連載　統計学のキー・ポイント―「検定」に焦点を当てて・1［新連載］

検定の基本的な考え方
Fundamentals of Statistical Hypothesis Testing

髙木廣文

看　護　研　究
第47巻 第1号 別刷
2014年2月15日発行
KANGO-KENKYU(*The Japanese Journal of Nursing Research*)
Vol.47 No.1/Jan.−Feb.2014

医学書院

［新連載］

統計学のキー・ポイント「検定」に焦点を当てて

第1回

検定の基本的な考え方

Fundamentals of Statistical Hypothesis Testing

髙木廣文　東邦大学看護学部長

はじめに

統計学は，人間を対象とする研究では必ず使わなければならないデータ解析ツールです。近代看護の祖であるナイチンゲールが統計学の専門家であったことは，よく知られた事実です。しかし，ナイチンゲールが活躍した頃には，まだ使うことができなかった便利な統計学的方法が，現在では誰でも使うことができます。それが「検定 test」です。正確には，「統計学的仮説検定」ですが，本稿では「検定」で済ませることにします。（何が便利なのかは，あとで説明しますので，少しお待ち下さい）

現在では，パソコン用の統計解析ソフトが普及していることもあり，あまり統計学の知識がなくとも，卒業研究でもやろうと思えば高度な解析方法を用いることも可能です。しかし一方では，解析はしたものの得られた分析結果をどのように解釈したらよいのかわからない，または，一応解釈したもののその解釈に自信がもてないなど，少なからぬ研究者が悩みを抱えていることもあるようです。

現在，ほとんどの大学や研究施設では，研究を開始する前に倫理審査を通過する必要があり，その場合に，研究デザインが問題になることが多くあります。倫理審査委員会で，なぜその研究では対象者数が100名ではなく200名必要なのか？など，標本数を決めた科学的根拠を示すことを要求されることがよくあります。

これらの問題をクリアするためには，どうしてもある程度は，検定についての正しい知識が必要です。さらに，用いる検定で必要とされる標本数がどのくらいになるのかという，研究デザインに直結する知識も要求されることでしょう。

この連載は，検定に関するこれらの点について，可能な限り平易に解説することを目的としています。ただし，統計学は実用の学問ですので，解説を読んだだけではピンとこない可能性が大きいので，できれば読者自身の行なっている研究に即して，実際に例題などに取り組んでみることをお勧めします。

●連載予定

0022-8370/14/¥360/論文/JCOPY

記述統計学と推測統計学

統計学は，研究のために集めたデータを，目的に応じて加工するための技術を提供するものですが，その目的は大きく2つに分類することができます。すなわち，「記述統計学 descriptive statistics」および「推測統計学 inferential statistics」です。

記述統計学とは，調査データなどをもとにして，ある現象について記述的な情報を提供するものです。例えば，ある病院の外来患者は1日何人来るのか，そのうち糖尿病患者は何人なのか，糖尿病患者のうち男女の割合はそれぞれ何%なのか，平均年齢は何歳なのか，などのような疑問に答えるための情報を提供するための方法を与えるものです。いわゆる「統計をとる」といわれる形で人数を数えたり割合を求めたりするような，単純な集計や方法が記述統計にあたります。

一方，推測統計学は，データに基づいて，そのもとになっている母集団の特性を推定したり，ある仮説の成立の有無を決定するために用いられる方法です。したがって，記述統計学に比べて，かなりその理論的背景は面倒なものになっています。

まず実際にデータを集める調査対象である「標本」と，その背景にある「母集団」の関係などを正しく理解する必要があります。さらに，標本から母集団に関するある特性の推定や仮説検定を行なう意味を正しく理解する必要もあります。これらの点をマスターするのはそれほど容易ではないためか，記述統計学の方法である集計方法を学習するのには困難を感じないのに，推測統計学の方法になると，統計嫌いになったり挫折したりする研究者の割合が増加するという，その原因になっているのではないかと考えられます。

実際の研究で必要とされる統計学的方法は，主に推測統計学の方法です。このことが，正しい研究デザインを構築したり，データ解析とその解釈を難しいものにしているため，統計学に関する多くの誤解のもとになっているように考えられます。もっとも，推測統計学の方法も，順番に学習を進めていけば決して難しいものではないことが理解できることを期待して，本稿を進めていきたいと思います。

母集団と標本

糖尿病患者の健康教育に興味があり，新しい教育内容と従来の方法とを比較する研究を行なうことにしたとしましょう。このような場合に，すべての糖尿病患者を研究の対象にするわけにはいきませんし，できるわけもありません。通常は，例えば200人の患者さんに研究内容を説明し同意が得られたとすると，無作為に100人ずつ新教育グループと旧教育グループに分け，教育開始前と教育終了後3か月後のグリコヘモグロビンを調べ，その平均値などを比較して，新旧の教育効果の差を検討することになります。このような方法では，すべての糖尿病患者を2グループに分けるわけではないのですが，新旧2つの教育方法による100人ずつのグループは，すべての糖尿病患者の2グループの代表になっているものと考えて，その比較を行なっていることになります。

理論的に正しい因果推論の方法では，すべての糖尿病患者が新しい教育方法で行なった場合の結果と，すべての糖尿病患者が旧教育方法で行なった場合の結果を比較しなければなりません〔カウンターファクチュアルモデル counterfactual model（高木，林，2006）〕。カウンターファクチュアルモデルを実施することは，実際には不可能ですので，現実には全体のほんの一部を使った研究が行なわれています。このとき，概念上の全体にあたる集合を「母集団 population」と呼びます（すべての糖尿病患者など）。また，実際に研究対象となる集合を「標本 sample」と呼びます（研究に参加した

200人の糖尿病患者など)。

母集団から標本を選ぶ方法は無作為(ランダムrandom)でなければなりませんが，これらの詳細については多くの統計学の成書で説明されており，ここで多くを述べる必要はないと思いますので，省略させていただきます。

推測統計学の方法は，ランダムに選ばれた標本データをもとにして，母集団についての特性値を求めたり，仮説を立証したりするためのものです。この点について，次節でもう少し詳しく説明します。

推定と検定

推測統計学の方法は，大きく2つに分けて考えることができます。すなわち，「推定estimation」と「検定」です。

推定は，日常的にも使われる言葉であり，調査などで得られた標本のデータから母集団のある特性値を推測するための方法です。例えば，電話調査で内閣の支持率を推定する場合や，TVの視聴率調査などを思い浮かべるとよいでしょう。全国の1000人に無作為に電話をかけ，現在の内閣を支持するかしないかを聞き取り，600人が支持すると回答した場合，全国の内閣支持率は60%である，というように推定することになります。ある特性の割合や平均値などは，後述する理由から，標本で求めた数値がそのまま推定値になりますので，あまり深く考えたり悩む必要もないので，多くの研究者も，ここで統計学につまずくことはないようです。

ただし，重要な点は，標本データの数値と一致していても，あくまでも求めているのは「母集団での特性値」であるという点です。これは検定でも同じことなので，この点についてはよくよく頭に入れておく必要があります。

一方，検定は，ある仮説の成立・不成立を決定するための統計学的な方法です。このために，判断の基準となる確率を計算する必要があります。このため，推定の方法とは大きく異なった感じを与えます。

統計学的な手続きとして，検定では，①仮説を立てる，②確率を計算する，③確率から仮説の成立・不成立(棄却)を決定する，という手順が必要になります。このために検定にまつわる問題として，仮説の必要性や意味がわからない，使用する計算式が理解できない，計算で求めた確率の意味がわからない，仮説の成立・棄却が何を意味しているのか解釈がわからない，などの多くの問題がしばしば起こります。これらの点が，現実に多くの統計学の学習者が挫折するところでもあると思われます。

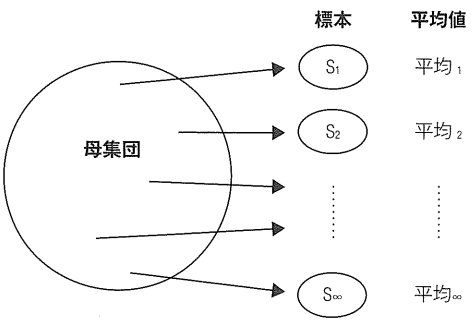
しかし，人間を対象とする研究について，現在の統計学がこれほどの必要性をもっているのは，少ない標本のデータから，母集団に関するさまざまな仮説を検定し，一般性のある結果を示すことができるからなのです。実際の検定に必要な考え方はそれほど難しいものではないので，焦らず解説していくことにしたいと思います。

期待値と中心極限定理

普通，我々は調査を行なう場合でも，何回も繰り返して行なうことはまれなので，どうしても調査は1回しか行なわないものであるように考えてしまいがちです。ところが実際には，母集団から標本を抽出する方法は無限にあるといってよいでしょう。例えば，100人から10人を選んで調査する場合の数は，簡単な計算により，

$$\text{組合せの数} = {}_{100}C_{10} = \frac{100!}{10! \times 90!}$$
$$= 17{,}310{,}309{,}456{,}440 \text{ 通り}$$

図1　無限の標本抽出と標本平均について



となります。すなわち，17兆3100億通り以上の標本の選び方があります。このように，ほんの小さな母集団からの標本抽出であっても，ほぼ無限といってよいくらいに膨大な数の標本の選び方があるのです。したがって，より人数の多い集団からの標本抽出調査では，標本の取り方は**図1**のようにほぼ無数あるということがわかると思います。

このことは，それぞれの調査では，ある特性，例えば内閣の支持率や血圧の平均値なども，それぞれの調査ごとに無数に存在することになります。当然ですが，それぞれの調査で得られた割合や平均値は，ある場合には一致するかもしれませんが，すべて完全に一致することはあり得ないことで，ほとんどの場合にはそれぞれ別々な数値をとることになります。

したがって，統計学的な推測を行なう場合には，このようなすべての標本の割合や平均値がどのように分布するのかを考えて，推定や検定の方法を構築する必要があります。自分が一度しか調査をしなくても，実際には可能であった調査は無数に存在することを考える必要があるのです。

我々に都合がよいことに，調査が無作為に行なわれたのであれば，標本から求められた割合や平均値については，可能なすべての標本から求めた割合の平均値や，平均値の平均値(期待値 expected value)は，母集団での割合(母割合，母比率)や母集団での平均値(母平均値)と一致することがわかっています。このため，無作為抽出による標本データから求めた割合や平均値が，それ以上は何の処理もされずに母集団の推定値として使用されることになるわけです。つまり，電話調査で1000人を対象に内閣の支持・不支持を聞いた場合，650人が支持すると回答したのならば，そのまま65%という数字が内閣支持率として，母集団(日本国民)の内閣支持率を示すものとして扱ってよいことになります。

平均値や割合の期待値については，それほど理解するのは困難ではないと思います。問題は，標本で得られた平均値がどのような分布になるのかという点です。結局のところ，多くの統計の検定理論は，ある統計量についての平均値と分散を求めて，その分布から確率計算を行なうという手順を踏みます。したがって，標本平均がどのような分布になるのかは，きわめて重要な問題になります。

統計理論，特に検定理論では，この標本データから求めた平均値の分布の問題に対する解答として，統計学の公理ともいうべき「中心極限定理 central limit theorem」(高木，2009)を与えています。

●中心極限定理とは

中心極限定理とは簡単にいうと，「分散が存在するような，ある分布に従う変数について複数個のデータの合計は正規分布に近似的に従う」という定理です。標本数が多くなれば，それだけ正規分布へのあてはまりがよくなります。

中心極限定理が重要なのは，極端な多峰性の分布でなく，一峰性の分布であれば，もともとが正規分布でない分布から得られたデータであっても，その合計の分布は正規分布に従うという点にあります。標本データでは，平均値と分散が求められないという場合はあり得ないでしょう。ただし，理論的な分布では，例えば，t分布で自由度が1の場合には，「コーシー分布」という特殊な分布になります。コーシー分布は，両裾が厚いために分散が発散して計算ができないという特徴があります。

通常は，標本から得られたデータの合計を標本数で割れば平均値になりますので，中心極限定理により，標本のデータから得られた平均値の分布は正規分布に近似的に従うことになります。多くの統計理論は，特定の変数の平均値と分散に基づいていますので，よほど特殊で変な分布でなければ，中心極限定理により，標本平均の分布の正規性が保証されることになります。

よくある誤解のひとつに，標本のデータが正規分布していないから，この検定は使えない，したがってノンパラメトリックでやらなければならない，という意見があります。確かに，もともと正規分布に従う変数から得られたデータであれば，その平均値の分布の正規性は保証されますが，そうでなくても，データの合計や平均値は中心極限定理により，正規分布に従うのです。データそのものの分布ではなく，データをまとめた統計量の分布についての定理である点を，正しく理解する必要があります。

統計学の発展期にはこのような議論が頻繁に行なわれ，2群の母平均値の差の検定などで使われるt検定は，用いるデータの正規性からの逸脱に対して，きわめて「頑健 robust」であることがわかっています。このように，中心極限定理の存在から，正規分布は統計理論の要になっているのです。

検定の考え方

さて，ここで検定の基本的な考え方を説明します。推定が，日常的に用いている「推定」という言葉と大差ない意味内容をもっているのに比べて，「検定」という言葉は，かなり異なった意味で統計学では使われているといえます。検定という用語は，普通は「能力検定」「英語検定」などのように，ある事柄について一定以上の力量をもっているか，知識をもっているかといったことを調べるために使われる言葉のように考えられます。検定は英語では test，つまりは「テスト」ということになります。テストというと，一般には「試験」という意味合いで使われているのではないかと思います。しかし，統計学的な意味でのテストとは，微妙にニュアンスが異なってきます。このように，同じ言葉が異なった意味で用いられると，混乱を来たすことがよくあります。このような問題を避けるためにも，統計学での検定の意味を，その手順にそって順次説明したいと思います。

いまはコンピュータを使えば，何も考えることなしに，データからさまざまな検定結果を得ることができますが，実際の(統計学的仮説)検定は，以下のような手順で行なわれます。

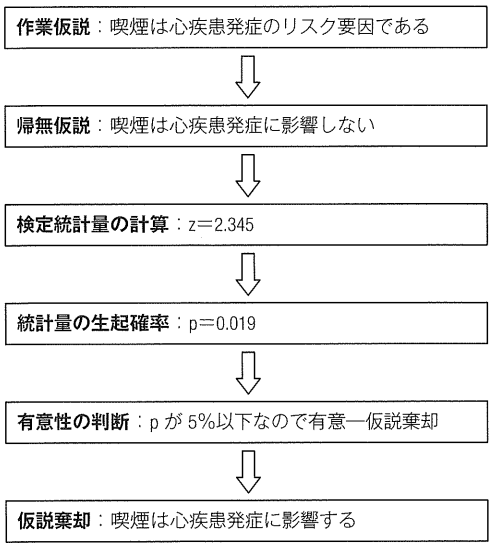
(1)帰無仮説を立てる
(2)検定のための統計量を求める
(3)検定のための統計量から観察されたデータの生起確率を求める
(4)有意性の判断をする
(5)有意性の判断結果から現実的な解釈をする

ここでやっかいなのは，「帰無仮説 null hypothesis」というあまり聞き慣れない用語です。ここを正確に理解できないと，その後に続く統計学の理論展開を正しく理解することができなくなります。検定の概略を**図2**に示しました。少し面倒ですが，きわめて重要な点ですので，図2を参照しながら，少し詳しく説明したいと思います。

普通は研究を行なう場合，研究上の仮説(作業仮説 working hypothesis)を設定します。例えば，喫煙が心疾患発症のリスク要因になるかどうかを研究する場合，研究の作業仮説としては，「喫煙は心疾患発症のリスクを高める」とか，より具体的に「喫煙は心疾患発症のリスクを2倍に高める」のように設定するのが一般的でしょう。

問題は，このように設定した作業仮説を，対象を観察したり調査したりして得たデータから直接立証するのは思っている以上に難しい，というこ

図2　仮説検定のプロセス



とです。例えば，多数の健常者を喫煙者と非喫煙者に分け，数年にわたって追跡調査を行なったところ，確かに喫煙者グループで心疾患の発生率が高かったとします。その結果から，作業仮説が証明されたといえるのでしょうか。研究者は作業仮説の通りの結果なのだから，これで喫煙の心疾患への影響は証明されたと結論したいところです。しかし，必ず次のような反論があります。すなわち，調査には必ず偶然の変動があるので，完全に2グループでの発生率が一致するよりも，どちらかでの心疾患の発生率が高く観察されることのほうがあり得ることである，というものです。

この反論に答える必要があります。すなわち，喫煙が心疾患のリスク要因になることを立証するためには，非喫煙者グループに比べて，喫煙者グループで心疾患の発生率が高かったのは偶然ではないことを，科学的に示す必要があります。

●作業仮説の検定法

このような目的のために考えられた方法が，統計学的仮説検定法です。まず，作業仮説を直接証明するのはきわめて難しいので，帰無仮説と呼ばれる仮説を設定します。最終的に帰無仮説は，否定されるためのものです。このため「帰無」という，通常では使用しない用語が使われています。例えば，作業仮説が「喫煙は心疾患発症のリスクを高める」ならば，帰無仮説としては「喫煙は心疾患の発症リスクを高めない」，言い換えると「喫煙の心疾患の発症リスクは1である」という仮説を設定します。

このように設定された帰無仮説のもとで，観察されたデータにみられるような関係や差が，偶然に生じる確率がどのくらいなのかを計算します。例えば，実際には喫煙の心疾患のリスクが3とデータから求められた場合，帰無仮説であるリスクが1の場合に，偶然にリスクが3と観察されるような確率を計算します。一般に，帰無仮説を仮定した場合に，観察されたようなデータを得られる確率を直接的に計算するのはきわめて困難です。このため，ある統計量を計算し，その統計量が正規分布やt分布などに従うという仮定のもとに確率を求めるのが，通常のやり方です。

検定の方法を難しく感じるのは，この確率計算のための統計量の計算式が，なぜこんな複雑な数式なのかよくわからなくて，計算式をみるだけでうんざりしてしまったりするからではないかと思います。

このようにして求められた確率(有意確率)が大きければ，調査によって得られたデータが示すような事態も偶然に生じることになります。そのような場合には，喫煙のリスクが仮に3と推定されていても，帰無仮説を否定する(棄却 reject)することはできません。したがって，喫煙が心疾患発症に影響する，ということはできません。ただし，影響しないともいえない点には注意が必要です。これは，有意確率が大きくても，帰無仮説が立証されたというわけではないからです。

一方，有意確率がうんと小さければ，起こらないはずのことが観察されたことになります。すなわち，偶然には観察されたような事態が起こるはずはないということを，示していることになります。起こるはずのないことが起きたというのはお

かしいので，これは確率計算のもとになった仮説が誤っていると考えて，仮説を棄却します。帰無仮説を否定することで，「喫煙の心疾患のリスクは1ではない」ことがいえます。必要ならば，信頼区間などを求めることで，喫煙により心疾患のリスクが高くなることを推論することができます。

帰無仮説を棄却するための有意確率の判断基準，すなわち帰無仮説を棄却するための基準値には5%が一般に使われています。求めた有意確率がこれより小さい場合に，検定結果は「有意であった」「有意差が認められた」「有意な相関が認められた」などのようにいわれています。検定のために用いられる確率は，このため有意確率と呼ばれるわけですが，最近では統計解析ソフトの出力で「p-value」とか「p値」と出力されるため，「p値(ぴーち)」呼ばれることも多くなっています。

統計学的な検定の方法は，あくまでも意思決定のために，ひとつの情報を与えているに過ぎないという点には注意が必要でしょう。有意確率の判断基準としてよく使われている5%という数値自体に，絶対的な科学的な根拠があるわけではありません。「まあこのくらいなら滅多にないのではないか」といった科学者集団でのコンセンサスとして，恣意的に決められている基準です。特に不便を感じないようなので広く使われていますが，5%を金科玉条のように考える必要はないのではないかと思います。研究内容や必要に応じて，10%や1%などでもよいはずですが，勝手に変更するにはそれなりの理由が必要です。論文を投稿する場合には，査読者を説得できる論理があるのならば，試してみる価値はあるかもしれません。

検定結果の現実的な解釈については，なかなか慣れないと難しい点もありますので，この連載を通じて解説していきたいと思います。少なくとも研究においては，統計学的な有意性を研究結果として示すことができた場合，それなりの現実的な解説が必要になります。単純に，「喫煙と心疾患には5%水準で有意な関係が認められた」だけではなく，「喫煙と心疾患には5%水準で有意な関係が認められ，喫煙により心疾患発生は3倍になることがわかった。したがって，心疾患の予防的行動として禁煙などの対策の一層の充実が必要である」くらいの記述は必要でしょう。

おわりに

きわめて簡単に，統計学で多用されている検定の概略を説明しました。現代では，パソコンの普及とともに統計解析用のソフトも出回り，誰でも簡単に統計解析ができるようになりました。便利になったのは疑いのない事実ですが，統計学の初学者であっても，また全く統計理論を知らなくても解析することが可能になりました。各検定の理論や高度な分析方法をすべて学ぶのは大変ですから，そういった研究者を非難したり否定したりするつもりはありません。ただし，当然ですが，ある程度自分が使っている解析方法にどのような理論的裏づけあるのかを知っておくのは，無駄にはなりません。統計解析の結果の解釈で思わぬミスなどをしないためにも，ある程度の知識は備えておくべきでしょう。この連載が少しでも役に立つように解説したいと考えています。

文献
• 高木廣文，林邦彦(2006)．エビデンスのための看護研究の読み方・進め方．中山書店，pp.45-51.
• 高木廣文(2009)．ナースのための統計学，第2版．医学書院，pp.50-51.

たかぎひろふみ●東邦大学看護学部
〒143-015　東京都大田区大森西4-16-20