研究員の眼

作られた有意性

前提に、手を加えられてはいないか?

保険研究部 主任研究員　篠原　拓也
(03)3512-1823　tshino@nli-research.co.jp

人々の健康状態や、病気について、統計学を用いて解明する学問分野として、疫学がある。疫学では、仮説検定が行われることが多い。その流れは、次の(1)～(5)のような手順となる。

手順(1) 示したいことと反対のことを、帰無仮説として設定する
手順(2) 帰無仮説を否定する(「棄却する」という) 確率の水準(「有意水準」という)を設定する
手順(3) 臨床試験や健康調査などを行って、データを収集する
手順(4) 帰無仮説を前提とした場合に、収集したデータが得られる確率(「P 値」という)を計算する
手順(5) 計算されたP 値と、有意水準を比較して、帰無仮説を評価する

手順(5)の評価で、P 値が有意水準未満の場合、帰無仮説が棄却される。即ち、統計的な論証が得られたことになる。この場合、示したいことが示されたとして、自説を展開することができる。

一方、P 値が有意水準以上の場合、帰無仮説は棄却されない。これは、統計学的には、帰無仮説が否定されなかったことを意味する。しかし、かといって、帰無仮説が肯定された訳でもない。つまり、帰無仮説は、否定も肯定もされない、中途半端な状態になったことを意味する。

通常、仮説検定を行う研究者は、自分の論説を展開する上で、ある命題を統計学的に示したいという強い意図を持っている。例えば、研究開発対象の候補薬には効果があるとか、喫煙とがん発症の間には関連性がある、などといったことだ。この状況を誇張して、やや口語的に言えば、研究者は、統計的検定を通じて、帰無仮説を棄却したくて、うずうずしている。検定の結果、もし、帰無仮説が棄却できないとなると、自説が論証できないため、困ってしまうことになるからだ。

帰無仮説が棄却できなかった場合に、そのデータを用いた論証をあきらめられればよいが、これはそう簡単な話ではない。通常、臨床試験の実施には、多額のコストや、多くの時間が費やされている。「臨床試験で、有意義な結果は得られませんでした」では、済まされないかもしれない。

そこで、何とかならないかと、研究者は考えてしまいがちだ。まず、すぐに思いつくのが、手順(2)で、有意水準を高めに設定し直すことだ。例えば、有意水準を 1%と設定して検定したところ、P 値が 3%となったため、有意水準を 5%に設定し直す。これは、スポーツで言えば、プレーの後に、ルールを変えて、正反対のジャッジをするようなもので、本末転倒と言える。このようなことを防止するために、臨床試験では、あらかじめ計画書を作成し、その中で有意水準を明記することとされている。

次に、手順(3)で、収集するデータの数を増やすことが考えられる。次の例を見てみよう。候補薬を投与された人の 70%、候補薬と形状や味がそっくりだが何も薬効のない対照薬(プラセボ)を投与された人の 63%が回復している。回復割合には、7%の差がある。しかし、P 値を計算してみると、32.4%もあり(*)、有意水準 1%で、帰無仮説は棄却されない。((*) P 値は、ピアソンのカイ二乗検定という検定の値[以下、同じ])

例 1. 候補薬を 80 人、プラセボを 100 人に投与　　⇒ P 値 = 32.4%

	症状回復	症状未回復	計
候補薬を投与	56 人　(70%)	24 人　(30%)	80 人　(100%)
プラセボを投与	63 人　(63%)	37 人　(37%)	100 人　(100%)

( )内は、横占率［以下、同じ］

そこで、試験データの数を、10 倍に増やしてみることにした。症状の回復、未回復の人数も、それぞれ 10 倍となったとする。即ち、候補薬やプラセボを投与された人の回復割合は、70%、63%のままだったとしよう。この例で、P 値を計算してみると、0.2%となり、有意水準 1%で、帰無仮説は棄却される。つまり、データを増やすことで、有意でなかったものが、有意に変わったのである。

例 2. 例 1 の 10 倍の人に投与（回復割合は変わらず）　　⇒ P 値 = 0.2%

	症状回復	症状未回復	計
候補薬を投与	560 人　(70%)	240 人　(30%)	800 人　(100%)
プラセボを投与	630 人　(63%)	370 人　(37%)	1,000 人　(100%)

ただし、いつもこの例のように、思惑通りに、事が進むとは限らない。例えば、次の例のように、候補薬の回復割合が低下し、一方、プラセボの回復割合が上昇する、といった事態が生じるかもしれない。この例で、P 値を計算してみると、2.6%となり、有意水準 1%で、帰無仮説は棄却されない。

例 3. 例 2 で回復割合が変化（候補薬の回復割合は低下、プラセボの回復割合は上昇）　　⇒ P 値 = 2.6%

	症状回復	症状未回復	計
候補薬を投与	552 人　(69%)	248 人　(31%)	800 人　(100%)
プラセボを投与	640 人　(64%)	360 人　(36%)	1,000 人　(100%)

このように、データを増やしても、P 値が低下しないことがある。そこで、P 値が 1%未満となるまで、更に、データを増やすことになる。例えば、データ量を、例 3 の 2 倍にして、回復割合が変わらなければ、P 値は 0.2%となって、有意水準 1%で、帰無仮説は棄却される。しかし、こうした行為は、研究者の恣意によるもので、公正さを欠く。このようなことを防ぐために、臨床試験の計画書では、あらかじめデータ数も明記することとされている。

疫学に限らず、一般に、統計学の仮説検定の仕組みには、研究者が恣意的に結果を操作できてしまう可能性がある。仮説検定の結果を把握する際には、検定方法の説明書類(臨床試験の計画書等)を通じて、恣意的な設定がないかどうかを、よく確認することが必要と考えられるが、いかがだろうか。

