研究のデザイン

Minato NAKAZAWA, Ph.D.
<minato-nakazawa@people.kobe-u.ac.jp>

観察研究のデザイン

- 現状の記述を目的とした観察研究のデザイン
  - 母集団から適切なサンプリングを行うことに尽きる
  - サンプルサイズが大きいほど精度が高まるがコストがかかる
  - 稀な疾患の場合はある程度観察期間が必要（過去の記録を含む）
- 仮説検証型観察研究のデザイン
  - 横断的研究
    - この場合も適切なサンプリングに尽きる
    - 調べる項目が多いほどサンプルサイズも大きくするべき
  - 症例対照研究
    - 原則として症例は定点で全数把握。いかに適切な対照を選択するかが鍵。
    - 基本的に，症例群と対照群で過去の曝露を比較するデザイン
- コホート研究
  - 曝露群と非曝露群を追跡し，疾病発生などのアウトカムを比較するデザイン

実験計画法

- 実験的研究は，どんなものであれ，注意深くデザインされねばならない。
- 実験計画法は，R.A. Fisher がロザムステッドで行った農学研究に始まる。
- 保健医療分野では，この種のデザインは，毒性試験や臨床試験で用量反応関係を分析するために必須。
- 倫理審査に提出する書類には，適切なサンプルサイズの設定だけでなく，適切な研究デザインが記述されねばならない。

Fisher の 3 原則

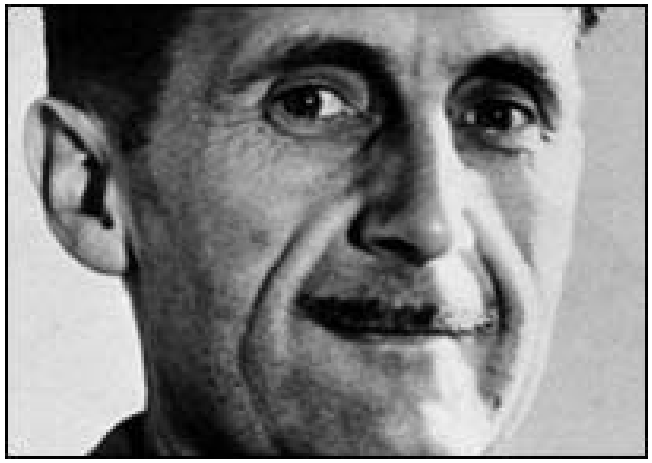
- 繰り返し（ Replication ）: 各処理について最低2回以上の繰り返し測定が必要。
- 無作為化（ Randomization ）: 実験の順序や区画（農業試験の場合）は無作為に割り付けねばならない。
- 局所コントロール（＝ブロック化 : blocking ）: 大規模な実験の場合，サンプル全体の無作為化は不適切であり，代わりにいくつかのやり方で局所のブロックを作り，各ブロック内で無作為割り付けをすることで，ブロック間変動として，偏りを除去することができる。

実験計画の起源に関わる，ある伝説

- ケンブリッジのある晴れた日，多くの教授がアフタヌーンティーを楽しんでいた。と，ある婦人が，自分はミルクティーを飲めば，それがミルクが先か紅茶が先のどちらで淹れたものか判定できると主張した。
- 教授たちの間で，彼女の主張を廻って大激論勃発。
- そこで R.A. Fisher 曰く「実験したらどうだい？」
- この能力は，ミルクを先に入れて作ったミルクティーと紅茶を先に淹れて作ったミルクティーを何杯か用意して，無作為な順番で飲んで貰えば検定可能。
- 能力の判定条件を考える必要あり。何回必要？

（参考）
ミルクと紅茶の順番は本当に味に影響する？

- George Orwell 「完璧な紅茶を淹れる 11 の法則」
  - 10: ミルクを紅茶に加えるのだ。逆ではいけない。
- 英国の王立化学会が George Orwell の生誕 100 年を記念するパーティを開いたとき，Dr. Andrew Stapley (2003) 曰く
  - 冷えたミルクをカップの底に入れておいてから，熱い紅茶を注ぐのが良い。こうするとミルクが紅茶を冷ますことができる。逆だと熱い紅茶がミルクの温度を急に上げるのでミルクの風味が損なわれる。
  - 情報源 : BBC News
  - (http://news.bbc.co.uk/2/hi/uk/3016342.stm)

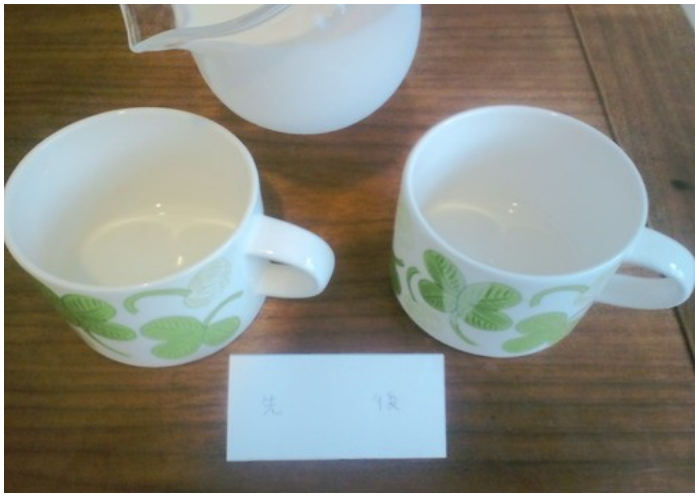


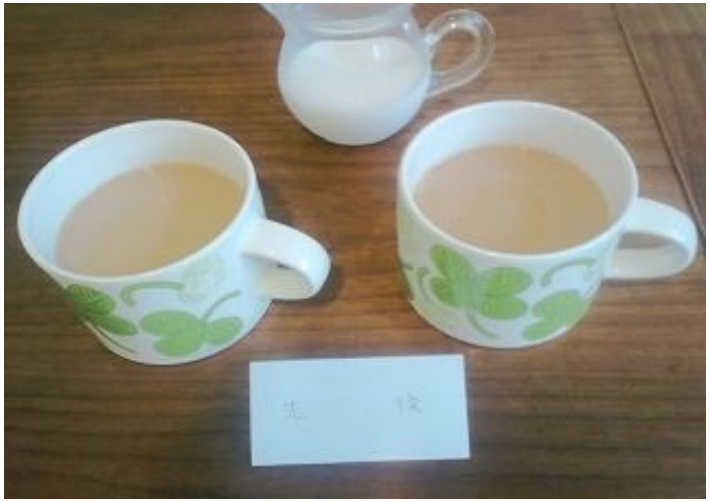
George Orwell's 11 rules

1. Use tea from India or Ceylon (Sri Lanka), not China
2. Use a teapot, preferably ceramic
3. Warm the pot over direct heat
4. Tea should be strong - six spoons of leaves per 1 litre
5. Let the leaves move around the pot - no bags or strainers
6. Take the pot to the boiling kettle
7. Stir or shake the pot
8. Drink out of a tall, mug-shaped tea cup
9. Don't add creamy milk
10. Add milk to the tea, not vice versa
11. No sugar!

完璧な一杯の紅茶の真実（続き）

- 本当はどちらが先だとより美味しいのか，試してみた日本人ブロガーがいた。130cc の紅茶と 30cc のタカナシ低温殺菌牛乳を使用した（高温殺菌とかロングライフのミルクでは違いが分からないらしい）。
- この人の主観的判断では，「ミルクが先」が美味。
  http://blog.livedoor.jp/teatime312/archives/3134982.html





白黒付けるには何杯飲めばいい？

- １杯飲んで正しく判定する確率は 50%.
- ２杯飲んで2杯とも正しく判定する確率は 25%
- ３杯飲んで3杯とも正しく判定する確率は 12.5%
- ４杯飲んで4杯とも正しく判定する確率は 6.25%
- ５杯飲んで5杯とも正しく判定する確率は 3.125 %
  - 本当は判別能力が無いのに偶然5回連続で正解する，3.125% という値は，偶然で片付けるには稀すぎる。通常，この判定基準は 5% を切るかどうかにおく。これが有意水準（Fisher 流）である。帰無仮説「彼女は判定能力をもっていない」が有意水準 5% で棄却される。
- つまり，この仮説を検証するには，最低５杯飲まなくてはいけない

有名な実験計画デザインと検定法

- 単一群，事前 - 事後デザイン：対応のあるｔ検定
- 完全無作為化法（平行群間比較試験）：量的データならｔ検定または分散分析，割合ならカイ二乗検定またはフィッシャーの直接確率
- 乱塊法：完全無作為化法と同様だが，ブロック効果を考慮した分析を行うこともある
- 要因配置法：通常は分散分析
- ラテン方格法：通常は分散分析
- クロスオーバー法：対応のある分散分析で，個人内の差がゼロかどうかを，処理順序を調整して評価

単一群，事前 - 事後デザイン

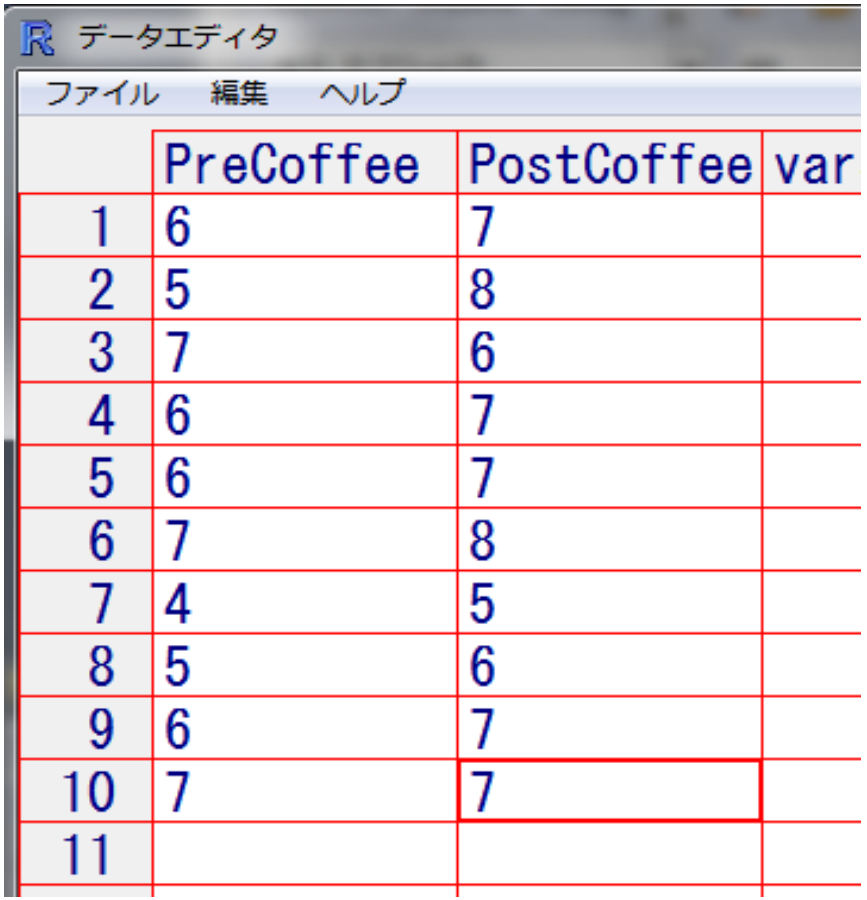
- 個々の対象者について処理前後に同じ精度で測定された値を比較する
  - （例１）関節リウマチ患者について手術前後で血清コルチゾル濃度を比較する
  - （例２）うつ患者に対して音楽療法の前後でうつ質問紙の得点を比較する
  - （例３）健康なボランティアに対して，珈琲を飲む前後で単純計算の得点を比較する
- 統計的検定法は，対応のあるｔ検定が普通
- 帰無仮説：差の平均値がゼロ

EZR での対応のある t 検定の実行例

- デザインは単一群事前事後比較デザインではなく，横断研究だが，ここでは MASS パッケージに含まれている survey というデータフレームを使う
- EZR では，「ファイル」「パッケージ内のデータを読み込む」と選び，MASS パッケージ，survey データフレームの順で選択する
  - survey はアデレード大学の 237 人の学生に対して多くの質問をした回答である (Venables and Ripley, 1999) 。
  - 変数には，字を書く手とそうでない手について，掌を広げたときの親指の先と小指の先の距離を示す変数（前者が Wr.Hnd，後者が NW.Hnd ）が含まれる
- 「統計解析」「連続変数の解析」と選んで，「対応のある2群間の平均値の比較 (paired t 検定) 」を選ぶ。第1の変数として NW.Hnd，第2の変数として Wr.Hnd を選んで OK ボタンを
- クリックすると検定が実行される

練習問題

- 「コーヒーを飲むと計算能力が変わるか」を調べるため，10 人のボランティアに単純計算をしてもらい，その得点をコーヒー飲用前後で比較
- EZR では，まずデータを入力
- 「ファイル」「新しいデータ」を選んでデータセット名を入力し，右の図のようにデータを入力する
- 「統計解析」「連続変数の解析」から「対応のある2群間の平均値の比較（paired t 検定）」を選び，第1の変数として PreCoffee，第2の変数として PostCoffee を選び，OK ボタンをクリック
- [t = -2.862, df = 9, p-value = 0.01872] と結果が出る。有意水準 5% で統計学的に有意な差があるといえる



	PreCoffee	PostCoffee	var
1	6	7	
2	5	8	
3	7	6	
4	6	7	
5	6	7	
6	7	8	
7	4	5	
8	5	6	
9	6	7	
10	7	7	
11			

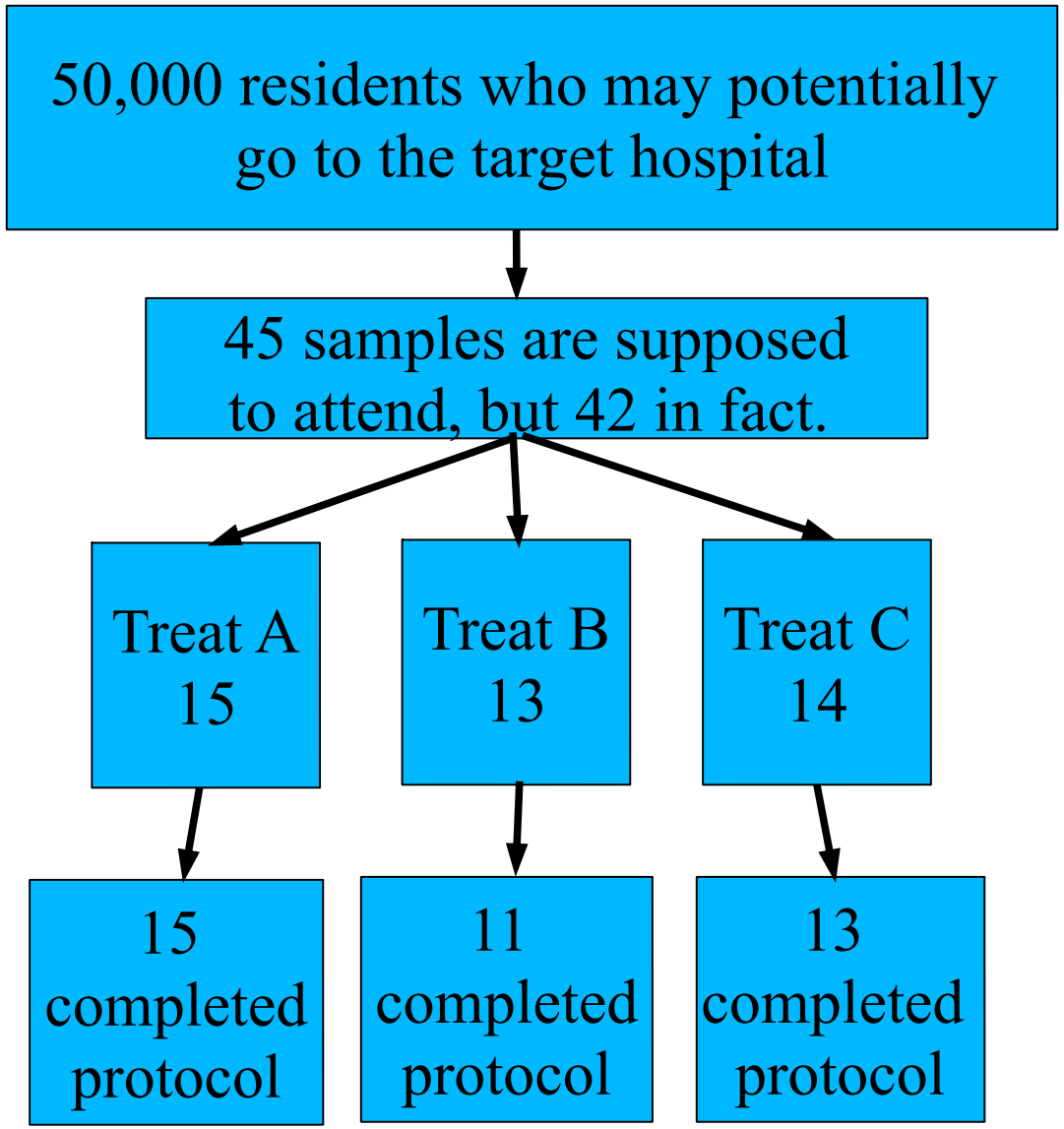
平行群間比較試験(完全無作為化法)

- きわめて単純。インフォームドコンセントを得た対象者を完全にランダムに(行き当たりばったりではなく), いくつかの処理(曝露)群の1つに割り付ける
- ランダム化の方法にはいくつかある。Fleiss JL (1986) は,「臨床試験のデザインと解析」という本で, 乱数表の代わりに乱数順列表を使うことを薦めている
- しかし今ではコンピュータソフトを使えば数表を使う必要は無い。45 人に3つの処理のどれかを割り付けたければ, R では,
- matrix(sample(1:45, 45, replace=FALSE),3) と打てば良い。

```
> x <- matrix(sample(1:45, 45, replace=FALSE),3)
> x[1,]
 [1] 34 13 37 17  3 11  4 23 18 39 42 40 36  8 25
> t(apply(x,1,sort))
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15]
[1,]    3    4    8   11   13   17   18   23   25    34    36    37    39    40    42
[2,]    1    6    7   10   14   15   16   19   26    29    31    33    41    43    44
[3,]    2    5    9   12   20   21   22   24   27    28    30    32    35    38    45
```

このデザインの記述方法

- 注意すべきは，脱落が起こることである
- このデザインは通常，右図のようなフローダイアグラムで示される（募集段階で３人脱落し，割り付け後にも３人脱落）。
- この図のような場合，統計解析は，アウトカムが連続量なら一元配置分散分析，アウトカムがカテゴリならカイ二乗検定になる
- 群間でサンプルサイズに差があると統計学的検出力が低下する



乱塊法

- 何かの理由で研究が予定通り完了しないと，サンプルサイズがアンバランスになる可能性がある
- この欠点を克服するのが乱塊法。もし各群のサンプルサイズが 15 で，処理が3通りなら，乱塊法では，{A, B, C}, {A, C, B}, {B, C, A}, {B, A, C}, {C, A, B}, {C, B, A} という塊を 15 回ランダムにサンプリングする。そうすれば，もし途中で研究が中断しても，群間のサンプルサイズの違いは1人以内に抑えられる
- 記述と分析は基本的に完全無作為化法と同じだが，ブロックの効果を考慮した分析もされることがある
- サイズのバランスを保つ方法はもう1つ。「最小化法」では，サンプリング時点ごとに群間のサンプルサイズの違いが最小になるように制約してランダム割り付け

要因配置法

- 複数の要因の影響を調べるには一般的
- ２ｘ２の要因配置法の例を示す
- McMasterら(1985)：乳がんの自己触診の教材として，小冊子とテープ／スライドを評価する無作為化試験を実施
  - ２種類の教材があるので，２×２の組合せで，４種類の処理がある。
  - 逆手にとれば，４種類の平行群間比較試験と考えることができる
    1. 小冊子もテープ／スライドも与えない（対照群）
    2. 小冊子のみ与える
    3. テープ／スライドのみ与える
    4. 両方与える
- こう考えれば，小冊子，テープ／スライドとその交互作用についての教育効果は分散分析で評価できる

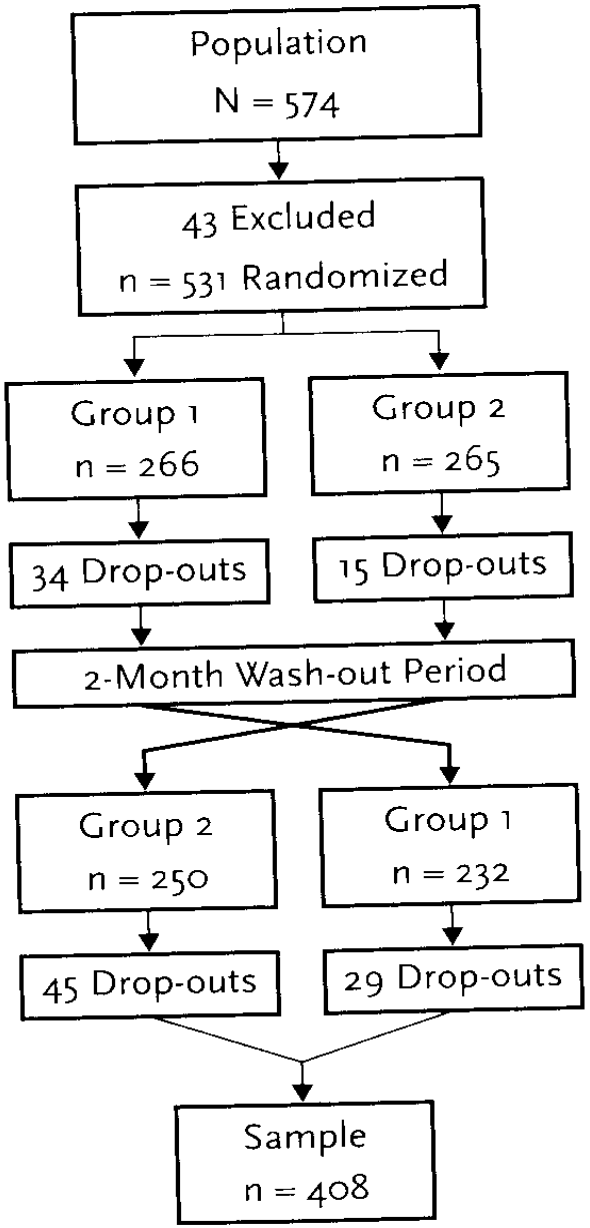
ラテン方格法

- p（≧２）個の水準をもつ1つの処理Ａの効果を調べる実験において，やはりp個ずつの水準をもつ，その効果を歪める変数Ｂ，Ｃがあるとき，このデザインが有用である。
- 名称は古典的なパズルからきている。
- pを３と仮定すると，ラテン方格は右図
- 第1グループはn1人について処理（a1, b1, c1）の組合せ，第2グループはn2人について処理（a1, b2, c3）の組合せというように実験を進める
- こうすることで，分散分析におけるアウトカム測定値に対するＢとＣの効果をコントロールできる。

	c1	c2	c3
b1	a1	a2	a3
b2	a2	a3	a1
b3	a3	a1	a2

Crossover Design（クロスオーバー法）

- 対象者はランダムに2群に分けられ，それぞれ2種類の介入を，適当な間隔（前の処理のキャリーオーバーを避けるためのウォッシュアウト期間と呼ばれる）をおいて，群ごとに異なる順番で受ける。
- （例は右のフローチャート）
  Hilman BC et al. “Intracutaneous Immune Serum Globulin Therapy in Allergic Children.”, JAMA. 1969; 207(5): 902-906.
- 統計解析は，プレテストとして (1) キャリーオーバー効果が無視できるか検定（2群それぞれの2つの測定値の和の平均値の差の検定），それが確認できたら，(2) 2群それぞれの差の平均値の差の検定
- (Wellek S, Blettner M: Dtsch Arztebl Int. Apr 2012; 109(15): 276–281. doi: 10.3238/arztebl.2012.0276)



実験結果の比較のタイプ

参考：https://www.igaku-shoin.co.jp/paper/archive/y2012/PA02971_04

- 優越性試験 (Superiority trials)：新しい処理の効果が，対照より統計学的に有意に優れているかどうか
  - 「差がない」帰無仮説を検定して，p 値が有意水準より小さければ帰無仮説を棄却する
- 同等性試験 (Equivalence trials)：新しい処理の効果が，対照と同等かどうか
  - 検定でなく，“十分なサンプルサイズ”で正確に同等だというために信頼区間を用いる。事前に決めた「±○× %の差であれば臨床的に同等」とみなす同等性の許容範囲（同等性マージン，研究計画書にも記載）内なら同等とみなす
- 非劣性試験 (Non-inferiority trials)：同等性試験の特別な場合。ジェネリック医薬品の試験などでよく用いられる
  - 信頼区間を片側にすることでサンプルサイズを節約
- R では ThreeArmedTrials というパッケージを使うと，試験群，参照群，プラセボ群がある場合の非劣性試験に便利
  https://cran.r-project.org/web/packages/ThreeArmedTrials/vignettes/ThreeArmedTrials.html

効果量

- 心理学分野での統計改革（Cohen 1994 がきっかけ）
  - APA Publication Manual 6th Ed. (2009) は，信頼区間をブラケットで示すことや p 値の後に効果量を記載することが APA の論文誌では最低限必要と書かれている
  - 心理学の他には，教育学，社会調査，疫学の一部で使われる
- 定義「効果の大きさをあらわす統計的な指標」
  - 帰無仮説が正しくない程度を量的に示す
  - 統計量はサンプルサイズの関数と効果量の関数の積なので，効果量は，検定統計量からサンプルサイズに依存する部分を除去したものともいえる
- d 族効果量（群間差の大きさを表す。effsize パッケージ cohen.d 関数）
  - Cohen の d：標本の平均値の差を 2 群をプールした標準偏差で割る
  - Hedges の g：標本の平均値の差を 2 群に共通な母分散を標本から推定する際の不偏推定量で割る
  - https://minato.sip21c.org/ebhc/dfamefs.R 参照
- r 族効果量（変数間の関係の大きさを表す。相関係数など）