EBNのための研究計画

統計学的推論：推定と検定

林　邦彦・高木廣文

EBNURSING Vol.3 No.4 2003〔別刷〕 中山書店

EBNのための統計の読み方 12

EBNのための研究計画
統計学的推論：推定と検定

林　邦彦 Hayashi Kunihiko
群馬大学医学部保健学科

高木廣文 Takagi Hirofumi
新潟大学医学部保健学科

「ハリー，君は何かを学ばなかったかね。我々の行動の因果というものは，常に複雑で，多様なものじゃ。だから，未来を予測するということは，まさに非常に難しいことになるのじゃよ」　（ダンブルドア校長）[1]

統計学的推論

前号では，調査や実験などで収集したデータについて，その「要約」や「データ構造」の記述を目的とした解析法の話であった．今号では，単に手元にあるデータの様子を把握するだけではなく，そのデータを利用して，背後にある自然法則を推論し，母集団や未来における予測に役立てることを考えてみる．

◆「推定」と「検定」

自然法則を推論する統計学の代表的な方式に，「推定（estimation）」と「検定（test）」がある．また，自然法則をつかさどるものや，集団の真の特性を示すものを「母数（parameter）」という．日常会話で，「ヨーグルトが6割の人で効くっていっていたけれど，母数が5人の調査じゃ信用できないよね」などというように，データ数や分数の分母（denominator）のことを「母数」という人がいるが，統計学の用語としては誤用といえる．自然法則をつかさどる母数の値は，残念ながら直接に測定したり観察したりすることができない．そこで，母数の値を手元のデータから何とか推し測ろうとするのが「推定」である．

一方，相反する2つの仮説を用意して，手元のデータから正しそうな仮説を選択するという，二者択一的な方式で自然法則を推論するのが「検定」である．

◆「母集団」と「標本」

ここで，統計学の教科書でまず登場する，「母集団（population）」と「標本（sample）」について考えてみよう．

総勢520人の看護師が勤務している病院看護部で，看護師が使う手洗い用洗浄液として製品Aと製品Bのどちらを使うか，意見を聞いて決めようと思う．全員に聞くのは困難なので，全体の1/20の26人に意見を聞くことにした．

この場合，「母集団」は520人の看護師全員と明示的に表すことができる．意見を聞くのに選ばれた26人が「標本」である．この標本のデータから統計学的推論を用いて，全員の様子を推測し，一般化する．そのため，標本26人は看護部全員の様子を代表するように選ぶ．全員から無作為に抽出したり，病棟ごとの看護師数に応じて代表者を選ぶなどの工夫をする．世論調査や一般住民対象の有病調査の場合などには，選挙人名簿や住民台帳から母集団を明確に定義して，層化多段階無作為抽出法などで標本を選び出す．

しかし，医学分野の多くの研究では，母集団から抽出し，母集団の特徴を保持した「標本」に基づくデータとは単純にはいかない．たとえば，米国のH大学病院に入院した26人の肺癌患者を対象に，治療法Aと治療法Bのどちらが効果で優れているかを調べ，統計学的推論を用いた研究があるとしよう．標本数は前例と同じ26人だが，この研究の母集団は一体，何だろう．対象になった入院患者のほかに，H大学病院に外来通院中の肺癌患者や，昨日新たに肺癌と診断された患者は母集団に含まれるのか．また，米国内の他病院や日本の病院に入院している肺癌患者などは，どうであろう．前述の520人の母集団の例とは違い，大変にあいまいである．母集団が明示できないので，標本を無作為に抽出することもできない．統計学的推論さえ使えば，すぐさま何にでも推論結果を一般化できるわけではないことに注意してほしい．

◆統計学的推論における一般化

医学研究では，推論の一般化について，次のように段階的に集団を分類して考えることが多い[2]．「白人の肺癌患者」といった研究結果を適用できると考えるいくつかの集団群を「想定母集団群（target populations：標的集団）」という．想定母集団群のなかで「2003年にH大学病院に入院した肺癌患者」といった実際のデータ源となる集団を「データ源集団（source population）」という．データ源集団のなかの「PSが2で，80歳以下の転移がない肺癌患者」といった研究対象の基準を満たす集団を「対象適格集団（eligible population）」，そして対象適格集団のなかでも「研究参加に同意し実際にデータを提供する患者」が「研究参加者集団（participant population）」となる．もう，population（集団）だらけである．結果を一般化しようとする想定母集団群と，実際にデータが収集される研究参加者集団との間の各段階で，特徴の保持をゆがめる可能性があることがわかるであろう．前述の母集団と標本のような単純な関係ではない．しかし，統計学的推論ではそれを前提としている．

推定の考え方

◆分布の性状

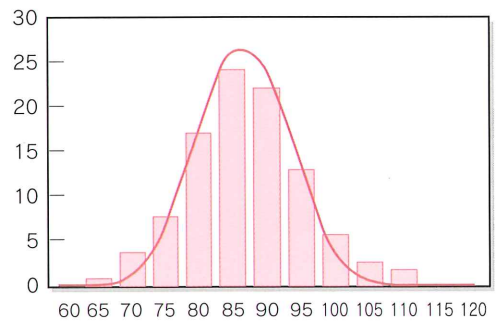
さて，実際に手元にあるデータを使って，「母数」を推定してみてみよう．

糖尿病や過度の肥満の人を除いた健康な30歳代女性100人の空腹時血糖値のデータを利用して，30歳代の一般健康女性集団（母集団）では空腹時血糖値はどんな値となるか推測する．❶に，標本100人の空腹時血糖値の度数分布を示す．

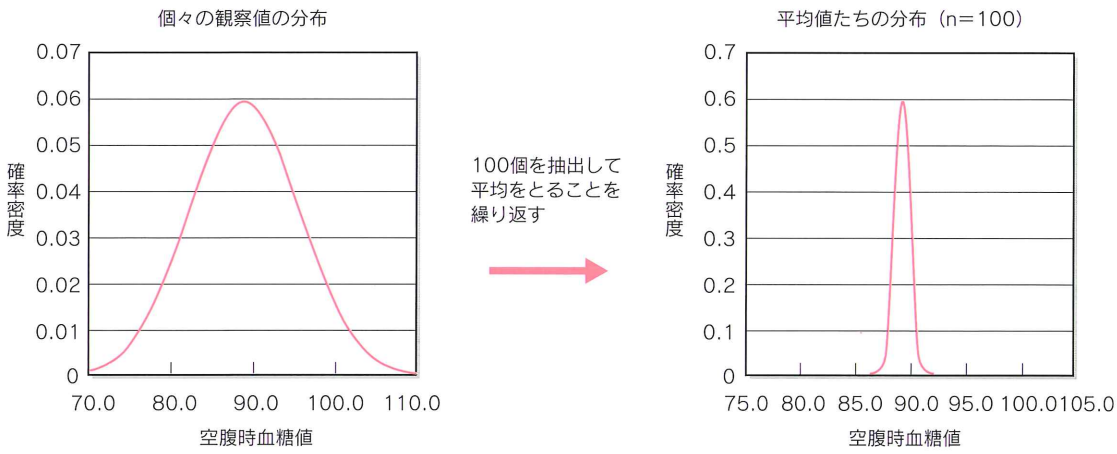
「標本データ」の平均値 $\bar{x}$ は88.97（標本平均），不偏分散 $\bar{s}^2$ は$(6.73)^2$である．これらを使って「母集団」での値の分布の様子を推測するには，母集団での平均である「母平均 μ」を「標本平均 $\bar{x}$」で，母集団での「母分散 σ^2」を「不偏分散 $\bar{s}^2$」で推定する．σが標準偏差である（前号を参照）．もし，母集団で各人の空腹時血糖値が，❶の曲線のように左右対称釣鐘型の正規分布（normal distribution）としてばらつくならば，その様子は

確率密度関数

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



❶ 空腹時血糖値の度数分布と正規分布近似



❷データの分布とその平均値の分布

と表すことができる．ただし，πは円周率を示す．これは，そもそも天文学での星の観測の誤差を説明するために，Gaussが提案したので，ガウス分布ともよばれる[3]．位置を決める「母平均μ」と，ばらつき度合いを決める「母分散σ^2」（もしくは標準偏差σ）の2つの未知母数を推定できれば，この分布の性状を決めることができる．

実際のデータから推定した分布が❷の左図である．横軸にx，縦軸に$f(x)$をとってある．曲線下の全面積はちょうど1となる．また，$\mu \pm \sigma$（$=88.97 \pm 6.73$：82.24～95.70）の間の曲線下面積は0.682，$\mu \pm 1.96\,\sigma$（$=88.97 \pm 13.19$：75.78～102.16）の間の曲線下面積は0.950となる．つまり，平均値から標準偏差の大きさを上下にとった区間に全体の68.2％が，標準偏差の1.96倍の大きさを上下にとれば全体の95％が分布する．

◆標準誤差と信頼区間

次に，この母集団の分布から，今回の標本数である100個のデータを選んでは平均値を計算することを繰り返し行ったことを考えてみよう．毎回算出する平均値は，100個のデータの選び方次第で少しずつ異なるだろう．平均値たちを無数に算出できたとして，その平均値たちはどのように分布するかを，正規分布を用いて推測したものが❷の右図である．その分布の中心の位置は，やはり母平均の推定値である標本平均88.97にくるだろう．

左図と異なる点は，ばらつき度合いである．平均値たちの分散は，母分散σ^2を標本数nで割った$\frac{\sigma^2}{n}$である．その平方根をとった$\frac{\sigma}{\sqrt{n}}$を「標準誤差（SEM：standard error of mean）」とよぶ．

100人の空腹時血糖データからは，平均値の標準誤差は$6.73/10 = 0.673$である．標準誤差の式からわかるように，同じ母集団から抽出した標本による平均値でも，その標準誤差の値は標本数によって異なることがわかる．標本数が2倍になると，標準誤差は$\frac{1}{\sqrt{2}}$倍になり小さくなる．標本数が大きなデータ・セットから推定したほうが平均値たちのばらつきは小さくなり，❷の右図での分布の山はより尖ったものになる．つまり，標本数が大きいほど算出した平均値たちの揺るぎが少ないと考えていいだろう．この平均値たちの揺るぎの程度，別の表現をとれば，算出した平均値の確信の程度を示す方法に，「信頼区間（CI：confidence interval）」がある．

$(100-\alpha)$％信頼区間は，その区間に母数の真値が$(100-\alpha)$％の確率で含んでいることを示し，

$$\bar{x} - Z(\alpha\%)\cdot \mathrm{SEM} \leq \mu \leq \bar{x} + Z(\alpha\%)\cdot \mathrm{SEM}$$

によって求められる．(100 − α)%は「信頼係数」といい，95%がよく用いられる．$Z(\alpha\%)$ は標準正規分布[注1)]の両裾の面積がα%となる点の値を表し，$Z(10\%) = 1.645$，$Z(5\%) = 1.960$，$Z(1\%) = 2.576$ である．上記の100人の空腹時血糖値から推定した母平均の95%信頼区間は，

$$88.97 - 1.960 \times 0.673 < \mu < 88.97 + 1.960 \times 0.673$$

となり，87.65～90.29の間に95%の確率で母平均があると推定される．確信の程度が低くてよければ90%信頼区間を，高くしたければ99%信頼区間を計算すればよい．

このように，母平均の位置について，標本での平均値という一つの値でぴたりと推定する方法を「点推定」といい，95%信頼区間のような値の幅で推定する方法を「区間推定」という．

この信頼区間は，「母比率」の推定にも利用できる[4)]．先ほどの30歳代女性100人のうち，現在喫煙をしている人は12人であった．この標本での喫煙割合pは12/100 = 0.12(12%) であるが，この値は母集団での喫煙の「母比率π」の推定値となる．この「標本比率p」の分散は$p(1-p)$，標準偏差SE(p)は$\sqrt{\frac{p(1-p)}{n}}$ となる．そこで，母平均のときと同じように，母比率πの95%信頼区間は，

$$p - Z(\alpha\%)\cdot \mathrm{SE}(p) \leq \pi \leq p + Z(\alpha\%)\cdot \mathrm{SE}(p)$$

によって求められ，

$$0.12 - 1.960 \times 0.0325 < \pi < 0.12 + 1.960 \times 0.0325$$

と，母集団の30歳代女性での喫煙母比率は0.0563～0.1837（つまり，5.63%～18.4%の間）にあると，95%の確率で推定できる．

検定の考え方

◆帰無仮説と対立仮説

次に，もう一つの統計学的推論の方式である「検定」に話を移そう．

前述のH大学病院での肺癌治療法の研究の例で，治療法A・Bの違いが改善・悪化という病状の変化に関係があるか推論することを考えてみよう．このような研究での作業仮説は，今まで標準的に用いられてきた治療法Aに比べて，新治療法Bが優れていることを示すことなのかもしれない．しかし，「統計学的検定」では相反する2つの仮説，「帰無仮説H_0 (null hypothesis)」と「対立仮説H_1 (alternative hypothesis)」を用意する．2つの仮説を合わせると，すべての状況が記述されているように設定する．肺癌治療法の例であれば，

帰無仮説　H_0：治療法と病状は独立している
　　（治療法間で病状は同じ）
対立仮説　H_1：治療法と病状は従属している
　　（治療法間で病状が異なる）

とすれば，この2仮説以外には治療法と病状の関係を示す状況はなくなる．であれば，この2仮説のいずれか正しいほうを一定ルールのもとで選ぶことができれば，真実が推論できる．

仮説選択のルールは，あたかも二重否定のように回りくどい．まず，自分の主張したい仮説の逆である「帰無仮説」の立場をとってみる．帰無仮説H_0が正しいと仮定して，今回得られたデータはどの程度起きやすいのかを，「検定統計量」(t検定でいえばt値が検定統計量）からみてみる．算出した検定統計量が事前に設定していた限界値を超える，もしくは検定統計量から算出したp値が小さくて，今回のデータが示す状況は帰無仮説H_0のもとでは極めて起き難

注1）標準正規分布：正規分布のなかでも平均0，分散1の分布を「標準正規分布」(standard normal distribution）とよぶ．統計学の教科書の多くに，この標準正規分布のZ値とα%が，数表として載っている．

	病状改善(人)	病状悪化(人)	計(人)
治療法A	3	10	13
治療法B	9	4	13
計	12	14	26

❸ 米国のH大学病院での肺癌治療の成績（仮想データ）

		検定の結果	
		H_0を棄却できず	H_0を棄却（H_1を採用）
真実	帰無仮説H_0	$1-\alpha$	α
	対立仮説H_1	β	$1-\beta$

α：第1種の過誤，β：第2種の過誤

❹ 検定における2種類の過誤の確率

いこと（たとえば，起こる確率が5％未満）と考えられた．しかし，それは大変まれなことが起きたと考えるよりは，前提としていた帰無仮説H_0が誤っていたとして捨て去る（棄却：rejectという）ほうが合理的であろう．帰無仮説H_0を捨てると生き残った仮説は今や対立仮説H_1のみとなって，自分の言いたかった「治療法間で病状が異なる」という対立仮説H_1が正しいことが検証される，という筋書きである．ややこしい筋書きだが，まず自分の意見と反対の立場に立って考えてみることが，統計的検定だけでなく社会一般で重要なようである．

この論法を，肺癌治療の例（❸）でχ^2検定[注2)]を使って追ってみる．まず，帰無仮説が正しいとして，今回の状況はどの程度起きやすいのか調べる．Yatesの補正したχ^2値は，

$$\chi_0^2=\frac{26\left(|3\times4-10\times9|-\frac{26}{2}\right)^2}{13\times13\times12\times14}=3.869$$

となる．この値は，事前に設定した自由度1の有意水準5％であるχ^2値の限界値 $\chi^2(1, 0.05)=3.841$ を超えていた．また，χ_0^2値からp値を計算した場合でも$p=0.0492$となり，これは帰無仮説の下では5％未満でしか起きない状況であることがわかる．そこで，前提としていた帰無仮説H_0が誤っていたとして棄却する．すると，対立仮説H_1が残り，治療法間で病状は統計学的に有意に異なると推論することになる．

これで，H大学病院では，胸を張って治療法Bを勧めるようになるかもしれない．

しかし，この論法の弱点は，帰無仮説が棄却できないときに表われる．二重否定的な論法になっているので，否定できなかった（帰無仮説H_0を棄却できなかった）場合には，論理が途中で止まり，態度保留となってしまう．帰無仮説H_0が棄却できないからといって，積極的に帰無仮説H_0が正しいとはいえないのである．

◆検定で起こり得る2種類の過誤

実際，帰無仮説H_0が棄却できないときには，真実H_0が正しい場合と，実は対立仮説H_1が正しいのにそれが検出できない場合がある．❹を参照してほしい．

検定には，2種類の推論の誤りが起こり得る．第1種の過誤（type-I error）の確率αは，帰無仮説H_0が真実正しいにもかかわらず，誤って棄却してしまう確率である．検定で仮説選択の基準とする5％といった有意水準（significant level）は，この確率の大きさのことである．このように，第1種の過誤の大きさは通常は5％未満におさえられていることになる．

検定でのもう一方の推論の誤りが，第2種の過誤（type-II error）である．第2種の過誤の確率βは，

注2）χ^2検定：2×2分割表において独立性の検定として用いられる．n数が小さな場合には，フィッシャーの直接確率法が用いられるが，n数がある程度以上あれば，Yatesにより考案された連続修正（χ^2の式の分子の$-\frac{n}{2}$の部分）したχ^2検定が用いられる．
また，1×m分割表での自由度は$(l-1)\times(m-1)$であり，2×2分割表では自由度は1となる．

対立仮説H_1が真実正しいにもかかわらず，帰無仮説を棄却できない確率となる．あまりにも標本数が小さな研究や，測定や観察が不正確な研究では，この見逃しの確率βが大きくなってしまう．真実は対立仮説H_1が正しいときに，検定で帰無仮説H_0を棄却し，対立仮説H_1が採択できる確率$1-\beta$を「統計学的検出力（statistical power）」という．研究計画では，この検出力が80～90％程度になるように設計される．

◆両側検定と片側検定

また，検定には「両側検定」と「片側検定」がある．前述の癌治療法の例では，研究の作業仮説は，治療法Aに比べて新治療法Bが優れていることであった．しかし，設定した対立仮説は「治療法間で病状が異なる」である．「治療法間で病状が異なる」には，「治療法Bのほうが治療法Aより病状がよい」と「治療法Bのほうが治療法Aより病状が悪い」の2つのケースがある．両方のケースを含んで，対立仮説を「治療法間で病状が異なる」とすれば，それは「両側検定（both sided test）」をすることになる．新治療Bのほうがずいぶんよいだろうと思って実施した研究でも，実は今までの治療法Aのほうがよいということも起こり得る．多くの場合，両側検定が使われる．しかし，理論的に片方のケースしか想定できない場合には，対立仮説を「治療法Bのほうが治療法Aより病状がよい」，もしくは「治療法Bのほうが治療法Aより病状が悪い」のいずれかにして「片側検定（one sided test）」を行うことになる．

おわりに

統計学的推論として，「推定」と「検定」の考え方を解説した．医学論文や学会発表の図表で，検定結果を有意水準5％で有意であった個所に＊，1％で有意を＊＊，0.1％で有意を＊＊＊と表示し，「差は小さくても，統計学的に有意」とか，「星3つです．きわめて有意です」など，まるで食通のレストラン情報のように使われることもある．しかし，看護や医学の研究では，有意か否か二者択一方式の推論より，差はどの程度といった大きさに関する母数の推定値のほうが重要な情報となることも多い．

また，検定では，帰無仮説が棄却できて初めて有効な仮説選択ができ，帰無仮説が棄却できないときは選択保留とならざるを得ない．同じであることを積極的に検定で示すには，「同等性の検定」や「非劣性の検定」とよばれる特殊な仮説構造が必要となる．ぜひ「推定」と「検定」という2つの異なる統計学的推論の長短所を理解して，上手に使い分けてほしい．

文献

1）ローリング J.K.：Harry Potter and the prisoner of Azkaban．松岡祐子訳：ハリー・ポッターとアズカバンの囚人．静山社；2001．
2）林邦彦：コホート研究の読み方・チェックリスト．EBM ジャーナル　2001；2(5)：60-63．
3）Walker AM：Observation and inference－An introduction to the methods of epidemiology．Epidemiology Resources Inc.，MA；1991．丸井英二，中井里史，林邦彦訳：疫学研究の考え方・進め方－観察から推測へ．新興医学出版；1996．
4）柳井晴夫，高木廣文編：新版看護学全書基礎科目 統計学．メヂカルフレンド社；1995．