NCNP-TMC 第5回CRT実践講座ワークショップ

臨床試験における サンプルサイズ計算と解析

東京大学大学院医学系研究科 公共健康医学専攻 生物統計学分野
国立精神・神経医療研究センター 情報管理・解析部 生物統計アドバイザー
柏原　康佑

2016.1.22

連続量データ

- 解析方法の種類は数えきれない
- 本日は最も基本的な連続量データの解析（検定）と症例数設計に焦点
  - 根本的な考え方はすべてのエンドポイントで共通
    - 2値データ、生存時間、…
- 本講義の目的
  - 連続データに限らず「統計学の考え方」

構成

- 検定と信頼区間のしくみ
- 症例数設計

臨床試験の例

- 例:
  大うつ病患者に対するSSRIの効果が知りたい

Patient	大うつ病患者
Intervention	SSRI（レクサプロ）
Comparison	プラセボ
Outcome (+ Time)	投与8週後のMADRS変化量

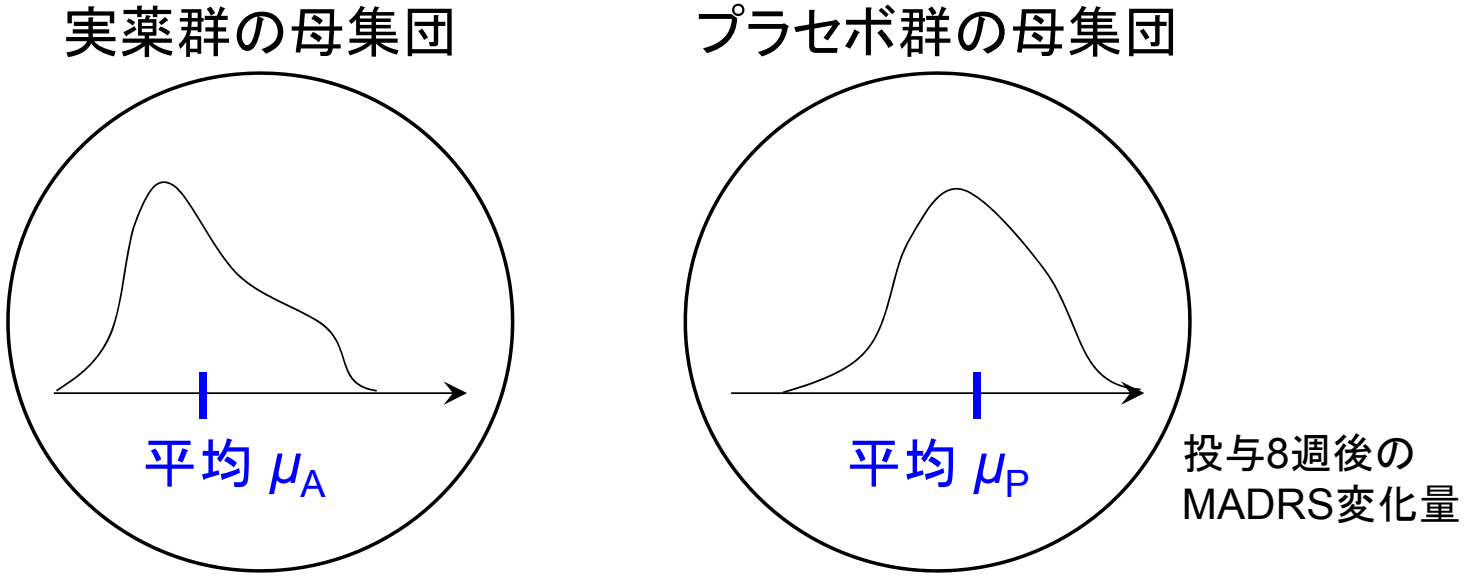
SSRI: 選択的セロトニン再取り込み阻害薬

レクサプロ第3相臨床試験

- 目的
  - 大うつ病性障害患者に対するエスシタロプラムのプラセボに対する優越性を検証
- 対象
  - DSM-IV-TRによる主診断が大うつ病性障害
  - DSM-IV-TRの大うつ病エピソードが4週以上継続
  - MADRS合計点が22点以上
    - MADRS:モントゴメリー/アスベルグうつ病評価尺度
  - CGI-Sの評価が4点以上
- エンドポイント
  - 8週投与後の投与開始時からのMADRS変化量

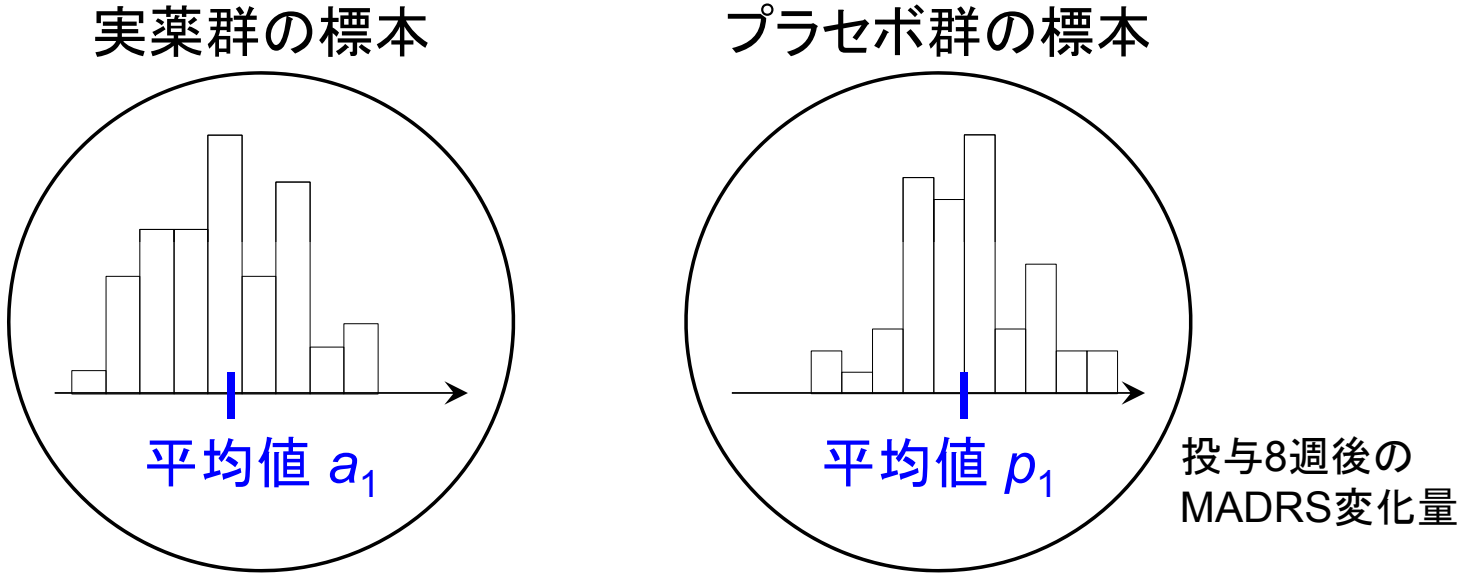
http://www.pmda.go.jp/drugs/2011/P201100076/index.html

最も観察したい値は何か?



- 最も観察したい値
  - 研究目的を定式化したもの
  - これが明確に分かるように PICO + T を記載すべき
- レクサプロ試験: 母集団間の平均値差 $\delta = \mu_A - \mu_P$

実際に観察できるのは?



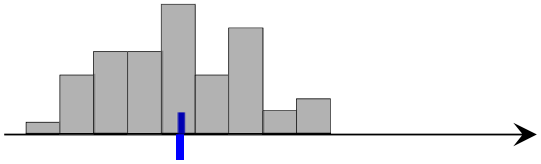
- 母集団からのランダムサンプリング
- レクサプロ試験では群間差 $d_1 = a_1 - p_1$
  - ここから $\delta = \mu_A - \mu_P$ に関する推測を行いたい

データ解析における2つの目的

- 記述統計学 descriptive statistics
  - 目的: 得られたデータ(標本)の特徴を提示すること
  - 方法: グラフ、要約統計量
  - 母集団に対する言及は行わない
  - 難解な確率の理論は必要ない
- 推測統計学 inferential statistics /
  統計的推測 statistical inference
  - 目的: 標本から母集団の特徴を推測すること
  - 方法: 検定、信頼区間
  - こちらが臨床研究の目的であることが多い
  - 統計学の本領

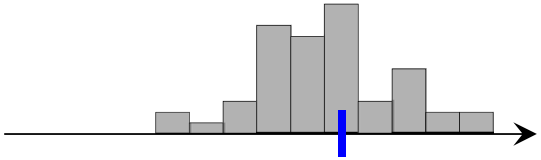
検定 test

実薬群



平均 a_1

プラセボ群



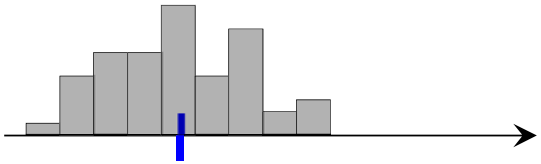
平均 p_1

- 実薬は有効か?
  - 二者択一の意思決定
  - 定性的
- (未知の)真実
  - 母集団の平均値の差 < 0 ⇒ 有効
  - 母集団の平均値の差 ≥ 0 ⇒ 無効
- (観察可能な)事実
  - 群間差 < 0 ⇒ 有効?
  - 群間差 ≥ 0 ⇒ 無効?

群間差は信頼できる?

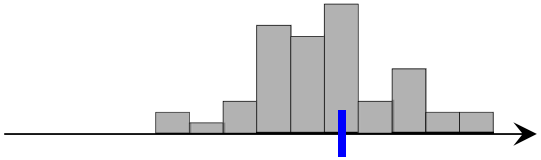
推定 estimation

実薬群



平均 a_1

プラセボ群

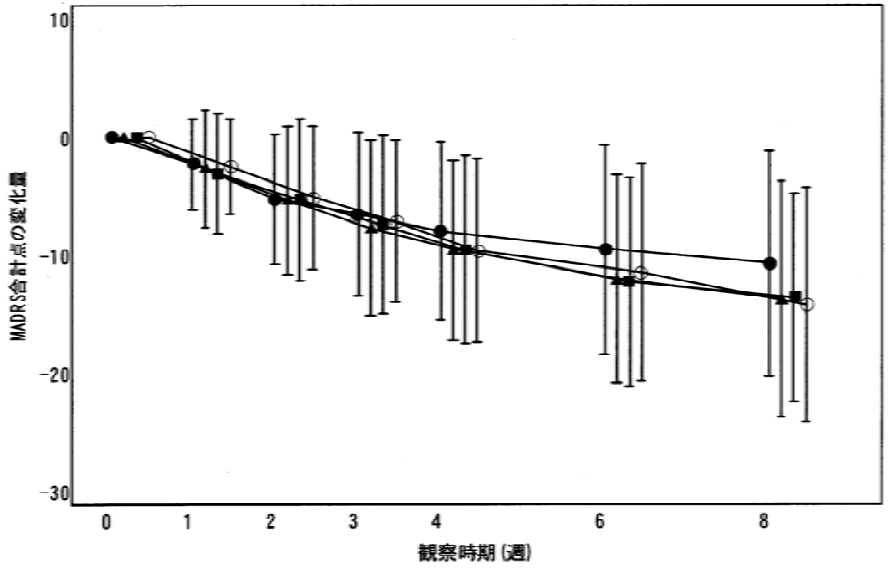


平均 p_1

- 実薬の効果の大きさは?
  - 定量的
- (未知の)真実
  - 母集団の平均値の差: $\delta = \mu_A - \mu_P$
- (観察可能な)事実
  - 群間差: $d_1 = a_1 - p_1$
    $\Rightarrow$ 群間差 = 効果の大きさ?

群間差は信頼できる?

試験の結果



●●● プラセボ群（124 例）
■■■ MLD-55 20 mg 群（119 例）
▲▲▲ MLD-55 10 mg 群（120 例）
○○○ PAR 群（121 例）
（平均値±標準偏差）

- 8週時MADRS変化量
  - 群間差: 3
  - 標準偏差: 10
  - p = 0.006

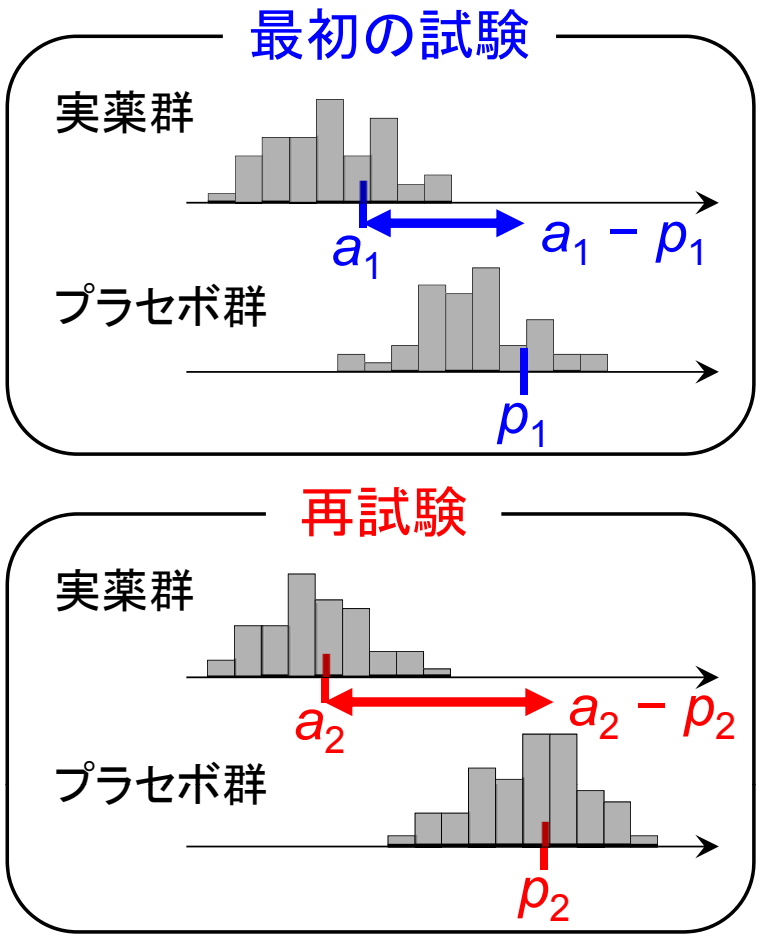
「理想的な結果」

http://www.pmda.go.jp/drugs/2011/P201100076/index.html

得られた群間差は偶然では?

- 群間差に比べてデータのバラツキは非常に大きい
- 同一プロトコルの試験を
  - もう一度実施したら?(再試験)
  - 他の研究チームが実施したら?(追試験)
- 結果は再現されるのか?

2種類のバラツキ



- データのバラツキ
  - e.g. 患者間のバラツキ
  - 標準偏差 standard deviation; SD
    - 記述統計学の範疇
- 統計量のバラツキ
  - e.g. 試験間のバラツキ
    - 実薬群の平均 $a_1 \neq a_2$
    - プラセボ群の平均 $p_1 \neq p_2$
    - 群間差 $a_1 - p_1 \neq a_2 - p_2$
  - データのバラツキに起因
  - 標準誤差 standard error; SE

検定統計量 test statistic

- 得られた群間差 ／ 得られた標準誤差
  - d_1 / (群間差のSE)
  - （シグナル ／ ノイズ）
- 個々の臨床試験の状況に依存しない便利な統計量
  - 例えば、MADRSとHAM-Dでは標準偏差が異なる
  - しかし、いずれに対しても *t* 検定統計量 を適用可能

t 検定統計量

- 得られた群間差 ／ 標準誤差の推定値

$$t = \frac{d_1}{d_1\text{のSEの推定値}} = \frac{a_1 - p_1}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_P}\right)\hat{\sigma}_C^2}}$$

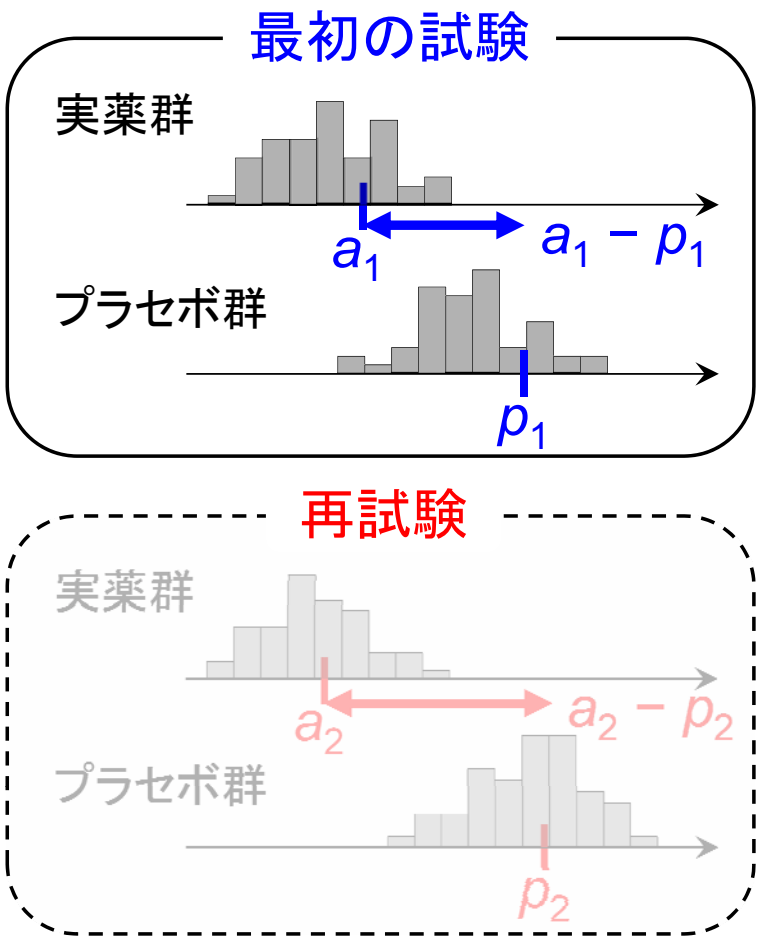
ただし、$\hat{\sigma}_C^2 = \dfrac{\sum_{i=1}^{n_A}\left(x_{Ai} - a_1\right)^2 + \sum_{i=1}^{n_P}\left(x_{Pi} - p_1\right)^2}{n_A + n_P - 2}$

x_{Ai} は実薬群の i 番目の対象者のデータ、a_1 は実薬群の平均値
x_{Pi} はプラセボ群の i 番目の対象者のデータ、p_1 はプラセボ群の平均値

統計的推測の「核」

- 統計量の分布を考慮した推論を行うこと
  - 標準誤差を考慮した推論を行うこと
  - 「もしも同一試験を繰り返したら…」
- 試験を繰り返せば
  - 平均値も分布する
  - 検定統計量も分布する

どうやって統計量の分布を考慮する?



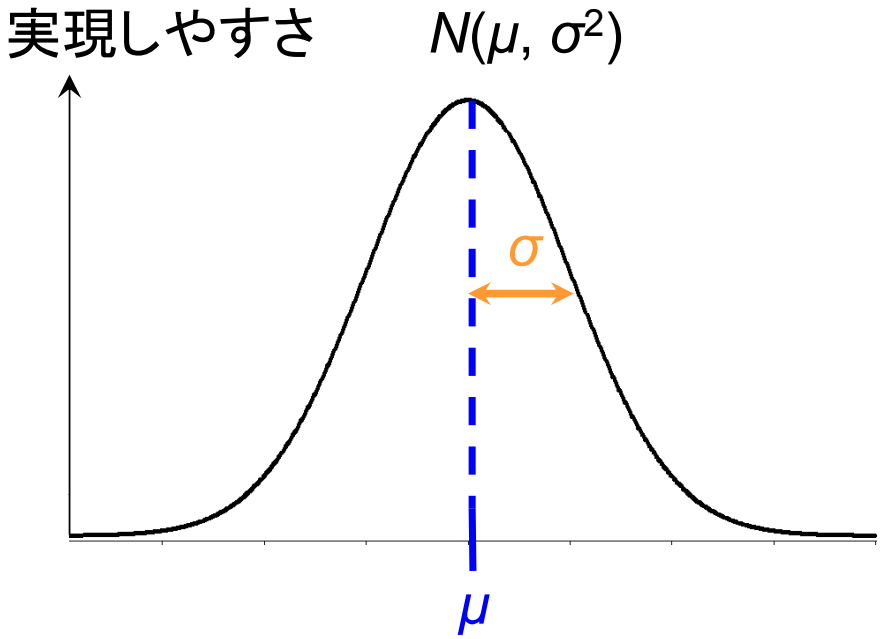
- 試験を繰り返す?
  - 試験の数だけ群間差
    a_1-p_1, a_2-p_2, …
  - 倫理的問題
  - 時間・費用の問題

臨床研究は1回限り

そこで、統計学: 分布

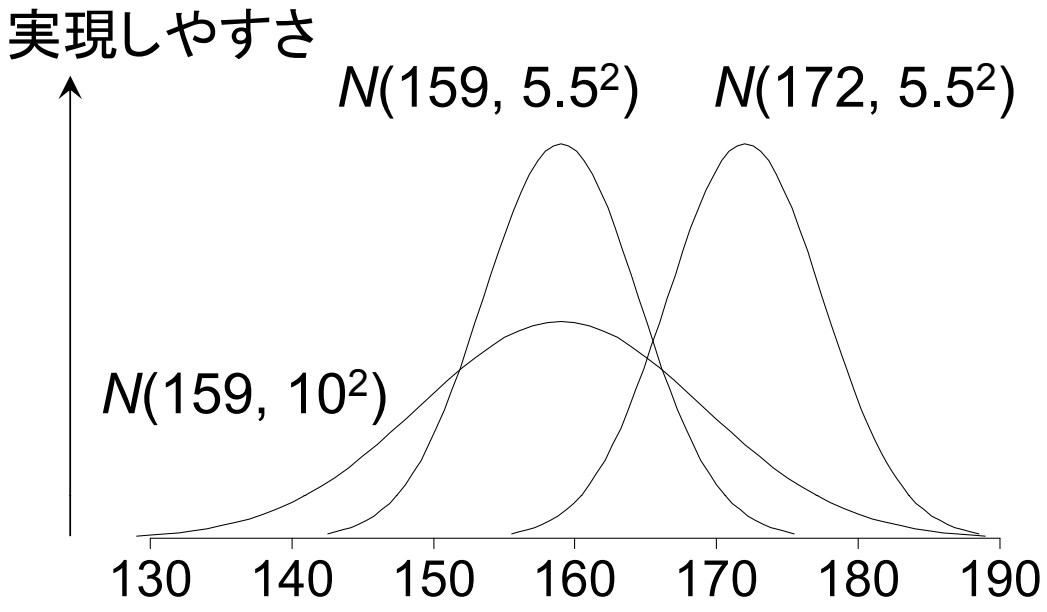
例えば

- 正規分布 Normal distribution
  - パラメータ
    - μ: 期待値
    - σ^2: 分散
      - σ: 標準偏差

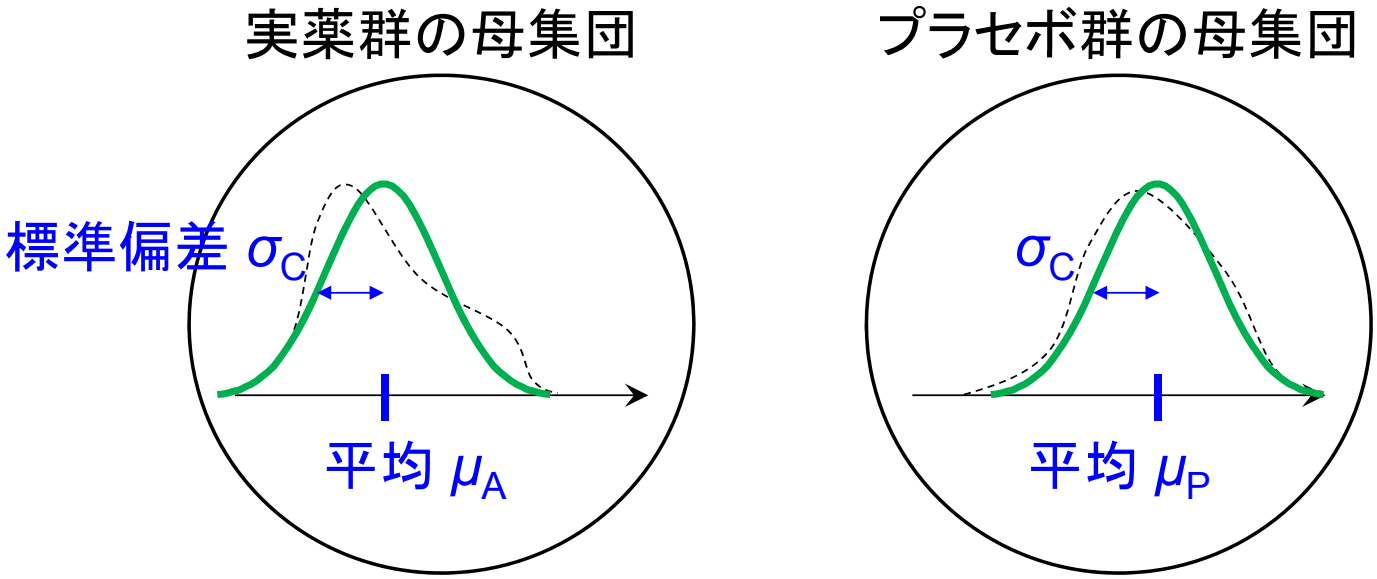


分布の利点

- パラメータの変化によってさまざまな分布を表せる
  - 正規分布 $N(\mu, \sigma^2)$

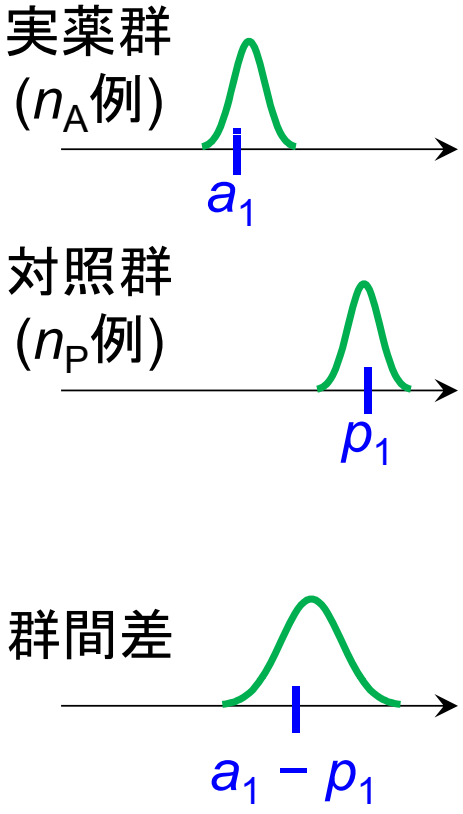


母集団に正規分布を仮定する



- 仮定: データが正規分布に従う
  - 実薬群は $N(\mu_A, \sigma_C^2)$
  - プラセボ群は $N(\mu_P, \sigma_C^2)$
    - μ_A, μ_P, σ_C は定数（真値という）

正規分布の仮定により



試験の繰り返しによる統計量の分布

- 実薬群のデータの平均は

$$N\left[\mu_P, \frac{\sigma_C^2}{n_P}\right]$$ に従う

- プラセボ群のデータの平均は

$$N\left[\mu_A, \frac{\sigma_C^2}{n_A}\right]$$ に従う

- 群間差は

$$N\left[\mu_A - \mu_P, \left(\frac{1}{n_A} + \frac{1}{n_P}\right)\sigma_C^2\right]$$ に従う

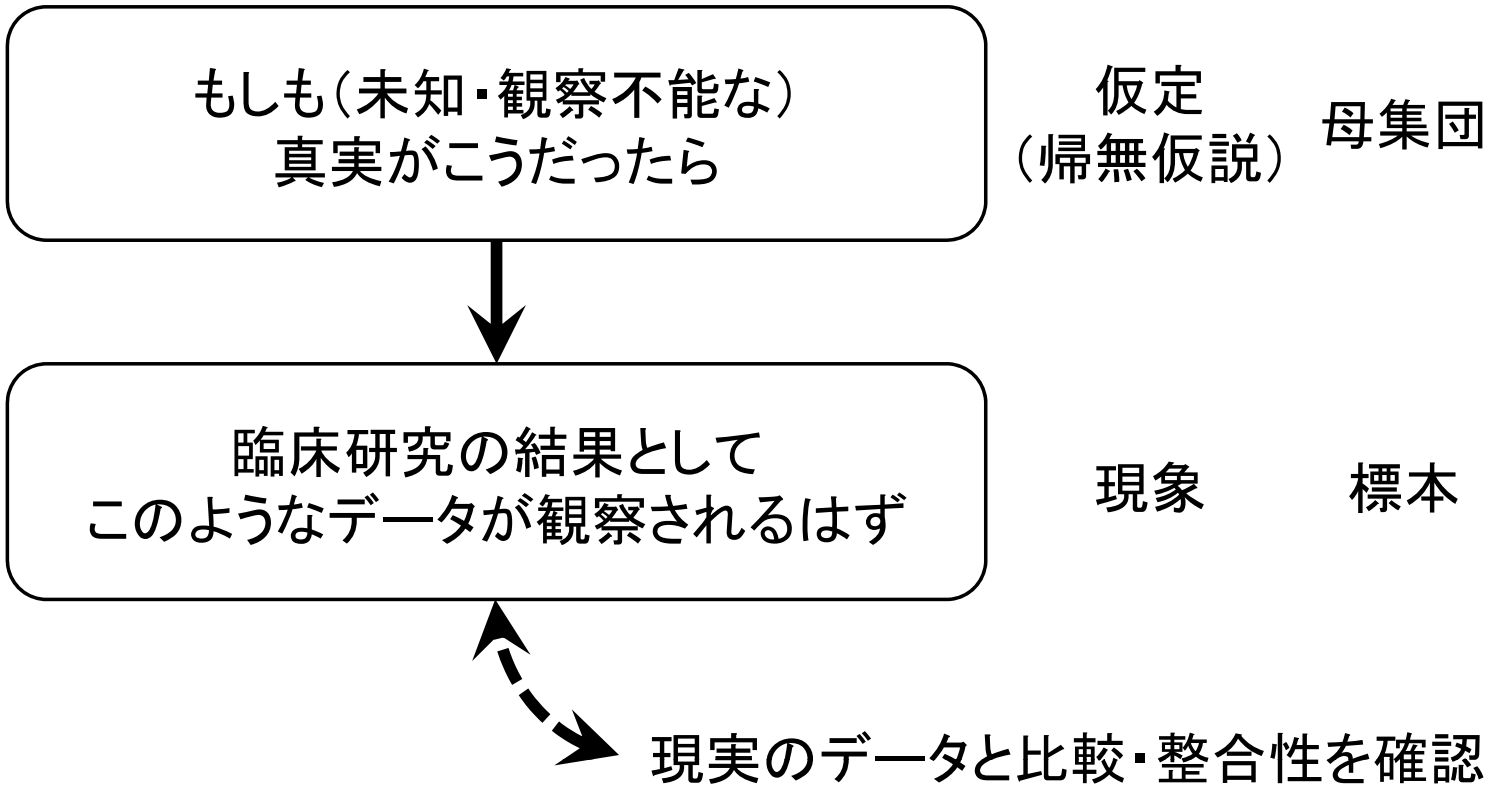
しかし、まだ問題が…

- 群間差の分布

$$N\left[\mu_A - \mu_P, \left(\frac{1}{n_A} + \frac{1}{n_P}\right)\sigma_C^2\right]$$

  - $\delta = \mu_A - \mu_P$ は研究目的そのものであり、未知
  - σ_C も未知

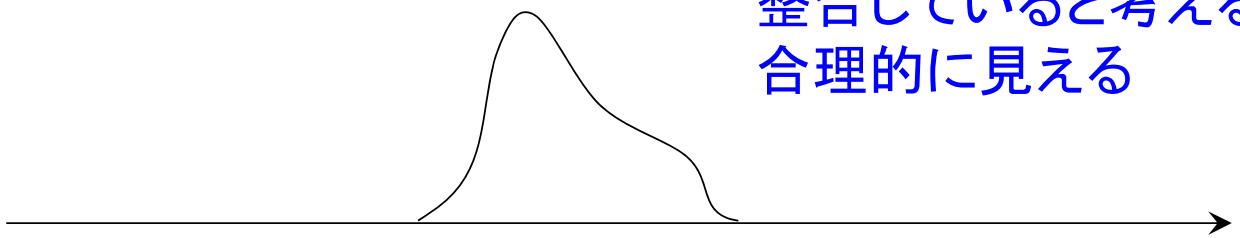
統計学の論法


もしも（未知・観察不能な）
真実がこうだったら
仮定
（帰無仮説）
母集団
臨床研究の結果として
このようなデータが観察されるはず
現象
標本
現実のデータと比較・整合性を確認

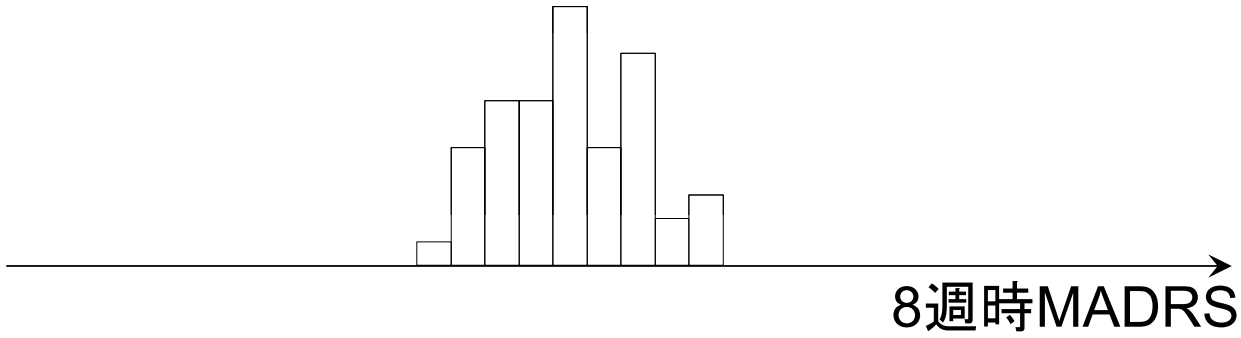
母集団と標本の整合性

- 実薬群の母集団

整合していると考えるのが
合理的に見える



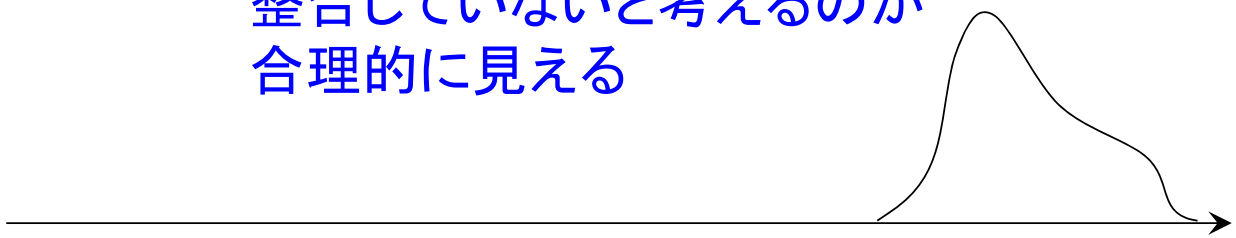
- 実薬群の標本



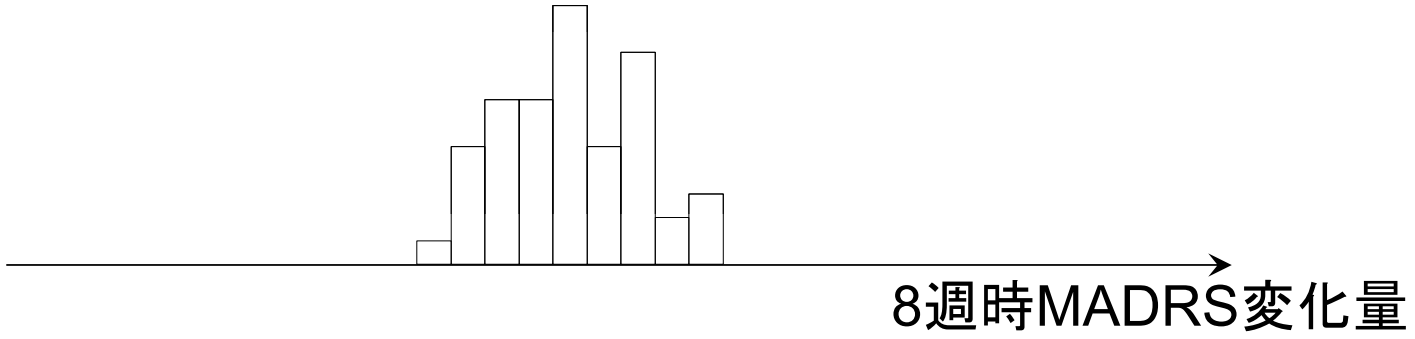
母集団と標本の整合性

- 実薬群の母集団

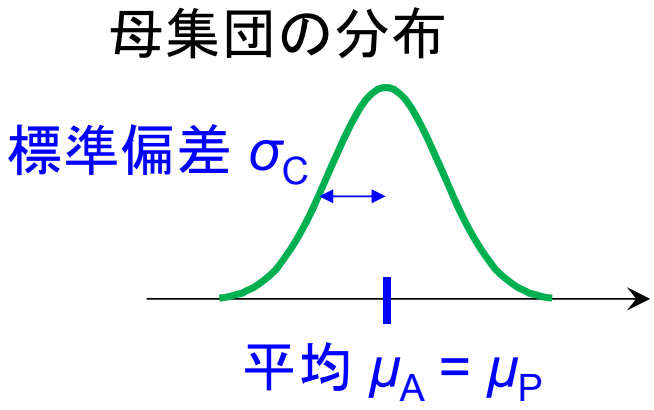
整合していないと考えるのが
合理的に見える

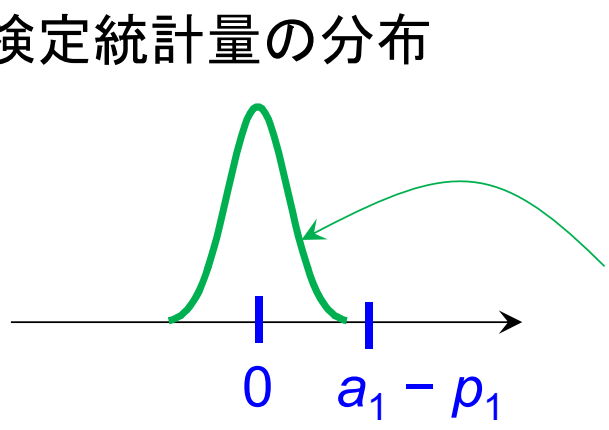


- 実薬群の標本



母集団の分布を決めてしまう



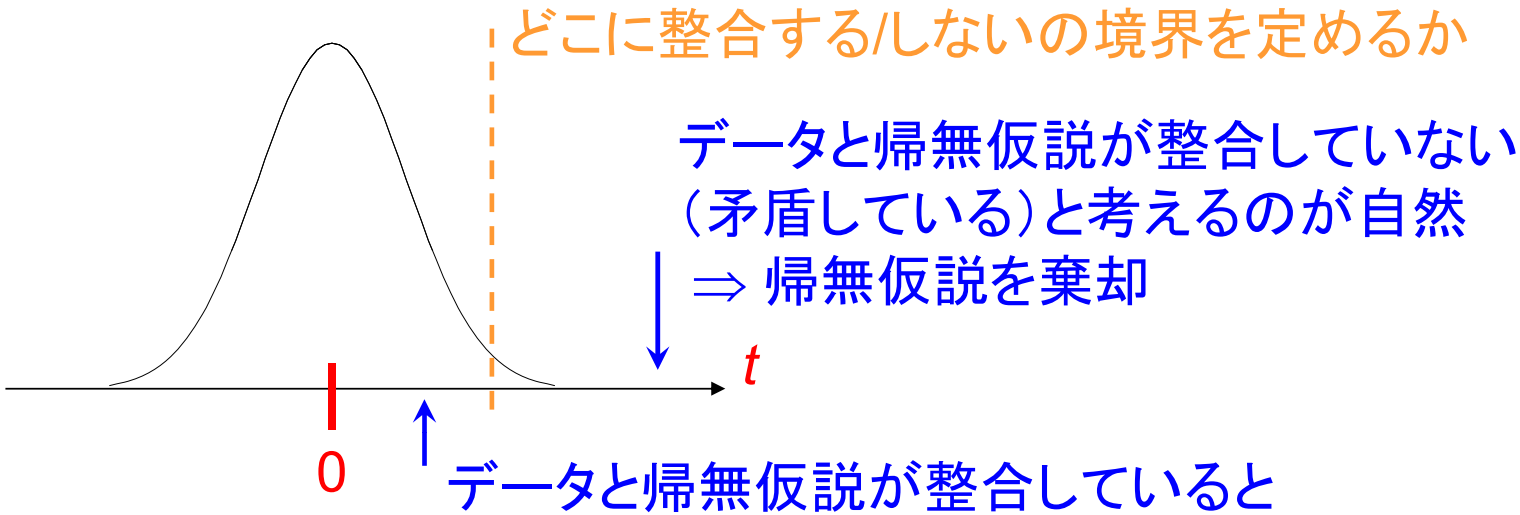


- 適当な値
  - e.g.「もしも実薬は無効ならば」
    - $\delta = \mu_A - \mu_P = 0$ を与える
  - これを帰無仮説（H_0）という
- そうすれば
  - 検定統計量の分布が定まる
  - 帰無仮説とデータの整合性を測ることは可能
  - 整合性が低い ⇒ 帰無仮説を棄却
    - 「与えた値が間違っていた」と推論

t 分布と有効性の判定

- 帰無仮説 $H_0 : \delta = 0$
  - この下でt 検定統計量は自由度 $n_A + n_P - 2$ のt分布に従う

t 検定統計量の分布

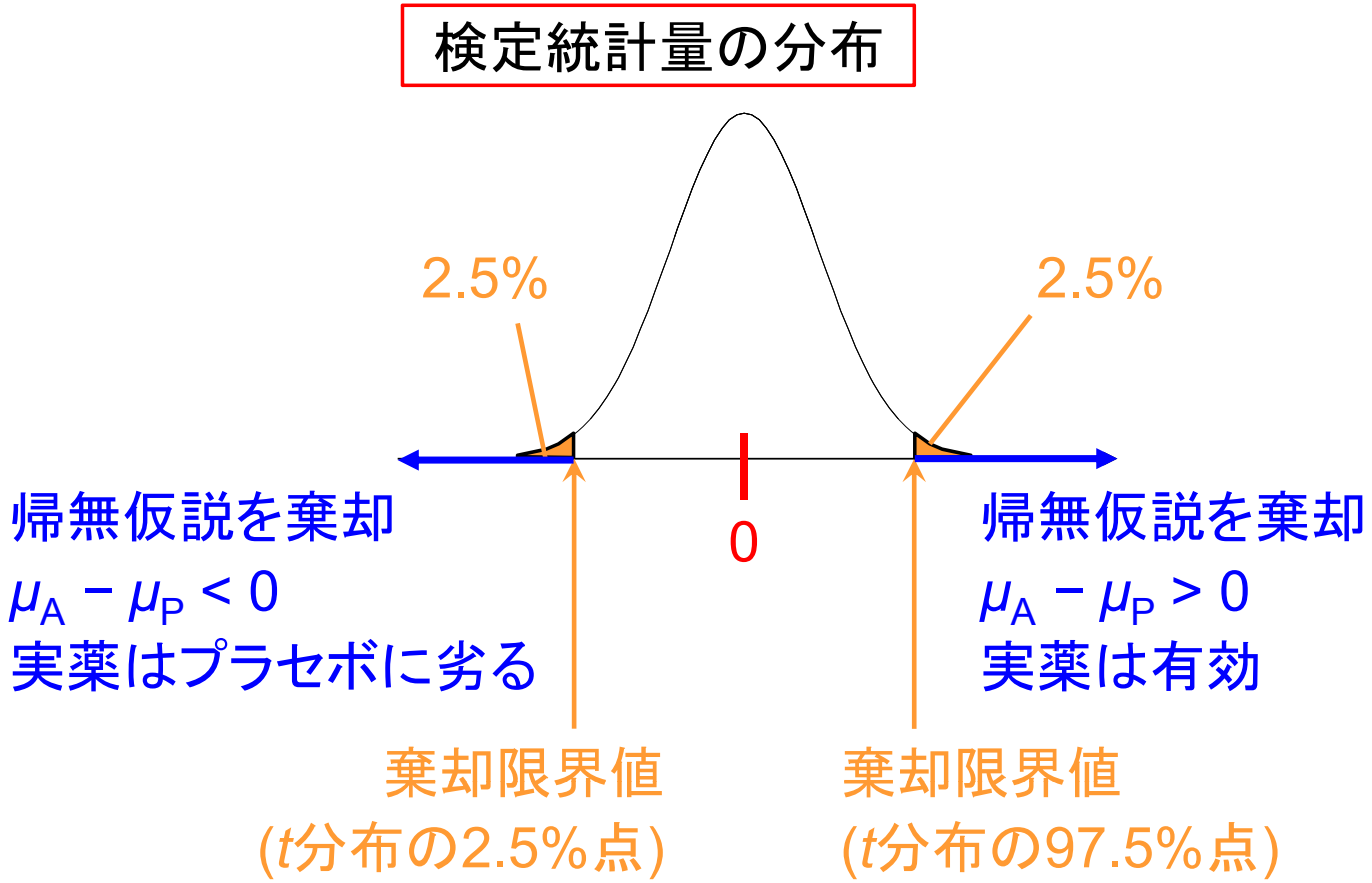


第1種の過誤と有意水準

- どんなに極端に大きな群間差が得られても帰無仮説を100%否定することはできない
  - あらゆる可能性は否定できない
- 第1種の過誤 type I error
  - 帰無仮説が正しいときに帰無仮説を棄却する間違い
    - 実薬が無効なときに有効と結論づける間違い
    - 偽陽性 false-positive
- 有意水準 significance level
  - 有意水準5%: 第1種の過誤を5%まで許容する
    - 無効な治療のうち5%が有効と判断されることを許容

29

有意水準が両側5%の場合



t 検定

- *t* 検定統計量

$$t = \frac{d_1}{d_1\text{のSEの推定値}} = \frac{a_1 - p_1}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_P}\right)\hat{\sigma}_C^2}}$$

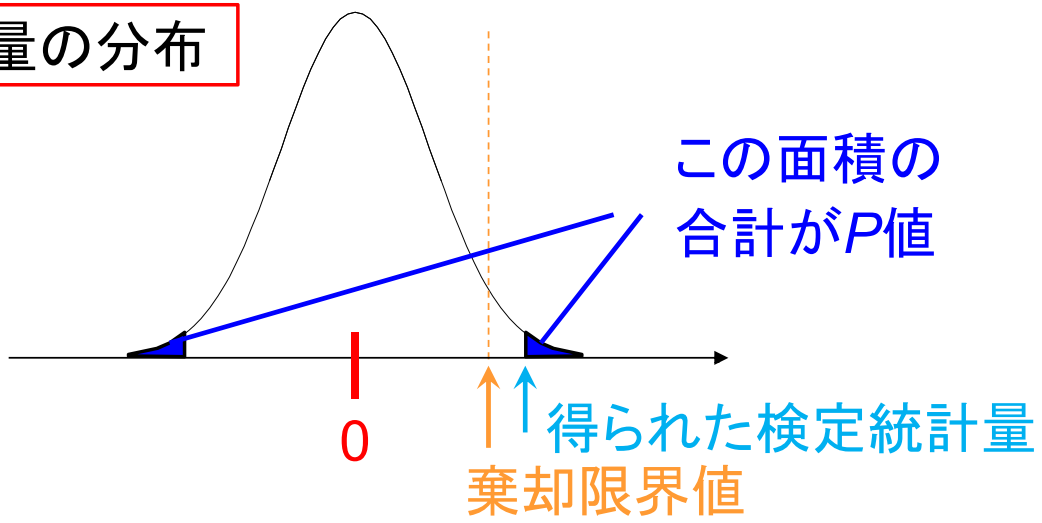
- 有意水準両側5%の検定
  - $t \geq$ (*t*分布の97.5%点) ⇒ 実薬はプラセボに勝る（有効）
  - $t \leq$ (*t*分布の2.5%点) ⇒ 実薬はプラセボに劣る

*P*値 *P*-value

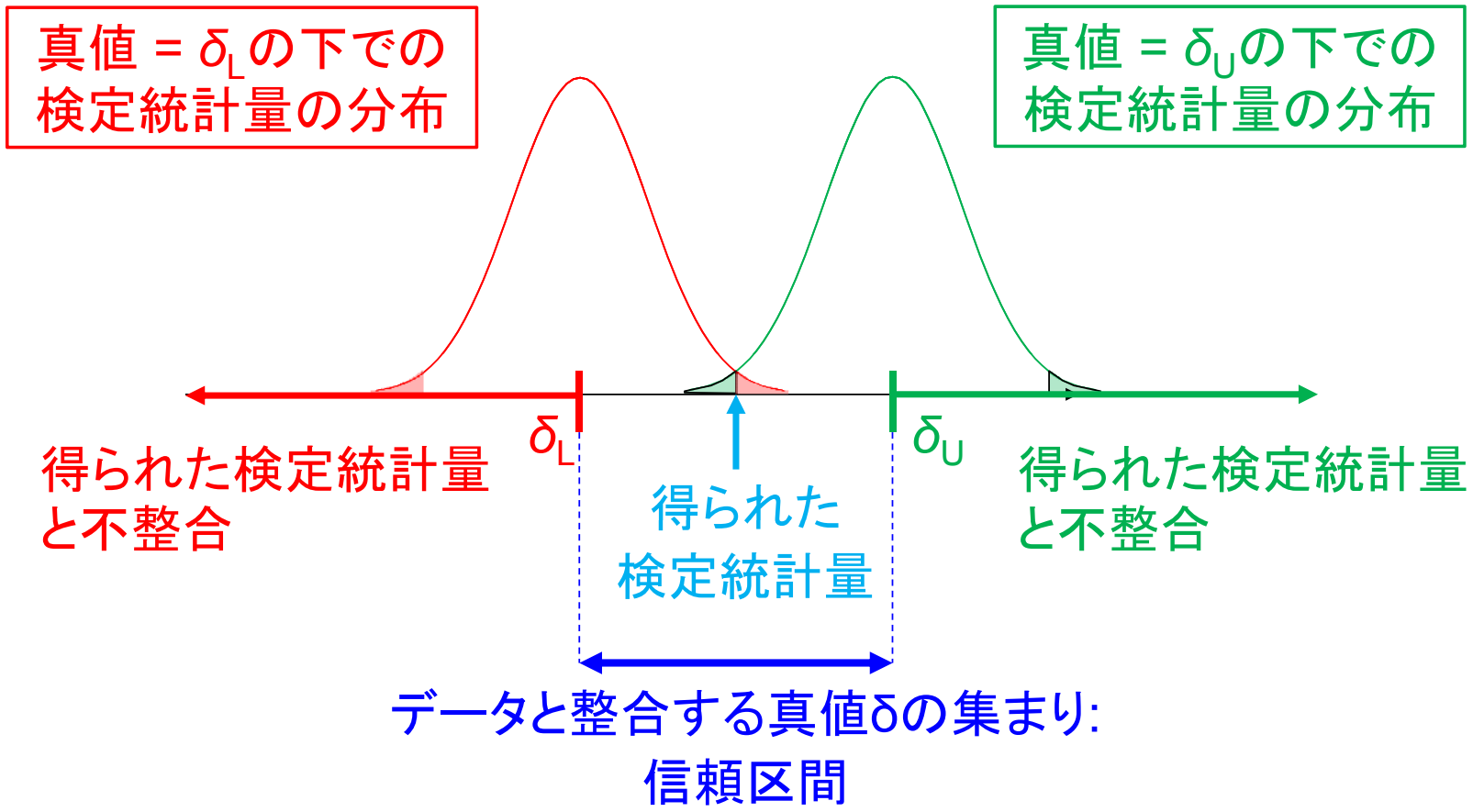
定義：帰無仮説が正しい下で、検定統計量が観測された値かあるいはそれより極端な値をとる確率

- データと帰無仮説の整合性を直接測る指標
- *P*値 ≤ 有意水準　⇒　帰無仮説を棄却

検定統計量の分布



データに整合する帰無仮説は無数



信頼区間 confidence interval

- 得られたデータと整合するすべての真値の集まり
  - 多くの場合に「真値の集まり」は「区間」
  - 有意水準5%の検定を用いれば、95%信頼区間

データ型による解析方法の分類

解析目的	データ型		
	2値	連続	生存時間
単純な群比較*	χ^2検定; Fisher正確検定	*t* 検定; Wilcoxon検定	Log-rank検定
	リスク差, 比; オッズ比	平均値の差	率差, 率比（ハザード比）
層別解析*	CMH検定	分散分析	層別log-rank検定
	標準化; MHリスク差, 比; MHオッズ比	分散分析	MH率差, 比
モデルによる解析	ロジスティック回帰	回帰分析 共分散分析	Cox回帰

* 上段は検定、下段は推定; CMH: Cochran-Mantel-Haenszel; MH: Mantel-Haenszel

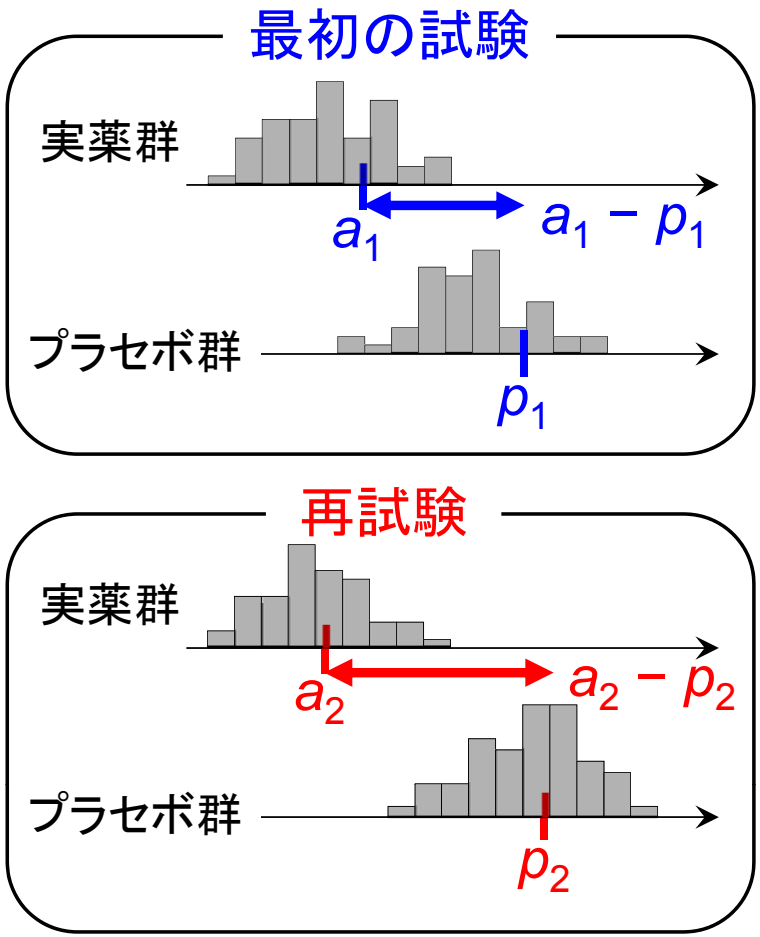
ここまでのまとめ

- 臨床研究の目的は数式で書けなければならない
  - 誰でも同じ式にたどり着けるように PICO + T を書く
- 統計学の論法は背理法
  - もしも母集団の分布がこうならば（帰無仮説）
    ⇒ 標本の統計量はこうなるはず
  - 帰無仮説とデータと整合しなければ帰無仮説を否定する
- データに整合する帰無仮説は無数にある
  - 特定の帰無仮説が唯一正しいと言い切ることは不可能
    - 検定が有意でなくても「治療効果がない（$\delta = 0$）」とは言い切れない

構成

- 検定と信頼区間のしくみ
- 症例数設計

2種類のバラツキ（再掲）



- データのバラツキ
  - e.g. 患者間のバラツキ
  - 標準偏差 standard deviation; SD
    - 記述統計学の範疇
- 統計量のバラツキ
  - e.g. 試験間のバラツキ
    - 実薬群の平均 $a_1 \neq a_2$
    - プラセボ群の平均 $p_1 \neq p_2$
    - 群間差 $a_1 - p_1 \neq a_2 - p_2$
  - データのバラツキに起因
  - 標準誤差 standard error; SE

2種類のバラツキの特徴

1. データのバラツキ
   - データとして観測可能
   - 標準偏差は症例数を増やしても増加も減少もしない
     - ヒストグラムの山が高くなるだけ
2. 推定値のバラツキ
   - 同一の研究を繰り返さない限り観測不可能
   - 標準誤差は症例数を増やした分だけ減少する
     - 症例数を増やせば、推定値の精度（精密度）が上がる
       (= 信頼区間幅が狭くなる)

- 臨床研究で本質的に制御したいバラツキは 2
  - そのために1も可能な限り制御

症例数設計の原理

- 症例数を増やせば推定・検定の精度が向上する
- 目標とする精度を得られる症例数を計算する
  - 目標とする推定精度（信頼区間幅）が得られる症例数
  - 目標とする検定の精度（検出力）が得られる症例数

- 主な症例数設計の方法論
  1. 推定精度ベース
  2. 確率ベース

推定精度ベースか確率ベースか?

- 研究の主解析が推定であれば推定ベース
  - e.g. 早期の臨床試験
- 研究の主解析が検定であれば確率ベース
  - e.g. 後期の臨床試験

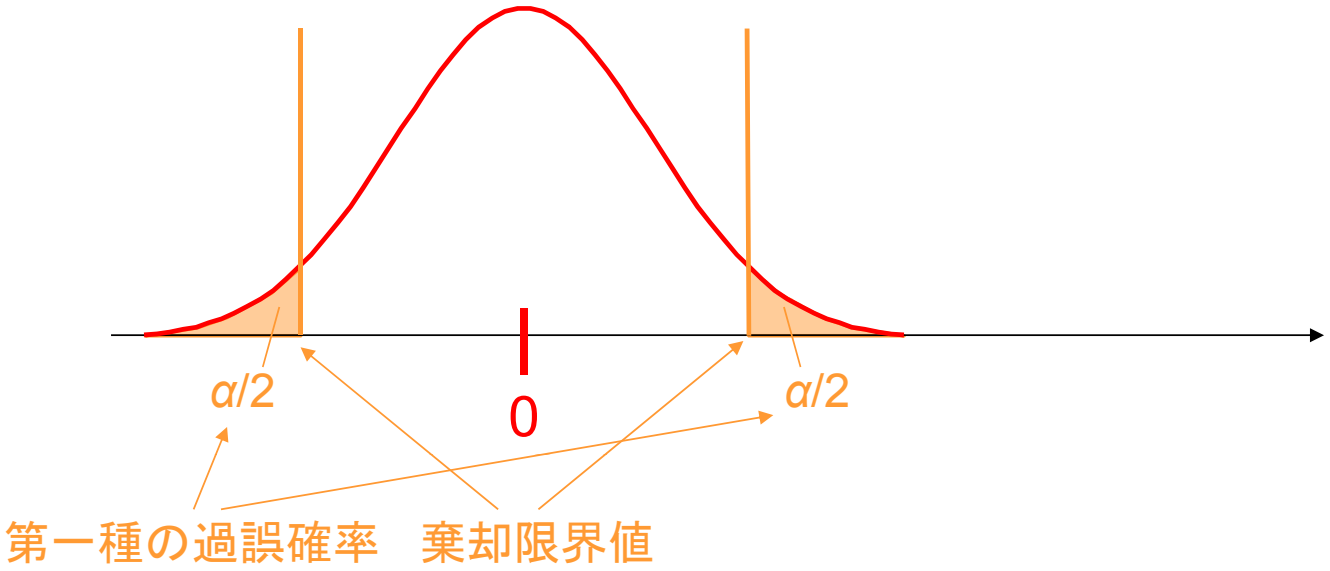
検定のしくみ

- 帰無仮説の下で検定統計量の分布を考えた
  - 帰無仮説 null hypothesis
    - 群間差が 0 (実薬が全く効果を有しない)
- 第一種の過誤確率が有意水準以下になるように棄却限界値を定める
  - 第一種の過誤 type I error
    - 帰無仮説が正しいときに帰無仮説を棄却する間違い
    - αエラー
  - 有意水準 significance level
    - 通常、両側5%

検定のしくみ（両側検定）

- 帰無仮説を棄却するか否かは、帰無仮説の下における検定統計量の分布と有意水準のみで決定される

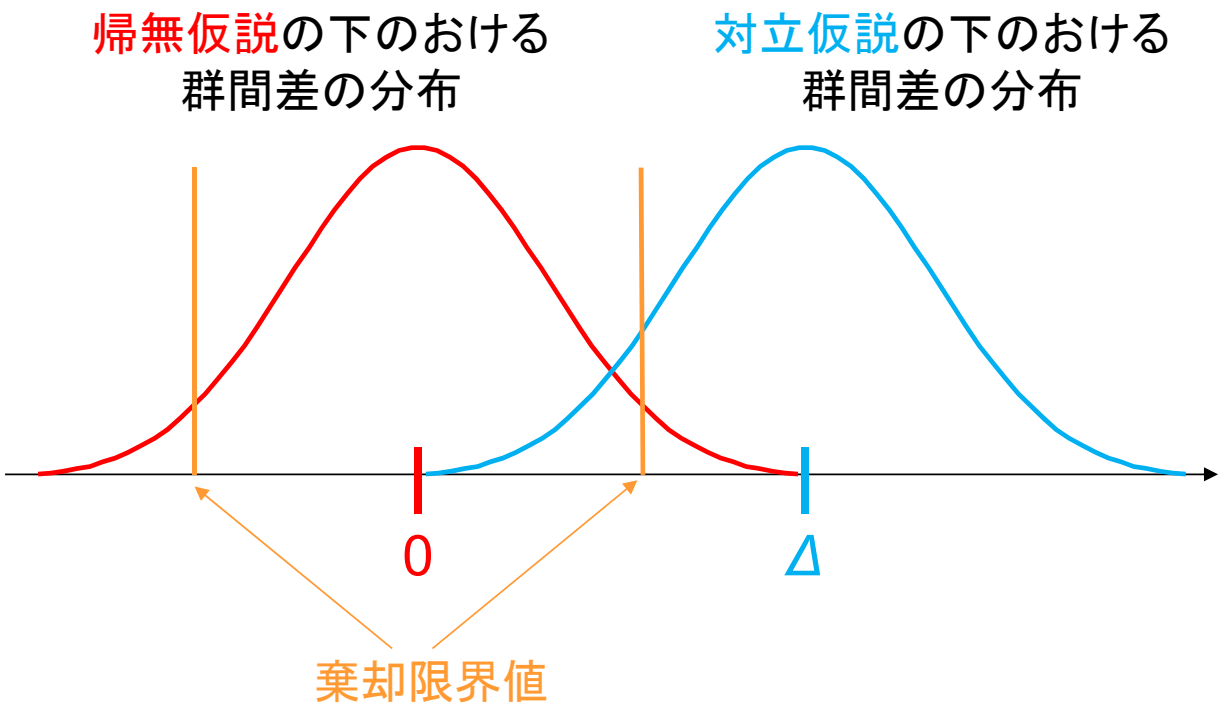
帰無仮説の下のおける
検定統計量の分布



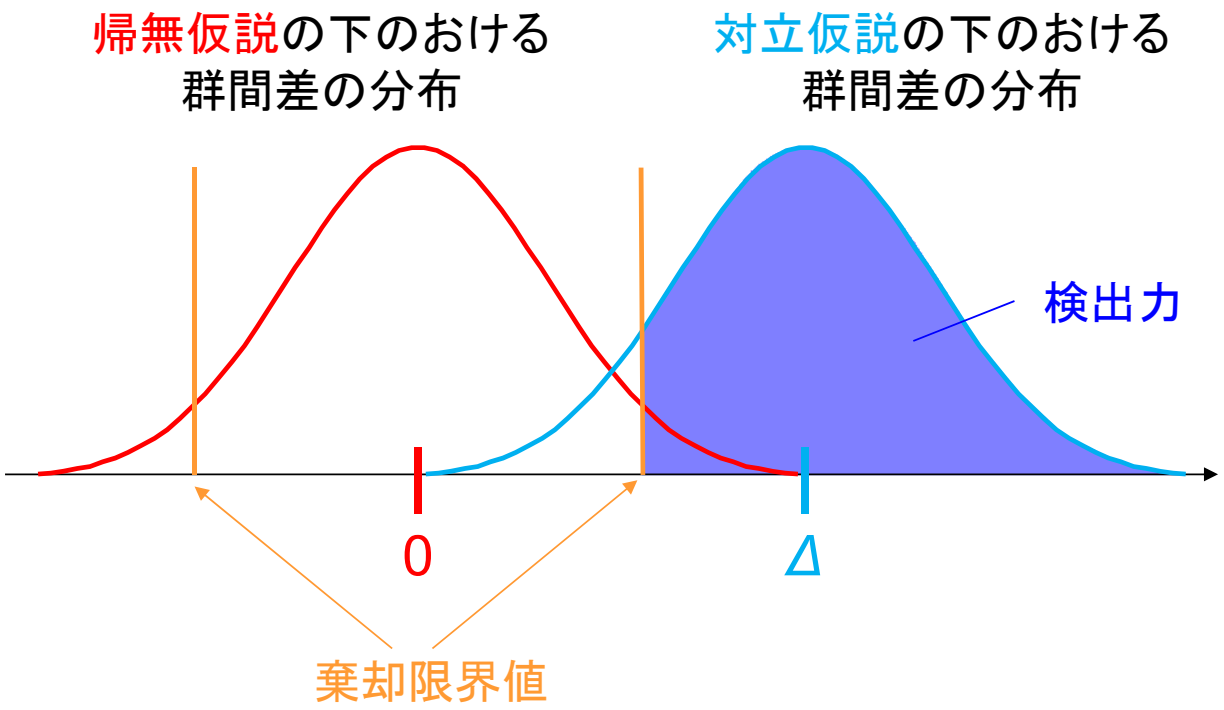
症例数設計のしくみ

- 対立仮説の下で検定統計量の分布を考える
  - 対立仮説 alternative hypothesis
    - 母集団の平均値の差が Δ
- 第二種の過誤確率が一定以下になるように症例数を定める
  - 第二種の過誤 type II error
    - 対立仮説が正しいときに帰無仮説を棄却できない間違い
    - β エラー
  - 検出力 power
    - 対立仮説が正しいときに帰無仮説を棄却する確率
    - 「試験が成功する確率」
    - 1 − 第二種の過誤確率

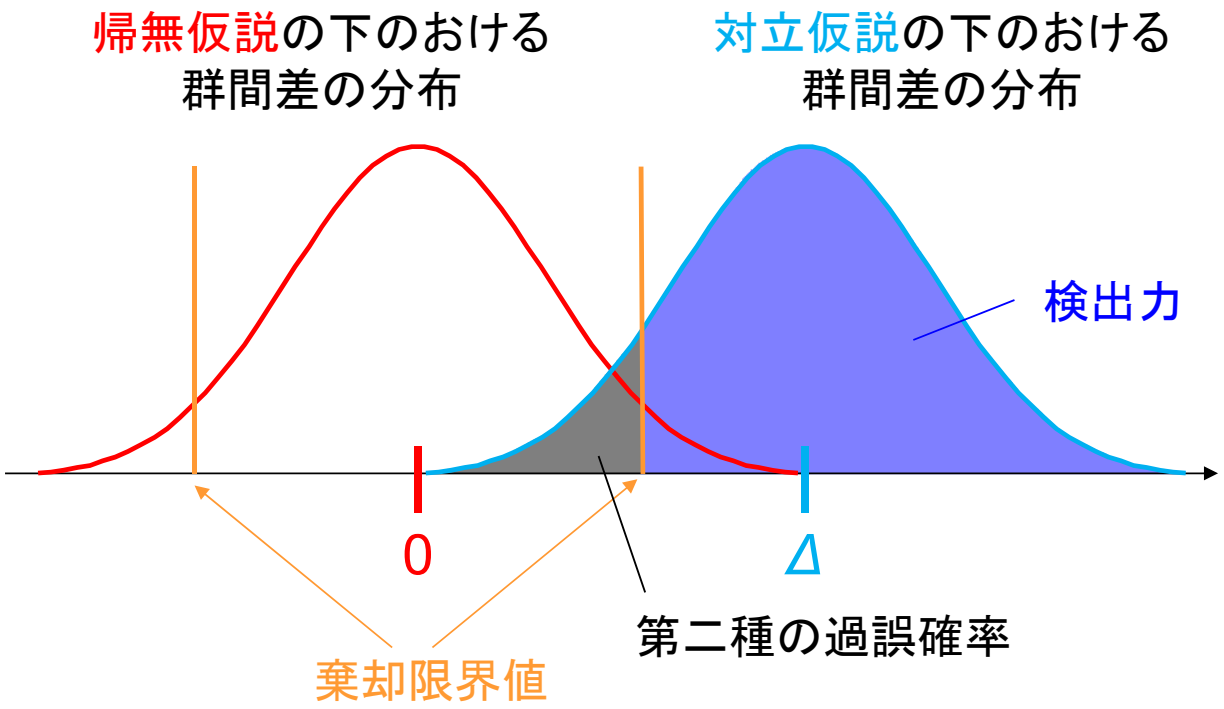
第二種の過誤と検出力



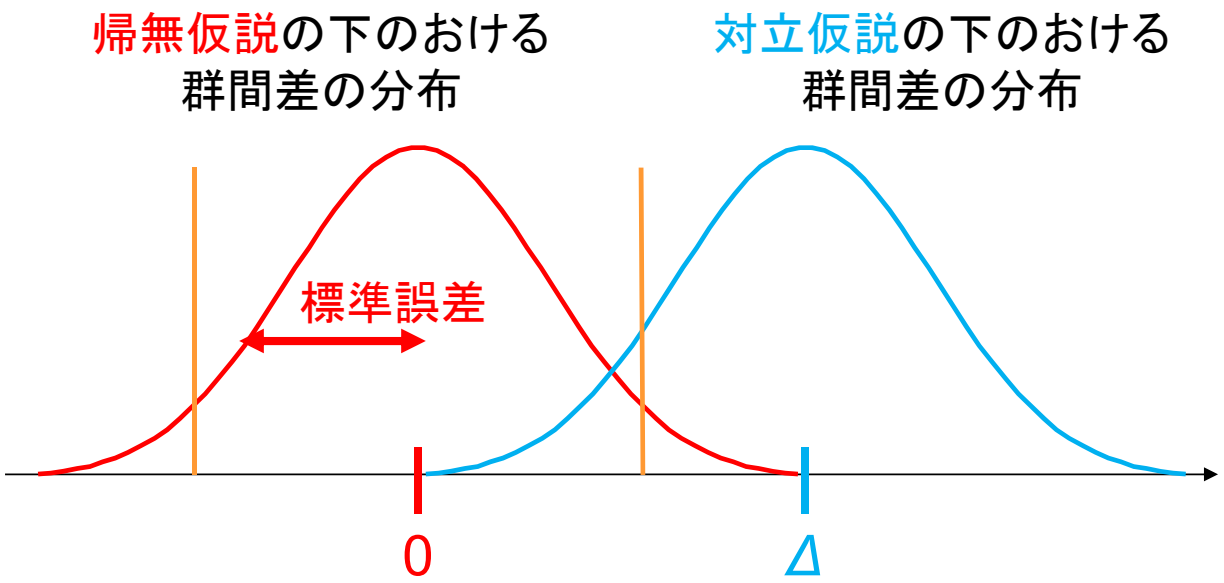
第二種の過誤と検出力



第二種の過誤と検出力

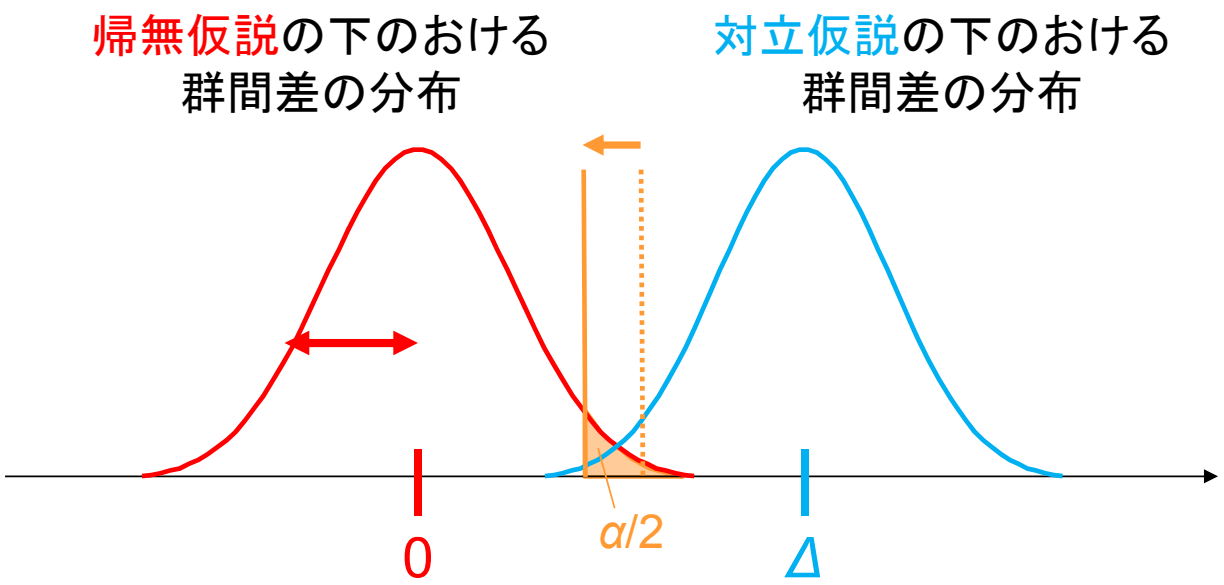

帰無仮説の下のおける
群間差の分布
対立仮説の下のおける
群間差の分布
検出力
0
Δ
第二種の過誤確率
棄却限界値

検出力ベースの症例数設計のしくみ



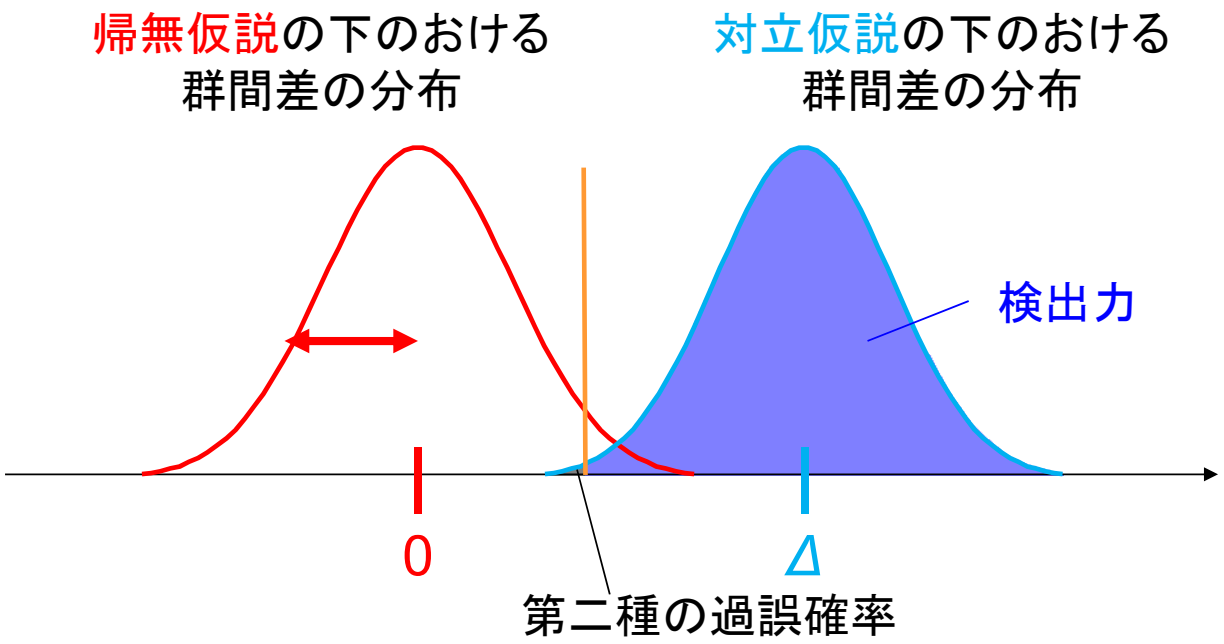
- n が増えると
  - 分布の位置は不変, 標準誤差が減少
  - 有意水準は不変, 検出力が向上

検出力ベースの症例数設計のしくみ



- n が増えると
  - 分布の位置は不変, 標準誤差が減少
  - 有意水準は不変, 検出力が向上

検出力ベースの症例数設計のしくみ



- 症例数が増えると
  - 分布の位置は不変, 標準誤差が減少
  - 有意水準は不変, 検出力が向上

レクサプロの例

- 平均値の2群比較では標準誤差は

$$SE = \sqrt{\left(\frac{1}{n}+\frac{1}{n}\right)\sigma_C^2}$$

- よって症例数は次の等式の解

$$\Delta = \left(z_{\alpha/2} + z_{\beta}\right) \cdot \sqrt{\left(\frac{1}{n}+\frac{1}{n}\right)\sigma_C^2}$$

- 例えば
  - 対立仮説 $\Delta = 4$、標準偏差 $\sigma = 10$
  - 有意水準両側5%($z_{0.025} = 1.96$)、検出力90%($z_{0.2} = 1.282$)
  - 片群 $n = 132$

症例数設計の特徴

- 効果の大きさと症例数の2乗が反比例
  - 効果の大きさ Δ が2倍になれば症例数 n は1/4倍
  - 効果の大きさ Δ が1/2倍になれば症例数 n は4倍

対立仮説 Δ の設定方法

- 「臨床的に意味のある（最小）差 Δ_{MID}」を用いる
  - Minimally important difference; MID
- 効果の大きさ Δ に対して検出力が $1-\beta$ なので
  真の効果の大きさがもし Δ より大きいならば
  その効果の大きさに対し $1-\beta$ 以上の検出力が保証
  - 症例数設計に対する科学的要求を満たす
- 無意味に小さな効果の大きさを検出しない
  - 症例数を多くしすぎない
  - 症例数設計に対する倫理的要求を満たす
- しかし、症例数が実施不可能な規模になることもある

簡易的な症例数設計ツール

- Southwest Oncology Group (SWOG)の WEBページ上で簡易的な症例数設計ができる
  - http://www.swogstat.org/statoolsout.html
- 計算結果は信頼できる
  - 各症例数設計で「Help Document」に参考文献など採用されている方法論の説明が記載
  - ただし、基本的に正確さよりも計算負荷が小さい簡易的な近似計算法を採用していることに注意
- 概算・目安として利用しましょう


54
SWOG Statistical Cente
Two Arm Normal
www.swogstat.org/statoolsout.html
Statistical Tools
These tools are written in JavaScript and may be copied (File | Save As ...) to your workstation to be run without an Internet connection.
WEBをダウンロードして使用も可
Design
One Arm Binomial
One Arm Survival
One Arm Normal
One Arm Non-Parametric Survival
Two Stage
Two Arm Binomial
Two Arm Survival
Other
Two Arm Normal
Binomial Interaction
Survival Interaction
Survival Equivalence
Expected Deaths
Analysis
Frequency Stat
Binomial Confidence Interval
Fishers Exact Test
Probabilities
Normal
Binomial
Poisson
Chi-square
Back To Public Resources
Return To SWOG Stat Center Home Page

まとめ

- 症例数設計の原理
  - 症例数が多いほど標準誤差は小さくなる
  - 目標の精度を達成するように症例数を決定
- 症例数設計の方法論
  - 推定精度ベース
  - 確率ベース（検出力ベース）
- 大事なこと
  - 試験の段階（早期・後期）に関わらず統計学的な根拠をもったサンプルサイズ設計が必要