統計数理 (1994)
第42巻 第2号 215-224

統計的推測における手法の妥当性

統計数理研究所 柳 本 武 美

(1994年1月 受付)

1. 序

統計学特に近代の統計学を端的に述べれば，帰納的推論のための計量的方法，といえよう．帰納的推論は演繹的推論とは異なって，論理に飛躍を伴う．しかしながら，実験あるいは現象の観察を通して新しい知識を得ようとすれば，帰納的推論でしかあり得ない．

K. Pearson, Fisher, Gosset らを中心に開発された統計的検定は，現在の統計的方法の核心を成している．統計的検定は一面では数理的側面が注目され，多くの研究を促し，今日でも極めて活発に研究されている．しかし統計的検定の遙かに強いインパクトは，実証的研究に新しい視点を与えたことにある．今日治療効果，薬効評価，農事試験，計量経済学をはじめとする極めて広範な分野に深く根づいているのは，研究者に精密なデータに基づいた仮説の提唱，検証への標準的な規準を与えた所にある．尤もらしい論理の展開よりも，実証的研究を優先させるインセンティブを与えた．

統計学を帰納的な推論の計量的方法ととらえれば，大局的には最善の方法が存在するとは考え難い．ある目的，例えば新しく開発した薬物の評価，のための最適なデータのとり方とか，得られたデータの最善の解析方法は想像の埒外である．最善とか最適の語はより限られた範囲で用いられ得る．

本稿の目的は統計的検定，あるいはそれに関連した，区間推定，実質的安全量の推定等の手法が帰納的推論のための一つの妥当な推論形式を与えるが，その最適性を論じることは殆ど不可能であることを強調することにある．統計的検定の論理が，理論統計学の周辺の研究者にもなじみ難い点があるとすれば，その論理が一見曖昧な点にある．この曖昧さが数理的研究にとっては理論的展開の足枷となる．通常の数理科学のように評価関数を最大あるいは最小にする論理の方が簡明である．問題はそのような評価関数が確かに最終的な目的に適うか否かにある．

与えられた条件の下で最善を尽くすことは，数理科学にとどまらず，より普遍的な規準である．従って数学的な素養を身につければ，この規準は理解し易い．数学的定式化が容易であれば，理論的展開が望まれる．更に最善を尽くすという姿勢が前向きなイメージを与える．企業の組織でいえば，統計的検定は監査部，検査部のイメージを与えるのに対して，最適性は企画部，開発部のイメージを与える．しかし実証的研究における真実の探求は，地道に忍耐強く理詰めで論証する必要がある．一般に最善を尽くすことは必要である．しかし同時に主張する言明が支持するにたる証拠を備えているかを評価することがむしろより重要である．

本稿で論じる内容は，もとより通常の学術論文のような意味での絶対的な正しさを主張するものではない．また厳密な意味での独創性を主張する訳でもない．しかし本題について改めて論じた稿は見当たらない．本稿は著者による報文 (柳本 (1981, 1991)) の延長上にある．

2. 統計的決定理論

統計的決定理論はWald (1950) によって提唱された．その骨子は以下のようである．標本 $\boldsymbol{x}$ がある分布 $p(\boldsymbol{x};\ \theta),\ \theta\in\Theta$ をもつ母集団から得られたとする．この標本に依存して，行動 $a(\boldsymbol{x})$ を選択する方法を評価する．真の母数が θ のときに，行動 a を選択した場合の損失関数を $L(\theta,\ a)$ とする．このとき危険関数 $r(\theta,\ a)$ を損失関数の真の分布の下での期待値，$E\{L(\theta,\ a(\boldsymbol{x}))\mid p(\boldsymbol{x};\ \theta)\}$ で定義する．そして良い行動の選択方式とは危険関数 $r(\theta,\ a)$ を小さくする方式であるとみなす．

統計的推定問題は上記の決定問題にごく自然に含まれる．実際，行動 $a(\boldsymbol{x})$ として Θ の一点を選択する．損失関数としては θ と $a(\boldsymbol{x})$ の距離をとれば良い．決定理論では行動 $a(\boldsymbol{x})$ の空間は Θ に限らないからより一般的である．行動空間と損失関数を柔軟に選ぶことによって，広い範囲の問題を扱うことができる．しかし決定理論では，正しい θ に関した推測を企図していない．

一方統計的検定を正規母数集団を仮定した下での2標本の平均値の検定を例にして，その骨子を記述すると次のようになる．標本 $\boldsymbol{x}=(x_1,\ldots,x_n),\ \boldsymbol{y}=(y_1,\ldots,y_m)$ が各々正規母集団 $N(\mu_1,\ \sigma^2),\ N(\mu_2,\ \sigma^2)$ からの大きさ $n,\ m$ の標本ベクトルとする．帰無仮説を

$$H_0:\ \mu_1=\mu_2$$

対立仮説を

$$H_1:\ \mu_1\neq\mu_2$$

とする．水準 α に対して棄却域 R を $P(R\mid H_0)=\alpha$ として $P(R\mid H_1)$ が大きくなるように選ぶ．もし $(\boldsymbol{x},\ \boldsymbol{y})\in R$ であれば，2つの平均は統計的に有意に異なっていて，帰無仮説は棄却される．逆に $(\boldsymbol{x},\ \boldsymbol{y})\notin R$ であれば帰無仮説は棄却されない．通常この問題では棄却域 R は

$$R=\{(\boldsymbol{x},\ \boldsymbol{y});\ nm(\bar{x}-\bar{y})^2/(n+m)s^2>F_{1;n+m-2}(\alpha)\}$$

と選ばれる．ただし $\bar{x},\ \bar{y}$ は標本平均，$s^2=\{\sum(x_i-\bar{x})^2+\sum(y_i-\bar{y})^2\}/(n+m-2)$ である．

Waldが強調したように，統計的検定は形式的には決定理論の枠組みに入れることができる．実際，行動関数 a は標本が棄却域に入るとき1，そうでないとき0とする．真の分布が帰無仮説に含まれているとき $a=0$ ならば損失関数は0とし，$a=1$ ならば1とする．逆に帰無仮説に含まれていないときには，$a=1$ ならば損失関数は0とし，$a=0$ ならば1とする．そして帰無仮説が正しいとき，危険関数が α となる決定方式を選択する．

確かに決定理論の枠組みには入るが，単に形式的でしかない．実際，帰無仮説が正しいとき危険関数を α にするという，特別な制約条件を付与する．そして決定理論からみた場合，良い統計的検定とはこの制約条件の下で検出力が高い検定とみなす．何故にこのような制約条件を置くかが不明である．このように決定理論からみた統計的検定は大変奇妙な定式化となっている．さらにこの定式化では，帰無仮説と対立仮説の2つのモデルの選択問題と理解されていることである．決定理論では帰無仮説が棄却されなかったとき，単純に帰無仮説を選択する行動をとる．一方統計的検定では帰無仮説が棄却されなかった場合にも，必ずしも帰無仮説を採択することには繋がらない，例えばMainland (1963)．統計的検定はデータに基づいて，主として対立仮説が正しいことを支持する証拠を与えるための論理である．統計的検定では決定理論と異なり，データに基づいた推測と，その推測に依存した行動の選択を明確に区別する．統計的検定は前者を目的としている．そして最適な推測を行うことを企図する以前に，妥当な帰納的

推測を行うことを企図した推論形式である．

統計的検定では損失関数の導入を必ずしも必要としない．標本が棄却域に含まれるか否かによって，判断を区別している．しかもその判断は2つの仮説を対等とみなさない．対立仮説が正しそうだと判断することはあっても，帰無仮説が正しそうだと判断することはない．標本から確実で妥当な推論の規準を与える．損失関数を導入しないために，推論に損得あるいは善悪の規準が明瞭には現れない．

3. 推論の妥当性

統計的検定が帰納的推論において妥当な推論形式であることを述べた．妥当性は別の面では最善を尽くしていないというニュアンスがある．統計的検定への批判の1つは実用性への批判であり，このことはデータの情報を十分に利用していないという認識があるように思われる．確かに統計的検定の結論は大変禁欲的であって，もう少し明快な答が期待されるのも無理がない．

統計的検定の結論が大変に控え目なのは，帰納的な推論からくる制約による．実際，新しい品種の小麦が従来種よりも単収が多いかを実験で確かめる場合を想定する．前節で述べた2標本の検定が使える設定である．帰納的な推論では，例えば栽培試験での結果から品種の収量を比較する．試験結果という現象から，本当の収量の多寡という本質が断定的には導けないのは自明の理である．だから限られたデータから真実を知ろうと企てる以上，どのような手法を用いようとも断定的な結論が下せる訳がない．データから結論がどの程度の確証が得られるか，に満足するほかはない．水準 α は推論の確証の程度と関連する．だから統計的検定の結論が回りくどく，明確でないのは極く理に適っている．

1つの実験から断定的な結論が下せないので，実際には結論の確かさを高めるために，様々な情報を得ることが多い．古くは病原菌を特定するためのコッホの4原則が有名である．また公衆衛生の分野ではHillの9つの視点，米国衛生局諮問委員会の規準がよく知られている．多角的な検証に加えることによって，仮説は遙かに確実になる．しかしながら改めて述べる迄もなく，仮説が完全に証明されることはない．現在真理として人々に信じられている法則・学説であっても，改めて検証するとそれ程確かめられていないことが多い．実際，新しい事実の観察がその学説の修正を迫ることが多い．

帰納的な推論が論理として不完全であったとしても，実証的な研究が科学の根幹であったことは疑いがない．データに基づいて確実な知識を増やす論理の整備こそが，統計学の基礎である．

問題が重要で基本的であればある程，行動の選択を導出した最適性よりも，対象を認識するための推論の妥当性がより大切になる．基本的な知見については科学的に丹念な検証が必要と

表1．最善の判断と妥当な判断の比較．

	最善の判断	妥当な判断	裏づけの規範
科学	知見の探索	知見の検証	単純な自然観 ケチの原理
司法	社会悪の絶滅	刑法の厳格適用	基本的人権
議員の選出	有能な人を任命	選挙による選出	民意の反映
治療	疾患の徹底治療 患部とその周辺全摘	患者としての治療 患部の最小限摘出	Q.O.L.

なる．表1にも示すように，司法とか治療においても当面の最適性よりも，より大局的な妥当性の方が重要になる．当面の目的とは一面では相対立するより重要な規範があるからである．

4. 最適な解の限界

前節の準備の下に統計的決定理論における最適な解の限界を論じる．

まず決定理論ではデータから帰納的な推論を行う目的を放棄している．良い行動の選択ではあっても，それによって直接新しい知見を得るとか，既存の仮説がより確からしいと再認識されるようになる訳ではない．

決定理論の立場からすると，事実の認識は行動の選択の手段であると考えられよう．現実問題では単なる科学的知見の獲得は当面何の役にも立たない．もし本当に行動の選択が必要なときには，行動の選択を最適化する立場は現実的対応であり，一つの見識である．現実の問題では実際に利用できる情報を使って行動の選択の最適性を計らなければならないからである．しかしながら真実を認識するためには，行動の選択とは離れて，あくまでも対象を誤りなく認識する努力をすると共に，その限界を押さえておくことが重要である．

モデルと実際とのずれも十分に考慮する必要がある．数学的な定式化のためには問題の簡略化が必須である．実際確率分布 $p(\boldsymbol{x};\theta)$ を仮定したとき，この仮定が本当に成り立つかと問うと，答は常に否定的である．実用的な分布は現実を良く近似した分布ではあっても，現実と厳密に一致する分布ではない．また大きさ n の標本を仮定しても，本当に同じ分布から独立にとられたデータであることは有りえない．仮定自体が近似であることを前提にしているから，与えられた規準での最適な解であっても，実際には近似的な最適解でしかない．近年のいわゆるノンパラメットリック回帰モデルでの進歩は，確率分布の仮定に関連して新しい行き方を示している．複雑な関数系を仮定しても，良い回帰モデルの推定はできない．現実にモデルを適合させようとすると，母数の数が不必要に多くなってしまうからである．むしろ滑らかさを事前情報として仮定する方が旨く行く，例えば柏木・柳本 (1990)．機械的な処理が最も得意な分野と考えられる情報処理技術分野で，柔らかな情報処理に基づいた，「それなりに妥当な解」が提唱されている（新情報処理技術調査研究委員会 (1992)）．これらの事実は厳密さのみを追求したデータ処理の限界を示唆している．

次にモデルと現実とのずれを無視したとしても，損失関数の決定は容易ではない．実際に行動の選択は損失関数にセンシティブなことが多いから，損失関数の決定は重要な問題である．一方では損失関数を旨く選ばなければ，数学的な展開を難しくする．だから損失関数は理論的な都合を考慮して決めるしかない．実際，理論的に都合の良い損失関数を仮定すると，様々に理論的に綺麗な結果が得られる．この事実こそが決定理論，さらには数理モデル，の研究が今日隆盛している理由と思われる．しかし一方では実用的でない理由でもある．

例え損失関数が定義できたとしても，それでも問題は残る．本当の目的は危険関数を小さくすることであるが，これを実現することは難しい．話を簡単にするために推定問題を例にして考えよう．従って，行動空間は母数空間となる．ところが危険関数を小さくするという規準では，推定量を構成することは困難で，例えば損失関数を小さくする推定値を選ぶといった方法しかないことである．実際危険関数は推定量と母数との関数である．推定量を決めるためには，各標本に対して推定値を定義しなければならないので，危険関数を小さくする推定量を具体的に構成することは，実際には困難である．一方損失関数 $L(a(\boldsymbol{x}),\theta)$ は観測値と母数の関数だから，これを最小にする規準で推定値を決めることはできる．損失関数が特に複雑でなければ，解

析的にあるいは数値計算によって十分に正確に求めることができる．しかし損失関数を小さくすることが元来の目的ではない．推定量の良さを調べる規準はあくまでも危険関数の小ささである．この一見混同し易い点に決定理論に対する過信の一因があると思われる．損失関数と危険関数の違いは他の面でも微妙な，しかし重要な違いとして現れる．損失関数は推定値と母数との一種の距離である．ところが実際の推定問題では，標本と推定値の距離が小さくなるように推定値を選ぶ．この距離を小さくすると，結果として危険関数を大きくしてしまうことが多い事実を注意する．

この事情をもう少し分かり易く述べるために，最尤推定量 $\hat{\theta}$ について論じる．分布が指数分布族であれば，任意の標本 $\boldsymbol{x}$ とその推定値 $\hat{\theta}(\boldsymbol{x})$ に対して

$$\log \frac{p(\boldsymbol{x};\ \hat{\theta}(\boldsymbol{x}))}{p(\boldsymbol{x};\ \theta)} = KL(\hat{\theta}(\boldsymbol{x}),\ \theta)$$

が成り立つ（Kullback (1959)）．ただし KL は $\hat{\theta}$, θ から誘導される（標本空間と同じ次元の）分布間の Kullback-Leibler 損失関数である．上式から分かるように，この損失関数を最小にする推定値は $\hat{\theta}(\boldsymbol{x})$ であって，しかも尤度を最大にする．しかし最尤推定量 $\hat{\theta}$ がこの損失関数に基づく危険関数を最小にするとは限らない．むしろ最小にする場合が例外的である．反例となる良く知られた例としては，Stein 推定量がある（James and Stein (1961)）．従って尤度を最大にするという，直観的には魅力的な規準であっても，この規準についての最適な解は良い解とは限らない．結局原点に戻って危険関数を小さくするように推定量を選ぶしかない．そうすると数学的に扱い難くなってしまう．久保川 (1993) が解説しているように，Stein 問題，即ち多次元正規分布の平均値の推定でも十分に満足できる推定量を見つけることは容易ではない．

本節を要するに，決定理論ではデータ解析で帰納的な推論という重要な役割を放棄したが，実際に最善の解を得ることは困難なことが多い．従って決定理論の枠組みで統計的検定に代替し得る，あるいは完全に凌駕する，推測方式が得られることは想像の埒外である．

5. 統計的検定の発展

いわゆる近代推測統計学の核心的な手法として，統計的検定は理論統計学の研究者によって，研究されてきた．今日でも専門雑誌に多くの研究が発表され，またその適用が広がっている．この節ではもう少し大局的な面から，統計的検定が発展している道筋をたどる．統計的検定と同じ考え方の方法は，最善の推測をするというより簡明で妥当な帰納的推測の方法であることが分かる．

5.1 区間推定

標本がランダムな変動に支配されているとすれば，母数の推定を一点で推定するよりも，区間で推定した方が良いことがある．記述を簡単にするために，母数は一次元とする．適当な信頼区間 $I(\boldsymbol{x})$ を

$$\Pr\{\theta \in I(\boldsymbol{x})\} = 1-\alpha$$

を満たすようにとる．これは $100(1-\alpha)\%$ 信頼区間と呼ばれる．統計的検定での棄却域との関係については，「細かい点を除けば $\theta \in I(\boldsymbol{x})$ と $\boldsymbol{x} \in R$ とは同値である」が成り立つ．文献にある具体的な手法の中で，この同値関係が成り立たない場合は殆ど近似法など理論的に不備な場合である．

同値関係が成り立つ場合には，信頼区間を得れば同時に統計的検定が容易にできることになる．更に信頼区間をみれば，検定結果とは別の情報を得られる．こうした次第で，近年では統計的検定の結果を示す場合に，信頼区間を同時に与えることが多くなった．

区間推定の場合も，水準 α の決め方には何らの方法も提案されていない．しかし水準と確率分布を与えたときの信頼区間の決め方については膨大な研究がある．

5.2 同等性の検定

2節で述べた2つの平均値の差の検定問題の場合，本当の所は2つの平均が等しいという帰無仮説は常に誤っていると考えて良い．実際2つの平均が全く同一とは考えられないからである．従って帰無仮説が誤っていて，対立仮説が，これも厳密には誤っているが，より真実に近いのは疑いがない．だからデータが有っても無くても帰無仮説が採択されないのは当然である．

しかし現実問題として μ_1 と μ_2 との差が小さいなら，2つの平均値を同一視してもおかしくはない．実際薬の製剤法を変えたり，より廉価な素材に変えた製品が，以前のものと同等かと問うとき，厳密に同じであることを求める訳ではない．実際従前の製品間にも変動がある．この場合予め決めた区間 I に $\mu_1-\mu_2$ の信頼区間が含まれるか否かで判断することが一つの方法である．区間 I が $(-a, a)$ であれば，帰無仮説 $\mu_1-\mu_2=-a$ あるいは $=a$ が各々対立仮説 $\mu_1-\mu_2>-a$ あるいは $<a$ として，各々水準 $\alpha/2$ とした片側検定が共に有意になることと同等である．この方法では，μ_1 と μ_2 が極めて近い場合には，標本サイズを大きくすると殆ど1の確率で同等と見なすことになる．

薬効評価の分野では，片側の同等性の検定と呼べる検定が採用されるようになった．その背景には新薬とプラセーボでなく新薬と標準薬と比較する傾向が強くなったこと，副作用の強弱などがある．我国では柴田・開原 (1981) にその嚆矢をみ，現在の薬効評価統計ガイドライン (新医薬品統計解析検討委員会 (1992)) にも採用されている．この場合予め決める区間は上半開区間 $(-a, \infty)$ になる．しかしながら本小節で述べた同等性は，厳密な意味の同等性とは異なった概念だから，両者を明確に区別する新しい用語が必要とされている．次節との関係で言えば，ここでの同等性は実質的同等（あるいは仮想的同等）と呼べる．

統計的検定で帰無仮説が棄却されなかったとき，帰無仮説が正しいとみなすには無理がある．本小節での方法は1つの妥当な方法とみられる．妥当とは思われるが，この方法が最善かというと何もそのような結果は得られていない．同等性に関する推測のような基礎的な問題では，最も良い方法を構成することは極めて困難 —— おそらくは不可能 —— である．

5.3 実質的安全量

化学物質が安全であるとは，ヒトに対して全く害作用を及ぼさないことである．しかしそのようなことは，特にその物質が何らかのアクティブな作用があれば，考え難いことである．まして本当に安全であることをデータで証明することは全く不可能である．

化学物質の作用の強さは用量 d に依存する．用量 d のとき害作用を起こす確率を $P(d)$ とする．$P(d)=0$ であれば，用量 d のとき本当に安全である．$P(d)$ は単調増加と仮定する．そこでデータから $P(d)$ の信頼区間を $(0, \hat{P}(d))$ とする．予め決めた危険水準 ρ に対して，上側信頼限界 $\hat{P}(d)$ が ρ と等しいときの d を $\hat{d}$ として実質的安全量 (Virtually Safe Dose) と呼ぶ．用量 d が $\hat{d}$ より小さいとき安全とみなす．この方法は Mantel and Bryan (1961) で提案された．データから危険度を高めに推定して予め定めた危険水準と比較している．毒性が計量値として正規分布で表される場合の方法は竹内 (1973) に与えられている．

実質的安全量の考え方は，従来の安全性の評価と大きく違う．従来は絶対的な安全量を議論

表2．信頼限界の推定と安全性評価及び信用調査の論理．

	上側信頼限界	安全性評価	信用調査
目的	母数の妥当な上限値	安全な量の推定	信用度の判定
提出する仮説	母数がある値以下	(安全である) 危険性がある値以下	(信用できる) 危険性がある値以下
問題の提出者	仮説の提出者	研究者・企業	信用を受ける人
挙証責任	同上	同上	同上
証拠が全くないとき	∞	安全とは見なされない	信用できるとは見なされない
判断の規準	上側信頼限界が小さい	実質的安全量 (V.S.D.)	審査規準に合格
許容する誤りの程度	水準 α	信頼水準 α 危険水準 ρ	担当者の判断
判断の誤り	真の値が大きい	危険な物質の利用	不良債権の発生
	真の値が小さい	有用な物質の禁止	金融機能の制約

していた．危険水準を導入することによって，計量的に安全さの程度を評価している．丁度推測統計学では誤差を積極的に評価することによって，従来の常に誤差を小さくするという考えを脱却して，データ処理の新しいパラダイムを切り拓いた事情に対応している．

実質的安全量の考え方は前小節の同等性の検定と同じ構造をしている．帰無仮説 $P(d)=a$, 対立仮説 $<a$ において帰無仮説が棄却されることと，d が実質的な安全量より小さいことと同等である．従って統計的データ処理としてはごく自然な考え方である．しかもこの考え方は，必ずしも統計解析に固有の特殊な考え方ではない．一例をあげれば金融での危険性を回避するための信用調査とは極めて近い．信用供与を与えるためには，いわゆるこげつきを回避する必要がある．しかし絶対に安全な取引先はない．現実の社会では信用供与をルーズに行った方が，一見社会が円滑に動くようにみえる．しかしルーズな信用供与は膨大な不良債権の発生となってはね返る．表2ではその対応を与えている．信用調査でも最善の方法は考えるべくもない．

統計的検定理論の展開は，帰納的推論あるいは実証的研究の根幹の分野で，目立たないながらも着実に進歩している．

6. 結　　語

実証的研究の重要性が疑いなく受け入れられるようになって，実証的な研究を支えるデータ処理技法の向上が求められている．データ処理の目的は観測に基づく推論であるから，なんらかの意味での帰納的推論である．統計的検定は帰無仮説の下でその仮説を棄却する確率を水準以下にするという新しい手法であった．帰納的な推論のための手法としては最初に登場した．現実に実証的研究，特にデータの裏づけの無い雑多な仮説を淘汰する上で輝かしい役割を担った．また一方で無名で権威のない研究者による一見奇抜なアイディアによる仮説であっても，データによって自己の仮説の正当性を主張できるルートを拓いた．実際，治療評価，薬効評価，品質管理，育種の分野では基本的な方法となった．

統計的検定の論理はごく常識的で妥当である．しかし繰り返し強調したように，最適な方法ともみなされないことも事実である．また結果の記述が回りくどく，またデータに含まれる情報を完全に汲み取っていない物足りなさが残るようである．更に2つの仮説を判別するとか，評

価関数を最大にするという，日常に親しんでいる論理とは違うための違和感も否めない．

統計的決定理論，あるいはこれに似たゲーム理論，モデル選択理論，判別分析，は評価関数の定義が容易である．数学的構造が簡単で研究成果が上がり易く，とっつき易くもある．しかしながら本稿で考察したように，決定理論は帰納的推論に直接役立つ論理ではない．また評価関数を最大にする規準も実際にはその規準が実現される訳でもない．

良い行動の選択を論じることで真実に迫ることができる，という考えがある．実際科学的知見の獲得も，実際に行動の選択に反映されなければ，役立たずである．また真実の裏づけのない行動の決定は見込みがなさそうに感じられる．しかし一方では良い行動の選択は必ずしも真実に迫らないこともまた事実である．患者とか医者が「効果がある」と直観した治療法は関係者には良い行動であると判断されるが，実際には殆ど効果が認められなかった例は多い．この事実こそが統計的検定が広く普及した理由である．一部の人が主張し，多くの数理科学者が漠然と考えているような，決定理論が統計的検定を無用にするという事はありそうにもない．

主として統計的検定を批判した方法の1つに，探索的データ解析がある．データにありのままに語らせようとする．その考え自体は尤もであるが，実際には手持ちのデータから，推論もなく自己の主張したい仮説を支持するための証拠を探す手段に急速に傾斜している．データが仮説を支持することを積み重ねることは，Popper (1957) が強調するように，実証的研究にとって有用な方法ではない．

決定理論は，数学的な簡明さのため，著しい理論的進歩をとげた．統計的検定のカバーしない分野で貢献するに違いない．しかしその方法は現実的には理論的結果のようには最善の手順を与える訳ではない．結局統計科学の方法の有用性は実証的研究にどれだけ有効に役立つかにある．

謝辞と附記

査読者のコメントは原稿を改訂するのに大変役立ちました．厚くお礼を申し上げます．また山本英二氏（岡山理科大）と佐藤俊哉氏（統数研）からもコメントを頂いたことに感謝します．

査読者の指摘した大局的なコメントとして

1. 「妥当性のような主観的あるいは感覚的な概念が科学を論じる場合に適切かどうか……」
2. 「統計的検定の定義が不明である．Fisher 流の統計的検定なのか，Neyman-Pearson 流なのか，それとも……」

の2点がありました．こうしたコメントには，査読者との基本的見解の違いがあって答えにくい．しかし査読者の不満も理解せざるを得ないので，短く著者の意見を述べる．

1のコメントに対して：論文及び4節のタイトルを始め原稿に手を入れた．勿論最適性よりも妥当性が良い規準であることを主張した訳ではない．4節でも述べたように，著者の意とする所は合理的な規準によって最適な解を得ても必ずしも妥当な推論を導かないことを強調することにある．一方統計的検定の論理は最適性から導かれないが，妥当な推論形式である．

2のコメントに対して：2つの流儀の違いが強調されすぎていると考える．Neyman-Pearson はあくまで統計的検定の枠組みの中で理論的整備を行ったにすぎない．しかし決定理論では行動空間を導入して，推測を行動の決定と定式化している．この理論的整備が，一方で帰納的な推論形式からの乖離をもたらしたと考える．

参 考 文 献

James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proc. Fourth Berkeley Symp. on Math. Statist. Prob.*, Vol. 1, 361-380, Univ. of California Press, Berkeley.

柏木宣久，柳本武美 (1990). 平滑化法とそのソフトウエア， 応用統計学，**19**, 79-92.

久保川達也 (1993). 統計的推定理論の最近の展開，日本統計学会誌，**22**, 257-271.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.

Mainland, D. (1963). *Elementary Medical Statistics*, Saunders, Philadelphia (柏木 力 訳『医学統計の基礎』，岩波書店，東京，1971).

Mantel, N. and Bryan, W.R. (1961). "Safety" testing of carcinogenic agents, *Journal of the National Cancer Institute*, **27**, 455-470.

Popper, K.R. (1957). *The Poverty of Historicism*, Routledge & Kegan Paul, London (久野 収 他訳『歴史主義の貧困』，中央公論社，東京，1961).

柴田義貞，開原成允 (1981). 試験薬が標準薬と同等またはそれ以上の有効率をもつことの判定法，臨床薬理，**12**, 421-426.

新医薬品統計解析検討委員会 (1992). 臨床試験の統計解析に関するガイドライン，厚生省新薬第 20 号.

新情報処理技術調査研究委員会 (1992). 委員会報告書， 通商産業省機械情報産業局.

竹内 啓 (1973). 許容基準の定め方 —— 汚染に対する安全基準の問題，応用統計学，**3**, 1-13.

Wald, A. (1950). *Statistical Decision Functions*, Wiley, New York.

柳本武美 (1981). 臨床試験における統計的検定，『薬効の評価基準』(七川歓次 編)，73-80，永井書店，大阪.

柳本武美 (1991). 統計的検定における帰無仮説の理解，応用統計学，**20**, 97-108.

Validity of a Procedure in Statistical Inference

Takemi Yanagimoto

(The Institute of Statistical Mathematics)

The important role of the theory of the statistical test in statistical inference is discussed in comparison with that of the statistical decision theory. The statistical test is advocated as a way to ensure sound reasoning in induction. Meanwhile, the selection of an optimum action under a suitable loss function is the primary goal in the statistical decision theory. It is emphasized that the statistical test cannot likely be taken place by other decision theoretic methods induced from the statistical decision theory and related theories such as model selection and the game theories. Recent developments in the statistical test are reviewed from this perspective.

This paper follows up the author's two previous articles.

Key words: Criterion for an estimator, optimum procedure, induction, statistical test.