依 頼 論 文

今日から使える統計学講座：歯学統計チェックリスト

新谷　歩

Primer of Statistics in Dental Research

Ayumi Shintani, PhD, MPH

抄　録

論文投稿からアクセプトまでのプロセスはEBM医療が盛んになる近年厳しさを増す一方です．医学雑誌の最高峰であるNew England Journal of Medicineでは，年間約5000本の論文が投稿され，最終関門の統計査読に到達した5%の論文の中から更に20%が統計手法の不備でリジェクトされます[1]．査読者が何を見ているのかが事前に分かれば，過酷な査読地獄を生きぬくための強力な武器となるのではないでしょうか．British Dental Journalが提供している統計チェックリストを紹介しながら，歯学研究を行う際に留意したい10個のトピックについて解説します[2]．

和文キーワード

統計，統計テストの選択法，多変量解析，多重性，リピート解析

トピック 1．グラフと記述統計量

1. 平均値を示す時はエラーバーなどばらつきを表す指標も共に示す．
2. エラーバーが何に基づくか表示はされているか？
3. データのばらつきを示す時は標準偏差を用いる．
4. 平均の精度（正確性）を示すとき標準誤差（SE）を用いるが，平均± SEのエラーバーは67%の信頼区間を示すため混乱を招く．平均± 2SEは95%信頼区間を示すので，こちらを使用すること．
5. 同じ患者から異なる時間に測定されたデータはそれが容易にわかるように線で結ぶと良い．
6. 散布図のように個々のデータ値を示すグラフを使用すると良い．

上の1から4は，データをグラフ化する際にばらつきの指標を用いることの重要性を説いています．グラフでデータのばらつきを表す時にエラーバーをよく用いますが，通常の統計ソフトではエラーバーは，標準偏差，標準誤差，信頼区間の3つについて描くことができます．どの指標についてエラーバーが用いられたかによって統計的有意差への判断が変わるので，エラーバーを用いるときはそれぞれが何かをまず表記し，それぞれの指標の使い方を知ることが大切です．

標準偏差（Standard Deviation, SD）は各データ値から平均までの差の平均です．5，6，10の3つの値の標準偏差は2となり，“5と平均の7，6と平均の7，10と平均の7の差の平均”で計算します．標準偏差はデータが正規分布に従うとき，集めてきたデータの95%が平均プラスマイナス2× 標準偏差の範囲に入るというデータの記述に用いることができます．例を挙げると，100人の被験者の年齢の平均が50歳，標準偏差が10歳の場合，95%の被験者の年齢が30歳から70歳の間にあると解釈できます．このように標準偏差はデータの記述には便利な指標ですが，薬剤の比較など統計的な推定には直接用いません．平均プラスマイナスで描かれた図1Cのエラーバーは重なっていますが，2群間に有意差が存在します．標準偏差が用いられている場合は，エラーバーが重なるかどうかということで有意差の判断はできません．これに対して標準誤差（Standard Error, SE）は，推計したい統計

ヴァンダービルト大学医療統計学部
Associate Professor, Department of Biostatistics, Vanderbilt University Medical Center
Director of Biostatistics, Center for Health Services Research, Vanderbilt University Medical Center

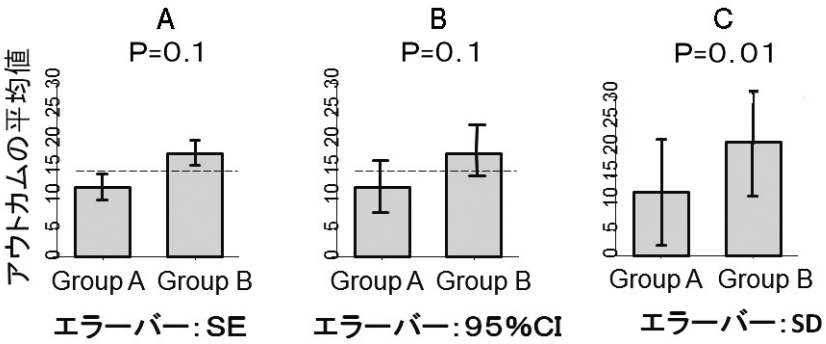

図1　エラーバーと統計的有意差の関係

量（平均など）の分布を表す指標で，見たい暴露（薬剤など）の効果推定の精度を表します．2群間に差があるかどうかという統計的推計には標準誤差を用いますが，ここで注意が必要です．平均プラスマイナス標準誤差のエラーバーが重なるかどうかで有意差有無の判断はできません．有意差の判断は，標準誤差を約2倍したエラーバーを用い，これを95%の信頼区間（Confidence Interval, CI）と呼びます．図1Aは二つのエラーバーは重なっていませんが，有意差はありません．それに対しエラーバーが2倍の標準誤差（95%の信頼区間）で描かれている図1Bでは，エラーバーが重ならない時は群間に有意差があると判断できます．（注：少し重なる場合は有意差が出ることもあるので注意が必要です．そのためBDJでは2群間の比較はそれぞれの群の平均に信頼区間を計算せず，群間の差に信頼区間を計算するよう勧めています―トピック3参照）．この様にエラーバーはグラフとp値をつなぐ非常に役立つ指標です．多くの論文ではエラーバーは平均とともに多くのグラフで用いられていますが，エラーバーが何を示すかですら記述されていない論文が非常に多いのが現状です．群間比較に用いるエラーバーは必ず信頼区間を用い，患者背景などデータの記述には標準偏差を用いるようにしましょう．

トピック2．外れ値

- 大部分のデータとかけ離れた値を持つデータを外れ値とよぶ．データを除く場合はいかなる場合でも表記する．
- 正当な理由がない限り，外れ値を解析から除いてはならない．
- データの排除は解析結果に影響するため，外れ値を除く前後で解析結果がどう変わるかを調べる感度解析が効果的である．

最近データの捏造などをよく耳にしますが，研究者に好ましい結果が出るようにデータを削除することは捏造と同様固く禁じられています．また，結果によらなくても外れ値だからと言ってデータを削除するのも“外れ値が完全なデータエラーによる”など正当な理由がない限りタブーです．外れ値が生じた場合は，ノンパラメトリック法など外れ値に左右されないような統計手法を用いる，ログなどの数学変換で外れ値の影響を小さくするなどの工夫をすれば，正しく解析に用いることができるので解析から外す必要はありません．どうしても除く必要がある場合は，外れ値を除く前と後で解析結果を比べるなど感度解析が重要です．

トピック3．信頼区間とp値

- p値が5%未満で有意差が出たからと言って差が本当にあるというわけではない．p値が小さいということは偽陽性の確率が少ないというだけのことである．
- 比較研究では，p値のみでなく信頼区間を用いることが望まれる．
- 信頼区間はそれぞれの群について別々に計算するのではなく，群間の違いについて計算することが望まれる．

統計解析に必ず出てくるp値は研究対象要因（例：薬剤）にまったく効果がないのに，“観測された差，またはそれ以上の差が全くの偶然で起こる確率”であり，言い換えれば“間違って差を検出する偽陽性の確率”と考えられます．つまりp値は小さければ小さい程間違いの確率が下がるので良いというわけです．このp値を大きくする要因は観測データの（1）差が小さい（2）症例数が少ない（3）バラつきが大きい事が挙げられます．そして，慣習的にp値が5%より小さいと，観測された差は偶然によるものではなく，本当に差が

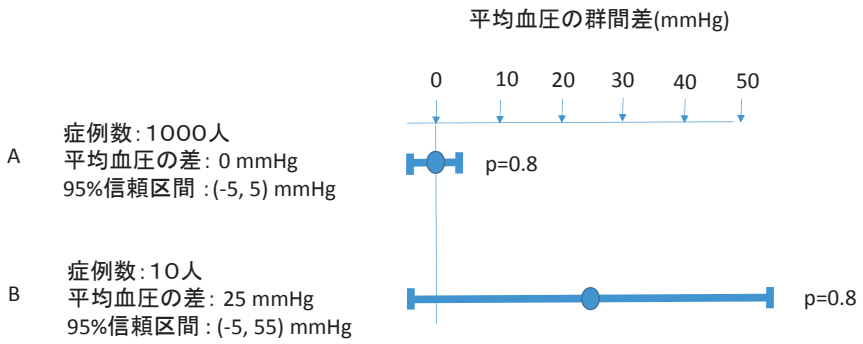

図２　信頼区間とｐ値の関係

あったのだろうという結論を導きます．一方，ｐ値が5% より大きい場合は，“差があるというエビデンスがない”ということにすぎません．これを間違ってｐ値が５％より大きいと“同じである”と解釈すると大きな間違いを犯すことになります．例えば，２つの薬剤の同等性を示す研究でそれぞれの群で３人の被験者からデータを採ったとします．症例数が小さいので間違いの確率であるｐ値は当然大きくなります．ｐ値が大きいことは差が無いこと（同等）からくるのではなく，単に症例数（エビデンス）不足からくるわけですから，当然同等性の判断にも症例数が不足，つまりエビデンスが欠落しているということになり，大きいｐ値をもとに同等性をいうことはできません．ｐ値が5% より大きいからと言って,“同じだ”ということはやめましょう！（でもこれは意外に難しいのです.）

ここでｐ値に替わってよく使われる信頼区間について解説します．図２は血圧の平均を新薬と既存薬間で比べた研究で，症例数の大きい研究（A）と小さい研究（B）のそれぞれの研究から得られた結果を信頼区間とｐ値を使って示しています．信頼区間の真ん中の値が実際に観測された血圧の平均値の群間差を示し，信頼区間の幅がその推定の正確性を示しています．95%信頼区間に“差がない値，この場合はゼロ”を含めば，ｐ値は５%より大きくなり，有意差は認められないと解釈できます．データにバラつきが多い，または症例数が小さいと区間の幅が大きくなり，ゼロを含みやすくなるので有意差が出にくくなるわけです．

A 研究でも B 研究でもｐ値は 0.8 だったので，差があるというエビデンスはどちらの研究でも認められませんでした．ここで問題です．もし差が出ればこれらの薬が認可されるとすると，あなたは A 研究で使った A の薬剤を製造した A の会社の株を買うか，B 研究で使った B の薬剤を製造した B 社の株のどちらを買うでしょうか？

それでは信頼区間を見てみましょう．A と B ではｐ値は同じですが，信頼区間は明らかに違いますね．

A の場合は症例数は十分大きかったけれど，もともと２群間の差がゼロで，信頼区間の真ん中にゼロが来ているので，これ以上症例数を増やして研究をつづけたとしても，信頼区間はゼロを中心として小さくなるだけなので，差は一向に出ません．一方 B は血圧の平均の差が 25mm Hg と大きいけれど症例数不足で区間の幅が広くなったせいで，ゼロを含んでしまったけれど，症例数を増やして研究をやり直せば，区間が狭くなり，ゼロを含まなくなるので，差が出る可能性が高い．と，私は B 社の株を買うことにしました．

この様に，ｐ値のみを用いると差が出なかったことは，実際に差が無いのか，症例数不足からくるのかわかりませんが，信頼区間ではそれが区別できるようになるのです．ｐ値には必ず信頼区間を添えて発表するようにしましょう！

トピック４．同等性，一致性

- 有意差が出なかったイコール違いがないと解釈してはならない．（有意差が出ない＝偽陽性の確率が高い＝差があるというエビデンスがない）
- 統計的有意差を同等性の考察に用いてはならない．
- 対応のあるｔ検定や独立なｔ-検定から得られたｐ値を用いて一致性の考察はできない

同等性を示すにはｐ値を用いず信頼区間を用いるようにします．同等性は研究プロトコルにあらかじめ定義されている同等性マージン，例えば血圧の差がプラスマイナス５mmHg 以内であれば同等だと認めると決め，95%の信頼区間がすっぽりとそのマージンに入っている場合にのみ認められます．それに対して統計的な有意差は，信頼区間に“差が無いことを示す値”が入っているかどうかのみで決められます．図 3C を見ると，

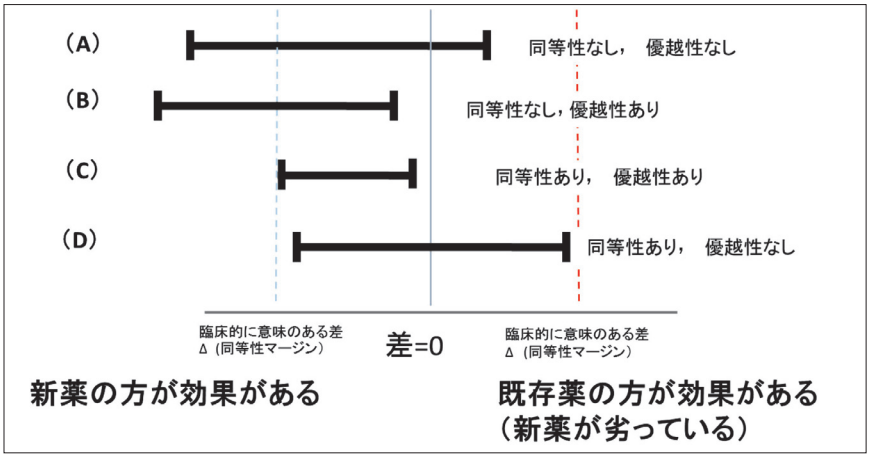

図3　群間差の信頼区間と結果の解釈例[3)]

信頼区間はゼロを含まず，と同時に，同等性マージンに入っているので，統計的には差が認められましたが，臨床的には同等であることがわかります．統計的な有意差は，差がゼロでないかどうかを判断しているにすぎず，それが臨床的にどういう意味を持つかは統計的な判断ではないのです．p値に惑わされないで，臨床的な差に注目しましょう！

トピック5．仮説と多重検定によるp値の補正

- アプライオリ（事前計画されていた）仮説の数は多すぎないか？
- ポストホック（事前計画のない）仮説に対して多重検定の補正は行われたか？
- ポストホック仮説よりもアプライオリ仮説により重きを置く．ポストホック仮説結果は探索的な仮説として提示し，確証的エビデンスとはしない．
- 多重検定が年齢のように順序のあるグループで行われているときは，トレンドテストを考慮する．

見すぎによる出過ぎは，比較群が多い時の対比較（A群とB群，B群とC群，A群とC群のそれぞれ2つずつの比較）など，一つの研究で有意差検定が幾度も行われる場合に問題となります．それぞれの比較で計算されるp値は“間違って差があるとみなす確率”なので，下手な鉄砲も数撃ちゃ当たる方式で，比較を何回も行うと，差が無いのにあるとしてしまう1型エラーの確率がぐんぐん高くなります．そのようなエラーの増加を防ぐために，解析が多く行われた時（例：p値が複数計算されたとき）は，それぞれの解析で有意差が出にくくなるように調整しようというのがボンフェローニ法などの多重検定によるp値の補正です．（各調整法の詳細は医学界新聞2941号多重検定を参照ください．）日本補綴歯科学会の第122回学術学会の抄録集を読むと，皆さん多重検定の補正はきちんとなされているようで，いかにこの問題に苦労されているかが実に良く伝わってきます．

それでは，皆さんがすでに行っている多重性の補正はこれで十分と言えるでしょうか？　歴史上，アウトカムが連続変数で，比較群が3つ以上の場合に多重性の補正法についての多くの開発がなされたことを背景として，現在，多重性の補正は分散分析のオプションとしてソフトウェアに取り入れられています．そのせいで，“アウトカムが連続変数で比較群が3つ以上の研究”で多重性の補正は非常に多く行われているようですが，多重性の問題はそれ以外の場面でも起こりえます．以下に多重性の問題がおこるシナリオをリストアップしています[4)]．

- ●主要評価項目が多重
- ●比較群が多重
  - ▲多重用量をプラセボと比較
  - ▲説明変数が多重
- ●単一評価項目につき解析集団が多重
- ●単一評価項目につき統計手法が多重
- ●中間解析
- ●メタ解析
- ●特定の患者に絞って比較をやり直す層別解析
- ●時間ごとの解析

このように，多重性の問題は一つの研究でp値を二つ以上計算するようなありとあらゆる場面で出くわしてしまうのです．同一研究内でp値が複数個計算された場合は必ず補正し，そのたびにp値をどんどん膨らませていくと，統計的有意差などめったにでそうもあ

りませんね．ここで問題となるのが補正のし過ぎによる“差があってもないとしてしまう”２型エラーの膨張です．はたして，多重検定による p 値の補正は必ず行わなければならないのでしょうか？

多重性のため p 値（または有意水準）を補正するか否かは専門家の間で大変な議論を呼んでいます．歴史的に有名な疫学者の Rothman は“補正すべきではない，なぜなら，現在多用されている補正法は，ある特定の患者群でさえも治療に効果がないという非現実的な仮定の下でのみ成り立つ”[5] イギリスの有名な統計家である Senn は“全ての検定を結果の有意差によらず，事前プロトコルに規制された順序ですべて報告するのであれば多重性の補正は行わなくてもよい”[6]．“一般に，1 型エラーの膨張は検定間の相関に依存する．臨床的なアウトカムが多重の場合は，それらのアウトカムは互いに相関することが多いので，ボンフェローニ法など一般的に多く用いられている補正法では，補正のし過ぎが問題となる”[6]，と述べています．これらの論文を引用して，補正をしないと言い切っている論文をレビューしたことも多くありますが，だからと言って，多重性の問題を無視してい良いというのも，なかなか同意できず，私はいつも複雑な思いで“それでも何らかの考慮をしてほしい”とコメントをしています．この時の私の心情を言い当てた米国衛生局のガイドラインからの引用を，ここに紹介します．“多重検定の補正をいつどのように行うか確固としたルール作りは困難であるが，何らかの基本的なガイドラインを作成することは可能である．しかしそのガイドラインは単なる提案に過ぎず，どうしても守るべきルールとして用いられてはならない(つまりケースバイケースである)．しかし，私たち統計家は，多くの p 値が計算されたため有意差がでたというような研究に対して，平常心ではいられない．したがって多くの統計家がほとんどの研究で多重比較の補正を行う．現実的な解決策として，彼らと我々のちょうど中間的な視点で論じる必要があるのではないか”[7]．

補正が必要と言ったり，必要でないと言ったり，いったいどうすればよいのでしょうか？残念ながら，上の引用からも言われているように，この問題は現在でも議論の真っただ中で，クリアーなガイドラインが存在しません．多重性の問題を完全に取り去るには，ベイズ法や，尤度法のように仮説検定および p 値を用いない解析法を模索する必要があり，最近の統計学界のトレンドはもうそちらに向かって進んでいるようです．

見すぎによる出過ぎの問題は，「とりあえずデータを採って，差が出たものだけを発表してしまえ」，というような事前計画のない研究で，より問題となります．このような解析は“データドレッジング”，“チェリーピッキング”，“フィッシング”と呼ばれ，他の研究者が同様の研究を繰り返したときに同じ結果がでないという再現性に問題が起こります．再現性のない研究は科学とは呼べません．そもそも p 値が 5％以下ならば有意差があるとしてよしという，マジックナンバーの“5％”は，今から 90 年以上前に RA Fisher が単なる思い付きで科学的な裏付けもなく考えたのですが[8]，その時代は一つの検定をするのに，何時間（または何日？）もかけて計算を行っていました．計算を行う前，もしくはデータを採取する前に仮説の絞り込みが，かなりに入念に行われていたことは明らかです．Fisher が今のようにコンピュータが簡単に何万もの p 値を数秒で計算するのを見れば p 値は 5％で評価すべきとは言わなかったでしょうね．

BDJ のチェックリストでは，解析が事前計画されたものかそうでないものかということと多重検定の問題を組み合わせて述べています．きちんと生物学的な背景を踏まえて少数に絞り込まれた仮説が事前に計画されているかどうか，事前計画のない（ポストホック）仮説は，多重検定による p 値の補正を行い，解析結果は探索的な仮説として提示し，確証的エビデンスとはしないなどです．見すぎによる出過ぎの問題を防ぐために，“すべて”の解析は事前計画の時点で熟考しましょう！

それでは見すぎによる出過ぎの問題を防ぐために，いったいどうすればよいのでしょうか？私は教科書として利用している，Rosner のアドバイスを引用して，“多重検定の補正は比較群の数が比較的多く，対比較が事前に絞り込まれていない場合はなされるべきであるが，群数が比較的少ない場合，また対比較が事前に絞り込まれている場合は，ANOVA など比較群の少なくとも一つが違っている場合に有意差を検出できるようなグローバルな検定で有意差が確認された場合は，それぞれの対比較を補正せずとも良い”[9] と授業では教えていますが，これもまたケースバイケースなのです．

ここで最後に，欧州医薬品庁（European Medicines Agency, EMEA）や日米 EU 医薬品規制調和国際会議（International Conference on Harmonization；ICH）から出しているガイドラインの中で多重検定の補正を必要としない研究例を紹介します．

アウトカムが２つ以上の解析で多重性の補正を必要としない例[4]

□　複数のアウトカムのうち，事前プロトコルで臨床

的な意義に沿ってその重要性の順位が，主要，2次，3次などと表記されている場合は多重性の補正は行わずともよい．この場合，上位にランクされたアウトカムで有意差が出なかったときは，それ以降にランクされたアウトカムでの有意差は確証的なエビデンスとしては用いられない．

□　研究対象要因の評価のため複数のアウトカムで全て有意差が出る必要性があるとき．

□　一方が主解析，他方が感度解析と表記されている場合．

比較群が3つ以上の解析で多重性の補正を必要としない例[4]

□　3-arm Gold-Standard デザインでは，新薬群と既存薬群とプラセボ群の対比較で優越性が認められ，同時に，既存薬を新薬間での非劣性が成り立つことが新薬の効果のエビデンスとして必要であるがこの場合，多重性の補正は行ずとも良い．

□　探索研究で2剤の混合処方の優越性を検証する場合，混合薬剤の優越性が双方の薬剤について成立されなければならない．この時多重性の補正は行わずとも良い．

□　用量の違いなどで比較群が多重である場合

○　有効性及び安全性を保障する用量を確定する場合は，グローバルテストや容量反応を見るトレンドテストなどで有意差が確認された後に，参照群とそれぞれの用量の対比較で多重性の補正を行う必要がある．

○　効果的な用量の決定を目的とするのではなく，効果と用量の相関関係の検出を目的とする場合はトレンドテストなどで相関関係を確認した後それぞれの用量と参照群間の対比較を行うが，この場合多重比較による補正は必要ではない．

トピック6．正しい統計テストの選択

- 全ての解析法が正しく表記されているか？
- 複雑な解析法が用いられている場合，その使用の意義が参考文献と共に表記されているか？
- 不適当な解析法が用いられていないか？
- 解析法の前提条件は満たされているか？
- ログ変換などでデータ変換が行われているか？
- ノンパラメトリック法が用いられた場合，それを正しく表記する．

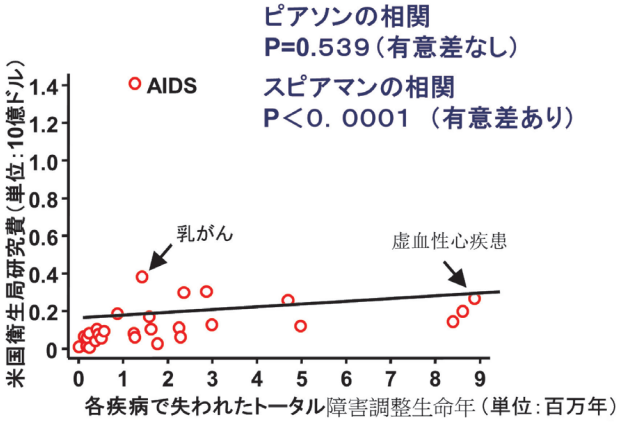

図4　異なる統計テストで結果が変わる例[10]

図4は米国衛生局の研究費を各疾病ごとに縦軸に表わしたもので，横軸は各疾病で失われた障害調整生命年のトータル年数を示しています[10]．変数の正規性を前提条件とするピアソンの相関テストではp=0.539と明らかに相関なしとなり，分布を仮定しないノンパラメトリックテストのスピアマンの相関では，p値は0.0001より小さく，かなり強い相関があることがわかります．通常コストや年数のデータは，小さい値にデータが集中してはいますが，かなり高い値も少数ではあるが存在するような左に偏ったデータなので，正規分布など左右対称の分布を前提条件とする検定を使うことはできません．この場合は分布を想定しないスピアマンの結果が正しいことがわかります．

単変量解析では，多くの検定でノンパラメトリック法が存在するので，できればそちらを用いるようにしてください．分布によらないということは正規分布でも正しい結果が出るということですから，ノンパラメトリック法を用いるといちいち分布を調べなくても済むのでとても便利です．

しかし，線形回帰モデルのように回帰分析となると，ノンパラメトリック法が簡単には使えないことが多く，その場合は，ログ，ルート，累乗などの数学変換でアウトカムを正規分布に変えてから解析を行うことがよくあります．ちなみにコストや，年数，日数のように左にゆがんだデータでは，ログ変換をするとデータを正規分布にすることができます（注：データがゼロを含むとログ変換後に欠損となるので，その場合はすべての値にゼロと最低値の中間の値を足すなどの工夫が必要です）．

スピアマンの検定では米国衛生局は人命を救うために研究費を使っているという結果がでて，ピアソンの検定ではそうでないという結果になりました．このよ

表 1　統計手法を選択する際のポイント[12)]

差 / 相関	比較データ間の対応性	変数の種類（正規性）	比較する群の数	サンプル数	適切な統計手法
差	対応なし	連続変数（正規分布）	2	総数 30 以上	スチューデントの t 検定
			＞2	1 群 15 以上	一元配置分散分析
		連続変数（非正規分布）/ 順序変数	2	制限なし	マン・ホイットニーの U 検定* ウィルコクスンの順位和検定*
			＞2	制限なし	クラスカル・ウォリス検定*
		2 値変数	2	総数 20 未満	フィッシャーの正確確率検定*
			≧2	総数 20 以上	ピアソンのカイ 2 乗検定
		打ち切り例のある 2 値変数（生存時間解析）	≧2	イベント総数 10 以上	ログランク検定
	対応あり	連続変数（正規分布）	2	15 組以上	対応のある t 検定
			＞2	15 組以上	反復検定による分散分析
		連続変数（非正規分布）/ 順序変数	2	制限なし	ウィルコクソンの符号順位検定*
			＞2	制限なし	フリードマン検定*
		2 値変数	2	制限なし	マクネマー検定
相関（関連性）		連続変数（正規分布）		総数 20 以上	ピアソンの相関係数
		連続変数（非正規分布）/ 順序変数		制限なし	スピアマンの順位相関係数*
		2 値変数		制限なし	ケンドールの順位相関係数* カッパの相関係数（一致性）

＊ノンパラメトリック検定，それ以外はパラメトリック検定を示す。

うに同じデータでも統計テストを替えるとこんな違いが出るなんて，驚きですね．大変苦労して集めたデータを最終関門である統計テストで間違って解析してしまうなんて，非常に残念です．

英国の著名な統計家であり医師でもある Douglas Altman 氏は「誤った解析手法を故意にまたは知らず知らず使う，正しい解析手法を誤った方法で使用する，解析結果を間違って理解し発表するなどにより不当な結果を導く．このようなことが数え切れないほど多くの一般的，専門的な医学研究論文で日常的に行われていて，これは間違いなく医療スキャンダルだ．」と言っています[11)]．皆さんの大切なデータを正しく解析し，必要としている治療がそれを待ち望んでいる患者さんに届けられるよう，正しい統計テストを選択する際のポイントを以下に挙げています（表 1）．表 1 はポイントを左から順に答えていけば最終的に正しい答えに到達します．

ポイント 1．差を見るのか，相関を見るのか？（例：比較群が 2 つ以上の場合は差）
ポイント 2．群間の差を見るとき比較群が対応しているか？一人の患者から 2 回以上測定されているか？
ポイント 3．アウトカムの変数の種類は何か？（例：連続変数，2 値変数，順序変数）
ポイント 4．アウトカムが連続変数の場合，分布にゆがみはあるか？
（例：歪みがない時はパラメトリックテスト，歪みがあるときはノンパラメトリックテスト）
ポイント 5．群間の差を比較する場合，比較群の数は 3 つ以上か？
（例:2 群の場合は t 検定，3 つ以上は分散分析）
ポイント 6．症例数は十分大きいか？
（例：大きい場合はパラメトリックテスト，そうでない場合はノンパラメトリックテスト）

トピック 7．解析ユニットとリピートの解析

- 一人の口腔内から繰り返しとられたデータは，別々の患者の口腔内からとられたデータに比べて類似しているので，この様にデータに対応がある場合は，対応のある t 検定のように対応（繰り返し）を考慮した統計テストを用いる．

統計テストの選択ポイントの一つに比較群の対応というのがありました．データが各患者さんで時間をおいて 2 回以上計測されているとき，または各患者さんから複数の歯のデータを解析するなど，各患者さんから 2 つ以上の計測値を使って解析が行われるときにデータは対応していると考えられます．データを親子からとる，または，患者さんの特性に応じて似た特性を持つコントロールと比較するマッチング研究でもデータの対応が起こります．この他にも，同じ医師に治療された患者さんの予後が他の医師に治療された患者さんのデータに比べ類似しているなど，歯科医師間の技術的な差もデータに対応を与えることになります．このように，歯学研究ではデータの対応がいたるところで起こっているのです．

p値とはデータを基に計算された確率です．データが対応しているかどうかによって確率計算が変わってくるのでデータ間に対応があるかどうか，用いる統計手法はそのデータの対応を考慮して確率計算を行っているかどうかがとても重要です．それでは，実際に計算してみましょう．

例えば，AとBの2人の研究者がいます．Aさんも Bさんもそれぞれが来月アメリカに行く確率が10%だとします．それでは両者が来月アメリカに行く確率を計算してみましょう．これは簡単ですね，10%×10%=1%です．それではAさんとBさんが結婚していて2人がいつもアメリカ出張に一緒に行くとしたらどうでしょうか？Aさんが行けばBさんも行き，同様にBさんが行けばAさんも行くので，両者が行く確率は10%×100%+10%×100%=20%です．データの対応が確率に影響するのが分かりますね．

次は各患者さんの口腔内から複数本の歯のデータを採る研究で考えてみましょう．Aの薬剤で処理した100本の歯と，Bの薬剤で処理したもう100本の歯を比べる臨床実験を考えます．A剤では虫歯の総数は20本，B剤では10本でした．B剤が効いたかどうかのp値を次の3つのシナリオのそれぞれについて計算してみましょう．

シナリオ1．A剤もB剤も100人に1本ずつデータを集め，A剤では20人，B剤では10人が虫歯になった．
シナリオ2．A剤もB剤も10人に10本ずつデータを集め，A剤では10人全てが2本，B剤では10人全てが1本ずつ虫歯になった．
シナリオ3．A剤もB剤も10人に10本ずつデータを集め，A剤では2人10本全部が虫歯でほかの8人は虫歯なし，B剤では1人が10本すべてが虫歯でほかの9人は虫歯なしだった．

ある患者さんの歯はその他の患者さんの歯と関連しないので（あくびはうつるといいますが，虫歯はうつらないと考えます．）他人同士は全く関連せず独立です．つまりAさんが虫歯になる確率はBさんに無関係で，AさんBさんのそれぞれが虫歯になる確率は10%だとします．でも，Aさんの歯が虫歯になればAさんの他の歯も虫歯になる確率は10%よりかなり高いはずです．（私の娘2人は何年か前のハロウィーンの後チョコレートを自分の部屋に隠しこっそり食べて，1か月ほどでそれぞれ5，6本虫歯ができてしまいました．1本の歯がむし歯になると他の歯も虫歯になる確率は私の経験からしても10%よりずっと高いはずです）．混合効果モデルなどデータの対応を考慮した回帰モデルは，この確率の違いを考慮してp値を計算します．

シナリオ1は一人の患者さんから歯は一本ずつなので，歯のデータ間に対応が無いので，スチューデントのt検定を行います．シナリオ2，3は一人の患者さんから複数の歯のデータを集めているので，解析は混合効果モデルを用いてp値の計算をしました．すべてのシナリオでA剤とB剤で虫歯の総数は変わりませんが，A剤とB剤の違いを表わすp値を計算すると，シナリオ1はp=0.052，シナリオ2はp<0.0000001，シナリオ3はp=0.54となりました．かなり違いますね．対応のあるデータには対応を考慮した統計手法を用いましょう．

トピック8．患者背景と交絡の補正

- 比較群の患者背景にバランスがとれているか？
- 比較群の患者背景が偏っている場合，それが解析で考慮されているか？
- 偏りのある変数がアウトカムに影響を及ぼす場合それら全ての変数が多変量回帰分析で補正されているか？
- 多変量回帰モデルを用いて補正を行うとき補正された変数は全て表記されているか？
  またそれらの変数はどのように選択され，どのように解析に用いられたか表記する．

心臓病リスクの高い患者を対象に，アスピリンが死亡率を軽減するかどうかを調べた前向きのコホート研究で，6000名の患者をアスピリンを使用した群とそうで無い群に分け，平均3年後の死亡率を比べました．アスピリン使用による死亡リスクを表わすハザード比は1.08でp値が0.5という結果になりました[13]．この研究結果を見て，あなたならアスピリンは処方しないという結論を出すでしょうか？これが比較群の特性が揃ったランダム化臨床研究であればそういう結論になるかもしれませんが，これは観察研究です．比較群が揃っていない可能性が大なので，どういう患者さんがこの研究に入っていたかを調べてみましょう．すると，アスピリンを使用した群では平均年齢が62歳と，そうで無い群に比べ6歳高く，男性の割合も77%対56%，既存の心臓病歴は70%と20%と，アスピリン使用群の方が予後に関して不利な人々が多く入っていたようです．ここであなたならアスピリンはまだ効果がないと結論付けますか？アスピリンを使っていた人々は予後の悪い人たちだった，つまり理由があってアスピリンを使っていたということです．予後がもと

もと悪かったのにもかかわらず，3年後の生存率が変わらなかったということは，アスピリンが効いていたからではないでしょうか？このように，年齢，性別，既存の心疾患のような他のリスクファクターの効果と混ざりあってしまい，本当に調べたい暴露の効果が，分からなくなってしまうことを交絡と呼んでいます．この場合の有効な調整法が，重回帰分析による交絡因子の補正です．死亡までの時間をアウトカムとした生存率解析で群間の違いを補正した(考慮に入れた)結果，調整ハザード比は0.5となり，アスピリンによって死亡リスクが半減することがわかりました．重回帰分析では，年齢，性別，既存の心疾患の違いからくる死亡率への影響を考慮し，それを差し引いたうえでアスピリンの効果を調べることができます．このような重回帰分析による補正は，考え方として少しも難しいものではなく，皆さん知らず知らずに行っているごくごく日常的な考え方です．

私の父は40代始めのころまで，自分は歯磨きを一度もしたことがないのに虫歯は一本もないと自負していました．これに対して母は歯磨きを一生懸命するのにもかかわらず虫歯がありました．幼心に私は歯磨きをすれば虫歯になると信じていたでしょうか？幸いにも，頭の中で多変量解析を行って，父に虫歯が無く，母にあるのは，歯磨きのせいでなく，歯の丈夫さなど他要因が影響したのだろうと補正を行っていたので，歯磨きに手を抜くことはありませんでした．父はさすがにこの後歯周病に苦しむことになり，私は歯磨きの重要性を再確認したのでした．

補正されていない単変量の解析は交絡が存在する場合まったく信頼性がなく，ランダム化の行われていない観察研究では交絡は必ずと言ってよいほどおこるので，観察研究では解析は必ず重回帰分析を用いましょう！

トピック9．因果関係

- 統計的相関のみで因果関係を直接的に示唆できない．ランダム化比較試験では因果関係の示唆は容易であるが，観察研究では，因果関係は統計以外の概念も重視して考察されなければならない．

私がアメリカに来て間もないころ，泌尿器科の医師から，生魚を食べると胃がんが増えるから気を付けるよう言われました．理由は，日本人は他の国に比べて魚を生で食べることが多く，胃がんも多いからだそうだとか．この他にもアイスクリームを食べると人を殺したくなるなんてデータも出ています．これらはエコロジカルファラシーと呼ばれ，国や地域ごとの統計量を見るときによく起こる間違いですが，夏場にアイスクリームの消費が増えるのと同時に殺人事件の件数も増えることから，データを誤って解釈してしまった例です．本当に生魚を食べると胃がんになるかどうかを調べるためには，生魚を食べる人と食べない人にランダムに振り分け，10年後の胃がんの率を比べる実験を行えばよいわけですが，観察研究などではこのように実験はできないので，様々なバイアスが生じます．バイアスには選択バイアス，情報バイアス，交絡と大きく分けて3つあります．例えば胃がんと生魚の研究をアンケート調査をもとに行った場合，アンケートの回答者には健康意識の高い人が多く，その中には胃がんの家族歴を持つ人も多かったとします．そしてそれらの人々は健康に留意し肉より生魚をよく食べていた．これは選択バイアスと呼ばれ，このバイアスに気づかずデータを解析すると，生魚と胃がんの間に本当はまったく相関が無くても，生魚を食べるとまるで胃がんになるような間違った結果が出てしまいます．

また，ケースコントロール研究と呼ばれる別の研究手法をもちいて胃がんになった患者さんとそうでない人に聞き取り調査で過去に生魚を食べていたかどうかを調べたとき，胃がんになった人のほうが癌になる前に何を食べていたのかよりよく覚えていたということが多いといったことがよくあります．この場合も両者に相関が無くても，生魚と胃がんの間に間違った相関が出てしまいます．これが情報バイアスです．

交絡は，アスピリンの例でも説明しましたが，生魚を食べない人の群には野菜も食べない人が多く，胃がんになるかならないかは生魚ではなく野菜を多く食べるかどうかに関連していたという場合に，両者に相関が無くとも，生魚を食べると胃がんが軽減されるというような結果になりえます．この3つのバイアスのうち，最後の交絡はデータ解析で調整ができますが，最初の二つは解析では対処できません．解析で何とか調整ができる問題は，解析をやり直して再度投稿してくださいということになりえますが，解析でどうにもならない“取り返しのつかない”バイアスについては即リジェクトになることが多いのです．いろいろな場面で遭遇するバイアスを熟考し研究計画を組み立てましょう！

トピック 10．相関と一致性

- ピアソンやスピアマンの様な相関に関する検定で一致性を考察できない．

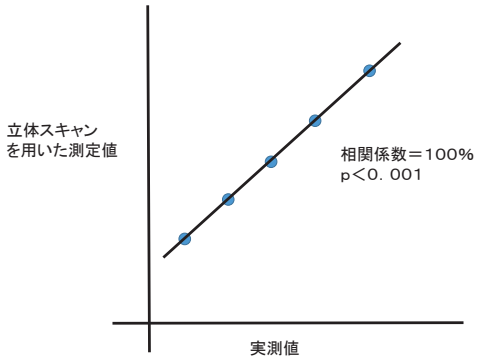

図5　相関係数と一致性の関係

二つの値が一致しているかどうか調べる研究では，ピアソンやスピアマンの相関テストは用いません．例えば，図5ではA（例：立体スキャン）とB（例：実測値）の二つの異なった方法で測定したあごの大きさの測定値をグラフで示しています．AとBで測った長さが丁度1 cmずつ全ての患者さんでずれていますが，相関は100%となります．データは完璧に一致していないのに相関100%なんておかしいですね．p値も無限にゼロに近い小さい値になりますが，同じ患者さんを二つの方法で計測しているので，関連があって当然，つまりp値が小さくても当然意味を成しません，p値が小さくて当然なのです．

別の例で説明します．ある疾患の有無をA歯科医とB歯科医が診断し，それら診断の一致性を比べる場合，症例数が十分なのにもかかわらずp値が大きいつまり相関がないということはどういうことでしょうか？同じ患者さんを診ているのに，A医師の診断とB医師の診断がまるで別々の患者さんをそれぞれが見ているかのように違っている．つまり，どちらかまたは双方がでたらめに診断を行っている？どうやら彼らの歯科医師免許を調べる必要がありそうです．どうでしょう，もう納得していただけましたか？ここでもp値は小さくて当然なのです．p値が小さいことを示しても“So, what?”と言われて一笑に付されるだけなのです．この様な理由で一致性の解析にp値は使いません．

連続変数の一致性を見るには，相関係数でなく，クラス内相関係数（Intra-Class-Correlation，ICC）を用います．ICCは患者間のデータのバラつき（分散）がデータのトータルなバラつきに占める割合で計算されます．データのトータルなバラつきは患者間からくるバラつきと患者内で起こるバラつきを足したものなので，患者内でのバラつきが小さい，つまり一人の患者で測定した患者内でのデータが一致していればしているほどICCは1に近づきます．

$$\text{クラス内相関係数，ICC} = \frac{\text{患者間のデータのバラつき}}{\text{トータルのバラつき（患者間のデータのバラつき＋患者内のデータのバラつき）}}$$

ICCのほかに，連続変数の一致性には同等性の解析と同様，差の信頼区間（p値は使いません）を用いることもできます．Periagoらの発表した論文では立体CI画像を用いて計測したNaとANSという頭がい骨上の2骨点の長さが実測値と一致するかどうかを，対応のあるt検定を用いて調べました[14]．立体CT画像の測定値のほうが0.83 mm短いという結果が出ました．95%信頼区間は（0.55 mm，1.1 mm）でした．信頼区間がゼロを含まない，つまりp<0.05なので著者たちは一致しないという判断をしたようですが，これは間違いです．トピック4でも触れましたが，この場合はp値ではなく立体CT画像と実際の差の信頼区間の大きい方の値が，1.1 mm，というのが臨床的に同等とできるマージンに入っているかどうかで一致性を見ていきます．

二つ目の例のA歯科医とB歯科医の診断で疾患有り無しの2値変数の一致性を調べるにはカッパ係数（κ）を計算し，p値ではなくカッパの係数が0.75以上でExcellent，0.4以上0.75未満はGood，それ以下はMarginalと一致性の良しあしを判断します[9]．

いかがでしたか？駆け足で，統計の基本的な10個の概念を紹介しました．統計学って意外と簡単でしょう．私たち人間が日常頭で行っている複雑な分析を数式に置き換え確率計算をしているだけなのですから，人間の思考をすこし観察してみると統計も見えてくるのです．折角苦労して集めたデータです．研究最終ステージのデータ解析法を正しく理解し，研究成果をどんどん世に出していただきたいと思います．その為に，今後もできるだけ分かりやすく皆さんに統計学を紹介していきたいと思います．もっと深く学びたい方には，http://blog.livedoor.jp/shintaak/ で私の講義ビデオを閲覧できますので，皆さんの研究に役立てていただけましたら心より嬉しく思います．

文　献

1) Harrington, D., *Improving Your Chances of Success at New England Jouranl of Medicine (NEJM): Some Guidelines for Statistical Reporting*. October 5, 2011: Vanderbilt University, Weekly Seminar, Department of Biostatistics.
2) British Dental Journal Statistical Guidelines, http://www.nature.com/bdj/authors/Guidelines/statistics.html.
3) 新谷　歩．同等性と非劣性の解析．週刊医学界新聞第2971号；2012年.
4) EMEA, *The European Agency for the Evaluation of Medical Products: Evaluation of Medicine for Human Products. Points of Consider on Multiplicity Issues in Clinical Trials*. 2002; London. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002989.pdf
5) Rothman, K., *No adjustments are needed for multiple comparisons*. Epidemiology, 1990; 1: 43-46.
6) Senn, S., *Statistical Issues in Drug Development*. 2008, Chichester, England; John Wiley & Sons Ltd.
7) Proschan, M.A.W., *Practical Guidelines for Multiplicity Adjustment in Clinical Trials. Office of Biostatistics Research, National Heart, Lung and Blood Institute*. Control Clin Trials 2000; 21: 527-539.
8) Fisher, R.A., *Statistical Methods for Research Workers*. Statistical Methods for Research Workers. 1925, Edinburgh: Oliver and Boyd.
9) Rosner, B., *Fundamentals of Biostatistics*. 7 ed. 2010: Cengage Learning.
10) Gross, C.P., G.F. Anderson, and Powe, N.R., *The relation between funding by the National Institutes of Health and the burden of disease*. N Engl J Med 1999; 340(24): 1881-1887.
11) Altman, D.G., *The scandal of poor medical research*. BMJ 1994; 308(6924): 283-284.
12) 新谷　歩．統計テストの選び方．週刊医学界新聞第2927号；2012年.
13) Gum, P.A., et al., *Profile and prevalence of aspirin resistance in patients with cardiovascular disease*. Am J Cardiol 2001; 88(3): 230-235.
14) Periago DR, Scarfe WC, Moshiri M, Scheets JP, Silveira AM, Farman AG. *Linear Accuracy and Reliability of Cone Beam CT Derived 3-Dimensional Images Constructed Using an Orthodontic Volumetric Rendering Program.* Angle Orthodontist 2008; 78(3).

著者連絡先：新谷　歩
Department of Biostatistics, Vanderbilt University Medical Center 2525 West End Ave.; Ste. 11000 Nashville TN
Tel: (615) 322-1357
Fax: (615) 343-4924
E-mail: ayumi.shintani@vanderbilt.edu

「本論文は，Journal of Prosthodontic Research 58(1) に掲載された論文 “Primer of Statistics in Dental Research Part I”，Journal of Prosthodontic Research 58(2) に掲載された論文 “Primer of Statistics in Dental Research Part II” をもとに，日本補綴歯科学会誌用に書き直されたものである」