医療機器_統計解析基礎講座

株式会社バイオスタティスティカルリサーチ
古川敏仁

本セミナーの目次と内容

I. 治験実施計画書と総括報告書での統計に関する記述
–記述の範囲と役割–

II. 統計とは　平均値とばらつきを評価するもの
–統計の原理と記述法–

第1部　治験実施計画書と総括報告書での統計に関する記述

- 関連する章、事項
- 記述の目的
- 記述例

治験実施計画書と報告書の統計に関連する章

- 試験の目的(Objective of the Clinical Study)
- 試験デザイン(Study Design)
  - 試験デザイン(Study Design)
  - 主要評価項目(Primary Endpoint)
  - 副次評価項目(Secondary Endpoint)
- 症例数設定とその根拠(Sample Size)
- 統計解析(Study Design)
  - 試験対象集団　ITT、FAS、PPS
  - 症例、データの取り扱い
  - 結果の判断基準、p値の有意水準
  - 統計解析方法の詳細

- 統計解析の結果と結論（STED、総括報告書）

なぜ治験実施計画書とSTEDには統計に関連する章が必要なのか

- このような統計学的な要素を加味した記述が必要なのは、データから得られた結論の科学的な妥当性を保証するため。

- データにはばらつき、バイアスが存在する。ゆえにデータから得られた結論はこれら要素を加味した上でも、結論が正しいといえる論理的な根拠が必要＝統計の役割

- バリデーションとは、データから得られた結論が、変わらないことを保証すること

なぜ治験実施計画書と報告書には統計に関連する章が必要なのか

- バリデーションとは、データから得られた結論が、変わらないことを保証すること

- そのためには下記の要素が試験実施計画書、報告書には必要となる。

- 科学的な妥当性
- 一貫性(Consistency)

科学的な妥当性

- 適切な結論を得るために
  - 解析対象被験者の定義は適切か
  - 欠測値やTime windowの設定は適切か
  - 試験の成否を判断する判断基準は適切か
  - 症例数は評価に対して十分であるか
  - 統計解析手法は適切であるか

一貫性(Consistency)

- 適切な結論をえるために、各記述は整合が取れているか

- 適切な結論をえるために、時間経過に対して、矛盾なく論理的な推論が行われているか

=試験実施計画書にはどこまで詳しく統計的事項を記述すべきか

試験実施計画書にはどこまで詳しく統計的事項を記述すべきか

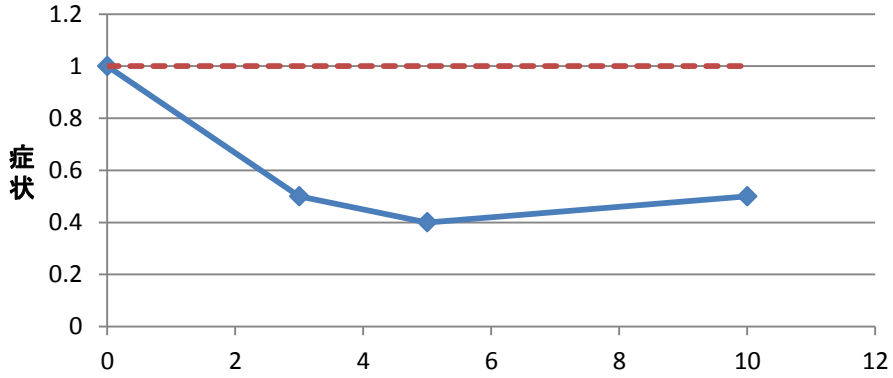
- 一貫性の問題

- データにはばらつきが存在する
- データを見てから自分の都合の良いようにルールを定めれば、どんな結論でも導くことができる
- それは、データから、科学的に、客観的に結論を導くということに反する
- ゆえに、結論に影響を与えるようなルールは、事前に、文書で、宣言しておく必要がある
- つまり、試験実施計画書には、結論に影響を与えるようなルールはすべて、記載する必要がある。

科学と錯覚の違い
科学的な結論

- 宣言（試験実施計画書への記載） 明確
  例：この薬剤は3日後に効き始め、以降効果は持続する

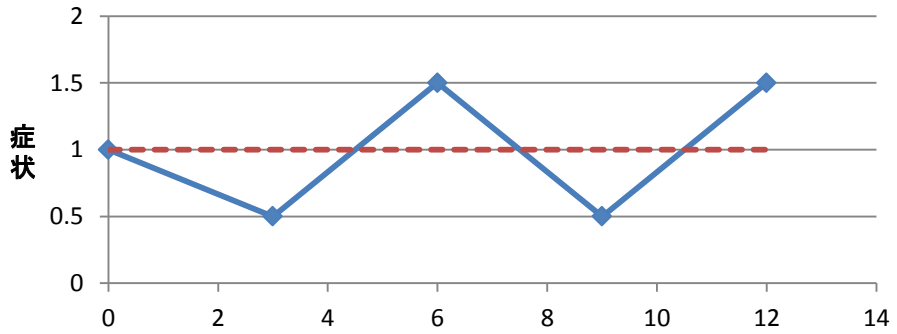
- 結果
  例：確かにこの薬剤は3日後に効き始め、以降効果は持続している



- 結論
  - 例：宣言通りのことが、科学的な観察で観測されたので、この結果は偶然ではなく、薬剤の作用であると推測できる。
  - 宣言通りのことが観測されたとき、その結論の妥当性は強い

科学と錯覚の違い
錯覚あるいは嘘

- 宣言（試験実施計画書への記載）は曖昧
  例：この薬を飲めば効くはずである
- 結果
  例：見て、3日と9日は、症状が低下しているじゃない



- 結論
  - 例：それで、薬の効果があるというのは、納得できない
  - 6日、9日のように悪化しているときもある
  - 自分の都合の良いように言うのは、錯覚か嘘か悪意、科学ではない

正しい記載のポイント　1
試験の目的(Objective of the Clinical Study)

- 試験の最終的なGOAL、すなわち、何を用いて機器の評価を測り、どのような結果になった時に試験は成功したと結論付けれるのかを明確に記載する必要があります。

- 何を用いて→　主要評価項目の定義
- どのような結果になったら→成功の定義
- 例：6か月後のSBPがBaselineよりも30mmhg以上、統計学的に有意に低下した場合、本機器の性能は示されたと定義する

- 成功の定義の判断基準も重要
- 例：現在主流の薬剤治療では、SBPの低下は10mmhg以上は見込めず、臨床的に30mmhg以上の低下があれば、心疾患リスクを60%低減できるため、30mmhg以上のSBPの低下があれば、本機器の性能は臨床的に満足するものであると考えられたため

正しい記載のポイント　2
試験デザイン(Study Design)

- 目的とする機器の性能、効果が適切に評価できる試験であることを記載

- 例：無対照、単群、多施設共同試験

- 試験デザインの妥当性の記述も重要
- 例：本機器の性能はすでに海外の無作為化比較試験によって効果が証明されている。この海外試験の結果をもとに、従来法では到達できない30mmhg以上のSBPの低下が認められれば、日本においても十分性能が科学的に評価できる。
- ゆえに、2群比較ではなく、すべての症例を本機器の使用対象とすることが安全性例数を増やすことになる。本機器の性能評価にはもっとも適切な単群無対照試験を設定した。

正しい記載のポイント　3
試験例数の設定(Sample Size)　(1)

- 設定した例数で、データのばらつきを考慮しても、試験が的確に実施され予想通りの成績であったならば、試験の成功が一定以上の確率で保証される例数でることを示します。また、被験者の未知のリスクを最小限にするために、前期範囲の中で最小の例数であることが望まれます。

- 例：　単群、baselineからの変化を検定で証明する場合
- 試験例数　1群　XX例
- 脱落の考慮　脱落率　5%
- 検定方法　　片側有意水準2.5%の1標本t検定
- 検出力　　0.8（設定した条件のもとでの試験の成功確率）

正しい記載のポイント　3
試験例数の設定(Sample Size)　(2)

- たとえ結論を検定による証明を用いないで判断する場合でも、その試験例数で何がいえるかを記述する必要があります

- 例：本試験は安全性の確認試験である。もし、30例の試験で1件のMACEも観測されなかった場合、MACEは統計学的に95%の確率で9.8%以下の発現率であることが示される。ゆえに、30症例以上の試験であれば、10%以上のMACE発現事象は観測することができ、ISOガイドラインで示される試験基準を満たすことから、本試験例数は30例で妥当であると考えた

正しい記載のポイント　4
解析対象集団（被験者）　(1)

- 主要評価項目を評価する解析対象集団は明確に定義する。

- 例：プライマリエンドポイントの評価における解析対象集団は、主解析はITT、副次的にPPS　など

正しい記載のポイント　4
解析対象集団（被験者）　(2)

- 解析対象集団の定義はできる範囲で明確に記載する

- 例：ITT（Intention-to-Treat)
- 試験に登録され、治験機器の埋植を試みたすべての症例、ただし、試験に登録され、割り付けられたが、手術時の判断で機器の施行を見合わせた症例は除く
  なお、試験機器の埋植を試みたというのは、付属のA機器を血管に挿入した症例を意味する。

- 医療機器において、ITT、FAS（最大の解析対象集団）、安全性解析対象集団は、同義となることが多い

正しい記載のポイント　4
解析対象集団（被験者）　(3)

- PPS（試験実施計画書に準じた集団）などの定義は明確に記載するが、あまりに細部まで規定すると、試験中にいろいろな例外が発生するとルールが維持できなくなる可能性がある。
- 結論が変わらないことを保証する範囲を明確にして、方針を記述することが重要

- 例：PPSは、ITT集団から以下の症例を除外した集団、ただし、重要なプロトコール違反の内容は、試験終了後、医学専門家により解析前に再確認する。
  - ベイルアウト症例
  - 3週間以上のフォローアップがない、あるいはデータが存在しない症例
  - 機器の設置が適切に行われなかった症例
  - 事後的に試験対象ではないと判明した症例
  - 重要なプロトコール違反を認めた症例

正しい記載のポイント　5
欠測値やtime window

- 例えば、6か月後までのイベントの発生率などの計算では、どのような症例を分母に含めるかで結果が大きく変わる可能性があります

- 例：100例の試験で、10例に6ヶ月までにイベントがあり、10例がイベントが観察されないで脱落となりました。
- 脱落者を分母に含む場合　　発生率=10/100=10%
- 脱落者を分母に含めない場合　発生率=10/90=11.1%

- このような場合、欠測の取り扱い、すなわち、脱落者を評価項目の分母に含めるのか、含めないのか、それともKM法などで脱落を考慮した方法で発生率を求めるのかは明記する必要があります。
- また、欠測値を欠測前の値で補完（LOCF）する場合や、一定の時点範囲で観測されてない症例は解析から除く場合など、事前に、明確な定義が必要です。

正しい記載のポイント　6
統計解析

- せめて、主要評価項目の解析方法と、結果の判定方法は明確に書く
- 基本的には、結論に影響を与える事項は明確に事前に記載することが必要なので、副次評価項目やその他の解析計画も試験実施計画書に記載することが望ましい。

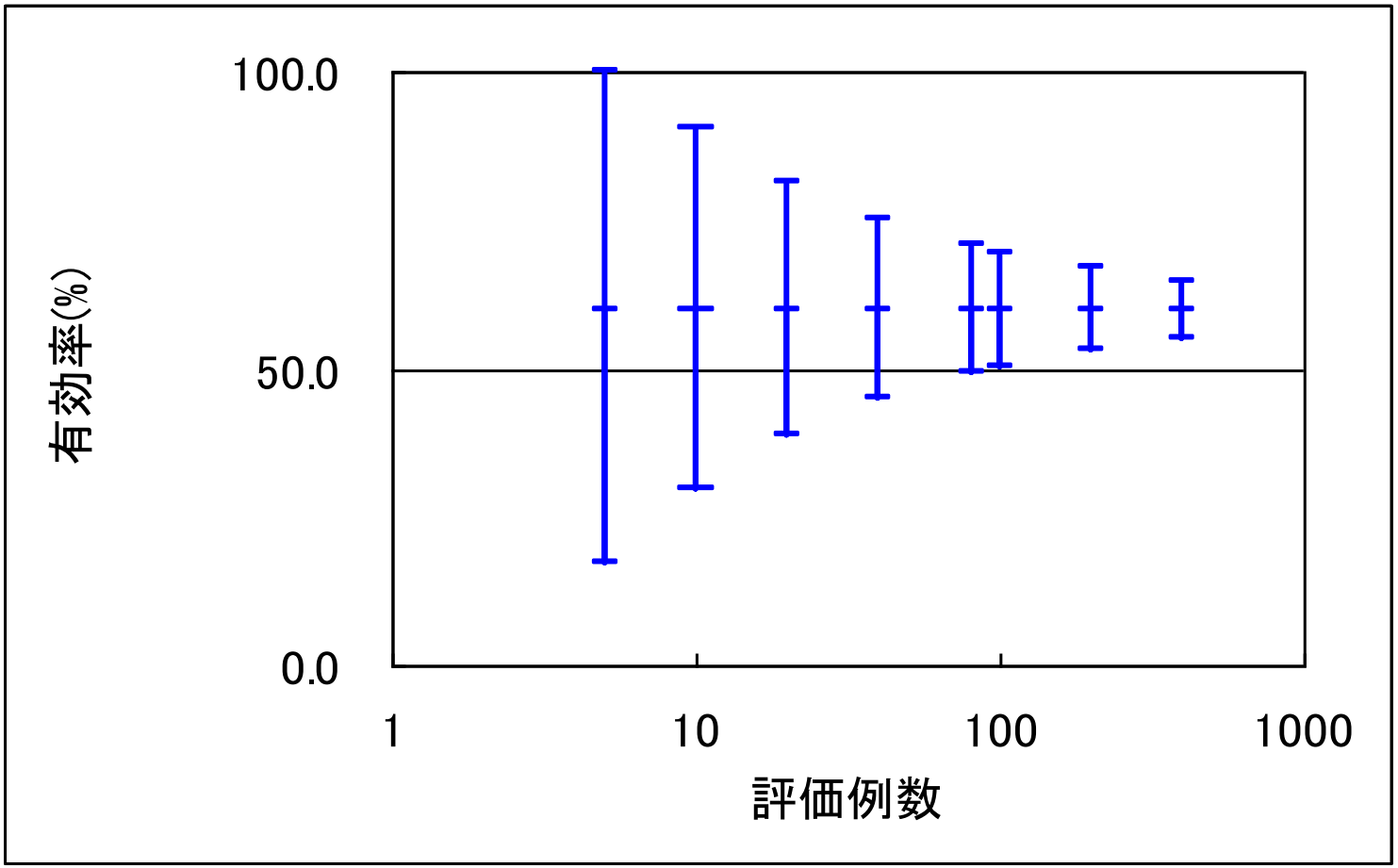
第Ⅱ部　そもそも　統計とは

- 平均値をバラツキとともに評価すること
- 例えば、60%の有効率といっても、例数が違うとその意味は違ってくる

信頼区間と例数の関係

有効例数	評価例数	有効率	95%信頼区間 下限	95%信頼区間 上限	50%との差 正規検定
3	5	60.0	17.1	100.0	0.6481
6	10	60.0	29.6	90.4	0.5186
12	20	60.0	38.5	81.5	0.3613
24	40	60.0	44.8	75.2	0.1967
48	80	60.0	49.3	70.7	0.0679
60	100	60.0	50.4	69.6	0.0412
120	200	60.0	53.2	66.8	0.0039
240	400	60.0	55.2	64.8	0.0000

信頼区間と例数の関係



ばらつきを考慮した割合平均の表し方

- 評価例数が15例以下→例数を表示
  例　3/7、12例（14例の評価例数）

- 評価例数が16～30例→例数と%を表示
  例　7/20 35.0%

- 評価例数が30例以上の場合
  →例数と%と信頼区間を表示
  例　24/40　60.0%　（44.8%～75.2%）

区間推定値：95%信頼区間

- 無限に実験（データ抽出）を繰り返した時、それぞれの試験の信頼区間に真の値が入ることが95%の確率で存在するであろう区間
- （つまり、100回実験を行えば、それぞれの試験の95%信頼区間に95回は真の値を含む区間）

- 例60.0（60/100）の95%信頼区間：50.4～69.6

- 注意：95%信頼区間はあくまで今回の実験により得られたデータに関する指標であり、この区間に95%の確率で真の値が含まれるわけではない

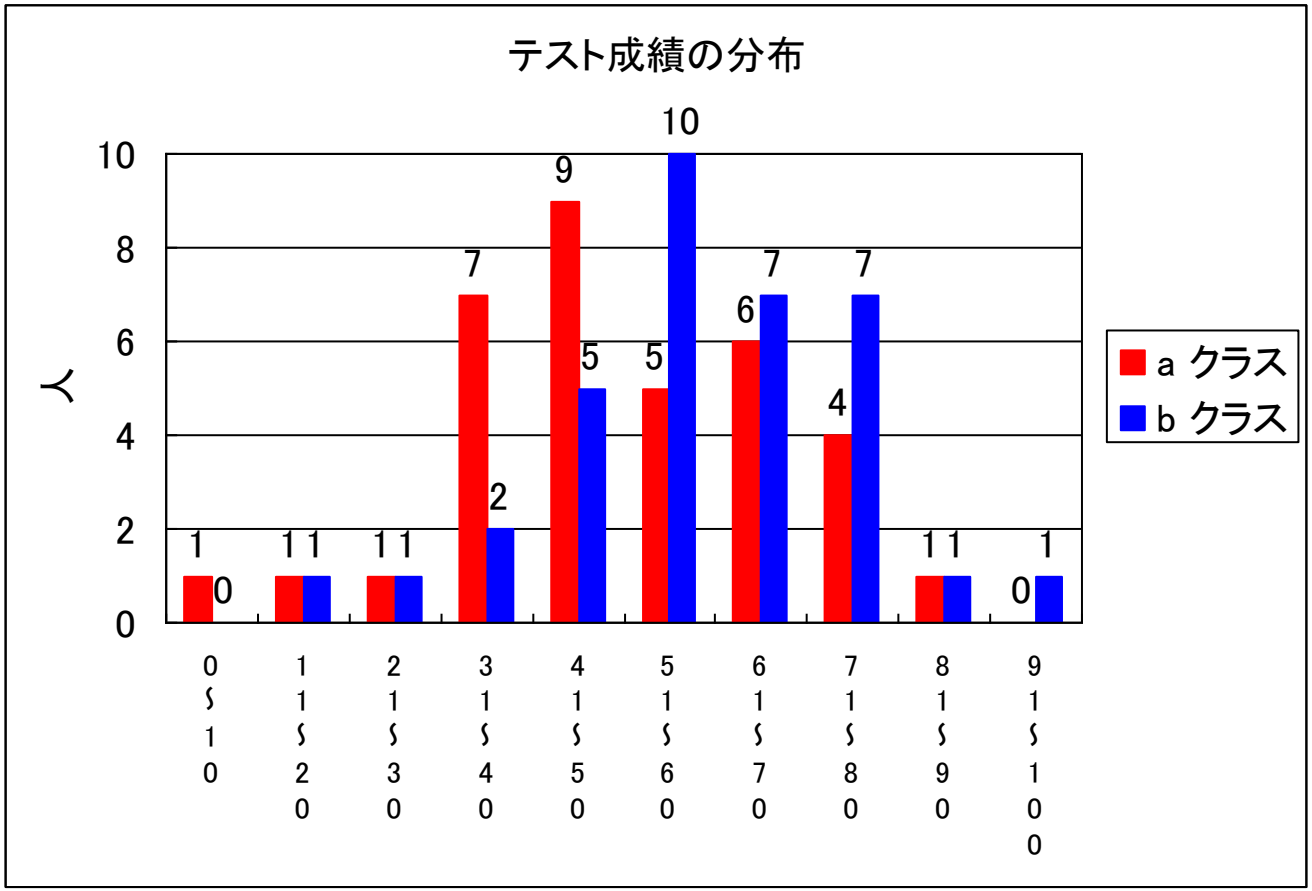
検定とは

- 左記の例で、100例以上であれば、95%信頼区間が50%より高くなる、つまり、95%以上の確率で、有効率60%は50%より高い有効率だといえる。
- それは、検定でp<0.05となる。（母比率50%の正規検定）

- 検定：実際に起こった事象が、偶然かどうかを判定する手法
- 事象が差がない状態から95%以上の確率で起こりえない現象を差があると定義する
- 検定：確率のみ　実際の有効率は評価できない
- 重要なのは結果の平均値　検定は平均値を補完する手法
- 重要なのは、有効率　60%　or　70%
- 検定のp値はあくまで補足的
- 注意：p値は例数に依存する。

分布と検定

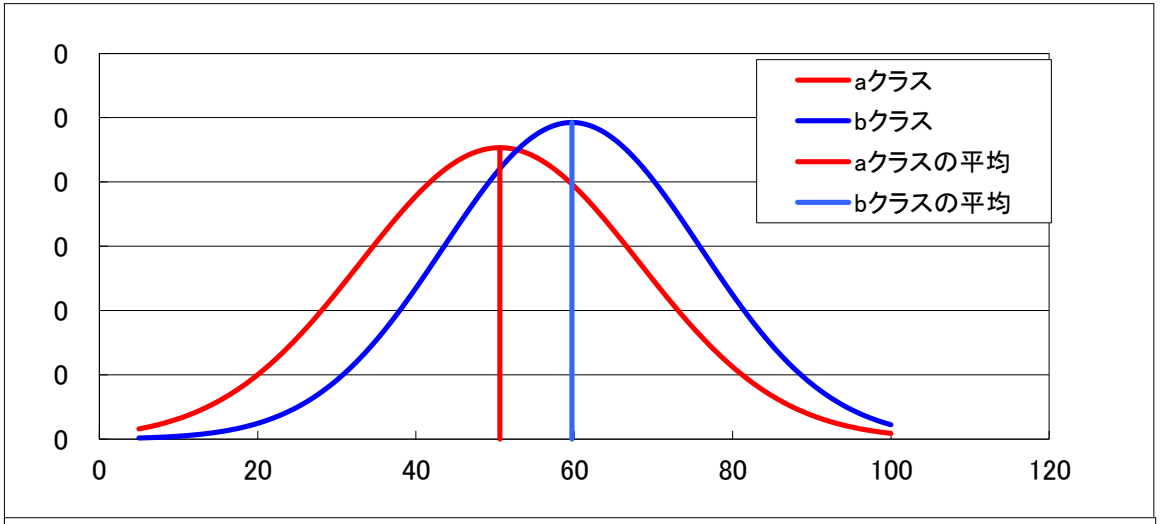
赤クラスと青クラスのテストの平均値に差があるか？

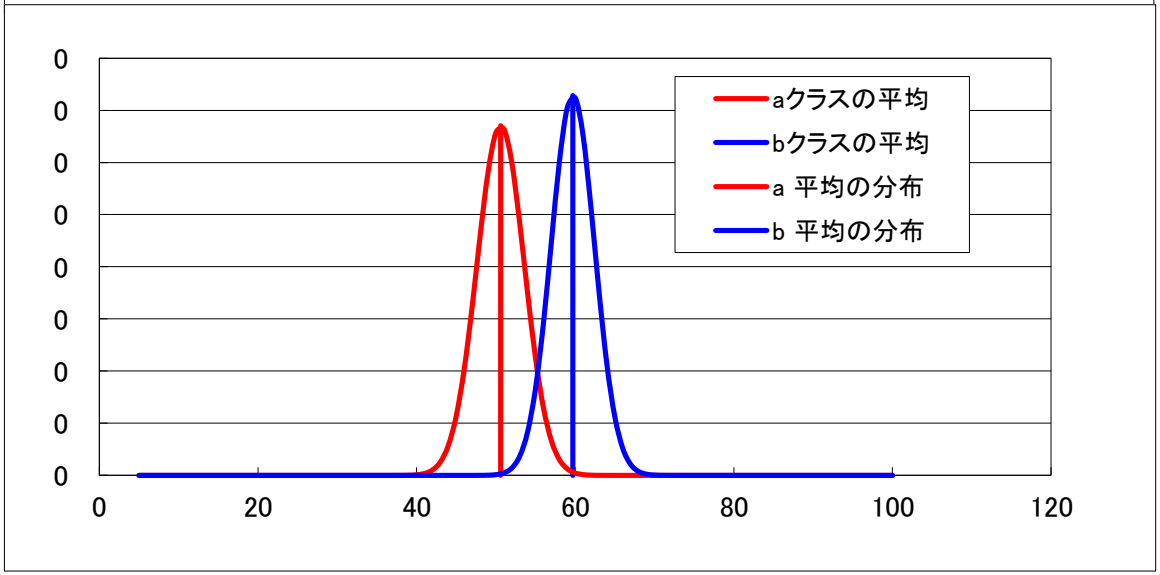


分布　ばれのあるデータの要約

表 1.　生物統計テスト成績の要約

	a クラス	b クラス
例数	35	35
平均値	50.6	59.7
標準偏差	17.6	16.2
最小値	5	14
中央値	49	60
最大値	90	97
歪み	-0.222	-0.418
尖り	0.475	1.236





- 生データの分布
平均±1.96×標準偏差

- 平均の分布
平均±1.96×標準誤差

- 例数が多くなると、平均値の重なり合いは小さくなる
- ＝2群は区分できる
- ＝統計学的に有意に差があるとなる

t検定で差を確認
p=0.0133

$$t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\left(\frac{1}{n_a} + \frac{1}{n_b}\right)\left(\frac{(n_a - 1)s_a^{\,2} + (n_b - 1)s_b^{\,2}}{n_a + n_b - 2}\right)}} \qquad (1)$$

$$= \frac{50.6 - 59.7}{\sqrt{\left(\frac{1}{35} + \frac{1}{35}\right)\left(\frac{(35-1)(17.6)^2 + (35-1)(16.2)^2}{35 + 35 - 2}\right)}}$$

$$= -2.27$$

ただし

$\bar{X}_a$: aクラスの平均点、$\bar{X}_b$: bクラスの平均点

s_a : aクラスの標準偏差、s_b : bクラスの標準偏差

n_a : aクラスの例数、n_b : bクラスの例数

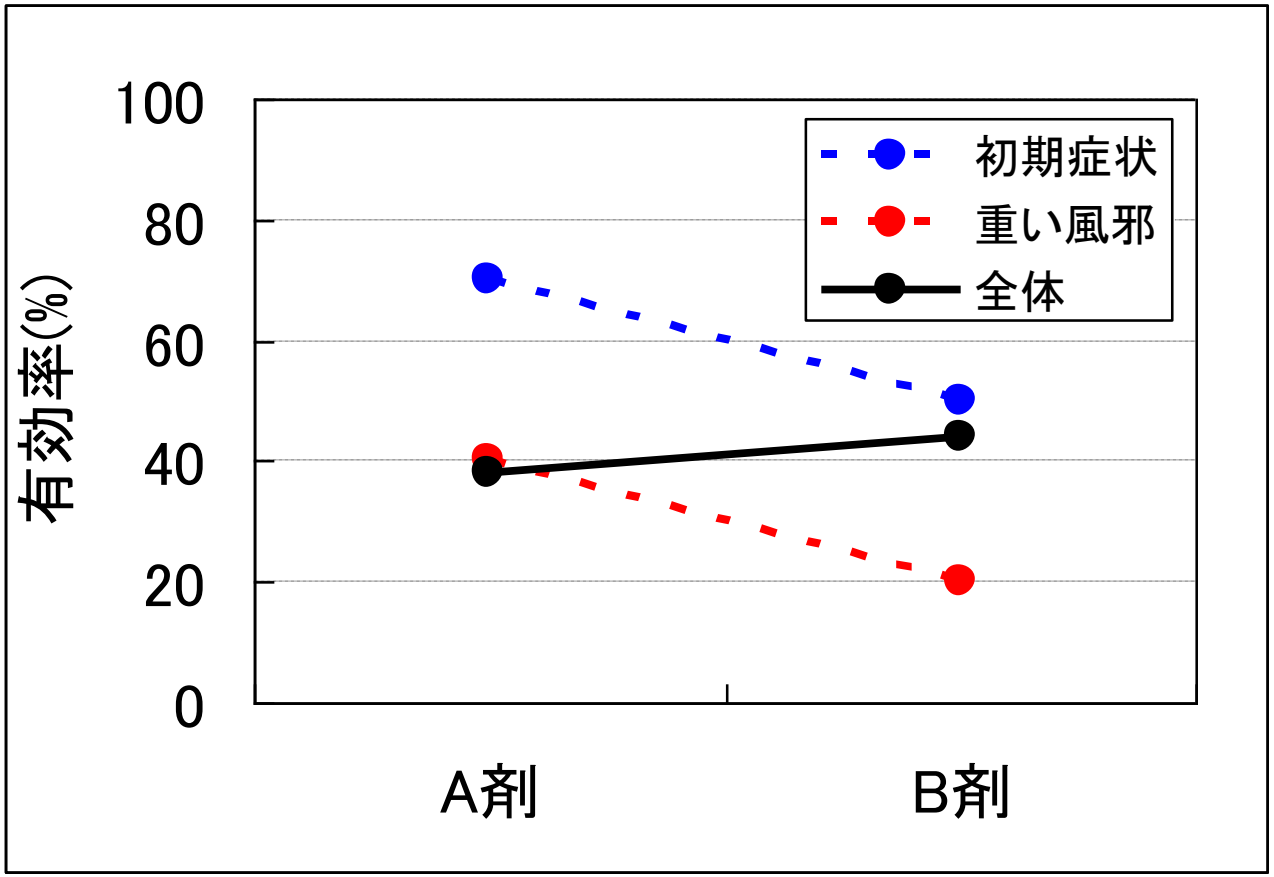
平均値に影響を与える2つの要因
交絡とバイアス

- **交絡（confounding）**
  - 結果（outcome）と関連があり、調査したい要因（暴露因子）とも相関がある因子（交絡因子）が存在すると、暴露因子の評価に影響を与える
  - 統計解析で取り除くことが可能
- バイアス（bias）
  - 結果（outcome）と暴露因子に影響を与える被験者背景の潜在的、顕在的な偏り
  - 統計解析ではとりのぞくことは不可能
  - 試験デザインでバイアスを最小とする必要がある

シンプソン・パラドックス
（Simpson's Paradox）

	風邪薬の有効率(%)			
	A剤		B剤	
初期症状	70%	(14/20)	50%	(40/80)
重い風邪	30%	(24/80)	10%	(2/20)
全体	38%	(38/100)	42%	(42/100)

シンプソン・パラドックス
（Simpson's Paradox）

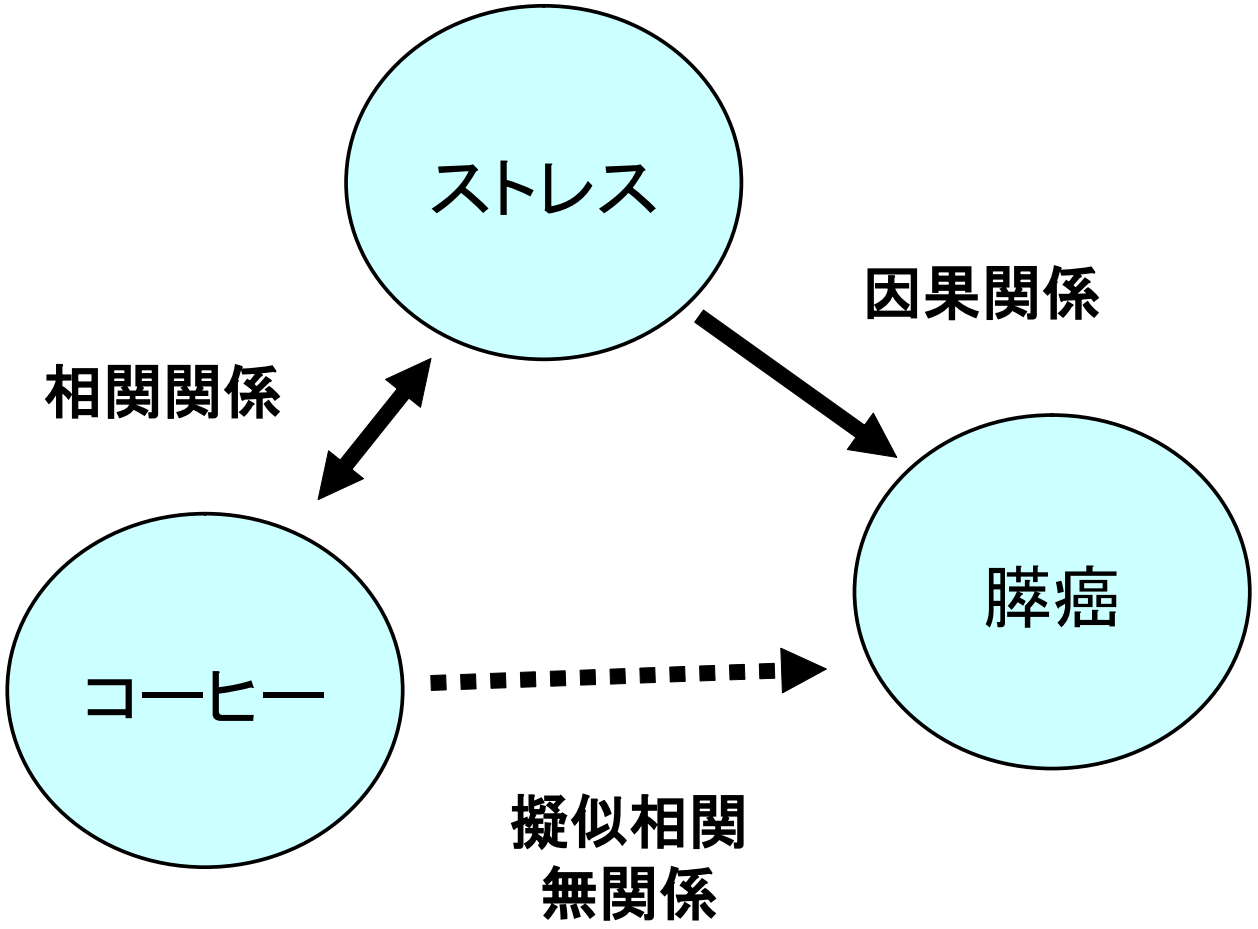


交絡とバイアス

- **交絡（confounding）**
- 1981年、The New England Journal of Medicine（NEJM）の特集
- 膵癌とコーヒーには因果関係があり、コーヒーを多飲すると膵癌になりやすいというショッキングな論文
- 実はコーヒーと膵癌には因果関係はなく、コーヒーを良く飲む人はストレスが強く、ストレスと膵癌には因果関係があるため、あたかもコーヒーと膵癌に因果関係があるかのような現象が観察された？
- 1)　B MacMahon, S Yen, D Trichopoulos, K Warren, and G Nardi ‘Coffee and cancer of the pancreas’, The New England Journal of Medicine,

  304:630-633, 1981

交絡（confounding）



交絡を統計解析上除外する方法

- 多変量解析
- Propensity Score解析

多変量解析による交絡の補正

薬物常習者治療データ(Applied Logistic Regression 2)

単変量解析

		95%信頼区間		
変数	オッズ比	下限	上限	p値
治療	1.55	1.06	2.26	0.024
年齢/10歳	1.20	0.89	1.62	0.236
過去の使用歴	0.93	0.88	0.97	0.002
静脈注射かつて	0.62	0.37	1.04	0.708
静脈注射最近	0.46	0.30	0.70	0.011
白人	1.58	1.05	2.39	0.030

多変量解析

		95%信頼区間		
変数	オッズ比	下限	上限	p値
治療	1.55	1.05	2.29	0.028
年齢/10歳	1.67	1.19	2.34	0.003
過去の使用歴	0.94	0.89	0.99	0.014
静脈注射かつて	0.55	0.32	0.97	0.410
静脈注射最近	0.47	0.29	0.76	0.041
白人	1.23	0.80	1.90	0.347

バイアス (系統的誤差)

- 選択バイアス
- 測定バイアス

選択バイアス

- 選択バイアス
- 風邪薬開発のような場合、新しい風邪薬の方が効くはずだからという期待感から、重い症状の症例に新しい風邪薬A剤を割付け、B剤には軽い症状の症例を割付ける

- **所属集団バイアス**
- 特定の集団、例えば、大学の運動部、ベジタリアン、医師のボランティア集団などは、一般的な集団とは違った健康度を示すことにより評価が偏ること。

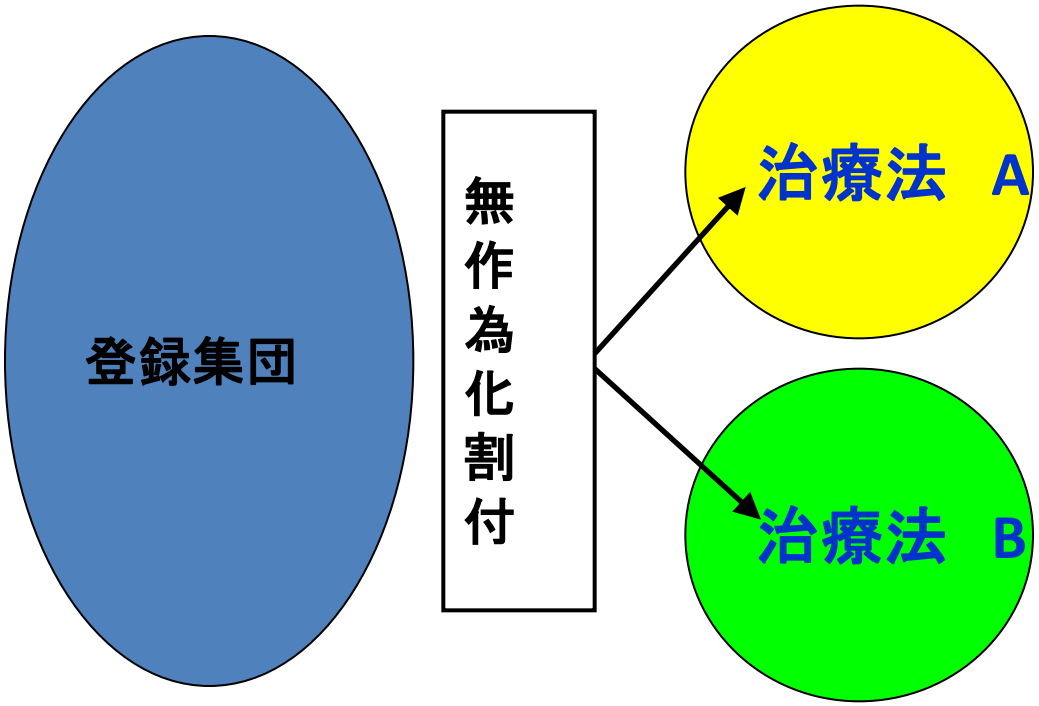
選択バイアス・交絡を避ける

- 無作為化割付

- 例数さえ多ければ、無作為化割付を行えば、理論的に治療効果意外の背景因子のバランスが、実験群、対照群で等しくなる。
- 無作為化割付：調査可能な背景因子以外の潜在的な因子のバランスも等しくする。

無作為化比較臨床試験

- 遺伝的要因、年齢、性別、重症度、生活環境、・・・etc　これらの影響をすべて正確に評価することはできない
- 無作為化割付　A、B2群の背景因子らが均一なら、治療効果だけは比較できる



背景因子の均一化

遺伝的要因

年齢

性別

病気の重症度

生活環境

・・・・・・

治療効果の差だけは評価できる

医療機器の臨床試験デザイン

- 選択バイアスを避ける手段として、有効性を評価するためにはほとんどすべての臨床試験で無作為化割り付けは必要です。

測定バイアス　A:評価者バイアス
医療機器臨床試験の問題

- 臨床試験で治療法を評価する医師が、評価症例の治療法を知っていたら、医師は恣意的、あるいは潜在的に自分の期待する治療法の評価に良い値をつけてしまう可能性がある。
- 評価書の癖、訓練などによって、評価値が違ってくる可能性がある。

測定バイアス　想起バイアス
アンケート調査に潜む問題

- イベントを経験した人は、しなかった人にくらべて暴露要因を報告しやすい、あるいは些細な暴露の経験でも思い出だす。

- 例えば、1980年代、テレビ画面の発する電磁波と流産に関する研究
- 多くのアンケートを基にした研究で、テレビ画面（パソコン）の近くで従事する女性は、そうでない女性と比較して流産の可能性が高いとの報告

- アンケート用紙に「テレビ画面と流産の関係を調査する」ということが記載されており、その結果、流産した女性は注意深くテレビ画面の前に座ったことを思い出そうとし、しなかった女性は、あまりテレビ画面のことには気にも留めなかったので、流産経験者の方が多くテレビ画面の前にいたという結果

測定バイアスを避ける＝盲検化

- 調査者、被験者に対して治療群、対照群が分からないようにする。

- 医療機器
- 調査者、被験者とも治療群が何であるのかは盲検化できない場合が多い
- 評価者盲検
- このような場合、有効性の評価だけはビデオ、あるいは治療法を知らされていない第三者による評価
- 第三者による評価はできれば複数が望ましい

one pointアドバイス　問題1（医学論文のエビデンス）

- ある癒着防止材を使用すると、心臓再手術の平均手術時間が癒着防止効果により、使用しない場合384分から157分に短縮する？という希望があった

	Treatment Group (n = 21)			Control Group (n = 23)		
	平均	中央値	四分位	平均	中央値	四分位
Age(days)	92	35	7-67	894	614	194-1112
体重(kg)	3.9	2.9	2.6-3.9	13.3	8.4	6.3-14.5
手術時間(min)	157	130	79-214	384	314	285-475

- この問題点は何か？　むろん、手術時間は統計的に有意（$p<0.0001$)に減少している

one pointアドバイス　問題1
（医学論文のエビデンス）

- 答え：被験者の年齢、体重からすると、Treatment群は生後30日前後の新生児であり、対照群は生後2歳前後の幼児
- 　術式、あるいは、時代も違うため、手術時間の比較自体意味を持たない
- つまり、比較を行うためには被験者背景が等しくなくては比較にならない
- 被験者背景さえ記載されていない論文は、全く信用できない

	Treatment Group (n = 21)			Control Group (n = 23)		
	平均	中央値	四分位	平均	中央値	四分位
Age(days)	92	35	7-67	894	614	194-1112
体重(kg)	3.9	2.9	2.6-3.9	13.3	8.4	6.3-14.5
手術時間(min)	157	130	79-214	384	314	285-475

one pointアドバイス　問題2の延長（よくSTEDにみる問題　2つの試験結果を検定で比較したがる）

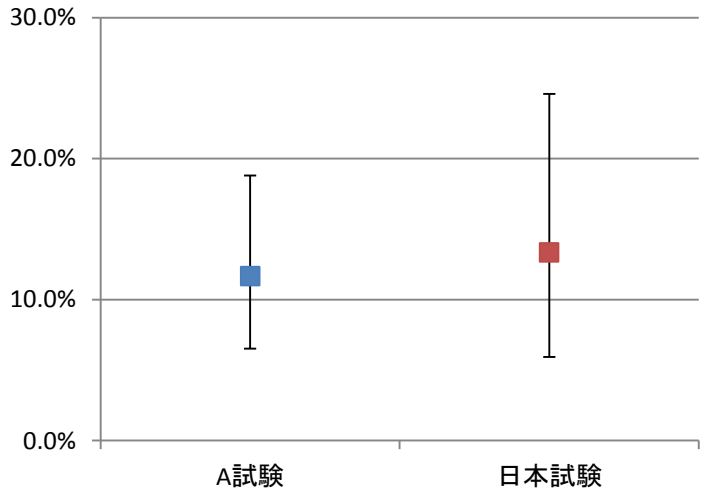
- よく、海外で開発されたDESの結果を、海外試験と日本試験の結果を列記し、p値をつけることがありますが、p値には比較可能性がありませんから、除外しましょう
- 過去の製品との比較でも同じです　p値は意味がありません
- P値が意味があるのは同時比較臨床試験だけです。

術後10カ月のTVR					
					Fisher
		例数	TVR	TVR率	p値
A試験	A　DES	120	14	11.7%	0.8106
日本試験	A　DES	60	8	13.3%	

one pointアドバイス　問題2の延長（よくSTEDにみる問題　正しい方法）

- それぞれの試験成績と信頼区間を記載し、その分布を臨床的な見地から妥当性を論議しましょう
- くれぐれも検定は使わないように

術後10カ月のTVR						
					95%信頼区間	
		例数	TVR	TVR率	下限	上限
A試験	A　DES	120	14	11.7%	6.5%	18.8%
日本試験	A　DES	60	8	13.3%	5.9%	24.6%



one pointアドバイス　STEDの記載
平均値とその分布の論議が重要

- 最悪な記載（p値しか書いていない）
- 海外でのA試験の術後10カ月のTVR発生率と、の本試験のTVR発生率はp=0.8106と有意差はなく、海外成績と日本成績が同等であると示された。

- 良い記載
- 海外でのA試験の術後10カ月のTVR発生率は14/120、すなわち11.7%(6.5%～18.8%　95%信頼区間）であり、一方、日本試験では、8/60、13.3%（5.9%～24.6%）であり、図に示すように両者の分布は平均値を中心に重複しており、海外成績と日本成績は臨床的に同等であると示された。

医療機器の臨床データ

医療機器データは複雑です

- 各種臨床試験デザイン、例数設計
- 中間解析やアダプティブデザイン
- ベイズ法を利用した解析
- 観察データの背景調整のためのPropensity解析

など個別の問題は生物統計家にご相談されるのが最も効率的な方法だと思います

Q&A time