

エビデンスのための統計学の基礎

高 木 廣 文 (東邦大学看護学部)

第59回日本小児保健協会学術集会 教育講演

エビデンスのための統計学の基礎

高 木 廣 文 (東邦大学看護学部)

I. はじめに

臨床医学の分野を中心に、EBM (Evidence Based Medicine) や EBN (Evidence Based Nursing) という用語で、研究にも臨床実践にも「エビデンス」が求められている。ある現象に対して科学的な証拠となるような研究成果をもたらすような研究や、科学的な証拠に基づく医療を提供するためのエビデンスという意味である。わざわざエビデンスなどと言うのは、これまでの医療は科学的な根拠なしに行われてきたのかと疑ってしまうことにもなりかねないが、研究や臨床実践にエビデンスが求められるのは、悪いことではない。EBM や EBN という用語は比較的最近のものであるが、そのために用いられる方法は、生物統計学や疫学を専門とする研究者の間では以前から使われていた方法である。

ここではエビデンスのための考え方、とくに統計学の視点からエビデンスをもたらす研究に必要なことについて、以下の4点から簡単な解説を行いたいと思う。すなわち、(1) 統計学とエビデンスの関係、(2) 統計学の特徴と科学性、(3) Counterfactual Model と研究デザイン、そして (4) 検定と標本数の関係、を中心に解説する。

II. 統計学とエビデンス

現在では研究の結果が「エビデンス」となるためには、データ解析に統計学を用いなければならない。人間を対象とするような研究において、統計学的な分析のない研究は、科学的な研究と見なされないといって

も過言ではないだろう。統計学的検定を用いて何らかの現象について「統計学的有意差」を見い出す必要があり、有意差のない研究は科学的なエビデンスをもたらさない。

しかし、よく考えてみると「科学研究」とは何かを明確にせずに話を進めるのでは、統計学の立場も不明確になる。ここではまず、科学とはどのようなことなのかを代表的な科学論などを紹介しながら考えていきたい。

1. Popper の反証主義

「科学とは何か」と問われると、多くの人は、「科学とは真理を明らかにする」とか「科学とはある現象の因果関係を明らかにする」とか答えるのではないだろうか。しかし、現実の科学者が行っている研究は、必ずしも物事の真理や因果関係を明らかにするためにやっている訳ではない。

Popper (1902~1994) は、科学と非科学の境界を設定するために、「反証主義 falsificationism」を提唱した。これは、『あらゆる科学理論は経験的科学体系にとっては反駁されることが可能でなければならない』という主張である。この理由は、科学理論の経験的検証は不可能であるという事実に基づいている。科学理論は、これまで起きた特定の現象を上手く説明できる必要があるのだが、さらに同様の現象の将来における予測が可能でなければならない。例えば、平成24年に日本では皆既日食が起きたが次回の皆既日食がいつ起きるかを予測することが、現在の物理学や天文学の理論では可能である。このように、科学理論は「普

遍言明」として『全ての時点において、全ての～について～が成り立つ』という形式をとる。しかし、個別の経験（単称言明）からは未来の出来事を含んでいる普遍言明を検証することはできない。この点から、帰納に基づく研究は科学的ではないと Popper は指摘している。逆に、普遍言明に合致しないような現象が一つでもあれば、その言明（科学理論）は誤りであると反証することができる。

この反証主義により、科学理論と非科学理論の区別が可能である。例えば、宇宙がどのようにして生まれたかについての理論として、A『宇宙は神が無から創造した』と B『宇宙は無からビッグバンにより創造された』という、よく似た2つの理論がある。『太陽はなぜ輝き続けているのか』という質問に対して、Aの理論に基づけば、『それは神の思召し』であるということで片がついてしまうので、反証することができない。一方、Bの理論ではまず宇宙がビッグバンを開始するための理論が必要であり、その詳細は省くが、その理論から恒星の内部では水素がヘリウムになる核融合により熱と光を発しているとの説明が可能である。さらに、ビッグバン後約137億年たった現在の宇宙空間の温度（背景輻射）が約3°Kであることが理論から予測できる。実際にこの温度を測り予測値とのずれを検証することで反証は可能である（実際に、理論値と観測値はよく一致していた）。

現実には、必ずしも正しくないと言われた科学理論であっても、すぐに捨て去られる訳ではなく、必要とされる修正などを行うことで、実際に役に立つことも多い。例えば、ニュートンの万有引力の法則は、アインシュタインの相対性理論の前に、理論としては誤りであることがわかったが、いまだに習うべき基礎的な科学理論として知られている。

実際の科学的な営為は、すでにアリストテレスが指摘しているように、ある現象について観察や実験などによる事実に関する知識に基づいて、その現象を説明する理論を構築し、次にその理論に基づいてその現象の説明が合致しているかを検討・修正するという一連の再帰的・螺旋的なプロセスであると考えられる。この点から、Popperの反証主義だけでは現実の科学研究の実情を的確に説明しきれていない。

2. Kuhn のパラダイム論

Kuhn (1962, 1970) によれば、『科学研究とは専門

化された各科学集団の中で、それぞれの「パラダイム paradigm」に従って問題を設定し、それらの解答を与えるという一種のゲームであり、科学者集団が行っている研究が科学である』ということになる。ここでパラダイムとは、『一般に認められた科学的業績で、一時期の間、専門家に対して問い方や答え方のモデルを与えるもの』であり、研究論文は科学者仲間のこのようなルールに従って記述されたものである。

現在では、科学はますます細分化し、科学的知識は膨大に蓄積されたが、科学者といえども専門外のことは理解できなくなるという事態を招来している。相手の用いている用語などを自分の意味理解に基づくと、相手は意味不明のことを述べていることになってしまう。このような事態を Kuhn は「通訳不可能性 incommensurability」と呼んでいる。

Kuhn のパラダイム論は、実際の科学者集団が行っている研究の実情をよく表しているのだが、『科学研究とはどのような研究なのか』という問いに対するわれわれの希望する形式の解答にはなっていない。

3. 構造主義科学論

池田 (1990) によれば、『科学とは同一性の追求である』。すなわち、科学とは現象を上手く説明するために、少数の同一性を用いて構造化することである。科学理論は、ある現象に関して、できるだけ少数の共通する要因を用いて説明できる必要がある。科学理論の良し悪しは、その現象説明力で評価されるので、現象をうまく説明できない理論は修正、もしくは他の理論と置き換えられることになる。例えば、ニュートンの万有引力の法則は、2つの物体の持つ特性のうち、色でも大きさでも形でもなく、「質量」という同一性を用いて定式化され、アインシュタインの相対性理論の登場まで、天体の運行を説明し予測する理論として、物理学の基礎となっていた。

構造主義科学論によれば、全ての科学的な研究には現象解明の基本となる構造、すなわち仮説、モデル、システムなど、要因間の関係についての仮定がある。逆にそのような構造のない研究は、科学研究ではない。これは、科学研究は演繹的であらねばならないとの Popper の主張に合致するものであるが、実際には帰納的研究と演繹的研究の螺旋的な研究活動により、特定の現象についての理解が進んでいくものと考えられる。

Ⅲ. 統計学の特徴と科学性

研究の結果が「エビデンス」となるためには、データ解析に統計学を用いなければならない。それでは、なぜ統計学が必要なのだろうか。まず、統計学ではあらゆる現象を数量化してデータとする。そして、統計学によりその現象の偶然による変動を評価できるからである。以下に、これらの点について説明しよう。

1. 統計学的なものの捉え方

ある現象を数量化することで、研究対象を極めて客観的に把握できる。各種臨床検査値はもちろんのこと、患者の臨床的症状や精神状態など、主観的で質的なことでも“量”として把握する。このような現象の数量化は、他の研究者による再調査や、これまでの研究との比較を極めて容易にする。研究結果の「再現性」や「比較可能性」は、科学的な研究には不可欠である。さらに、一般性（普遍性）の欠如した主観的な自己満足的な研究にならないためには、統計学的視点からの解析が必要とされる。

科学的研究は客観的であるといわれるように、あらゆる現象は数値としてデータ化される。例えばリンゴをどのように扱うかは、いろいろと考えることができる。色、匂い、味といった感覚的な特性でリンゴを記述できる。一方では、直径や周囲を測ったり、重さを量ったりもできる。誰がやってもほぼ同じ結果が得られるのは、後者の測定方法だろう。数値で表された特性は、多くの人たちの共通理解を得ることができ、このため客観的であると考えられている。

これらのことをまとめると、統計学は特定の現象を、(1) 具体的に現実的に把握する、(2) 現象を操作的に扱うために「数値」で表現する、という方法を用いる。

このように量的なデータを収集して、統計学的な解析を行うのだが、特定の現象について、(3) 個々の対象の測定値は変動する、しかし、(4) 集団の中では一定の傾向がある、という考えをする。統計学は個々の変動を考慮しながら、この「ある傾向性」を見つけるための方法であるといつてよい。例えば、平均値、割合、リスク、相関係数などをデータをもとに計算するのは、この傾向性を求めるためである。

2. 推測統計学と記述統計学

人間集団のように個体差の大きな対象を扱う場合、

比較グループ間で対象の選び方による偏りがないように、まず研究デザインを正しく設定する必要がある。正しい研究デザインのもとで行われた調査であれば、ある仮定のもとで2グループの差が偶然により観察される確率を求めることで比較することができる。そのような方法は、「推測統計学 inferential statistics」と呼ばれている。この方法がエビデンスのための基本的方法論となっている。

一方では、例えば2型糖尿病患者の性別年齢分布はどのようになっているのか、その平均年齢はどのくらいなのか、都市部と農村部での発症率はどのくらいなのか、といった問いに答えるための情報を提供する必要がある。このような、ある現象に関する客観的な記述を行うことを目的とする「記述統計学 descriptive statistics」もあり、現実の研究では両方の方法が併用されている。

3. データの標準化と検定のための理論的背景

統計学がこれほど普及し、人間集団を対象とする研究で用いられているのは、推測統計学の威力によるといえよう。実際のデータ解析において用いられる統計学的方法の基礎となる考え方について説明する。

①データの標準化

身長、体重、血圧などのデータを考えてみればわかるように、それぞれの変数の単位や平均的な数値の大きさは異なるのが当たり前である。当然、各変数の平均値や標準偏差の単位も異なる。身長ならばcm、体重ならばkgである。このように変数の単位が異なっていると、全ての変数に使用できる理論を構築するのは難しい。そこで、すべての変数のデータを平均0、分散1（標準偏差1）となるように、データを変換する。このような操作を「データの標準化 standardization」と呼んでいる。

推測統計学の基礎として、標準化した変数の分布は「標準正規分布 Standard Normal Distribution」に従うという仮説がある。正規分布は、誤差の分布としてGaussが発見したといわれており、「ガウス分布」とも呼ばれている。標準正規分布については極めてよく研究されており、例えば、平均0の±1の区間内にはそれぞれ34.1%、2以上もしくは-2以下には2.3%の割合になっている、などがわかっている。検定や推定でよく使われるマジックナンバーである1.96以上は2.5%の確率になっている。

②中心極限定理

上記の標準正規分布が検定や推定で用いられるのは、多くのデータについて「中心極限定理」が成り立つからである。

通常、われわれは何度も同じ調査を繰り返して行うということはないのだが、母集団から標本を無作為抽出した場合、構成人員が有限の母集団であっても、実際の標本の選び方はほとんど無限にあるといってもよいだろう。それぞれの標本抽出により収集されたデータには、異なる平均値や異なる分散があり得る。そのため、標本ごとの各平均値の平均値や平均値の分布といったものが考えられる。あらゆる標本の平均値の平均値は「期待値」と呼ばれており、標本が無作為に選ばれているのなら、母集団の平均値である「母平均」に一致する。同様に、各標本の平均値の分布は、正規分布に従う。より一般的には、『独立な同一の分布に従う複数の確率変数の和の分布は、もとの確率変数に標準偏差が存在するならば、もとの分布の形状に関係なく、その個数が多くなるほど正規分布に収束する』という中心極限定理が成り立つ。これが、統計学における推定や検定の理論的な根拠である。もとの変数の分布が正規分布でなくても、平均値などのデータの和の分布は正規分布に従うという点が極めて重要である。

③分布と検定

標準正規分布に基づく検定は「 z 検定」と呼ばれることもある。標準正規分布から抽出された変数を2乗したものの合計（2乗和）の分布は、その個数を自由度とする「カイ2乗（ χ^2 ）分布」となる。そして、2つのカイ2乗分布に従う変数の比は、「F 分布」に従う変数となる。このため、F 分布は自由度を2つ持つ。一方、標準正規分布に従う変数を標本抽出した場合、その母分散が不明な場合、母分散を不偏分散で推定したとき、その平均値の分布は「標本数-1」を自由度とする「 t 分布」により近似できる。したがって、 t 分布はその自由度が大きくなると標準正規分布に限りなく近づく。推測統計学では、 z 検定、 t 検定、カイ2乗検定、F 検定などの名称が用いられることもあるが、それらは確率計算にどの分布を用いているかを示している。

特定の変数について、母集団での分布を仮定して検定を行う方法は、「パラメトリック検定」と呼ばれており、通常は正規分布を仮定することが多い。一方、

特定の分布を仮定しない方法は「ノンパラメトリック検定」とか「分布によらない方法」と呼ばれている。ノンパラメトリック検定では、順位に基づく方法がよく用いられている。

IV. Counterfactual Model と研究デザイン

エビデンスを得るための研究は、正しく研究デザインを設定し、適切な分析方法を使う必要がある。そのためには、Counterfactual Model の考え方を理解することが重要であり、科学的な因果推論の方法を理解すべきである。

A さんに対して、喫煙が肺がん発症の原因であるかを調べるにはどうすればよいだろうか。簡単ようで実はかなり大変なことである。われわれは遺伝的要因の他にさまざまな環境要因の影響を受けているので、何かの病気になっても何が原因なのかを即断する訳にはいかない。タバコを吸っていて肺がんになったからといって、喫煙が原因とは言えない。これは、梅干しを食べていたから胃がんになった、といった論理が成り立たないのと同じである。一方、A さんが喫煙者でないのに肺がんになったとすると、喫煙以外に何か原因があったことになるという推論は正しい。それでは、喫煙と肺がんに関する正しい因果推論はどのように行えばよいのだろうか。このためには、①タバコを吸う A さんを追跡調査する。②タバコを吸わない A さんを追跡調査する。①と②を比較し、A さんが肺がんになったか否かを比較し、もしも①では肺がんになり、②では肺がんにならなかった場合のみ、A さんにとって喫煙が肺がんの原因であったと結論付けることができる。ただし、①と②の A さんは喫煙習慣の有無以外はすべて同じ状況になければならない。これが、「Counterfactual Model（半事実モデル）」と呼ばれる考えである。これを集団に適用して、喫煙の肺がん発症のリスクを調べることが可能になる。すなわち、①集団 A の構成員全員が喫煙者である状況で追跡調査を行う。②集団 A の構成員全員が非喫煙者である状況で追跡調査を行う。①での肺がん発生率 p_1 と②での肺がん発生率 p_0 を求め、その比 $R = p_1/p_0$ を計算すると、これがその集団での喫煙の肺がんに対する正確なリスク比となる。

実際に、この Counterfactual Model を使った研究を行うためには、タイムマシンや多重世界観測装置といった SF でお馴染みの装置が必要であり、残念ながら

らまだ開発されてはいない。したがって、正しい因果推論を行うための次善の研究デザインが必要である。

現実世界では、ケースとコントロールを比較するという方法によって研究が行われている。Counterfactual Model の要点は、原因と考えられている要因以外は、ケースとコントロールですべて同一であるということである。したがって、研究対象としては、遺伝的に同一な一卵性双生児をケースとコントロールにすると、遺伝的要因を排除できる。また、双生児ならば家庭環境も同一と考えられるので、その点からも都合がよい。一卵性双生児がそれほど集められなければ、二卵性双生児、兄弟姉妹などを対象とした研究も遺伝や家庭環境の同一性がかなり保証できるのでよい方法である。しかし、そのような対象者を多数集めて研究を行うのは、通常はそれほど簡単ではない。

一般には、研究対象者に適格と考えられる人に、研究内容の説明を行い同意が得られた人のみを対象に行わざるを得ないのが現状である。したがって、悪影響があるような要因（喫煙、放射線など）を摂取させたり、曝露させたりするわけにはいかない。例えば、効果があることがほぼ確実な糖尿病の新薬と現在使われている経口剤などを比較するなど、対象者の不利益にならない倫理的な問題のない研究デザイン以外は実施できない。

研究に参加することになった対象者を、ケースかコントロールのどちらにするかを、コンピュータなどを用いて無作為にグループ分けするのが普通である。このような方法は、「無作為割付 random allocation」と呼ばれる方法である。

無作為割付によって行われた研究で、ケースとコントロールの比較には以下のような特徴がある。

- ① 割り当てたグループでの各種特性は、平均的には（多数の調査結果をまとめた場合には）同一になると期待できる。
- ② 1回の割り当てでは、必ずしも同一特性をもった集団に分割されとは限らない。
- ③ 対象者数が多いほど、対象集団の各特性が母集団特性に近くなるものと期待できる。
- ④ 各種の推測統計学の手法が適用可能である。

上記のように無作為割付による研究であっても、1回だけの調査では必ずしもケースとコントロールの基本属性が完全に一致することは、ほとんど期待できない。したがって、結果に影響する可能性のある性や年

齢などの交絡を除くために、層別解析などが必要な場合も多い。また、類似の目的で同一の研究デザインで実施された研究が多数ある場合には、メタアナリシスを行うことで、より正確なリスク評価を行うことができる。

V. 検定と標本数の関係

実際の研究では、標本数を適切に定める必要がある。標本数の決定方法に関しては、統計学的推定と検定についての理解が必要である。

統計学的推定は、標本データから母集団の特性値（母平均、母割合（母比率）、母相関など）を文字通り「推定 estimation」の方法である。これには、一つの数値で推測する「点推定 point estimation」とある信頼性を付加した区間である「信頼区間 confidence interval」を用いた「区間推定 interval estimation」による方法があり、理解は容易である。

一方、統計学的検定では研究上の証明したい「作業仮説 working hypothesis」ではなく、「帰無仮説 null hypothesis」と呼ばれる特殊な仮説を設定し、この帰無仮説を否定することで、作業仮説が成り立つことを統計的に推論する、というプロセスをとる。例えば、「新しい糖尿病の経口剤が現在の標準的治療薬よりも効果がある」ということを調べるために、新薬群と標準薬群の2群を無作為割付で設定し、実際に調査を行ったとしよう。作業仮説通りに新薬群での有効者の割合が85%と標準薬群の75%に比べて、10%ほど有効率が高かったとする。この結果から、新薬の方が標準薬よりよく効くことが実証されたとは、すぐに言うわけにはいかない。2群しかなければ、どちらかの数値がたまたま良かったということは大いにあり得ることなので、それが偶然の結果ではないことを示す必要がある。

まず、帰無仮説として「新薬と標準薬の効果は等しい」を設定する。この帰無仮説のもとに実際に観察されるような有効率の差が生じる可能性、すなわち確率を求める。求めた確率が大きければ、偶然にそういった差が生じることもあり得ることになり、有効性に違いがあるとは言えないことになる。しかし、求めた確率がある基準（通常は5%）より小さければ、そのような差が生じるようなことは偶然ではないことになる。偶然とは考えにくい事象が生じたのであるから、確率計算のもととなった帰無仮説はおかしいので否定する。すなわち、「新薬と標準薬の効果は等しくない」、

「統計学的有意差がある」といった表現で検定結果を示すのが普通である。この確率計算のために、標準正規分布やt分布、カイ2乗分布、F分布が用いられ、そのための検定統計量が計算される。統計学が苦手な研究者は、この統計量の計算式の煩雑さとその意味が理解できないので、嫌になるのではないだろうか。現在では、PC用の統計ソフトを使えば面倒な計算のみならず、検定のための確率計算もしてくれるので、われわれは出力された「有意確率 (p 値)」を見て、その有意性を判断すればよいので、計算式に拘る必要はなくなった。

ただし、統計学的検定結果は標本数に影響されるので、研究計画を立てる場合には、有意な結果を得るためには、想定される結果に見合った標本数を設定する必要がある。

例えば、食塩摂取量と収縮期血圧の関係について、成人男性20人を対象に調査したところ、両者の単相関係数は0.25であったとする。この場合、「無相関の検定」をするとt値は1.095となりp値は0.288となる。したがって、5%よりp値が大きいので有意な相関とは言えない。しかし、これが100人の調査結果だった場合には、t値は2.556となり、p値は0.012となるので、「有意な相関が認められた」と結論付けられる。

帰無仮説の形式は、「2群の母平均値は等しい」、「2群の母割合は等しい」、「母相関は0」などのように、

「 $A = B$ 」や「 $H = 0$ 」のような形式をしている。したがって、ほんの僅かでも違いがあるのならば、また0でないのならば、数千人、数万人を対象にすれば、そのような差を統計的有意差として認めることが可能である。多くの標本を対象とすればそれだけ多くの予算が必要となるので、必要最小限の標本で調査を行うのが实际的である。

必要な標本数の計算や、エビデンスのための統計的方法を正しく理解するためには、実践的に研究を行いながら、必要なことを覚えていくのが近道だろう。

文 献

- 1) 中山康雄. 科学哲学入門—知の形而上学. 勁草書房, 2008.
- 2) Kuhn TS. The Structure of Scientific Revolutions. the University of Chicago, 1962, 1970 / 中山 茂訳. トーマス・クーン 科学革命の構造. みすず書房, 1971.
- 3) 池田清彦. 構造主義科学論の冒険. 1990. 毎日新聞社; 講談社学術文庫, 1998.
- 4) 高木廣文, 林 邦彦. エビデンスのための看護研究の読み方・進め方. 中山書店, 2006.
- 5) 高木廣文. ナースのための統計学, 第2版. 医学書院, 2009.