0.05 (1.96)	0.15 (1/6.8)	75	31	中程度
		50	13	
		26	5	
0.03 (2.17)	0.095 (1/11)	75	22	中程度
		50	9	
		33	5	
0.01 (2.58)	0.036 (1/28)	75	10	中程度～強い
		50	3.5	
		60	5	
0.001 (3.28)	0.005 (1/216)	75	1	強い～非常に強い
		50	0.5	
		92	5	

5. 結びに代えて

「p 値の文化」が医療分野に及ぼす影響の大きいことが、前谷³の指摘の動機になっていそうである。統計学の教科を受けたことのある多くの者（統計家だけでなく実質科学分野の研究者も含む）が陥る誤解は、統計解析の目標が p 値を求めることであるとの信念をもたられ、この値が十分に小さければ、その結果あるいは結論を記載した論文が著名な論文誌に採択されることである（Nelder²⁸、後藤²⁹）。すなわち、そこで本来の目標である実験あるいは研究の価値を p 値が左右することになる。そして、仮説検定や有意性検定だけが、統計科学の代表であるかのような誤解を与えている。確かに、論旨の応否を 2 値論証で済ませることのできる簡便性は利用者にとって魅力的である。そのことが、信頼区間や Bayes 流接近法などに基づく定量的な解釈を要する方式よりも馴染みやすいものになっている。ただし、医学や医療の研究で必要なことは、情報を累積していく科学的な方法であり、「p 値の文化」それ自体はそこへ何の貢献も果たしていない。²⁹ Cox & Snell³⁰ も「p 値だけでなく、効果の大きさ、いわゆる信頼限界を計算することが必要である。このことはきわめて重要である。p 値だけで重要な研究を要約することは非常に悪しき慣行である」と警告を発している。

他方、吉村³¹は「実際、もし決定をくだす人が、p 値を正確に理解する能力をもっているならば、それを記しておくほうが、単に検定で有意であったかどうかを記しておくより、よい決定をくだすことができる。しかし p 値の意味を正確に理解していない場合には、それが恣意的な結論を生む

ための口実になるから、ことは単純でない。データ解析というのは、ある意味で結論の単純化であるから、情報の減少がないことを重視して、結論のおろしくさを無視するわけにはいかない。p 値は、一つの情報として、記録として残しておくべきだが、二者択一問題としては、それを補助的に使うのがよいのでなからうか」と主張している。おそらく、これが統計家の無難な代表意見であろう（丹後³²も参照）。

p 値の代替測度としては、前谷³の提案以外にも、先述の Sackrowitz & Samuel-Cahn²²の期待 p 値（EPV）がある。これは

$$EPV(\theta) = \Pr(T^* \geq T)$$

で定義される。ここに、 T は検定統計量であり、帰無仮説 H_0 のもとでの T の分布関数が $F_0(\cdot)$ 、対立仮説 H_1 のもとでの T^* の分布関数が $F_\theta(\cdot)$ である。また、 θ は対立仮説を規定する分布のパラメータである。 $F_0^{-1}(\cdot)$ は $F_0(\cdot)$ の逆関数であり、 $0 < \gamma < 1$ の場合に $F_0(F_0^{-1}(\gamma)) = \gamma$ である。p 値は、 T よりも極端な値を観測する（確率）統計量であるから、確率変数

$$X = 1 - F_0(T)$$

で定義される。 H_0 のもとで $F_0(T)$ は $[0, 1]$ 上で一様分布に従うことはよく知られている。実際に T が F_θ に従って分布するとき、 T に基づく有意水準 α での検定の検出力は $P_\theta(X \leq \alpha)$ と表される。これは、対立仮説のもとで p 値が α 以下となる確率になっている。すなわち

$$P_\theta(X \leq \alpha) = 1 - F_\theta(F_0^{-1}(1 - \alpha))$$

である。 α を 0 から 1 まで動かすことで、対立仮説のもとでの p 値 (X) の分布関数を得る。この分布の期待値をとることにより、 $EPV(\theta)$ が上式のように与えられる。

EPV()が小さいほど, 検出力が高いことを意味する。因に, 通常の p 値は, $T = t$ が観測されたときの $\Pr(T^* \geq t | T = t)$ の場合に相当する。なお, このEPVは連続分布の場合だけでなく, 離散分布の場合にも定義される。

最後に, 「 p 値を一人歩きさせない」ために二三の留意点をあげ結びに代える。

- (1) 実験研究と観察研究: p 値を論じる前に留意したいのが, 研究の型と姿勢である。 p 値の提示が要請されるのは仮説の検証を必要とする実験研究の場合が多い。その折でも, p 値だけに注意を払うのではなく, 実験研究のデザイン, 確率化操作, 標本サイズなどが結果の解釈でより重要である。これについては, Jeffers³³ と Finney³⁴ が参考になる。
- (2) p 値の記法: p 値を大文字と小文字のいずれで記すのか, 諸文献に戸惑いが見られる。多くの文献では, 医療分野では大文字の P 値, 統計科学分野では小文字の p 値(ローマン体)が用いられている。このような点に関して, 統計科学でも標準化が必要である。因に, Simon & Wittes³⁵ では, p 値を小文字で記すように奨めている。
- (3) 統計科学では, 実際に臨場感のあるデータを扱うときに, その方法論が活きることになるが, 臨床試験のデータのとり扱いでは, デザインの時点できり決めた解析以外は, 何もしないとった形骸的姿勢が目につく。 p 値もそのような場面で頻用される代表格である。統計的データ解析の過程では, 今後の研究につなげる生産的知見の発掘を狙って, 事後検討を徹底すべきである。

その過程で, p 値の形骸的な解釈をおさえるべきであろう。Driscoll³⁶ は, いくらかの比喩を込めて「統計的推測の十戒」を与えている。

汝, 統計的有意性を猟銃で射ることなかれ。

汝, 実験デザインなくして推測の処方の谷間にはいることなかれ。

汝, モデルなくして統計的推測を行うことなかれ。

汝, 汝のモデルの仮定を崇めよ。

汝, 有意な結果を得んために汝のモデルを汚すことなかれ。

汝, 汝の仲間のデータを無闇に欲することなかれ。

汝, 汝の対照群に偽証することなかれ。

汝, 0.05 有意水準を崇拜することなかれ。

汝, 大標本近似を妄りに適用することなかれ。

汝, 因果関係を統計的有意性から推測することなかれ。

謝辞

貴重な論文(前谷^{2,3})を送付いただき, 本小文の寄稿をすすめていただいたこと, また統計科学における日常の道具や手法に鋭い問題提起と勇気ある提案をされていることに対して, 日頃のご指導と併せて前谷俊三先生に深甚の敬意と謝意を捧げます。

参考文献

1. Day S. Changing times in pharmaceutical statistics: 1980-2000. *Pharmaceutical Statistics* 2002;1:9-16.
2. 前谷俊三 [天理よろづ相談所医学統計解析グループ]. Neyman-Pearson 統計学から新しい臨床統計学へ: エビデンスより説明責任を. *天理医学紀要* 2002;5:100-116.
3. 前谷俊三 [天理よろづ相談所医学統計解析グループ]. 医療におけるエビデンスとP値. *天理医学紀要* 2003;6:54-66.
4. Ingeltinger JA, Mosteller F, Thibodeau LA, et al. What are p values? *Biostatistics in Clinical Medicine* (2nd edition), chap. 1987;7:151-167. Macmillan [土屋佳英・渡辺秀章・後藤昌司]. p値とは何か. *講読速報* No.321, シオノギ解析センター1990.].
5. Freeman PR. The role of p-values in analyzing trial results. *Statistics in Medicine* 1993; 12: 1443-1452.
6. Fisher R. *Statistical Methods and Scientific Inference*. 3rd ed. Macmillan, New York. 1973.
7. Goodman SN. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine* 1999;130:995-1004.
8. Hung HMJ, O'Neill RT, Bauer P, et al. The behavior of the p-value when alternative hypothesis is true. *Biometrics* 1997;53:11-22.
9. Hubbard R, Bayarri MJ. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician* 2003;57(3):171-182.
10. Gibbons JD, Pratt JW. P-values: Interpretation and methodology. *The American Statistician*, 1975;29(1): 20-25. [後藤昌司・土屋佳英. p値: その解釈と方法論. *講読速報* No.4(改訂版) シオノギ解析センター].1983.
11. Berkson J. Tests of significance considered as evidence. *Journal of the American Statistical Association* 1942;37: 325-335.
12. Pearson E. 'Student' as a statistician. *Biometrika* 1938;38:210-250.
13. Royall RM. The effect of sample size on the meaning of significance tests. *The American Statistician* 1986;40:313-315. [栗林和彦 (2003). 標本サイズの有意性検定の意味に及ぼす影響. *医学統計研究会・2003年度第12回定例会資料*2003;9;13(2)].
14. Bulpitt CJ. Confidence intervals. *The Lancet* 1987;28:494-497.
15. Evans SJW, Mills P, Dawson J. The end of p-values? *British Heart Journal* 1988;60: 177-180.
16. Gardner MJ, Altman DG. Confidence intervals rather than p-values: estimation rather than hypothesis testing. *British Medical Journal* 1986; 292:746-750.
17. Rothman KJ. A show of confidence. *The New England Journal of Medicine* 1978;299(24): 1362-1363.
18. Simon R. Confidence intervals for reporting results of clinical trials. *Annals of Internal Medicine* 1986;105:429-435.
19. Weinberg CR. It's time to rehabilitate the P-value. *Epidemiology* 2001;12(3):88-290.
20. Poole C. Low p-values or narrow confidence intervals: Which are more durable? *Epidemiology*, 2001;12(3): 291-294.
21. Anscombe FJ. The summarizing of clinical experiments by significance levels. *Statistics in Medicine* 1990;9:703-708. [河合統介・後藤昌司. 有意水準による臨床試験の要約. *統計科学講究録* 2003-G9-2. 大阪大学 大学院基礎工学研究科 数理科学領域].
22. Sackrowitz H, Samuel-Cahn E. P values as random variables: Expected p values. *The American Statistician* 1999;53(4):326-332. [河合統介・後藤昌司 (2003). 確率変数としてのp値: 期待p値. *統計科学講究録* 2003-G9-3. 大阪大学 大学院基礎工学研究科 数理科学領域].
23. Bhattacharya B, Habtzgh D. Median of the p-value under the alternative hypothesis. *The American Statistician* 2002;56(3):202-206.
24. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine* 1999;130:1005-1013.

25. Goodman SN. Of P-values and Bayes: A modest proposal. *Epidemiology* 2001;12(3): 295-297.
26. Goodman SN. Author's reply: Letter to the editor: A comment on replication, p-values and evidence. *Statistics in Medicine* 2002;21:2445-2447.
27. Senn S. Letter to the editor: A comment on replication, p-values and evidence. *Statistics in Medicine* 2002;21:2437-2444.
28. Nelder JA. Statistics for the millennium: From statistics to statistical science. *The Statistician*, 1999;48(2):257-269. [後藤昌司・高瀬貴夫(2000) .新たな千年紀に向けての統計学:統計学から統計科学へ. 統計科学講究録2000-G6-1, 大阪大学 大学院基礎工学研究科 情報数理系専攻 数理科学領域].
29. 後藤昌司. 生存時間解析過程における忘れもの. *癌臨床研究・生物統計研誌* 2000;20(1):1-9.
30. Cox DR, Snell EJ. *Applied Statistics: Principles and Examples*. Chapman and Hall 1981. [後藤昌司・土屋佳英 (医学統計研究会) (1985) .応用統計実践教本, MPC].
31. 吉村 功 .p値をどう見る, どう使う? 医薬安全性研究会・第39回会合1990;1:30 .
32. 丹後俊郎. p-value (p値)とは? 誤解を防ぐために. 医薬安全性研究会第42回会合1990;4:7 .
33. Jeffers RE . Check list of experimental design. Institute of Terrestrial Ecology, Merlewood Research Station, Change-over-sands, Cumria LA116JU.1978.
34. Finney D. The questioning statistician. *Statistics in Medicine* 1982;1: 5-13. [土屋佳英・後藤昌司 (1983). 統計実践訓:統計家の自問自答. SHI-Seminary Note No.62, シオノギ解析センター].
35. Simon R, Wittes RE. Methodologic guidelines clinical trials. *Cancer Treatment Reports* 1985;69(1):1-3. [土屋佳英・後藤昌司 (1986) .臨床試験の報告書の作成指針: 編集委員会の提言. SHI-Seminary Note No.27, シオノギ解析センター].
36. Driscoll MF. The ten commandments of statistical inference. *American Mathematical Monthly* 1977;84:628.