統計学の落とし穴

- 標準偏差と標準誤差
- 有意性検定と統計的仮説検定
- パラメトリック手法とノンパラメトリック手法
- ハンディキャップ方式の検定

標準偏差はデータのバラツキ具合を表す要約値

偏差(バラツキの定義)：$d_i = x_i - m$

平方和：$SS = S_{xx} = \sum_{i=1}^{n} d_i = \sum (x_i - m)^2$

分散：$V = \frac{SS}{n} = \frac{\sum d_i^2}{n} = \frac{\sum (x_i - m)^2}{n}$

標準偏差：$s = SD = \sqrt{V} = \sqrt{\frac{SS}{n}} = \sqrt{\frac{\sum d_i^2}{n}} = \sqrt{\frac{\sum (x_i - m)^2}{n}}$

標準偏差の特徴

- データ1個あたりのバラツキ具合を表す
- 正規分布では平均値から分布の変曲点までの距離になる
- 正規分布では平均値±標準偏差の間に全データの約68%が含まれる
- 正規分布では平均値±2×標準偏差の間に全データの約95%が含まれる

標準誤差は推測統計学独特の要約値

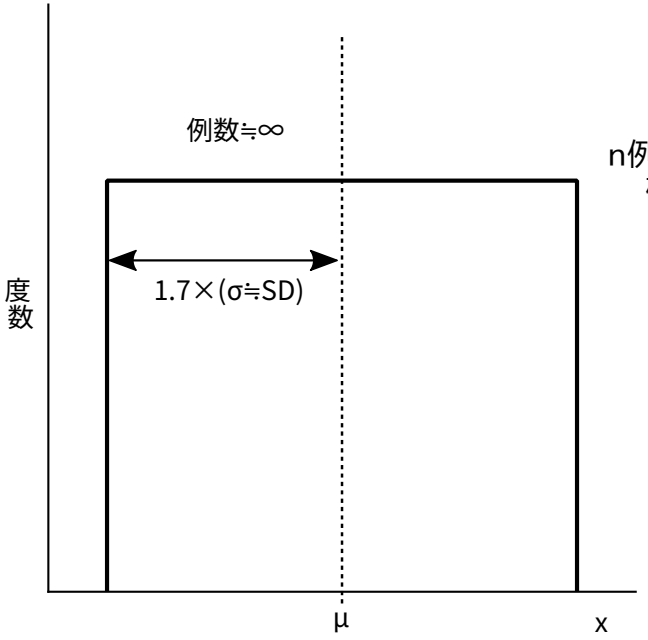


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

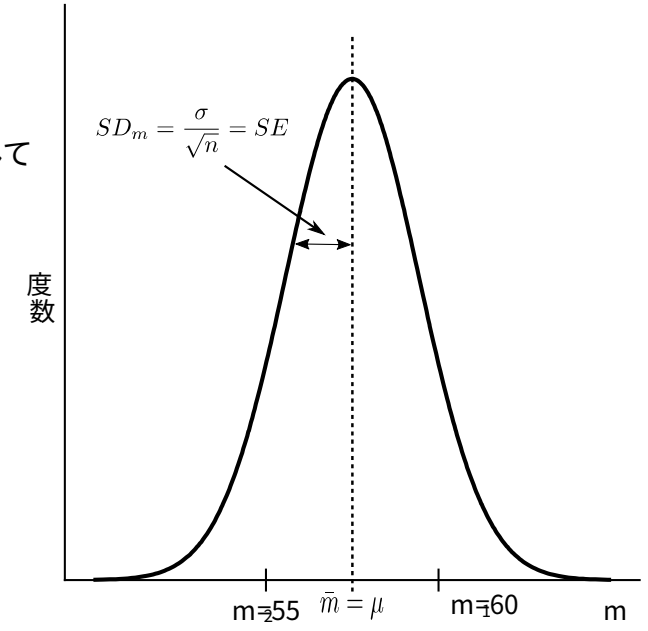


図1.4 標本平均値の分布

標準誤差の求め方

1.母集団からn例の標本集団を無作為抽出する

2.標本平均値を求め、それをm_1として標本平均値の度数分布図にプロットする

3.n例の標本集団を母集団に戻す

4.1番から3番を無限回繰り返す

5.最終的に図1.4のような標本平均値の度数分布ができあがる

標準誤差は標本平均値の誤差を表す要約値

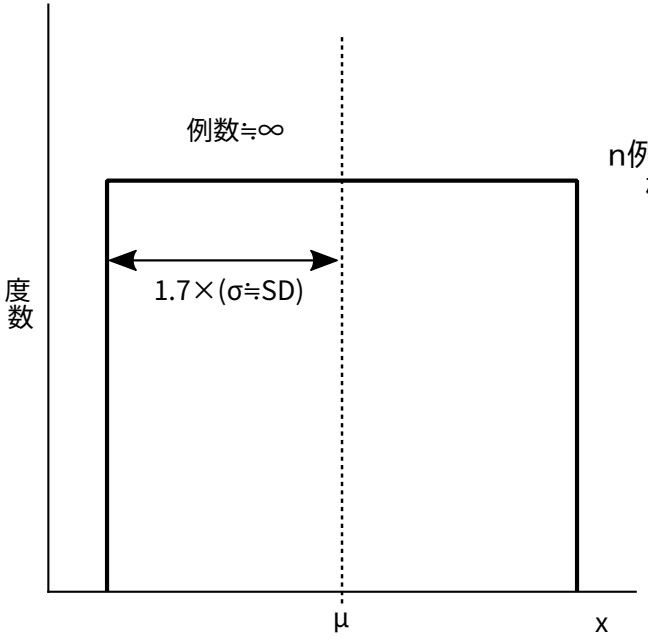


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

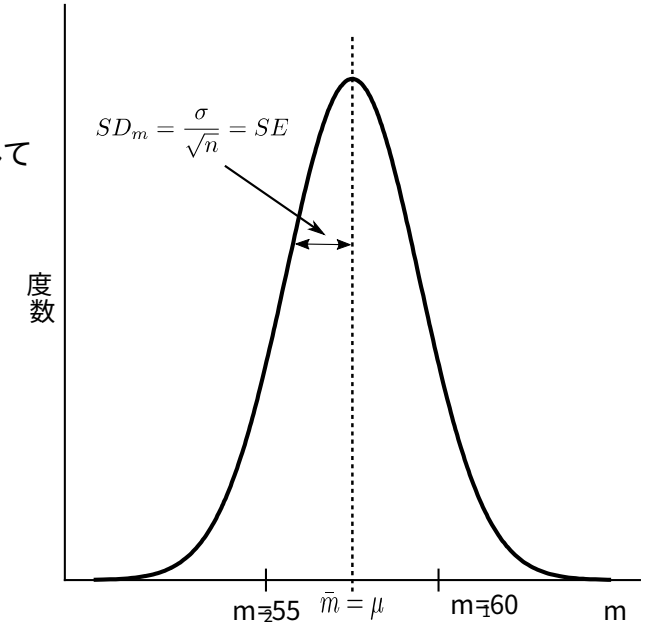


図1.4 標本平均値の分布

標本平均値の度数分布の特徴

- 母集団がどんな分布をしていても近似的に正規分布になる(nが多いほど近似が良い)
  → 中心極限定理(推測統計学の基本定理)
- 標本平均値の平均値は母平均値と一致する
- 標本平均値の標準偏差は次のような値になる → 標準誤差

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

標準誤差と標準偏差の使い分け

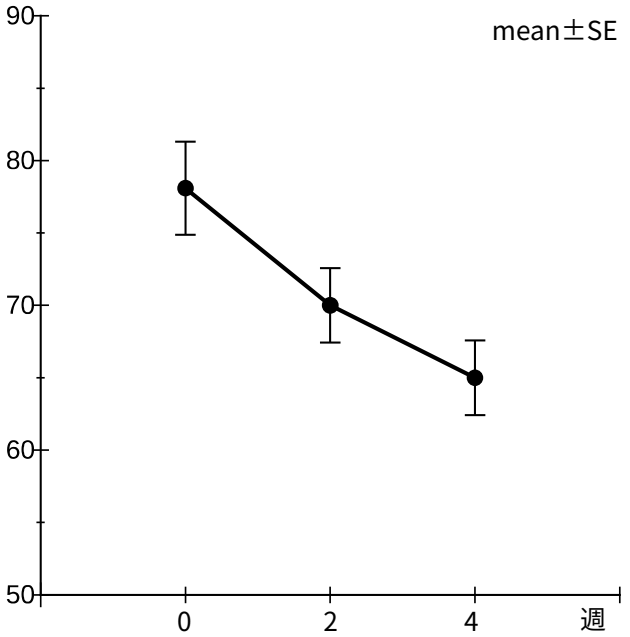


図1.5 体重の推移

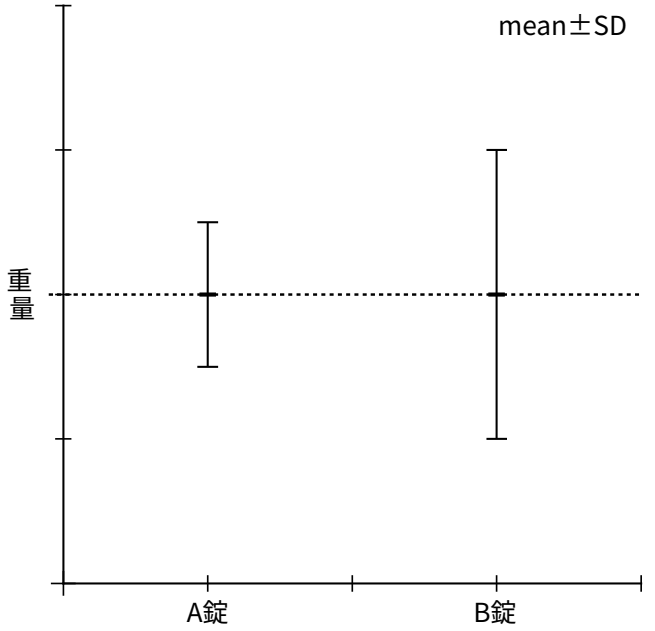


図1.6 錠剤の重量

母平均値の変化とその推測誤差を
表したい時は標準誤差

データのバラツキ具合を
表したい時は標準偏差

μ=m±SE: μをmで推測するとSE程度の推測誤差がある

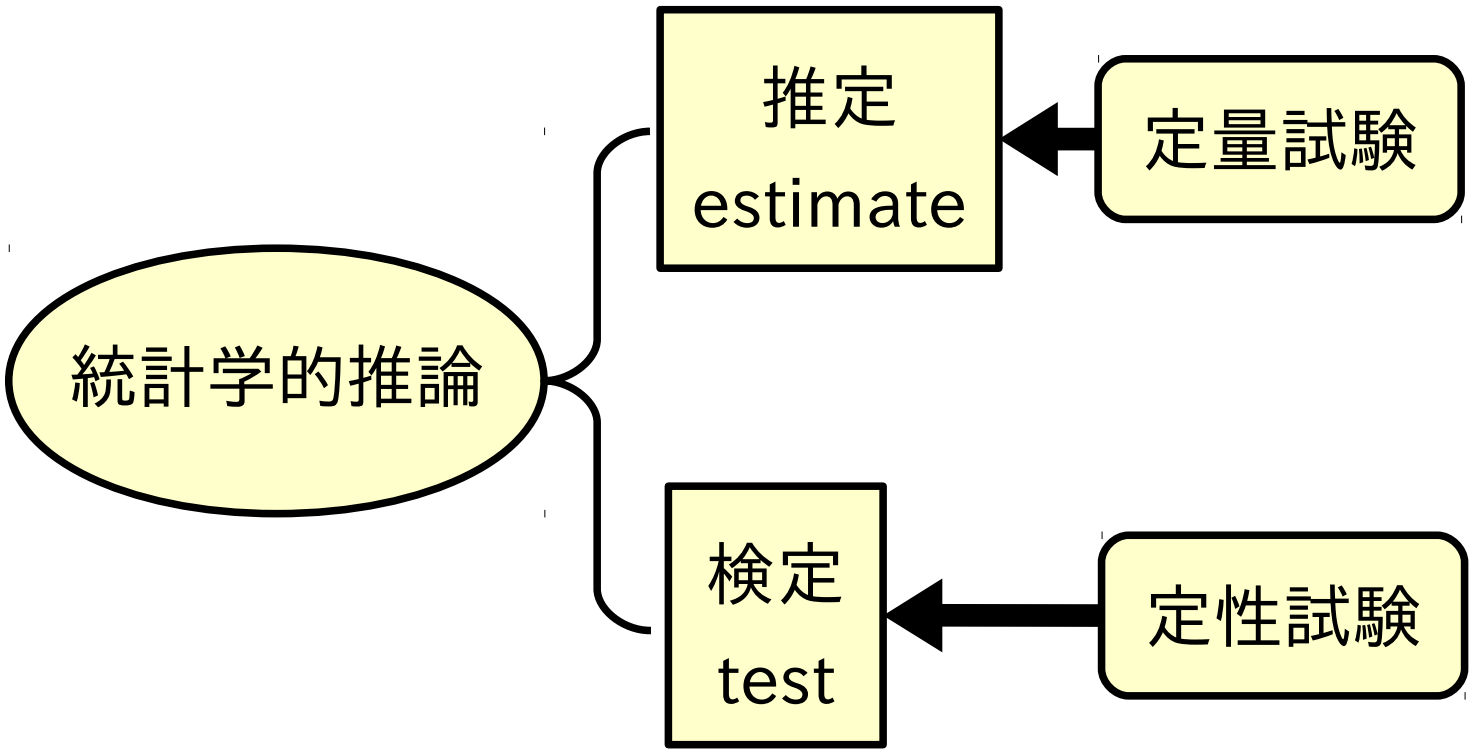
※m±2×SE(95%信頼区間)を描く方が合理的 ← リスク比やオッズ比はこれが普通

m±SD:データはmを中心にしてSD程度のバラツキがある

統計学の落とし穴

- 標準偏差と標準誤差
- 有意性検定と統計的仮説検定
- パラメトリック手法とノンパラメトリック手法
- ハンディキャップ方式の検定

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
〇×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

しかし研究現場や厚労省では検定が偏重されている

∵〇×式の方が採点が楽!

点推定はピンポイントの推測

母平均値：$\mu \approx m$：標本平均値

母標準偏差：$\sigma \approx SD = \sqrt{V} = \sqrt{\frac{SS}{n-1}}$：不偏標準偏差

点推定(point estimation)

母平均値を標本平均値で、母標準偏差を不偏標準偏差でそのまま推測

母標準偏差は点推定が普通

区間推定は幅を持たせた推測

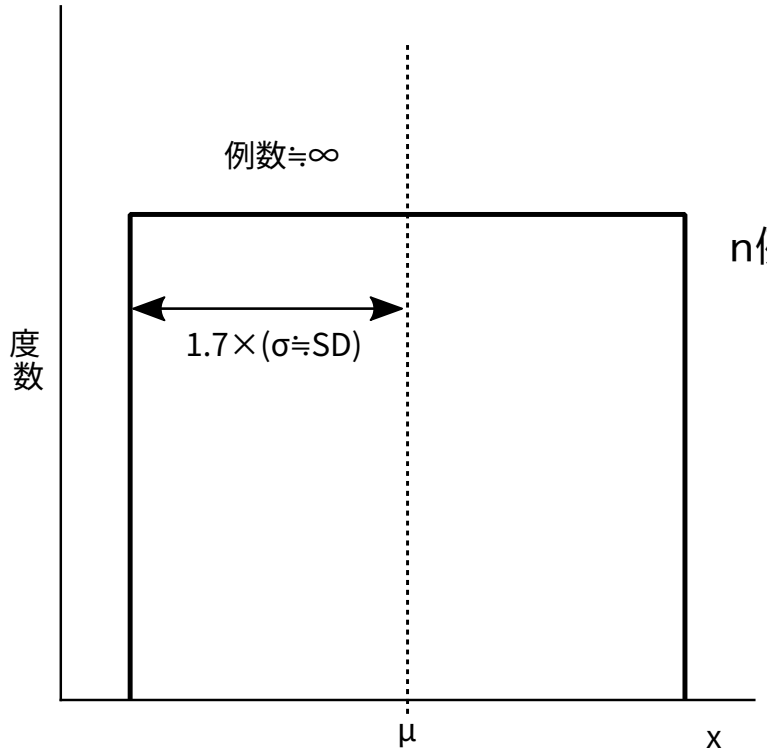


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

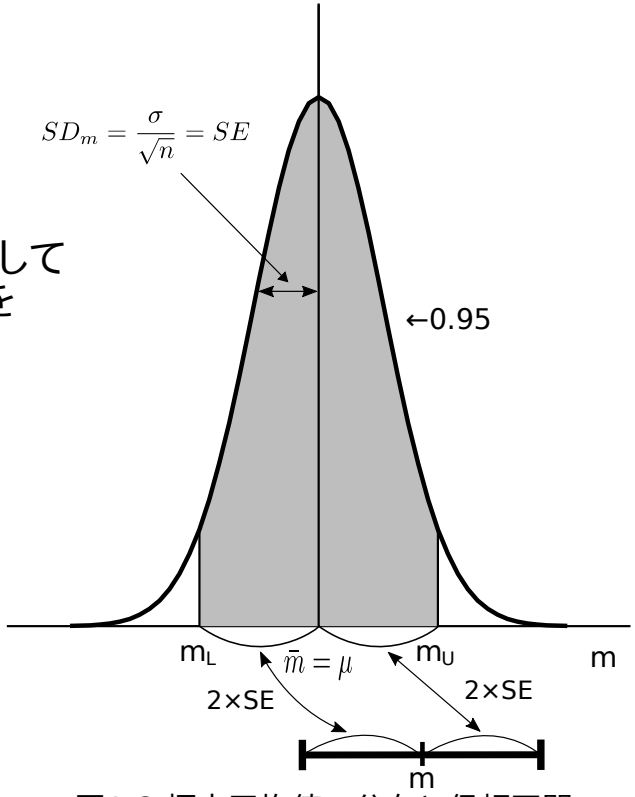


図1.8 標本平均値の分布と信頼区間

区間推定(interval estimation)

ある程度の幅を持たせて母数を推測する

母平均値は区間推定が普通

母平均値の区間推定法

標本平均値mの分布は近似的に正規分布になる

↓

mの平均値は母平均値 μ に、mの標準偏差は標準誤差SEになり
$\mu \pm 2 \times SE$ の範囲に約95%のmが含まれる

↓

ある標本平均値mが $\mu \pm 2 \times SE$ の範囲に含まれる確率は約95%

↓

逆に $m \pm 2 \times SE$ 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)：母平均値が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

点推定と区間推定

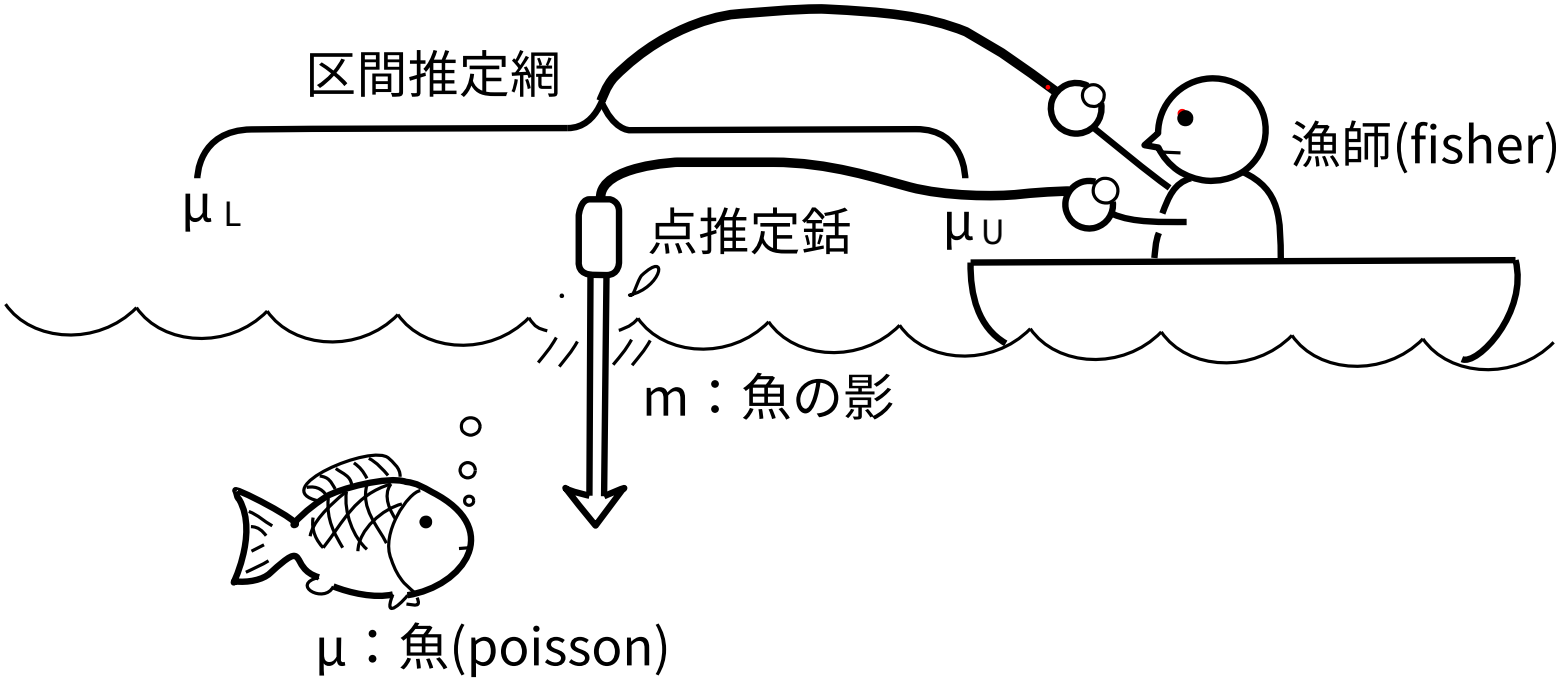


図1.9 点推定と区間推定

推定は漁師(Fisher)が
水面に映った魚(Poisson)の影mを見て魚μを捕まえるようなもの
点推定は銛で一突き、区間推定は投網を打つことに相当
普通は点推定を用い、重要な時だけ区間推定を行う

有意性検定は基準値を用いた〇×式の定性試験

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 ← 問題の答えは〇

対立仮説H_1：日本人の平均体重は50kgではない ← 問題の答えは×

有意性検定の手順

1. 問題を設定する → 医学的意義のある基準値μ_0を決める
2. 問題の答を〇×式で設定する → 帰無仮説H_0と対立仮説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

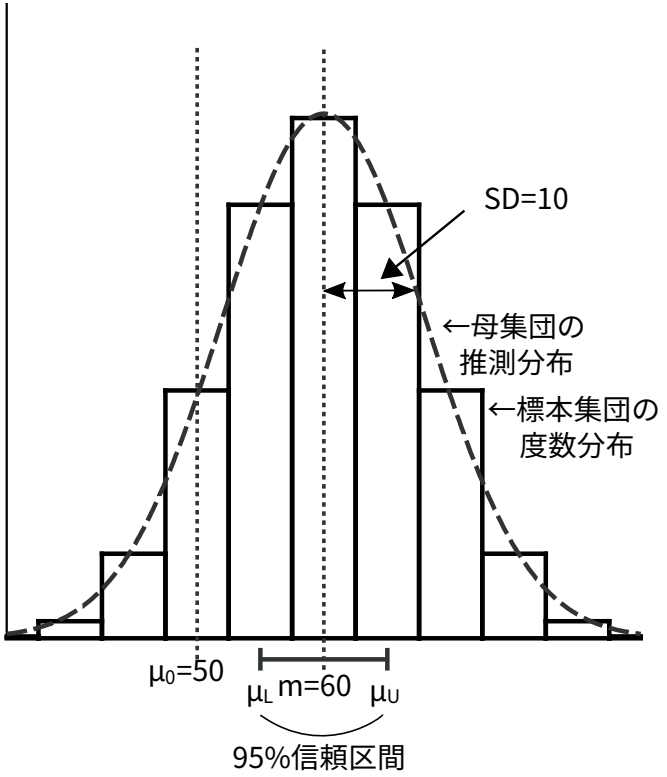


図1.10-a 信頼区間と有意性検定

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

母平均値は95%の確率で58～62kgの間にある

→ 母平均値は95%以上の確率で50kgではない

母集団から見た有意性検定

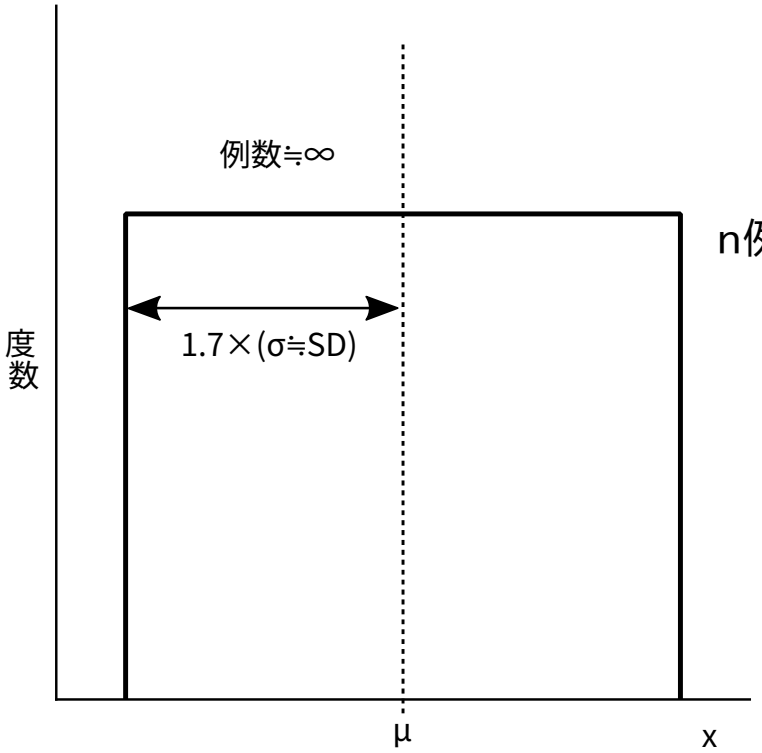


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

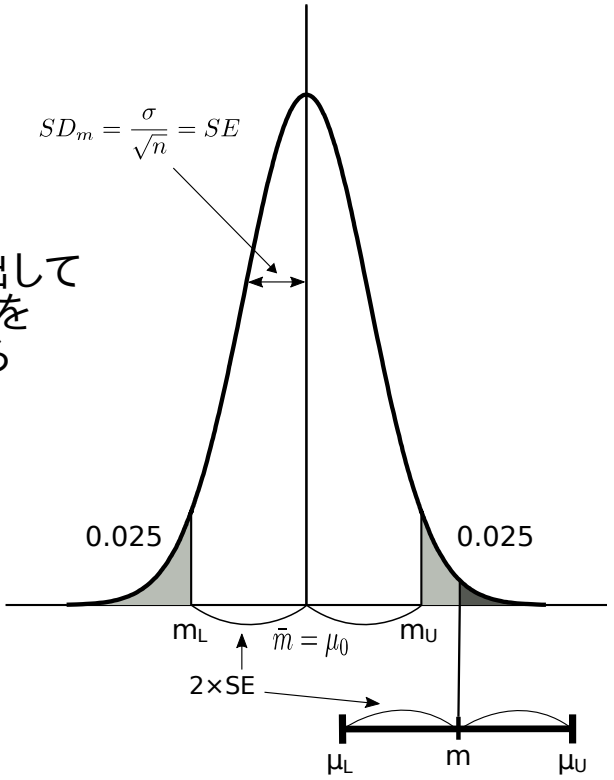


図1.5.3 標本平均値の分布と信頼区間

μ_0=50の時(帰無仮説が正しい時):標本平均値の95%が含まれる区間=50±2=48～52

m_L=48:下側棄却域の上限　　m_U=52:上側棄却域の下限＜60:標本平均値

標本平均値が棄却域に入っている=95%信頼区間に基準値μ_0が入っていない

=標本平均値mから右側の面積(図1.5.3の濃い灰色の部分)×2=p値(有意確率)が0.05以下

→ 母平均値は95%以上の確率で50kgではない

有意性検定は対立仮説だけを採用する

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

→ 統計学的結論が間違っている確率が5%程度ある

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

統計学的結論:日本人の平均体重は50kgではないと断定できない ← 問題の答えは保留

「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

これは帰無仮説「日本人の平均体重は50kgである」の採用ではないことに注意!

※不確かなデータから得られた結果を解釈する時は確定的なことを断言する方が「非科学的」

※得られたデータから結論できる限界を明確にするのが「科学的」

「有意差あり」は「実質科学的に差がある」という意味ではない

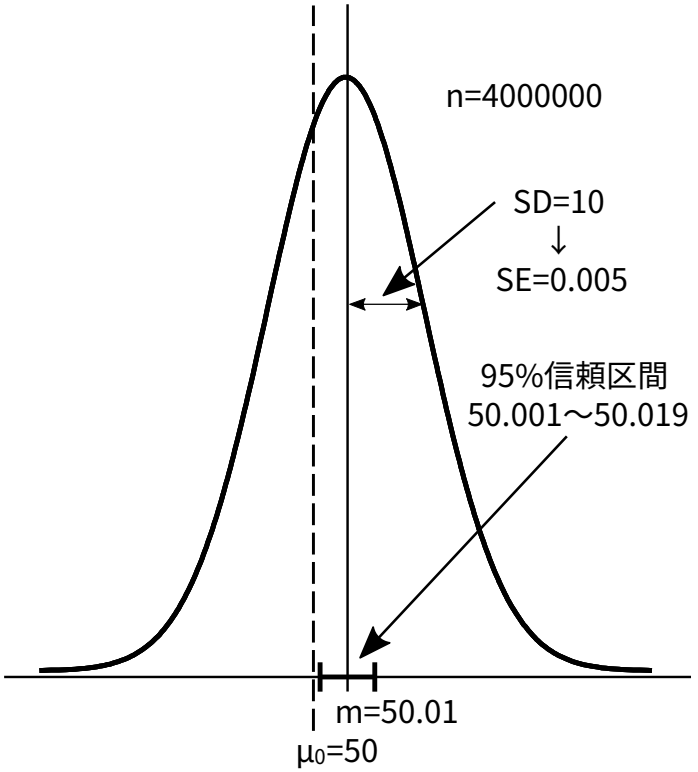


図1.11 有意でも実質科学的には
無意味な差

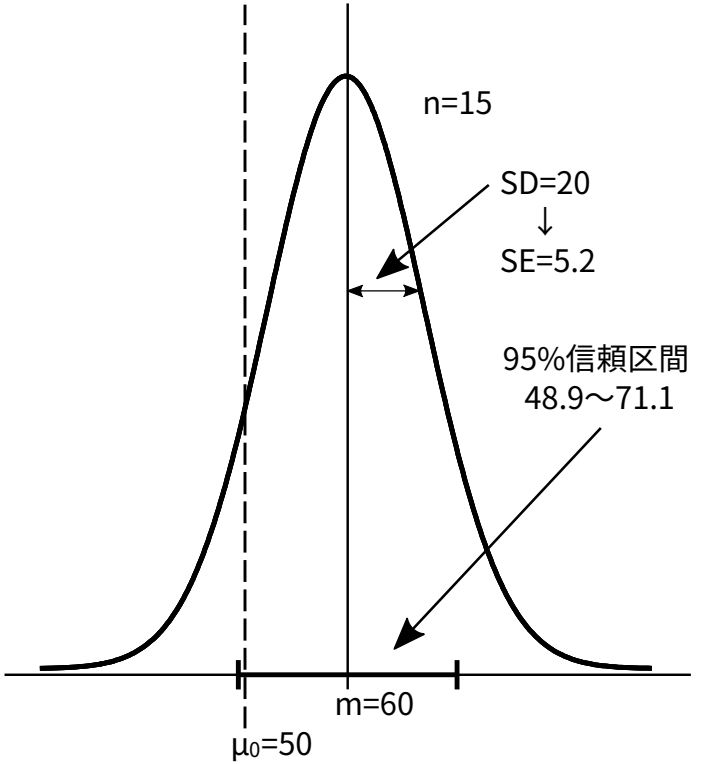


図1.12 実質科学的に意味があっても
有意ではない差

有意：数学的に意味が有る → 統計学的結論の信頼性が高い

有意ではない：数学的に意味が無い

→ 統計学的結論の信頼性が低い → 統計学的結論を保留する

※母平均値が基準値とぴったり一致することは現実には有り得ない

→ わざわざ検定する必要はない！ → 検定廃止論

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

統計的仮説検定は有意性検定の欠点を是正した手法

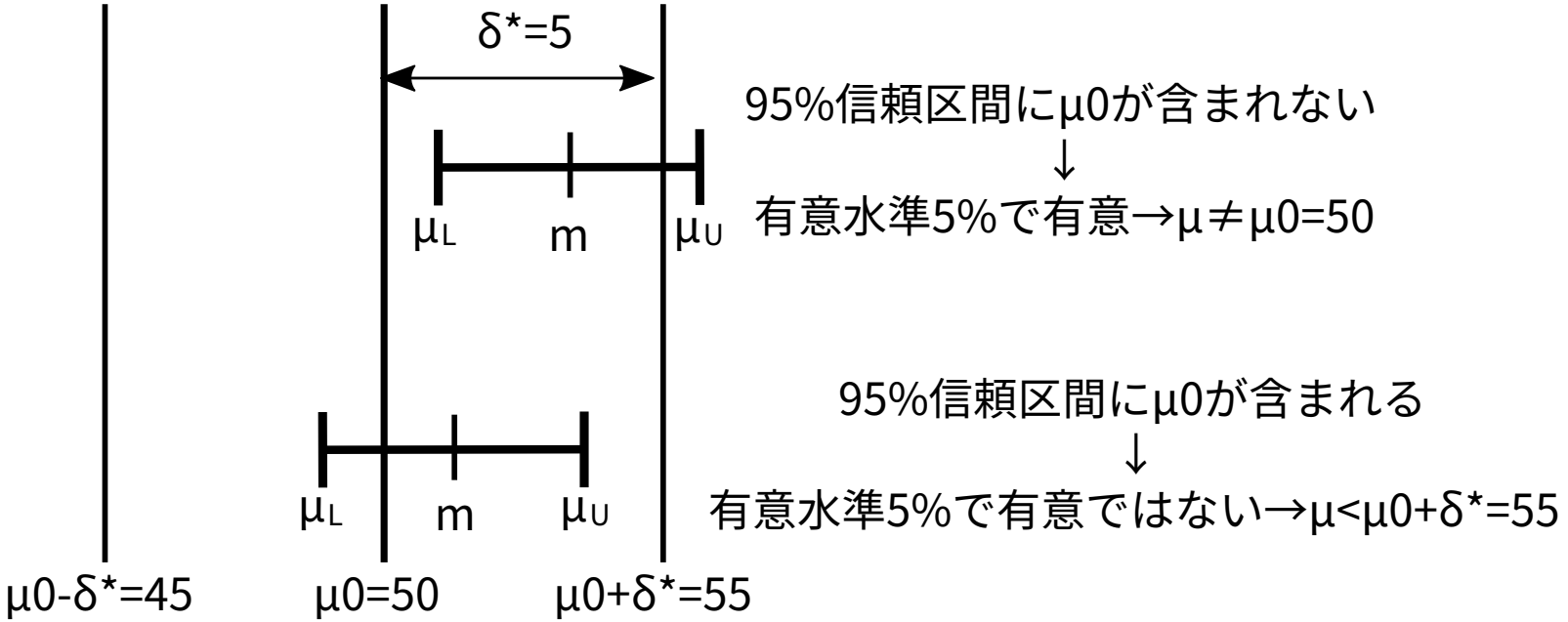


図1.13 信頼区間と統計的仮説検定

δ*=±5kg:(最小)検出差(scientific significant difference)≒医学的な許容範囲

95%信頼区間：$\mu = 51 \pm 2 \times \frac{10}{\sqrt{100}} = 51 \pm 2 \rightarrow \mu = 49 \sim 53$

母平均値は95%の確率で49～53kgの間にある

→ 母平均値は95%以上の確率で45kgよりも重く55kgよりも軽い

統計的仮説検定は基準値と許容範囲を用いた〇×式の定性試験

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 ← 問題の答えは〇

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と許容範囲δ^*を決める

2.問題の答を〇×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

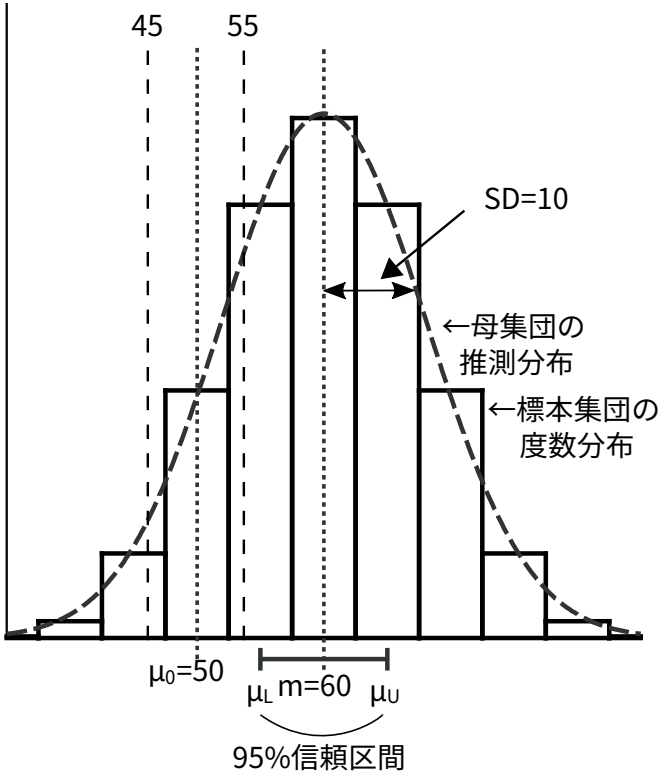


図1.10-b 信頼区間と統計的仮説検定

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

母平均値は95%の確率で58～62kgの間にある

→ 母平均値は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

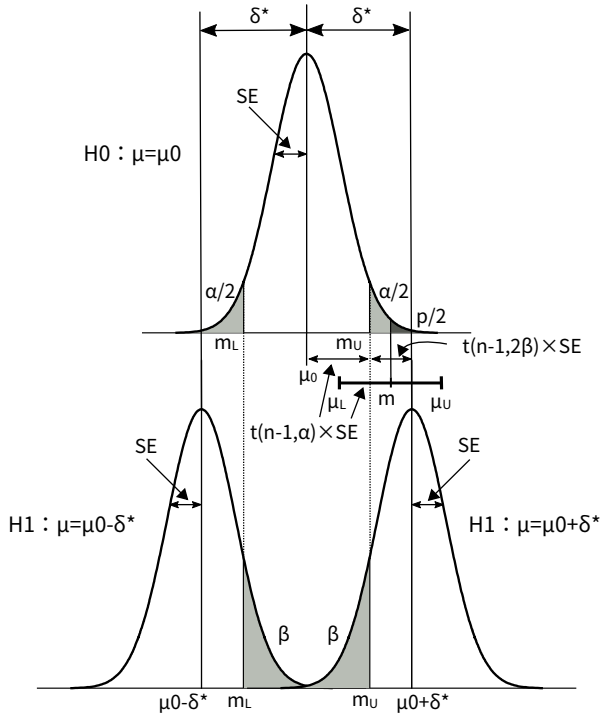


図1.14 統計的仮説検定の模式図

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

μ_0-δ*=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μ_0+δ*=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%

統計的仮説検定は有意ではない時も結論を採用する

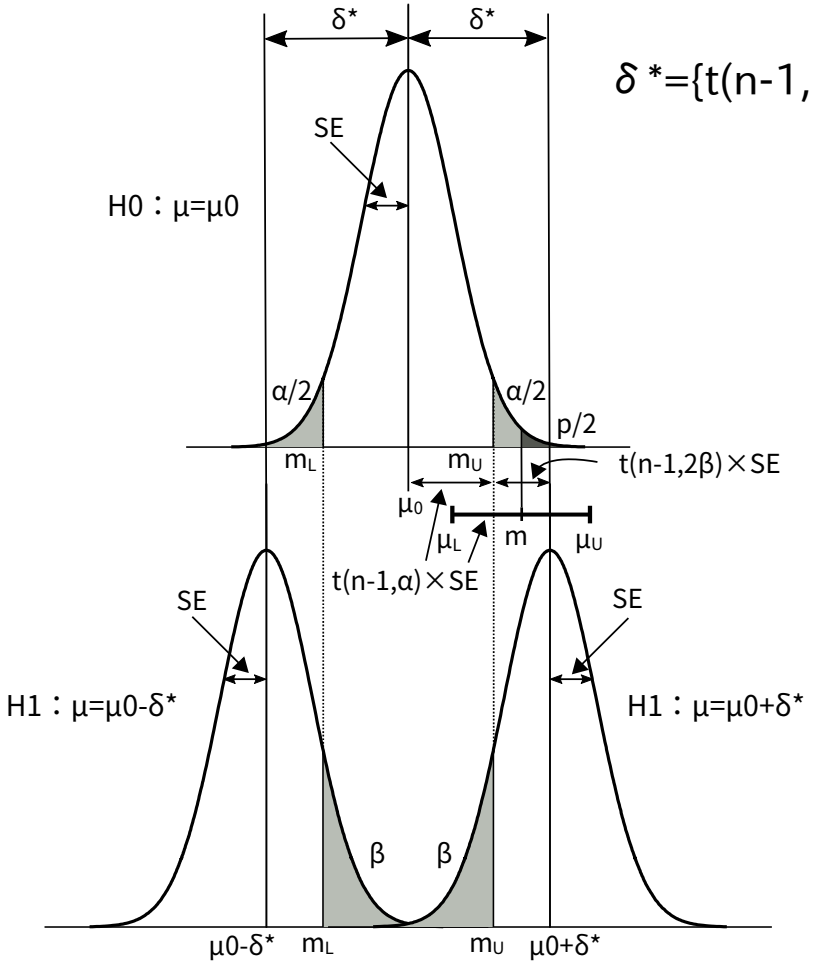
棄却域に標本平均値が入っている時＝95%信頼区間に基準値が入っていない時
統計学的結論:日本人の平均体重は50kgではない ← 問題の答えは×

「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



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

図1.14 統計的仮説検定の模式図

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

信頼区間を許容範囲よりも小さくする必要がある

検定結果だけから実質科学的な判断をするのは危険

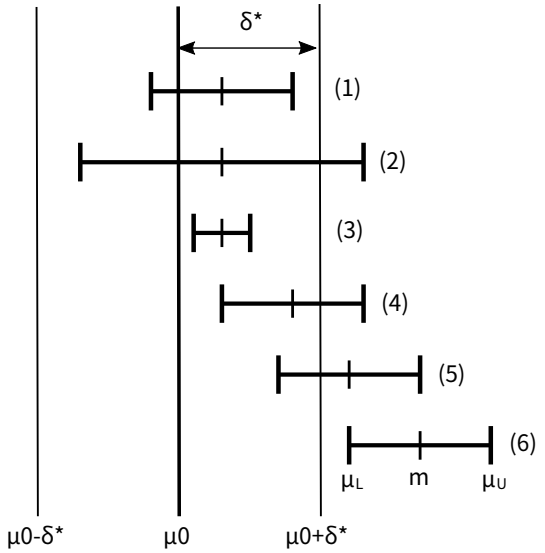


図1.15 検定結果と信頼区間

	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

# p<0.001になっても結果の信頼性は95%

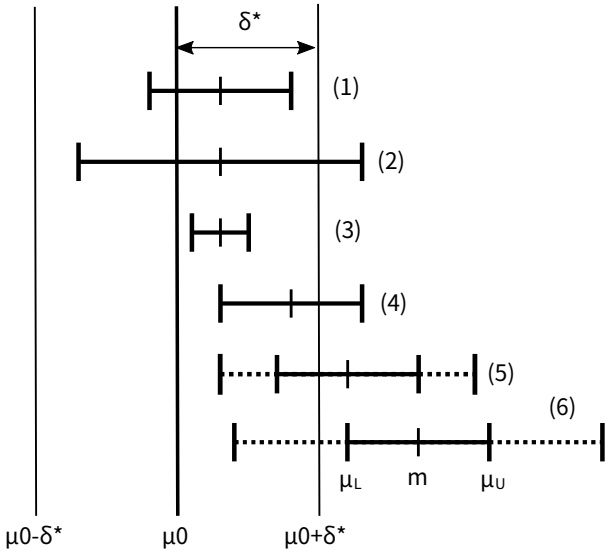


図1.20 検定結果と信頼係数を変えた信頼区間

(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

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

統計学の落とし穴

- 標準偏差と標準誤差
- 有意性検定と統計的仮説検定
- パラメトリック手法とノンパラメトリック手法
- ハンディキャップ方式の検定

パラメトリック手法とノンパラメトリック手法の特徴

統計手法

パラメトリック手法
母数に依存した手法
→ 数学的モデルを利用

要約値：平均値、標準偏差
検定・推定：平均値の検定(t検定)

- 母集団のデータが特定の分布に従うと仮定する
- 要約値がデータの分布状態を反映する
  → データの分布状態によって要約値の値が変化する
- 結果の精度が高い
- 結果の普遍化が容易

ノンパラメトリック手法
母数に依存しない手法
→ 数学的モデル利用せず

要約値：中央値、順位平均値、割合
検定・推定：順位和検定、出現率の検定(χ^2 検定)

- 母集団のデータがどんな分布をしていても良い
- 要約値がデータの分布状態を反映しにくい
  → データの分布状態によって要約値の値が変化しにくい
- 結果の精度が低い
- 結果の普遍化は困難

データの分布状態で適用する手法を選択してはいけない

データが正規分布しない時はノンパラメトリック手法を適用せよ

※データが正規分布しない時はパラメトリック手法が使えないのではなく

検定と推定の効率がノンパラメトリック手法よりも悪くなる時がある

要約値に関する科学的な考察を無視した乱暴な主張だから鵜呑みにしてはいけない

※腹痛で胃腸薬を求めたら「今は効果の弱い胃腸薬しかないので

代わりに効果の強い降圧剤を使った方が良い」と助言されるようなもの

統計手法は要約値の数学的な信頼性を評価するためのもの

統計手法と要約値は1対1で対応している

統計手法を決定する最も重要な要因は要約値の科学的な妥当性

標本平均値の分布

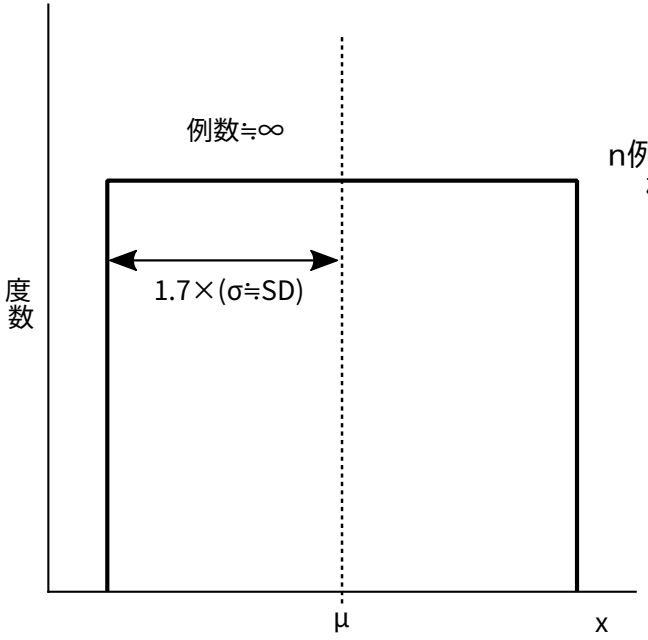


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

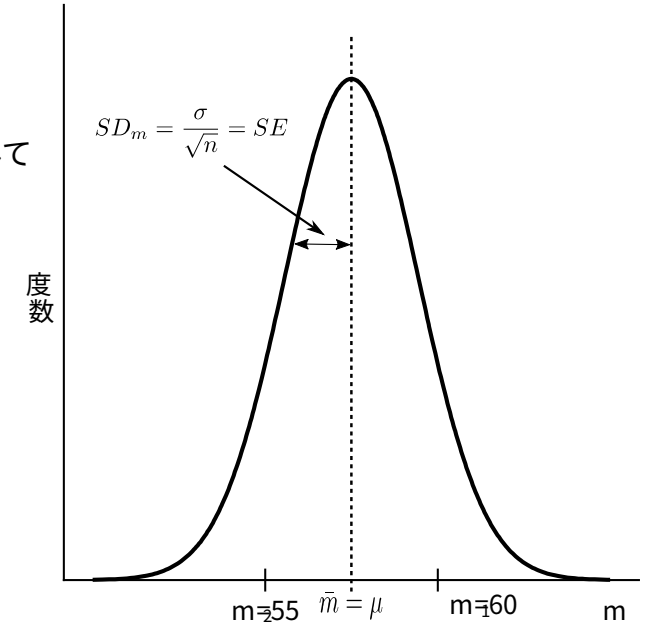


図1.4 標本平均値の分布

標本平均値の度数分布の特徴

- 母集団がどんな分布をしていても近似的に正規分布になる(nが多いほど近似が良い)
  → 中心極限定理(推測統計学の基本定理)
- 標本平均値の平均値は母平均値と一致する
- 標本平均値の標準偏差は次のような値になる → 標準誤差

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

順位平均値の分布

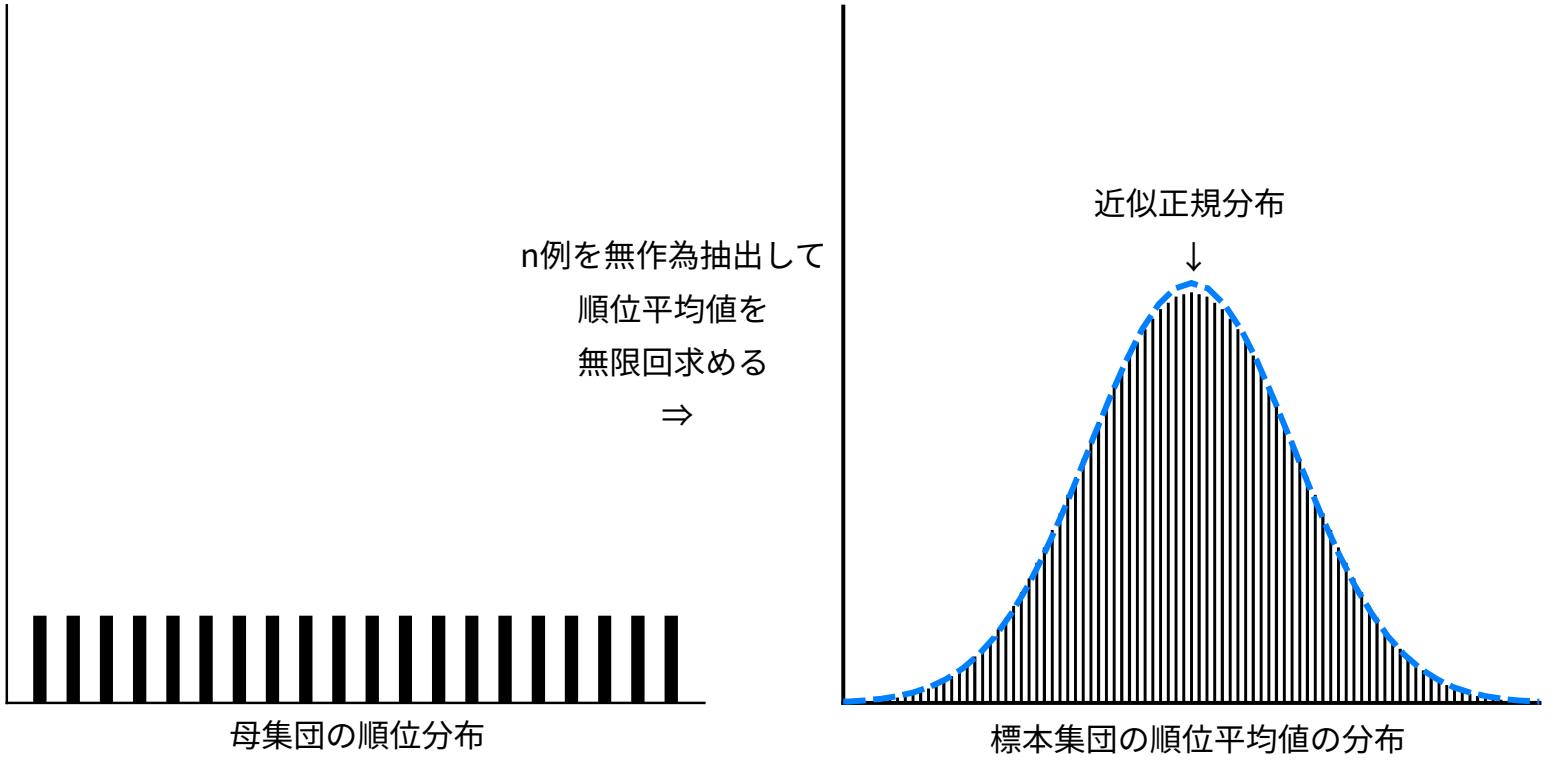


図4.10 順位分布と順位平均値の分布

- 図4.10の左のグラフ:同位のデータがない時の母集団の順位分布
- 図4.10の右のグラフ:順位平均値の理論分布
- 順位平均値の分布は中心極限定理によって近似的に正規分布になる

※パラメトリック手法もノンパラメトリック手法も要約値の近似正規分布を利用している!

外れ値がある時のデータの分布と順位の分布

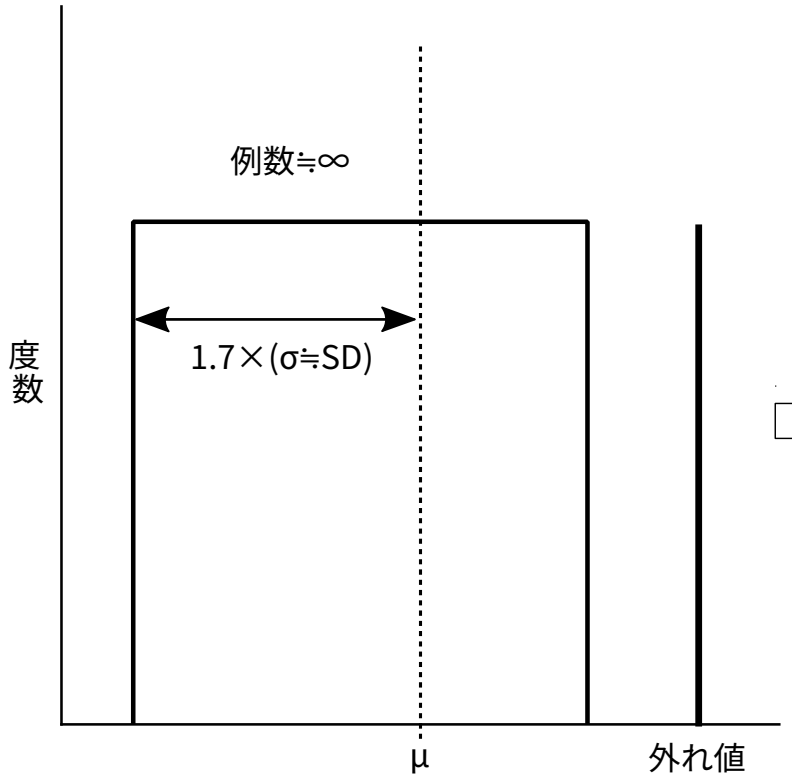


図1.18 外れ値がある母集団

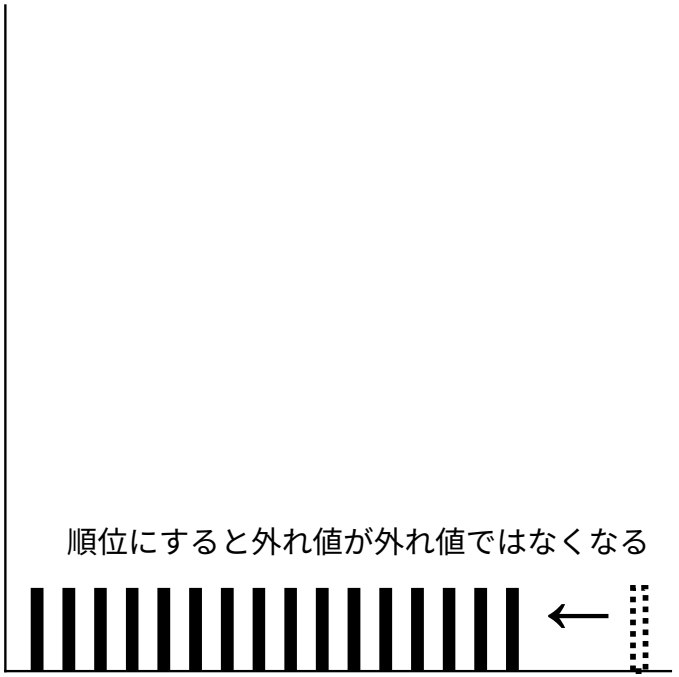


図1.19 外れ値がある母集団の順位分布

外れ値がある時

データの分布の標準偏差は大きくなるが順位分布の標準偏差は不変

→ 標本平均値の標準誤差は大きくなるが順位平均値の標準誤差は不変

∴順位平均値よりも平均値の方が推定と検定の効率が悪くなる時がある!

中央値の場合

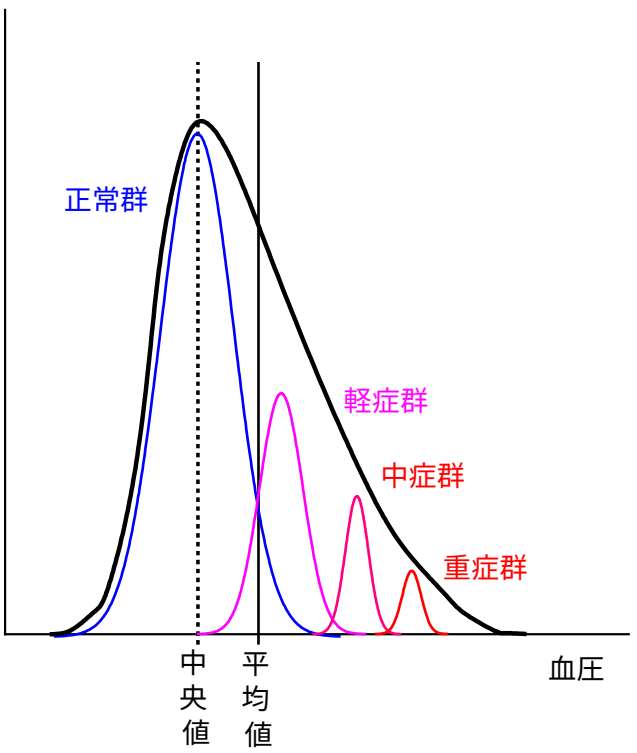


図3.1 投与前の血圧分布

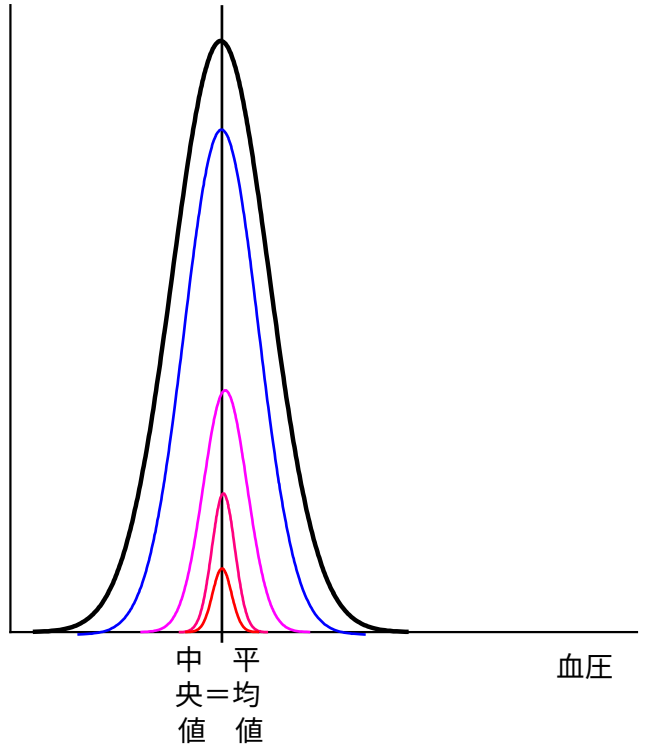


図3.2 投与後の血圧分布

図3.1の左側の集団に降圧剤を投与したところ、高血圧患者だけ血圧が低下して図3.1の右側のようになった

→ 大多数の正常者は血圧が不変のため平均値は低下、中央値は不変

降圧剤の効果があったと評価するべきか、なかったと評価するべきか？

→ 医学的に妥当な評価指標は平均値か中央値か？

順位和検定の場合

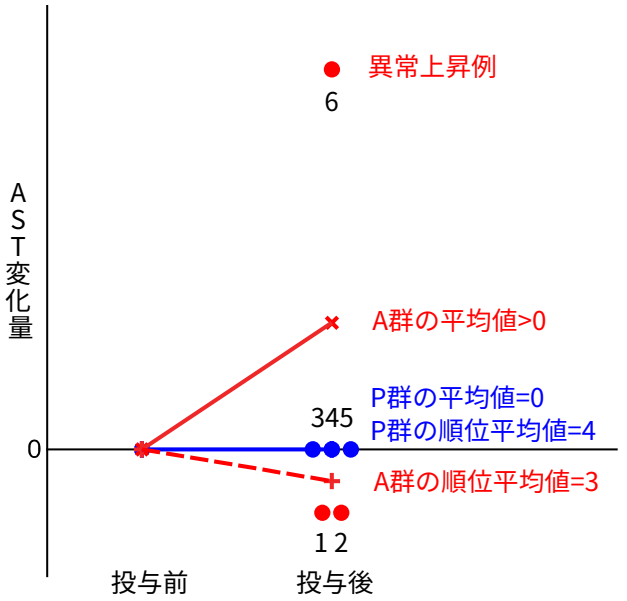


図3.3 平均値と順位平均値が矛盾する例

A群は副作用によるAST異常上昇例が発生し、それ以外はわずかに低下、P群は全例不変
変化量の平均値：A群は上昇、P群は不変 → A群はP群よりも上昇
変化量の順位平均値：A群は低下、P群は不変 → A群はP群よりも低下

副作用があったと評価するべきか、なかったと評価するべきか？
→ 医学的に妥当な評価指標は平均値か順位平均値か？
※医学分野では疾患による異常値や副作用による異常値が必然的に発生する

※順位和検定はデータの実測値には科学的意義がなく、順位だけが科学的意義を持つ時に適用すべき手法

順序尺度のデータを計量尺度扱いした方が良い例

＜薬剤の効果判定－1＞

群	著明改善	改善	不変	悪化	著明悪化	計
薬剤1投与群	0	40	40	0	0	80
薬剤2投与群	40	0	0	40	0	80
計	40	40	40	40	0	160

・Mann-WhitneyのU検定(Wilcoxonの2標本検定)：正規分布 z_o=0　有意確率 p=1

・著明改善～著明悪化を1～5と数量化して計量尺度扱いした時(リッカート尺度)

薬剤1投与群の平均値=2.5　薬剤2投与群の平均値=2.5　2標本t検定：t_o=0　p=1

＜薬剤の効果判定－2＞

群	著明改善	改善	不変	悪化	著明悪化	計
薬剤1投与群	0	40	40	0	0	80
薬剤2投与群	40	0	0	0	40	80
計	40	40	40	0	40	160

・Mann-WhitneyのU検定(Wilcoxonの2標本検定)：正規分布 z_o=0　有意確率 p=1

・著明改善～著明悪化を1～5と数量化して計量尺度扱いした時(リッカート尺度)

薬剤1投与群の平均値=2.5　薬剤2投与群の平均値=3.0　2標本t検定：t_o=2　p=0.0338 *

医学的に見ると計量尺度扱いした方が妥当な結果

統計学の落とし穴

- 標準偏差と標準誤差
- 有意性検定と統計的仮説検定
- パラメトリック手法とノンパラメトリック手法
- ハンディキャップ方式の検定

ハンディキャップ方式の検定の模式図

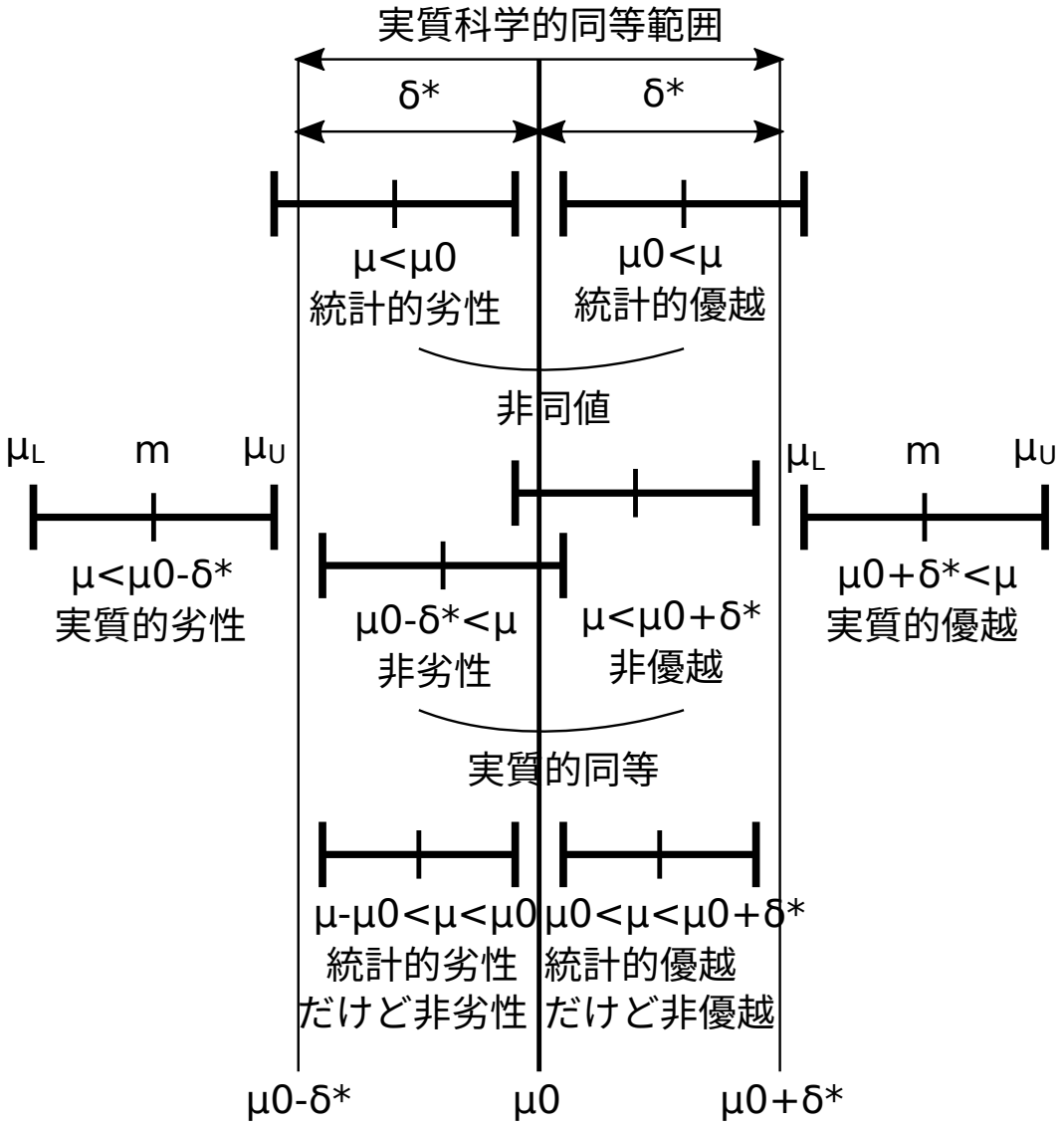


図3.13 ハンディキャップ方式の検定

ハンディキャップ方式の検定の名称

- 非同値検定：μ_0を基準値にした有意性検定

  $\mu_0 < \mu_L$(95%CIの下限) → 有意：統計的優越＝非同値

  μ_U(95%CIの上限)$< \mu_0$ → 有意：統計的劣性＝非同値

- 同等性検定：μ_0を基準値にし95%CIの幅をδ^*以下にした統計的仮説検定

  $\mu_0 < \mu_L$(95%CIの下限) → 有意：統計的優越＝非同値

  μ_U(95%CIの上限)$< \mu_0$ → 有意：統計的劣性＝非同値

  $\mu_L < \mu_0 < \mu_U$(95%CI内にμ_0が含まれる) → 有意ではない：実質的同等

- 優越性検定または非優越性検定：$\mu_0 + \delta^*$を基準値にした有意性検定

  $\mu_0 + \delta^* < \mu_L$(95%CIの下限) → 有意：実質的優越

  μ_U(95%CIの上限)$< \mu_0 + \delta^*$ → 有意：実質的非優越

- 非劣性検定または劣性検定：$\mu_0 - \delta^*$を基準値にした有意性検定

  $\mu_0 - \delta^* < \mu_L$(95%CIの下限) → 有意：実質的非劣性

  μ_U(95%CIの上限)$< \mu_0 - \delta^*$ → 有意：実質的劣性

非劣性検定は優越性検定とペアで使用しなければならない

非劣性検定を同等性検定の代わりに使用してはいけないし

非同値検定を優越性検定の代わりに使用してもいけない

標準薬に対する非劣性検定の非合理性

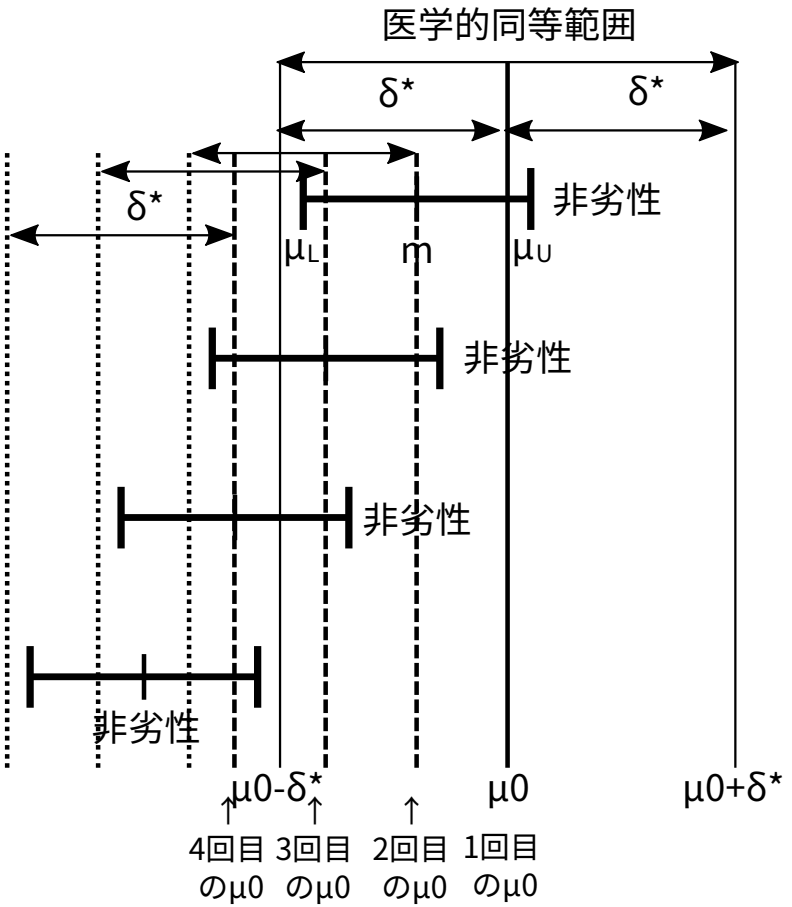


図3.14 標準薬に対する非劣性検定の非合理性

同等性検定の代わりに非劣性検定を行い、非同値検定を優越性検定の代わりに使用するのは欺瞞
→ 新薬が標準薬と実質的に同等でも「新薬にはメリットがある」という結論が導けてしまう
→ 数回後には最初の標準薬と比較すると実質的に劣性な新薬が許可されてしまう

本日の結語

ノンパラメトリック手法と

非劣性検定はできるだけ使わず

推定結果を重視しましょう!

ご清聴ありがとうございました