統計学から見た臨床試験と臨床研究のツボ

寺子屋・統計庵　其之三

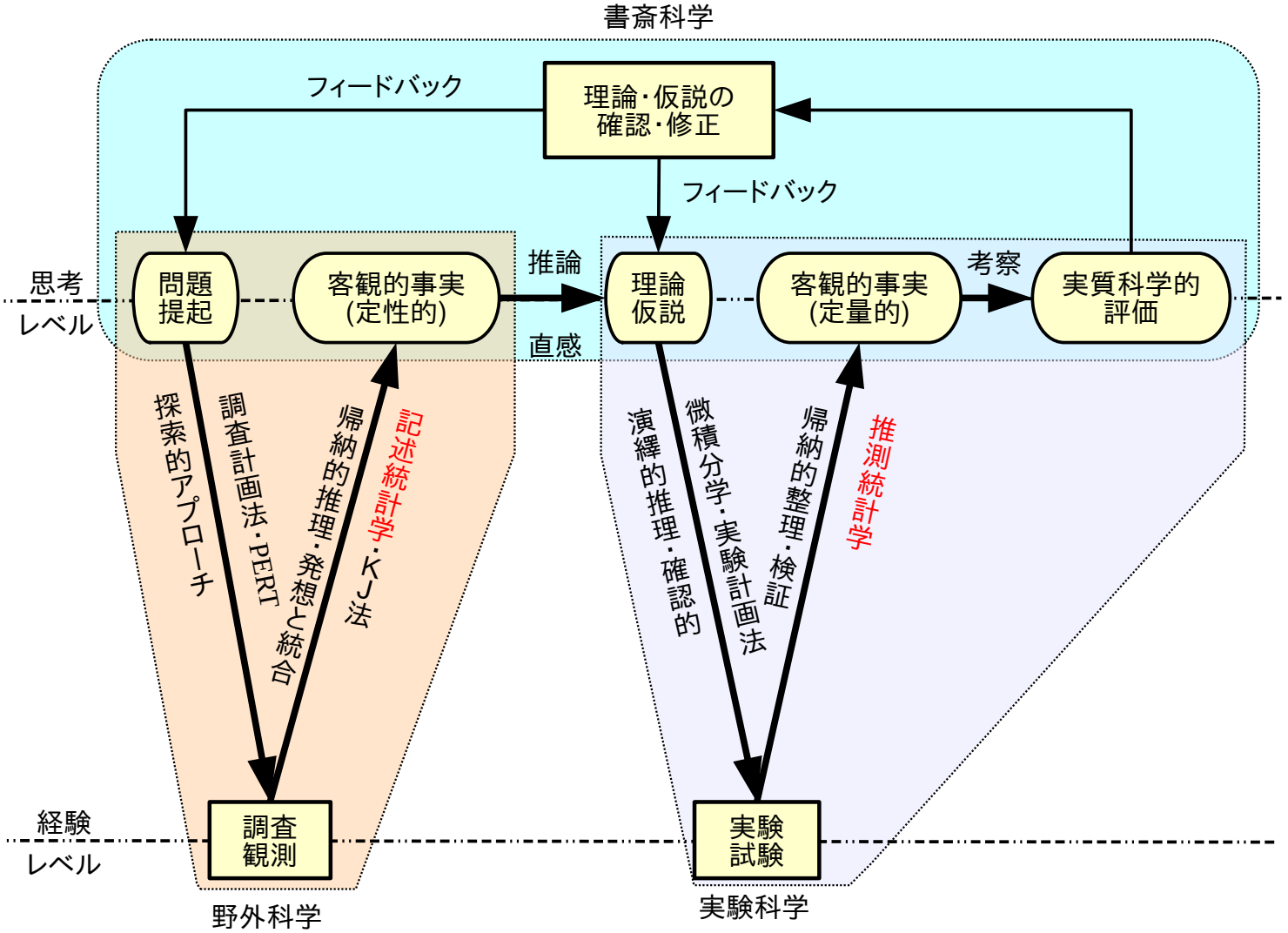
杉本解析差有比数

杉本典夫

統計学から見た臨床試験と臨床研究のツボ

- 科学的研究の進め方–仮説演繹法
- 科学的研究の種類とデザイン
- 推定と検定
- 推定の原理
- 統計的仮説検定の原理
- 必要例数の計算と検出力分析
- 検定と推定と科学的判断の関係–検定廃止論

科学的研究の進め方–仮説演繹法



W型解決法の応用による仮説演繹法の手順

書斎科学：主として頭の中で行う作業だけで成立する科学

野外科学：現場の調査や観測が中心になる科学–探索的

実験科学：実験や試験を中心にした科学–検証的

PICO(ピコ)とPECO(ペコ)

P	Patients Population Problem	対象	誰に対して?
I/E	Intervation Exposure	介入 原因(暴露)	何をすると? 何によって?
C	Comparison	比較	何と比較して?
O	Outcom	結果(帰結)	どうなるか?

問題提起:CQ(Clinical Question)→RQ(Research Question)→PICO/PECO

探索型研究:主としてPECO　検証型研究:主としてPICO

FINERチェック

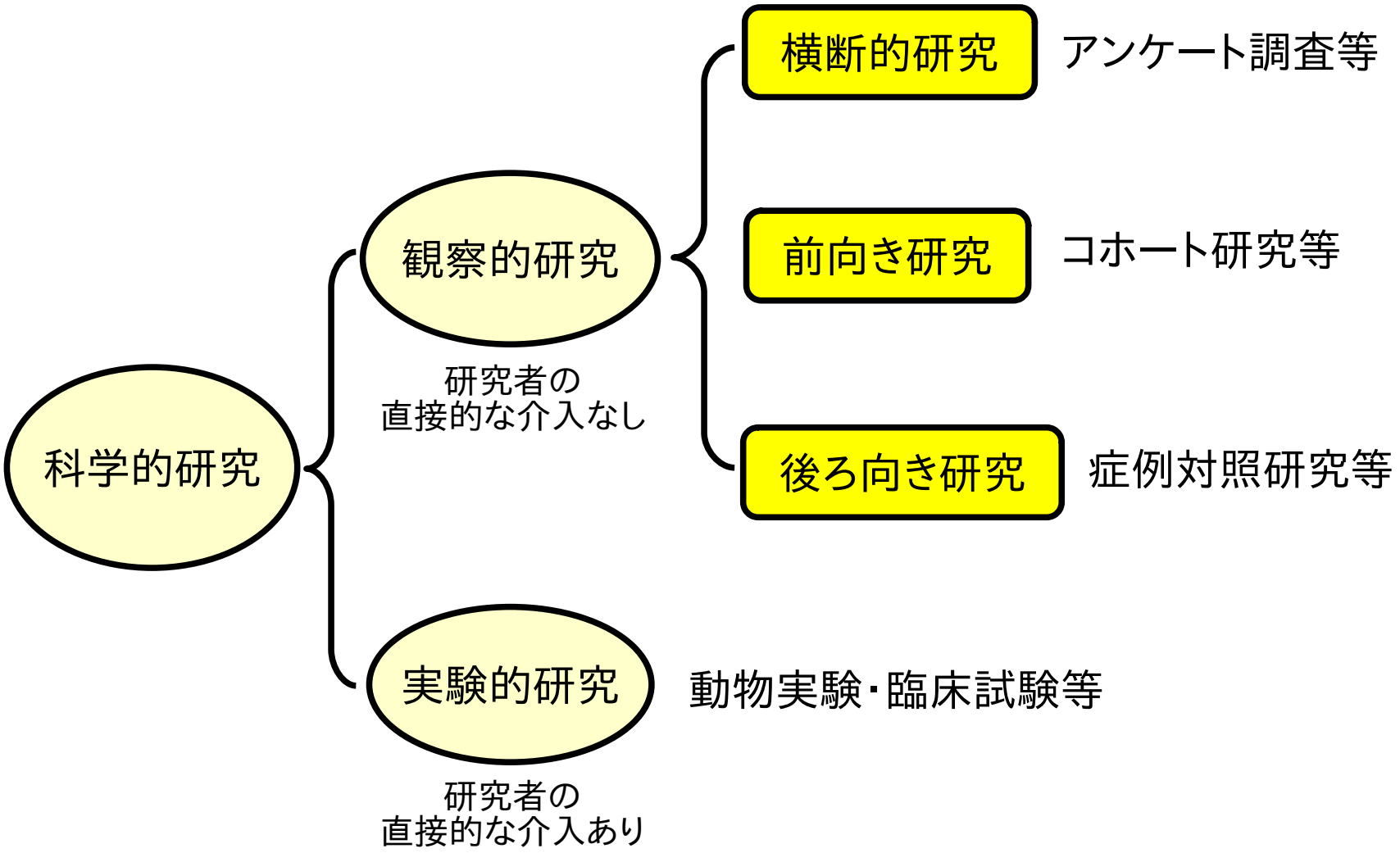
F	Feasible	実現可能か？
I	Interesting	科学的関心(学術的価値)は高いか？
N	Nobel	新規性はあるか？
E	Ethical	倫理性に配慮されているか？
R	Relevant	社会的必要性があるか？

FINERチェックによって研究計画を最終確認する

統計学から見た臨床試験と臨床研究のツボ

- 科学的研究の進め方–仮説演繹法
- 科学的研究の種類とデザイン
- 推定と検定
- 推定の原理
- 統計的仮説検定の原理
- 必要例数の計算と検出力分析
- 検定と推定と科学的判断の関係–検定廃止論

科学的研究の種類とデザイン



研究デザインを原因と結果の2×2分割表で理解する

	結果:無	結果:有	計
原因:無	a	b	(a+b)
原因:有	c	d	(c+d)
計	(a+c)	(b+d)	N

原因:タバコなどの危険因子

結果:肺癌などの疾患

横断的研究

	肺癌：無	肺癌：有	計
タバコ：無	a=55(55%)	b=5(5%)	(a+b)=60(60%)
タバコ：有	c=25(25%)	d=15(15%)	(c+d)=40(40%)
計	(a+c)=80(80%)	(b+d)=20(20%)	N=100(100%)

ある時点におけるデータを原因も結果も指定せずに横断的に観測する研究法

全体の例数Nを指定し、ある時点におけるタバコと肺癌の有無を調べてa、b、c、dを観測する

→N以外のデータは誤差を含む(値が変動する)

※割合を求める時は定数を分母にする

- 比較的手軽で迅速に実施できる
- 因果関係の検証はできない
- 主として探索的研究や予備調査に用いられる
- アンケート調査やスクリーニング調査が代表例

横断的研究

	肺癌：無	肺癌：有	計
タバコ：無	a=55(55%)	b=5(5%)	(a+b)=60(60%)
タバコ：有	c=25(25%)	d=15(15%)	(c+d)=40(40%)
計	(a+c)=80(80%)	(b+d)=20(20%)	N=100(100%)

喫煙率(危険因子の出現率)：$p_R=\frac{c+d}{N}=\frac{40}{100}=0.4$

肺癌の有病率(prevalence)：$p_d=\frac{b+d}{N}=\frac{20}{100}=0.2$

四分点相関係数(ファイ係数)：$\phi=\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}=\frac{15\times55-25\times5}{\sqrt{40\times60\times20\times80}}\approx0.357$

クラメールの連関係数：$V=\theta=\frac{|ad-bc|}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}=\frac{|15\times55-25\times5|}{\sqrt{40\times60\times20\times80}}\approx0.357$

四分点相関係数(ファイ係数)：2×2分割表における相関係数←この検定がMantel-Haenszelの検定
クラメールの連関係数：2種類の分類データの関連性の指標←この検定がχ^2検定(連続修正なし)
※連関係数は1度数あたりの実現度数と理論度数の食い違い量を表す

前向き研究

	肺癌:無	肺癌:有	計
タバコ:無	a=40(80%)	b=10(20%)	(a+b)=50(100%)
タバコ:有	c=20(40%)	d=30(60%)	(c+d)=50(100%)
計	(a+c)=60(60%)	(b+d)=40(40%)	N=100(100%)

原因の有無を指定し、ある時点から未来に向かって結果を観測する研究法

タバコ無の例数(a+b)と有の例数(c+d)を指定し

それらの群について肺癌発症の有無を調べてa、b、c、dを観測する

→(a+b)と(c+d)とN以外のデータは誤差を含む(値が変動する)

- 実施に手間と時間がかかる
- 因果関係の検証が可能
- 主として検証的研究に用いられる
- コホート研究が代表例

前向き研究

	肺癌:無	肺癌:有	計
タバコ:無	a=40(80%)	b=10(20%)	(a+b)=50(100%)
タバコ:有	c=20(40%)	d=30(60%)	(c+d)=50(100%)
計	(a+c)=60(60%)	(b+d)=40(40%)	N=100(100%)

非喫煙群における肺癌の発症率:$p_- = \frac{b}{a+b} = \frac{10}{50} = 0.2$

喫煙群における肺癌の発症率:$p_+ = \frac{d}{c+d} = \frac{30}{50} = 0.6$

リスク差(絶対危険度): $RD = p_+ - p_- = \frac{d}{c+d} - \frac{b}{a+b} = 0.6 - 0.2 = 0.4$

リスク比(相対危険度): $RR = \frac{p_+}{p_-} = \frac{d(a+b)}{b(c+d)} = \frac{0.6}{0.2} = 3$

非喫煙群における肺癌オッズ:$O_- = \frac{b}{a} = \frac{10}{40} = 0.25$

喫煙群における肺癌オッズ:$O_+ = \frac{d}{c} = \frac{30}{20} = 1.5$

オッズ比:$OR = \frac{O_+}{O_-} = \frac{d/c}{b/a} = \frac{ad}{bc} = \frac{30 \times 40}{10 \times 20} = 6$

リスク差=出現率の差←この検定がリスク差のχ^2検定

※Fisherの正確検定は本来は片側検定用

リスク比=出現率の比←この検定がリスク比のχ^2検定

※これはリスク差のχ^2検定とは別の手法

※出現率が10%未満の時、出現率は指数関数的に変化する

→対数出現率の差(対数リスク差)を指標にする

オッズ(見込み比):有の確率と無の確率の比

オッズ比:ある群のオッズと別の群のオッズの比

※この場合はタバコと肺癌の関連性の指標

※出現率が10%未満の時、オッズ比はリスク比に近似する

後ろ向き研究

	肺癌：無	肺癌：有	計
タバコ：無	a=40(80%)	b=20(40%)	(a+b)=60(60%)
タバコ：有	c=10(20%)	d=30(60%)	(c+d)=40(40%)
計	(a+c)=50(100%)	(b+d)=50(100%)	N=100(100%)

結果の有無を指定し、ある時点から過去にさかのぼって原因を観測する研究法

肺癌無の例数(a+c)と有の例数(b+d)を指定し

それらの群についてタバコの有無を調べてa、b、c、dを観測する

→(a+c)と(b+d)とN以外のデータは誤差を含む(値が変動する)

- 心筋梗塞のような稀な疾患の研究に適している
- 因果関係の検証はできない
- 主として探索的研究に用いられるが、稀な疾患では検証的研究に用いられることもある
- 症例対照研究が代表例

後ろ向き研究

	肺癌:無	肺癌:有	計
タバコ:無	a=40(80%)	b=20(40%)	(a+b)=60(60%)
タバコ:有	c=10(20%)	d=30(60%)	(c+d)=40(40%)
計	(a+c)=50(100%)	(b+d)=50(100%)	N=100(100%)

喫煙率の差$=\frac{d}{b+d}-\frac{c}{a+c}=\frac{30}{50}-\frac{10}{50}=0.6-0.2=0.4$

感度:$sn=\frac{d}{b+d}=0.6$　特異度:$sp=\frac{a}{a+c}=0.8$

陽性尤度比:$LR_{+}=\frac{sn}{1-sp}=\frac{d(a+c)}{c(b+d)}=\frac{30\times50}{10\times50}=3$

陰性尤度比:$LR_{-}=\frac{1-sn}{sp}=\frac{b(a+c)}{a(b+d)}=\frac{20\times50}{40\times50}=0.5$

正常群における喫煙オッズ:$O_c=\frac{c}{a}=\frac{10}{40}=0.25$

肺癌群における喫煙オッズ:$O_d=\frac{d}{b}=\frac{30}{20}=1.5$

オッズ比:$OR=\frac{O_d}{O_c}=\frac{d/b}{c/a}=\frac{ad}{bc}=\frac{30\times40}{10\times20}=6$

喫煙率の差はリスク差に相当する指標
※リスク差のχ^2検定と同じ手法で検定可能
陽性尤度比と陰性尤度比はリスク比に相当する指標
※リスク比のχ^2検定と同じ手法で検定可能

この場合のオッズ比はタバコの有無に関するオッズ比だが
結果的に肺癌の有無に関するオッズ比と一致する
→オッズ比の不変性
※不変性:データの取り方によって値が変化しない性質
※有病率が10%未満の時、オッズ比はリスクに近似する
→後ろ向き研究から前向き研究のリスク比を推測できる

実験的研究

	肺癌:無	肺癌:有	計
タバコ:無	a=40(80%)	b=10(20%)	(a+b)=50(100%)
タバコ:有	c=20(40%)	d=30(60%)	(c+d)=50(100%)
計	(a+c)=60(60%)	(b+d)=40(40%)	N=100(100%)

研究者が原因の有無に介入するため必ず前向き研究になる

RCT(Randomized Controlled Trial、無作為化比較対照試験)

被験者をタバコ無群(対照群)と有群(処置群)に分けて結果の有無を比較する試験方法

二重盲検法(Double Blind Method)

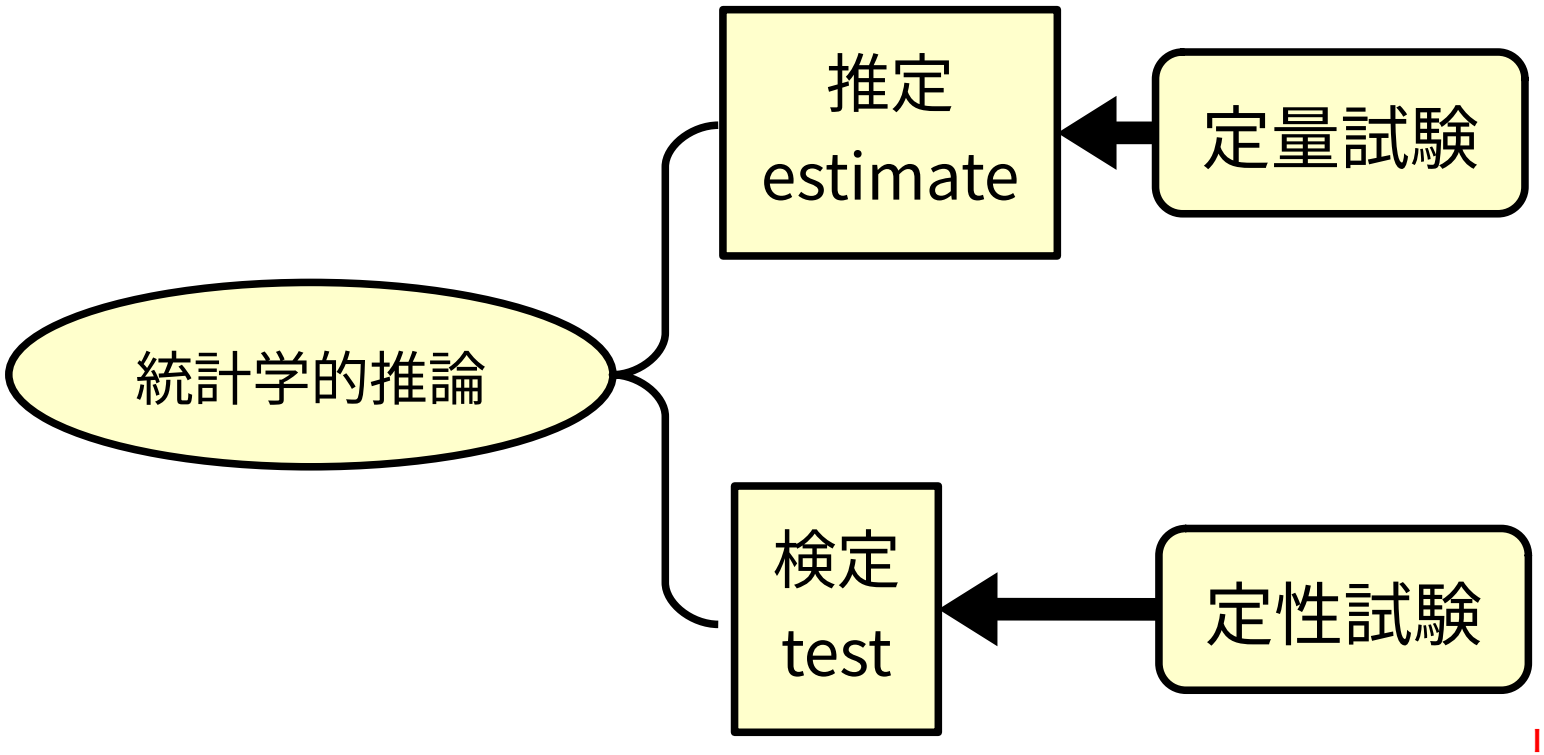
プラセボ効果を均等にするために被験者と評価者の両方にブラインドをかける試験方法

- 実施に手間と時間がかかる
- 因果関係の検証が可能
- 主として検証的研究に用いられる
- 動物実験や臨床試験が代表例

統計学から見た臨床試験と臨床研究のツボ

- 科学的研究の進め方–仮説演繹法
- 科学的研究の種類とデザイン
- 推定と検定
- 推定の原理
- 統計的仮説検定の原理
- 必要例数の計算と検出力分析
- 検定と推定と科学的判断の関係–検定廃止論

推定と検定



母数を推測
↓
pH計に相当

母数が基準値と
等しいかどうかを
○×式で推測
↓
リトマス試験紙に相当

検定よりも推定の方が重要

ところが研究現場や厚労省では検定が偏重されている
↑
○×式の方が採点が楽！

推定と検定の基本原理−中心極限定理

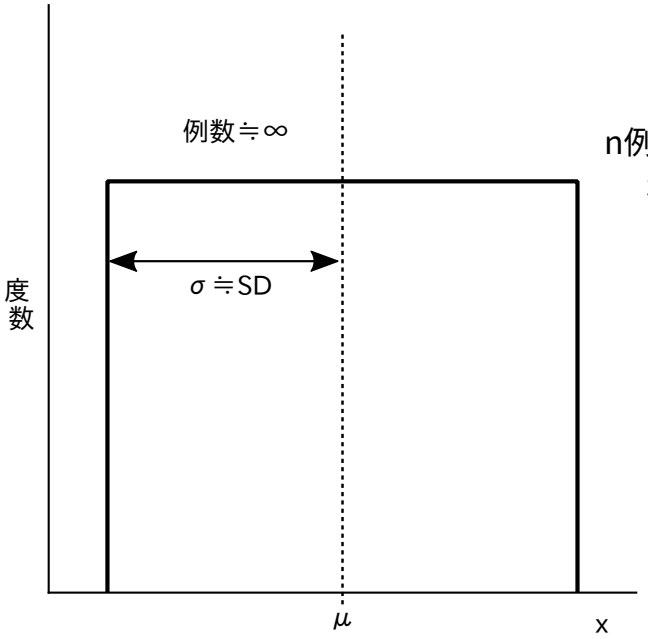


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

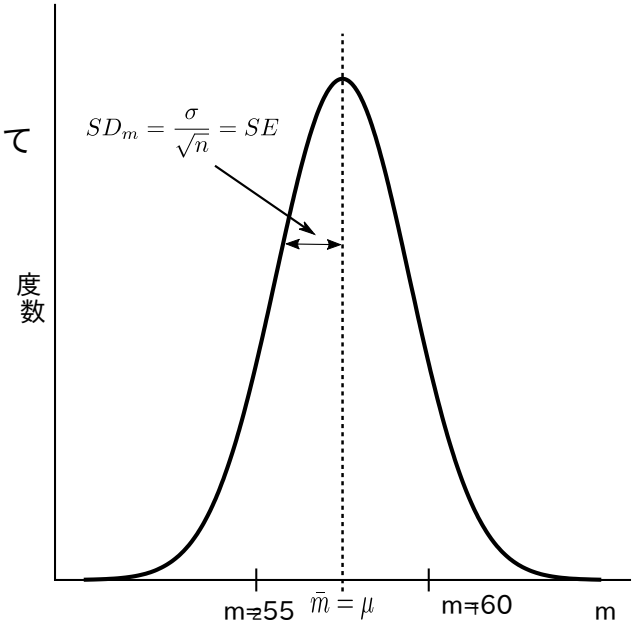


図1.4 標本平均値の分布

標本平均の分布の特徴

1. 母集団がどんな分布をしていても、漸近的に(nが多いほど)正規分布に近似する
   →中心極限定理(推測統計学の基本定理)
2. 標本平均mの平均値$\bar{m}$は母平均μと一致する
3. 標本平均の標準偏差SD_mは次のような値になる→標準誤差SE

$$SD_m = \frac{\sigma}{\sqrt{n}} \fallingdotseq \frac{SD}{\sqrt{n}} = SE$$

σ：母標準偏差　　n：標本集団の例数

SD：標本集団から求めたσの推測値

統計学から見た臨床試験と臨床研究のツボ

- 科学的研究の進め方–仮説演繹法
- 科学的研究の種類とデザイン
- 推定と検定
- 推定の原理
- 統計的仮説検定の原理
- 必要例数の計算と検出力分析
- 検定と推定と科学的判断の関係–検定廃止論

推定の原理

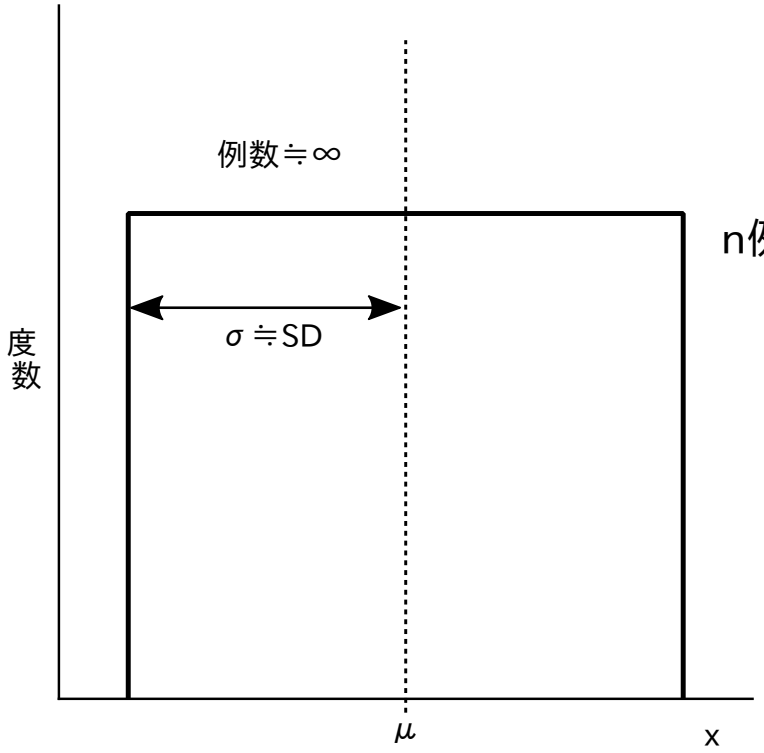


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

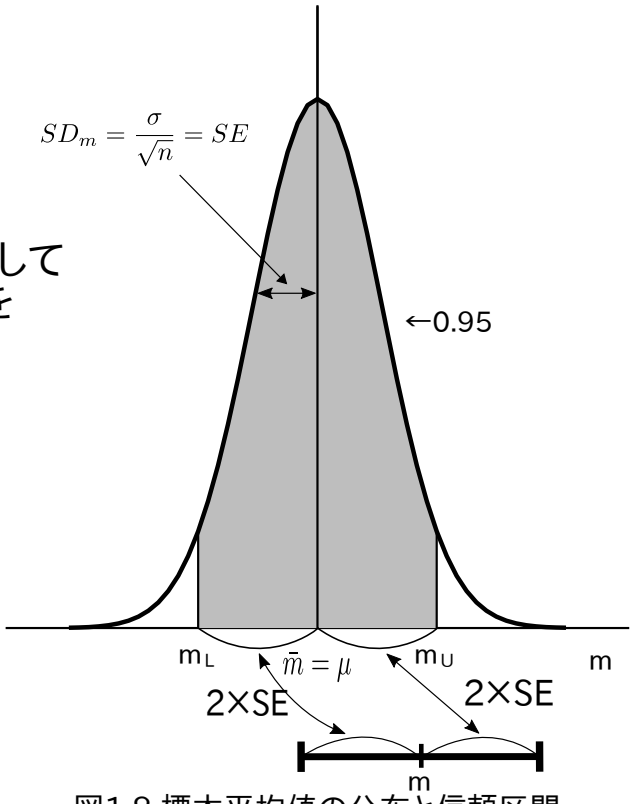


図1.8 標本平均値の分布と信頼区間

点推定(point estimation)：母数(母平均)をピンポイント(標本平均)で推測

区間推定(interval estimation)：母数をある程度の幅(信頼区間)を持たせて推測

母平均の区間推定法

標本平均mの分布は近似的に正規分布になる

↓

mの平均値は母平均μになり、mの標準偏差は標準誤差SEになる
μ±2×SEの範囲に約95%のmが含まれる

↓

ある標本平均mがμ±2×SEの範囲に含まれる確率は約95%

↓

逆にm±2×SEの範囲にμが含まれる確率も約95%

↓

95%信頼区間：$\mu \fallingdotseq m \pm 2 \times SE \rightarrow \mu_L = m - 2 \times SE \quad \mu_U = m + 2 \times SE$

95%信頼区間(95%CI)：母平均が95%の確率で含まれる区間、95%信頼限界(95%CL)

μ_L：信頼区間下限　　μ_U：信頼区間上限　　95%：信頼係数

統計学から見た臨床試験と臨床研究のツボ

- 科学的研究の進め方–仮説演繹法
- 科学的研究の種類とデザイン
- 推定と検定
- 推定の原理
- 統計的仮説検定の原理
- 必要例数の計算と検出力分析
- 検定と推定と科学的判断の関係–検定廃止論

統計的仮説検定の原理

問題：日本人の平均体重は50kgか？

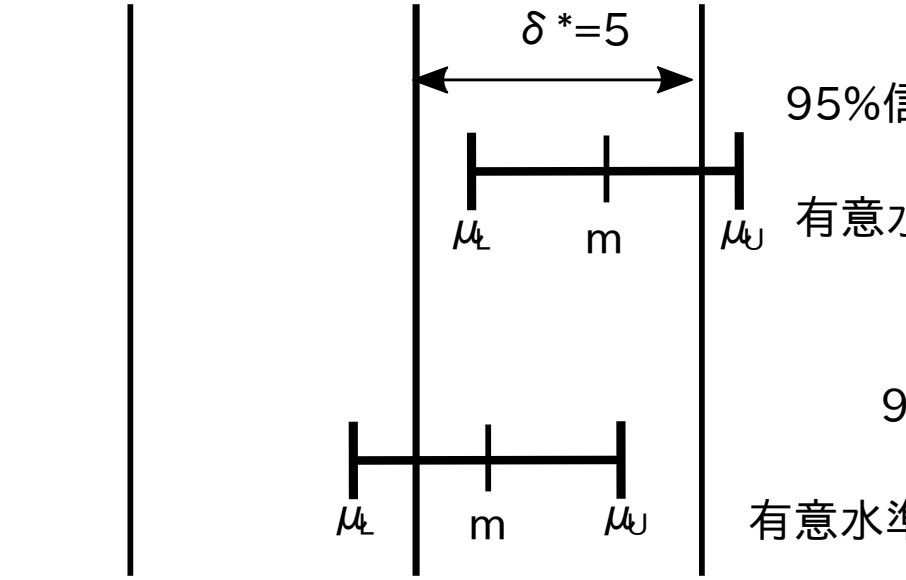
帰無仮説H_0:日本人の平均体重は50kgである←問題の答えは○

対立仮説H_1:日本人の平均体重は45kgまたは55kgである←問題の答えは×

統計的仮説検定の手順

1. 問題を設定する→医学的意義のある基準値μ_0=50と許容範囲(検出差)δ^*=±5を決める
2. 問題の答を○×式で設定する→帰無仮説H_0と具体的な対立仮説H_1を設定する
3. データに基づいて仮説の妥当性を判定する

検定は定性試験だから定量試験である推定結果から判定可能



95%信頼区間にμ_0が含まれない
↓
有意水準5%で有意→$\mu \neq \mu_0=50$

95%信頼区間にμ_0が含まれる
↓
有意水準5%で有意ではない→$45<\mu<55$

図1.13 信頼区間と統計的仮説検定

標本集団：n=100例　標本平均m=60kg　標準偏差SD=10kg　標準誤差SE=1kg

95%信頼区間: $\mu = 60 \pm 2 \times \frac{10}{\sqrt{100}} = 60 \pm 2 \rightarrow \mu = 58 \sim 62$

母平均は95%の確率で58～62kgの間にある
→母平均は95%以上の確率で50kgではない

母集団から見た統計的仮説検定

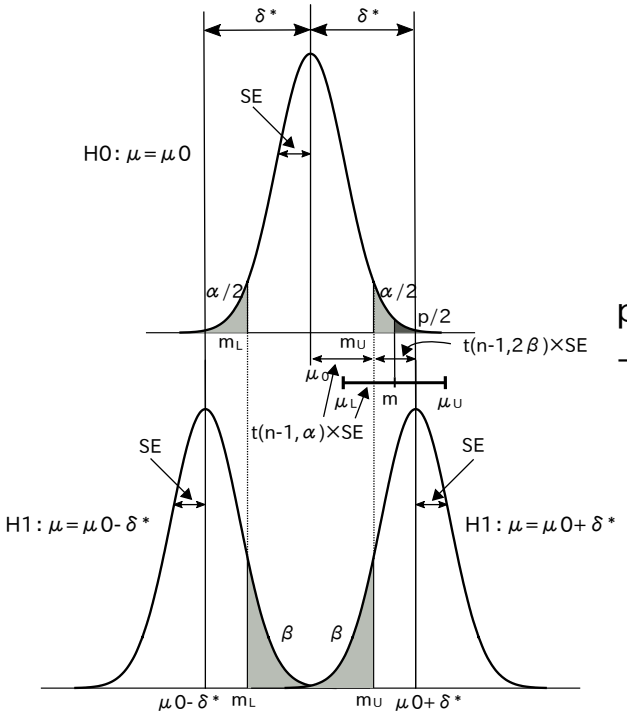


図1.14 統計的仮説検定の模式図

p/2<α/2 の時 mは棄却域に入っている
→ p<α の時 mは棄却域に入っている
p:有意確率

μ_0=50の時(帰無仮説が正しい時):標本平均値の(1-α)が含まれる区間=50±2×SE=m_L～m_U
$\mu_0-\delta^*$=45の時(対立仮説が正しい時):標本平均値の(1-β)が含まれる区間=-∞～m_L
または$\mu_0+\delta^*$=55の時(対立仮説が正しい時):標本平均値の(1-β)が含まれる区間=m_U～∞
m_L:下側棄却域の上限　　m_U:上側棄却域の下限

αエラー=アワテの言い過ぎ:帰無仮説が正しい時にアワテて$\mu \neq \mu_0$=50と結論する確率
βエラー=ボンヤリの見逃し:対立仮説が正しい時にボンヤリして45＜μ＜55と結論する確率
※(1-β):検出力=対立仮説が正しい時に$\mu \neq \mu_0$=50と結論する確率 ← 医学分野では80%

統計的仮説検定結果の解釈

棄却域に標本平均値が入っている時＝95%信頼区間に基準値が入っていない時
統計学的結論:日本人の平均体重は50kgではない ← 問題の答えは×

「有意水準5%で有意」または「危険率5%で有意」と表現する
これは「日本人の平均体重は45kgまたは55kg」の採用ではないことに注意!
※統計学的結論が間違っている確率は5%(α=0.05)

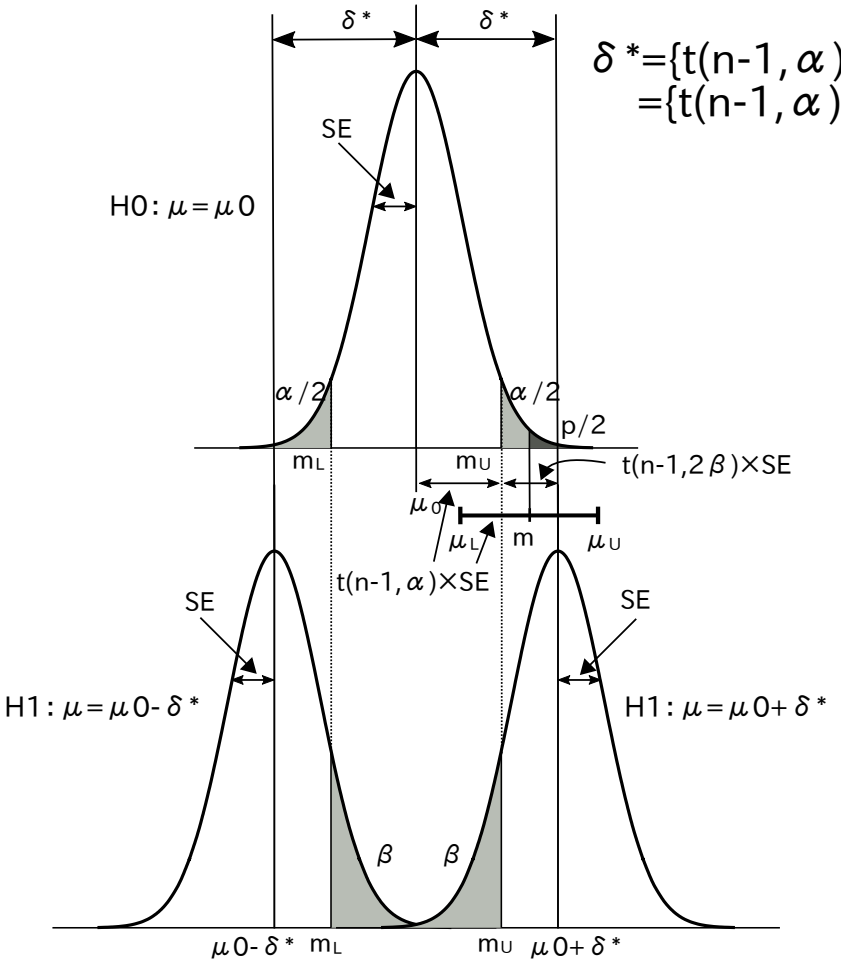
棄却域に標本平均値が入っていない時＝95%信頼区間に基準値が入っている時
かつ信頼区間が許容範囲内に収まっている時
統計学的結論:日本人の平均体重は45kgよりも重く55kgよりも軽い ← 問題の答えは△

「有意水準5%で有意ではない」または「危険率5%で有意ではない」と表現する
これは「日本人の平均体重は50kgである」の採用ではないが、実質的には同じ意味
※統計学的結論が間違っている確率は20%(β=0.2) ← $\alpha=\beta$にするのが理想

統計学から見た臨床試験と臨床研究のツボ

- 科学的研究の進め方–仮説演繹法
- 科学的研究の種類とデザイン
- 推定と検定
- 推定の原理
- 統計的仮説検定の原理
- 必要例数の計算と検出力分析
- 検定と推定と科学的判断の関係–検定廃止論

統計的仮説検定は事前に試験の必要例数を計算する



$\delta^* = \{t(n-1,\alpha)+t(n-1,2\beta)\}SE$
$= \{t(n-1,\alpha)+t(n-1,2\beta)\}(\sigma/\sqrt{n})$

図1.14 統計的仮説検定の模式図

必要例数の計算式: $n = \left[\{t(\infty,\alpha)+t(\infty,2\beta)\}\frac{\sigma}{\delta^*}\right]^2$

信頼区間を許容範囲よりも小さくする必要がある

統計的仮説検定は事後に試験の検出力分析を行う

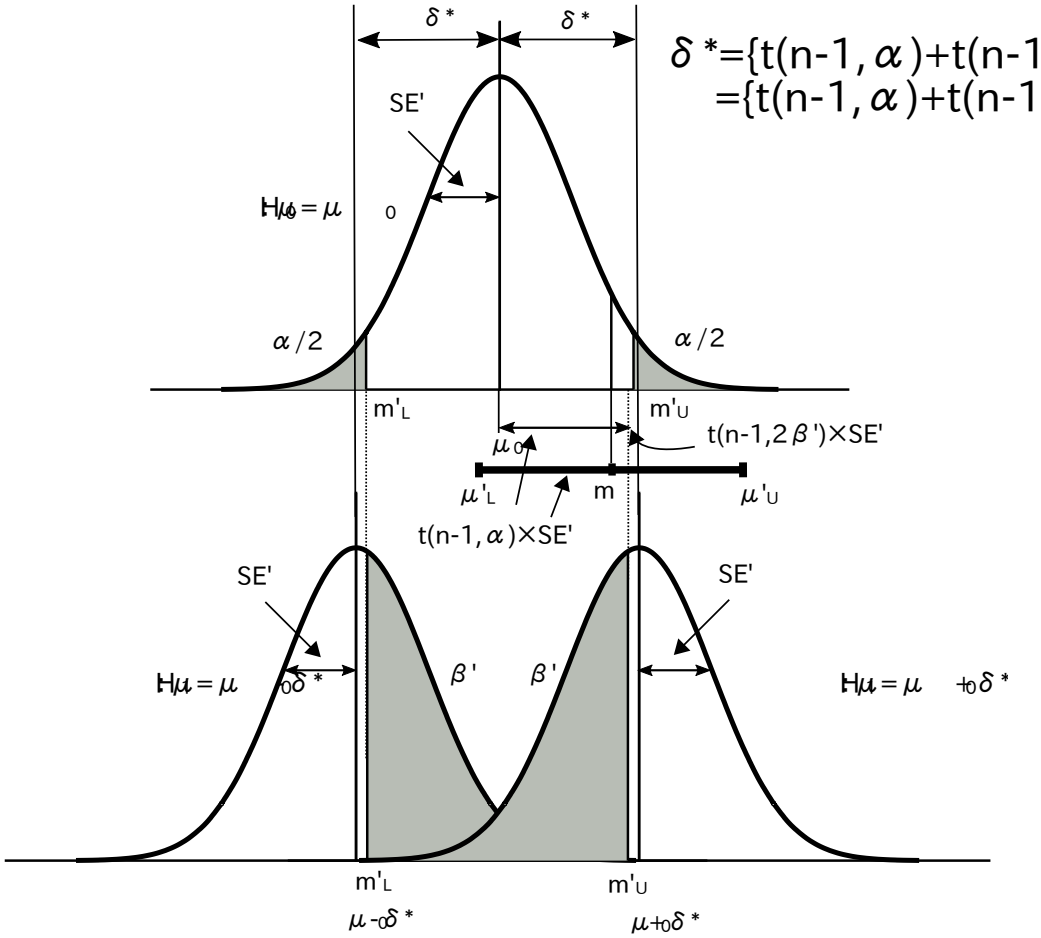


図1.6.8 統計的仮説検定の実際の模式図

実際のnとSDを用いた検出力：$\delta^*=\{t(n-1,\alpha)+t(n-1,2\beta')\}(SD/\sqrt{n})$

実際の検出力$(1-\beta')$が事前に設定した$(1-\beta)$以上であることを確認する

統計学から見た臨床試験と臨床研究のツボ

- 科学的研究の進め方–仮説演繹法
- 科学的研究の種類とデザイン
- 推定と検定
- 推定の原理
- 統計的仮説検定の原理
- 必要例数の計算と検出力分析
- 検定と推定と科学的判断の関係–検定廃止論

検定結果だけから科学的な判断をするのは危険

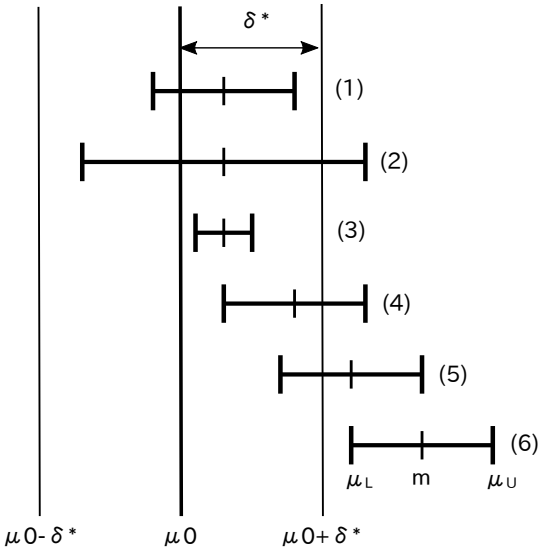


図1.15 検定結果と信頼区間

	検定結果	推定結果	実質科学的な判断
(1)	有意ではない	$\mu \fallingdotseq \mu_0$	母平均値は基準値とほぼ等しい
(2)	有意ではない	$\mu = \mu_0 \sim \mu_0 + \delta^*$	この結果だけでは判断できない 信頼区間をもっと狭くする必要がある(例数を増やす)
(3)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(4)	有意	$\mu \fallingdotseq \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない可能性が高い
(5)	有意	$\mu \fallingdotseq \mu_0 + \delta^*$	母平均値は基準値よりも大きい可能性が高い
(6)	有意	$\mu_0 + \delta^* < \mu$	母平均値は基準値よりも大きい

※生物学的同等性試験では推定結果を重視し、検定結果は参考程度 → 検定廃止論

例数が多いと検定結果は科学的な判断には使えない

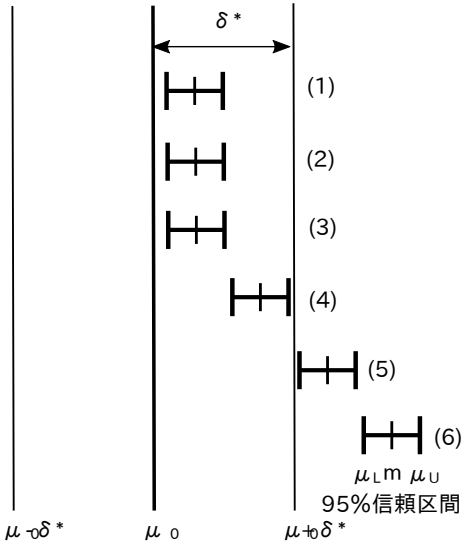


図1.7.6 例数が多い時の検定結果と信頼区間

	検定結果	推定結果	実質科学的な判断
(1)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(2)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(3)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(4)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(5)	有意	$\mu_0 + \delta^* < \mu$	母平均値は基準値よりも大きい
(6)	有意	$\mu_0 + \delta^* < \mu$	母平均値は基準値よりも大きい

※生物学的同等性試験では推定結果を重視し、検定結果は参考程度 → 検定廃止論

# p<0.001になっても結果の信頼性は95%

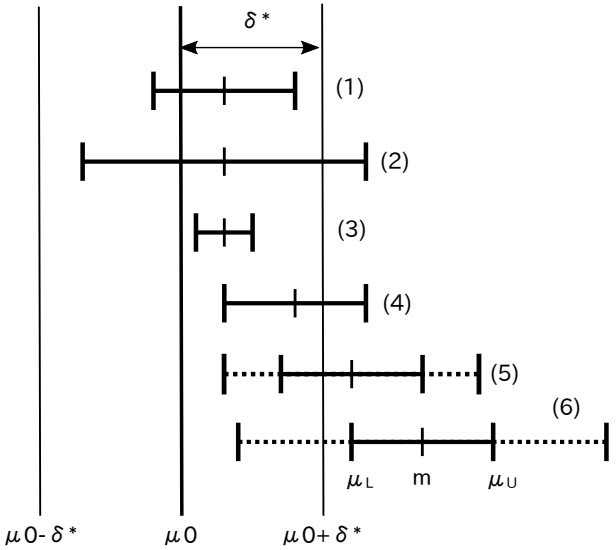


図1.20 検定結果と信頼係数を変えた信頼区間

(1)と(2):p>0.05　(3)と(4):p<0.05　(5):p<0.01　(6):p<0.001になった時

(5)を「有意水準1%で有意」と表現すると99%信頼区間が対応

(6)を「有意水準0.1%で有意」と表現すると99.9%信頼区間が対応

→ 信頼区間の幅が広くなって($\mu_0+\delta^*$)が入ってしまい、結論が曖昧になる

p値はmが棄却域に入っているかどうかを判定する目安にすぎない!

必要例数を計算した時の有意水準(αエラー)によって結果の信頼性が決まる

有意症・有意症症候群は難治性疾患

有意症(significantosis)・有意症症候群(significant syndrome)

「有意差あり＝実質科学的に有意義な差がある」とか

「有意差なし＝実質科学的に有意義な差がない」と誤解する難治性の疾患

医学界や厚生労働省等で大流行中！

有意症・有意症症候群の予防策

- 「有意差あり」や「有意差なし」という用語を使わず
  「差は有意である」や「差は有意ではない」という用語を使う
- 推定結果を重視する

本日の結語

検定結果ではなく

推定結果に基づいて

科学的判断をしましょう！

ご清聴ありがとうございました