原稿作成日：　2017 年 3 月 31 日

検定と P 値：統計的エビデンスとは

〈教材提供〉
AMED 支援「国際誌プロジェクト」　提供

草案
新谷歩　大阪市立大学医学研究科医療統計学講座教授
加葉田大志朗　大阪市立大学医学研究科医療統計学講座特任助教

査読
大門貴志　兵庫医科大学医療統計学教授
山中竹春　横浜市立大学医学部臨床統計学教授
市川家國　信州大学特任教授
山本紘司　大阪市立大学大学院医学研究科医療統計学講座准教授
石原拓磨　大阪市立大学大学院医学研究科医療統計学講座特任助教

目次

はじめに

医学研究においては、母集団全員を対象としてデータを収集することはできないため、それに代わって母集団を代表するような人々をランダムに抽出して集団を構成し、それを研究対象としてデータを収集します。このランダムに抽出して得られる集団を標本（Sample）と呼び、この標本を対象として収集されたデータを標本データ（Sample data）と呼びます。標本データをもとに、母集団（Population）に関してどのようなことがいえるか、を推測することを統計的推測（Statistical Inference）と呼びます。統計的推測する統計的推定（Statistical Estimation）（「正しいデータの記述の仕方：記述統計量とグラフの描き方」を参照）と統計的仮説検定（Statistical Hypothesis Testing）に分類することができます。

ここでは、後者の「仮説検定」について学んでいきます。例えば、日本で開発され、使用されているAという治療法を従来のBという治療法に比べて効果があることを示したい場合、治療を受けた全員の患者さんからデータを集めてくることは、ほとんど不可能です。そこで標本データをもとに、母集団での仮説「治療法Aが治療法Bに比べて効果がある」の成否を推測するわけですが、これを「仮説を検定する」又は「仮説検定を行う」といいます。標本データは多ければ多いほど、仮説検定に基づく判断は真実に近づきます。

仮説検定は、

① 「仮説を立てる（Establishing a hypothesis)」
② 「仮説を棄却する（Rejecting the hypothesis)」

という２段階を設けて行いますが、最終的に仮説を棄却するか否かを判断するのにはP値を用います。ここでは、仮説検定とともにこのP値についても学びます。

学習目標

本単元を通じてあなたが修得を目指すものは：

- 仮説検定の手順を説明できる
- P値の意味を説明できる
- 仮説検定の結果の解釈を説明できる

統計的なエビデンスとは？

なぜ仮説を棄却するのか

例えば「水は100度で沸騰する」という仮説を証明するやりかたには、大きく分けて二つが考えられます。

① 「水は100度で沸騰する」という仮説を支持する.

② 「水は 100 度で沸騰しない」という仮説を棄却する.

科学的エビデンスを得る際は、②の「〇〇しない」という「仮説を棄却する」、というダブルネガティブの方法を使います。これは例えば「私はリンゴが好き」、というために、「私はリンゴが嫌いでない」というようなものです。①の「支持する」ための仮説は対立仮説（Alternative Hypothesis)、②の「棄却する」ための仮説は「無に帰する仮説」という意味で帰無仮説（Null Hypothesis）と呼びます。①の対立仮説を支持して「水は 100 度で沸騰する」、「私はリンゴが好きです」とストレートに言いたいのはやまやまですが、科学的エビデンスとは②の帰無仮説を棄却して、それぞれ「水は 100 度で沸騰しないはずはありません」、「私はリンゴが嫌いではありません」といった周りくどい方法を使います。
なぜこんなまわりくどいことをするのでしょうか。

例えば「水が 100 度で沸騰する」という仮説を証明したい場合、どうしたらよいでしょうか。東京でまず水を沸騰させたら 100 度で沸いたとします、次は大阪で沸騰させたらまた 100 度で沸きました、次に北海道でも、ニューヨークでも、試したすべての場所で水は 100 度で沸きました。これでエビデンスとしては十分でしょうか？　まだまだ足りません。次はアラスカで、次はイギリスで、という具合に世界中 1 万箇所で沸かしてみました。すべて 100 度で沸きました。これで十分でしょうか？10001 か所目に富士山の上で沸かしたところ 88 度で沸きました。この最後の一つのエビデンスによって「水は 100 度で沸騰する」という仮説が棄却されます。そして次に「水はいつも 100 度で沸くのではなく、沸点は海抜に依存する」というように、当初の仮説を棄却したうえで、新たな仮説を立ててそれを証明しようとするプロセスを繰り返す。この過程を通じて科学的エビデンスが次第に構築されていきます。

この例でも明らかなように、仮説を支持するためのデータはいくらあってもきりがありませんが、仮説を棄却するには一つのデータで十分です。これを踏まえて科学的エビデンスを得るためには「帰無仮説を棄却する」というダブルネガティブな手法を用います。

それでは、例えば、臨床的な仮説「薬剤 A には薬剤 B よりも日本人成人男性の血圧を下げる効果がある」ことを証明することを考えてみましょう。薬剤 A および薬剤 B をそれぞれ 50 人の患者へ投与後、その血圧を観測した結果をもとに、薬剤 A が投与された患者群（薬剤 A 群）と薬剤 B が投与された患者群（薬剤 B 群）とで血圧の平均値（平均血圧）を比較します。この研究において主に評価するデータは、薬剤を投与した後の平均血圧であり、こうした注目する結果の値のことをアウトカムと呼びます。この研究において、次の帰無仮説を立ててみました。

帰無仮説「薬剤 A 群の平均血圧と薬剤 B 群の平均血圧の差はゼロである」

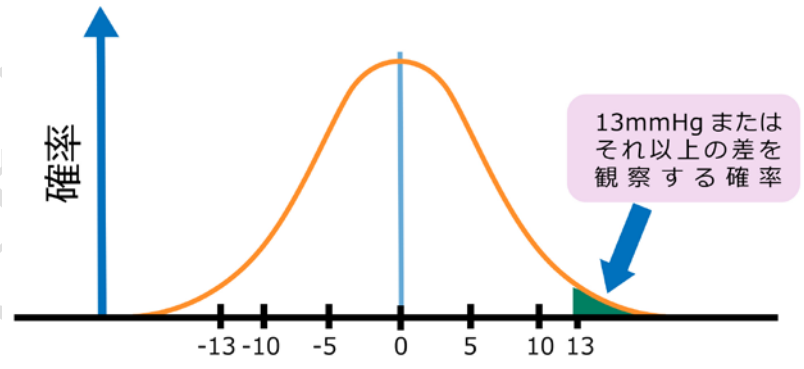
この帰無仮説はどのような基準をもって棄却できるのでしょうか？　まずこの帰無仮説が日本人の成人男性という母集団において正しいと仮定します。そのもとで実際にデータを集めてみると、薬剤 A 群では薬剤 B 群に比べて、明らかに血圧が下がる人が多かったとします。この結果を踏まえて二通りの判断が考えられます。

① 帰無仮説が正しい、すなわち、薬剤 A には薬剤 B と比べて効果がないのに、今回だけ偶然このような結果が起こったと考える
② 今回だけ偶然にこのような結果が起こったのではなく、実は、帰無仮説が正しいと仮定したことが間違っている、すなわち、薬剤 A には薬剤 B と比べて効果がある、からだと考え

る

帰無仮説を棄却するときの判断に利用するのがP値です。P値は確率（Probability）の実現値の略称で、「帰無仮説が母集団において正しい時に標本データで観測された事象、またはそれよりも更に仮説から外れた事象が起こる確率の実現値」を意味します。上述の例に準えて考えると、P値は「薬剤Aには薬剤Bに比べて効果がない時に、偶然に観測された差以上の差が起こる確率の実現値」ということになります。すなわち、P値が小さいということは、帰無仮説が正しいと仮定したことが間違いであることを示唆します。では、このP値がどの程度小さければ帰無仮説を棄却すると判断してよいのでしょうか。当然のことながら、この判断のための基準値は、場当たり的に決められてしまうと、帰無仮説の棄却を都合よく行えてしまいます。それ故、この判断のための基準値は、事前に宣言しておくことが必要になります。この判断のための基準値は有意水準（significance level）と呼ばれ、慣例的には5%が用いられます。したがって、P値が有意水準よりも小さければ、②のように考えることになります。このとき、「(有意水準 5%で）統計的に有意な差が示された」と記述されます。一方で、P値は、「薬剤Aには薬剤Bと比べて効果がない時に、標本データをもとに効果があると「間違って判断してしまう確率の実現値」ともいえます。この間違って判断してしまう確率の実現値は小さい方が望ましいことはいうまでもありません。先ほどの慣例の5%を用いて、P値がこの5%より小さいとき「帰無仮説を棄却する」と判断する、とした場合、帰無仮説を棄却すると誤って判断してしまう確率を最大でも5%に抑えることができることになります。「帰無仮説を棄却する」と誤って判断してしまう確率をこのように小さい確率（ここでは5%）に抑えることができるのであれば、②のように考えてよいということになります。

それでは、上述の例に基づいてP値の計算方法を示しましょう。薬剤A群と薬剤B群の平均血圧がそれぞれ100mmHgと113mmHg（すなわち、差は13mmHg)、両群の平均血圧の標準誤差が5mmHgだったとします。



母集団から無作為に
集めたサンプルデータ
の2群の平均の差

帰無仮説のもとで、不特定多数の研究者がそれぞれにデータを収集し、薬剤A群と薬剤B群の平均血圧を比較する研究を実施したとします。それぞれの研究から平均血圧の差が得られます。そうした場合、一つ目の研究で観測された平均血圧の群間差は、薬剤群Aが薬剤B群より10mmHg高い一方、二つ目の研究では、逆に薬剤群Bの方が5mmHg高く、群間差はマイナス5mmHgであっ

たといった具合に、研究の数だけ様々な平均血圧の群間差が記録されます。このとき、その無数に存在すると想定される平均血圧の群間差の分布を考えてみましょう。「薬剤 A と薬剤 B に効果の差がない」、とする帰無仮説が正しい場合は、平均血圧の群間差は 0 の場合が最も多く、0 から離れるにしたがって頻度が減ってくると考えられます。

上の図の横軸は、母集団からランダムに抽出された標本データから計算される平均血圧の群間差を示しています。縦軸は、帰無仮説が正しい場合に、平均血圧の群間差が観測される確率（統計学では、厳密にはこれを「確率密度」と呼びます）を示しています。この分布は 0 を中心とした左右対称の形をとると仮定します。分布を表す赤い曲線下の面積は確率を表しています。例えば、横軸が 0 以上の大きな値を示すときの曲線下面積は 0.5、確率で言うと 50%を表します。これは、母集団において帰無仮説が正しい時、標本データで計算される平均血圧の群間差が 0 以上になる確率はちょうど 50%であることを意味します。同じ図の曲線下面積を青色で示した部分は、母集団において帰無仮説が正しい場合、標本データで平均血圧の群間差が偶然に 13mmHg またはそれより大きな値になる確率を示しています。ここで P 値の定義を思い出しましょう。P 値は「帰無仮説が母集団において正しい時に、母集団の一部からなる標本データにおいて観測された事象、ないしそれよりも更に仮説から外れた事象が起こる確率の実現値」ということでした。「標本データにおいて観測された事象」とは、この例では「標本データで計算された平均血圧の群間差」であり、それは 13mmHg ですから、上の図の曲線下面積を青色で示した部分は、まさしく P 値に対応しています。この部分の面積が仮に 3%としましょう。この例での P 値は、「帰無仮説が正しい場合、今回観察されるアウトカムの平均（平均血圧の差）が 13mmHg またはそれ以上の値である確率の実現値」、として 3%であるということになります。すなわち、薬剤 A と薬剤 B の効果に真に全く差がなくとも、偶然に今回のような差が観測される確率は 3%であるということになります。この 3%という確率が 5%よりも小さければ、偶然にしては小さすぎる、すなわち、偶然でない、として、帰無仮説が正しくないものとして棄却し、冒頭の臨床的な仮説「薬剤 A には薬剤 B よりも血圧を下げる効果がある」に対してエビデンスを得ることになります。

片側検定（One-sided test）か両側検定（Two-sided test）か

両側検定

上述の例では、帰無仮説を「薬剤 A 群の平均血圧と薬剤 B 群の平均血圧の差は 0 である」として、薬剤 A には薬剤 B よりも血圧を下げる効果がある場合の仮説検定のみを考えていました。これを片側仮説検定と呼びます。しかしながら、そのような場合を期待していたにもかかわらず、逆に薬剤 B を用いたときに薬剤 A を用いた時よりも血圧が下がってしまう場合も考えられます。薬剤 A 群の平均血圧を M_a、薬剤 B 群の平均血圧を M_b としましょう。薬剤 A には薬剤 B よりも血圧を下げる効果がある場合とその逆の場合も考えるときの仮説検定では、

帰無仮説は　　$M_a = M_b$
対立仮説は　　$M_a \neq M_b$

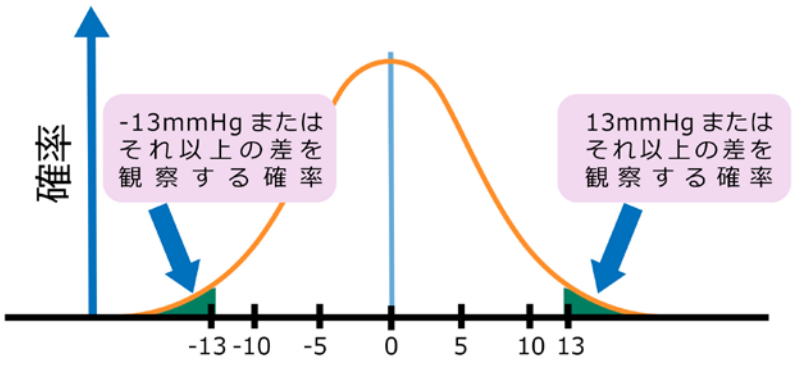
となります。これを両側仮説検定と呼びます。

両側仮説検定では、対立仮説が $M_a \neq M_b$（薬剤 A 群と薬剤 B 群とでの平均血圧が同じではない）ということは　$M_a > M_b$（薬剤 B 群の方が薬剤 A 群の平均血圧よりも低い）、という場合と、$M_a < M_b$（薬剤 A 群の方が薬剤 B 群の平均血圧よりも低い）という両方向の仮説を包含しています。つまり、

$M_a = M_b$　の　帰無仮説の対立仮説は　$M_a > M_b$　または　$M_a < M_b$

したがって、$M_a < M_b$ が支持される場合だけでなく、$M_a > M_b$ が支持される場合でも帰無仮説は棄却されるはずです。これを受けて一般的に多く用いられる P 値は、上述の両方向の仮説を考慮した両側仮説検定に対して計算を行います。この場合の P 値を両側仮説（検定）に対する P 値、「両側の P 値」などと、その意味を込めて呼ぶことがあります。例えば、上述の例で考えると、標本データで薬剤 A 群の方が薬剤 B 群よりも平均血圧が 13mmHg 低かったとしても、それと同じ確率で薬剤 A 群の方が薬剤 B 群よりも平均血圧が 13mmHg 分だけ高い可能性がある、と想定して P 値の計算を行います。すなわち、$M_a < M_b$ のみの片方向の仮説を考慮した片側仮説（検定）に対する P 値（片側の P 値）は 3%でしたが、両側の P 値は 3%の 2 倍の 6%となります。

$M_a < M_b$ のみの片方向の仮説を考慮する際の有意水準を 5%と決めていた場合、片側の P 値（3%）はそれよりも小さいので、統計的に有意な差があると言えます。しかしながら、両方向の仮説を考慮する場合の有意水準を 5%と設定したとき、両側の P 値（6%）はそれよりも大きいので、統計的に有意な差があるとは言えません。このように片側の P 値の方が両側の P 値に比べて明らかに小さくなることから、仮説の方向を明示せずに一律に有意水準を 5%とすれば統計的に有意な差が得られやすくなります。慣例としては、特別な例を除いて、通常、仮説検定では両側の P 値を用います。片側仮説検定を用いる場合は必ずその理由を明記することが必要です。しかし、実際、研究をデザインする際、標本サイズを設計するにあたり、そのサイズを小さく済ますべく片側仮説検定に基づいて決め、最終的な解析では両側仮説検定で行うことになって、統計的に有意な結果が得られなかったなどという研究が散見されます。これは、仮説の方向性（片側か両側か）に関してデザインと解析との間で整合性がとれなかったための大きな失敗です。こういった誤りを生じないように、整合性をとることを心がけてください。



母集団から無作為に
集めたサンプルデータ
の 2 群の平均の差

片側検定

片側検定が許容される可能性のある研究としては、研究薬剤が確実に血圧を下げるものである場合など、その方向性が決まっているものが挙げられます。例えば 100 人の研究者が 100 通りの研究をした場合でも、測定誤差を含め、それらの研究のうち一つも薬剤 A 群の平均血圧が薬剤 B 群の平均血圧よりも高くならないような場合を指します。測定誤差やミスは研究につきものなので、そういった事はあまり実現性がないことが分かります。

片側検定が許容されるもう一つの例は、「一方向の効果しか考慮する必要のない非劣性（Non-inferiority）を目標とした研究」があります。非劣性研究とは、「薬剤 A は薬剤 B に負けていない」ことを示す研究です。例えば薬剤 A の有効性は既存薬 B と同等か、少し劣っている（負けている）かもしれないが、あまり大きく負けていなければ、すなわち、薬剤 A の薬剤 B に対する非劣性を示せれば、安全性や利便性を考えて薬剤 A の効果を認めるということを主張したい場合です。非劣性を目指す場合、研究対象である新薬が既存薬に比べて優れている分にはいくら優れていても良いので、仮説は一方向（優れているかはさておいて、劣っているかどうか）のみに注目します。

例えば、薬剤 A 群の平均血圧を Ma、薬剤 B 群の平均血圧を Mb とします。このとき、Ma ― Mb は薬剤 A 群と薬剤 B 群の平均血圧の差を表します。薬剤 A の方が平均血圧をより低く抑えることができている、つまりこの差の値が負であれば薬剤 A が優れているという解釈ができますが、いま薬剤 A の方が劣っているかもしれないので、この値は正になる可能性もあります。ここで、劣っていないと判断する範囲のことを非劣性マージンといいます。例えば、薬剤 B に対して薬剤 A の平均血圧が 10mmHg 高くてもそれは劣っていないとみなす研究の場合、この 10mmHg が非劣性マージン（Non-inferiority margin）といえます。

このとき、薬剤 A 群の平均血圧（Ma）が、薬剤 B 群の平均血圧（Mb）よりも、劣っていないかどうかを調べる非劣性試験の仮説は非劣性マージンを用いて

帰無仮説　　Ma ― Mb　＞　10
対立仮説　　Ma ― Mb　≦　10

となります。棄却する目的で立てられる帰無仮説は「薬剤 B に対して薬剤 A は平均血圧において 10mmHg よりも劣っている」と解釈でき、対立仮説は「薬剤 B に対して薬剤 A は平均血圧において 10mmHg 以上は劣っていない」と解釈できます。このときの仮設検定には片側の P 値を使います。

P 値を用いた結果の解釈

P 値の解釈としてよくある間違いに、P 値が 0.05（5%）以上で帰無仮説が棄却されなかった場合、帰無仮説を支持して、「薬剤 A を使用した時も薬剤 B を使用した時も平均血圧は同じである」または「薬剤 A には効果がない」と言い切ってしまうことが挙げられます。P 値を用いて「同じである」という判断をするのはご法度です。例えば、ジェネリック薬品と呼ばれる後発医薬品を開発する場合など、薬剤の血中濃度などを比較して「後発医薬品は先発医薬品と生物学的に同等である」というエビデンスを確認する臨床試験があります。このように違いではなく、同じであ

ることを証明する研究を同等性試験（Equivalence Trial）と呼びます。

P値を大きくすることは非常に簡単です。P値は研究対象者数によって大きく左右されるので、研究対象者数を少なくすればいくらでもP値を大きくすることは可能です。

帰無仮説を棄却しない（P値が0.05以上）ことによって有意差が確認されなかったときに、同等性を示唆するという間違ったやり方は1980年代までは実際に用いられ、New England Journal of Medicineなどの質も注目度も高いジャーナルなどに発表された例もありました。これによって多くの薬剤が同等性を売りにしてマーケットに流出してしまったという今では信じられない歴史があります。日本でも旧厚生省による旧統計ガイドライン（1992年）が発行される以前は、そのような研究発表が多くありました。P値は実際に差があるか、ないかのみでなく、研究対象者数によっても左右されるので、P値が0.05以上で有意差がなかったからといってそれはあくまでも「薬剤Aを使用した時も薬剤Bを使用した時も平均血圧は同じである」または「薬剤Aに効果がない」という帰無仮説を棄却することができなかったというだけに過ぎません。棄却できなかった理由は、本当に効果がないのか、ただ単にデータ（研究対象者数）が不足していたからなのかはわからないのです。以上のことから、有意差が確認できた場合もできなかった場合も、「効果があった」、「効果がなかった」と結論するのではなく以下のような正しい表記を心掛けてください。

P値による結果の解釈（例）

帰無仮説　「薬剤A群の平均血圧と薬剤B群の平均血圧の差はゼロである」

P値が0.05以上だったので両側5%の有意水準を適用すると「統計的有意差は確認できなかった。」
つまり、薬剤Aと薬剤Bにおいて効果に差があることは示唆されなかった。

P値が0.05未満
両側5%の有意水準を適用し、「統計的有意差があった」。つまり、薬剤Aと薬剤Bにおいて効果に差があることが示唆された。

このように帰無仮説を棄却できなかった場合には、「薬剤Aには効果がない」というような帰無仮説が正しかったという表現ではなく、「薬剤Aには効果がある」とは言えなかった、とか示唆されなかった、という表現をするようにしてください。

P値の落とし穴

P値の計算には「比較したい群間のアウトカムの差（効果）」の他に、研究対象者数が大きく影響しています。P値のみに頼って判断した場合、P値が0.05以上の場合、「臨床的に効果がない」のか「臨床的に効果はあるが、研究対象者数が足りなかったからである」のか判断がつきません。同様にP値が0.05未満の場合、「臨床的に効果がある」のか「臨床的には（実際となると、ばらつきが大きいなどで）効果を余り期待できないが、研究対象者数が多すぎたために統計的に有意

になった」のかの区別がつきません。
何万人もの対象者からのデータを使った研究では臨床的に殆ど意味のない違い（例：薬剤の効果）でも、P 値はとても小さくなることがあるので、研究の結果を P 値のみで判断するのは大変危険です。この場合、P 値と並行して結果の判断に用いられるのが信頼区間です。

P 値と信頼区間

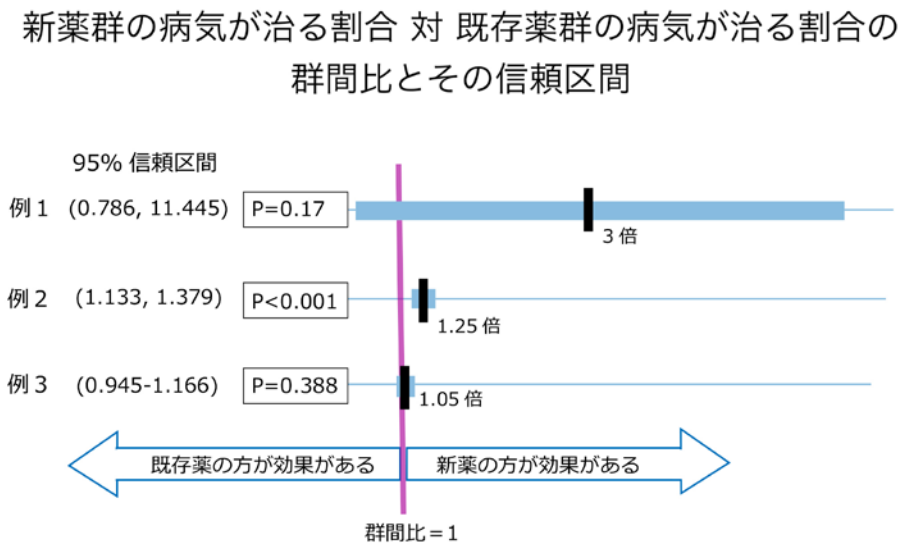
統計的有意差があるかどうか、P 値が 5%未満かどうかは信頼区間によっても示すことが可能です。

例 1
新薬を使用した 10 人の研究対象者のうち 6 人、既存薬を使用した 10 人の研究対象者のうち 2 人で病気が治癒したとします。新薬群と既存薬群の病気が治る割合はそれぞれ 60%と 20%です。新薬群の病気が治癒する割合 対 既存薬群の病気治癒する割合の群間比は 0.6/0.2 = 3 となります。すなわち、新薬の方が 3 倍病気が治癒するといえます。また、この群間比の 95%信頼区間は（0.786 – 11.445）となります。

例 2
新薬を使用した 1000 人の研究対象者のうち 500 人、既存薬を使用した 1000 人の研究対象者のうち 400 人で病気が治癒したとします。病気が治癒する割合の新薬群の既存薬群に対する群間比は 0.5/0.4 = 1.25 となります。すなわち、新薬の方が 1.25 倍病気が治癒するといえます。また、この群間比の 95%信頼区間は（1.133 – 1.379）となります。

例 3
新薬を使用した 1000 人の研究対象者のうち 420 人、既存薬を使用した 1000 人の研究対象者のうち 400 人で病気が治癒したとします。病気が治癒する割合の新薬群の既存薬群に対する群間比は 0.42/0.4 = 1.05 となります。すなわち、新薬の方が 1.05 倍病気が治癒するといえます。また、この群間比の 95%信頼区間は（0.945 – 1.166）となります。

例 1 では 95%信頼区間は、群間で差がないこと意味する「1」（＝群間比）という値をこの区間の中に含んでいます。この場合は、P 値が 5%以上であることを示唆しており、新薬の方が既存薬より 3 倍も効果があると計算されたにも関わらず、統計的有意差が確認されません。一方、例 2 では、95%信頼区間は「1」という値をこの区間の中に含んでいません。この場合、P 値が 5%未満だったことを示唆しており、新薬の方が既存薬よりも 1.25 倍しか効果がないと計算されたにもかかわらず、統計的有意差が確認されます。これは、例 2 では、例 1 と比べて群間比は 1.25 と小さかったものの、研究対象者数が多いため、95%の信頼区間が小さくなり、「1」という値を含まなかったためです。



上の図は例 1 から例 3 の 95%信頼区間を図示したものです。例 1 も例 3 も P 値が 5%より大きいので統計的有意差は確認されませんが、例 1 と例 3 では有意差の出なかった理由が異なることがわかります。例 1 は、新薬群の方が病気の治る割合が 3 倍も既存薬群に比べ高かったのにも関わらず、研究対象者数が各群 10 例と少なかったために信頼区間の幅が広くなってしまい、その区間の中に差がないことを意味する「1」という値を含んでしまい、統計的有意差が確認されません。一方、例 3 では、新薬群と既存薬群の病気の治る割合がそれぞれ 42%と 40%とほとんど変わらなかったため、研究対象者数がかなり大きくても信頼区間が 1 を含んでしまい、統計的有意差が確認されません。

信頼区間を用いた研究結果の解釈

上の例のように信頼区間を用いると、P 値を用いなくても仮説検定を行うことができます。また P 値だけでは見えなかったことも見えてくるので、解析結果をまとめるときには P 値のみでなく、効果に関する群間の違い、そしてその信頼区間もあわせて記載することを心がけて下さい。

新薬と既存薬を比較する臨床試験では以下の三つの場面が考えられ、信頼区間を用いて試験結果を解釈できます。

① 新薬の効果は既存薬の効果に勝っている　（**優越性**試験, Superiority Trial）
② 新薬の効果は既存薬の効果に負けていない　（**非劣性**試験, Non-inferiority Trial）
③ 新薬の効果は既存薬の効果と同じである　（**同等性**試験, Equivalence Trial）

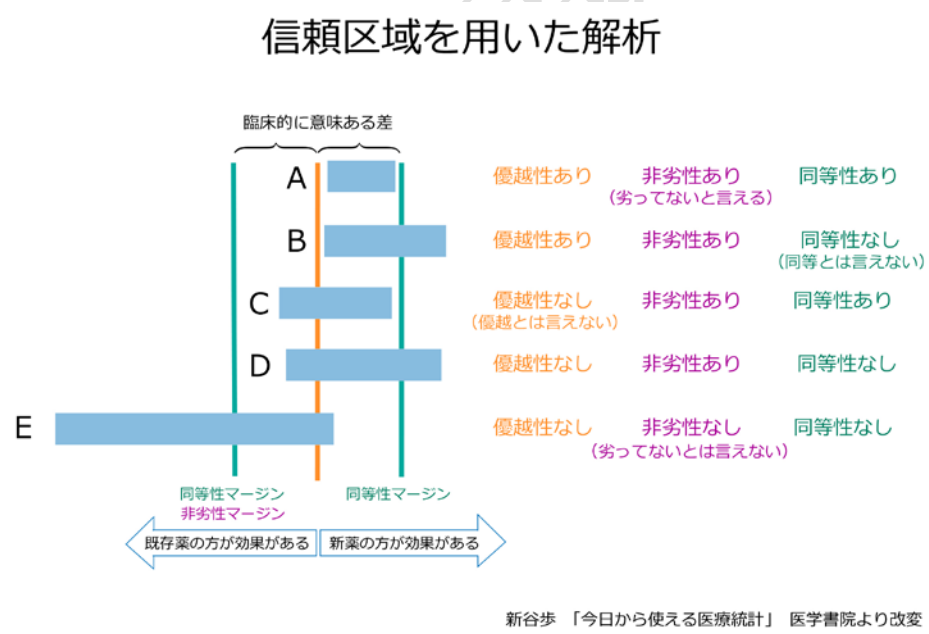
優越性を証明しようとする場合

上記①の優越性試験では、95%信頼区間に違いがないという値を含んでいなければ、それに対応する P 値は 0.05 未満であり、統計的有意差はあるといえます。群間の違いは大きく分けて差

か比で表されます。例えば、薬剤 A を投与された 50 人の平均血圧が 100mmHg、薬剤 B を投与された群の平均血圧が 113mmHg だったとします。平均血圧の差は 113-100=13mmHg ですが、平均血圧の比は 113÷100＝1.13 となります。差が用いられる場合、群間で効果に違いがないことは平均血圧の差が 0g の時を指すので、信頼区間が 0 を含んでいなければ優越性が証明されることになります。一方、比が用いられる場合は比が 1 の時に群間で効果に違いがないことを指すので、信頼区間が 1 を含んでいなければ優越性が証明されることになります。

非劣性、同等性を証明しようとする場合

②の非劣性試験でも、また③の同等性試験でも、信頼区間を用いるとわかりやすいでしょう。「新薬を用いたときも既存薬を用いたときも平均血圧は同じである」という同等性を証明しようとする試験では、例えば「新薬と既存薬を用いたときの平均血圧の群間差が 5mmHg を超えなければ同等と定義する」というような同等性を示すマージン（Δ）を試験開始前に設定します。新薬と既存薬を用いたときの平均血圧の群間差の信頼区間がこのマージンに入っていれば、同等性が統計的に証明されることになります。　同様に新薬が既存薬に劣っていないという非劣性を証明しようとする試験では、劣っていないことを示す信頼区間の上限値または下限値（下の例では下限値）が設定されたマージンを超えていなければ非劣性が統計的に証明されることになります。



新谷歩 「今日から使える医療統計」 医学書院より改変

それでは改めて先ほどの結果の例を見てみましょう。ここでは同等性・非劣性を示すマージンを試験開始前に±6%、－6%とそれぞれ設定したとします。

例 1

新薬を使用した 10 人の研究対象者のうち 6 人、既存薬を使用した 10 人の研究対象者のうち 2 人で病気が治癒したとします。新薬群と既存薬群の病気が治癒した割合はそれぞれ 60%と 20%です。新薬群の病気が治癒する割合　対　既存薬群の病気が治癒する割合の群間比は 0.6/0.2=3

となります。すなわち新薬の方が、3 倍病気が治癒するといえます。また、この群間比の 95%信頼区間は（0.786 － 11.445）となります。

➡ 同等性・非劣性マージンは±6%、－6%とそれぞれ設定されており、例 1 は信頼区間が（0.786－11.454）なので、これは差がない値（1）を含みます。したがって、優越性は示されていません。また信頼区間の下限値が 0.786 なので、これは既存薬の効果を 1 とすると新薬は 21.4%も既存薬よりも効果が低く 6%既存薬よりも効果が低いことを許容する同等性および非劣性のマージンを超えているので、同等性も非劣性も示されていません。（上図 E のパターン）

例 2

新薬を使用した 1000 人の研究対象者のうち 500 人、既存薬を使用した 1000 人の研究対象者のうち 400 人で病気が治癒したとします。病気が治癒する割合の新薬群の既存薬群に対する群間比は 0.5/0.4=1.25 となります。すなわち新薬の方が、1.25 倍病気が治癒するといえます。また、この群間比の 95%信頼区間は（1.133 － 1.379）となります。

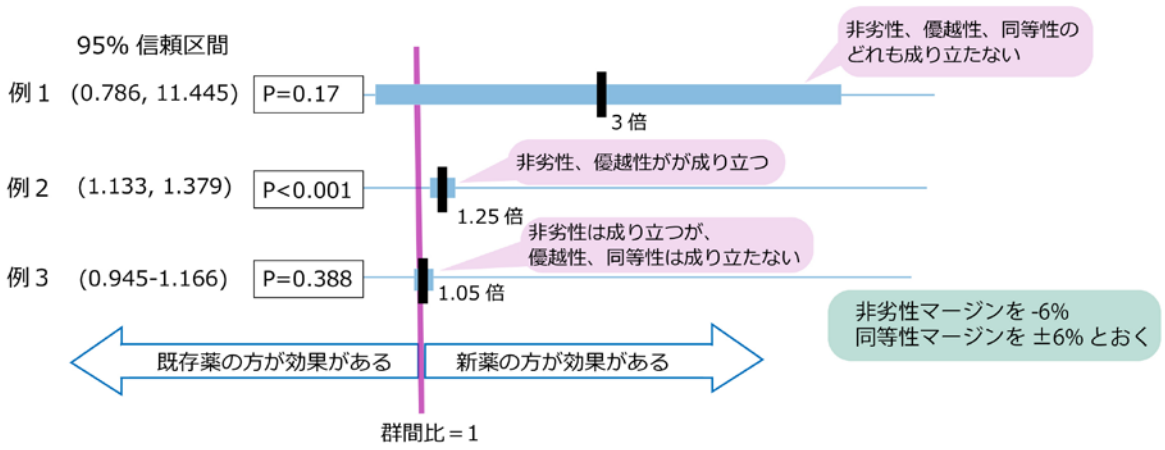
➡ 信頼区間に 1 が含まれないので優越性が示されています。また、優越性が示されているので、非劣性も示されています。（上図 A のパターン）

例 3

新薬を使用した 1000 人の研究対象者のうち 420 人、既存薬を使用した 1000 人の研究対象者のうち 400 人で病気が治癒したとします。病気が治癒する割合の新薬群の既存薬群に対する群間比は 0.42/0.4=1.05 となります。すなわち新薬の方が、1.05 倍病気が治癒するといえます。また、この群間比の 95%信頼区間は（0.945 － 1.166）となります。

➡ 下限値が 0.945 なので非劣性は示されていますが、優越性及び同等性は示されていません。（上図 D のパターン）

新薬群の病気が治る割合 対 既存薬群の病気が治る割合の群間比とその信頼区間



この単元に関係するビデオ教材

仮説検定とＰ値

本単元は日本医療研究開発機構:研究公正高度化モデルである「医系国際誌が規範とする研究の信頼性にかかる倫理教育プログラム」（略称：AMED 国際誌プロジェクト）によって作成された教材です。作成および査読等に参加した専門家の方々の氏名は、冒頭に掲載されています。

この単元に関係する国際誌におけるチェックポイントをいくつか紹介します。
（内容は解釈を助けるために一部意訳している部分もあります）

①Nature
（http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/2010128212513557501.pdf; visited on 2018.02.11）
②New England Journal of Medicine （http://www.nejm.org/page/author-center/manuscript-submission#electronic; visited on 2018.02.11）
③Science （http://www.sciencemag.org/authors/science-editorial-policies; visited 2018.02.11）
④The EMBO Journal （http://emboj.embopress.org/authorguide#embargopolicy; visited on 2018.02.11）
⑤JAMA （http://jamanetwork.com/journals/jama/pages/instructions-for-authors; visited on 2018.02.11）

① Nature
- 全ての検定において有意水準を記載すること（例：5%）
- 両側検定か片側検定かのどちらを用いたか記載すること
- 主解析についての実際のP値を記載すること

② New England Journal of Medicine
- 非劣性試験のような片側検定が必要とされるような研究デザインを除いて、すべてのP値は両側のものを使用する。0.01より大きいP値については小数点以下2桁まで、P値が0.001から0.01までの間の値を採れば小数点以下3桁まで、P値が0.001より小さい場合はP<0.001と記載すること。臨床試験の早期中止ルールにP値が用いられる場合や、ゲノムースクリーニング研究にP値が用いられる場合はこの通りでない。
- ランダム化試験の群間を比較する背景表には、表の凡例部分に統計的有意差を示唆する結果（例：P<0.05）を記載し、表の中にはP値を載せるべきでない。

③ Science
- それぞれの検定で統計的有意差の判定に用いられた基準について（片側検定か両側検定かなど）についても記載すること。通常は両側検定を用いるべきであるが、片側検定を用いる場合はその正当性の理由づけを行うこと。
- 論文の結論が適切な統計解析によって得られたものであることが分かるように、解析結果については詳細まで記載しておくこと。またその際の制限が生じた事項などについても正直に記載するべきである。
- それぞれの統計検定の結果については、十分に統計量とP値をあわせて記載し、有意差があった・なかったのみに言及しないこと。遺伝解析のように多重性の補正を試みなければならないというような状況でない限り、P値については3桁以上の有効数字を表記する必要はない。

④ The EMBO Journal
- 行ったすべての仮説検定についてはその検定の名称と、利用した各時点の症例数、結果として得られたP値などを記載する必要がある。このときP値については統計的に有意か否かではなく、P値そのものを記載する。

⑤ JAMA
- 仮説検定を行ったときには、P値のみでは実際の差などの重要な情報が得られないため、P値のみの記載は避けること。

■ P 値のみでは実際にどの程度の差があったのかなどの情報が得られないため、頻度や差などの比較結果とあわせて記載することが勧められる（例：0.8%（95%CI; 0.2% to 1.8%）; P=0.13）。

■ P 値を報告する場合、0.001 未満の P 値は"P<.001"と記載し、0.001 から 0.01 の P 値は小数点第 3 桁まで、0.1 以上の場合は小数点以下第 2 桁まで、0.99 より大きい場合は"P>.99."と記載すること。遺伝解析などで得られるような極端に小さい P 値については、$P = 1 \times 10^{-5}$ などと表記することも可能。原則として t 統計量・F 統計量・χ^2 のような統計量や自由度については記載の必要はない。

■ 複数群を比較する無作為化比較試験において群間の背景比較において、P 値を計算することは科学的に妥当ではないため避けること。比較群間での背景情報の差を表す場合は P 値ではなく、臨床的に重要な偏りであるかどうか、またその偏りを多変量解析などで考慮したか否かなどを報告すべきである。
