EBM とサンプルサイズ

高木廣文

新潟大学医学部保健学科　教授

１．はじめに

EBM の枠組みの中では，無作為化臨床試験(Randomized Clinical Trial, RCT) は，エビデンスとして高い地位を与えられている。しかし，RCT を折角実施して，新薬の有効性を確認しようとしても，その結果が統計学的に有意でなければ，エビデンスにはならないことになる。統計学的仮説検定では，標本数の問題を避けて通るわけには行かない。

実際の様々な臨床試験や研究を実施する場合，「この研究にはどのくらいの標本数が必要なのか」という疑問，質問が必ずでてくる。このような標本数に関する質問の背景には，統計学的仮説検定は標本数によって，有意になったりならなかったりするという事実があるからである。　例えば，心筋梗塞の予防薬として新治療薬アタックＡと従来の治療薬アタック１の RCT を行う場合，新薬の効果が認められなければ，開発した企業にとっては死活問題になるだろう。

よく知られているように，比較研究において，統計学的に有意差が認められなかったからといって，それが２つの治療法の効果に完全に差がないことを示すものではないのである。その研究で用いた標本数の２倍の標本数で検定すれば有意になるかもしれないし，それでもだめなら，さらに２倍の標本数で研究すれば，統計学的に有意な結果を得られるかもしれないのである。

統計学的仮説検定の結果は，有意水準と検出力の影響を大きく受けており，標本数の不足や偶然の変動により，期待した結果が得られないこともある。この点から，適切で必要最小限の標本を確保するように，研究計画を立てる必要がある。

一方では，実際のすべての研究が，必ずしも統計学的仮説検証的な研究ばかりではない。研究の多くは，実態調査的であったり，仮説探索的なものである。そのような研究でのサンプルサイズは，どのように決めればよいのだろうか。

ここでは，検定と標本数の関係をまず簡単に解説し，その後，研究に必要なサンプルサイズについて,いくつかの場合に分けて説明する。

２．統計学的仮説検定と標本数の関係について

まず，統計学的仮説検定が標本数の影響を，どのように受けるのかを理解するために，簡単な例を示そう。

心筋梗塞の予防薬として，新薬アタックＡと偽薬（プラセボ）の効果を比較するものとしよう。心筋梗塞のリスクの高い高齢者 40 例を，各 20 例ずつに無作為に分けたものとしよう。観察期間中にアタックＡ群では，20 例中５例（25%）が心筋梗塞を発症し，プラセボ群では 20 例中 10 例（50%）が発症したものとしよう。簡単な計算から，リスク比=(5/20)／(10/20)＝0.5，となる。したがって，アッタクＡにより心筋梗塞発症のリスクは，プラセボに比べ，半分に減らすことができるものと推定できるだろう。

しかし，偶然の変動によってこのような結果が得られたのではないことを示すためには，統計学的仮説検定を行う必要がある。

検定の帰無仮説は「母集団での心筋梗塞の発症は，新薬アッタクＡとプラセボの服用による差はない」というものである。これは統計学的には，帰無仮説「母集団でのリスク比（母リスク比）＝１」を検定することである。計算は，２群の母比率（割合）の差の検定を使うか，四分表のカイ２乗検定を用いればよい。イェーツの修正カイ２乗値 χ_Y^2 を求めると（計算方法略，著者のホームページ http://www.clg.niigata-u.ac.jp/~takagi/，などで簡単に計算できる），$\chi_Y^2=1.707$，となる。この値は，自由度１のカイ２乗分布に従うので，有意確率ｐ（ｐ値）を求めると，ｐ＝0.191 となる。ｐ値は５％よりも大きいので，この結果は有意ではない。すなわち，帰無仮説を棄却することができないので，アタックＡによる心筋梗塞の発症を予防する効果について,統計学的にはあるとは言えないことになる。ただし，よくある解釈上の誤りとして，帰無仮説を棄却できないことを，帰無仮説が正しいものと考えて，「アタックＡには心筋梗塞発症の予防効果がないことが証明された」とするものがある。しかし，この結論完全に間違いなので注意が必要である。

さて，各 20 例ずつの RCT では，統計学的に有意差を認めることができなかったが，それでは各 40 例ずつの RCT を行い，同様に，アタックＡ群では，40 例中 10 例（25%）が心筋梗塞を発症し，プラセボ群では 40 例中 20 例（50%）が発症したものとしよう。リスク比=(10/40)／(20/40)＝0.5，と前例と同様である。

この場合，検定のためにイェーツの修正カイ２乗値 χ_Y^2 を求めると，$\chi_Y^2=4.320$，となり，ｐ＝0.038，となる。すなわち，ｐ値は５％よりも小さく，前例とは異なり帰無仮説は棄却される。したがって，帰無仮説を否定し，「アッタクＡは心筋梗塞発症の予防効果がある」という結論を得ることができる。

これらの例のように，比較するグループ（母集団）間での差や変数間の関係が完全にない場合（差が０，相関係数が０など）以外は，標

本数さえ大きくすることができれば，統計学的仮説検定では必ず有意な結果を得ることができるのである。

一般に，全く差がないとか，相関が完全に０であるような健康事象は希なので，検定結果が有意でないというのは，単に標本数が少ないためであるということになる。ただし，統計学的な有意性が必ずしも，現実的に意味のあることを示す訳ではない点に注意が必要である。すなわち，無闇に標本数を増やそうと努力するのではなく，その研究に適切な標本数を収集するように，研究計画を作成する必要がある。

３．研究目的とサンプルサイズの関係について

それぞれの研究に必要な標本数は，その研究によって検証したい仮説に依存している。すなわち，新薬の効果を調べる場合のように，研究目的が特定の仮説検証にある場合には，その仮説に応じた適切な標本数を求めなければならない。一方，コホート研究で生活習慣の何が心筋梗塞発症に影響するのかを調べるような場合のように，研究上の仮説や理論構築のための研究の場合，必ずしも単独の統計学的仮説検定のみに興味があるわけではない。この点から，研究に必要なサンプルサイズは，研究目的が，(1)単独の仮説検証，(2)仮説探索／理論構築，かによって分けて考える必要がある。

(1)単独の仮説検証的研究で必要なサンプルサイズ

統計学的仮説検定には，２種類の誤りがある。すなわち，「第１種の誤りα」と「第２種の誤りβ」である。通常，第１種の誤りは「有意水準」と呼ばれているものであり，通常の検定で使用される検定のための偶然性の判断基準である。一方，第２種の誤りは，１－β，を考えることで，統計学的仮説検定の「検出力 power」として知られている。

検出力は，「帰無仮説が正しくない場合に，帰無仮説を棄却する確率の大きさ」を示すものである。したがって，仮説検証的研究では，帰無仮説が誤っている場合に，必ず帰無仮説を棄却できるようにするためには,検出力を十分大きくするように標本数を決めねばならない。しかし，一般の検定においては，対立仮説を明確に定義できないため，第２種の誤りβの大きさを定めるのが難しい。そのため，検出力も求めにくい場合が多い。

コーヘン Cohen(1969) によれば，βはαの約４倍程度が望ましいものとされている。それに従うと，有意水準が５％であれば，第２種の誤りは 20%（検出力は 80%），有意水準が１％であれば，第２種の誤りは４％（検出力 96%）と設定することとなる。

実際のサンプルサイズを求める式は，やや複雑であり，ここでは簡単な例のみにとどめるが，どのような場合であれ，

(a)有意水準αの大きさ
(b)検出力１－βの大きさ
(c)対立仮説での効果の大きさなど
を研究計画時に決める必要がある。

上記の(a)，(b)，(c)うち，有意水準や検出力は，５％と 80％などと，研究者によって任意に定めることができるが，問題は(c)である。

通常の統計学的仮説検定では，帰無仮説は「母相関係数＝０」，「２群の母平均値は等しい」，「喫煙グループの心筋梗塞発症の母リスク比は１」などのように，相関の大きさやグループ間の差の大きさなどを定めずに，全く存在しないものとして検定する。ところが，検出力の計算のためには，対立仮説として「母相関係数＝0.2」，「２群の母平均値の差＝10」，「喫煙グループの心筋梗塞発症の母リスク比は３」などのように，実際的な数値を推定しなければならない。

さらに，実際の計算によっては，検定に使用する変数の母分散の情報が必要な場合もある。これらの数値がわからない場合には，事前に研究に必要な標本数を求めることはできないのである。

２グループの比較研究に必要な標本数を求めるための近似的な計算式として，以下のような簡単だが優れた式が知られている（竹内，1975）。

２グループＡ，Ｂの母平均値の差の検定のためには，まず２つの母平均値の差をΔとする。２グループの母分散を等しいものと仮定し，それをσ^2とする。このとき，検定の有意水準を両側２α（片側α），検出力を１－βとすると，各グループの標本数ｎは，

$$(1) \qquad n = 2(z_\alpha + z_\beta)^2 \times \frac{\sigma^2}{\Delta^2} + \frac{z_\alpha^2}{4}$$

によって近似的に求めることができる。ここで，z_αとz_βは，それぞれ標準正規分布の上側α点と上側β点の値である。

なお，上記の(1)式は，２群の特定事象の割合の差（母比率の差）の検定にもそのまま応用できる。すなわち，２グループのある特性の割合（例えば，心筋梗塞発症の割合）をそれぞれp_1とp_2はとした場合，２グループの割合の差Δ（＝$p_1 - p_2$）を定めることができる。全体での母割合をｐとすると，分散$\sigma^2 = p(1-p)$，$p = (p_1 + p_2)/2$，と置くことで，(1)式を用いることができる。

計算に必要とされる，λ＝Δ／σ，は検定すべき差が標準偏差の何倍であるかを示す値であり，「エフェクト・サイズ effect size」と呼ばれることがある。エフェクト・サイズの値が大きいほど標本数は少なくてよく，逆に小さな場合には多くの標本数が必要になる。

通常の検定では，有意水準は片側 2.5%（両側５%），βをその４倍の片側 10%（両側 20%）とする。このとき，計算に必要な標準正規分布での各数値は，$z_{0.025}=1.96$，$z_{0.1}=1.282$，となる。$z_{0.025}$はほぼ２なので，(1)に各数値を入れて計算すると，

(2)　　$n = 21.02\times\sigma^2/\Delta^2+1$

と必要な標本数を近似的に求めることができる。

例えば，高血圧グループと健常者グループでの食塩摂取量が，それぞれ平均値 17ｇ／日，14ｇ／日であるとすると，この３ｇ／日の差が偶然でないことを検定するのに必要な標本数を求めるためには，まず標準偏差の推定値が必要になる。かりに，食塩摂取量の標準偏差を５ｇ／日と推定したとすると,各グループで必要な標本数ｎは(2)式を使うと，

$n=21.02\times(5/3)^2+1=59.4$

となり，各群 60 例の標本が必要であることがわかる。

前述の心筋梗塞発症の予防の例では，アタックＡ投与群での発症割合は 50%（$p_1=0.5$），プラセボ群では 25%（$p_2=0.25$）であった。この 25%の発症割合の差が，統計学的仮説検定で有意となるために必要とされる標本数を求めてみよう。

有意水準は両側５%，検出力 80%の場合，上記の(2)式を用いれば，簡単に必要とされる各群の標本数が計算できる。まず,母割合$p=(0.5+0.25)/2=0.375$，として，エフェクト・サイズλは，

$\lambda^2=(p_1-p_2)^2/[p(1-p)]=(0.5-0.25)^2/[0.375\times0.625]=0.2667$

である。したがって，

$n=21.02/\lambda^2+1=21.02/0.2667+1=79.8$

となるので，各群 80 例の標本数が必要ということになる。

前述の例では，各 40 例で検定結果は有意となっているのだが，この計算式からは２倍の標本数が必要と言うことに結果になった。おかしな気もするのだが，この理由は，通常の検定では，検出力は 50%に設定されているということによる。すなわち，$z_{0.5}=0$，なので，この

値を(1)式に代入して，必要な標本数を計算すると，

$$n = 2 \times (1.96 + 0)^2 / 0.2667 + 1 = 29.8$$

となり，各群 30 例でよいことになる。各群 30 例の場合，実際に観察される人数に端数はないので，かりにアッタクＡ群とプラセボ群の心筋梗塞が７人と 15 人観察された場合,仮説検定のための補正なしのカイ２乗値は 4.593 となり，ｐ＝0.0321，で５％で有意となる。しかし，アッタクＡ群で心筋梗塞の発症者が１人増えて８人観察された場合には，カイ２乗値は 3.455 となり，ｐ＝0.063 で，有意とはならない。ただし，実際の検定で使用するイェーツの修正カイ２乗値では，どちらの場合も有意にはならないのだが。

現実の調査研究では，偶然性の変動の影響があるために，仮定した割合通りに心筋梗塞が各グループで発症するわけではない。このため，偶然性を考慮して，検出力を高く設定しないと，思ったような結果を得られないことがあり得る。したがって，すでに観察されたデータによる検定のように，検出力を 50%と設定している単純な検定式が必要とする標本数よりも，より多くの標本数が必要となる点に注意しなければならない。

(2)探索的／理論構築のための研究で必要とされる標本数

サンプルサイズの問題は，統計学的仮説検定の要求によって生じる問題であることは，ここまでの説明で理解できたものと思う。

それでは，検証すべき数量的な仮説がない研究では，どのように研究上の標本数を決めればよいのだろうか。

残念ながら，数値目標のない研究のためには，一般的な標本数を適切に設定する便利な公式は存在しない。それでは，全く適当に標本数を決めてよいかというとそうでもなく，ある程度は考慮すべき点はいくつかある。以下に基本的な考え方を説明しよう。

まず第一に，

(a)標本数は多い程よい，

ということである。すなわち，30 例の研究よりも 300 例の研究の方がよいし，それよりも 3000 例の研究の方がもっとよいということである。ほとんどの統計学的推定の計算式に基づけば，推定値の信頼性は，ほぼ標本数の平方根に比例している。この意味は，母平均値や母割合の推定値の信頼区間を 1/10 の幅にするためには,10 倍ではなく約 100 倍の標本が必要であることを示している。

現実には，それほど多くの標本調査は不可能なことも多いが，質問紙調査などでは，

(b)調査項目数の２倍程度の標本数，

は最小限欲しいところである。標本数は多いほどよいのだが，逆にあまりに少なすぎると，実施した調査によって得られたデータが，一般的な回答パターンを示しているという保証がなくなってしまう。

ここではあまり説明していないが，一度に多数の変数間の相関関係を考慮して解析するために，分析に多変量解析を使用する場合には，理論的に解が得られなくなってしまうことがある。それを避けるためには，(c)多変量解析の手法を用いる場合，少なくとも分析に用いる変数の個数より多い標本数を確保する必要がある。できれば，変数の個数の２倍以上の標本数が最低でも必要だろう。

このように，唯一の数量的な仮説検証を目的とするのではない研究では，少なくとも調査項目数の２倍程度の標本数を最低でも集めるようにすべきだろう。ただし，これは目安であり，あくまでも標本数は多いに越したことはないのである。

４．最後に

今回は，研究で必要とされるサンプルサイズについて，仮説検証のための標本数を中心として，簡単に説明した。紙面の都合や，複雑な計算式を避けたために，多くの検定方法に関しての詳細な説明はしていない。より詳しい説明に興味がある読者は，成書や下記の参考文献などを参照していただきたい。

参考文献

1)Cohen, J: Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York, 1969

2)Fleiss, JL 著；佐久間昭訳：計数データの統計学，東京大学出版会（東京），1975

3)林邦彦，高木廣文：EBN のための研究計画－統計学的推論：推定と検定，EBNursing，3(4)，82-87，2003

4)高木廣文，林邦彦：コントロールや無作為化はなぜ必要なのか，EBNursing，2(4)，102-107，2002

5)竹内　啓：確率分布と統計解析，107，日本規格協会（東京），1975

6)椿広計，藤田利治，佐藤俊哉編：これからの臨床試験，171-172，朝倉書店（東京），1999

7)上坂浩之：臨床試験における被験者数設計の視点と方法，計量生物学，24(1)，17-41，2003

8)柳井晴夫，高木廣文：臨床試験における標本数の定め方，医学のあゆみ，102(13)，
881-886，1977

9)柳井晴夫，高木廣文　編著：新版看護学全書　統計学，88-105，メディカルフレンド　社（東京），1995