日病薬誌　第50巻6号（729-732）2014年

第２回　お馴染みp値，見慣れぬ信頼区間

横浜薬科大学医療統計学
奥田千恵子　Chieko OKUDA

検定と区間推定

第２回は主役，推測統計の登場である。

前回述べたように，推測統計では，データは母集団（population）から抽出された標本（sample）として扱われ，統計学的仮説検定（以下，検定），あるいは区間推定（interval estimation）と呼ばれる多種多様な手法で解析される。個々の手法の特徴や手法の選択法については次回に述べる。

どの手法を用いても，結果には必ず不確かさが残る。推測の「不確かさの度合」を数値で表すためにp値（p value）や信頼区間（confidence interval：以下，CI）が用いられる。

p値は確率

p値はすでに「統計用語」として医療従事者に浸透してしまっているので，確率（probability）の略語であることに気づかずに使っている人が多いのではないだろうか？

確率とは，事象の起こりやすさの程度を０～１（０～100％）の数値で表したものである。歪みのないサイコロを何度も振れば，どの目も1/6の確率で出る。降水確率80％とは，気象台の観測値や全体的な気圧配置などの過去の多くの記録をパターン化した資料があらかじめ作成されており，特定のパターンに当てはまっていると，これまで５回に４回は雨だったので，今回も同じ確率で雨が降ると予想したものである。

このような確率計算に関する知識がなくても，降水確率は大きければ大きいほど雨が降りやすいと素直に解釈すればよい。80％と出ていれば，たいていの人は傘をもっていこうとか，洗濯物は部屋干しにしようとか，実用的な判断材料にすることができる。

ところが，p値のほうは一筋縄ではいかない。p値が0.8（80％）と書かれていても，業務として日常的に科学論文に接していない限り，情報として利用できる人はそれほど多くはない。

検定は自己否定から始める

p値には，統計学習者の理解を阻む２つの壁がある。１つ目は論理学の壁。検定（t-検定やχ^2検定など）では，論理学で「背理法」と呼ばれる間接的証明方法を用いる。

まず，２つの仮説を設定する。

帰無仮説（null hypothesis）：自分の説を否定した説

対立仮説（alternative hypothesis）：自分の説

そして，帰無仮説が間違っていることを統計学的に証明して，自分の説が正しいと主張する。私たちの日常感覚ではこのような回りくどい考え方はしないが，例えていうなら，ミステリ小説で最初に疑われる人はたいてい濡れ衣で，後に残ったのが真犯人という筋運びに近いかもしれない。

不確かさの残る結論

２つ目の壁は，論理学の教科書にもあまり登場しない奇妙な結論の仕方にある。

数学的な命題の場合は，例えば，以下のような仮説を立てる。

帰無仮説：素数の個数は有限である［Aである］

対立仮説：素数の個数は無限である［Aでない］

これを背理法により証明するには，

［Aである］か，［Aでない］か，どちらかである。

［Aである］は間違っている（100％確実）。

故に，［Aでない］。

と，一分の隙もない結論を得なければならない。この論法なら数学が苦手という人でも比較的受け入れやすい。ところが，統計学的な証明のほうは，

［Aである］か，［Aでない］か，どちらかである。

［Aである］は間違っている（95％以上確実）。

故に，たぶん……，［Aでない］。

と，数学が得意と自認する人にとっては気持ちが悪くなるような結論になる。

p<0.05とは

このような結論の不確かさはp値で表される。p値の定義は「帰無仮説が正しいにもかかわらず棄却してしまう確率（α過誤率）」である。p値は小さければ小さいほど，「自分の説が正しい」と自信をもって主張できる。

p<0.05とは，例えば，「群間に差がない」という帰無仮説が正しいにもかかわらず，得られたデータ間に差が出るのは，調査や実験を何回も繰り返したとすると，20回に1回以下しか起こらない稀なことだということである。通常の感覚では偶然の産物とは考えにくいので，不確かさを残したまま，「群間に差がある」という対立仮説を採択する。

しかし，0.05（5％）という有意水準（significance level）の数値に理論的根拠はない。医療分野では慣習的にそうしているだけである。その研究領域で合意が得られるのであれば，どんな値に設定してもよい。誤りを恐れず果敢にということであれば，有意水準を10％あるいは20％と設定することもできる。逆に，1％あるいは0.5％と，極めて慎重な態度を取ることもできる。

p値の計算

p値を正確に計算するのはコンピューターにとってさえ負担である。医療研究で使われる検定法のほとんどは電卓すらなかった頃に考え出されたものであり，正規分布を初めとする様々な確率分布を利用した近似計算をせざるを得なかった。まず，データから検定統計量（z値や，t値，F値，χ^2値など）を計算して，棄却限界値と比較するという方法が用いられてきた。統計学の教科書の巻末に今でも時々載っている棄却限界値表には，5％，1％，0.5％など，要所要所のp値しか書かれていない。そのため，p<0.05なる不等号式が使われるようになったのである。

統計ソフトを用いて検定する場合，棄却限界値との比較は行わず，得られた検定統計量に対応する値をそのまま求めている（例：p=0.035）。特定の確率分布を利用している以上，近似計算であることに変わりはない。

コンピューターの性能が飛躍的に進歩した現在，2項検定やフィッシャー直接確率法では，確率分布を利用せずに，「事象の起こりやすさの程度」（前頁の確率の定義，参照）として直接p値を計算している。従って，これらの手法では，近似値ではなく正確なp値が求められる。

p>0.05は「情報なし」

これまで述べてきたことからもわかるようにp値は決して単純明快な指標ではない。にもかかわらずp値が多用されるのは研究者が長らく馴染んできたからにほかならない。一方，統計学者の間ではp値はあまり評判が良くない。何故なのか？

例えば，新しい降圧薬X群と対照薬群との間で，血圧の変化量の平均値を比較した結果，有意差なし（p>0.05）だったとしよう。

帰無仮説が棄却できない場合の結論は「差がない，従って，同等である」ではなくて，「差があるのかないのか，どちらともいえない」である。つまり，p>0.05という検定結果から得られる情報はほとんど皆無である。

この状況は以下のように分けられる。

(1) 実際に差がない。
(2) 実際には小さな差（1～2 mmHg）があるはずだが，有意差が出なかった。
(3) 実際にはかなり大きな差（10～20mmHg）があるはずだが，有意差が出なかった。

新薬がほんの1～2 mmHg，対照薬より余分に血圧を下げるということを統計学的に証明できたとしても臨床的な意味はほとんどない。データ数を増やすなどして研究を継続すべきかどうか判断するには(1)と(3)を区別できればよいのだが，p値から「どの程度の差」があるのかという情報を得ることができない。

臨床的に意味のある差

臨床試験では，事前に，臨床的見地からだけでなく，科学的な価値，費用対効果比などを検討したうえで，ターゲットとする差を設定して，それを統計学的に検出するにはどの程度のデータ数が必要かを算出しておくことが求められている[1,2]。

研究目的によっては「差がない」ことを示したい場合もある。新しい降圧薬Xのほうが，副作用が少ない，あるいは，飲みやすいなどのメリットがあれば，必ずしも降圧効果が対照薬に比べて有意に優れている必要はなく，「劣っていない」ことを検証すればよいとされている。

これを証明するには，仮に劣っていたとしても臨床的に許容できる最小の差を設定して，非劣性の検定（non-inferiority test）と呼ばれる特別な方法を用いる必要がある[2]。

区間推定

医療研究では，結果（例：群間での血圧の変化量の平均値の差）が臨床的な意味をもつのかどうかに関する推定の不確実さを，データの測定尺度（例：mmHg）で直接示すことができる指標が必要なのである。

そこでp値の代わりとして用いられるのがCIである。CIは母平均値などの母数が存在すると思われる範囲を表す値であり，データからCIを求めることを区間推定という。母数とは，もし母集団のすべてのデータがあったとしたら求まるはずの「真の値」と考えることができる。CIにどの程度の確率で母数が含まれるかを表す値が信頼係数（confidence coefficient）である。医療研究では95％がよく用いられる。

p値は，有意である/有意でないと，情報を２値化して利用されることが多いのに対して，CIは，研究結果を元々の測定尺度で直接示すことができ，推定の不確実さを，CIの位置と広がりという２つの情報で示すことができる。

CIの使い方

p値と同様，CIも直観的に理解できる指標ではないが使い方はそれ程むずかしくはない。一般的な統計ソフトを用いると，幾つかの検定法では，p値だけではなく，CIも同時に得られる。降圧薬Xと対照薬の区間推定の結果は以下のように表される。

例）降圧薬X群と対照薬群における血圧変化量の平均値の差：15.5mmHg（95％ CI：−2.2～33.2mmHg）。

区間推定の結果から，⑴降圧薬Xが臨床的に意味のある降圧作用（15.5mmHg）を示す傾向があること，および，⑵95％ CIに０mmHgが含まれているので，両群の血圧変化量の平均値に差がないという可能性が排除できないことがわかる（CIの下限値が負の値を示すことから，むしろ対照薬のほうが降圧効果が優っている可能性もある）。

95％ CIとは

95％ CIとは，95％の確率で母数（真の値）が含まれている区間を表す。95％の確率で含まれているとは，「同じ調査や実験を20回繰り返したら，19回分のCIに母数が含まれている」ということである。母数は実際に行われた研究で手に入れた１つの標本から推定することになる。１回しか行われない研究が，偶然「真の値」を含んでいないこともあり得る（図）。

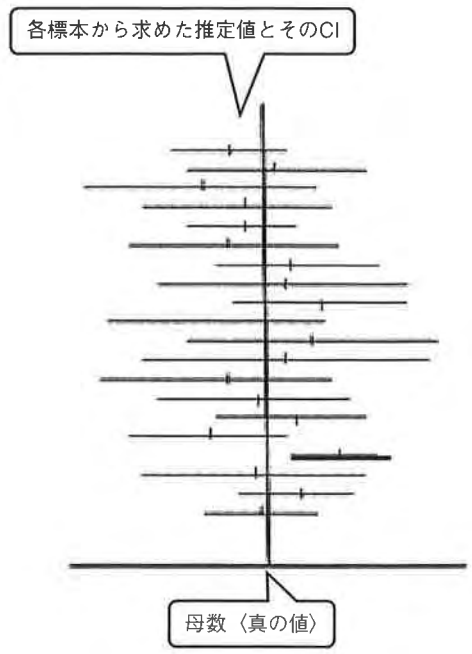

図　95％ CIのイメージ図

降圧薬Xと対照薬の血圧変化量の平均値の差は「真の差」の推定値である。同じ15.5mmHgという値であっても，それがばらつきの多い少数のデータから得られた場合と，粒のそろった多数のデータから得られた場合で，信頼性が異なることは容易に想像がつくであろう。

CIとSEの関係

CIは推定値の信頼性の指標として用いられる。CIは標準誤差（standard error：以下，SE）と深い関係にある。一般には，SEの計算はかなり複雑な式を用いる必要があるが，１群のデータの平均値のSEは標準偏差（以下，SD）/$\sqrt{n}$という簡単な式で求まる。この式から類推できるように，データのばらつきの指標であるSDが小さいほど，また，データ数（n）が大きいほど，SEは小さくなる。SEもまた推定値の信頼性の指標であり，値が小さければ小さいほど得られた推定値（データの平均値）は信頼できる。SEからさらに一歩進めて，その推定値が従うとされている確率分布（平均値の場合はt分布）を用いてCIを計算する。

平均値だけでなく，比率（割合），相関係数，回帰係数，中央値，分散など，どのような統計量でもCIを求めることができる。２群間でデータを比較する場合は，平均値の差や比率（割合）の差，あるいは，リスク比やオッズ比などのCIを求めることもできる。

CIの性質

(1) データ数を大きくすると，CIは狭くなる：データ数が小さい場合，極端な値をもつデータの影響を受けて，母数とはかけ離れた推定値が得られる可能性があるが，データ数を増やせばそのようなデータの影響が薄まり，その分信頼度が高まると考えることができる。

(2) 信頼係数を大きくすると，CIは広くなる：1つの統計量に対して，95% CIと99% CIの両方を求めた場合，95% CIには20回に1回，母数が含まれない可能性があるが，99% CIにはその可能性は100回に1回しかない。つまり，慎重に区間推定を行いたい場合は信頼係数を大きくして，CIを広めにとっておけばよい。

(3) 同じ確率分布を利用して母数のCIを求めれば，検定と同じ結論が得られる。すなわち，有意水準5%で平均値の*t*-検定を行うことと，*t*分布を利用して母平均値の95% CIを求めることは同等である。同様に，有意水準1%で検定を行うことと，99% CIを求めることは同等である。

p値とCIの報告の仕方 To Do & Not To Do

p値はあらかじめ有意水準を決めず，そのままの値を報告するよう推奨されている（例：$p<0.05$ではなく，$p=0.035$）。$p>0.05$の場合も，not significant (N.S.) とは書かずp値を報告する。p値が非常に小さい場合は従来通りでよい（例：$p<0.001$）。帰無仮説を棄却するか否か，微妙な場合は，例えば「増加する傾向がある（$p=0.065$）」として，過去の類似の研究の結果などを示したうえで，最終的な判断を論文の読者に委ねるべきである。

検定統計量（*z*値や，*t*値，*F*値，χ^2値など）はp値を求める手段であって，それ自体が別の情報をもっているわけではないので報告する必要はない。ただし，雑誌によっては，査読者や読者がp値を検証することができるように，検定統計量や自由度（degree of freedom：df）を併記することを求められる場合がある。

CIは±を用いず，下限と上限を報告する（CIは常に推定値を挟んで対称とは限らない）。多くの医療分野の雑誌で，できるだけp値とCIを併記することが推奨されている。

例）降圧薬X群と対照薬群における血圧変化量の平均値に差がある（$p=0.023$）。血圧変化量の平均値の差は15.5mmHg（95% CI：3.5～27.5mmHg）。

引用文献

1）山口拓洋："サンプルサイズの設計"，健康医療評価研究機構，東京，2010.

2）丹後俊郎，上坂浩之："臨床試験ハンドブック－デザインと統計解析－"，朝倉書店，東京，2006.

参考文献

1．M.J. Gardner, D.G. Altman："信頼性の統計学"，舟喜光一，折笠秀樹訳，サイエンティスト社，東京，2001.

2．奥田千恵子："医薬研究者の視点から見た道具としての統計学　第2版"，金芳堂，京都，2011.