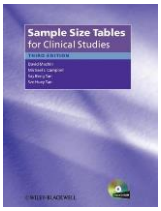
臨床試験

第6回サンプルサイズ設計
臨床統計学/臨床統計家育成コース

2017年5月17日
田中司朗

Clinical Biostatistics Course

アウトライン



- 仮説検定
- 基本公式$m = \frac{21}{\Delta^2}$の導出
  - 21はどこから来たのか
- 割付比$\varphi = n/m$が1でないとき
- 2値データ・生存時間データ
- 実例
  - クラスターランダム化試験・3群比較試験
  - 要因計画試験・非劣性試験
  - がん単群第II相試験（Simonの2段階デザイン）
  - 希少疾患の第II相試験

Machin, et al. Sample Size Tables for Clinical Studies, 3rd Edition. 2009

イカサマコイン

- コイン投げで, 表または裏が何度も連続で出たら, イカサマが疑われる
- では, 何回連続だったら, 怪しいと思い始めますか?

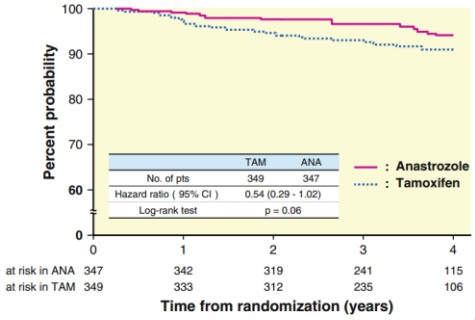
イカサマコインのロジック

- コイン投げで6回連続で表が出た
- 表が出る確率=1/2という仮説の下で
  - 6回連続で表の確率は$(1/2)^6$=0.0156
  - 6回連続で裏の確率は$(1/2)^6$=0.0156
- すなわち, このような極端なデータが得られる確率はp=0.0312と極めて低い
- よって, このコインにはイカサマがある

p値

- 帰無仮説の下で, データより極端な差が生じる確率
- p値が大きい = 当たり前のことが起きた
- p値が小さい = 極端なことが起きた
  = 帰無仮説が間違いかも

サンプルサイズが足りないと



Aihara, et al. Breast Cancer Res Treatment 2010

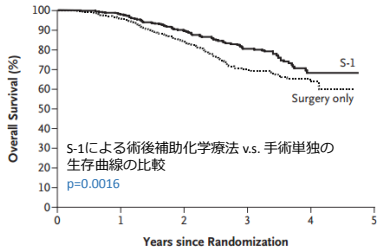
例1. 早期胃癌臨床試験

- 対象
  - ステージIIまたはIIIでD2廓清手術を受ける胃癌患者
- 治療
  - S-1による術後補助化学療法 v.s. 手術単独
  - 治療期間1年
- 主要エンドポイント: 全生存期間
- 有効性の主たる解析対象集団
  - 全ランダム化集団（適格性に関わらず）

Sakuramoto, et al. New Engl J Med 2007

7

例1. 早期胃癌臨床試験



No. at Risk					
S-1	529	515	370	196	46
Surgery only	530	504	352	163	40

Sakuramoto, et al. New Engl J Med 2007

8

仮説検定は二者択一の問題

- 「S-1の生存曲線は, 手術単独と差がない」
- 「S-1の生存曲線は, 手術単独を上回る」

9

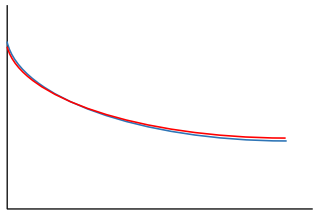
二つの仮説は不平等

- データに基づく論理では「差がないこと」は証明不可能
- 一方, 「生存曲線に差があること」は証明可能
- そこで, 前者を「帰無仮説」と呼び, データが帰無仮説に矛盾するかどうかに注目

10

仮説検定の手続き(1)
帰無仮説の設定

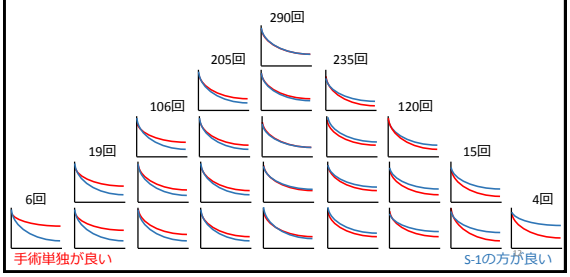
- 「S-1の生存曲線は, 手術単独と差がない」ときの生存曲線



11

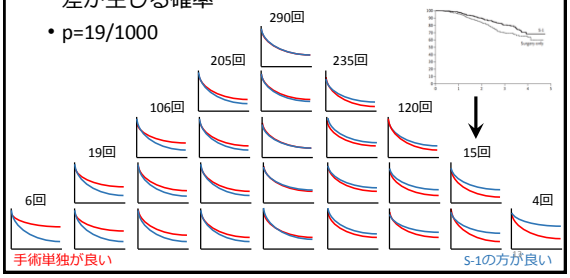
仮説検定の手続き(2)
帰無仮説の下で分布を調べる

- 対象者1059人の試験を1000回繰り返す



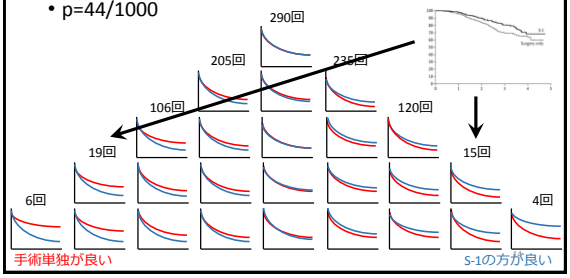
仮説検定の手続き(3)
分布とデータを比べp値を計算

- 観察されたS-1と手術単独との差以上に大きな差が生じる確率
- p=19/1000



仮説検定の手続き(3)
分布とデータを比べp値を計算

- 両側検定の場合, 左裾と右裾の両方を数える
- p=44/1000



仮説検定の二つのエラー

- αエラー
  - 帰無仮説が正しいのに, それを棄却する誤り
- βエラー
  - 帰無仮説が間違いなのに,それを保留する誤り
  - 検出力=1-β

	真実	
試験の結果	差がない（帰無仮説）	差がある（対立仮説）
p≥0.05（有意差なし）		βエラー
p<0.05（有意差あり）	αエラー	検出力 = 1-β

15

消費者リスクと生産者リスク

- 薬の有効性を検証する臨床試験では
  - αエラー = 無効な薬が市販される消費者リスク
  - βエラー = 有効な薬を見過ごす生産者リスク
  - 消費者リスクを5%以下にコントロールすべき
- 薬の安全性を調べる研究では
  - βエラー = 副作用を見過ごす消費者リスク
  - こちらを優先すべきかもしれない

16

サンプルサイズと
二つのエラーの関係

- サンプルサイズを大きくすれば, αエラーとβエラーは低くなる
- αエラーとβエラーはトレードオフの関係で, サンプルサイズが固定されたとき, 両方を同時に低くすることはできない
- 仮説検定では, 有意水準を通常5%に設定することで, αエラーを5%より低くコントロールする
- βエラーは, サンプルサイズを十分大きくすることでコントロールする

17

サンプルサイズに関する質問

- 臨床試験はさまざまな規模で行われている
- もっとも大きいのは何人?
- 小さいのは何人?

18

N=1,000,000の試験

Vitamin A supplementation every 6 months with retinol in 1 million pre-school children in north India: DEVTA, a cluster-randomised trial

Shally Awasthi, Richard Peto, Simon Read, Sarah Clark, Vinod Pande, Donald Bundy, and the DEVTA (Deworming and Enhanced Vitamin A) team

Summary

Background In north India, vitamin A deficiency (retinol <0·70 μmol/L) is common in pre-school children and 2–3% die at ages 1·0–6·0 years. We aimed to assess whether periodic vitamin A supplementation could reduce this mortality.

Methods Participants in this cluster-randomised trial were pre-school children in the defined catchment areas of 8338 state-staffed village child-care centres (under-5 population 1 million) in 72 administrative blocks. Groups of four neighbouring blocks (clusters) were cluster-randomly allocated in Oxford, UK, between 6-monthly vitamin A (retinol capsule of 200 000 IU retinyl acetate in oil, to be cut and dripped into the child's mouth every 6 months), albendazole (400 mg tablet every 6 months), both, or neither (open control). Analyses of retinol effects are by block (36 *vs* 36 clusters). The study spanned 5 calendar years, with 11 6-monthly mass-treatment days for all children then aged 6–72 months. Annually, one centre per block was randomly selected and visited by a study team 1–5 months after any trial vitamin A to sample blood (for retinol assay, technically reliable only after mid-study), examine eyes, and interview caregivers. Separately, all 8338 centres were visited every 6 months to monitor pre-school deaths (100 000 visits, 25 000 deaths at ages 1·0–6·0 years [the primary outcome]). This trial is registered at ClinicalTrials.gov, NCT00222547.

Awasthi et al. Lancet 2013

N=4の試験

【臨床成績】

1. 国内の医師主導治験

国内の脂肪萎縮症患者4例を対象に，本剤（0.01～0.08mg/kg）を1日1回5ヵ月間連日皮下投与したときのHbA1c（JDS値），トリグリセライドの経時変化を表2に示す。HbA1cは投与前に比べ，すべての症例で低下した。トリグリセライドも投与前に比べ，正常値まで低下したが，症例No. 4では副腎皮質ステロイド投与により一時的に上昇した。なお，症例No. 3は投与前後ともに正常値であった。また，糖尿病治療薬及び（又は）高脂血症治療薬が本剤投与開始前から投与された3例の患者では，投与開始2ヵ月以内にそれら治療薬の投与が中止された[1]。

表2　国内の医師主導治験でのHbA1c及びトリグリセライドの経時変化

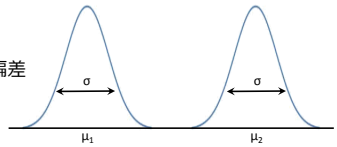
症例No.	年齢	HbA1c（%）（JDS値）				トリグリセライド（mg/dL）			
		投与前	3ヵ月	4ヵ月	5ヵ月	投与前	3ヵ月	4ヵ月	5ヵ月
1	18歳	8.6	5.5	4.8	4.8	210	55	55	62
2	23歳	7.7	5.6	5.9	6.4	246	51	144	204
3	11歳	5.8[注1]	5.1	5.1	5.4	59	46	60	77
4	6歳	5.8	5.0	5.1	5.2	180	83	131	382

注1：登録時にはHbA1c＝6.1

Ito et al. Clin Eval 2014

サンプルサイズの基本公式

- 個人の測定値は正規分布に従うと仮定
- 群1
  - 平均μ_1
  - 標準偏差σ
- 群2
  - 平均μ_2
  - 標準偏差σ
- 効果の指標
  - 平均の差/標準偏差
  - $\Delta = \frac{\mu_1 - \mu_2}{\sigma}$



サンプルサイズの基本公式

- $m = \frac{21}{\Delta^2}$
  - 2群の人数が等しいときの1群の必要サンプルサイズ
  - 検出力90%
- 必要サンプルサイズは，効果の二乗に反比例
- 試験の計画段階で効果を設定する必要がある

サンプルサイズの基本公式

- $m = \frac{21}{\Delta^2}$
- たとえば糖尿病治療薬
  - HbA1Cの標準偏差は1%
  - A1C改善効果が0.5%だとm=84
  - A1C改善効果が1%だとm=21
- 実際の臨床試験では?

デベルザ錠20mg

(1)プラセボ対照二重盲検比較試験[注2]

食事療法・運動療法にて血糖コントロールが不十分な2型糖尿病患者を対象に，プラセボ，トホグリフロジン10mg，20mg又は40mgのいずれかを1日1回24週間経口投与した。24週時（最終評価時）における結果は次表のとおりであった。HbA1cの投与前からの変化量において，トホグリフロジン群はプラセボ群と比べ有意な低下が認められた。

表　プラセボ対照二重盲検比較試験(24週時)の結果

投与群及び投与前HbA1c(NGSP)[#1](%)	主要評価項目		副次的評価項目	
	HbA1c(NGSP)[#1](%)		空腹時血糖[#1](mg/dL)	食後2時間血糖[#1](mg/dL)
	投与前からの変化量	プラセボとの差	投与前からの変化量	投与前からの変化量
プラセボ 8.41±0.76 n=56	-0.028±0.083	—	-8.561±2.278	-33.5±17.6 n=48
トホグリフロジン10mg 8.43±0.75 n=57	-0.797±0.083	-0.769* [-1.000, -0.538]	-31.868±2.338	-63.5±19.0 n=53
トホグリフロジン20mg 8.34±0.83 n=58	-1.017±0.082	-0.990* [-1.220, -0.759]	-35.899±2.337	-71.0±13.7 n=56
トホグリフロジン40mg 8.37±0.72 n=58	-0.870±0.082	-0.842* [-1.072, -0.612]	-32.327±2.337	-60.2±17.2 n=53

#1：平均値±標準偏差

#2：LOCF(Last observation carried forward)法を適用した。

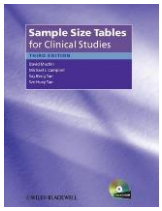
最小二乗平均±標準誤差，[95%信頼区間]

*P<0.0001(共分散分析)

これから解説を行う範囲

- 2群比較のランダム化臨床試験を想定
  - 群1の人数　：m
  - 群2の人数　：n
  - 割付比　：$\varphi = n/m$
  - 全体の人数　：$N = m + n$
- アウトカムの型
  - 連続データ
  - 2値データ
  - 生存時間データ
- あらゆる状況をカバーできるわけではないが，特殊なケースについて実例を通じて学ぶ

アウトライン

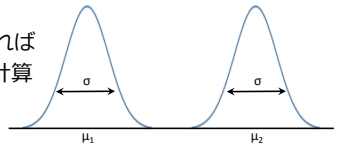


- 仮説検定
- 基本公式$m = \frac{21}{\Delta^2}$の導出
  - 21はどこから来たのか
- 割付比$\varphi = n/m$が1でないとき
- 2値データ・生存時間データ
- 実例
  - クラスターランダム化試験・3群比較試験
  - 要因計画試験・非劣性試験
  - がん単群第II相試験（Simonの2段階デザイン）
  - 希少疾患の第II相試験

Machin, et al. Sample Size Tables for Clinical Studies, 3rd Edition. 2009

正規分布の検定

- 個人の測定値は正規分布に従うと仮定
- さらに以下の二つを仮定
  - 標準偏差σが共通で既知
  - $m = n$
- 平均の差の推定値について検定を行いたい
  - $\hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1$
- 確率分布が分かれば p値や検出力を計算できる



正規分布の検定

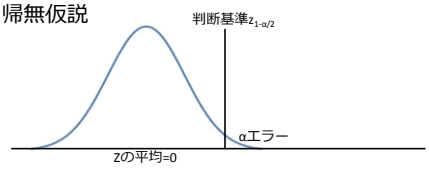
- 平均の差の推定値$\hat{\delta}$は, 帰無仮説が正しければ, 正規分布$\mathrm{N}(0, \mathrm{SE}^2)$に従うことが示される
  - $\mathrm{SE} = \sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}$
    $= \sigma\sqrt{\frac{2}{m}}$
- 正規分布の検定では, Zが標準正規分布に従うことを利用する
  - $Z = \frac{\hat{\delta}}{\mathrm{SE}}$

正規分布の検定

- 検定統計量Zを, 標準正規分布のパーセント点$z_{1-\alpha/2}$と比較することで, 有意水準αの両側検定を行うことができる
  - $Z \geq z_{1-\alpha/2}$
  - $Z \leq -z_{1-\alpha/2}$
- 対立仮説$\delta > 0$の下では
  - $\mathrm{E}(Z) = \frac{\delta}{\mathrm{SE}}$
    $= \frac{\delta}{\sigma\sqrt{2/m}}$
    $= \frac{\Delta}{\sqrt{2/m}}$

検出力（片側のみ図示）

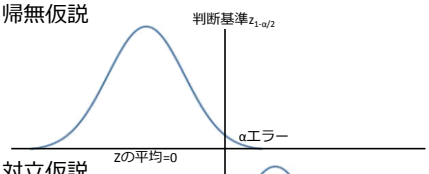
- 帰無仮説



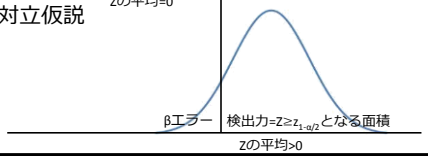
- 一般に正規分布のパーセント点をz_xで表す
  - 片側α=0.025に対応する点　: $z_{0.975} = 1.96$
  - 1-β=0.9に対応する点　: $z_{0.9} = 1.28$

検出力（片側のみ図示）

- 帰無仮説

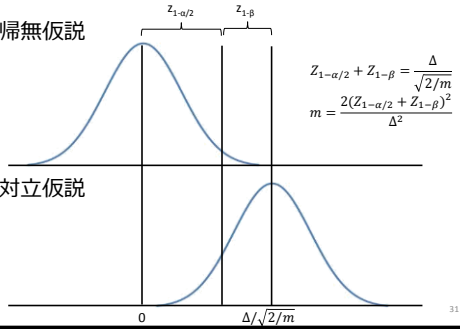


- 対立仮説



検出力（片側のみ図示）

- 帰無仮説



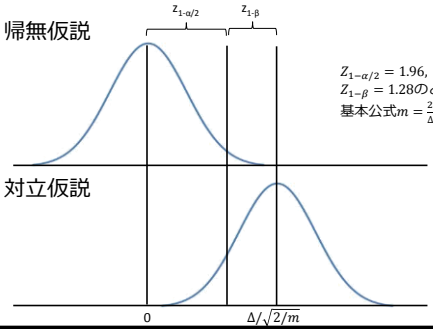
$$Z_{1-\alpha/2} + Z_{1-\beta} = \frac{\Delta}{\sqrt{2/m}}$$
$$m = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\Delta^2}$$

- 対立仮説

31

検出力（片側のみ図示）

- 帰無仮説



$Z_{1-\alpha/2} = 1.96$,
$Z_{1-\beta} = 1.28$のとき
基本公式$m = \frac{21}{\Delta^2}$

- 対立仮説

32

αエラー, βエラーごとの参照値

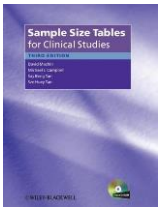
Table 2.3 Values of $\theta(\alpha, \beta) = (z_{1-\alpha/2} + z_{1-\beta})^2$.

Two-sided	One-sided β				
α	0.05	0.10	0.15	0.20	0.50
0.001	24.358	20.904	18.723	17.075	10.828
0.005	19.819	16.717	14.772	13.313	7.879
0.01	17.814	14.879	13.048	11.679	6.635
0.02	15.770	13.017	11.308	10.036	5.412
0.05	12.995	10.507	8.978	7.849	3.841
0.1	10.822	8.564	7.189	6.183	2.706
0.2	8.564	6.569	5.373	4.508	1.642
0.4	6.183	4.508	3.527	2.833	0.708

Machin, et al. Sample Size Tables for Clinical Studies, 3rd Edition. 2009

33

アウトライン



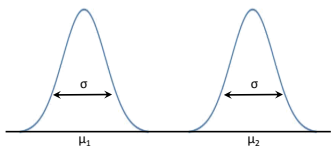
- 仮説検定
- 基本公式$m = \frac{21}{\Delta^2}$の導出
  - 21はどこから来たのか
- 割付比$\varphi = n/m$が1でないとき
- 2値データ・生存時間データ
- 実例
  - クラスターランダム化試験・3群比較試験
  - 要因計画試験・非劣性試験
  - がん単群第II相試験（Simonの2段階デザイン）
  - 希少疾患の第II相試験

Machin, et al. Sample Size Tables for Clinical Studies, 3rd Edition. 2009

34

割付比$\varphi = n/m$が1でないとき

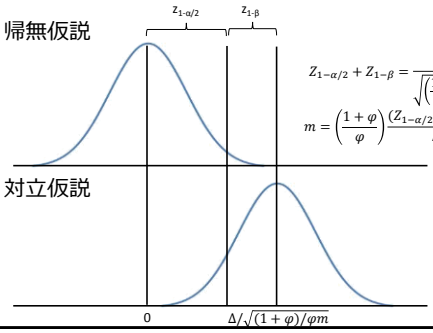
- 平均の差の推定値$\hat{\delta}$は, 帰無仮説が正しければ, 正規分布$\mathrm{N}(0, \mathrm{SE}^2)$に従うことが示される
  - $\mathrm{SE} = \sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}$
    $= \sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{\varphi m}}$
    $= \sigma\sqrt{\left(\frac{1+\varphi}{\varphi}\right)\frac{1}{m}}$



35

割付比$\varphi = n/m$が1でないとき

- 帰無仮説

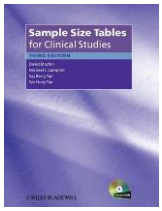


$$Z_{1-\alpha/2} + Z_{1-\beta} = \frac{\Delta}{\sqrt{\left(\frac{1+\varphi}{\varphi}\right)\frac{1}{m}}}$$
$$m = \left(\frac{1+\varphi}{\varphi}\right)\frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\Delta^2}$$

- 対立仮説

36

アウトライン



- 仮説検定
- 基本公式$m = \frac{21}{\Delta^2}$の導出
  - 21はどこから来たのか
- 割付比$\varphi = n/m$が1でないとき
- 2値データ・生存時間データ
- 実例
  - クラスターランダム化試験・3群比較試験
  - 要因計画試験・非劣性試験
  - がん単群第II相試験（Simonの2段階デザイン）
  - 希少疾患の第II相試験

Machin, et al. Sample Size Tables for Clinical Studies, 3rd Edition. 2009

正確検定と近似検定

- 統計学では, 確率を正確に計算することもあれば, 近似解を求めることもある
- 標準偏差σが共通で既知のとき
  - 正規分布の検定は正確
- 標準偏差σが共通で未知のとき
  - 正規分布の検定は, サンプルサイズが大きいときにだけ正しい（近似検定）
  - t検定は正確
  - t検定のサンプルサイズの公式
    - $m = \left(\frac{1+\varphi}{\varphi}\right)\frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\Delta^2} + \frac{Z_{1-\alpha/2}{}^2}{2(1+\varphi)}$

2値データ・生存時間データ

- 正確な検定は計算が難しい
- たとえば, Fisherの正確検定は2値データのときの正確検定だが, サンプルサイズの公式はない
- そこでどうするか
  - 近似検定の公式の利用（たとえばχ^2検定）
  - 乱数を用いたシミュレーション

2値データにおける正規近似

- 群1
  - 確率π_1, 人数mの二項分布を仮定
- 群2
  - 確率π_2, 人数nの二項分布を仮定
- 効果の指標
  - 対数オッズ比
  - $\log(\mathrm{OR}) = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$

2値データにおける正規近似

	イベント発生		
	あり	なし	合計
群1	A	C	m
群2	B	D	n

- オッズ比 $= \frac{A/C}{B/D} = \frac{AD}{BC}$
- 対数オッズ比の標準誤差の公式
  - $\mathrm{SE} = \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}$

2値データにおける正規近似

- π_1とπ_2の平均を$\bar{\pi}$で表す
  - $\bar{\pi} = (\pi_1 + \varphi\pi_2)/(1+\varphi)$
- $\bar{\pi}$がπ_1とπ_2に近い値であれば, $m = n$のとき

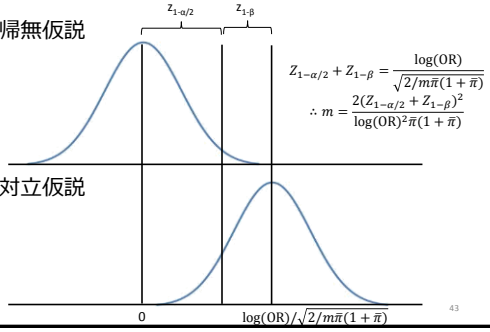
$$\sqrt{\frac{1}{\mathrm{E}(A)}+\frac{1}{\mathrm{E}(B)}+\frac{1}{\mathrm{E}(C)}+\frac{1}{\mathrm{E}(D)}} = \sqrt{\frac{1}{m\pi_1}+\frac{1}{n\pi_2}+\frac{1}{m(1+\pi_1)}+\frac{1}{n(1+\pi_2)}}$$

$$\approx \sqrt{\frac{2}{m\bar{\pi}(1+\bar{\pi})}}$$

- 対立仮説の下で, 検定統計量Zの平均は
  - $\mathrm{E}(Z) \approx \frac{\log(\mathrm{OR})}{\sqrt{2/m\bar{\pi}(1+\bar{\pi})}}$

検出力（片側のみ図示）

- 帰無仮説



- 対立仮説

2値データのときの公式

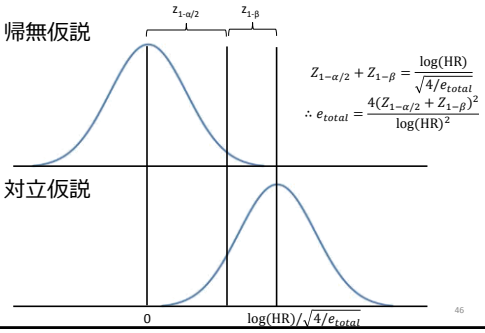
- $m=\left(\frac{1+\varphi}{\varphi}\right)\frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\log(\mathrm{OR})^2\bar{\pi}\,(1-\bar{\pi})}$
  - 2×2分割表のオッズ比=1の検定（χ^2検定）から導出
  - $\bar{\pi}=(\pi_1+\varphi\pi_2)/(1+\varphi)$
- 確率π_1とπ_2を見積もる必要がある

生存時間データにおける正規近似

- 対数ハザード比の推定値が, サンプルサイズが大きいとき, 正規分布$\mathrm{N}\{\log(\mathrm{HR}),4/e\}$に従うという性質を用いる
  - 効果の指標
    - 対数ハザード比log(HR)
  - 2群合わせたイベント数: e_{total}
- 対立仮説の下で, 検定統計量Zの平均は
  - $\mathrm{E}(Z)\approx\frac{\log(\mathrm{HR})}{\sqrt{4/e_{total}}}$

検出力（片側のみ図示）

- 帰無仮説



- 対立仮説

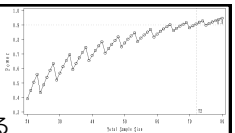
生存時間データのときの公式（Schoenfeldの公式）

- 群1の必要イベント数
  - $e=\frac{2(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\varphi\log(\mathrm{HR})^2}$
- 群1の必要サンプルサイズ
  - $m=e/\pi_1$
  - イベント発生確率π_1を見積もって割ればよい

イベント数について

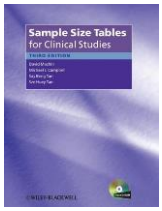
- 生存時間データでは, 検出力は人数ではなくてイベント数によって決まる
- 特に注意すべきこと
  - 見積りよりもイベント発生率が低くはないか
  - 死亡などの競合リスクは多くはないか
  - 追跡期間は十分か

計算精度について



- いずれの公式にも誤差がある
  - 正確検定ではなく近似検定であり，サンプルサイズが小さいとき近似精度が低い
  - 2値・生存時間データでは，イベント数が整数という制約があるので，設定したαエラー・βエラー（名義水準）に正確に一致するように，サンプルサイズを決めることができない
- いずれの公式も机上の計算に過ぎない
  - 臨床試験を実施すると様々な問題が発生
- これらを考慮するため，予定登録数は必要サンプルサイズより多く設定することが普通

アウトライン



- 仮説検定
- 基本公式$m = \frac{21}{\Delta^2}$の導出
  - 21はどこから来たのか
- 割付比$\varphi = n/m$が1でないとき
- 2値データ・生存時間データ
- 実例
  - クラスターランダム化試験・3群比較試験
  - 要因計画試験・非劣性試験
  - がん単群第II相試験（Simonの2段階デザイン）
  - 希少疾患の第II相試験

Machin, et al. Sample Size Tables for Clinical Studies, 3rd Edition. 2009

例1. クラスターランダム化試験

- 非精神病性の未治療単極性大うつ病患者に，治療を施設ごとにランダムに割り付けるクラスターランダム化非盲検臨床試験
- 対象
  - 非精神病性の未治療単極性大うつ病患者
- 治療
  - セルトラリンを25mg→50mgに漸増する群
  - セルトラリンを25mg→100mgに漸増する群
- アウトカム
  - 第1週～3週の自記式うつ尺度PHQ9の変化

Furukawa, et al. Trials 2011

例1. クラスターランダム化試験

- 個人ではなく，施設や地域などのクラスター単位で，介入をランダムに割り付けるデザイン
- サンプルサイズの計算や統計解析では，クラスター内相関を考慮しなければならない
  - $m = \left(\frac{1+\varphi}{\varphi}\right)\frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\Delta^2}$
  - $m_{cluster} = \{1 + (k-1)\rho\}m$
    - クラスター内相関係数: ρ
    - クラスター数: k

問題1. 公式を用いて必要サンプルサイズを求めよ

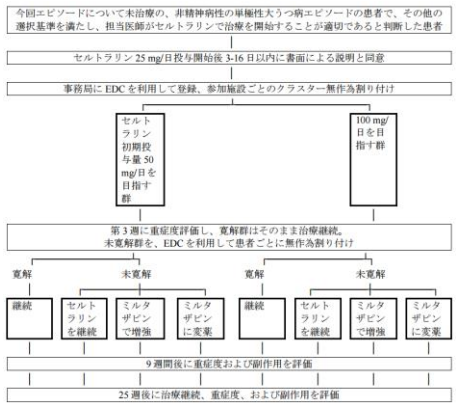
- $m = \frac{16}{\Delta^2}$
- $m_{cluster} = \{1 + (k-1)\rho\}m$
  - クラスター内相関係数: ρ
  - クラスター数: k

16. 統計学的事項

16.1. サンプルサイズの設定とその根拠

Step I のサンプルサイズ

ステップ I でのクラスター内相関係数を 0.05 と仮定し[45, 46]、有意水準（alpha）が 5%になるように設定し，検出力を 80%とするとエフェクトサイズ 0.2、つまり PHQ9 の得点で平均点で 1 点の差（SD=5）を検出しようとすると、30 のクラスター（施設数）で各[redacted]名をリクルートすることが必要、すなわち[redacted]人のサンプル数が必要である。



例2. 3群比較試験

- 非精神病性の未治療単極性大うつ病患者に, 治療を施設ごとにランダムに割り付けるクラスターランダム化非盲検臨床試験
- 対象
  - 非精神病性の未治療単極性大うつ病患者
- 治療
  - 第3週でセルトラリンを継続する群
  - 第3週でミルタザピンで増強する群
  - 第3週でミルタザピンに変薬する群
- アウトカム
  - 第3週～9週の自記式うつ尺度PHQ9の変化

Furukawa, et al. Trials 2011

55

問題2. 効果の大きさについて妥当な設定かどうか考察せよ

Step II のサンプルサイズ

Step II の臨床疑問が本研究の最も主な仮説である。第 3 週でセルトラリンを継続 vs ミルタザピンで増強 vs ミルタザピンに変薬の 3 群間での比較を行う。大うつ病の急性期治療の先行研究[47-49]によると、平均して PHQ9 得点は治療によって、ベースライン得点 15 点(SD=5)から治療後得点 10 点 (SD=6)へと減少し、差の得点の平均は 5 点(SD=5)であった。治療前後の減少量に 20%（1 点）の差が見られることを今回の研究で検出したい臨床的に有意義な効果の差として期待する。そこで、3 群間の比較において、比較する全体の有意水準（alpha）が 5%になるように設定し、検出力を 80%とすると、PHQ9 のベースラインとの差の群間差 1 点（SD=5）を検出するには、1 群 522 人、step II の全体（3 群）で 1566 人必要である。さらに、Step I での非寛解率 90%、脱落率 20%と想定すると、step I でエントリーに必要な人数が 2175 人となる。

PHQ9 の得点の減少について 1 点の差はエフェクトサイズにして 1/5=0.2 に相当する。本研究はアクティブな治療方法のあいだでの比較であること、抗うつ剤の対プラセボのエフェクトサイズが 0.31 であること[50]、コクランライブラリーに系統的レビューとして収載されたすべての健康介入の真のエフェクトサイズは 0.3-0.4 程度であると推定されること[51]を考慮すると、今回の研究で検出するに足る臨床的に有意義な差であると考えられる。また、エフェクトサイズの 0.2 は、NNT に換算すると、対照群での発生率が 50%程度のアウトカム（例えば、うつ病で言えばうつ病重症度半減で定義される「反応」）であれば 10%ポイント近くの差（NNT にして約 10）、対照群での発生率が 20%程度のアウトカム（例えば、うつ病の寛解）については 5%ポイント程度の差（NNT にして約 20）に相当し、この観点からも臨床的に有意義な差であると考えられる[52]。

なお、パイロット研究終了時にもサンプルサイズの再検討を行う。

割付比について

- 実薬低用量, 実薬高用量, プラセボの3群だと
- 割付比1:1:1
  - 合理的でもっとも一般的
- 割付比1:1:2
  - もっとも検出力が高い
    - 実薬低用量とプラセボ, 実薬高用量とプラセボの比較のみを想定
  - 安全性情報はあまり得られない
- 割付比2:2:1
  - プラセボが許容されづらい理由があれば, 選択肢の一つかもしれない

57

例3. 要因計画試験

- 漿膜浸潤胃癌で根治切除後の患者に, 以下の治療をランダムに割り付ける2×2要因計画試験
  - A群: UFT単独療法
  - B群: S-1単独療法
  - C群: パクリタキセル→UFT逐次療法
  - D群: パクリタキセル→ S-1逐次療法
- 二つの仮説
  - パクリタキセル→フッ化ピリミジン逐次療法（C群+D群）は, 単独療法（A群+B群）に優るかどうか
  - UFTベース補助化学療法（A群+B群）は, S-1補助化学療法（C群+D群）に比べ, 劣らないかどうか

Tsuburaya, et al. Lancet Oncol 2014

58

問題3. 表を参照して必要サンプルサイズを求めよ

2.5.3. 臨床的仮説と登録数設定根拠

標準治療群である UFT あるいは TS-1 による治療群（A+B 群）の 3 年無病生存割合を 40%～50%と仮定すると、A+B 群に対する Paclitaxel→UFT あるいは Paclitaxel →TS-1 による治療群（C+D 群）の治療効果がハザード比で 0.80（リスクリダクション 20%）あれば、臨床的にも充分有用で、標準としても妥当と考えられる。この時、C+D 群の 3 年無病生存割合は 48.1%～57.4%と予想され、優越性試験デザインとした場合、登録 3 年、追跡 3 年、αエラー = 5%（両側）、検出力 = 90%のもとで有意差を得るためには A+B 群および C+D 群それぞれで [redacted] 例が必要となる。

59

Table 8.2 (*continued*): Number of subjects for comparison of survival rates (Logrank test). Each cell gives the number of subjects for each group, m. Hence, the total sample size for the study is $N = 2m$.

	Two-sided $\alpha = 0.05$; Power $1 - \beta = 0.9$									
	First proportion, π_1									
π_2	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	332	–	–	–	–					
0.15	117	645	–	–	–					
0.2	67	198	947	–	–					
0.25	46	105	273	1222	–					
0.3	36	68	137	338	1464					
0.35	29	50	86	165	395	1666	–	–	–	–
0.4	25	39	61	101	189	441	1823	–	–	–
0.45	22	32	46	70	113	207	475	1938	–	–
0.5	20	28	38	53	77	122	220	499	2008	–
0.55	18	24	31	42	57	82	128	227	510	2030
0.6	17	21	26	34	44	61	86	133	230	512
0.65	16	20	24	28	35	46	62	87	132	229
0.7	15	17	21	24	31	36	47	63	85	130
0.75	14	16	19	21	26	30	38	48	60	83
0.8	14	15	18	21	24	27	31	38	46	58
0.85	13	14	17	17	21	24	26	30	35	44
0.9	14	14	15	16	19	20	22	26	28	34
0.95	12	13	14	15	18	19	20	22	24	30

2.5.3. 臨床的仮説と登録数設定根拠
標準治療群である UFT あるいは TS-1 による治療群（A+B 群）の 3 年無病生存割合を 40%～50%と仮定すると、A+B 群に対する Paclitaxel→UFT あるいは Paclitaxel →TS-1 による治療群（C+D 群）の治療効果がハザード比で 0.80（リスクリダクション 20%）あれば、臨床的にも充分有用で、標準としても妥当と考えられる。この時、C+D 群の 3 年無病生存割合は 48.1%～57.4%と予想され、優越性試験デザインとした場合、登録 3 年、追跡 3 年、αエラー = 5%（両側）、検出力 = 90%のもとで有意差を得るためには A+B 群および C+D 群それぞれで [redacted] 例が必要となる。

60

2.5.3. 臨床的仮説と登録数設定根拠

標準治療群である UFT あるいは TS-1 による治療群（A+B 群）の 3 年無病生存割合を 40%～50%と仮定すると、A+B 群に対する Paclitaxel→UFT あるいは Paclitaxel →TS-1 による治療群（C+D 群）の治療効果がハザード比で 0.80（リスクリダクション 20%）あれば、臨床的にも充分有用で、標準としても妥当と考えられる。この時、C+D 群の 3 年無病生存割合は 48.1%～57.4%と予想され、優越性試験デザインとした場合、登録 3 年、追跡 3 年、αエラー = 5%（両側）、検出力 = 90%のもとで有意差を得るためには A+B 群および C+D 群それぞれで 606～708 例が必要となる。上記、無病生存割合および治療効果の予測には不確実なところがあるため、再発イベント発現状況について下記に記すとおりにモニタリングを行う予定である。試験途中で実際に発現したイベント数と予想されたイベント数との間に大きな乖離が見られた場合には集積症例数の見直しを検討する。多少の不適格例の発生を考慮して本試験計画時における予定登録数を A、B、C、D 群各 370 例、計 1,480 例とした。また 1,480 例の集積もとでは、非劣性のマージンをハザード比で 1.33 とした場合、88%の検出力で A+C 群に対する B+D 群の非劣性を証明できると見積もられる。非劣性の検討に関してもαエラーは両側 5%とする。

標準治療群における3年無病生存割合	試験治療群における3年無病生存割合	A+B 群と C+D 群の3年無病生存割合差	ハザード比	リスクリダクション	A+B 群あるいは C+D 群の必要例数
40.0%	45.9%	5.9%	0.85	15%	1,308
40.0%	48.1%	8.1%	0.80	20%	606
40.0%	50.3%	10.3%	0.75	25%	436
50.0%	55.5%	5.5%	0.85	15%	1,124
50.0%	57.4%	7.4%	0.80	20%	708
50.0%	59.5%	9.5%	0.75	25%	372

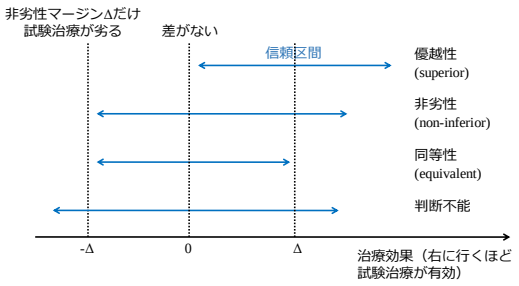
61

例4. 非劣性試験

- 非劣性であるためには, 一方の治療がわずかに劣っていても良いが, 臨床的に意味のある差があってはならない
- 非劣性試験では, 非劣性マージンという値により, どの程度劣っていてよいのかを設定する
  - SAMIT試験ではハザード比1.33
- 一般に, 効果はゼロ（ハザード比1）とし, 両側でなく片側検定で, 必要サンプルサイズを計算
  - $m = \left(\frac{1+\varphi}{\varphi}\right)\frac{(Z_{1-\alpha}+Z_{1-\beta})^2}{\Delta^2_{margin}}$

62

非劣性の判断基準



例5. がん単群第II相試験

- 対象
  - 根治照射不能なIIIA-IV期または術後再発の非小細胞肺癌（非扁平上皮癌）
  - 年齢20歳以上, ECOG PSは0～2
  - 前治療は1～2レジメン
- 治療
  - エルロチニブ（150mg/day連日）とベバシズマブ（15mg/kg^{3}週毎）の併用
- 主要エンドポイント
  - 腫瘍縮小効果（奏効率）

Tanaka, et al. Trials 2011

64

例5. がん単群第II相試験

- がん第II相試験の目的はスクリーニング
  - 多くの候補薬をランダム化第III相試験に進むかどうか選別
- 奏効率が閾値より有意に大きいかどうかで判断
  - 閾値=帰無仮説
    - これよりも低ければ価値がないと判断される奏効率の値
  - 期待値=対立仮説
    - 実際に試験治療を行ったときに期待できる奏効率の値
- 無効中止を目的とした中間解析を1回行う二段階デザインを用いるのが普通

65

Simonの2段階デザイン

- 第一ステージ
  - N_1人を登録し, R_1人以上で反応が見られたら, N_2人を追加
- 第二ステージ
  - $N=N_1+N_2$人の内, R人以上で反応が見られたら第III相へ

		最適デザイン（期待症例数を最小に）		ミニマックスデザイン（最大症例数を最小に）	
閾値	期待値	R_1/N_1	R/N	R_1/N_1	R/N
0.05	0.20	2/21	5/41	2/29	5/38
0.10	0.25	3/21	11/66	4/31	10/55
0.20	0.35	9/37	23/83	9/42	22/77
0.30	0.45	14/40	41/110	28/77	34/88
0.40	0.55	20/45	50/104	25/62	46/94
0.50	0.65	23/42	61/105	29/57	55/93

66

例5. がん単群第II相試験

- パラメータ
  - 閾値奏効率: 20%
  - 期待値奏効率: 35%
  - 検出力: 90%
- 必要サンプルサイズ
  - 中間解析時点で42人
  - 全体で77人
  - 不適格や評価不能を考え, 予定登録症例数は80人

例6. 希少疾患の第II相試験

メトレレプチン皮下注用11.25mg「シオノギ」
Metreleptin（SHIONOGI）

【臨床成績】

1. 国内の医師主導治験

国内の脂肪萎縮症患者4例を対象に，本剤（0.01～0.08mg/kg）を1日1回5ヵ月間連日皮下投与したときのHbA1c（JDS値），トリグリセリドの経時変化を表2に示す。HbA1cは投与前に比べすべての症例で低下した。トリグリセリドも投与前に比べ，正常値まで低下したが，症例No. 4では副腎皮質ステロイド投与により一時的に上昇した。なお，症例No. 3は投与前後ともに正常値であった。また，糖尿病治療薬及び（又は）高脂血症治療薬が本剤投与開始前から投与された3例の患者では，投与開始2ヵ月以内にそれら治療薬の投与が中止された[1]。

表2　国内の医師主導治験でのHbA1c及びトリグリセライドの経時変化

症例No.	年齢	HbA1c（%）（JDS値）				トリグリセライド（mg/dL）			
		投与前	3ヵ月	4ヵ月	5ヵ月	投与前	3ヵ月	4ヵ月	5ヵ月
1	18歳	8.6	5.5	4.8	4.8	210	55	55	62
2	23歳	7.7	5.6	5.9	6.4	246	51	144	204
3	11歳	5.8[注1]	5.1	5.1	5.4	59	46	60	77
4	6歳	5.8	5.0	5.1	5.2	180	83	131	382

注1：登録時にはHbA1c = 6.1

Ito et al. Clin Eval 2014

例6. 希少疾患の第II相試験

10.1　目標症例数

本治験の目標症例数は3例とする。
ただし、登録期間中は3例を超えても登録を継続する。

〔設定根拠〕

対象疾患である脂肪萎縮症は100万～500万人に1人とも言われる稀少疾患である。本邦における直接的な統計データは無いが、昭和61年に厚生省特定疾患、難病の疫学調査研究による全国調査で見出された症例も31例であり、我が国においても、海外の報告と同様に、数百万に1人の稀少疾患であると考えられる。さらに、すべての脂肪萎縮症患者が本治験における血中レプチン濃度などの適格基準を満たすわけではないことからも、多数の症例を確保することは困難である。実際、本邦における臨床試験においても2002年のスタート時より精力的に症例リクルートに努めてきたが、これまでの約7年間で登録可能であった症例は12例であった。これは単純計算で1年間あたり1.7例であり、本治験で計画している登録期間1年8か月を考え合わせると、リクルート可能な症例数は3例が現実的である。

Ito et al. Clin Eval 2014

Table 2. Characteristics of Pivotal Preapproval Trials of Orphan and Nonorphan Cancer Drugs

Characteristics	No. (%)[a] Orphan Drug Pivotal Trials (n = 23)	Nonorphan Drug Pivotal Trials (n = 15)	P Value
Enrollees, median (interquartile range)	96 (66-152)	290 (185-394)	<.001
Randomized, multigroup	7 (30)	12 (80)	.007
Comparator			
Active	4 (17)	7 (47)	.007
Supportive care	2 (9)	1 (7)	
Placebo	1 (4)	4 (27)	
None	16 (70)	3 (20)	
Blinding			
Double-blind	1 (4)	5 (33)	.04
Single-blind	1 (4)	0	
Open-label	21 (91)	10 (67)	
Primary trial end point reported[b]			
Disease response[c]	17 (68)	4 (27)	.04
Disease progression[d]	4 (16)	6 (40)	
Overall survival	2 (8)	4 (27)	
Other	2 (8)	1 (7)	

[a] Data are reported as No. (%) unless otherwise indicated.
[b] Two orphan trials included co–primary end points of disease response and disease progression.
[c] Disease response included hematologic response (6 pivotal trials of orphan drugs), cytogenetic response (3 pivotal trials of orphan drugs), or change in tumor burden (8 pivotal trials of orphan drugs and 4 pivotal trials of nonorphan drugs).
[d] Disease progression end points include time to tumor progression (4 pivotal trials of nonorphan drugs) and progression-free survival (4 pivotal trials of orphan drugs and 2 pivotal trials of nonorphan drugs).

Kesselheim, et al. JAMA 2011

主たる解析とは別に考慮すべき事項について

- 希少疾患かどうか
- 試験の実施可能性
  - 登録期間
  - 施設数
  - コスト
- 副次的な目的
  - 副次エンドポイントやサブグループ解析の検出力
  - 安全性情報の収集

公式のまとめ

- 連続データ
  - $m = \left(\frac{1+\varphi}{\varphi}\right)\frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\Delta^2}$
- 2値データ
  - $m = \left(\frac{1+\varphi}{\varphi}\right)\frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\log(\mathrm{OR})^2\bar{\pi}\,(1-\bar{\pi})}$
- 生存時間データ
  - $e = \frac{2(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\varphi\log(\mathrm{HR})^2}$
  - $m = e/\pi_1$
- クラスターランダム化
  - $m_{cluster} = \{1+(k-1)\rho\}m$
- 非劣性試験
  - $m = \left(\frac{1+\varphi}{\varphi}\right)\frac{(Z_{1-\alpha}+Z_{1-\beta})^2}{\Delta_{margin}^2}$

Key words & further reading

- 仮説検定の用語
  - αエラーとβエラー
  - 検出力（power）
  - 帰無仮説（null hypothesis）と（alternative hypothesis）
  - 片側（one-sided）と両側（two-sided）
  - 有意水準（significance level）
- サンプルサイズ設計のパラメータ
  - 検出力（power）
  - 検出したい治療効果の大きさ（size of effect）
  - サンプルサイズ（sample size）

- Machin, et al. Sample Size Tables for Clinical Studies, 3rd Edition. Wiley.