

臨床試験における必要症例数の計算

医学統計セミナー2017 第5回目

下川敏雄

(代) 谷岡健資

医療データサイエンス

(附属病院臨床研究センター)

11月28日(火)	診断・検査データの解析：ROC曲線を中心に
12月22日(金)	多変量解析：重回帰・ロジスティック回帰・比例ハザードモデル
1月26日(金)	傾向スコアによる解析：マッチングによる解析を中心に
2月23日(金)	メタアナリシス：2値・量的・生存期間データに対するメタ回帰分析
3月23日(金)	臨床試験における必要症例数の計算

目次

□はじめに

□症例数設計の基本

□症例数設計のプロセス

□症例数設計の事例

□サンプルサイズを減らすために

□症例数設計のためのフリーソフトウェア

□おわりに

必要症例数

臨床的仮説(Clinical Question)に信頼のおける解答を与えられるような被験者数のことを指す

研究計画書を策定する上で必要症例数を記載することが「当たり前」となっている

必要症例数を計算するために必要な項目

必要症例数を通常の方法に基づいて決定するために必要な項目は下記の通り(ICH E9より)

1. 主要変数
2. 検定統計量
3. 帰無仮説
4. 選択された用量での対立仮説
 - その用量と選ばれた対象集団で検出すべき又は棄却すべき試験治療の差を考慮することも含めて
5. 誤って帰無仮説を棄却する確率（第一種の過誤）
6. 誤って帰無仮説を棄却できない確率（第二種の過誤）

症例数とp値の関係（症例数は検定結果に影響を及ぼす）

胃がんに対する二つの臨床研究があったとする。これらは、同一対象患者に対して、同一レジメンの術後補助化学療法が実施された。なお、これまでの当該患者に対する奏効率は40%だった。

Study A

症例数：40例, 奏効例：20例

奏効割合： $\hat{p} = 0.5$ (50%)

p 値(閾値0.4)でのp値 = 0.130 (n.s.)

Study B

症例数：80例, 奏効例：40例

奏効割合： $\hat{p} = 0.5$ (50%)

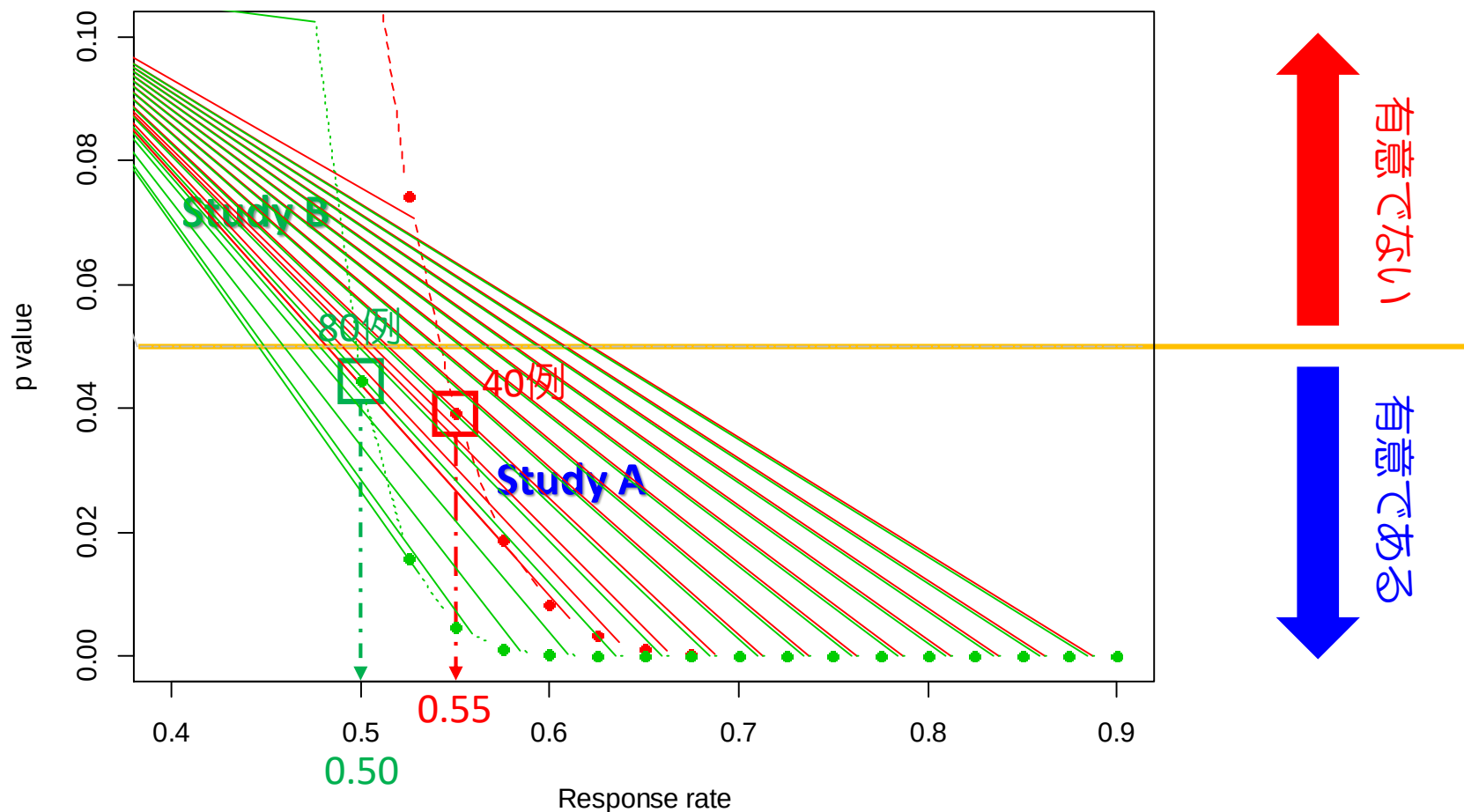
p 値(閾値0.4)でのp値 = 0.045*

全く同じ奏効割合(0.50)であったとしても、症例数の違いから、Study Aではnegative studyとなり、Study Bではpositive studyとなる。

Study AがPositive studyになるにはどのような状況である必要があるか？

症例数決定の基本

Study A と Study Bで有意になる奏功割合



80例の場合には、奏効割合が0.5 (50%)以上の奏効割合で有意になるが、40例の場合には0.55 (55%)にならないと有意にならない。

臨床試験では、新薬(治療)で期待される効果量を予め規定し、試験の結果、期待される効果が得られたときに、有意差が得られるような最小症例数を計算する。

	検定により評価	標本サイズで規定
	H_0 が正しい [H_1 : 偽]	H_1 が正しい [H_0 : 偽]
H_0 を棄却 [H_1 を受容]	第1種の過誤 α (H_0 が正しいのに棄却)	検出力 $1-\beta$ (H_0 が正しいので受容)
H_0 を受容 [H_1 を受容]	正しい判断 $1-\alpha$ (H_1 が正しいので棄却)	第2種の過誤 β (H_1 が正しいのに受容)

- 第2種の過誤が一定水準未満(β 未満)になるように、標本サイズを規定する (本当はpositiveなのに本試験を通じてnegativeと判断する可能性を低くする)。
- 統計解析の結果、第1種の過誤が一定水準未満(α 未満)であるか否かを確認する (本当はnegativeなのに本試験結果を通じてpositiveと結論付けるエラーがかなり低いことを確認する)。

対立仮説のもとでの検定統計量の分布

臨床的に有効（あるいは安全）であると判断される
治療効果の大きさ（エフェクトサイズ）によって設定される

エフェクトサイズの設定

統計学的な観点ではなく，臨床的な観点から設定しなければならない

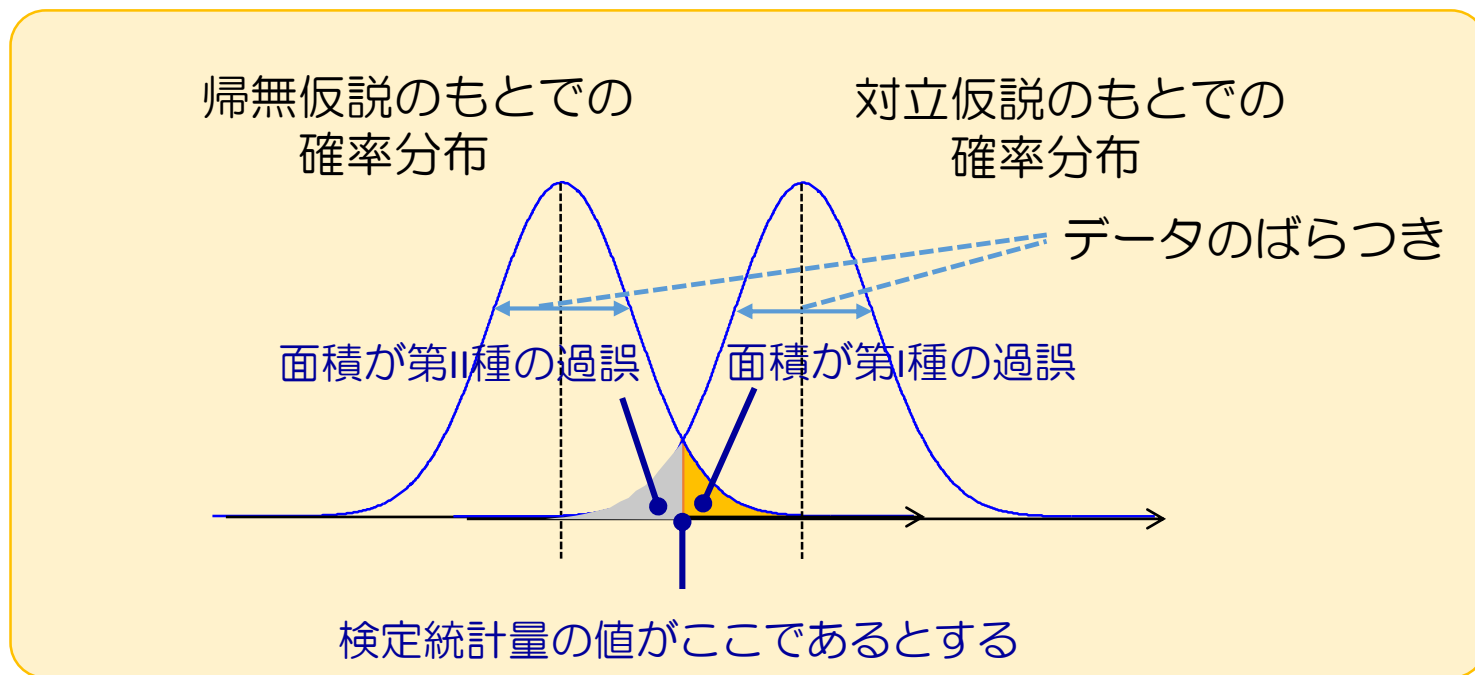
例1) 統計学的観点のみではどんなに小さな差であろうともサンプルサイズを大きくすれば統計的検定によって棄却できてしまう

例2) 「対照治療での中央全生存期間24.5カ月に対して，試験治療では24.6カ月であるため，試験治療は有効である」と言い切ることが難しい

例3)
エフェクトサイズが大きい場合 ⇒ 必要サンプルサイズが小さくなる傾向
エフェクトサイズが小さい場合 ⇒ 必要サンプルサイズが大きくなる傾向

第II種の過誤を小さくするためには

- ①対立仮説で想定される治療効果（**エフェクトサイズ**）を大きくすることで対立仮説のもとでの検定統計量の分布を右側に移動させる
- ②標本サイズを大きくすることで「データのばらつき」を小さくする



単アーム試験の際の効果量を計算するのに必要な情報

	1標本（単アーム試験）	
	必要な情報	対応する検定
比率 (奏功割合等)	閾値（比率） 期待値（比率）	母比率の検定
平均値 (検査値等)	閾値（平均値） 期待値（平均値）	母平均の検定
生存曲線 (全生存期間等)	閾値（中央生存期間 or 年次生存率） 期待値（中央生存期間 or 年次生存率）	1標本ログランク検定 信頼区間

閾値：試験治療が上回りたい（否定したい）主要評価項目の値

期待値：試験治療によって期待される主要評価項目の値

例）期待値以上の試験結果が得られた際，下記の帰無仮説を棄却し，対立仮説を採択することができる。

帰無仮説：「試験治療による真の治療効果が閾値である」

対立仮説：「試験治療による真の治療効果は閾値を上回る」

単アーム試験の際の効果量を計算するのに必要な情報（対応あり）

	対応のある場合	
	必要な情報	対応する検定
比率	試験治療が有効で対照治療が無効の比率 試験治療が無効で対照治療が有効の比率	McNemar検定
平均値	被験者ごとのアウトカムの差の平均値 被験者ごとのアウトカムの差の標準偏差	対応ありのt検定

例）対応ありのMcNemar検定の際，下記の帰無仮説に対して症例設計が行われる。

帰無仮説「試験治療が有効で対照治療が無効の比率と
試験治療が無効で対照治療が有効の比率の差が0である」

対立仮説「試験治療が有効で対照治療が無効の比率と
試験治療が無効で対照治療が有効の比率の差が0でない」

例）対応ありのMcNemar検定の際，下記の帰無仮説に対して症例設計が行われる。

帰無仮説「被験者ごとのアウトカムの差の平均値が0である」

対立仮説「被験者ごとのアウトカムの差の平均値が0でない」

症例数決定の基本

ランダム化比較試験の際の効果量を計算するのに必要な情報

	2標本（ランダム化比較試験）	
	必要な情報	対応する検定
比率 (奏功割合等)	対照群での比率，試験治療での比率 (2群の比率⇒オッズ比でも可能)	カイ二乗検定 Fisherの正確確率検定
平均値 (検査値等)	対照治療での平均値，試験治療での平均値 共通の標準偏差	2標本t検定
生存曲線 (全生存期間等)	対称治療の中央生存期間 or 年次生存率 試験治療の中央生存期間 or 年次生存率 (2群の生存率⇒ハザード比でも可能) 登録期間および追跡期間	1標本ログラंक検定 信頼区間

主要評価項目が比率の2標本(ランダム化比較試験)の症例数設計

下記に対して症例数設計を行う

帰無仮説：「試験治療と対照治療の真のオッズ比は1である」

主要評価項目が平均値の2標本(ランダム化比較試験)の症例数設計

下記に対して症例数設計を行う

帰無仮説：「試験治療と対照治療の真の平均値の差は0である」

主要評価項目が生存曲線の2標本(ランダム化比較試験)の症例数設計

下記に対して症例数設計を行う

帰無仮説：「試験治療と対照治療の真のハザード比は1である」

症例数設計を実施するための必要な情報がない場合は？

先行研究（論文）や当院のデータがなく，症例数設計を実施することができない場合は小規模な**パイロット試験**を実施する

-パイロット試験では症例数が少ないため，データがばらつかないような手順により注意を払う必要がある

-パイロット試験では複数の主要評価項目となる候補を副次評価項目としてデータを集め，次の試験へつなげることできる

連続変数の平均値や中央値は分かるが，標準偏差が分からないときは，連続変数に2値化等の変換を適用してカイ自乗検定によりサンプルサイズを見積もる

Step1 主要評価項目を考える

クリニカルクエスチョンに対応する適切な主要評価項目を選定する必要がある

例)

抗癌剤に対する第II相試験の場合には、奏功割合、生存期間

抗癌剤に対する第III相試験の場合には、生存期間

以下の点に留意する必要がある。

- (1) クリニカルクエスチョンに應えるものか（主要評価項目の妥当性）
- (2) 第三者的にも適切であるか（アウトカムの信頼性）
- (3) アウトカムを取得できるか（実施可能性）

PICO 【胃癌に対する新しい手術が既存手術より根治切除率が高い試験の例】

P: patient (誰に, どのような患者に)

【胃癌患者に対して】

I : intervention (介入 (試験治療) を行うと)

【新たな手術法を施行すると】

C : comparison (何と比較して)

【従来の手術法に比べて】

O : outcome (どのような結果になるのか)

【根治切除が期待できる】

Step2 臨床的な仮説を設定する

ここで、エフェクトサイズ（臨床的に有効であると判断できる治療効果）を設定しなければならない。

⇒「臨床的有效性を統計的有效性のもとで検証する」観点から設定することが重要

- 結果として統計的に有効ならば臨床的有效とは限らない
（当該差に臨床的意味がなければ小さい差があると主張しても意味がない）
- 強気な（臨床的効果の見込みとして無理のある）エフェクトサイズを避けた方がよい
（十分な検出力が保証されないことからnegative studyになりやすくなる）

Step2 臨床的な仮説を設定する

単アーム試験の場合

エフェクトサイズを設定するうえで、ヒストリカル・コントロール（閾値）を決めなければならない。

主要評価項目が奏効率の場合

試験治療が「臨床的に有効であると判断されるための奏効率」がヒストリカル・コントロールになる。このとき、試験治療がヒストリカル・コントロールを統計学的に上回ることを期待するため、片側対立仮説がよく用いられる。

主要評価項目が奏効率の場合の例

閾値奏効率を**40%**に設定するとき、 p 値が有意水準 α を下回ったときに「試験治療における奏功割合は、**40%**を上回る」と判断される。

上記の40%は試験実施前に臨床的な観点より定める必要がある

Step2 臨床的な仮説を設定する

ランダム化比較試験の場合

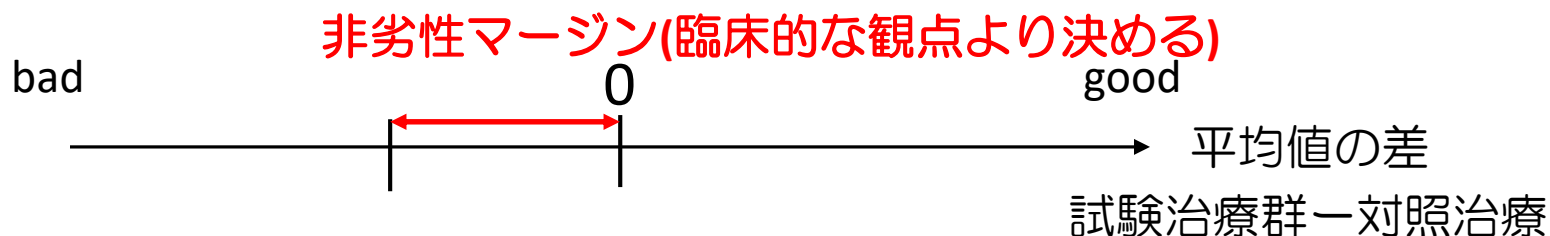
エフェクトサイズを設定する前に、臨床試験の形式を決めなければならない

優越性試験

試験群が対照群に比べて主要評価項目において勝ることを統計学的に検証するデザイン

非劣性試験

試験治療が対称治療に比べて主要評価項目以外でのベネフィット（安全性、治療方法の簡便性、医療経済的なベネフィット）が明らかな場合に、試験治療の主要評価項目でのアウトカムが対称治療に比べて劣らないことを示す試験



同等性試験

試験群が対照群の主要評価項目でのアウトカムが同等であることを検証する試験

例) ジェネリック医薬品の試験等

Step3 統計的な判断基準を設定する；有意水準および検出力

症例数設計には**有意水準**および**検出力**を設定しなければならない
ICH E9 「臨床試験のための統計的原則」等や実際には下記のように定めることが多い

第I種の過誤： $\alpha = 0.05$ and 第II種の過誤： $\beta = 0.20$ or 0.10

症例数設計には**有意水準**および**検出力**を設定しなければならない

例) 抗癌剤試験における単アームでの臨床試験

第I種の過誤： $\alpha = 0.10$ and $\beta = 0.10$ (検出力0.90)を用いる場合もある

留意点

エフェクトサイズを強気に責めて、サンプルサイズを見積もるより、第I種の過誤を上げて、サンプルサイズを減らした方がまだよい
理由はエフェクトサイズは評価することが難しいが、第I種の過誤は統計的観点より評価することができるからである

論文での記述例

Phase II study of trastuzumab in combination with S-1 plus cisplatin in HER2-positive gastric cancer (HERBIS-1)

The required sample size was estimated based on a threshold RR of 35% and an expected RR of 50%, 80% power, and an alpha value of 0.1 (one-sided) using the binomial test. Given 2% of ineligible patients, the target sample size was determined to be at least 50 patients.

(Kurokawa et al., British Journal of Cancer, 110, 1163-68, 2014)

■ 上記記載の要約

試験デザイン：Phase II Study (単アーム)

主要評価項目：奏効割合 (RR: Response Rate)

閾値奏効率：35%， 期待奏効率：50%

臨床的判断

第1種の過誤(type I error, 有意水準)： $\alpha = 0.1$

1－第2種の過誤(1-type II error, 検出力)： $1 - \beta = 0.8$ (80%)

統計的判断
必要症例数 = 50

H_0 : 奏効割合は35%である. H_1 : 奏効割合は35%を上回る.

試験後のはなし

検定により評価

H_0 が正しい [H_1 : 偽]

第1種の過誤 α
0.10 (片側)

正しい判断 $1-\alpha$

試験前のはなし

標本サイズで規定

H_1 が正しい [H_0 : 偽]

検出力 $1-\beta$
0.80

第2種の過誤 β

試験後の評価

試験の結果, 奏効率が50%のとき,
有意水準 α のもとで有意になる.

試験計画

必要症例数 = 50

■ 閾値(帰無仮説)には根拠が必要

単群試験では、閾値に対して臨床的な根拠が必要であり、統計解析は閾値よりもポジティブな結果か否かが評価される。ランダム化比較試験では、帰無仮説は「差がない(or $HR=1$)」になる。

HERBIS-1では、ToGA試験(Y-J Bang, et al. *The Lancet*, 28, 687 - 697, 2010.)の結果を参考に行っている。ToGA試験では、化学療法群(5-FU /CDDP あるいはCapecitabine /CDDP)での**奏効率が35%**であった。

■ 期待値(対立仮説)には実現可能性及び臨床的有効性に基づいて設定

期待値は、臨床的に合意されるような有効性(エフェクトサイズ)として設定される。したがって、期待値が本臨床試験のエンドポイントの目標値となる。

HERBIS-1では、ToGA試験では、(1) Trastuzumab併用群の**奏効率が47%**であること、(2) HER2による選択基準(IHC 3+ or FISH positive + IHC 2+)、(3) 本邦の状況を鑑みて、閾値奏効率を50%と設定している。

サンプルサイズを最小にして、パワーを最大にする方法

質的変数ではなく連続変数を用いる

アウトカムとして連続変数を用いることができれば、質的変数を用いる場合よりもサンプルサイズは少なくて済む

例)

6ヶ月間栄養サプリーを投与された高齢者介護施設入所者を対象に以下を比較する臨床試験を想定

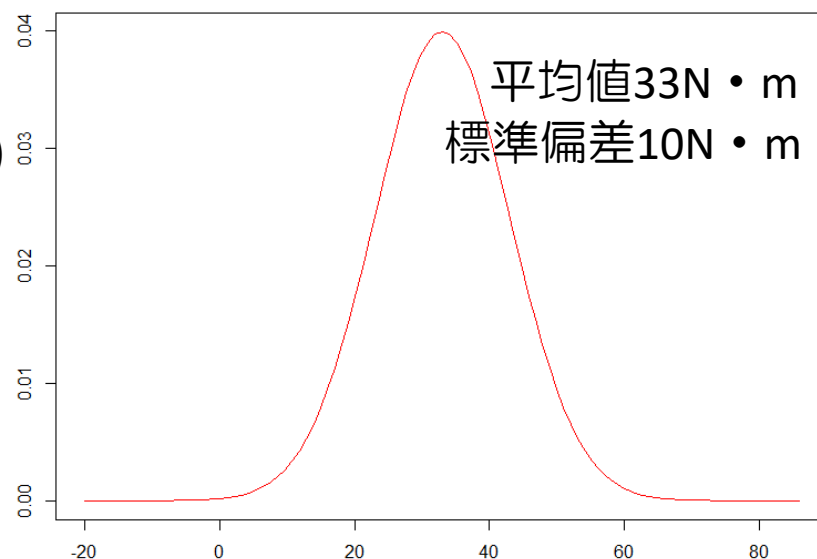
介入群：栄養サプリー投与群

対照群：プラセボ投与群

分かっている情報

- (a) 高齢者の10%が筋力が非常に弱い(<20N・m)
- (b) 6ヶ月間栄養サプリーを投与すると
5N・mの平均筋力の増加が期待できる
- (c) 非常に筋力の弱い高齢者が5%減少する
((b)によって)

四頭筋筋肉（最大トルク値）



サンプルサイズを減らすために

サンプルサイズを最小にして、パワーを最大にする方法

質的変数を用いた場合

(1)帰無仮説：介入群と対照群における筋力の非常に弱い被験者の割合は同じである

対立仮説：介入群と対照群における筋力の非常に弱い被験者の割合は異なる

(2)介入群における筋力の非常に弱い被験者の割合: 0.10

対照群における筋力の非常に弱い被験者の割合: 0.05

(3) $\alpha=0.05$, $\beta=0.2$ とすると必要サンプルサイズは**各群473例**

量的変数を用いた場合

(1)帰無仮説：介入群と対照群における筋力の非常に弱い被験者の平均四頭筋筋力は同じである

対立仮説：介入群と対照群における筋力の非常に弱い被験者の平均四頭筋筋力は異なる

(2)期待される平均値の差：5N・m 四頭筋筋力の標準偏差：10 N・m

(3) $\alpha=0.05$, $\beta=0.2$ とすると必要サンプルサイズは**各群64例**

サンプルサイズを最小にして、パワーを最大にする方法

変化量の使用

最終観測時点の量的アウトカムを用いるよりも、ベースライン時点と最終観測時点の量的アウトカムの変化量を用いた方がサンプルサイズが減少する傾向にある

例)

サルブタノールによる治療がイプラトロピウムによる治療と比較して、喘息患者における1秒肺活量(FEV1)を200ml改善できるか

介入群：サルブタノール投与群

対照群：イプラトロピウム投与群

情報

- (a) 治療2週間後の平均値の差の期待値0.2 L
- (b) 治療2週間後のFEV1の標準偏差 1.0 L
- (c) 治療2週間後のFEV1の変化量の平均値の差の期待値0.2L
- (d) 治療2週間後のFEV1の変化量の標準偏差0.25

サンプルサイズを減らすために

サンプルサイズを最小にして、パワーを最大にする方法

治療2週間後の値を用いた場合

(1) 帰無仮説：介入群と対照群におけるFEV1の平均値は同じである

対立仮説：介入群と対照群におけるFEV1の平均値は異なる

(2) $\alpha=0.05$, $\beta=0.2$ としてt検定を用いると必要サンプルサイズは**各群394例**

ベースラインと治療2週間後の変化量を用いた場合

(1) 帰無仮説：介入群と対照群におけるFEV1の変化量の平均値は同じである

対立仮説：介入群と対照群におけるFEV1の変化量の平均値は異なる

(3) $\alpha=0.05$, $\beta=0.2$ とすると必要サンプルサイズは**各群26例**

変化量を用いた方がサンプルサイズが減少する傾向にある理由

ベースライン時に各被験者のFEV1が揃っていない場合が多いため、
変化量を用いることで必要サンプルサイズを減らせる場合がある

サンプルサイズを最小にして、パワーを最大にする方法

不均等サンプル比の使用

検定の検出力は一般的に各群の人数が均等な場合に最大となるが、
「一方の群の対象数」の方が「他の群の対象数」よりも集めやすい場合、
サンプルサイズを不均等することがある

一般に一方の群のサンプルサイズを2倍や3倍にすると検出力の増加を見込むことが可能となる。

例) ケースコントロール研究

ケースコントロール研究で患者（ケース）の数は少ないが、コントロールは
たくさん集められる

サンプルサイズを減らすために

サンプルサイズを最小にして、パワーを最大にする方法

問題)

家庭内で使用される殺虫剤が再生不良性貧血のリスクファクターとなるかどうかを研究する場合を考える。ケース：コントロール比が1:1のときに必要なケース数は25例であったが、18例程度しか集められなさそうとする。この場合どのようにすればよいか

解答)

ケース1名につき、複数のコントロールを設定することを考える。
コントロールとなる、白血病でない子供をたくさん集められとして、
ケース1名につき3名のコントロールを収集する場合、必要なケース数は
17名となる。

サンプルサイズを最小にして、パワーを最大にする方法

より頻度の高いアウトカムを用いる

アウトカムが質的変数である場合、アウトカムの頻度が高いほど検出力が高くなる傾向にあるため、そのようなアウトカムに変更した方がよい

アウトカムの頻度を上げるには、対象とするアウトカムのリスクが大きい人々を対象に選ぶことができれば理想的（乳癌の家族歴を有する女性等）だが、リサーチクエスチョンが変わってしまう可能性があるので注意が必要である

サンプルサイズを減らすために

サンプルサイズを最小にして、パワーを最大にする方法

問題)

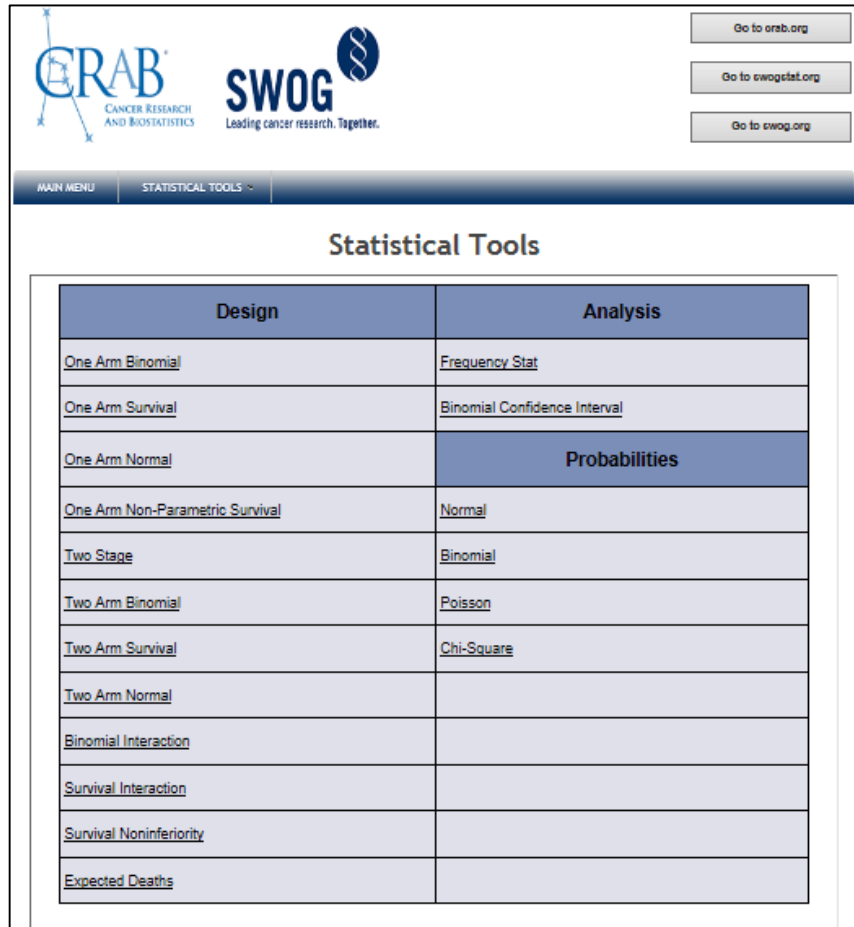
抗菌薬入りのうがい薬とプラセボの上気道感染症予防効果を比較したいとする。はじめ200名の大学生で研究を計画していたが、その集団ではフォローアップ期間3カ月中における上気道感染症の発生率はせいぜい約20%と予想されたため、サンプルサイズが不十分と考えられた。どのように修正すればよいか

解答)

- (1) 対象者を大学生よりもはるかに上気道感染症を起こしやすいと考えられる小児科の研修医に切り替える
- (2) 上気道感染症が起こりやすい冬に研究を実施する
- (3) 観察期間を6ヶ月から12カ月に延長する

これらはリサーチクエッションが少し変わるが、抗菌薬入りのうがい薬の効果をみるという観点から大きくはずれたものではない

Web Site



Design	Analysis
One Arm Binomial	Frequency Stat
One Arm Survival	Binomial Confidence Interval
One Arm Normal	Probabilities
One Arm Non-Parametric Survival	Normal
Two Stage	Binomial
Two Arm Binomial	Poisson
Two Arm Survival	Chi-Square
Two Arm Normal	
Binomial Interaction	
Survival Interaction	
Survival Noninferiority	
Expected Deaths	

<https://stattools.crab.org/>

単アーム試験

- 母比率 (One Arm Binomial)
- 母平均 (One Arm Normal)
- 生存時間 (One Arm Survival)
(One Arm Non-parametric / KM推定に基づく)
- 2段階デザイン(母比率のみ, Two Stage)

2アーム試験

- 母比率 (Two Arm Binomial)
- 母平均 (Two Arm Normal)
- 生存時間 (One Arm Survival)

その他の試験

- 母比率に対する交互作用 (Two Arm Binomial)
- 生存時間に対する交互作用 (Two Arm Normal)
- 生存時間での非劣勢 (Survival Noninferiority)
- 期待死亡数 (Expected Deaths)

The Chinese University of Hong Kong

Sample Size Estimation

Introduction: In general, sample size calculation is conducted through a pre-study power analysis. Its purpose is to select an appropriate sample size in achieving a desired power for correctly detection of a pre-specified clinical meaningful difference at a given level of significance. This study/tool will provide statistical procedures for determination of sample size required not only for testing *equality*, but also for testing *non-inferiority/superiority, equivalence* (similarity), some *survival analysis* and *phase II clinical trials designs* that are commonly employed at various phases of clinical development. As a handful tool, it is very useful for clinical scientists and biostatisticians in the pharmaceutical industry, regulatory agencies, academia, and other researchers.

This program helps users to determine sample sizes and confidence intervals for a wide range of statistical techniques including means, proportions, survival analysis, phase II clinical trials, epidemiological studies and some general cases).

DIRECTIONS:

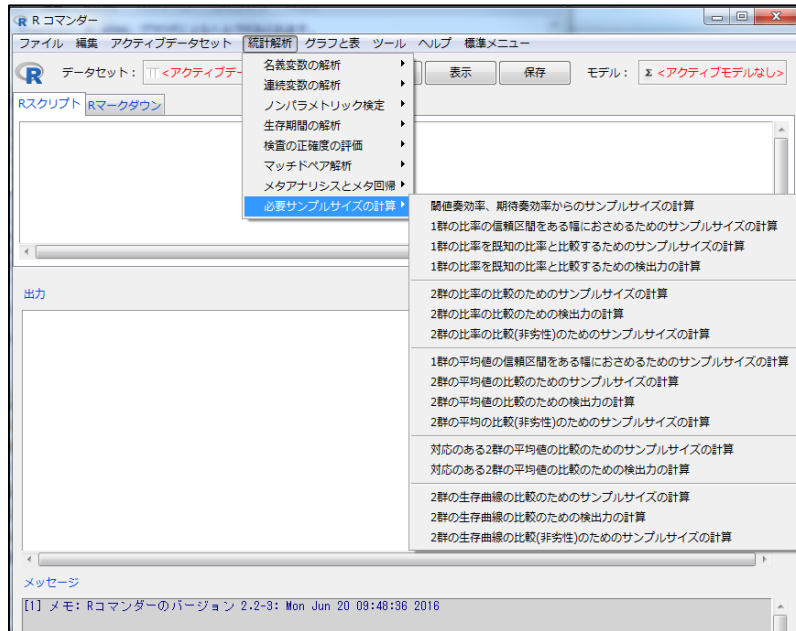
- Choose one of techniques from the following table
- From the chosen table, please select one of specified case that you are interested
- A new window will pop up and please read the instructions on how to use the calculation tools.
- If you don't know which program to use, please feel free to check : [User guidance, definition and terminology](#)

Means	Proportions	Survival Analysis	Phase II Clinical Trials	Epidemiological Studies	Confidence Intervals	Others
-----------------------	-----------------------------	-----------------------------------	--	---	--------------------------------------	------------------------



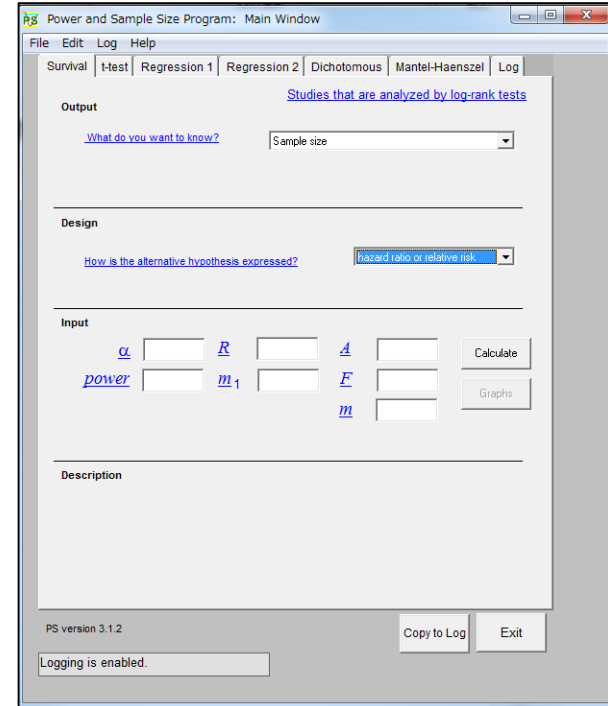
あまり知られていないか、
比較的豊富な試験デザインが含まれている。

<https://www2.ccrb.cuhk.edu.hk/stat/>



統計ソフトウェアR(統計学分野のデファクトスタンダード)をR言語が使えない利用者のために修正した統計ツール

<http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html>



ヴァンダービルド大学が開発した標本サイズを計算するためのソフトウェア

<http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize>

有償ソフトウェア



Obtaining a Sample Size in PASS

1 Select the Procedure

PASS Home

Select a Procedure

Two-Sample T-Tests Assuming Equal Variance (Enter Means)

2 Enter the Values

Design

Test: Two-Sample T-Tests Assuming Equal Variance (Enter Means)

Parameters:

- Mean: 0.00
- SD: 0.00
- Sample Size: 20
- Group Allocation: 10 to 10

Results:

- Power: 0.80
- Alpha: 0.05
- Standard Deviation: 0.00

View the Result

Two-Sample T-Tests Assuming Equal Variance

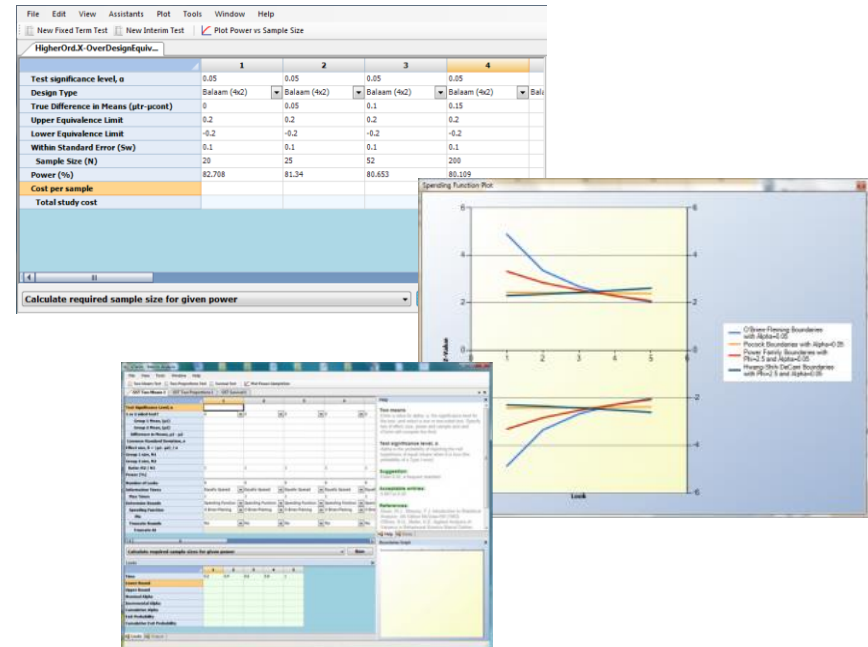
Results:

Parameter	Value
Mean	0.00
SD	0.00
Sample Size	20
Group Allocation	10 to 10
Power	0.80
Alpha	0.05
Standard Deviation	0.00

Two-Sample T-Tests Assuming Equal Variance (Enter Means)

Results:

Parameter	Value
Mean	0.00
SD	0.00
Sample Size	20
Group Allocation	10 to 10
Power	0.80
Alpha	0.05
Standard Deviation	0.00



臨床試験における必要症例数の設計方法について、必要症例数設計のための統計的背景とともに、症例数設計に必要な情報について触れた

留意点

「必要症例数の根拠は臨床的観点から決定されるものであり、統計的評価はそれを裏付けるものにすぎない」

これから

本セミナーでは基本的な内容について紹介させて頂きましたが、昨今Bayes流試験デザインや多変量解析を用いた症例数計算等様々なデザインがあるため、当該内容については臨床研究センターにお問合せ頂ければ幸いです。

Thank you for your kind attention

