治験における統計:主要概念

はじめに

統計的手法は、患者の治療に対する反応における変動要因を形式に基づいて説明するものとなります。統計の使用によって、臨床研究者は収集された情報から合理的で正確な推論を行い、不確実性があっても正しい判断を下すことができます。統計は臨床研究においてエラーやバイアスを防止するカギとなります。ここでは統計の主要な概念をいくつか取り上げ、治験での応用方法について説明します。

仮説検定

仮説とは 1 つまたは複数の仮定のことで、a) 科学的調査の指針となる見解を示しながら、仮定に基づいて何らかの主張を行うか、b) 確立された事実に照らして、可能性が非常に高いものを裏付けます。

ここでは、何らかの主張を行う仮説について取り上げます。たとえば、「ある病気のための新しい治療法は既存の標準の治療法よりも優れている」といったものです。新しい治療法を「B」、標準の治療法を「A」とすると、「B」は「A」よりも優れているという仮説を立てることになります。

このような仮説の証明に、科学者が着手するものと考えがちですが、実際は異なります。この目標には、間接的なアプローチを使用します。仮説 B の証明を試みるのではなく、標準の治療法と新しい治療法には相違点がないという事実 A が真であるとする科学的方法を仮定します。これは「帰無仮説」として知られています。科学者はそれから A が誤りであることを証明することを試みます。これは、帰無仮説が偽であるという証明とも考えられています。仮説 A が偽であると証明できれば、標準の治療法は新しい治療法よりも優れていないということになり、B が真ということになります。つまり、新しい治療法は標準の治療法よりも優れているということになります。

この手法が使われる理由とは？

理由について単純に答えることはできませんが、これは近代科学で生まれ広く使われるようになった手法であり、法的類推の活用に役立つ場合があります。帰無仮説は現在の状況や知識に基づいて立てられ、十分な証拠が得られない限りは、それを信頼する必要があるということです (法廷に例えると「被告人は無罪」ということになります)。ただし、「対立仮説」を証明する (つまり「帰無仮説」を反証する) ことで、「被告人は有罪」ということになります。

この仕組みは、アルベルト・アインシュタインが残した次の言葉を聞けば理解しやすいでしょう。

「どんなに実験を繰り返しても私が正しいことは証明できないが、私が間違っていることを証明するのはたった一度の実験で事足りる。」

これは、帰無仮説が偽、つまり誤りであることを証明する試みのほうが、対立仮説が正しいことを証明する試みよりも厳密であり、達成できる可能性が高い目標であることを示唆しています。これは科学の世界でこのアプローチが採用されている理由を適切に説明したものとはいえませんが、わかりにくい概念を理解しやすく、受け入れやすくするのに役立ちます。

I 型エラーと II 型エラー

以下の表は、I 型エラー（偽陽性）と II 型エラー（偽陰性）の相違点をまとめたものです。

	帰無仮説が真である	帰無仮説が偽である
帰無仮説を棄却	I 型エラー 「偽陽性」	正しい結果 「真陽性」
帰無仮説を棄却できない	正しい結果 「真陰性」	II 型エラー 「偽陰性」

この構造は非常に混乱しやすいため、以下のように極端な例を使って、シンプルな言葉で説明します。

- I 型エラーでは、患者が死亡するおそれもあります。たとえば、標準的なケアが新しい治療より優れないとする誤った所見を研究によって得られた場合、それを受けて新しい治療を患者に実施すると、重篤な結果を生じる可能性があります。I 型エラーを起こすと、実際には存在しない効果が誤って検出されることになります。
- II 型エラーは、潜在的に有益な研究が無駄になることを意味します。その研究は非常に有益であった可能性はありますが、続行されないため、患者に対する危害は生じません。II 型エラーを起こすと、実際には存在する効果を検出できないことになります。

患者に対する影響という意味では、明らかに I 型エラーのほうが II 型エラーよりも深刻です。

有意水準

有意水準とは、I 型エラーを起こす可能性です。これはサンプルのサイズと、テストの「統計的検出力」による影響を受けます。

統計的検出力

統計的検定の「検出力」とは、帰無仮説を正しく棄却できる可能性、つまりある効果が実際に存在する場合に、その効果を検出する検査能力のことです。検定の「検出力」は、II 型エラーを引き起こさない可能性と言い換えることもできます。

P-values

p 値（probability value: 確率値）とは、証拠の強さを 0~1 の尺度で量ったものです。p 値が小さい場合（通常は 0.05、または 5% 未満）は、帰無仮説を 反証する 強力な証拠が存在することを示し、帰無仮説を棄却する根拠となります。p 値が大きい場合（0.05 を超える）は、その逆です。

相関関係と因果関係

試験の結果を分析する場合は、相関関係と因果関係が異なるものである点に留意することが重要です。相関関係とは 2 つの変数が何らかの形で関連していることを意味しますが、一方が他方の原因になっているわけではありません (両変数の関連性は存在します)。相関関係の例としては、ホルモン補充療法 (HRT) と冠動脈性心疾患 (CHD) の関係が挙げられ、HRT を受けた女性は CHD のリスクが低くなります。ただし、これは実際の HRT のプロセスによるものではなく、HRT を受けている患者群が平均以上に優れた食事と運動の習慣を備えた、より高い社会経済的グループに属している傾向があるためです。

因果関係とは、ある要因が結果を引き起こすときに見られます。原因要素は、多くの場合、結果の部分的な原因です。相関関係と因果関係を区別するには、試験の参加者に関する情報をできるだけ多く記録することが重要です。また、治験のデザイン段階で科学的方法論を慎重に適用し、試験で考えうるバイアスを評価することも必要です。

データ操作

データ操作とは、データを不正確に報告したり、誤った結果を作成したりする選択的な行為です。データ操作の例としては、予期した結果と一致しないデータを意図的に破棄し、主張した仮説を後押しする結果の割合を押し上げる行為が挙げられます。外れ値 (隣接する結果と比較して極端に大きい、または小さい値を持つ結果) を結果から排除する場合は、そのような値が真に外れ値であり、単に予期される (あるいは望まれる) 結果と異なるだけでないことを確認することが重要です。データ操作の別の例としては、データ収集者が、1 人の患者から収集した測定値を使って、データ全体をランダムに生成するケースが挙げられます。

データ変換

データ変換とは、試験から得られた何らかのデータに数式を適用することです。多くの場合、データが明確になるように体裁を整えたり、理解しやすくしたりする目的で使用します。たとえば、自動車の燃費性能を測定する場合、測定値は「km/ℓ」という形式で表すのが一般的です。ただし、走行距離を延長する目的で追加する燃料の量を特定する場合は、「ℓ/km」として表されます。このようなケースで誤った数式を適用してデータを別の体裁に整えると、試験の結果全体に影響が及ぶことになります。

データ結合

データ結合とは、状況を理解しやすくするために、複数の試験から収集したデータを組み合わせることです。データ結合の最も一般的な形式としては、メタ分析が挙げられます。メタ分析では、公開されている複数の試験の結果をまとめて集計し、比較します。メタ分析を実行する場合は、各試験の手法が同じか同等であることを慎重に確認することが重要です。根本的に異なる変数 (交絡変数) が存在することがないように、デザイン上のいかなる相違点も考慮する必要があります。間違ったデータ結合の例としては、動物試験として異なる種のマウスを使った複数の試験のデータを集計してしまったというケースが挙げられます。

□□□□

A2-4.33.3-V1.1