企画特別講演（第144回東邦医学会例会）

総　説

医療統計，最初の一歩

村上　義孝

東邦大学医学部社会医学講座医療統計学分野教授

要約：医療統計学の初学者を対象に医学研究で用いられる主な統計手法を説明し，統計的仮説検定とその限界，区間推定（95% 信頼区間）とそのメリット（効用）を解説した．はじめにエンドポイントの変数型（二値/カテゴリカル/連続量/打ち切りのある時間）と対応する検定手法（カイ二乗検定/t 検定など/ログランク検定）についてまとめ，研究成果の図表イメージと検定手法が対応していることを説明した．次に統計的仮説検定についてその論理的枠組を概観し，検定が意志決定の道具であること，臨床研究の大部分を占める医療技術の定量的評価には区間推定の方が適していることを説明した．最後に医学データ解析で現在頻繁に用いられる区間推定（95% 信頼区間）の考え方について簡単に解説した．

東邦医会誌 61(5)：238–243，2014

索引用語：生物統計学，医療統計学，統計的仮説検定，95% 信頼区間

人間集団を対象とした医学研究はデータに基づき検討することが多いため，研究成果をまとめる際に統計学の知識や考え方が必要となる．医学分野への応用を目的とした統計学は生物統計学（biostatistics），医療統計学（medical statistics）とよばれ，いまや統計学の中で確固たる地位を占めている．本稿では医療統計学の初学者を対象に，医学分野で用いられる統計手法，考え方，注意すべき点についてまとめたものである．はじめに医学研究で用いられる主な統計手法について概観し，つぎに医学データ解析で頻出する統計的仮説検定とその限界について説明し，最後に区間推定（95% 信頼区間）の考え方とそのメリット（効用）の説明をする．

医学研究で用いられる主な統計手法

医学研究においてエンドポイントが決定されると同時に，使用すべき統計手法はおのずから決まっていく．Fig. 1に示したのはよくある統計相談の1例であり，「A，Bの2群間で60と50という値を比較したいが，何検定を使用したらよいか？」という質問である．これに対しては比較したいエンドポイントが，体重など平均（もしくは中央値）でまとめられる（要約できる）ものであれば t 検定（t test）などを，病気の発生など割合（パーセント）で要約できるものであればカイ二乗検定（chi-square test）を，生存時間など打ち切りをともなった時間でまとめられる場合はログランク検定（log-rank test）が通例である．このことからわかるように要因と結果との関連の検討ではエンドポイントを意識して，解析手法を選択することが重要となる[1–3]．**これらエンドポイントの決定・確認は通常は研究計画もしくは解析段階で明らかになるが，いずれにせよ研究目的に照らし合わせてエンドポイントを決定するのは研究者の責任である．**

Table 1にエンドポイントの変数型と使用すべき検定手法についてまとめた．変数の型としては以下の4つを示した．①二値変数（例：発生あり/なし，死亡あり/なし），②カテゴリ変数［例：総合評価（無効/有効/著効），満足度（不満/ふつう/満足）］，③連続量（例：収縮期血圧，体重），④打ち切りをともなう生存時間（例：発生・死亡までの期間，入院期間）．統計解析ではデータの記述（図表）とデータの分析（検定を含む）の2つが重要である．①の二値変数では2×2分割表（2 by 2 table）で記述し，カイ

〒143-8540　東京都大田区大森西5-21-16
受付：2014年9月1日

東邦医学会雑誌　第61巻第5号，2014年9月1日
ISSN 0040-8670，CODEN: TOIZAG

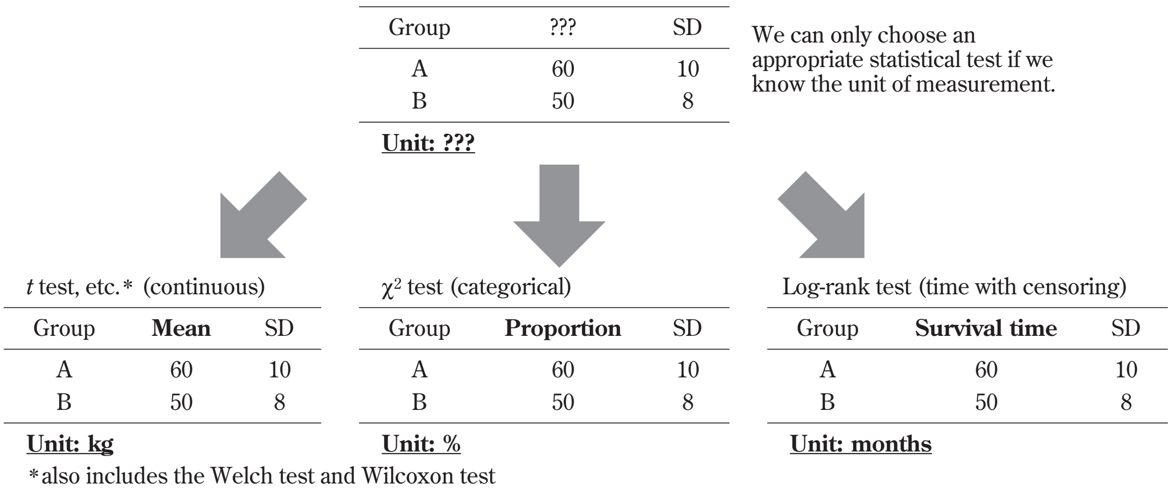

Fig. 1　Choice of statistical tests in various research contexts.

Table 1　Classification of statistical tests

Purpose	Types of outcome variables			
	Dichotomous (Yes/No)	Categorical	Continuous	Survival time
Description	2×2 Table	2×c Table	Box-and-whisker plot	Kaplan-Meier curve
Statistical test	χ^2 test	χ^2 test	*t* test, Wilcoxon	Log-rank test
Statistical modelling	Logistic regression	Logistic regression	Multiple regression*	Proportional hazards model (Cox regression)

*includes multiple regression, analysis of variance, and analysis of covariance

二乗検定で検定する．②のカテゴリカル変数では 2×2 分割表の拡張である 2×c 分割表（2 by c table）でまとめ，カイ二乗検定で検定する．③の連続量では散布図や箱ひげ図（box and whisker plots）でまとめ，*t* 検定などで検定する．④の打ち切りをともなった生存時間ではカプランマイヤー曲線でまとめ，ログランク検定を用い検定する．なお要因と結果との関連を歪める第 3 の因子のことを交絡因子とよぶが，この交絡の影響を除外する方法の 1 つとして医学分野では統計モデリングが用いられる．使用する統計モデルは上記の変数の型に対応しており，二値変数，カテゴリカル変数ではロジスティック回帰（logistic regression）を，連続量の場合は重回帰（multiple regression），分散分析（analysis of variance），共分散分析（analysis of covariance）のいずれかを，生存時間の場合は比例ハザードモデル（コックス回帰）（proportional hazards model, Cox regression）を使用する．研究を行う初歩の段階では研究目的に合致した統計手法の選択に悩む場合がある．**その場合は基本に立ち返り，研究成果の発表イメージがどのようになるか，図表としてどうまとまるか考えてみるとよい**[1]**.**

Table 1 をみると連続量をエンドポイントとした場合の検定手法の選択肢として，*t* 検定，ウェルチ検定（Welch test），ウィルコクソン検定（Wilcoxon test）が併記されている．これは検定する際の前提条件の違いであり，*t* 検定の場合，2 つのグループの分布が左右対称で 2 群の分散が等しい必要がある．左右対称であっても分散が等しくない場合はウェルチ検定を使用する．また分布が左右対称でない場合ウィルコクソン検定を使用する．分布が左右対称か否かの判断については，通常ヒストグラムを作成し解析者が判断することが多い．分散が等しい（等分散）かは等分散性の検定（F 検定）により吟味でき，多くの統計ソフトウェアでは *t* 検定の結果とともに出力される．なお等分散性の検定は，*t* 検定を実施する前の予備的検定なので有意水準 15〜20% を使用するのが一般的である．

医学分野におけるデータ解析のもつ意味

医学データ解析に対する大きな誤解として，「統計的仮説検定は必須である」,「統計的な有意差がないと研究論文として成立しない」というものがある．またそれに関連して，「統計的に有意な結果＝よい結果，有意でない結果＝悪い結果」という先入観がある．そもそも，いい結果/悪い結

果という表現は「自分にとって都合のよい」というニュアンスを含み，科学的客観性から乖離しており恣意的でもある．医学分野で必要なことは差あり/なしという判断ではなく，どのくらい差があったかという定量的評価であり，このことは次に述べる統計的仮説検定の限界と区間推定の有効性の議論につながっていく．

統計的仮説検定とは？

統計的仮説検定は医学研究で頻回に利用される一方で，その論理的枠組の難しさや数理統計的な側面から十分な理解がされていない方法の１つでもある．これは統計的仮説検定を理解する際に，①確率変数，確率分布の理解，②確率分布を使用した計算法，③仮説検定における論理（背理法）の利用，など要素が複雑に絡み合っているためだと思われる．ただ現在では，②の確率分布を使用した計算（p値の計算）はコンピュータから自動的に出力され，①確率変数，確率分布の理解についても本質的には生物統計学の教科書の範疇であり，実践的には前述した統計手法の使い分けがわかれば十分である．ただ統計の実践でも避けて通れない部分として③の検定の論理があるので説明する．

背理法とは「ある事柄Ｈを証明するために，Ｈの否定である H_0 を仮定したときに矛盾が起きることを利用した証明手法」である．統計的仮説検定では「H_0：群間に差がない」と仮定し，その仮定のもとに「実際のデータもしくはそれ以上の差が生じる」確率を計算し，得られた確率が極めて小さい場合には，H_0 を否定し「H_1：群間に差がある」を採択する，という一連の論理判断をとる．なお H_0 を帰無仮説，H_1 を対立仮説，H_0 を否定することを棄却，逆に計算した確率が小さくなく H_0 を否定しきれない場合を受容，保留などと呼ぶ．背理法は論理的な矛盾に基づいて事柄を証明しようとするが，仮説検定では論理的矛盾でなく解析者の判断（確率が極めて小さいことから H_0 を否定し）に基づいて事柄を証明する．統計的仮説検定において解析者の判断が入ることは，同時に判断が時として誤っている可能性を意味している．Table 2にデータから検証できない真実とｐ値による判断の組み合わせと，判断の正誤を示したマトリックスを示す．統計的仮説検定においては，真実では差がなくても「たまたま」今回得られたサンプルで差があった場合，仮説検定では差ありと誤って判断する（Table 2のαエラー）．通常，検定の有意水準を5%とするのは，この誤りを最小にしようという意図からである．逆に真実は差があっても，たまたま今回得られたサンプルで差がなかった場合，仮説検定では差なしと誤って判断する（Table 2のβエラー）．**統計的仮説検定では有限（にすぎない）集団から，真実に何が起こっているか意志決定（判断）するプロセスである以上，このαエラーとβエラーから逃れることはできない．**

Table 2　Truth, human decisions, and α and β errors
Hypothesis testing is a method of making decisions when data are limited. This decision process (with limited data) sometimes (or often) leads to incorrect decisions (α error and β error) ! This is inevitable.

	Decision	
Truth	Different	Not Different
Different	Correct (power)	Wrong (β error)
Not Different	Wrong (α error)	Correct

医学研究における統計的仮説検定の限界

統計的仮説検定で注意すべき点として，検定結果はサンプルサイズによって変化し，同様の傾向を示すデータであってもサンプルサイズの多少により判断が変化することがある．Fig. 2に例を示す．Study 1ではA，B群おのおの300人の集団の疾病発生割合の比較を行っている．その結果，A群では2%（6/300），B群では0.66%（1/300）が観察された．これをカイ二乗検定するとｐ値は5.7%と有意水準5%では有意でなく，解釈として「A群とB群の疾病発生割合は異なるとはいえない」となる．このStudy 1

Study 1

Group	Events	No events	Total
A	6	294	300
B	1	299	300

p value = 0.057
Risk ratio (95% CI): 6.0 (0.72-49.5)

Study 2

Group	Events	No events	Total
A	60	2940	3000
B	10	2990	3000

p value <0.001
Risk ratio (95% CI): 6.0 (3.1-11.6)

Fig. 2　Hypothetical example of two statistical tests, in studies with identical risk ratios and different sample sizes
The sample size in Study 1 is ten times larger than that of Study 2.
CI: confidence interval

を10倍の規模で実施した仮想例がStudy 2である．これをみるとA群では2%（60/3000），B群では0.66%（10/3000）とStudy 1と同様の結果が観察された．しかしながらこれをカイ二乗検定するとp値は0.1%より小さくなった．この解釈としては「A群の方がB群に比べて多く疾病が発生していた」となる．Study 1では「差があるとはいえない」，Study 2では「差があるといえる」という正反対な結論となったのは，検定結果がサンプルサイズに依存するためである．統計において推測精度はサンプルサイズに依存し，数が少ない場合は精度が低く明確なことはいえず，逆にある程度以上になると，明確な結論が得られるというのが，その根底にある．ただ統計的仮説検定では「差あり」「差なし」の2つの判断しかないため，**Study 1の例では，サンプルサイズが少なかったから明確な結論「差がある」がいえなかったのか，それとも本質的に「差がない」のかの判別ができない．**これは統計的仮説検定を用いた判断の限界であり，治療法や危険因子の定量的評価を目的とした医学研究においては，次に述べる区間推定（95%信頼区間）という統計手法の方が，研究目的に即した方法といえる．なお医学研究で「検定を用いるべきか？推定を用いるべきか？」という議論は1990年代に主にアメリカを中心になされたが，治療薬や技術の評価（医療技術評価）を目的とした比較研究では95%信頼区間を示すことで合意がなされている[4]．

区間推定（95%信頼区間）とその効用

区間推定について簡単に説明する．データの平均値をもって代表値とするように，1点（1つの値）で推定を実施することを点推定といい，データの存在する範囲を推定することを区間推定，構成される区間を信頼区間と言う．仮説検定を対応させて表現すると，信頼区間とは「データに矛盾せずに仮説の存在しうる範囲」のことであり，推定された値の精度を表すものである．仮説検定においては帰無仮説・有意水準を固定し帰無仮説の棄却/受容を議論したが，区間推定においては有意水準を固定したもとで帰無仮説を変化させ，帰無仮説が棄却されない範囲を探索するという方法をとる．詳しくは文献2）もしくは文献4）を参考にしてほしい．なおサンプルサイズが十分な場合は正規近似により信頼区間は計算され，この場合95%信頼区間はpの標準誤差をSE（standard error）と表すと，$(p-1.96\times SE,\ p+1.96\times SE)$となる．

Fig. 2をみると，Study 1ではリスク比は6.0，95%信頼区間が0.72～49.5となっている．これは真のリスク比が「小さくてせいぜい0.72くらいだが，大きい場合は49.5をとることも考えられる」ということを示している．またStudy 2ではリスク比は6.0，95%信頼区間が3.1～11.6と幅がせまくなっていることがわかる．**解釈としてはStudy 1ではリスク比が1，つまりA群とB群の疾患発生割合が「等しい」ことを否定できないが，Study 2ではリスク比1の可能性は「ほぼない」と言ってよく，**A群の疾患発生割合がB群のそれより大きいことが示されている．区間推定ではFig. 2の例のようにリスク比（点推定値）の値も同時に記載されるので，治療法の効果など直接的に解釈できる数値が，その精度（95%信頼区間）とともに表示される．この統計的仮説検定にはない定量的評価が可能となることは，医学研究において重要であり，今では海外の一流学術雑誌においては95%信頼区間を用いた結果提示が要求されている．

おわりに

本稿では，医学研究で用いられる統計手法の使い分けから説明し，統計的仮説検定とその限界，区間推定（95%信頼区間）の考え方を説明した．研究で使用されるエンドポイントから統計手法は定まっていくように，研究目的（リサーチクエスチョン）と統計手法は密接な関連にある．今回は紹介できなかったが，研究目的を立案するにあたってサンプルサイズ計算を始めとした統計の役割は大きい[5,6]．医療統計的な視点を入れた，医学の真の発展に寄与する研究計画，プロトコル作成が今求められている．

文　献

1）松山　裕：統計学再入門（第1回）データの要約：要約統計量と分布の視覚化．心身医　**53**：436-441, 2013
2）松山　裕：統計学再入門（第2回）統計的仮説検定と効果の推定．心身医　**53**：687-693, 2013
3）松山　裕：統計学再入門（第3回）研究デザインと統計解析手法の選択．心身医　**53**：764-770, 2013
4）Rothman KJ；矢野栄二，橋本英樹，大脇和浩（監訳）：ロスマンの疫学–科学的思考への誘い（2版），篠原出版新社，東京，2013
5）Altman DG；木船義久，佐久間昭（訳）：医学研究における実用統計学，サイエンティスト社，東京，1999
6）Green S, Benedetti J, Smith A, et al；福田治彦（訳者代表）：米国SWOGに学ぶがん臨床試験の実践（2版），医学書院，東京，2013

An Introduction to Medical Statistics

Yoshitaka Murakami

Professor, Division of Medical Statistics, Department of Social Medicine,
School of Medicine, Faculty of Medicine, Toho University

ABSTRACT: This article attempts to explain the most important statistical methods for medical researchers. I explain statistical hypothesis testing and interval estimation, including the advantages of 95% confidence intervals in statistical testing. First, I explain the types of outcome variables (dichotomous, categorical, continuous, survival time with censoring) and the corresponding statistical tests (χ^2 test, t test, log-rank test). I note that the form in which data are presented (*e.g.*, tables and graphs) is directly related to the statistical method used. Second, I summarize the logic of statistical hypothetical testing and conclude that statistical testing is a decision-making tool. I also discuss why it is better to use interval estimation for quantitative assessment in medical technology, which is a concern in most clinical research. Finally, I briefly explain the logic of interval estimation, which is routinely used in medical data analysis.

J Med Soc Toho 61 (5): 238–243, 2014

KEYWORDS: biostatistics, medical statistics, statistical hypothesis test, 95% confidence interval

5-21-16 Omorinishi, Ota, Tokyo 143-8540

Journal of the Medical Society of Toho University
61 (5), Sept. 1, 2014.　ISSN 0040-8670,　CODEN: TOIZAG

村上義孝教授　ご略歴

1993年 3月	東京大学医学部保健学科卒業
4月	東京大学大学院医学系研究科保健学専攻修士課程入学
1995年 3月	同　修了
	保健学修士（東京大学）取得
4月	東京大学大学院医学系研究科健康科学・看護学専攻博士課程入学
1998年 3月	同　修了
	博士（保健学）(東京大学)(博医第 1348 号)
4月	大分県立看護科学大学看護学部助手（人間科学講座・健康情報科学研究室）
2002年 5月	国立環境研究所　研究員（環境健康研究領域・疫学・国際保健研究室）
2005年 7月	滋賀医科大学　特任講師(社会医学講座・福祉保健医学部門)
2008年12月	同　准教授（社会医学講座・医療統計学部門）
2009年 2月	The George Institute for Global Health, visiting post-doctoral fellow
2012年 4月	滋賀医科大学附属病院臨床研究開発センター　副センター長（併任）
2013年 5月	滋賀医科大学アジア疫学研究センター　副センター長(併任)
2014年 4月	東邦大学医学部社会医学講座医療統計学分野　教授

受賞など

2006年10月	29th Japanese Society of Hypertension (JSH) Awards (Young Investigators Travel Awards)
2008年10月	第 31 回日本高血圧学会 Young Investigators Awards（YIA）最優秀賞
2009年10月	第 68 回日本公衆衛生学会学術総会優秀演題賞
2012年 1月	第 22 回日本疫学会学術総会ポスター賞
2012年10月	日本公衆衛生学会ベストレビュワー賞
2013年 1月	第 23 回日本疫学会奨励賞

学会の役職など

日本疫学会評議員，日本公衆衛生学会評議員，日本計量生物学会評議員