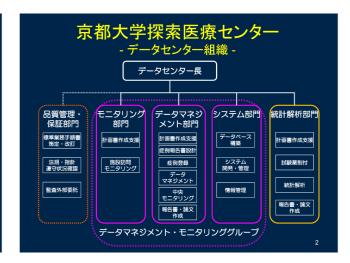


科学的実験と統計的実験の境界

手良向 聡

京都大学医学部附属病院探索医療センター検証部京都大学大学院医学研究科臨床試験管理学分野

2010.12.8 日本分子生物学会若手教育シンポジウム、神戸



科学的実験と統計的実験

- > 科学的実験
 - 人工的に作り出された純粋な条件のもとでの因果関係 を確定することを目的とする
 - 科学的法則の法則性を検証する
- > 統計的実験(技術的実験) by R.A.Fisher
 - 現実の場での因果関係を確かめることを目的とする
 - エ学、農学、医学等の現実の応用において、何をなす
 べきかを知る

竹内啓「統計学的な考え方」In: 統計学の基礎II、岩波書店、2003

基礎医学研究

- ▶ ばらつきを完全にコントロールできる場合▶ 科学的実験
- > ばらつきをコントロールできない場合
 - 統計的実験
 - シグナルとノイズを分離
 - ・再現性で結果を保証



データの統計的な解析と解釈が必要

統計ガイドライン (1)

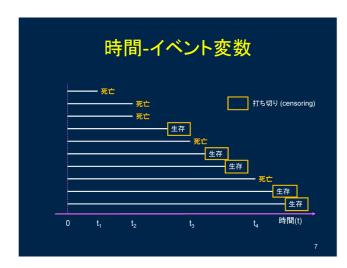
統計的検定を含むすべての報告されるデータの記述は、エラーバーとP値を生成するのに用いた統計 的検定名、各データ点の基礎となる独立した実験数 n(1標本の反復測定ではなく)、および各検定に関する実際のP値(単に「有意」あるいは「P<0.05」ではなく)を明記しなければならない</p>

Nature Cell Biology, Guide to authors and referees, 2009.12

統計手法の例

目的	連続変数	分類変数	時間-イベント変数
分布の記述	ヒストグラム、 箱ヒゲ図、散布図	ヒストグラム、 分割表	生存曲線 (Kaplan-Meier法)
要約統計量	平均、分散、中央値、 パーセント点、相関 係数	頻度、一致度、 相関係数	x年生存割合、 中央生存期間
検定(単純)	t検定、分散分析、 Wilcoxon検定	χ²検定、 Fisher正確検定	logrank検定
検定(層別)	分散分析	Mantel-Haenszel検 定	層別logrank検定
回帰モデル	分散分析、 重回帰分析	logistic回帰分析	Cox回帰分析

6





統計ガイドライン (2)

- 記述統計量は、明確に標識された中心の尺度(平均値あるいは中央値のような)、および明確に標識されたばらつきの尺度(標準偏差あるいは範囲)を含むべきである
- ➢ 小さなデータセットに関して、範囲は標準偏差あるいは標準 誤差よりも適切である
- ▶ データを対照と比較する際には、標準誤差あるいは信頼区間が適切である
- ▶ グラフは明確に標識されたエラーバーを含むべきである
- » 著者は±記号に続く数が<u>標準誤差s.e.m.</u>あるいは<u>標準偏差</u> s.d.のどちらであるかを明記しなければならない

Nature Cell Biology, Guide to authors and referees, 2009.12

9

記述統計学と推測統計学

- ➤ 記述統計学(descriptive statistics)
 - 集団としての特徴を記述するために、データを整理・要約すること
 - 位置の尺度
 - 算術平均、メディアン(中央値)、幾何平均
 - ばらつきの尺度
 - 範囲、四分位範囲、分散、標準偏差
- > 推測統計学(inferential statistics)
 - データを分析し、母集団についての推測をすること
 - 推定
 - 標準誤差(推定量の標準偏差)、信頼区間
 - 仮説検定
 - P値

推測統計学の標準: 頻度論アプローチ - Fisher-Neyman-Pearson理論 -

- ▶ 確率は「仮想的繰り返しに基づく相対頻度」
- > パラメータは「未知の定数」
- > 仮説検定
 - 帰無仮説、対立仮説
 - 第I種の過誤(αエラー)、有意水準
 - 第II種の過誤(βエラー)、検出力
 - P値
- ▶ 推定
 - 信頼区間

Jerzy Neyman (1894-1981) 著作権保護のため画像削除 Egon S. Pearson (1895-1980) 著作権保護のため画像削除

仮説検定の例

	転		
	生存	死亡	計
治療群	12 (57%)	9	21
対照群	0 (0%)	21	21
計	12	30	42

- 仮説の設定
 - 帰無仮説:2つの群の生存率は等しい
 - 対立仮説:2つの群の生存率は等しくない
- 有意水準:0.05
- 検定統計量
 - カイ二乗検定
- 統計的有意性の評価
 - P値:0.0007
 - P値<有意水準 → 帰無仮説を棄却(対立仮説を採択)

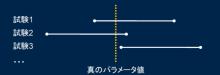
P値とは?

- ▶ 帰無仮説とデータの乖離(ギャップ)の指標
 - 20名にある治療を行って14名が成功した
 - 帰無仮説:真の成功確率 (p) = 0.5



信頼区間とは?

- > 信頼係数95%の信頼区間:95%信頼区間
- 観測値からある計算式に従って区間を求めたとき、 それが真のパラメータ値を含む確率が0.95である



ある1つの区間には、真のパラメータ値が含まれるか、含まれないかのどちらか 仮想的繰り返しの長い系列の中で、真のパラメータ値が含まれる相対頻度が0.95

1

信頼区間の例

- > 20名にある治療を行って14名が成功した
 - ・成功確率の点推定値:14/20 p̂=0.70
 - 成功確率の区間推定値(正確法)
 - 95%信頼区間: 0.46~0.88

近似法:95%信頼区間

 $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 2 \times (p \mathcal{O}$ 標準誤差)

90%信頼区間:0.49~0.86



90%信頼区間の下限がわずかに0.5より小さい

帰無仮説(p=0.5)の検定の片側P値:0.058に対応している!

Nature Cell Biology, Guide to authors and referees, 2009.12

統計ガイドライン (3)

- 複雑な生物学的実験に関して、測定の独立した反復数は実際的理由から制限されなければならないことが多いため、非常に小さいnを用いた統計的尺度が日常的である
- しかしながら、あまりに小さい標本サイズに適用された統計 的尺度は有意にならず、誤った有意水準を示唆する



無理に検定をしない! 記述と推定(信頼区間)を中心に!

統計ガイドライン (4)

- この理由のため、n<3に関してエラーバーは提示すべきでない。そのかわりに、各実験からの実際の個別データをプロットすべきである
- n<5のときは、個別データをエラーバーの横にプロットすべきである



n=2のときに標準偏差を計算してはいけないという理論的根拠はない

Nature Cell Biology, Guide to authors and referees, 2009.12

18

統計ガイドライン (5)

- n、P、および適用された検定の基本的な記述は、図の説明において提供すべきであり、統計的方法論のさらなる考察は方法の節において提供すべきである
- 著者はある特別な検定の利用を正当化し、自分たちのデータがその検定の仮定に従うかどうかについて説明しなければならない
- ▶ 以下の3つの誤りが特によく見られる
- 多重比較:単一のデータセットについて複数の統計的比較を 行う際、著者は第I種の過誤率の上昇を防ぐために、α水準 をどのように調整したかについて説明すべきである、あるい は(一連のt検定ではなく分散分析のような)複数の群に関し て適切な統計的検定を選ぶべきである

Nature Cell Biology, Guide to authors and referees, 2009.12

10

多重性の問題

同じデータに対して、検定を繰り返し行うと、検定の回数に比例して、第1種の過誤(差がないものをあると言う誤り)が大きくなる、という問題

例:20本に1本の割合で不良品のワインが混入している貯蔵庫からランダム に2本ワインを選んだとき、2本のうちどちらかが不良品である確率は?

答え:2本ともまともなワインである確率 = (1本がまとも) × (1本がまとも) = 0.95 × 0.95 = 0.9025 どちらか1本が不良品の確率 = 1 - (2本ともまとも)



= 1 - 0.9025 = 0.0975

もし、40本に1本の割合で不良品が混入していたら?

20

例:40本に1本の割合で不良品のワインが混入している貯蔵庫からランダム に2本ワインを選んだとき、2本のうちどちらかが不良品である確率は?

答え:2本ともまともなワインである確率 = (1本がまとも) × (1本がまとも) = 0.975 × 0.975 = 0.950625 どちらか1本が不良品の確率 = 1 - (2本ともまとも)



= 1 - 0.950625 = 0.049375

不良品の割合=有意水準

有意水準0.05/2=0.025で2回検定すると 全体としての有意水準は約0.05になる (ボンフェローニの方法)

その他多くの調整方法(テューキー、ダネット、シェフェ、...)

١,

統計ガイドライン (6)

正規分布:多くの統計的検定は、データがほぼ正規分布していることを要求する。従って、これらの検定を用いるとき、著者はデータの正規性についてどのように確かめたかについて説明すべきである。データがその検定の仮定に合っていないならば、ノンパラメトリックな代替法をかわりに用いるべきである

ヒストグラムや箱ヒゲ図で分布の形をみる



大きく右に裾を引いている、あるいは大きな外れ値があれば、対数変換 それでも左右対称にならなければ、ノンパラメトリック手法

Nature Cell Biology, Guide to authors and referees, 2009.12

統計ガイドライン (7)

小さい標本サイズ:標本サイズが小さい(約10未満)とき、著者は小さい標本に対して適切な検定を用いる、あるいは大標本検定の利用を正当化しなければならない。



無理に検定をしない! 記述と推定(信頼区間)を中心に!

この指針は臨床データに最も適切であるが、著者が統計的誤りを最小限にすることを手助けするチェックリストは、以下から入手可能である(次ページ)

Nature Cell Biology, Guide to authors and referees, 2009.12

Checklist of statistical adequacy

- > 試験開始時および各解析に関するnが報告されている
- 標本サイズ計算あるいは正当化が提供されている
- 統計的方法が十分に記述されているとき、それらのうち明らかに不適切なものはあるか?
- > すべての統計的検定に関するαが提供されている
- 検定が片側であるか両側であるかが特定されている
- データがその検定の仮定に合っているかどうかが記載されている
- > 主要な解析に関して実際のP値が報告されている
- 統計的尺度(平均、標準誤差、標準偏差など)が報告されており、それらは明確に標識されているか?
- > すべての比較において解析の単位が明確に記載されている
- 非正規分布の可能性のあるデータセットを記述するのに、あるいは標本サイズが非常に小さいときに、平均と標準偏差が用いられている?
- > 一般的でないあるいは複雑な統計的方法の説明
- ▶ もしあれば、データ除外の説明
- > 最初のnと各解析に関するnの乖離についての理由が説明されている
- > 治療割付(もしあれば、ランダム化)の方法が説明されている
- > 多重検定に関する調整が考察されている

http://www.nature.com/ncb/pdf/checklist_of_statistical_accuracy.pdf

:

質問

- 反復データ(replicates)と独立繰り返し実験(independently repeated experiments)の関係について
- > 反復データと独立繰り返し実験の境界は、生化学的な実験では比較的明確です。 これに対して、細胞や個体レベルの実験では、生物学的ばらつきや実験的なばらつきが大きい上に、現実的な測定回数が限られることなどから、反復データと独立繰り返し実験の境界をどのように扱えばよいのか戸惑うことがあります。
- 測定値が、マウス体長のような場合と、マウスを屠殺して計測するような場合、さらには、各マウスからMEFを確立してその活性値を測定するような場合とでは、独立した実験としては要求される水準が異なると考えられますが、このような場合の「反復データ」と「独立線」)及し実験」の取扱に関して、考慮すべき点や犯してはいけない注意点などをご教授頂きたいと思います。
- このことは、統計学における「独立性」の定義は非常に明白であるにも関わらず、 生物学実験において「独立」であるという確定が困難であるため、各研究領域に おける現実に即したコンセンサスがある</u>と考えておりますが、そのような解釈でよ いのでしょうか?

仮想例

例えば、KOマウスとWTマウスの比較実験で、1回の実験につきWTとKOマウスそれぞれ10匹ずつを測定したとします

1回目 (n=20)		2년 (n=	目 20)	3回目 (n=20)		
WT	KO	WT	KO	WT	КО	
150	145	151	146	149	144	
140	135	141	136	139	134	
140	135	141	136	139	134	
130	125	131	126	129	124	
130	125	131	126	129	124	
130	125	131	126	129	124	
130	125	131	126	129	124	
120	115	121	116	119	114	
120	115	121	116	119	114	
110	105	111	106	109	104	

<u>方法1</u> 方法2 <u>方法3</u> 各実験の平均値を代 表値として、比較する 第1回目の結果のみ 全てのデータを合わせ (他の2回も同様で て解析する あったと併記) 各国の平均値をとって比較 n=3 n=10 n=10 n=30 n=30 n=3 n=3 独立実験(別の実験) 同一実験(反復データ) 独立実験の結果を併合 とみなす とみなす 可能とみなす (ただし、各実験内のバ ラツキを無視)

25

「同一実験(反復データ)」とみなす場合(前頁の方法2)

1元配置分散分析(2群の場合はt検定と同じ)

要因	自由度	平方和	平均平方	F値	P値
群(WT vs. KO)	1	375.0	375.0	3.004	0.088
誤差	58	7240.0	124.8		
全体	59	7615.0			

「独立実験」とみなして併合する場合

2元配置分散分析(乱塊法)

要因	自由度	平方和	平均平方	F値	P値
群(WT vs. KO)	1	375.0	375.0	2.917	0.093
ブロック(実験)	2	40.0	20.0	0.156	0.856
誤差	56	7200.0	128.6		
全体	59	7615.0			

参考文献

- > Cumming G, et al. Error bars in experimental biology. Journal of Cell Biology 2007;177:7-11.
- > 山中伸弥、青井貴之、佐藤俊哉. 分子生物学、生化学、細胞生物学における統計のポイント. 蛋白質 核酸 酵素 2009;54:1792-1801.



29