生理心理学と精神生理学 22 (3): 275-290, 2004

評論

心理生理学データの分散分析

広島大学総合科学部　入戸野　宏

Analysis of Variance of Psychophysiological Data

Hiroshi NITTONO
Faculty of Integrated Arts and Sciences, Hiroshima University
1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521, Japan

2004.3.1 受稿，2004.5.31 受理

Abstract

Performing analyses of variance on data obtained from experimental designs with repeated measures is a common practice in psychophysiological research. However, investigators often find difficulty in selecting appropriate statistical methods. The present article describes statistically sound and easily performed procedures for conducting repeated-measures analyses of variance of psychophysiological data. Topics include comparison between univariate and multivariate analyses of variance, selection of error terms, tests for simple effects and interaction contrasts, multiple comparison, and effect size. A typical sequence of statistical tests is illustrated using a numerical example. (*Japanese Journal of Physiological Psychology and Psychophysiology*, 22(3): 275-290, 2004.)

Key words: statistical methodology, repeated measures design, *F* test, tutorial

【要 約】反復測定を含む実験計画で得られたデータに分散分析を実施することは，心理生理学の研究で広く行われている．しかし，研究者が適切な統計手法を選ぶことは往々にして難しい．本稿は，心理生理学データに反復測定の分散分析を行うときの，統計学的に妥当で，容易に実施できる手続きについて述べたものである．取り上げたのは，単変量・多変量分散分析の比較，誤差項の選択，単純効果と交互作用対比の検定，多重比較，効果量といった話題である．典型的な統計検定の流れを数値例を用いて解説する．

１．はじめに

心理生理学では，同じ実験参加者から多数のデータを同時的・継時的に記録することが多い．そのため，個々のデータは互いに関連しあい，初等統計学で想定するような単純な構造をしていない．海外の雑誌に論文を投稿すると，統計手法に関してクレームがつくことがよくある．しかし，統計学の専門書を調べても，数式が多く抽象的で，自分のデータにどう適用したらよいか分からないことが多い．

本稿では，心理生理学の研究でよく用いられる分散分析――具体的には，反復測定を含む分散分析とその下位検定，多重比較，効果量――について具体的に解説する．反復測定（repeated measures: 反復測度，繰り返し測定，繰り返し測度ともいう）を含むデータの分析は，統計学でもいまだ発展途上の分野であり，手法の改良と比較が続けられている（最近の総説として Keselman, Algina, & Kowalchuk, 2001）．筆者は統計学の専門家ではないので，学問的に厳密で十全な議論はできない．本稿で紹介するのは，現在の心理生理学において許容される，できるだけ簡単な手続きである．

反復測定データの分散分析とそれに関連した問題は，心理生理学に限ったものではない．しかし，比較的少ない実験参加者から多くのデータを収集する心理生理学では，分散分析を不適切に使用することで間違った結論が偶然得られる可能性が高く，その対処法について古くから議論されてきた（Wilson, 1967, 1974; Jennings & Wood, 1976; Keselman & Rogan, 1980）．統計検定法は最新の知見に基づき研究者の判断で慎重に適用すべきであるという主張は，この分野の主要学会誌である *Psychophysiology* の編集方針（Jennings, 1987; Miller, 2000）にも示されている．分散分析について正しく理解しておくことは，心理生理学の研究を発表する上で不可欠な要件である．

心理生理学の中でも事象関連電位の統計分析については，すでに投石(1997)による解説がある．本稿では，そこで十分に説明されていない事項（多変量分散分析，交互作用対比，多重比較，効果量）についても取り上げた．統計量の算出は，SAS（SAS Institute Japan Ltd.）や SPSS（SPSS Japan Inc.）などの統計解析ソフトで行うことを前提とした．そのため，実際の計算に用いる数式は示さなかった．また，分散分析を含むパラメトリック検定を行うときの前提条件として，(a) 無作為なサンプル抽出，(b) 母集団の分布の正規性，(c) 分散の等質性，がある．これらは成書に詳しいので，本稿では述べなかった（森・吉田, 1990; 橘, 1986）．

2. 反復測定データの分析

2-1. 定義と特徴

反復測定を含む実験計画とは，同じ参加者がすべての水準にかかわる参加者内変数（within-participants or within-subjects variable）を 1 つ以上含んだ実験計画である．2 つの条件の平均値の差を調べる *t* 検定では，対応のある（paired）場合に相当する．各水準に異なる参加者を割り当てる参加者間変数（between-participants or between-subjects variable）を含むことも含まないこともある．反復測定したデータは，水準間に対応があり，ある水準のデータが他の水準のデータと相関している．これに対して，反復測定を含まない実験計画（完全無作為計画）では，各水準のデータは独立しており，互いに相関はない．

水準間に相関があるデータに分散分析を行うと，タイプIエラー率（本当は差がないのに“差がある”と誤って判定する確率）が，設定した有意水準よりも大きくなる場合がある．有意水準を.05 に設定するとは，“差がある”と判定したときにそれが間違っている可能性が 5 %（20 回に 1 回）を超えることはないと宣言することである．さまざまな統計検定法は，この規則が成り立つように作られている．しかし，検定を行うときの前提条件が崩れるとタイプIエラー率が増えてしまうことがある．[1]

反復測定を含む分散分析では，ある要因の効果に各水準が等しく貢献していることを前提にしている．水準数が 2 のときは自動的にバランスがとれるが，3 以上になるとバランスがとりにくくなる．このバランスを球面性（sphericity）という．これは，対応のあるすべての水準対の“差”の分散が等しいことだとされている（Huynh & Feldt, 1970; 小牧, 1995）．通常の分散分析は，データに球面性があるという前提で p 値を計算する．しかし，このような球面性の仮定（sphericity assumption）が破られていると p 値は小さくなりすぎ，本当は差がないのに“差がある”と誤判定する確率が増えてしまう．

実際のデータで球面性が成り立つかどうかは球面性の検定で調べられる．しかし，このような検定をしなくても，心理生理学のデータでは球面性が成り立たないと考えた方がよい．たとえば，事象関連電位の P3（P300）を，前頭部（Fz），中心部（Cz），頭頂部（Pz）の 3 部位から記録し，その振幅を測定したとする．球面性が成り立つためには，Fz と Cz の振幅値の差の分散と，Cz と Pz の振幅値の差の分散，そして Fz と Pz の振幅値の差の分散がすべて等しいことが必要である．しかし，このようなケースはほとんどありえない．距離が離れると，ふつうは振幅差が大きくなり，分散も大きくなる．そのため，球面性が成り立たない可能性を考慮して分析するのが妥当である．

反復測定の要因が増えると，球面性の仮定の数も増える．要因数を T とすると 2^T-1 で，交互作用を検定するための仮定が増える．このことから，一般に，2 つ以上の反復測定要因を含む分散分析では，球面性の仮定が成り立たないと考えてよい（O'Brien & Kaiser, 1985）．

そこで，反復測定データの分析では，(a) ふつうの（単変量 univariate の）分散分析で得られた F 値を自由度を調整して検定することで，球面性が成り立たないときの歪みを補正する自由度調整法，(b) 球面性の仮定を必要としない多変量分散分析（multivariate analysis of variance: MANOVA，正確には general MANOVA: GMANOVA という）のいずれかを使用する（Jennings, 1987; Vasey & Thayer, 1987）．

2-2. 自由度調整法

反復測定を含む分散分析で，球面性の仮定が成り立たないときのタイプ I エラー率の増大を，自由度を小さく調整することで抑えようとする方法である．球面性からの逸脱の程度を表わす係数 ε（イプシロン）を，要因の効果の自由度と誤差の自由度の両方に乗算して自由度を調整する．水準数が 2（要因の効果の自由度が 1）のときは球面性が自動的に成り立つので，修正は行わない．

ε は，球面性が成り立つときの最大値 1 から，成り立たないときの最小値 1 / (水準数－1)までの値をとる．まず，ε の最大値を使って，自由度を調整せずに F 検定を行い，そこで有意にならなければ“差がない”と結論できる．次に，ε の最小値を使って，すべての検定で自由度を(1, サンプル数－

1)として F 検定を行い, 有意になれば“差がある”と結論できる（これを Geisser & Greenhouse [1958] の保守的検定 conservative test という）．自由度を調整しない検定で有意になり，保守的検定で有意にならないときは，実際のデータの分散-共分散行列から ε の推定値を算出し，その値を使って自由度調整した F 検定で結論を出す．この方法は Greenhouse & Geisser (1959)の 3 段階法と呼ばれ，以前はよく使われた．しかし，統計解析ソフトを使えば，実際のデータから ε の推定値をすばやく計算できるので，現在は第 3 段階だけを行うことが多い．

データから ε の推定値を算出するには，Greenhouse-Geisser（グリーンハウス-ガイサー）と Huynh-Feldt（フィン-フェルト）の方法がよく用いられる．Greenhouse-Geisser の ε は，ε が小さいとき（0.5 程度かそれ以下）には適しているが，サンプル数が少なく ε が 1 に近い（0.75 程度かそれ以上）ときには自由度を小さく調整しすぎる（有意性の判定が厳しくなりすぎる）というバイアスをもっている（Huynh & Feldt, 1976）．Huynh-Feldt の ε は，Greenhouse-Geisser の ε がもつこのようなバイアスを，サンプル数と水準数を使って修正したもので，Greenhouse-Geisser の ε よりも大きめに計算される（その結果，有意性の判定が甘くなる）．Huynh-Feldt の ε は 1 を超えることがあるが，その場合は自由度調整を行わない．

どちらの方法が望ましいという明確な指針はないが，サンプル数が十分に大きいときは Greenhouse-Geisser の ε を，サンプル数が小さいとき（10 名程度）は Huynh-Feldt の ε を使うことが多い．一つの分析の中で ε の値によって両者を使い分けることはほとんどないので，どちらを使うかはあらかじめ決めておく．なお，サンプル数が大きくなると，両者の ε の差は小さくなる．

Jennings & Wood (1976)やそれを引用した投石(1997)は，反復測定要因間の交互作用の ε_{AB} は，それぞれの要因の ε の積 $\varepsilon_A \cdot \varepsilon_B$ であると述べている．この計算式の出典はMcHugh, Sivanich, & Geisser (1961)である．確かに，ε の最小値を使うと，$\varepsilon_A \cdot \varepsilon_B$ は[1 / (要因Aの水準数－1)]×[1 / (要因Bの水準数－1)]となり, 交互作用の ε_{AB} の最小値 1 / [(要因Aの水準数－1)×(要因Bの水準数－1)]と一致する．しかし，この式で求めた ε は，交互作用のための分散-共分散行列をデータから新しく計算して求めた ε とは一致しない（小さくなることも大きくなることもある）．統計解析ソフトが利用できるときは, 実際のデータから計算した値を使うとよい. 利用できずに簡易式を使うときは, Jennings & Wood (1976)やMcHugh et al. (1961)を引用するとよいだろう．

自由度調整法を用いるときは,“ Huynh-Feldt の ε による補正を自由度が 1 より大きい反復測定の F 値の検定に用いた（The Huynh-Feldt ε correction was used to evaluate F ratios for repeated measures involving more than one degree of freedom）”などと方法に記載し，結果では，F 値と調整前の自由度，自由度調整後に得られた p 値，ε をこの順序で記載する．

2-3. MANOVA

統計解析ソフトで反復測定を含んだ分析を行うと, 上記の自由度調整法の結果とともに, MANOVA の結果も出力される．SAS や SPSS では，4 種類の統計量（Wilks' lamda［ウィルクスのラムダ］, Pillai's trace［ピライのトレース］, Hotelling-Lawley trace［ホテリング-ローリーのトレース］, Roy's Greatest Root［ロイの最大根］）が算出される．これらの値は，ある自由度を持った F 値に

近似的に変換できるので，それを使って検定を行う．計算される F 値と自由度が 4 種類の統計量で一致するとは限らない．

心理学の論文では，Wilks のラムダ（とその近似 F 値を求める Rao の方法）をよく見かける．しかし，Olson (1976)は，サンプル数が比較的小さいときは，仮定からの逸脱に頑健で検出力も高い Pillai のトレース（Pillai-Bartlett trace ともいう，記号 V）を使うことを薦めている．心理生理学の分野でも，この統計量が推奨されている（Keselman, 1998）．MANOVA を行うときは，“A が B に及ぼす効果は Pillai のトレースを用いた多変量分散分析で検討した（The effect of A on B was examined with a multivariate analysis of variance using Pillai's trace statistic）”などと方法に記載し，結果では F 検定の結果（F 値と自由度，p 値）を記載する．ε がないだけで，見た目は自由度調整法とほとんど変わらない．

2-4. 自由度調整法と MANOVA の比較

統計学的には，球面性を仮定しない MANOVA の方が優れている．MANOVA の短所は，サンプル数が水準数より少ないと（自由度より 1 以上大きくないと）計算できないことである．たとえば，条件(3 水準)×左右半球(2 水準)×部位(8 水準)の反復測定 3 要因の分析では，最低でも 15 名のデータがないと 3 要因交互作用の値を計算できない（自由度$(3-1)\times(2-1)\times(8-1) = 14$ よりも 1 以上大きければ計算できる）．また，計算できたとしても，サンプル数が水準数よりも十分に大きくないと，本当は存在する差を見逃すタイプ II エラー率が高くなる．

Vasey & Thayer (1987)は，Davidson (1972)の計算結果に基づいて，サンプル数が水準数より 20 以上大きいときは MANOVA を，サンプル数が水準数より 6 程度しか大きくないときは自由度調整法を用いるのがよいと述べている．心理生理学の実験では，水準数に比べて参加者数が少ない（10–20 名）ことが多いので，自由度調整法が今でもよく使われている．

しかし，このような一般的な検出力とは別に，Davidson (1972)は，単変量の分散分析では検出できないが，MANOVAでは検出できる差があることを指摘している．Table 1 にその例を示した．3 水準の平均値は類似しており，単変量の分散分析では$F(2, 18) = 0.43$ となる．F値が 1 より小さいので，自由度調整に関係なく有意にはならない．しかし，個人データを示したFig. 1 を見ると，水準X_1とX_2には小さいがほぼ一貫した差が認められる．このデータにMANOVAを適用すると，$F(2, 8) = 7.84$, $p = .0130$（Pillaiのトレース）となり，水準間の差を検出できる．

一般に，3 水準以上の反復測定の単変量分散分析では，他との相関が低く分散が大きい水準があると，誤差とみなされる変動が大きくなるので，小さな差を見逃しやすくなる．Davidson (1972)は，自由度調整法は MANOVA よりもやや優れることもあるが，ずっと悪くなることもあると述べ，サンプル数が水準数よりわずかしか大きくない場合でも MANOVA を行うことを薦めている．また，MANOVA が行えるように水準数を減らすことも提案している．

実践的なアドバイスとしては，両方の結果が出力されたらどちらも眺めて比較するとよい．自由度調整法で有意差が得られたのに，MANOVA で有意差がなければ，サンプル数が少ないことによる検出力不足と考えられる．反対に，MANOVA では有意なのに自由度調整法では有意でないときは，上述のような単変量の分散分析では検出できないデータの構造があるのかもしれない．統計学的には，

同じデータを 2 つの方法で分析し，都合のよい方を採用することは“ 検定の多重性（multiplicity）”として避けるべきである．しかし，実際の方便としては，結果として示したいことが伝わりやすい方法を一貫して使えばよいだろう. 少なくとも, 今後 MANOVA に慣れていくことは有益と考えられる．

2-5. 参加者間要因を含んだ実験計画の注意点

実験計画に参加者間要因が含まれるときは，これまで述べてきたことに加えて，多標本球面性（multisample sphericity）という新しい仮定が加わる．簡単にいえば，すべての群間で分散-共分散行列が等しいということである. 自由度調整法も MANOVA もこの前提に基づいて p 値を計算する．しかし，この前提が成り立たないときは，p 値が信頼できなくなる．

詳しい説明は千野(1995)に譲り，ここでは Keselman (1998)に基づいて，心理生理学のデータ分析にかかわる実際の注意点を述べる．結論をいえば，多標本球面性の仮定が成り立たなくても，サンプル数が群間で等しいときはその影響が小さい（計算された p 値を信頼できる）．しかし，サンプル数が群間で異なると影響が大きくなる（p 値を信頼できなくなる）．後者を検定するときは，プールしない（検定に直接関連したデータから求めた）誤差項を用いて自由度調整をした Welch-James 法を用いる．この指摘は，心理生理学の主要学会誌である *Psychophysiology* の現在の投稿規程（2004 年 41 巻）にも反映されており，看過できない．しかし，簡単なケース（等分散が仮定できない 2 群の平均値の差を検定する Welch の方法に相当する）を除くと，この方法は統計解析ソフトに精通していなければ使えない. 一般の研究者は, 実験を行う段階で“ 群間比較するときは参加者数をそろえる ”という方針を厳守するのがよいだろう．

3. 交互作用の下位検定

3-1. 分析の方針

2 要因以上の実験計画で交互作用が有意になったときは，2 つの方法で下位検定（下位効果検定 sub-effect tests）を行う．単純効果（simple effects）の検定と交互作用対比（interaction contrasts）の検定である．これらの計算には，大きく分けて 2 つの方法がある．最初に行った全体の分析の計算結果を利用する方法（プールした誤差項 pooled error term を用いる方法）と，下位検定に直接関連したデータだけで新たに分析を行う方法（水準別誤差項 separate error term を用いる方法）である．

前者は，“ できるだけ多くのサンプルから推測した方が精度が高くなる ”という統計学の原則に従っている．また，誤差の自由度が大きいので，後者より検出力が高くなることが多い．そのため，小牧(1995)や森・吉田 (1990)は，こちらの方法を薦めている．しかし，分析対象ではない（もしかすると異質かもしれない）データを一緒にして分析することが合理的かどうかについては，統計学者の間でも意見が分かれている．また，ケースごとに分析の手続きが異なり，誤差項や自由度を手計算で調整しなければならないこともある．

そこで，本稿では，宮本・山際・田中 (1991)や投石(1997)，Vasey & Thayer (1987)に従い，後者の方法を薦める．その理由として，分析方針がすべてのケースで一貫しており単純明快であること，統計解析ソフトのデフォルト設定で結果が得られること，下位検定で球面性が成り立っているかどうかを気にしないでよい（新たに ε が計算される）ことが挙げられる．後者の方法を用いるときは，“ 単

純効果の検定にはプールしない水準別誤差項を用いた（Nonpooled separate error terms were used for simple effect tests）”などと方法に記載する．

3-2. 単純効果の検定

分散分析における交互作用は，ある要因の効果が他の要因によって変わるときに生じる．交互作用が得られたときは，特定の要因の水準ごとに個別に分析を行うと，結果の見通しがよくなる．

2 要因の交互作用（two-way interaction または first-order interaction）が有意であったときは，一方の要因における各水準で他方の要因についての 1 要因分散分析を行う．これを単純主効果（simple main effect）の検定という．3 要因の交互作用（three-way interaction または second-order interaction）が有意であったときは，1 つの要因の水準ごとに他の 2 つの要因について 2 要因分散分析を行い，交互作用の有無を調べる．これを単純交互作用（simple interaction）の検定という．

一般に，p 要因の交互作用が得られたときは，そこに含まれる要因を 1 つ選んで，その水準ごとに(p－1)要因の分散分析を行う．このとき，直接関連するデータだけで分析する（水準別誤差項を用いる）という方針に従い，新しいデータセットを作って新規に分散分析を行えばよい．交互作用に含まれない要因については，全水準の平均値を使って分析する．最後は 1 要因の分析になるが，そこで有意差が得られたときは，水準数が 3 以上であれば，4 節で述べる平均値の多重比較を行う．

具体例で説明しよう．事象関連電位の N1 振幅について，刺激(低頻度 vs. 高頻度)×条件(注意 vs. 無視)×部位 (Fz, Cz, Pz) の反復測定 3 要因の分析を行ったとする．3 要因の交互作用が有意であれば，どれか 1 つの要因を選んで，その水準ごとに 2 要因分散分析を行う．要因の選択は，仮説やデータの構造（まとまり）を考慮して行う．この例では，N1 が刺激誘発性の電位であることを考慮して“刺激”を選ぶこともできるし，測定状況が異なることを考慮して“条件”を選ぶこともできる．また，両方とも行ってもよい．

刺激を選んだときは，各水準（低頻度と高頻度）について条件×部位の 2 要因の分散分析を行う．この分析で交互作用が得られたときは，さらに条件ごとに部位の効果（または，部位ごとに条件の効果）を調べる 1 要因の分析を行う．また，3 要因の交互作用が有意でなく，2 要因（たとえば刺激×条件）の交互作用が得られたときは，交互作用に含まれない“部位”の全水準を平均した値を使って，刺激ごとに条件の効果（または，条件ごとに刺激の効果）を調べる 1 要因の分析を行う．

原理はこのように単純だが，実際の分析でときどき気になるケースについて補足する．3 要因の交互作用があるだろうと思って分析したが有意差が得られず，2 要因の交互作用がいくつか有意になったとする．上の例では，刺激×条件×部位の交互作用が有意でなく，刺激×条件，刺激×部位，条件×部位の交互作用の一部または全部が有意になったとしよう．問題は，このときに刺激ごとに 2 要因の条件×部位の分析を行ってよいかである．下位検定の手順としてはこれは不適切だが，そうするのが妥当だという理由があれば行ってもよい．このときは，“まず，すべての要因を含んだ包括的分析を行い，その後，刺激ごとに条件×部位の分析を行った”などと方法に記載し，そのように分析することをあらかじめ明記しておく．ただし，交互作用が有意でないのに別々に分析し，ある要因の効果が一方で有意になり，他方で有意でなかったとしても，要因の効果が両者で異なると積極的に主張することはできない（Picton et al., 2000, p. 147）．

下位検定を行うときの有意水準は，全体のタイプ I エラー率の増大を抑えるために，同時に行う下位検定の数で割って小さくするのが正しい（Keselman, 1998）．n 回の下位検定を全体の有意水準を α にして行うならば，1 回あたりの有意水準を α / n にする（4-3 節で述べる Bonferroni [ボンフェローニ]の方法）．しかし，有意水準を調整せずに単純効果の有意性を判定することも慣例的に行われている（森・吉田, 1990; 投石, 1997; 橘, 1997）．原則を理解した上で，どちらかの方法を一貫して用いればよいだろう．

3-3. 交互作用対比の検定

単純効果の検定は，厳密にいうと，交互作用の効果を直接調べるものではない．水準を固定して単純効果の検定をすると，交互作用だけでなく主効果も含んで分析することになるからである．全体の分析での主効果が大きいと，単純主効果がすべて有意になることもある．その場合は，交互作用についての説明ができなくなる．

交互作用の効果を直接調べるには, 交互作用対比の検定を行う（Keselman, 1998; Lix & Keselman, 1996）．これは，一方の要因における 2 水準の平均値の差を，他方の要因の 2 水準間で比較することだと考えればよい．水準数が p のときに組み合わせられる水準対の総数は $p(p-1) / 2$ なので，2 要因 $p \times q$ の交互作用では，交互作用対比の数は$[p(p-1) / 2] \times [q(q-1) / 2]$になる．たとえば，ある事象関連電位の振幅について，条件(1，2，3)×部位(Fz, Cz, Pz)の反復測定 2 要因の分析を行ったところ，交互作用が得られたとする．この例では，3 水準×3 水準なので，$[3(3-1) / 2] \times [3(3-1) / 2] = 9$ の交互作用対比ができる．それぞれの対比について，それぞれに関連するデータだけで両側 t 検定を行う．複雑そうにみえるが，Fz と Cz の振幅差を条件 1 と 2 の間で比較するといった検定を 9 回行えばよい（6 節の計算例も参照）．

これは一種の多重比較であり，検定の繰り返しによるタイプ I エラー率の増大を避けるために，4 節で述べる Bonferroni の方法かその改良版で有意性の判定を行う．Bonferroni の方法では，9 回の比較を同時に行うと考えて，全体の有意水準が.05 のときは，比較あたりの有意水準を.05 / 9 = .0056 として検定する．

有意差が得られた対比が，(条件 1 の Fz－Cz) vs. (条件 2 の Fz－Cz) と (条件 1 の Cz－Pz) vs. (条件 2 の Cz－Pz) であれば，この交互作用は，Fz と Cz，Cz と Pz の関係が条件 1 と 2 の間で異なっていたために生じたと説明できる．このような要因の組み合わせ効果を調べることは，単純効果の検定ではできない．どちらの方法を使うのがよいかは仮説によるが，単純効果の検定でうまく説明できないときは，交互作用対比の検定で明快に説明できることもある．

4. 多重比較

4-1. 多重比較の定義

t 検定が 2 つの群（水準）の平均値を比較する方法であるのに対し，多重比較法は複数の群（水準）の平均値を比べるときに使う方法であると簡単に考えられがちだが，実際にはもっと複雑な手法である．多重比較法を適用できる前提条件を正確に示し，分析に先だってそれらが満たされているかどう

かを確認するように促した教科書はほとんどなかった．統計解析ソフトでは，方法を指定すれば，自動的に結果が計算される．そのため，心理生理学の論文にも誤用が多く認められる．

よくある誤りは，互いに独立した群間（参加者間要因）の平均値を比較するために考案された方法を，互いに相関のある水準間（参加者内要因）の平均値の比較にそのまま使ってしまうことである．本稿では，この問題を的確に指摘した永田・吉田(1997)の優れた教科書に従い，多重比較について解説する．

2 水準の平均値を比較するときは，まず帰無仮説（$\mu_1 = \mu_2$）を立てる．そして，データから計算された統計量がこの帰無仮説の下で得られる確率を計算し，それがかなり小さいならば，誤判定をする一定の危険率を含んだ上で，帰無仮説を棄却し，平均値に差があると結論づける．これが統計検定の原理である．

多重比較では，いくつかの帰無仮説（帰無仮説族またはファミリー family of null hypothesesという）について複数回の検定を行う．1 回の検定で間違える確率が 5%あると，2 回検定すればそのどちらかで間違える確率は $1-(1-0.05)^2 = 0.0975$（9.8%），3 回検定すれば $1-(1-0.05)^3 = 0.1426$（14.3%）となる．多重比較法とは，推測の対象とする帰無仮説ファミリーの少なくとも 1 つが誤って棄却されてしまう確率（ファミリーごとのタイプIエラー率 familywise type I error rate）が，公称の値（あらかじめ設定した有意水準）を超えないように，1 回 1 回の検定における棄却限界値を設定した方法である．ファミリーに含まれる帰無仮説の形と数を明確にしなければ，多重比較は行えない．

多重比較は分散分析の後で行うと書いてある教科書もあるが，多重比較と分散分析は別のものである．厳密にいえば，分散分析の後で多重比較を行うのは“検定の多重性”にあたる．しかし，これを問題にすることが少ないのは，検出力は下がるが，タイプ I エラー率は抑えられるので，致命的な誤りとはいえないからである．4-5 節では，最初に行う 1 要因分散分析の有意性を利用して，多重比較の検出力を上げる方法を紹介する．

4-2. 多重比較法の種類と前提条件

多重比較にはいろいろな方法がある（森・吉田, 1990; Seaman, Levin, & Serlin, 1991; 高橋・大橋・芳賀, 1989）．しかし，そのほとんどは互いに独立した群間の平均値を比較するために考案されたものであり，本稿で扱う反復測定データの分析には使えない．たとえば，最もよく使われる Tukey（テューキー）の HSD（honestly significant difference）法は，球面性が成り立たないデータに適用すると，タイプ I エラー率が設定した有意水準を超える可能性があることが指摘されている（Maxwell, 1980）．Tukey の HSD 法では，すべての平均値の対について，共通の基準を用いて有意性を判定する．これは，すべての水準間の平均値の差の分散が等しいと仮定することであり，球面性の仮定と同じである．最初の F 検定で球面性からの逸脱を補正しながら，多重比較で球面性を前提とした方法を用いるのは矛盾する．

反復測定データの多重比較には，Bonferroni の方法かその改良版を用いるのが安全で確実である．Bonferroni の方法は，汎用性が高く，簡便であるにもかかわらず，あまり丁寧に紹介されていない．本稿では，検出力を高めた改良版である Holm（ホルム）と Shaffer（シェイファー）の方法ととも

に詳述する．すべての水準間の平均値を比較するときに便利なように，水準数が 3，4，5 のときに用いる比較あたりの有意水準を Table 2 に示した．

4-3. Bonferroni の方法

正しくは，Bonferroni の不等式に基づく多重比較法であり，Dunn (1961)が紹介したことから，Dunn の方法ともいう．ここでは慣例に従い，Bonferroni の方法と呼ぶことにする．水準数が p のときに，すべての水準間で平均値の対比較を行うと，比較の総数 c は $p(p-1)/2$ になる．これらの比較を，有意水準をα / cとして同時に検定する．こうすれば，全体のタイプ I エラー率がαを超えることはない．3-1 節で述べた水準別誤差項を用いるので，関連するデータだけを使って対応のある t 検定を行う．差の方向に関係なく，差があるかどうかを検定するので，両側検定とする．

このように，すべての帰無仮説を共通の基準で一度に検定する方法をシングルステップ（single-step）法という．これに対して，基準を変えながら帰無仮説を順番に検定していく方法をステップワイズ（stepwise）法という．後者の方が一般に検出力が高い．

4-4. Holm の方法

Holm (1979)は，検定すべき帰無仮説の“数”に注目し，Bonferroni の方法をステップワイズ法に改良することで検出力を高めた．これを逐次棄却型多重比較法（Holm's sequentially rejective multiple test procedure）という．

4 水準の平均値（μ_1, μ_2, μ_3, μ_4）について，すべての水準間の対比較を考える．4(4－1) / 2 で計 6 対の比較を行うことになる．つまり，ファミリーに含まれる帰無仮説の数は 6 である（$\mu_1 = \mu_2$，$\mu_1 = \mu_3$，$\mu_1 = \mu_4$，$\mu_2 = \mu_3$，$\mu_2 = \mu_4$，$\mu_3 = \mu_4$）．それぞれについて対応のある両側t検定を行い，p値を小さい順に並べる．最も小さいp値については，Bonferroniの方法と同じく，$\alpha / 6$ で検定する．ここで有意差がなければ分析は終了する．有意差が得られたときは，帰無仮説の 1 つが棄却されたのだから，残る帰無仮説（間違って棄却してしまう可能性をもった帰無仮説）は 5 になる．そこで，2 番目に小さいp値は$\alpha / 5$ で検定を行えばよい．有意差がなければ分析を終了する．有意差があれば，残る帰無仮説はまた 1 つ減って 4 になる．検定を行うたびに帰無仮説の数はひとつずつ減る．比較の総数がcのとき，p値を小さい順に並べ，有意水準をα / c，$\alpha / (c-1)$，$\alpha / (c-2)$，・・・，αとして順番に検定していき，有意差が得られなくなった時点で終了する．Holmの方法は，Bonferroniの方法の代わりにいつでも使える便利な方法である．

4-5. Shaffer の方法

Shaffer (1986)は，Holm (1979)の方法を改良し，修正版逐次棄却型多重比較法（Shaffer's modified sequentially rejective multiple test procedure）を提案した．これは，検定すべき帰無仮説の“数”と“形”に注目し，さらに検出力を高めた方法で，帰無仮説の間に論理的な関連があるときに有効である．論理的な関連がないときは Holm の方法とまったく同じである．

先ほどと同様に，4 水準の平均値の対比較を考えてみよう．帰無仮説の数は 6 である（$\mu_1 = \mu_2$，$\mu_1 = \mu_3$，$\mu_1 = \mu_4$，$\mu_2 = \mu_3$，$\mu_2 = \mu_4$，$\mu_3 = \mu_4$）．これらは論理的に独立していない．たとえば，$\mu_1 \neq \mu_2$で

あると分かれば，$\mu_1 = \mu_3$と$\mu_2 = \mu_3$が同時に成り立つことはありえない．Shafferの方法では，全体の有意水準を，同時に成り立ちうる帰無仮説の最大数で割って，比較あたりの有意水準を決める．多重比較では, 帰無仮説ファミリーの少なくとも 1 つが“ 間違って ”棄却される確率を統制できればよい．正しく棄却される帰無仮説（つまり，論理的に成り立たないことが明らかな仮説）は，この勘定にいれなくてよいのである．

Holmの方法と同じく, 得られたp値を小さい順に並び替える. まず, 最小のp値を$\alpha / 6$ で検定する．ここで有意差が得られ，$\mu_1 = \mu_2$という帰無仮説が棄却されたとしよう．残る帰無仮説は 5 である（$\mu_1 = \mu_3$，$\mu_1 = \mu_4$，$\mu_2 = \mu_3$，$\mu_2 = \mu_4$，$\mu_3 = \mu_4$）．しかし，これらは論理的に独立していない．すでに行った検定で$\mu_1 \neq \mu_2$であると分かったので，$\mu_1 = \mu_3$と$\mu_2 = \mu_3$，$\mu_1 = \mu_4$と$\mu_2 = \mu_4$がそれぞれ同時に成り立つことはない．したがって，同時に成り立つ帰無仮説の最大数は 3 になる．そこで，2 番目に小さいp値は$\alpha / 3$ で検定する．さらに有意差が得られ，$\mu_3 = \mu_4$が棄却されたとする．残る帰無仮説は 4 つだが（$\mu_1 = \mu_3$，$\mu_1 = \mu_4$，$\mu_2 = \mu_3$，$\mu_2 = \mu_4$），すでに分かっている$\mu_1 \neq \mu_2$，$\mu_3 \neq \mu_4$という情報から，$\mu_1 = \mu_3$と$\mu_1 = \mu_4$，$\mu_1 = \mu_3$と$\mu_2 = \mu_3$，$\mu_1 = \mu_4$と$\mu_2 = \mu_4$，$\mu_2 = \mu_3$と$\mu_2 = \mu_4$がそれぞれ同時に成り立つことはないといえる．したがって，同時に成り立つ帰無仮説の最大数は 2 となる（$\{\mu_1 = \mu_3, \mu_2 = \mu_4\}$または$\{\mu_1 = \mu_4, \mu_2 = \mu_3\}$）．そこで 3 番目に小さい$p$値は$\alpha / 2$ で検定する．

このように，Shaffer の方法では，検定の途中で棄却される帰無仮説の組み合わせによって，同時に成り立つ帰無仮説の最大数が変わる．この手続きのおかげで検出力が高くなっているが，かなり煩雑である．そこで，Holland & Copenhaver (1987)は，すべての水準間で平均値の比較をするときのために“ 同時に成り立つ帰無仮説の最大数 ”の上限値が分かる数表を作成した．この数表は，検定の途中経過にかかわらず利用できる．Table 2 に示した有意水準は，この上限値を用いて算出したものである. ただし, 検出力を最大にするには, 上限値を用いずに帰無仮説を実際に数え上げた方がよい.

Shafferの方法には別解がある．もし 1 要因の分散分析が有意になったときだけ多重比較を行うのであれば，最初の分散分析によって帰無仮説$\{\mu_1 = \mu_2 = \mu_3 = \mu_4\}$が棄却されると考えられる．水準の中で少なくとも一組は平均値が同じでないといえるので, 上記の方法の 2 段階目と同じ論理構造になる．そのため，最小のp値と 2 番目のp値はどちらも同じ基準（4 水準のときは $\alpha / 3$）で検定してよい．それ以降は上記の方法とまったく同じである．Table 2 にはこの考え方に基づく数値を示した.

4-6. その他の方法

Bonferroni の方法を改良した多重比較法は，他にもある．Keselman (1998)は，Hochberg (1988) の方法を紹介している．これは，Holm や Shaffer の方法とは逆に p 値を大きい順に並べ，有意水準をα，$\alpha / 2$，$\alpha / 3$，･･･，α / cとして順番に検定していく．最初に有意差が得られるまでの帰無仮説を保留し（差がないと判定し），有意差が得られた以降のすべての帰無仮説を棄却する（差があると判定する）．この方法の検出力は高いが，全体のタイプ I エラー率を抑えられるという数学的な裏付けが不十分なので，反復測定の分析には使わない方がよいようである（永田・吉田, 1997）.

母集団が独立で分散が等しいと考えられる場合（群間の平均値の比較）には，Tukey の HSD 法を用いてもよい．計算が簡単で，全体のタイプ I エラー率が公称の有意水準以下に抑えられ，検出力もそれなりに高い．しかし，現在では，さらに検出力の高い Peritz（ペリ）の方法も利用できる．Fisher

（フィッシャー）の LSD（least significant difference）法や Newman-Keuls（ニューマン-コイルス）法は，水準数が 4 以上のときは，全体のタイプ I エラー率を統制できないので使ってはならない（永田・吉田, 1997; Seaman et al., 1991）．

5. 効果量

統計検定での有意性とは“差がないとはいいにくい”という判断である．p 値は“差がない”という帰無仮説が誤って棄却される確率を示したものである．サンプル数を増やせば，たいていの現象では p 値が小さくなっていき，帰無仮説が棄却される．p 値は，効果の大きさ（効果量 effect size）を直接表わすものではない．そのため，最近の雑誌の投稿規程では，p 値とともに効果量を示すことが推奨されている（たとえば，American Psychological Association, 2001, pp. 25-26）．

効果量とは，平均値の差の大きさを標準化したものと考えればよい．2 条件の平均値の差を表わすとき（t 検定における効果量）は d という指標を使う．d は，2 条件の平均値の差を 2 条件をプールした標準偏差で割ったものである．3 水準以上の分散分析でも，2 水準を取り出してその間の平均値の差について言及するときは，この d を使うことができる．この場合，どのように標準偏差をプールするかが問題になるが，さまざまなケースについての算出法が，Cortina & Nouri (2000)に示されている．d を計算しないときでも，各水準の平均値とばらつきの指標（標準偏差 standard deviation: SD か標準誤差 standard error of mean: SEM）は示しておきたい．

分散分析における要因の効果の大きさは，関連度（strength of association）で表わすことが多い（Maxwell, Camp, & Arvey, 1981）．これは，ある要因の効果がデータの変動のどれだけの割合を説明できるかを示したものである．関連度を特定のサンプルについて求めたものが，η^2（イータ二乗）である．反復測定の分散分析では，参加者の要因やその他の主効果・交互作用によって説明できる変動を除いてから，この割合を計算する．このときは偏η^2（partial η^2，η_p^2）という．偏η^2は，分散分析で用いたF値と自由度から，(効果Aの自由度×効果AのF値) / (効果Aの自由度×効果AのF値＋効果Aの検定に用いる誤差の自由度) として計算できる．たとえば，$F(2, 30) = 3.75$ の場合，$\eta_p^2 = (2 \times 3.75) / (2 \times 3.75 + 30) = .20$ となる．これは，サンプルに含まれる（他の要因では説明できない）変動の 20%が要因Aによって説明できることを示す．なお，偏η^2の計算と自由度調整のε とは無関係である（調整後に計算しても同じ値になる）．

η^2や偏η^2はサンプルから求めた記述統計量であるが，母集団推定値としてϖ^2（オメガ二乗）やε^2（イプシロン二乗：自由度調整のε とは違う）も提案されている．サンプル数も考慮に入れて計算し，η^2や偏η^2よりも小さな値になる．他の研究データと比較する場合には，母集団推定値を用いる方がよい．算出式と使用法については，Kirk (1995)，Maxwell et al.(1981)が参考になる．

6. 計算例

以上述べてきた方法で，反復測定 2 要因の分析を行った例を示す．Table 3 に，森・吉田(1990)の例題 3・2・5（pp. 116-121）のデータを示す．この程度のサンプル数で分散分析を行うのは一般的でないが，ここでは手順の説明が目的である．メインの分析（自由度調整法と MANOVA）には The SAS system Release 8.2 for Windows を用いた．細かな分析は Microsoft Excel で行った（t 検定に

よる p 値は TTEST 関数で計算できる）．統計検定を行う前に，Fig. 2 のような平均値のグラフを作成し，検定結果を予測しておくとよい．そうすることで，分析手順を間違えたときに，誤りに気づきやすくなる．

6-1. 全体の分析

3 水準以上の反復測定要因 B があるので，球面性が成り立たない可能性を考慮した分析を行う．サンプル数が少ないので Huynh-Feldt の ε で自由度調整した分散分析を行うと，A の主効果 $F(1, 4) = 8.10$, $p = .0466$，B の主効果 $F(3, 12) = 6.04$, $p = .0095$, $\varepsilon = 1.00$，A×B の交互作用 $F(3, 12) = 7.07$, $p = .0054$, $\varepsilon = 1.00$ となる．要因 A は水準数 2 で自動的に球面性が保たれるので自由度調整をしない．また，Huynh-Feldt の ε が 1 以上になったので，B の主効果や交互作用も自由度調整が不要である．Greenhouse-Geisser の ε で自由度調整すると，B の主効果は $p = .0173$，$\varepsilon = 0.79$，A×B の交互作用は $p = .0169$, $\varepsilon = 0.67$ となる．参考までに，MANOVA の結果をみると，A の主効果は $F(1, 4) = 8.10$, $p = .0466$ で単変量の分散分析と同じだが，B の主効果は $F(3, 2) = 1.81$, $p = .3755$，交互作用は $F(3, 2) = 16.53$, $p = .0576$ と，いずれも有意にならない．2-4 節で述べたように，サンプル数が少ないと MANOVA の検出力は自由度調整法に比べて低い．Fig. 3 に，この分析に用いた SAS プログラムとデータファイルを示す．

6-2. 下位検定

交互作用が得られたので，Fig. 2 の平均値のグラフを見ながら下位検定を行う．誤差項をプールせずに水準別誤差項を用いる．最初に，単純効果を検定する．まず，要因Aをa_1とa_2に分けて，それぞれで要因Bの主効果を調べる．a_1またはa_2のデータだけを使って新たに 1 要因分散分析を行うと，a_1では $F(3, 12) = 15.38$, $p = .0002$, $\varepsilon = 1.00$，a_2では $F(3, 12) = 0.23$, $p = .8747$, $\varepsilon = 1.00$ となり，前者でのみ有意差が得られる．次に，要因Bをb_1, b_2, b_3, b_4に分けて，それぞれで要因Aの主効果を調べる．各水準のデータだけを使って 1 要因分散分析（この場合は 2 水準なので対応のある両側 t 検定でもよい）を行うと，p 値はそれぞれ.0341，.7292，.1084，.0008 となる．1 回の検定あたりの有意水準を.05 のままにすると，b_1とb_4で要因Aの効果があったといえる．検定の繰り返しを考慮し，Bonferroniの方法を用いると，1 回の検定あたりの有意水準は.05 / 4 = .0125 となり，b_4だけで要因Aの効果があるといえる．

次に，交互作用対比を検定する．要因Bのすべての水準対の差をa_1とa_2で比較する．たとえば（b_1－b_2 at a_1）vs. (b_1－b_2 at a_2)という対比であれば，参加者ごとに“ b_1からb_2を引いた値 ”をa_1とa_2のそれぞれについて求め，対応のある両側 t 検定を行う．Table 4 に，すべての対比とそれぞれの p 値を示す．BonferroniやHolmの方法で検定すると，b_1－b_4だけが有意になる．このことから，交互作用は，b_1とb_4の関係がa_1とa_2で異なっていたために生じたといえる．

6-3. 多重比較

単純主効果が有意であった要因a_1については，平均値の多重比較を行う．反復測定なので，Bonferroniの方法かその改良版を使う．4 水準なので，すべての平均値の対比較では，4(4－1) / 2 = 6

回の検定を行う．すべての対比について，対応のある両側t検定で得られたp値をTable 4 に示す．有意差がある対は，Bonferroniの方法ではb_1-b_4，HolmやShafferの方法ではb_1-b_4とb_1-b_3，b_2-b_4であった．この違いは，方法による検出力の差を表わしている．全体のタイプIエラー率は.05 以下に抑えられているので，どの方法を用いてもよい．

もし，全体の分析において，要因Bの主効果が有意で，交互作用が有意でなければ，残りの要因の全水準（この場合はa_1とa_2）の平均値を使ったデータセットを新たに作り，それについて多重比較を行う．

6-4. 効果量

最後に，最初の分散分析の結果から，サンプルについての効果量の指標 η_p^2を算出する．Aの主効果は$F(1, 4) = 8.10$ なので，$\eta_p^2 = (1\times8.10) / (1\times8.10 + 4) = .67$ となる．同様にBの主効果は$F(3, 12) = 6.04$ なので，$\eta_p^2 = (3\times6.04) / (3\times6.04 + 12) = .60$，A × Bの交互作用は$F(3, 12) = 7.07$ なので $\eta_p^2 = (3\times7.07) / (3\times7.07 + 12) = .64$ となる．

6-5. 論文記載時の注意

このように得られた検定結果を，すべて論文に載せる必要はない．検定結果が羅列されると，たいていの査読者がクレームをつける．重要な数値だけを厳選して載せているつもりでも，研究結果を十分消化できていないという印象を与えてしまうからである．

Salovey (2000)は，結果を読みやすく書くコツを述べている．いくつか紹介すると，(a) 最も重要な発見から述べる，(b) 序論-方法-結果で記載の順序をそろえる，(c) 数値や統計量を示す前に言葉で明確に述べる，(d) 頻繁に要約を入れる，(e) 平均値などの記述統計量は F 値などの推測統計量よりも先に書く，(f) いつも正確な p 値を書かなくてもよい（小さい方が効果量が大きいと考えているという印象を与えてしまう），(g) 有意でないときは “marginally significant”，“just missed significance”，“trended in the right direction”などと防衛的な言い訳をしない（p 値を示せば有意でないことは分かる）．これらのアドバイスに加え，煩雑な統計結果は（どうしても必要なときは）本文中ではなく表にまとめて示すことも薦められる．

7. おわりに

本稿では，心理生理学の研究でよく行われる反復測定データの分析とその注意点について解説した．最適ではないかもしれないが，統計学的に妥当で，比較的容易に実施できる方法である．生理データに限らず，反復測定を行った主観・行動データの分析にも適用できる．

最後に，統計検定そのものについて 2 つコメントする．第 1 に，実際の研究場面で統計検定を行うときに感じる “ 不確かさ ” についてである．これは，現在の主流である統計検定法が，決して知ることのできない母集団の性質についてのさまざまな仮定の上に成り立っていることに由来する．このような足元の弱さをなくすために，母集団を仮定せず，無作為割り当てを前提として，シミュレーションにより実験要因の効果を検定する方法（確率化テスト randomization test）が提案されており，今

後の普及が期待される（橘, 1997）．統計手法に振り回されて，何を示したいのかが分からなくならないように，分析を行う前には仮説を明確にすることが重要である．

第 2 に，統計検定の使用と研究成果の信頼性との関係についてである．Cohen (1994)は，その論文“地球は丸い（$p < .05$）”の中で，ある事象が普遍的であることを示す過程で，統計検定は賢く使えば役立つが, 最後に頼れるのは（他の科学と同様に）結果が再現できるかどうかであると述べている．統計検定ではタイプ I エラーが一定の割合で生じるので，どんなに厳密に統制された実験でも，1 回の結果に基づいて結論は出せない．統計検定を厳密に行い有意性の有無にこだわるよりも，第 2，第 3 の実験で結果に再現性があることを示す方がずっと建設的である．また，統計検定は集団のデータを対象とするので，個人のデータを軽視しがちである．しかし，個人のデータをよく見れば，ある要因の効果がどの程度一貫して認められるかを知ることができる．このような視点は，心理生理学の研究成果を実際場面に適用していくときには，特に大切であろう．

脚 注

[1]これに対して，本当は差があるのに“差がない”と誤って判定する確率をタイプIIエラー率という．１からタイプIIエラー率をひいた確率，つまり，本当に差があるときに“差がある”と正しく判定できる確率を，検出力（statistical power）という．検出力は高いほど望ましいが，効果量（5 節参照）が同じときは，検出力を上げるとタイプIエラー率も増えてしまう．優れた統計検定法とは，タイプIエラー率を有意水準以下に抑えながら，検出力をできるだけ高くした方法である．

謝 辞

正木宏明先生（早稲田大学），小谷泰則先生（東京工業大学），城田愛先生（広島大学）には，草稿の段階で有益なコメントを多数いただいた．本稿を執筆できたのは，投石保広先生（朝日大学）の長年にわたる指導のおかげである．ここに記して感謝する．論文作成にあたり，科学研究費補助金 若手研究(B) 14710044 を受けた．

引 用 文 献

American Psychological Association. 2001 *Publication manual of the American Psychological Association (5th ed.)*. Washington, DC: Author.

千野直仁 1995 教育や心理の分野における ANOVA, MANOVA, GMANOVA 適用上の問題点 愛知学院大学文学部紀要, **25**, 71-96.

Cohen, J. 1994 The earth is round ($p < .05$). *American Psychologist*, **49**, 997-1003.

Cortina, J. M., & Nouri, H. 2000 *Effect size for ANOVA designs* (Sage Universitiy Papers Series on Quantitative Applications in the Social Sciences, Series No. 07-129). Thousand Oaks, CA: Sage.

Davidson, M. L. 1972 Univariate versus multivariate tests in repeated-measures experiments. *Psychological Bulletin*, **77**, 446-452.

Dunn, O. J. 1961 Multiple comparisons among means. *Journal of the American Statistical Association*, **56**, 52-64.

Geisser, S., & Greenhouse, S. W. 1958 An extension of Box's results on the use of the *F* distribution in multivariate analysis. *Annals of Mathematical Statistics*, **29**, 885-891.

Greenhouse, S. W., & Geisser, S. 1959 On methods in the analysis of profile data. *Psychometrika*, **24**, 95-112.

Hochberg, Y. 1988 A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800-802.

Holland, B. S., & Copenhaver, M. D. 1987 An improved sequentially rejective Bonferroni test procedure. *Biometrics*, **43**, 417-423.

Holm, S. 1979 A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.

Huynh, H., & Feldt, L. S. 1970 Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association*, **65**, 1582-1589.

Huynh, H., & Feldt, L. S. 1976 Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, **1**, 69-82.

Jennings, J. R. 1987 Editorial policy on analyses of variance with repeated measures. *Psychophysiology*, **24**, 474-475.

Jennings, J. R., & Wood, C. C. 1976 The ε-adjustment procedure for repeated-measures analyses of variance. *Psychophysiology*, **13**, 277-278.

Keselman, H. J. 1998 Testing treatment effects in repeated measures designs: An update for psychophysiological researchers. *Psychophysiology*, **35**, 470-478.

Keselman, H. J., Algina, J., & Kowalchuk, R. K. 2001 The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, **54**, 1-20.

Keselman, H. J., & Rogan, J. C. 1980 Repeated measures *F* tests and psychophysiological research: Controlling the number of false positives. *Psychophysiology*, **17**, 499-503.

Kirk, R. E. 1995 *Experimental design: Procedures for the behavioral sciences (3rd ed.)*. Pacific Grove, CA: Brooks/Cole.

小牧純爾 1995 データ分析法要説－分散分析法を中心に－ ナカニシヤ出版

Lix, L. M., & Keselman, H. J. 1996 Interaction contrasts in repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, **49**, 147-162.

Maxwell, S. E. 1980 Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics*, **5**, 269-287.

Maxwell, S. E., Camp, C. J., & Arvey, R. D. 1981 Measures of strength of association: A comparative examination. *Journal of Applied Psychology*, **66**, 525-534.

McHugh, R. B., Sivanich, G., & Geisser, S. 1961 On the evaluation of personality changes as measured by psychometric test profiles. *Psychological Reports*, **9**, 335-344.

Miller, G. A. 2000 Editorial. *Psychophysiology*, **37**, 1-4.

宮本友弘・山際勇一郎・田中敏 1991 要因計画の分散分析において単純主効果検定に使用する誤差項の選択について 心理学研究, **62**, 207-211.

森敏昭・吉田寿夫(編著) 1990 心理学のためのデータ解析テクニカルブック 北大路書房

永田靖・吉田道弘 1997 統計的多重比較法の基礎 サイエンティスト社

投石保広 1997 事象関連電位の分析における統計検定 丹羽真一・鶴紀子 (編著) 事象関連電位：事象関連電位と神経情報科学の発展 新興医学出版社 Pp. 113-124.

O'Brien, R. G., & Kaiser, M. K. 1985 MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, **97**, 316-333.

Olson, C. L. 1976 On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, **83**, 579-586.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. 2000 Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, **37**, 127-152.

Salovey, P. 2000 Results that get results: Telling a good story. In R. J. Sternberg (Ed.), *Guide to publishing in psychology journals*. Cambridge, UK: Cambridge University Press. Pp. 121-132.

Seaman, M. A., Levin, J. R., & Serlin, R. C. 1991 New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin*, **110**, 577-586.

Shaffer, J. P. 1986 Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, **81**, 826-831.

橘敏明 1986 医学・教育学・心理学にみられる統計的検定の誤用と弊害 医療図書出版社

橘敏明 1997 確率化テストの方法－誤用しない統計的検定－ 日本文化科学社

高橋行雄・大橋靖雄・芳賀敏郎 1989 平均値の比較 竹内啓(監修) SAS による実験データの解析 東京大学出版会 Pp. 336-344.

Vasey, M. W., & Thayer, J. F. 1987 The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology*, **24**, 479-486.

Wilson, R. S. 1967 Analysis of autonomic reaction patterns. *Psychophysiology*, **4**, 125-142.

Wilson, R. S. 1974 CARDIVAR: The statistical analysis of heart rate data. *Psychophysiology*, **11**, 76-85.

Table 1. 単変量の分散分析で差が検出できないケース（Davidson, 1972, Table 4, Case C）

参加者	X_1	X_2	X_3
1	52	50	71
2	56	46	91
3	66	62	45
4	40	30	15
5	42	36	39
6	46	48	67
7	46	44	5
8	52	42	27
9	68	62	85
10	62	50	55
平均	53.0	47.0	50.0

Table 3. 反復測定 2 要因のサンプルデータ（森・吉田, 1990, 例題 3・2・5）

	a_1				a_2			
参加者	b_1	b_2	b_3	b_4	b_1	b_2	b_3	b_4
1	3	4	6	5	3	2	3	2
2	3	3	6	7	5	6	2	3
3	1	4	6	8	2	3	3	3
4	3	5	4	7	4	6	6	4
5	5	7	8	9	6	4	5	6
平均	3.0	4.6	6.0	7.2	4.0	4.2	3.8	3.6

Table 4. 交互作用対比と多重比較

			比較あたりの有意水準		
	p	順位	Bonferroni	Holm	Shaffer[a]
交互作用対比					
(b_1－b_2 at a_1) vs. (b_1－b_2 at a_2)	.1836	4	.0083	.0167	–––––
(b_1－b_3 at a_1) vs. (b_1－b_3 at a_2)	.0506	2	.0083	.0100	–––––
(b_1－b_4 at a_1) vs. (b_1－b_4 at a_2)	.0016	1	.0083*	.0083*	–––––
(b_2－b_3 at a_1) vs. (b_2－b_3 at a_2)	.2658	6	.0083	.0500	–––––
(b_2－b_4 at a_1) vs. (b_2－b_4 at a_2)	.0614	3	.0083	.0125	–––––
(b_3－b_4 at a_1) vs. (b_3－b_4 at a_2)	.2262	5	.0083	.0250	–––––
多重比較（at a_1）					
b_1 vs. b_2	.0349	4	.0083	.0167	.0167
b_1 vs. b_3	.0090	2	.0083	.0100*	.0167*
b_1 vs. b_4	.0063	1	.0083*	.0083*	.0167*
b_2 vs. b_3	.1079	5	.0083	.0250	.0250
b_2 vs. b_4	.0123	3	.0083	.0125*	.0167*
b_3 vs. b_4	.1447	6	.0083	.0500	.0500

*全体の有意水準を.05 としたときに有意差があると判定された対比.

[a]対比（帰無仮説）間に論理的な関連がなければ，HolmとShafferの方法は一致する.

Table 2. すべての水準間で平均値の多重比較を行うときの比較あたりの有意水準

		全体の		小さい順に並び換えた p 値									
水準数	対比数	有意水準	方法	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
3	3	α	Bonferroni	α/3	α/3	α/3							
			Holm	α/3	α/2	α							
			Shaffer	α	α	α							
		.05	Bonferroni	.0167	.0167	.0167							
			Holm	.0167	.0250	.0500							
			Shaffer	.0500	.0500	.0500							
4	6	α	Bonferroni	α/6	α/6	α/6	α/6	α/6	α/6				
			Holm	α/6	α/5	α/4	α/3	α/2	α				
			Shaffer	α/3	α/3	α/3	α/3	α/2	α				
		.05	Bonferroni	.0083	.0083	.0083	.0083	.0083	.0083				
			Holm	.0083	.0100	.0125	.0167	.0250	.0500				
			Shaffer	.0167	.0167	.0167	.0167	.0250	.0500				
5	10	α	Bonferroni	α/10	α/10	α/10	α/10	α/10	α/10	α/10	α/10	α/10	α/10
			Holm	α/10	α/9	α/8	α/7	α/6	α/5	α/4	α/3	α/2	α
			Shaffer	α/6	α/6	α/6	α/6	α/6	α/4	α/4	α/3	α/2	α
		.05	Bonferroni	.0050	.0050	.0050	.0050	.0050	.0050	.0050	.0050	.0050	.0050
			Holm	.0050	.0056	.0063	.0071	.0083	.0100	.0125	.0167	.0250	.0500
			Shaffer	.0083	.0083	.0083	.0083	.0083	.0125	.0125	.0167	.0250	.0500

Note. 2 水準のデータだけを取り出して両側 t 検定を行い，得られた p 値を小さい順に並び替える($p_1 \leq p_2 \leq p_3 \leq \cdots \leq p_c$)．$p_1$ から順に表中の有意水準と比較し，有意でなくなった（値が大きくなった）時点で打ち切る．Shaffer の方法で用いる有意水準は，Holland & Copenhaver (1987)の上限値に基づき，なおかつ，最初の 1 要因分散分析が有意であったときの値を示した．

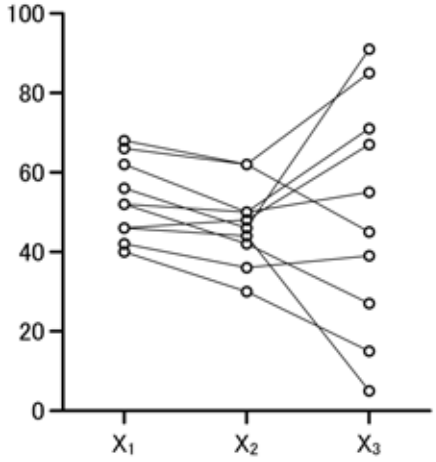

Fig. 1. 個人データのプロット

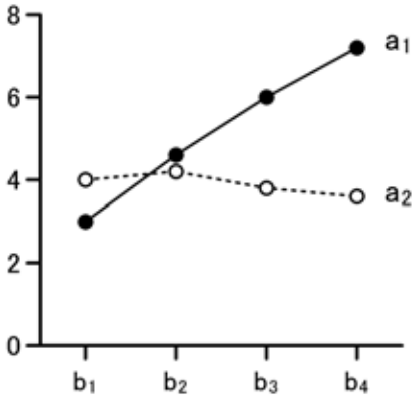

Fig. 2. サンプルデータの平均値

［プログラム］

```
data;
 infile"a:¥data.csv" delimiter = "," LRECL = 10000;
      データファイル名    区切り記号     1行の最大文字数

 do sub = 1 to 5;  参加者 1から5までのデータを読み込む.
  input d1-d8 @@; output;
 end;     d1からd8までの8個のデータ(観測値)を読み込み,
          分析用のデータセットを作る. 1行に複数の観測値
          があるときは @@ をつける.
proc glm;
 model d1-d8 = / nouni;  変数ごとの分散分析結果の省略
 repeated A 2, B 4;
          反復測定要因の指定(変数名  水準数)
run;
quit;
```

［データファイル］

各参加者のデータを横1列にする.
上位の要因から階層的に並べる.

$\overbrace{b_1\ b_2\ b_3\ b_4}^{a_1}\ \overbrace{b_1\ b_2\ b_3\ b_4}^{a_2}$

data.csv

```
3,4,6,5,3,2,3,2
3,3,6,7,5,6,2,3
1,4,6,8,2,3,3,3
3,5,4,7,4,6,6,4
5,7,8,9,6,4,5,6
```

Fig. 3. SAS による反復測定 2 要因の分散分析. 自由度調整法と MANOVA の結果が同時に出力される.