Argonauta 11: 3-16 (2005)

Type II error と Power analysis

大垣俊一

以前、私が参加したある研究会で次のようなやりとりがあった。発表者はある種のカニのデコレーティング（体表に藻類などを付着させる行動）を調べ、デコレーションのあるオスとないオスに対する雌の選好性を、十数例についての観察結果をもとに二項検定によって比較した。そして有意な差が認められなかったことをもって、デコレーティングはメスによるオス選択に影響を及ぼさないと結論した。これに対し参加者から、「差がない」と結論するためにこの標本数は十分なものか。極端に言えば標本数を減らして有意差を出にくくすれば、常にこうした結論が可能になるのではないか、といった疑問が呈された。こうしたやりとりはこの例に限らず、検定で有意差が出なかったときしばしば交わされるものである。

「差がない」ないし「同じである」という帰無仮説（H_0）を立てて検定を行ったとき、有意差が出れば、一定の危険率（α）のもとに「同じであるとはいえない」＝「差がある」と結論され、このときの α を Type I error（第 1 種の過誤）と呼ぶ。しかし有意差が出なかった場合、その結果は単に「差があるとは言えない」ことを示すにすぎず、「差なし」と結論することはできない、というのが、一般の統計学教科書が検定の基本として解説するところとなっている。もしも「差なし」と言いたいのであれば、それは検出力（statistical power ないし単に power）の問題になるのであり、十分な検出力を持った検定において初めて、「差なし」という結論が、一定の危険率（β）のもとに肯定される。この β が、今回テーマとする Type II error（第 2 種の過誤）である。この、Type I と II エラーの関係を定性的に示したのが、統計の教科書に出てくる error table（下表）と呼ばれるものである。

		検定結果	
		H_0 採択（＝差なし）	H_0 棄却（＝差あり）
自然の状態	H_0 真（＝差なし）	正しい判定	Type I error
	H_0 偽（＝差あり）	Type II error	正しい判定

自然状態において実際には差がないとき、検定結果に従い「差なし」とすればそれは正しい判定をしたことになる。しかし「差あり」とするとそれは誤りであり、これが Type I error である。逆に自然の実態として差があるとき、「差なし」としてしまうと Type II error が発生し、「差あり」とすれば正しい判定になる。

「差なし」と結論したいケースは、「差あり」と言いたいケースに劣らず生態学研究において頻発するように思われるが、実際に β を計算した例は極めて少ない。Peterman（1990）は水産生物関係の研究論文を review した結果、検定において有意差が出なかった 160 例中、83 例で著者は事実上「差なし」とみなしていたが、

検出力（ないし Type II エラー）を計算していたのはわずか 3 例だったと述べている。なぜこのような事態になっているのかについては、後に述べるように様々な理由が考えられるが、一つの背景として Type II エラー概念のわかりにくさがあると思われる。そこで本稿では Type II エラーの理論と検出力計算の実際について解説し、その問題点を検討してみる。

１．Type I error（第 1 種の過誤）

Type II エラーは、我々が通常の検定で問題にする Type I エラー（危険率）と、多くの共通点を持っている。そこでまず Type I エラーについておさらいしておく。Type I エラーの計算はあらゆる統計検定で行われており、具体的な計算方法は検定の種類ごとに異なるが、基本的な考え方は同じである。ここでは具体例として、ある種の捕食性巻貝 A 種の、フジツボ B 種に対する捕食圧を、対応 2 資料 t 検定で分析するケースを取り上げる。海岸で B 種の高密分布帯に網囲いの実験区とコントロール区を隣り合わせに 4 ペア設置し、区内すべての捕食性巻貝を除去した上、新たに実験区のみに同数の A 種を導入する。時間がたってから、4 組内での実験区とコントロール区の B 種の密度差（実験区における減少量）を調べたところ、30 個体／㎡（標準偏差 SD=20）であった。この結果から、A 種による B 種への捕食圧を考えるとすれば、帰無仮説（H_0）は

H_0：A 種の存在は B 種の密度を減少させない。

となる。これが事実とすると、この実験における B 種の変化量は平均値 0 の、図 1 のような分布を描くはずである。この分布は同じ実験をくりかえし無数に行った

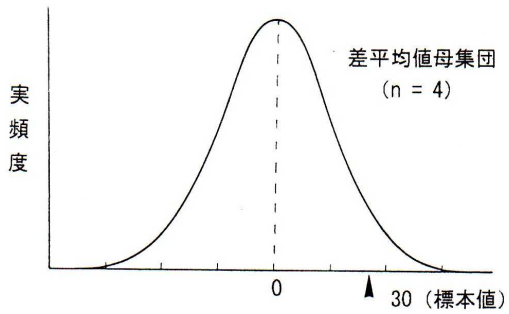

図 1. 野外実験における B 種減少量の実頻度分布（母集団）

ときの、B 種変化量平均値の母集団を表し、捕食の影響がなくても計数誤差などにより、0 の回りにばらつくと考えられる。なお実験は小標本であるから t 検定が妥当し、その条件として B 種変化量は正規分布することを要する。ここで次の t の値を計算する。

$t = (m - \mu) / s \times \sqrt{n}$（m, 標本平均；$\mu$, 母集団平均；s, 標本分散；n, 標本数）

これによって図 1 の実頻度分布を面積 1.0 の t 分布（自由度 n－1）に変換（標準化）し、得られた m の値の、0 からのズレを確率的に表現できるようにする。この例の場合 μ は 0 だから、$t=(30-0)/20\times\sqrt{4}=3.0$ となる。

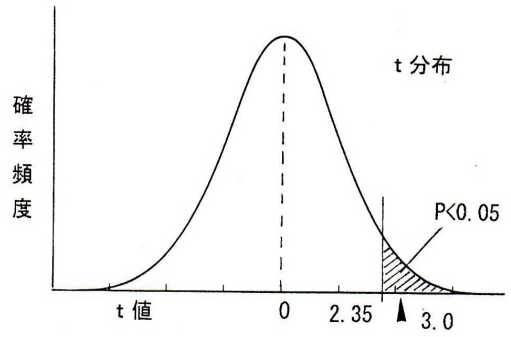

図 2. 図 1 の実頻度分布を変換した、自由度 3 の t 分布。

自由度 4－1=3 のもとでは、t=2.35 以上になる確率が 0.05 であるから、図 2 のように、得られた m は P<0.05 の棄却域に含まれる（A 種の存在が B 種を増加させることはほとんど考えられないので、片側検定とした）。この結果は、もし B 種の影響が 0 であるとすると極めて起こりにくい（偶然こうなったとすると 20 回に 1 回以下しか起こらないほどまれなことが起こった）という意味において、H_0 は真らしくないことを意味している。統計の言葉で言えば「危険率 5%で（ないし有意水準 95%で）H_0 は棄却された」ということである。従って「A 種の B 種に対する捕食圧は存在するらしい」と結論することになる。このときの 5%が Type I エラー（α=0.05）である。エラーとか危険という表現は、H_0 が実際に真であってたまたま 0 から飛び離れた値になったのかもしれず、それなのに誤って H_0 を偽とした確率が 5%残されている、という事情を背景としている。

このように Type I エラーを問題とする通常の検定操作においては、「差あり」ということを直接証明するのではなく、いったん「差なし」の仮説をたて、それを否定することによって（しかも一定の危険率を容認しつつ）対立仮説の正当性を主張する（つまりの A の否定 ≒ not A の肯定）という、屈折した論法を用いているところに、統計を学ぶ者がなじみにくさを感じる一つの要因があると思われる。しかし Type II エラーの場合は、このわかりにくさが文字通り倍増することになる。

２．Type II error（第 2 種の過誤）と Statistical power（検出力）

Type II エラーの概念をつかむためには、やはり具体例から入るのがわかりやすい。今、ある種の貝に X 型と Y 型の二つのタイプがあり、両者はサイズが異なっていて、それぞれのサイズ分布はよく調べられているとする。X 型は平均殻長 14.0mm (SD, 3.0mm)、Y 型は平均 17.0mm (SD, 3.0mm) の、共に正規分布を示

すとしよう。海岸のある地点から 9 個体のサンプルが得られ、その標本平均殻長は 15.0mm (SD, 3.0mm) であった（標準偏差は別に等しくなくてもよいが、ここでは計算の便宜上等しく設定してある)。これらのサンプルがすべてどちらか一方に属することは確実で、かつサイズ以外の判定基準はないものとすると、サンプルは X 型、Y 型どちらに属すると考えられるだろうか。標本平均値は X 型に近いので、帰無仮説は次のように設定する。

H_0：サンプルは X 型に属する（＝母集団平均殻長は 14.0mm である）

まず、H_0 の分布のみに注目して、通常の Type I エラーによる検定を行うと、この場合の $t=(15.0-14.0)/3\times\sqrt{9}=1.0$ となる。t 分布表で、自由度 9－1=8 における P=0.19 なので、危険率 α =0.05 のもとで、標本集団と X 型母集団との差は有意ではない。この場合、Y 型は X 型より大きいので、X 型母集団平均からのズレは、その右側（サイズの大きい側）だけ考慮すればよく、片側検定となる（Type II エラーの分析は、対立仮説の存在する側のみを意識するので、常に片側検定である)。

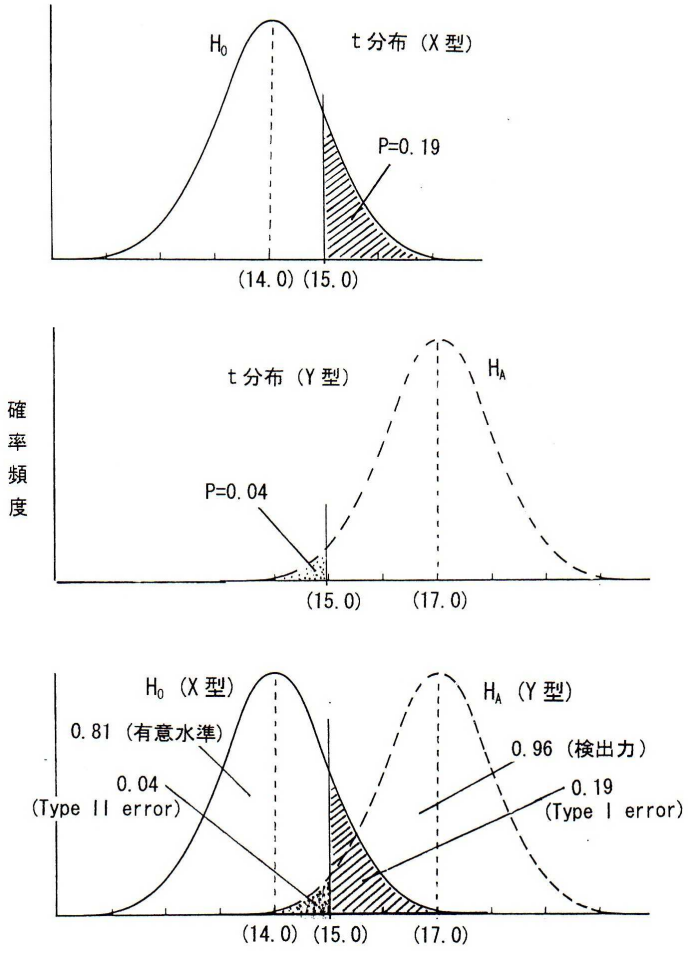

図３．Type II エラーと検出力。(14.0) は 14.0mm の t 変換値を示す。

以上の状況を図３上段に示した。しかしこの結果から、直ちにサンプルが X 型に属すると結論することはできない。Y 型の中で、特に小さいものを採集した可能性があるからである。そこで今度は対立仮説

H_A：サンプルは Y 型に属する（＝母集団平均殻長は 17.0mm である）

を立て、その真偽を検討する。$t=(17.0-14.0)/3\times\sqrt{9}=2.0$、自由度８における P=0.04 となり、Y 型平均値との差は有意である（片側検定）。これを図３中段に示す。この時の P=0.04 が Type II エラー β、$1-\beta=1-0.04=0.96$ が、検出力と呼ばれるものである。図３の上段と中段を合せたものが、検出力の解説で必ずと言ってよいほど出てくる下段の図となる。私は以前、Type II エラーを説明するこの図がわかりにくいので、あまり重視せずに言葉の説明に頼って文献を読んでいたが、そのほうがかえってわかりにくく、混乱する。従って文による説明を読むときも、この図を思い浮かべて、各記述がどこの面積を意味しているのかを考えることを勧めたい。ここで用語を整理しておく。

Type I error（α、危険率）：　実際には H_0 が正しいにもかかわらず、それを誤りと判定する確率。差がないのにあると判断してしまう危険の率、と言える。

Significance level（$1-\alpha$、有意水準）：　H_0 が正しいとき、それを棄却せずにすます確率。あるいは H_0 を棄却した場合の信頼度。差を有意と判定するときの水準、といったほどの意味か。

Type II error（β）：　実際には H_A が正しいにもかかわらず、それを誤りとみなす確率。これを H_0 の側から見ると、H_0＝ not H_A により間接的に、H_0 が誤っているにもかかわらず、それを正しいと認めてしまう確率（あるいは、H_0 を正しいと認めたときに、それが誤っている確率）となる。

Statistical power （ないし単に power、$1-\beta$、検出力）：　H_A が正しいときに、それを棄却せずにすます確率。これを H_0 の側から見ると、H_0＝ not H_A により、H_0 が誤っているとき、それを正しく認める確率となる。差の存在を正しく検出する力、と考えてよいだろう。別の見方をすると、検出力が大きいということは、Type II エラー β が小さいことである。β は上に述べたように「H_0 を正しいと認めたとき、それが誤っている確率」であるから、β が小さいということは「差なし」と判定したときの信頼度が高いことになる。つまり、「差なし」とか「同じ」と言うためには、検出力が十分高くなければならない。そしてこれが、実用上最も重要な検出力の意味である。検出力を計算することを、power analysis（検出力分析）と呼ぶ。

ここで注意しなければならないのは、Type I エラーによる検定が、H_0 の情報のみで実行できるのに対し、Type II エラーないし検出力の決定は、必ず対立仮説の特定を要するということである。検出力計算の理論的根拠である「H_0＝ not H_A」は、H_0 でなければ H_A しかありえないという状況でのみ成り立つ。H_A の分布の位置と形が決まらなければ β を計算できないことは、本節冒頭に示した具体例や、図３の分布図から明らかだろう。これについては、次のような比喩が可能かもしれない。ある学校の教室に消しゴムが落ちていて、A 君がそれを自分のものだと主張した（H_0）。それが誤りであることは、A 君とケシゴムについての情報のみから判断

できる。たとえばA君がその日学校を休んでいたとか、消しゴムにA君以外の名前が書かれていた、などである。しかしA君のものである、ということはA君の状況だけから判断できない。たとえばケシゴムにA君の名前が書かれていても、クラスに同じ名前の生徒があと二人くらいいるかもしれない。ケシゴムが落とされたと思われる時刻に教室にはA君とB君の二人しかおらず、しかもB君のものではないことがわかっている、というように周囲の状況がはっきりして初めて、ケシゴムがA君のものであることを確定できる。

Type IIエラーの確定に、Type Iエラーに比べて多くの情報を必要とするのは、ある意味では帰無仮説による論証の必然と言える。統計検定の論法としては「差なし」という帰無仮説を設定し、それが誤りであることを示すことによって逆に「差あり」と主張しようとする。したがって「差なし」を示そうとすれば、「帰無仮説（H_0）以外に1つしかない対立仮説（H_A）」という状況を想定し、そのH_Aを否定することによってH_0を肯定する、という形を取らざるをえない。Type Iエラーの項で指摘した「屈折した論法」がこのように二重に想定されているところに、Type IIエラー概念のわかりにくさがあると言えよう。またこのことは同時に、「帰無仮説以外の唯一の対立仮説」を想定できなかったとき、Type IIエラーの決定もまた不可能になることを意味している。

3．検出力に影響する要因

Type IIエラーや検出力は、前節で示した方法で計算できる。実際の計算方法は検定の種類によって変わるが、それは教科書類（Cohen 1977, Zar 1999）を見てもらうことにして、ここでは、一般的に検出力に影響を与える要因について検討してみる。図4は、H_0とH_Aに対応する確率分布の概念図である。

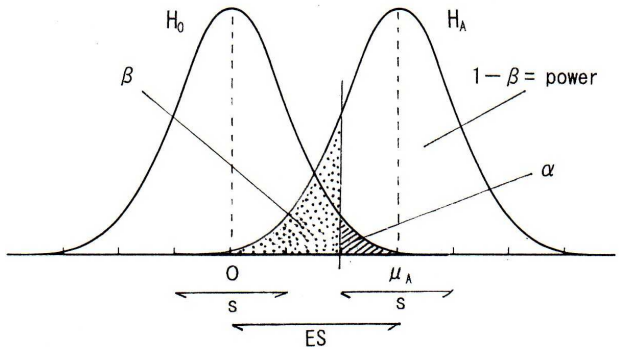

図4．検出力に影響を与える要因。

この図で検出力はH_Aの分布のうちの斜線部の面積で表され、それに影響を与える要因としては、Type Iエラー（α）、母集団標準偏差（σ、しかし通常は標本標準偏差sで代用）、H_0とH_Aの距離（これをeffect size, ESと呼ぶ）、そしてsに影

響を与える標本数（n）などが考えられる。図５に、これらの要因の影響のしかたを示した。

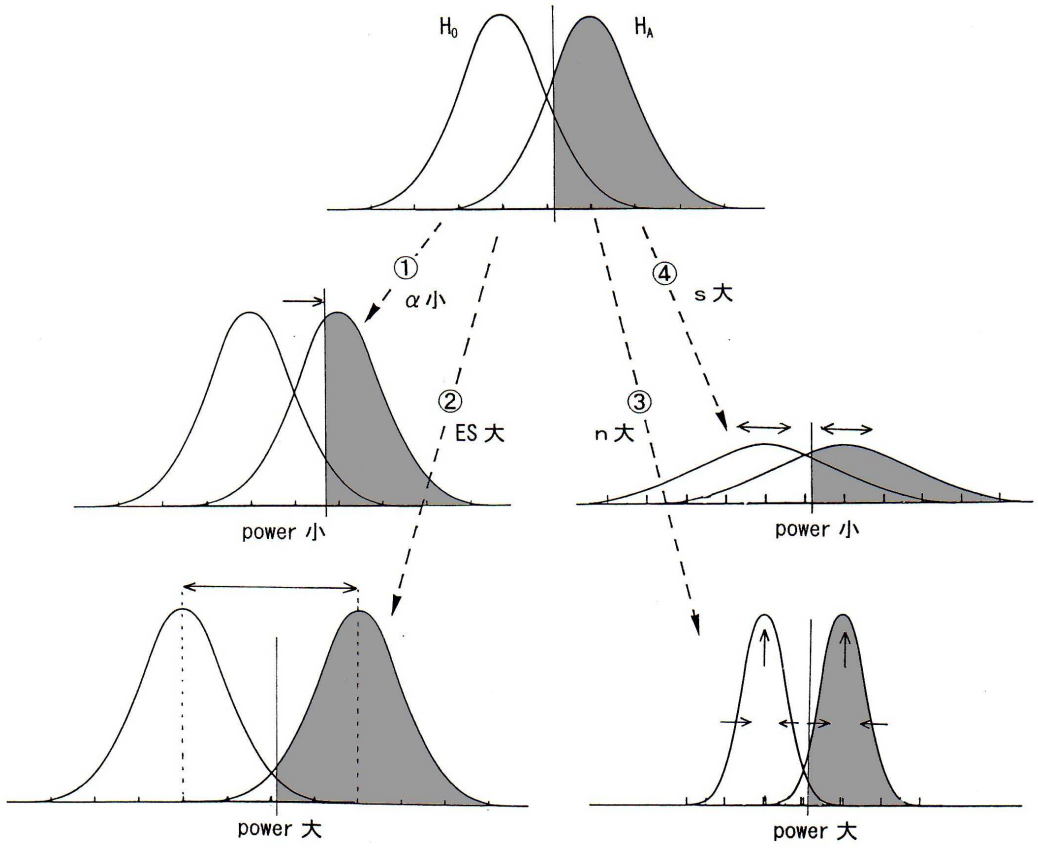

図５．４つのパラメータと検出力（灰色部）。

① **α（危険率）**：　H_0に対する有意水準を右にずらし、αを減少させると（0.05→0.01など）斜線部の面積は減少し、検出力は下がる。これがαとβのtrade-offといわれる関係で、「差あり」を確実に言おうとするほど、実際には差があるにもかかわらず、「差なし」と判断してしまう危険が増すことを示している。

② **effect size, ES**：　H_0とH_Aの母集団の位置の隔たりが大きいと2つの山が左右に離れるので、同じ値のαに対して斜線部分の面積が増大する。前節の実例では、X型の母集団平均14.0mm、Y型17.0mmで考えたが、もしもY型20.0mmなどであれば、平均値の標本集団をX型とみなしたときの信頼度が増すことは、常識的にも理解されるだろう。

③ **n（標本数）**：　nが大きいと、これを分母にもつsが小さくなり、分布の尖度（尖り具合）が増す。H_0とH_Aの分布はそれぞれ平均値の回りに集合して吊り上げられたようになり、重なりが小さくなる。そのため同じ値のαに対して斜線部の面積が増し、検出力は上がる。αを変えずに検出力を上げるには標本数を増やせばよ

い、とよく言われるのはこの点に根拠がある。

④ **s（標本標準偏差）**：　この値が大きいと H_0 と H_A の分布は横に広がり、重なりの部分が大きくなる。そのため同じ α に対して斜線部の面積が小さくなり、検出力が落ちる。s は標本数 n の関数なので n の影響も受けるが、それ以外に実験条件や対象種の性質によるばらつきもこれに加わってくる。

検出力は α、ES、s、n の 4 要素に依存することにより、一般的に

$\text{Power} = f(\alpha, \text{ES}, s, n)$

と書ける。この式から、power そのものを含めた 5 つのパラメータのうち、4 つが決まれば残り 1 つが計算できる理屈となる。5 つのパラメータの中では、この式のように power を α、ES、s、n から決定する場合と、n を power と α、ES、s から求める場合がよく解説され、前者を a posteriori（事後）、後者を a priori（事前）の検証と呼んでいる。前者を「事後」というのは、調査や実験が終わって検定を行い、特に有意差が出なかったとき、その枠組みでどれくらいの検出力があるかを調べるために行われることが多いからである。一方後者の計算は実験開始前に、有意差が出なかったときに備え、一定の検出力を得るためにどの程度の標本数が必要かを知るために用いられることが多いために「事前」という。

4．　Power analysis の問題点　－Effect size の怪

power は α, ES, s, n の 4 つのパラメータを決定することで計算できる。ではこれらの 4 変数の決定には、どのような問題があるだろうか。まず、危険率 α は研究者の判断により、0.05、0.01 などと任意に選べる。n は標本の実数だから、以上二つは問題ない。s（標本標準偏差）は理論上 σ（母集団標準偏差）であるべきだが、大標本であれば、標本分散の母集団分散からのずれはそれほど気にしなくてよいし、小標本の場合は t 検定となるので、母集団正規分布の条件のもとに標本分散を用いることが容認される。問題は effect size, ES である。

ES は理論的に言えば、帰無仮説 H_0 と対立仮説 H_A との距離、ないしそれを標準化したものである。教科書や総説では、'the degree to which the phenomenon is present in the population'（Cohen 1977）とか、'the magnitude of the true effect for which you are testing'（Peterman 1990）などと定義されている。しかし「現象が（実際に）母集団において存在している程度」とか 'true effect' というのは何だろうか。もしも H_A の位置が 'true' であって、それがあらかじめわかると言うのなら、H_0 を立てて検定するまでもない。この点について Rotenberry & Wiens（1985）は、power analysis のためには「効果がないと証明しようとしているまさにその現象に対して、効果を見積もらなければならない」と述べ、対立仮説の存在そのものに疑問を呈している。本稿 2 節で例とした、貝の X 型と Y 型のように、power analysis を矛盾なく行えるケースは、あるにはある。それは、「もし H_0 でないとすれば H_A 以外にありえない」という状況である。これを 2 標本集団間の平均値差の検定に当てはめると、「差（効果）はないかもしれないが、あるとすれば

この値以外に考えられない」ということである。生態学の調査、実験においてこのような状況は、ありえないとは言えないまでも極めて想定しにくい。そしてここに、Rotenberry & Wiens の本質的な疑問が発生するのである。

実はこの ES の性質をめぐり、power analysis についての統計学者の立場は 2 つに分かれている（Cohen 1977）。統計学の租 Fisher は、明確に位置を決められるのは H_0 のみであって、H_A の分布はその右側に連続的に変化して存在するとみなした。つまり H_0 とそれ以外、というくくり方である。このようにすると Type II エラーは計算できず、従って Fisherian に power analysis の概念はない。Sokal & Rohlf（1981）が「H_A は特定できるものではなく、β や $1-\beta$ は様々な値からなる連続体」と述べているのは、この立場に立つものと言える。ちなみに Sokal & Rohlf のテキストは概念としての Type II エラーを解説するものの、検出力計算の具体的手法は全く示していない。ホーエル（1981）においても、検出力分析の具体例は本稿 2 節の X 型 Y 型のように、H_A の位置づけが明快なものに限定されている（ホーエルは、発掘された人骨が 2 種のよく知られた化石人類のどちらに属するかを、頭骨のサイズから判定する例を示している。2 節の具体例はこれをアレンジしたものである）。一方、初期統計学のもう一人の大家 Pearson は H_A を特定できるものと捉え、こちらの流派において power analysis の計算方法が発展した。その場合、肝心の ES は当該研究に関連する理論からの推定値や、以前に行われた同種の研究、予備調査における値などから決めるという（Cohen 1977, Andrew & Mapstone 1987, Underwood 1997）。それらに依拠できない場合、実験によって得られた 2 標本値の差そのものを使っている場合もある（Zar 1999）。H_A として理論値を用いるのは、その理論を反証しようとする場合には意味があるだろう。しかし以前に行われた研究や予備調査における値は、状況が変わっている可能性があるから適当とはいえない。仮にそれが十分信頼できるのなら、そもそも新たな実験や検定を行う必要もないことになる。さらに、当該研究における 2 標本値の差を ES とするのは、本来理論的に無関係であるべき標本値差と母集団間の距離を混同している点で問題がある。母集団間距離が大きいときでも、実際の標本値差が小さくなることはありうる。このとき、標本地差が小さければ、それゆえ「同じ」である確率が高まるべきところ、標本値差を ES とすると、差が小さい時には必ず検出力が小さくなって「同じ」と言いにくくなる、というおかしな事態が生じる。これは結局、あるものを評価するのに、それ自身を評価基準に含めたことから来る混乱なのである。ES として標本値差を取るようでは、もはや使えるものは何でも使うといわんばかりであろう。

power analysis の変法として、先に述べた a posteriori analysis のように n を他の 4 つのパラメータから決めたり「検出可能な ES」（？）を他の 4 要素から求める計算も解説されている。たとえば後者の場合、$\alpha=0.05$ で $1-\beta=0.95$、標本数 30 ならどれくらいの効果（ES）が測定できるのか、という形になる。しかしこのようにして決定された ES とは何なのか。ES の概念そのものに疑問がある以上、形を変えてみても意味のある検討にはならないと思われる。

５．Type II error と Precautionary principle（予防原理）

一般に統計学の教科書には、Type I エラーを減らすために危険率を下げると Type II エラーを増すことになるので、それが不適当な場合、α を通常の 0.05 ではなく、0.10 などで検定することも考慮すべきであると書いてある（石居 1975 など）。たとえば、ある種の薬品の毒性を調べるために薬品投与群と非投与群を設け、「２群の差なし（＝毒性なし）」の帰無仮説を立てたとする。ここで危険率をきびしく（小さく）取ると「差あり（＝毒性あり）」と判定したときの信頼度は上がる。しかし有意差が出ずに「毒性なし（あると言えない）」と判定する確率も高まり、それにもとづいて薬品を製品化すると被害が出る可能性が大きくなる。このように、特に社会的影響のある研究や試験においては、「影響があるにもかかわらず、なしとしてしまう危険」つまり Type II エラーに配慮し、α の水準をゆるめたり（0.05 →0.10 など）、標本数を増やすなどで対応するべきだというのである。統計検定が重視される傾向の中、特に環境問題など社会的側面を持つ分野で、有意差が出なかったときにどのような判断をするかということが問題になってきた。汚染物質の影響が証明できなくても（統計的に有意でなくても）排出削減などの対策が取られるべきであるという主張を Precautionary principle（予防原理）といい、これについて議論が行われている（Gray 1990, Josefson 1990, Johnston & Simons 1990）。その中で、ただ「証明できなくてもその危険がある」と定性的に主張するのではなく、Type II エラーを具体的に決定することによって、影響なし（あるといえない）と判定したときの信頼度を定量化できるという主張がなされるようになった（Peterman & M'Gonigle 1992, Buhl-Mortensen 1996）。つまり、「毒性なし」の帰無仮説を棄却できない場合、もしも power が 0.80 とか 0.95 以上あるならば暫定的に毒性なしとする、ということで Precautionary principle を科学的に裏づけられるとする。1990 年代以降の Type II エラーをめぐる議論は、もっぱら Impact study（環境影響評価）における precautionary principle をめぐって行われており、（Fairweather 1991, Carey & Keough 2002, Underwood & Chapman 2003）、基礎研究分野では、議論も Type II error の計算もほとんど行われないという状況が続いている。もっぱら環境問題において検出力が議論されている背景としては、Type II error が時に深刻な社会的影響を引き起こすことのほかに、応用分野では検定の使用は実用を重視して、多少の理論的不備は甘受するという傾向があるからだろう。従って議論では power は計算可能という前提に立ち、理論の根幹に切り込むようなアプローチはみられない。Mapstone（1995）などは ES を定量化することをあきらめ、それに一定値を与えた上、Type I と Type II error の相対関係を知ることで、環境影響評価の decision making に役立てることができるとしている。裏技というか、苦肉の策というべきだが、ES の本質論をめぐって理論的な解決があるわけではない。

６．Power analysis をどう考えるか

検定結果が「非有意」と出たとき、それゆえ「差なし」「同じ」とみなしてよいかどうかは重要な問題であり、power analysis の理論的重要性は疑いない。ではなぜそれはほとんど行われていないのか。言われているのは次のようなことである。

i）研究者の無知。つまり Type II エラーの重要性に気づいておらず、計算方法を知らない（Peterman 1990）。一方、Type I エラーは効果を誤って特定する危険（うそをつく危険, false positive）であり、Type II エラーは効果を見逃す危険（無視する危険）であるから、特に基礎科学分野では研究者は心理的に Type I エラーのほうを重視するのだという見方もある（Schrader-Frechette & McCoy 1992, Buhl-Mortensen 1996）。

ii）実用上の困難。生態学研究において実際に検出力を計算した例では、十分な検出力を得るためには標本数や実験におけるレプリケート数が多くなりすぎてほとんど実行不能、という主張がなされている（Doherty & Sale 1985, Young 1988）。私も自分の行った研究に関連し、異なる二人の調査者の発見効率に差があるかについて Zar（1999）のテキストよって検出力を計算したところ、数百コドラートが必要という結果になって驚いた。

iii）ES 概念のあいまいさ。本稿で述べてきたことであり、私はこれが最大の要因であると考える。検出力を計算するには ES を特定しなければならず、研究者はここでそれをどう設定すべきかの判断を迫られる。Rottemberry & Wiens (1985) は power analysis が広まらない理由として、研究者の側に、よくわからない ES を仮定しなければならないことへの抵抗感があると指摘し、Young（1988）も ES をめぐり、power analysis はその哲学的妥当性に疑問があるとしている。power analysis が要求する無理な仮定に対し、ここまではついて行けないと感ずる研究者は多いのではないだろうか。

power analysis は理論的に問題があると考える研究者は、それでは検定において「有意差なし」という結果が出た場合、どのように対応すべきだろうか。無難なのは教科書類にあるように「差があるとはいえない」と表現することである。しかしこれは「何も言えない」と言うのと同じであり、せっかく実験して検定したのになんとも残念な結果といえる。では得られた結果から、「どのくらい同じらしいか」という、程度を測定することはできないだろうか。わかりやすいのは標本から算出された P 値をつかうことで、たとえば P=0.03 より P=0.08 のほうが、「より同じらしい」と考える。しかしこれは理論的に妥当でない。危険率 α（およびそれによって判定される、算出された P 値）は、４つのパラメータ（β, ES, n, s）の関数であって、β とのみ連動しているわけではない。たとえば標本数 n によっても影響されるので、n=10 での P=0.08 と n=100 での P=0.08 では重みがちがう。場合によっては P=0.08 での検出力が P=0.10 でのそれを上回ることも考えられるので、P=0.10 の場合のほうが P=0.08 よりも「同じらしさ（＝同じと判定したときの信頼度）」が大きいとは、一概に言えないのである。

私は「非有意」の結果を生かすには、次の２つの方法があると考えている。一つは「非有意」を以って、暫定的に「同じである」と主張することである。これはやってはいけないとされているが、反証主義的立場に立てば、必ずしも誤りとは言えない。差が有意でなかったということは、H_0が偽であるという反証に失敗したということである。何らかのテストにかけて生き残った仮説は、反証されるまでは真実の可能性あるものとして保持される、というのが反証主義の立場であるから、その意味において研究者は「差なし」と主張する権利を有する。ただし反証主義は「理論」の発展についての説明であり、検定のような「事実関係」（特定の条件における差など）に適用すると混乱を招くという指摘もある（ポパー 1978）。そこでここでは反証主義「的」と表現しておく。なお注意すべきことは、この時の「差なし」は、H_0が棄却された場合の「差あり」と同じレベルの信頼度を有していない（Toft & Shea 1983）ということである。否定されるまで暫定的に保持される仮説としての「差なし」なのであって、同じレベルを要求するなら検出力を定量化しなければならない。もちろん、標本数が少ないなど、検出力が小さいと見られる場合は「弱い仮説」となる。なお、しばしば有意差ありを「異なる」、有意差なしを「同じ」と断定し、同列に論じる例があるが不適切であり、表現や論証方法に配慮が必要だろう。

もう一つは、同じ枠組みで行った調査や実験において、「非有意」の結果が「有意」と並列されている場合である。たとえば、ある湾で湾外と湾内の水温を調べ、夏と冬で比較した。夏と冬で地点数、位置その他の方法が同じで、しかも冬の場合だけ内外で有意差が出た場合、夏より冬のほうが湾内外の水温差が顕著である、と結論することは妥当であると考える。あるいはある種のオスとメス同数の個体に同じ枠組みで実験処理を施し、オスでは有意差が出たが、メスでは出なかったという場合、この処理に対してメスよりオスのほうが sensitive であったと言ってよいだろう。このことの理論的根拠は、次のように考えられる。標本数等、実験の枠組みがそろっていることは、αに影響するβ以外のパラメータの相等性を示唆し、2つの検定の検出力が同等であることを推測させる。このことは、「同じ」と判定したときの Type II error の絶対値が小さくなることを意味するわけではないが、算出されたＰ値（Type I error）が、Type II error と連動する可能性を高め、2つの検定におけるＰ値の差をもって、「同じ」と判定したときの危険率（Type II error）の「差」に読み替えることができるであろう。このことから、次のようなデータ処理が考えられる。Ａ地点における n=100 の検定において有意差が出て、Ｂ地点の n=50 の検定では有意差が出なかったとき、このままではＢ点でも n=100 にすれば有意差が出るのではないかという疑いを払拭できない。ならばＢ点のサンプルをあと 50 増やせばよいわけだが、逆にＡ点のほうを、ランダムサンプリングによって 50 に減らしてもよい理屈である。それでも有意差が出たら、「Ａ点有意差あり」vs「Ｂ点有意差なし」の、その違いの部分において、有効な議論が成立するだろう。

power analysis は、理論の基本的な部分において問題をかかえており、全体的に見て信頼できる手法ではないと私は考えている。ただ、統計理論にはもともとあ

る程度のあいまいさがつきものであり、どの程度それを受け入れるかは使用者の判断にかかっている。問題があるとはいえ、対立仮説が明快であるとか、特定の理論を反証するような場合には使用が妥当と思われるケースもあるので、どのような場合にどのような問題が含まれるかについて、今後整理が必要だろう。

引用文献

Andrew NL, Mapstone BD 1987　Sampling and the description of spatial pattern in marine ecology.　Oceanogr. Mar. Biol. Ann. Rev., 25, 39-90

Buhl-Mortensen L 1996　Type-II statistical errors in environmental science and the precautionary principle.　Mar. Poll. Bull., 32, 528-531

Carey JM & Keough MJ 2002　The variabiligy of estimates of variance, and its effect on power analysis in monitoring design.　Envir. Monitoir. Assess., 74, 225-241

Cohen H 1977　Statistical power analysis for the behavioral scientists. 2nd ed. Academic Press

Doherty PJ, Sale PF 1985　Predation on juvenile coral reef fishes: an exclusion experiment.　Coral Reefs, 4, 225-234

Fairweather PG 1991　Statistical power and design requirements for environmental monitoring.　Aust. J. Mar. Freshw. Res., 42, 555-67

Gray JS 1990　Statistics and the precautionary principle.　Mar. Poll. Bull., 21, 174-176

Johnston P, Simons M 1990　Precautionary principle.　Mar. Poll. Bull., 21, 402

Josefson AB 1990　(Letter)　Mar. Poll. Bull., 21, 598

Mapstone BD 1995　Scalable decision rules for environmental impact studies: effect size, Type I, and Type II errors.　Ecol. Appl., 5, 401-410

Peterman R M 1990　Statistical power analysis can improve fisheries research and management.　Can. J. Fish. Aquat. Sci., 47, 2-15

Peterman RM, M'Gonigle M 1992　Statistical power analysis and the precautionary principle.　Mar. Poll. Bull., 24, 231-234

Rotemberry JT, Wiens JA 1985　Statistical power analysis and community-wide patterns.　Am. Nat., 124, 164-168

Schrader-Frechette KS, McCoy ED 1992　Statistics, costs and rationality in ecological inference.　Trend..Ecol. Evol., 7, 96-99

Sokal RR, Rohlf FJ 1981　Biometry, 2nd ed.　WH Freeman & Co.

Toft CA, Shea PJ 1983　Detecting community-wide patterns: estimating power strengthens statistical inference.　Am. Nat., 122, 618-625

Underwood AJ 1997　Experiments in Ecology.　Cambr. Univ.. Press

Underwood AJ & Chapman MG 2003　Power, precaution, Type II error and sampling design in assessment of environmental impacts.　J. Exp. Mar. Biol. Ecol., 296, 49-70

Young CM 1988　Larval predation by barnacles: effects on patch colonization in a shallow subtidal community.　Ecology, 69, 624-634

Zar JH 1999　Biostatistical Analysis 4th ed.　Prentice Hall

ホーエル 1981　初等統計学（第 4 版）浅井晃・村上正康訳　培風館（原著 1976）

石居進 1975　生物統計学入門　培風館

ポパー 1978　果てしなき探求－知的自伝　森博訳　岩波書店（原著 1976）