Japanese Journal of Clinical Oncology投稿に際しての統計解析結果のレポートに関するガイドライン

福田治彦[1]、大橋靖雄[2]

[1]国立がんセンター中央病院内科、同研究所がん情報研究部

[2]東京大学医学部保健学科　疫学・生物統計学教室

はじめに（日本語版作成にあたって）

本稿は本誌２７巻３号に掲載された英文のガイドライン「A Guideline for Reporting Results of Statistical Analysis in Japanese Journal of Clinical Oncology」の日本語版であり、英文同様、Editorial Officeの依頼により執筆したものである。英文は投稿論文に準じた体裁で本誌に掲載いただいたが、その日本語版である本稿はインターネットホームページで公開すること、および希望により郵送配布する小冊子として利用していただくことを目的として作成した。構成や基本的な内容は英文のガイドラインと同じであるが、直訳だと読みにくい部分を意訳し、かつ英文では表現しきれなかった箇所を追加の解説として加えたので、英文ガイドラインの和訳そのものではないことを予めお断りしておく。

ご助言・ご批判をJJCO事務局宛にお寄せいただきたい。

過去３年間の本誌（Japanese Journal of Clinical Oncology）をReviewしてみたところ、掲載されたoriginal articleのうち76%で何らかの統計学的手法が用いられていた。このことからも明らかなように医学論文における統計学の重要性が高いことは改めて言うまでもないが、実際に医学研究者の論文作成に際しての実用的な統計解析に関するガイドラインは非常に少ない（1-3）。本ガイドラインは、Editorial officeの依頼により、本誌に投稿される臨床研究論文の作成に際して、統計解析の結果をまとめる上での実用的な手助けとなることを目的として作成した。本ガイドラインは網羅的なものではないので、実際に統計学的手法を適用する際には、末尾で紹介したような教科書や雑誌論文等を参考としていただきたい。ただし、我が国は人的リソースの面で欧米と比べて状況は厳しいとは言うものの、やはり適切な研究デザインや統計解析を行う上での最良の方法は、研究の初期段階で経験のある生物統計家に相談することであり、もちろん本ガイドラインはその替わりとなるものではない。

本ガイドラインは以下の構成になっている。

１章では、本誌に投稿する上で記載されていることが必要な項目をあげ、若干の解説を加えた。２章ではより一般的な事項として、医学論文で見られる統計学に関する誤解について、臨床医の陥りやすい“落とし穴”として解説した。３章は本ガイドライン作成に当たって参考とした教科書や他のより詳細なガイドラインを推薦図書として紹介した。

＜目次＞

1 投稿論文における基本的事項（Basic Requirements for Study Reports）

1-1 Methods Section

1-2 Results Section

2 統計学におけるよくある落とし穴（Common Pitfalls in Statistics）

2-1 統計学的有意性

2-2 関連と因果関係

2-3 分布と前提となる仮説

2-4 カテゴリー変数

2-5 相関と回帰

2-6 生存時間解析

2-7 多変量的手法

2-8 最後に

3 推奨される教科書およびより詳細なガイドライン

3-1 英文

3-2 和文

4 引用文献

—

1 投稿論文における基本的事項（Basic Requirements for Study Reports）

JJCOに投稿される論文では以下の事項が記述されている必要がある。

—

１-１　Methods Section

１-１-１　研究の種類

臨床研究はTable Iに示すように分類されうるが、投稿される研究がどれに該当するかをabstractまたはtitleの中に示すことが望ましい（後者がより望ましい）。

Table I. Type of Study

1. observational study
    1) case series study
    2) cross-sectional study
    3) longitudinal study
        a) case-control study
        b) cohort study (prospective, retrospective, historical)
        c) nested case-control study
2. intervention study (trial)
    1) controlled study
        a) parallel (randomized, non-randomized)
        b) sequential (self-controlled, cross-over)
        c) historical control
    2) uncontrolled study

１-１-２　endpointの定義

“ endpoint ” は、研究の対象となる個人単位で観察された結果の指標として用いられる場合（individual endpoint）と、研究全体での結果を評価する上での指標（study endpoint）として用いられる場合がある。例えば個々の患者の “ 死亡 ” が前者であり、これをまとめた “ ○年生存率 ” が後者である。“ endpoint ” が “ 目的 ” の意味で誤用される例を多くみかけるが、臨床研究の領域では、元々の “ 終点 ” すなわち目的やゴールという意味ではなく、“ 結果の指標（outcome measure） ” の意味で用いられることが一般的である。

論文ではどのstudy endpointまたはどのindividual endpointを用いたかをMethods sectionで明確に記述する。研究目的によってさまざまなstudy endpointとindividual endpointの組み合わせが有り得るが、代表的なものをTable IIに列記した。

研究の主な目的を達成するための指標として、ひとつまたは少数の主要なstudy endpointを選び“ primary endpoint ”とする。残りのendpointは“ secondary endpoint ”として記述する。それぞれのendpointごとに定義、評価の方法、評価の時期および間隔を明確に記述する必要がある。

Table II. Commonly used individual endpoints and study endpoints

Individual endpoint	Study endpoint
death	proportion surviving (so-called "survival rate") (proportion of survivors) mortality
duration of survival / time to death	survival (distribution)
recurrence	proportion of recurrence (so-called "recurrence rate")
duration of relapse-free survival / time to recurrence	relapse-free survival (distribution)
(objective tumor) response	(objective tumor) response rate
toxicity	toxicity spectrum dose-limiting toxicity maximum tolerated dose
adverse event (complication)	morbidity
diagnosis of cancer	prevalence, incidence
test positive for diseased subjects test negative for non-diseased subjects	sensitivity specificity

１-１-３　データ収集の方法

データを収集した方法、例えば、カルテを順に検索したのか（survey）、特別なケースカード（Case report form/Case record form）を用いたかどうか、病院情報システムやその他のデータベースの検索によったのか、などを明記する。

１-１-４　対象の選択と除外の規準

対象の選択規準（selection criteria または eligibility criteria）は通常、inclusion criteria（組み入れ規準）とexclusion criteria（除外規準）から成る。inclusion criteriaは研究の外的妥当性（external validity）または一般化可能性（generalizability）を示す（研究の結果を適用または外挿できる範囲= target populationを示す）ために定められるものであり、exclusion criteriaはその研究におけるstudy endpointの評価を困難にするような対象を除外したり（がんとは関係のない合併症で死亡する可能性が高いと予測される症例は治療の延命効果の評価対象として不適当）、研究に組み入れること自体が安全でないような対象（患者）を除外する（心毒性が予期される試験治療の場合、元々心室性不整脈を有する患者は、通常では問題がない程度の毒性であっても致命的になることがあり得るため、適切に除外されなければならない）ために設定されるものである。

１-１-５　割り付けの方法

介入試験（Intervention study）においては、（治療の）割り付けに用いられた方法が詳細に記述されなければならない。例えば無作為割り付けをおこなったのかどうか、cross-overかどうかや、予め決められた順序で割り付けられた（sequential）のかどうかが示されなければならない。記述に際しては、「何が」「どうやって」割り付けられたかを具体的に記載する。

１-１-６　バイアスを避けたり研究の妥当性を保証するために取られた方法

バイアスを避けるためや、研究の妥当性を確保するためになされた無作為割り付けや盲検化（マスキング）、マッチング等の方法を用いた際は詳細に記述する。もし評価が何らかの盲検化（マスキング）の下になされた場合は、「誰が」「何について」マスクされたのかを明確に記述する。対照群（コントロール）がある場合は、それが「なぜ」「どうやって」選ばれたかについて記述する。

１-１-７　統計学的手法

原則として研究の中で用いられた統計学的手法はすべて参考文献とともに列記するが、以下にあげたような非常に一般的な統計学的手法については文献引用は必要ではなく、Results sectionで結果と共にどの方法を用いたかのみ示せばよい。（ただし、主たる解析に用いた場合は Methods section に簡単に述べておく方が望ましい）

ｔ検定

χ２乗検定（Fisher 直接確率検定）

Wilcoxon検定またはMann-Whitney U検定

相関と線形回帰

上記以外のすべての統計学的手法はMethods sectionで引用文献と共に記述されなければならない。ただし、スタディデザインや統計学的手法に関する引用においては、それらの手法が最初に発表されたoriginalな文献ではなく、むしろ教科書やReview articleで“標準的な手法”として紹介されたものを用いるべきである。

注）

1) １つ以上のオプションがあるような手法（t-test：paired or unpaired、分散分析、順位相関など）を用いる場合は、どのオプションを使ったかがわかるように明記されること。

2) 片側検定か両側検定かの区別が明記されること。

3) 対応のあるデータに対しては、対応のないｔ検定（unpaired t-test）ではなく対応のあるｔ検定（paired t-test）が用いられるように、対応のない（通常の）Wilcoxon検定の替わりに対応のあるWilcoxon検定が、通常のχ２乗検定の替わりにMcNemar検定が用いられること。

１-１-８　統計解析に用いたソフトウェア

例えばSAS（Statistical Analysis System）やSPSS（Statistical Package for the Social Science）のような統計解析用のパッケージソフトを使った場合は，そのソフトウェアの名称とそのバージョンを記載すること。その場合も用いた統計学的手法はそれぞれ明記されなければならない（SAS の場合はプロシジャ名、SPSS の場合はコマンド名とともに明記すべき）。

目次に戻る

—

１-２　Results Section

１-２-１　対象の記述

要約統計量（例：平均値、奏効率、〇年生存率など）を算出することは、いつの場合でも元のデータが持っている情報量のかなりの部分を捨てていることになる。従って解析結果のレポートでは，単に一部の要約統計量のみを示すだけではなく，可能な限り効果的に個々の対象のデータも示すようにこころがけるべきである。

省略されるべきでない重要なデータには以下のようなものがある。

解析対象の総数（分母）

医学論文において、推定値（割合、パーセンテージ、比率）の分子は常に示されるが、その時の分母が必ずしも常に明確に示されているとは限らない。すべての推定値のレポートにおいては常に分母が明確に示されなければならない（例：奏効率の分母が“すべての登録患者”なのか、“すべての適格患者”なのか、レジメンの“全コースの治療を受けた患者”なのか・・というような区別）。本誌では分母が明らかでない推定値はacceptされないので注意されたい。

解析から除外された対象

解析の過程における対象の除外は、どんな場合においても比較可能性や一般化可能性を損なうことになる。何らかの理由で解析から除外された対象がある場合は、解析から除かれた対象の数とその理由（例：登録後の不適格性判明、規定のプロトコール逸脱やプロトコール治療の途中中止、追跡不能など）は漏れなく記載（リストでもよい）されなければならない。

治療に際しての有害事象の詳細

治療法や診断法の評価は常にその有効性と安全性のバランスによりなされなければならないので、解析結果のレポートでは治療による有効性のデータだけでなく安全性のデータが注意深く詳細に報告されなければならない。安全性のデータとは有害事象や有害反応※（化学療法に際しての薬物有害反応や放射線治療

や外科切除術の合併症）の頻度と重篤度のことである。

※従来「副作用（side effect）」と呼ばれてきたが、“ side effect ”には主作用以外の好ましい作用も含まれるため今後は使われないようになる。従来の「薬剤の副作用」は「薬物有害反応（adverse drug reaction）」、それ以外の治療による有害な反応は「有害反応（adverse reaction）」、治療との因果関係の有無に関わらず治療に際して生じた好ましくない事象は「有害事象（adverse event）」と呼ぶことが、ICH（International Conference on Harmonization）関連の各種ガイドラインでも規定された。ただし日本語訳あるいは日本のガイドラインでは法制度の整合性上、「副作用」という言葉が「薬物有害反応」のかわりに使われている。

１-２-２　p値の示し方（Exact p-value）

統計学的に有意でなかった検定の結果を、一律に“ N.S. ”や“ not significant（有意差なし） ”と示すよりも、例えばP＝0.23のような実際のp値（Exact p-value）として示す方が望ましい。

表や図の中で示されるp値は本文の中で繰り返される必要はない。通常、p値は小数点以下３ケタ（P=0.025、P=0.003）まで示されるが、p値が0.1より大きい場合には、P=0.25のように小数点以下２ケタで充分である。非常に小さいp値の時はP < 0.001やP < 0.0001のように示して構わない。“ p ”でも“ P ”でもどちらでもよいがP（大文字イタリック）が国際標準である。

１-２-３　点推定値と信頼区間

さらにp値のみをレポートするよりも推定値（連続変量の差の平均値、カテゴリー変数の値の割合やオッズ比を含む相対危険度、ハザード比など）の信頼区間をいっしょに示すことが望ましい。

注)

1) さらに検定統計量の実際の計算値も示す方が望ましい。例えばｔ検定の場合は、変数の平均値、SD（標準偏差）、ｔ統計量、自由度をp値とともに示す。

2) 平均値は通常、元のデータを越える有効ケタ数を示す必要はないが、標準偏差や標準誤差は元のデータより１ケタ多く示す。割合をパーセントで示す場合、56.47%というように小数点以下２ケタまで示す必要があることは稀であり、特に例数が少ない時には38.6%のように小数点１ケタを示すことも必要でない場合も多い。極めて少ない例数における割合の表示でパーセントを用いることは避けるべきである（例：1/1=100%、2/2=100%）。ｔやχ^2、rのような検定統計量は小数点以下２ケタで充分である。ただし、これらの注意点は結果の表示の際の話であって、解析の途中では（切り上げ、切り捨て、四捨五入のような）まるめは行うべきではない。

目次へ戻る

２　統計学におけるよくある落とし穴（Common Pitfalls in Statistics）

このセクションでは、医学論文でよく見られる統計学の誤用、つまり研究者が陥りやすい“ 落とし穴 ”について述べる。

—

２-１　統計学的有意性

２-１-１　“ 有意差がない ” と “ 差がない ” は違う

医学論文でもっともよく見られる統計学的な基本概念についての誤解は “ 統計学的有意性（有意差） ” についてであるが、それは「統計学的有意差がない」ことを「差がない」つまり同等であると勘違いしていることである。

検定（統計学的帰無仮説検定）は「帰無仮説（例えば両群に差がない）が正しいという仮定の下で観察された差がもっともらしいかどうか」を評価するひとつの手法である。p値自体は、帰無仮説が正しいという仮定の下で観察されたデータ（またはそれ以上に偏った結果）が偶然に得られる確率を示すに過ぎない。“ 統計学的有意差 ” が得られた場合の解釈として２とおりの解釈が可能である。ひとつは「（差がないという）帰無仮説が正しくて、たまたま珍しい事象が観察された」であり、もうひとつは「（差がないという）帰無仮説が間違っていたと考え、差があるとみなす。」である。「統計学的有意差がある」すなわちp値が充分に小さいとき（$P < 0.05$や$P < 0.01$）に、「差がある」としてそれに基づく推論を進めるのは、ふたつの解釈のうちのひとつを便宜上選択する、つまり帰無仮説を棄却し対立仮説を採用する（差があると考える）ということであり、決して前者の「たまたま珍しい事象が観察された」ことを否定するものではない。

一方、充分に小さいp値が得られなかった（$P > 0.05$）場合、帰無仮説は棄却されないが、それは「差がない」ことを意味するのではなく、「差があるとは言えない」、つまり「見込んだ差があるのかないのかわからない」ということを意味するに過ぎない。

もし単に「有意差がない」ことで「差がない」と言えるとするなら、「差がない」ことを示そうとする場合はできる限り少数の例数で研究すればよいことになるが、これが誤りであることは明らかであろう。

「差がない」ことを積極的に示したい場合は「同等性の検定」と呼ばれる手法を用いる必要があるが、本ガイドラインの範囲を越えるので詳述しない。

２-１-２　p値と有意水準

P=0.05というのは、統計学的有意水準（type-I error level：α）としてもっともよく用いられるカットオフ値であるが、前項で説明した原則から考えて、例えばP=0.04の結果とP=0.06の結果に異なる解釈を加える理由はないことは理解されよう。つまり、（特に探索的なデータ解析においては）検定結果の解釈において盲目的に0.05に固執することに実は根拠はなく、それよりも実際のp値とそれぞれの推定値の信頼区間を示してp値の大きさに応じた解釈を加えることの方が望ましい。

２-１-３　統計学的有意性と臨床的意義

例数が充分に多い（sample sizeが大きい）場合は、小さい差であっても “ 統計学的有意 ” となりうる。仮想的には1500例以上を登録する無作為化比較試験では５年生存率で５％以下に相当する小さな（生存時間分布の）差を検出できることになる。こうした小さな差が仮に “ 統計学的に有意 ” であったとしても、個々の患者の治療においては臨床的に必ずしも重要であると言えない場合があるだろう。一方、例数が少なくて、非常に大きな差であっても “ 統計学的有意差 ” が得られない場合があるが、それが臨床的に重要な差であるとみなせるならば単に “ 有意差がない ” として無視されるべきではなく、さらなる検討の対象とされる（新たな臨床研究を行う）べきであろう。研究者は「統計学的有意差」と「臨床的な重要性」を明確に区別する必要があり、“ 臨床的に意味のある差 ” を検出し得るような研究デザインを組むべきである。

２-１-４　多重性

ひとつの研究の中で検証的手法とくに検定方法を繰り返し用いること、例えば、たくさんのサブグループ解析を行うこと、一つの問題にたくさんの検定手法を用いること、たくさんの変数（endpoint）を検定すること、および同じ変数を（時間の経過に沿って）何度も検定することは、それぞれ偽陽性（例：本当は差がないのに差があるとしてしまう）の結果を得る確率を高くするが、このことを（検定の）多重性と呼ぶ。従ってこのような状況下ではスタディデザインおよびp値の解釈において注意が必要である。

多重性を避けるための方法としては以下のような例がある。

１）より単純なendpointを用いる

検定の多重性を回避するもっともよいアプローチはたくさんの統計量（検定）の替わりに可能な限りひとつかごく少数の統計量（検定）を用いることである。例えば経時的に測定される連続変量の解析において、“ 最高値に達するまでの時間（time taken to reach a peak） ”（図1a）や、“ 一定の値以上であった時間（time above a given level） ”（図1b）などは有用なendpointであり得る。こうした手法は、よく知られている例として、薬理学的パラメータであるAUCやCmaxとして用いられているが、もちろん他の領域にも応用可能である。

図1a

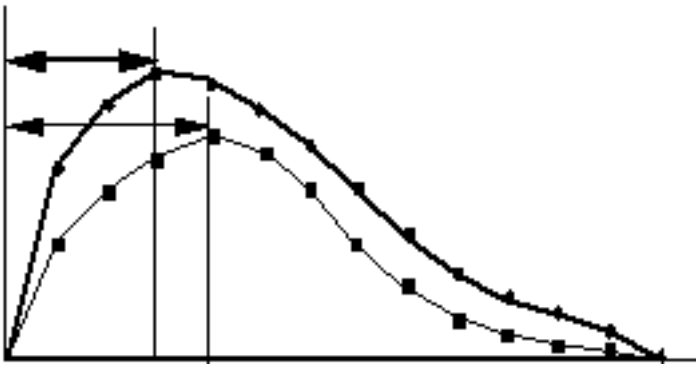
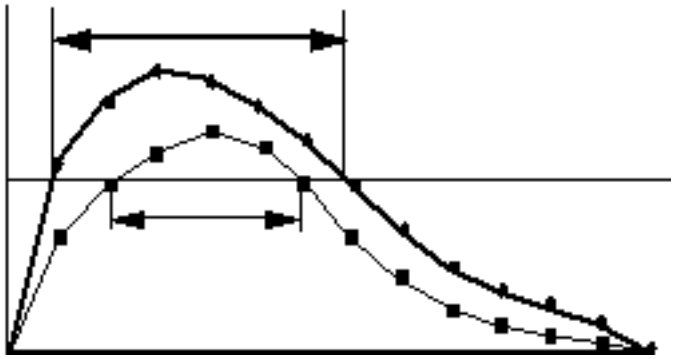


図1b



２）「有意」という言葉を用いずに控えめな解釈を加える

他のアプローチとしては、実際のp値（Exact p-value）をそれぞれ示し、個々に控えめな解釈を加えることである。この場合、P=0.05に固執するのは不適切なので“ 有意 ”とか“ 有意でない ”という言葉も必要でない。

３）有意水準を調整する

単純でad hocな（後知恵的な）方法として、Bonferroni correctionが用いられることがある。これは、もしn回の検定を行うとして、研究全体としての有意水準（type-I error level：α）を0.05に保ちたい場合は、ひとつの検定のp値がα／nを下回ったら“ 有意 ”とする、という方法である.。例えば５つのサブグループ解析を行う場合なら、p値が（0.05／５ =） 0.01を下回った場合のみ“ 統計学的に有意 ”と呼ぶ、と言うことになる。

この他にも、相関における多重性を回避する特殊な方法やいくつかのBonferroniタイプの有意水準の調整方法があるが、これらの特別な手法を用いる場合は、どの手法を選ぶのが適切かが仮説の構造や研究の目的によって異なるため、生物統計家に相談することをお奨めする。

２-１-５　確定的か探索的か？

確定的なデータ解析か、探索的なデータ解析かは常に区別されなければならない。

（データが得られる前に）予め設定された仮説と解析計画に従って行われた解析のみが“確定的データ解析”とみなされるが、それのみが結論的な結果を与えることができる。一方、それ以外の（予め仮説や解析計画が立てられて始められたのではない）すべての解析は“探索的データ解析”であり、それらは、さらにそれ以上の研究によって検討されるに値する新しい仮説を見つけることのみに用いられるべきものである。ほとんどの観察研究で行われるデータ解析は探索的と言える。

多くの研究者はこれら２つの主要なタイプの解析（または研究）の違いを誤解または混同している。検定で同じように“有意”な（$P < 0.05$）結果が得られても、それが解析（データ収集）の前に立てられた主要な仮説の検定の場合と、事後的に（ad hocに）行われたサブグループ解析における副次的な仮説の検定の場合とは全く違った意味合いを持つ。データ解析の中で得られたサブグループに対する解析の結果は、他の別の研究で確認（confirm）されるまでは常に注意して扱われなければならず、「結論が出た」かのように扱ってはならない。

探索的データ解析の弱点は先述した多重性である。すべての探索的データ解析の結論は「さらなる検討を行う必要があり、かつその価値がある」か、「これ以上検討の価値がない」かのどちらかである。いかに“高度に有意な”結果が得られようとも基本的には同じである。なぜなら、予め計画された以外の探索的な解析では、自由に何通り（何十通り、何百通り）もの計算が可能であるため、意図したかそうでないかに関わらず“差のありそうな”比較に注目して進められた（“fishing”と呼ばれる）可能性が常にあり、“たまたまその時だけ偶然に出た（有意）差”を拾っているに過ぎないことを否定できないからである。

探索的データ解析における結果の妥当性・再現性を高める上で、cross-validationは有効な方法である。もし例数が多い場合、初めからデータセットを（無作為に）２つに分け、ひとつのデータセットで探索的解析を行って仮説を探索し、その仮説がもう一方のデータセットでも成立するかどうかを確認する、という方法である。もし例数がそれほど多くない場合でも、全体で得られた結果を２つかまたはそれ以上のデータセットに（無作為に、あるいは時系列で）分割してみて、得られる結果が変わらないことを確かめておくことは結果の信頼性を高める上で重要（必要）である。

目次へ戻る

—

２-２　関連と因果関係

“統計学的な関連”はそれ自体因果関係の直接的な証拠となるわけではない。多くの医学論文では、結果の解釈において“統計学的な関連”が“因果関係”と混同されている。例えば壊血病とビタミンＣ欠乏のような直接的な因果関係が介入研究（無作為化比較試験）を経ずに証明されることは稀である。しかし、実際には多くの臨床上の問題や疫学的なテーマに対して、つねに介入研究が容易に行えるわけではなく、因果関係の多くは観察研究によって検討されることになる。観察研究において見られた統計学的な関連をもって因果関係を議論する際には、統計学的根拠によるのではなくむしろ非統計学的な根拠に基づいて総合的に考察される。因果関係を示唆する非統計学的な根拠についてはさまざまな提案があるが、ここではCampbell[4]の例を紹介する（Table III.）。

Table III　因果関係を示唆する非統計学的な根拠

---

1) 一貫性（consistency）

異なる集団を対象とした他の研究で同じ結論が導かれていること

2) もっともらしさ（plausibility）

結論が生物学的にもっともらしいこと

3) 用量反応関係（Dose-response）

用量-反応関係（強い曝露がより高いリスクに関連している）が見られること

4) 時間経過（Temporality）

曝露の増減が疾患の発生率の増減に先行して見られていること

5) 関連の強さ（Strength of the relationship）

大きな相対危険度、オッズ比、ハザード比であること

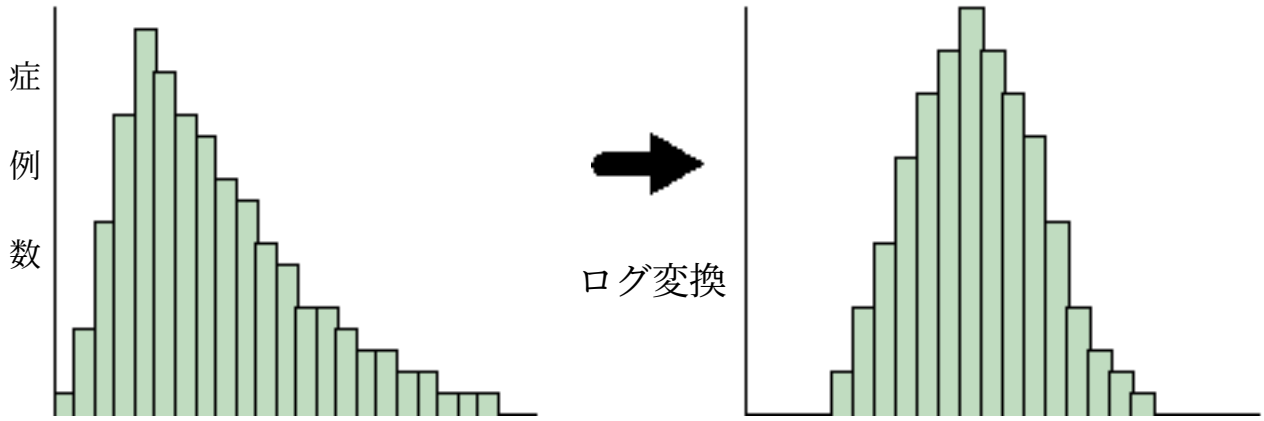
目次へ戻る

—

２-３　分布と前提となる仮説

例えばｔ検定を適用する際には、通常は、「概ね正規分布とみなせる」や「標準偏差で示される両群のバラツキがほぼ等しい」などの分布に関する前提を置く。このようにデータのバラツキをある確率分布で表して、この分布のパラメータに関する推論を行おうとするアプローチをパラメトリックな手法と呼ぶ。パラメトリックな手法を用いる場合は、もしこれらの前提が満たされない場合、検定結果の妥当性（αエラーで示される）や検出力（１-β）は保証されない。データが歪んだ（非対称的な）分布を示す場合や比較する群どうしでバラツキが大きく異なる場合は、検定の前に何らかの変数変換（transformation）を行って前提が満たされるようにしておいてからパラメトリックな手法を用いるか、替わりとなるノンパラメトリックな手法を用いることが推奨される。

図2）



serum level of CEA　　log(serum level of CEA)

多くの生物学的な変数は、例えば腫瘍径や血清腫瘍マーカーレベルなど、右にすそを引く（値が大きくなるほど分布がまばらになる）分布をすることが多い。経験的にこうした右にすそを引く分布を示す変数に対してはログ変換（logarithmic transformation）を行うことによって正規分布に近い分布になることが多い（図２）。ログ変換を行って正規分布とみなせる分布になり、かつ特にそれに何らかのもっともらしい生物学的な意味づけが可能（腫瘍径など）な場合には、ログ変換の後にパラメトリックな手法を用いればよい。変数変換を行った場合は、結果のレポートでは元の単位に変換し直して表示する方が望ましい（検定のp値は変換後の検定結果をそのまま使うが、平均値や標準偏差、信頼区間は変換後の計算で得られた点推定値を逆変換で元の単位に戻して表示してよい）。

２-３-１　要約統計量と分布

分布がほぼ左右対称である場合に限り、平均値や標準偏差を（その分布を代表する）要約統計量として用いることが妥当である。前項で述べたようなどちらかに歪んだ分布で平均値を用いることは適当ではない。左右対称でない場合は、替わりに中央値（メディアン）やパーセンタイル（例：４分点＝25パーセンタイル、メディアン、75パーセンタイル）、データのバラツキのうちのある割合（90%や95%）をカバーする値の範囲を用いる方がよい。

２-３-２　外れ値

外れ値（Outlier）とはデータの主要な固まりから大きく外れたデータのことを言う。解析の前に外れ値のチェックを行うことは極めて重要である（言葉を換えれば“外れ値のチェック”は解析の重要な一部である）。なぜなら外れ値はデータ入力のミス、元々のカルテ記載のミスやケースレポートフォームの記載ミスなどデータ収集のすべての段階で生じうる人為的なミスであり得るからである。これらの外れ値のチェックを行ってデータ入力過程のミスを修正した後に残った外れ値は、データ解析のすべての段階で注意深く扱う必要がある（簡単に削除しない）。

外れ値は特にパラメトリックな手法（ｔ検定、ピアソン相関や回帰分析）を用いる際は、推定値や検定結果に大きく影響するので特に注意が必要である。単に“外れ値検定”の結果のみに基づいて簡単に外れ値を解析から除外してはならない。どう考えてもそのまま解析に含めるのに妥当でないような外れ値がある場合（先天性の酵素欠損など、非常に稀な疾患の症例がたまたま混入したような場合）でも、その外れ値を含めた解析と含めない解析の両方の計算結果をレポートすることが原則である。

目次へ戻る

—

２-４　カテゴリー変数

カテゴリー変数には名義変数（nominal data）と順序変数（ordered）がある。前者は性別や血液型など、値に順序がないものであり、後者は病期（stage I, II, III, IV）や腫瘍縮小効果（CR, PR, NC, PD）など値に順序があるものである。順序変数において平均や標準偏差を求めることは妥当ではなく、それぞれの割合（%でよい）が示されるべきである。

２-４-１　分割表分析

２つのカテゴリー変数どうしの関係は分割表（contingency table）と呼ばれるクロス集計表の形で調べることができる。２変数の関連（association）を調べるのに、よくχ２乗検定（chi-squared test）が使われるが、χ２乗検定は概して臨床的に意味のある仮説を検定するには感度（検出力）が充分高いとは言えないオムニバスな検定手法であり、検定する仮説によっては（見たいことによっては）必ずしも適切でないこ

とがある。

例えば、胃癌の例で考えると、組織学的なサブタイプ（２値または３値以上の名義変数）の違いによって肝転移の割合が異なるかどうか（ありなしの２値の名義変数）という単純な検定を行うなら単純な（通常の）χ２乗検定でよいが、一方、壁深達度（m, sm, mp, ss, s）の順にリンパ節転移の割合が高くなる傾向があるかどうかというような、順序変数に従って変化する“ trend ”の有無を検定したければ、単純なχ２乗検定ではなく、マンテル検定（the Mantel test）のような“ trend test ”を用いる必要がある。

両方が順序変数の場合には順位相関に基づく検定（rank correlation）が推奨される。

２-４-２　背景因子の群間比較における同等性

多くの臨床研究のレポートにおいて、治療群間での背景因子やベースラインデータに偏りがないことを示すために群別の背景因子の表が示されることが多い。こうしたデータを示すことは比較したい群間の比較可能性を示す上で重要であるが、特にχ２乗検定のように検出力の低い検定手法を用いた群間比較での「有意差なし（P>0.05）」という検定結果は、必ずしも比較したい両群の背景因子に偏りがないことを意味するわけではない。特に例数の少ない研究においては、背景因子やベースラインデータの群間比較の「有意差なし」は多くの場合、群間の同等性ではなく単なる症例数の少なさを示すのみであることが多い。また、例数の多い研究においても、背景因子の小さな偏りが有効性や安全性のendpointの大きな差に結びつくこともあるので、背景因子の偏りについては検定によって一律に処理するべきではなく、それぞれの偏りがそれぞれのendpointに対してどういうバイアスとして影響し得るかを個々に検討し、解釈を加える必要がある。

しかしそうは言っても目安は必要と思われるので、一応の目安として我々の推奨するやり方を示す。

まず背景因子やベースラインデータの集計表では、“ N.S. ”や“ not significant ”という表現は使わずにχ２乗検定（やその他の適切な検定）の実際のp値を示し、少なくとも$p < 0.2$であった因子については、特にそれが研究のendpointに影響を及ぼしそうな場合はそれぞれ個々に追加のサブグループ解析やその他の統計学的な調整の対象とする、というものである。

目次へ戻る

—

２-５　相関と回帰

相関分析は２つの連続変量の間の関係を調べるのに有用な方法であるが、医学論文の中での誤用がもっとも多い手法でもある。相関を用いる場合は常に散布図を示すことが望ましく、特に相関の有無がその論文における重要なポイントとなっているような場合には、必ず散布図を示すべきである。回帰分析においては、散布図に加えて回帰曲線（直線）を共に示すことが望ましい。

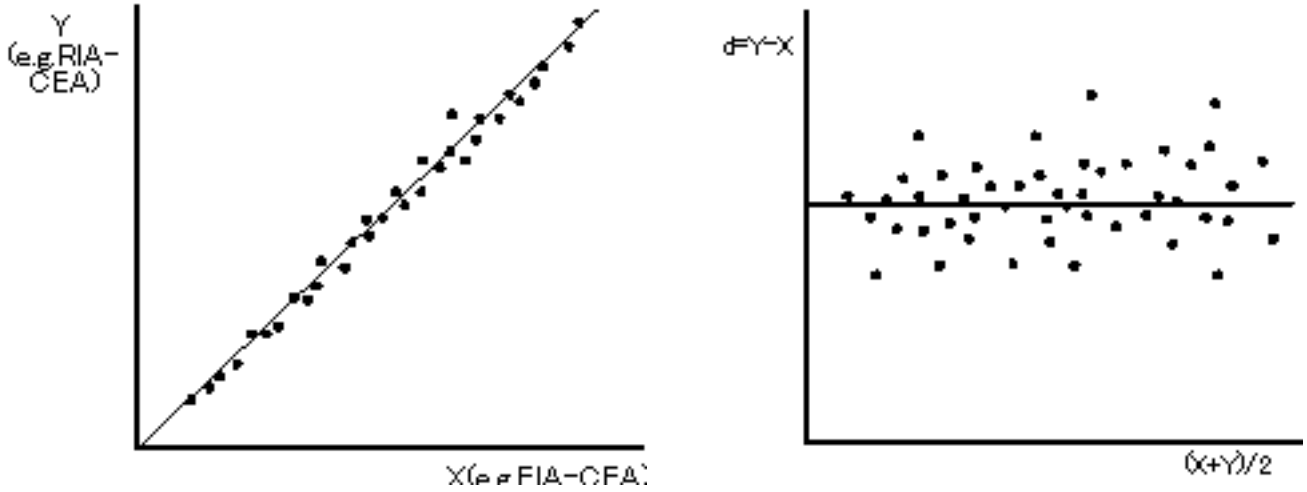
相関や回帰分析を用いる場合は以下のような点に注意が必要である。

２-５-１　診断方法の比較

同じ変数を別の測定法で測定したようなデータを検討する場合（例えば同一検体を用いて２つのアッセイ法でCEAを測定したような場合）には、相関係数や相関の有無の検定結果を示すだけでは不充分である。なぜならこの場合、同じものを測っているのだから両者が相関するのは当たり前であり、実際の関心の対象は「両者のデータがどの程度一致しているか」のはずだからである。このような場合は関連や相関について議論するべきではなく、個々の対象（検体）における両者（２つの方法によるアッセイ結果）の“

差 ” が許容範囲にあるかどうかを論じるべきである。この “ 許容範囲 ” は統計学的な観点から決められるものではなく、臨床的（医学的）観点からのみ判断される。推奨される方法は、２つの変数の散布図（図3a左）に加えて検体ごとの両者の値の差をプロットした散布図（図3a右）を示すことである。較正式（calibration formula）作成に当たっては、用いられるべき適切な手法は回帰分析ではなく関数関係解析（functional relationship analysis）または構造関係解析（structural relationship analysis）と呼ばれる方法である。回帰分析では片方の変数での測定誤差が考慮されないのに対し、これらの手法は双方の変数の測定誤差を考慮する。詳細は本稿の範疇を越えるので省略する。

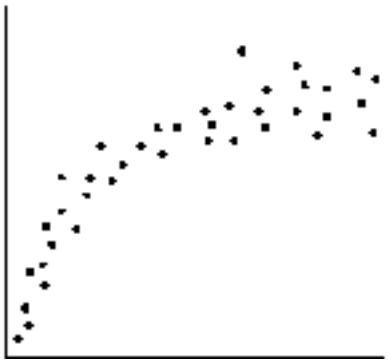
図3a



２-５-２　線形でない相関関係

例えばy = log xのように線形（直線的）でないが一様な関係が見られる（図3b）ような場合には通常用いられるピアソン（確率）相関係数は低く見積もられる傾向にあるため、機械的にピアソン相関係数を調べるだけでは重要な関係が見過ごされてしまうことがある。散布図はこうした非線形の相関関係を見つける上でも有用である。

図3b



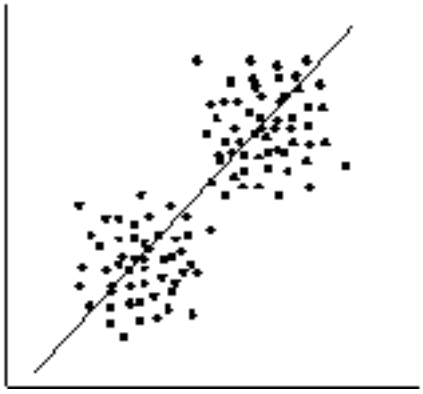
２-５-３　外れ値がある時

外れ値は相関係数や相関の有無の検定に大きな影響を及ぼす（外れ値の有無によって結果が大きく変わりうる）ため、外れ値がデータに含まれる場合は相関解析を行う前に充分に注意する必要があり、また外れ値を除く際にも慎重でなければならない。

２-５-４　複数のサブグループがある時

図3cで示すように、一つまたは両方の変数の平均値が異なるような２つ以上のサブグループ（固まり）が存在するような場合は、全体として相関係数を求めたり検定を行ったりすることは不適切である。（このような場合は、むしろ今まで知られていなかったような新しいサブグループの発見に結びつく可能性もあるので、その意味でも生データの散布図を描いてみることは極めて重要である。）

図3c



２-５-５　前提となる仮説

ピアソン相関の有意性検定における前提は、両者の変数がランダムに正規分布に従って分布しているが、一方の値を固定したとき、他方が正規分布することであり、不規則な分布をしている場合は順位相関（Spearmanのρ、Kendallのτ）を用いることが薦められる。

２-５-６　“予測”における有用性

相関や回帰分析の結果、２つの変数の間に単に統計学的に有意な関連が示されたからと言って、片方の変数がもう一方の変数の予測に“有用”であるとは限らない。予測における有用性を示すためには、回帰方程式による予測値がどの程度実測値と一致しているか（precision）を示すか、感度／特異度（sensitivity/specificity）を適切な方法で示す必要がある。

目次へ戻る

—

２-６　生存時間解析

２-６-１　平均より中央値

生存時間データは通常“打ち切り（censored）例”を含むことと、分布が正規分布でなく右にすそを引くことから、“平均生存時間”を求めることは適当ではない。生存時間分析においては中央値（median survival time）が平均値よりよい要約統計量である。

２-６-２　生存“率”の検定はしない

特に打ち切り例を含む場合や生存曲線の形が異なる場合には、例えば５年生存率をχ２乗検定で比較するように、ある時点での生存割合（いわゆる“生存率”）を直接比較することは適切とは言えない。通常その替わりにログランク検定や一般化Wilcoxon検定などのように生存時間分布全体を扱う統計手法が用いられることになる。

２-６-３　起算日とイベント

生存時間解析において起算日とイベントの定義は常に明確に示される必要がある。例えば「無再発生存時間は無作為割り付け日から次のいずれかのイベントが観察された日までの時間とした。[1] 生検で確認された術後最初の再発所見、[2] 二次がん、[3] 再発や二次がんの所見がない場合のあらゆる原因の死亡」などのように。

２-６-４　奏効者（responder）と非奏効者（non-responder）との比較

化学療法や放射線療法などのある治療に対して奏効した症例（responder）と奏効しなかった症例（non-responder）の生存時間を比較して前者が勝ったとしてもなんら結論的な結果を示したことにはならない。なぜなら前者は“奏効（response）”が観察されるまで生存していることが必要なので、治療による延命効果のあるなしに関係なく前者の生存時間は長くなる傾向にあるというバイアスを見ているに過ぎないことを否定できないからである（５）。本誌ではこうした誤用はacceptされないので注意されたい。

２-６-５　予後因子

予後因子（prognostic factors）とは他の因子による調整を行った上でなおかつ何らかの予後に対する影響を有していることが示されたものを呼ぶ。従って単変量解析においてのみ予後との関連が見られた因子は必ずしも“予後因子”とは呼ばない。予後因子と言えるかどうかを論じるには、多変量解析やその他の適切な層別解析またはサブグループ解析による検討が必要である。ただし、多変量解析の手法は、先述した検定の多重性から“甘い”（liberalな、false positiveの多い）推定を与える傾向があるので、特に探索的な解析において多変量の手法を用いる場合には結果の解釈を控えめ（conservative）にする必要がある。単にp値が0.05を下回ったからと言って機械的に“有意な予後因子”と決めつけないことである。多変量解析において有意水準を0.05とする根拠は実は乏しい。

目次へ戻る

—

２-７　多変量的手法

最近はコンピューターのハードウェア、ソフトウェアの発達に伴って、多変量解析の手法をサポートした強力なパッケージソフトが簡単に使えるようになってきた。多変量解析の手法は、複数の変数を同時に扱ったり多数の交絡因子（共変量）の調整を行う上で有用な手法である。しかし多くの多変量解析手法はパラメトリックまたはセミパラメトリックな手法であるために分布についての前提や、Cox回帰の際の比例ハザード性などの制限事項を有するが、多くの医学論文においてはこれらが無視され、また結果の解釈が誤っている場合も多い。多変量解析の手法は現実問題として専門家の手助けを必要とする専門的な手法であり、使う前に生物統計家に相談することを強く推奨する。多変量解析手法の用い方は本ガイドラインの範囲を越えるので詳述しない。

目次へ戻る

—

２-８　最後に

統計学において最も大事なことは解析結果をいかに示すかではなく、研究をいかにデザインするかである。よくデザインされたが誤って解析された研究は再解析により救済することが可能であるが、誤ってデザインされた研究はいかに洗練された手法をもって解析しても救うことはできない（６）。生物統計家に相談する場合は、データが得られてしまった後ではなく、必ず研究の初めに相談するべきである。

筆者らは本ガイドラインが、読者が統計学の勉強を始める手助けに、また投稿者と編集者が共通の認識のもとに議論できる手助けになることを希望している。
目次へ戻る

—

３　推奨される教科書およびより詳細なガイドライン

（　）内は筆者の感想

—

３-１　英文

1) Campbell, M. J. and D. Machin (1993). Medical Statistics. A Commonsense Approach. 2nd Edition, Chichester, John Wiley & Sons.

（臨床医向け。volumeも手頃で通読できる。）

2) Armitage, P. and G. Berry (1994): Statistical Methods in Medical Research. 3rd Edition, Oxford, Blackwell Scientific Publications.

（最も有名な標準的な教科書。引用には最適か？）

3) Ingelfinger, J. A., F. Mosteller, et al. (1994). Biostatistics in Clinical Medicine. 3rd Edition, New York, McGraw-Hill, Inc.

（臨床研究の事例が豊富）

4) Essex-Sorlie, D. (1995). Examination & Board Review : Medical Biostatistics & Epidemiology. London. Prentice-Hall International Inc.

（非常に基本的な事項から書かれている）

5) Gonick, L. and W. Smith (1993). The Cartoon Guide to Statistics. New York, Harper Perennial.

（基本的な事項がマンガを使って解説してある）

6) Thall, P. F. (1995). Recent Advances in Clinical Trial Design and Analysis. Norwell, Kluwer Academic Publishers.

（臨床試験における新しい方法論の解説）

～以下はJournalに掲載されているガイドラインである。本稿でも大いに参考とした～

7) Altman, D. G., Gore, S. M., and Gardner, M. J. Statistical guidelines for contributors to medial journals. BMJ. 286: 1489-1493, 1983.

8) Bailar III, J. C. and Mosteller, F. Guidlines for Statistical Reporting in Articles for Medical Journals. Amplifications and Explanations. Annals of Internal Medicine. 108: 266-273, 1988.

9) Murray, G.D. Statistical guidelines for The British Journal of Surgery. Br.J.Surg. 78: 782-784, 1991.

10) Lewis, J. A., Jones, D. R., and Rohmel, J. Biostatistical Methodology in Clinical Trials - European Guideline. Statistics in Medicine. 14: 1655-1657, 1995.

—

3-2　和文

1) 折笠秀樹.　「臨床研究デザイン　医学研究における統計入門」真興交易医書出版部1994.

（デザインの解説書なのでほとんど数式も出てこない。本稿でも大いに参考とした）

2) 久道茂.　「医学判断学入門　―われわれの判断や解釈はまちがっていないかー」南江堂1990.

（統計解析以前の医学研究の前提となる基本的なものの見方・考え方の解説書）

3) 駒澤勉、高木廣文、佐藤俊哉.　「別冊・医学のあゆみ　ヘルスサイエンスのための統計科学」医歯薬出版株式会社1996.

（読み物として読めるよう平易に書かれている）

4) 大橋靖雄、浜田知久馬.　「生存時間解析　SASによる生物統計」東京大学出版会1995.

（SASを使って生存時間解析を行う際のマニュアルであるが、教科書としても使える）

5) ラリ・ゴニック、ウールコット・スミス.　「マンガ確率・統計が驚異的によくわかる」白揚社.　1995.

（英文の5)の日本語訳）

6) 芝祐順、渡部洋、石塚智一編.　「統計用語辞典」新曜社1984.

（辞典。言うまでもなくわからない用語に出会った場合に引く。）

目次へ戻る

—

4　引用文献

1. Altman, D. G., Gore, S. M., and Gardner, M. J. Statistical guidelines for contributors to medial

journals. BMJ. 286: 1489-1493, 1983.

2. Bailar III, J. C. and Mosteller, F. Guidlines for Statistical Reporting in Articles for Medical Journals. Amplifications and Explanations. Annals of Internal Medicine. 108: 266-273, 1988.

3. Murray, G.D. Statistical guidelines for The British Journal of Surgery. Br.J.Surg. 78: 782-784, 1991.

4. Campbell, M. J. and Machin, D. Designed Observational Studies. In: Medical Statistics. A Commonsense Apploach, 2nd edition, pp. 125. Chichester: John Wiley & Sons, 1993.

5. Simon, R. Design and Conduct of Clinical Trials. In: V. T. DeVita, S. Hellman, and S. A. Rosenerg (eds.), Cancer: Principles & Practice of Oncology, 4th edition, pp. 420. Philadelphia: J.B.Looppincott Co.,, 1993.

6. Campbell, M. J. and Machin, D. Uses and abuses of medical statistics. In: Medical Statistics. A Commonsense Apploach, 2nd edition, pp. 2. Chichester: John Wiley & Sons, 1993.

目次へ戻る