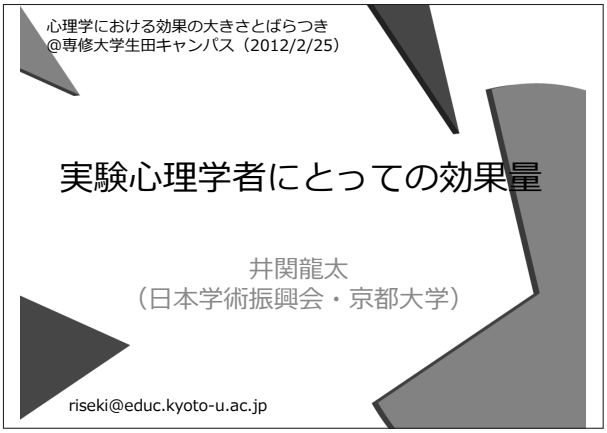
【講演1】

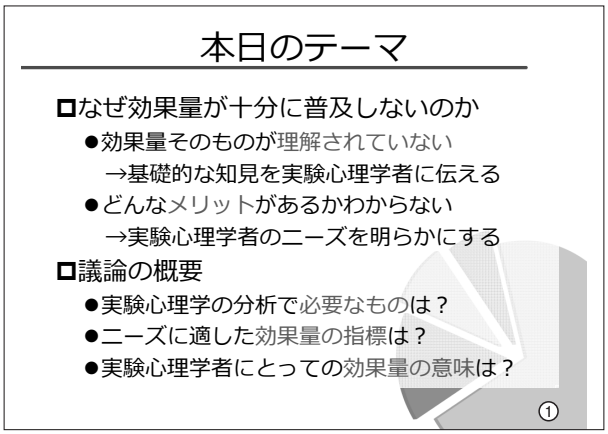
井関 龍太（日本学術振興会特別研究員（PD）・京都大学）

「実験心理学者にとっての効果量」



ご紹介いただきました井関です。このたびはこのようなお話をする機会を与えていただきまして，どうもありがとうございます。

今ご紹介いただきましたように，私は言語の記憶や理解に関する研究を主にやっていまして，どうしてこちらに呼んでいただくことになったかといいますと，おそらくは「ANOVA君」という分散分析のプログラムをつくっているためですね。その中に効果量のプログラムも含んでいて，そういうものをつくっているということで呼んでいただくことになったと思います。プログラムをつくった以上，もちろん効果量についてある程度知ってはいるのですが，統計学自体にそこまで詳しいというわけではありません。そこでどうしたらいいのかなと考えて，ここで実験心理学を専門とする方々あるいは実験心理学に興味のある方々と，統計学，特に今回の統計改革のあいだをつなぐようなお話ができればいいかなと考えています。



本日の大きなテーマとしてまず何を考えるかというときに，いちばん最初に思ったことはなぜ

効果量が十分普及していないのか，これについてお話しするのがいいのではないかと考えています。

どうして効果量が普及していないと考えているかといいますと，先ほど大久保先生からもお話がありましたように，実験心理学ではAPAマニュアルに書いてあるから，査読者に言われるから論文に書くという，そのくらいの認識で効果量をとらえている人が多いのではないか，さらに進んで言うと，査読者であっても本当に効果量についてよくわかっているのかちょっとあやしいところがあるなと，そういうふうに感じたこともあるからです。

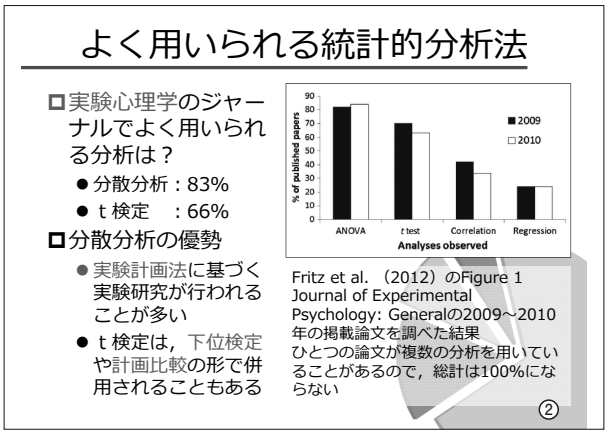
どうして普及していないのか，理由の1つとしてはまず効果量そのものが十分理解されていないからだ，ということがあると思います。この点に関しては基礎的な知見，統計学の効果量に関する基本的な情報をここで簡単にご紹介すれば少しはフォローになるだろうと思います。つまり，統計学から実験心理学者へと知識を伝える役回りとなります。

それからもう1つ，効果量というものがあることは知っているけれど，それにどういうメリットがあるのかよくわかっていない。それが伝わらないから広まらないということもあると思います。この点に関してはちょっとからめ手ですが，どういうメリットがあるかをそのまま答えるよりも，実験心理学者が分析において何を求めているか，そのニーズを明らかにすることから考えてみたいと思います。このトピックは，実験心理学者はこんなことを思っているよということを統計学のご専門の方にお伝えすることを意図しています。実験心理学者の側としては我が身を振り返るということにもなります。

流れとしては，最初にニーズの分析ということで，実験心理学の分析では一般にどういうものが必要とされていそうかということについてお話ししたいと思います。それからそのニーズになるべく適した効果量の指標とはどんなものなのかを考えたい。最後に実験心理学者にとって効果量はどんな意味があるのかという問題についてちょっと触れたいと思います。

実験心理学者が必要とする分析

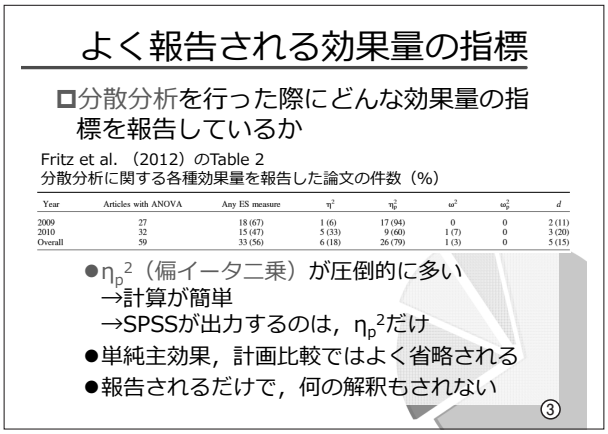
まず最初に実験心理学者のニーズはどんなものかということをお話ししましょう。ちょうど最近都合のいい論文が出ていまして，「Journal of Experimental Psychology: General」という実



験心理学の専門誌に載っています。著者のFritzという人たちは統計の専門家，おそらく効果量も専門としていて，いろいろなジャーナルについてどんな分析が実際に行われているのかということをあちこちで論文にしているグループです。今回の論文ではJEPのエディターから依頼を受けて，JEP: Generalの2009年と2010年の掲載論文について調査を行っています。

その結果の1つがこのグラフです（スライド2)。黒いほうが2009年，白いほうが2010年のデータで，掲載論文においてどんな分析が報告されているかをパーセンテージで示しています。1つの論文が複数の統計量を報告しているので，合計で100%にはなりません。これを見ていただいてどんな分析が多いかに注目すると，明らかに分散分析が多い。数値に直すとトータルで83%。続いてt検定が66%です。どうしてこういう結果になったかというと，おそらくは実験心理学のジャーナルでは実験計画法に基づく実験研究が行われることが多い。何らかの要因計画に基づく実験を設計しているので，基本的にそれを分散分析で分析して報告することになるのでしょう。

グラフからするとt検定もそこそこ多いように見えるのですが，下位検定とか計画比較のかたちで分散分析の後に使っているというパターンがけっこうあるので，実際はこのグラフの印象以上に分散分析がメインの分析として使われているのではないかと思います。ということで，ここからの話は分散分析を主に考えていきたいと思います。

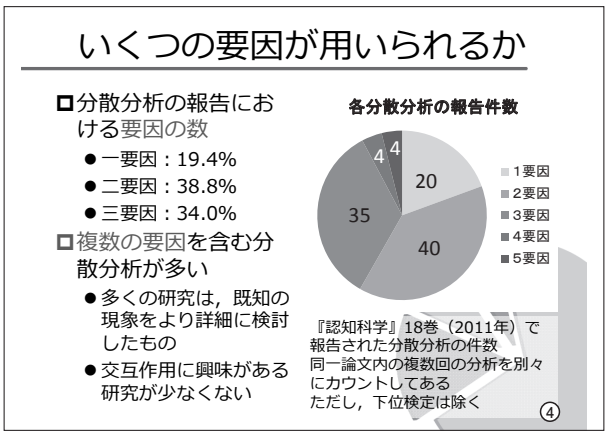


この表は今のFritzらの論文の続きですが，分散分析を行ったときに実際にどんな効果量の指標を報告しているかを示しています（スライド3)。3段目が2009年と2010年を統合した結果で，括弧内が全体のパーセンテージになっています。見ていただくとわかりますように，η_p^2（偏イータ二乗，パーシャルイータ二乗）が圧倒的に多くなっています。79%の研究がこれを報告していて，次に多いのがη^2（イータ二乗）で18%，ω^2（オメガ二乗）はほとんど報告されていない。どうしてこういう結果になったかというと，これはFritzらの考察ですが，1つの理由としてはη_p^2は計算がすごく簡単である。それからもう1つの大きな理由としては，SPSSが出力するのはこの中ではη_p^2だけだという事情があるのではないかと言っています。残りもけっこう重要なコメントだと思いますが，JEPの論文を分析した不満点として，単純主効果と計画比較ではよく効果量が省略されてしまうことが挙げられています。効果量自体は昔に比べると随分報告されるようになってき

ているけれども，なぜかこれらの分析についてはきちんと報告されないことが多い。それからもう1つの不満点として，これも先ほどのお話にあったと思いますが，効果量は報告はされている，でも報告されるだけで何の解釈もされていないじゃないかということについて述べています。

先走って効果量の話までしてしまったようなかたちですが，ちょっと話を戻して，どんな分析が用いられているかということについてもう少し考えてみたいと思います。

Fritzらの分析では，どんな分析を使うかというところに主な関心があって，分散分析の内容にはそれほど深く突っ込んでいませんでした。しかし，もう少しこれを詳しく見てみましょう。これから示すのは日本のジャーナル，『認知科学』の2011年の号で報告された分散分析の件数です。その件数をそれぞれ何要因の分散分析をしているかということについて調べました。ただこれから示すのは，このお話があってから私がものすごくざっと数えたものですので，あまり数値は信用しないでください。だいたいの目安として，雰囲気として受け取っておいていただくと助かります。

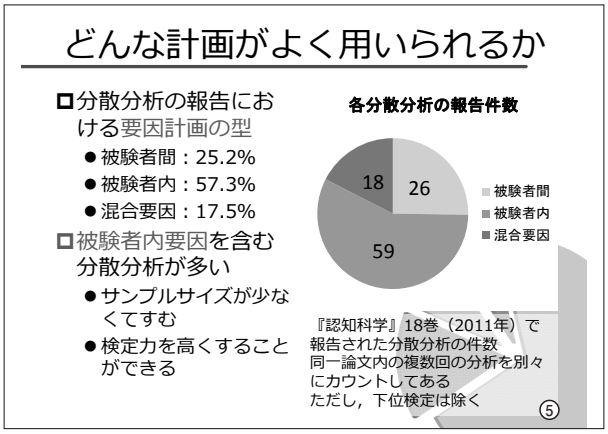


それぞれ見ていきましょう（スライド4）。まず1要因の分散分析はどれくらいあったか，これはざっとカウントすると20件くらいでした。次に2要因の分散分析は40くらい。それから3要因が35件。ちょっと驚いたのですが，4要因と5要因の分散分析も少しですが使われていました。これをパーセンテージにしますと，2要因と3要因の分析で70%以上とかなりの数を占めることがわかります。

ここから何が言えるかというと，分散分析を使ったというときには複数の要因，少なくとも2つ以上の要因の分散分析が多いということがわかると思います。これはどうしてかといいますと，おそらくほとんどの研究というのは2要因以上の計画を使いたいという要望がある。なぜかというと，実験心理学の分野でもまったく新しい現象を発見したという報告はすごく少ないわけです。ほとんどの研究は既存の現象をさらに詳しく調べたとか，あるいは応用的な関心から現実的な適用が有効な条件を調べたいという意図で行われている。だから基本的には多くの実験研究において2要因以上の計画になりやすい傾向があるのではないか。

今申し上げたこととも関連して，交互作用にそもそも興味がある，そういうタイプの研究もあるかと思います。成績の高い子どもにはこの学習法は特に有効でないけれども，成績の低い子に

は有効だ，そういう交互作用を検出することを目指している研究も少なくないのではないかと考えています。



同じデータをまた別の切り方で分析してみることもできます。これは先ほどとまったく同じデータですが，今度はどんな要因計画が使われているかで分類し直してみました（スライド5)。つまり被験者間か被験者内か，または混合要因かという観点です。これで見てみますと，被験者間の分散分析は四分の一くらい。被験者内計画がかなりを占めていまして半分よりも多い。混合要因計画の分散分析を使った分析はこのくらいです。同じようにパーセンテージに直してみますとこうなります。被験者内計画だけで50%以上を占めて57.3%。さらに混合要因計画は少なくとも1つ以上の被験者内要因を含むという意味で考えてみると，圧倒的多数が被験者内要因を含む要因計画を使っていることがわかります。これはなぜかといいますと，皆さんよくご存じのように，被験者内要因にしたほうが基本的にサンプルサイズが少なくてすむわけです。それから検定力も高くすることができる，分散分析で有意になりやすいということがあると思います。

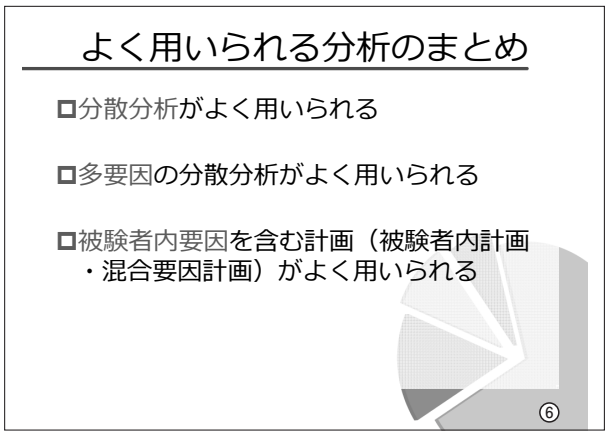
ここまでをまとめますと，まず実験系の心理学でよく使われる分析としては分散分析がよく使われている。それから『認知科学』に限定するのは代表性という点で妥当かどうかわかりませんし，また，もともと私自身この結論どおりの印象を持って行った分析だったので注意してほしいのですが，分散分析の中でも多要因の，2つ以上の要因を使った分析がよく使われている。それから被験者内要因を含む分析がよく使われている。これらの特徴が挙がってきます。そこで，効果量の指標を報告するときにもこれらの特徴に適したものを使うべきだといえるのではないでしょうか。

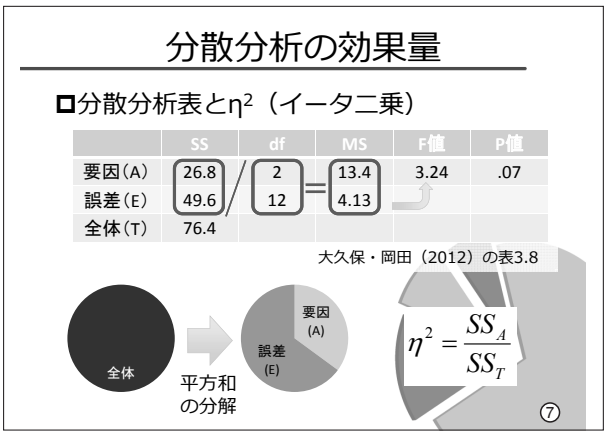
ニーズに沿った分散分析の効果量

そのようなわけで，ここからは分散分析の効果量についてお話しさせていただきたいと思います。

その前に，この後の話の内容に関係しますので，分散分析についてちょっとだけおさらいをさせてください。

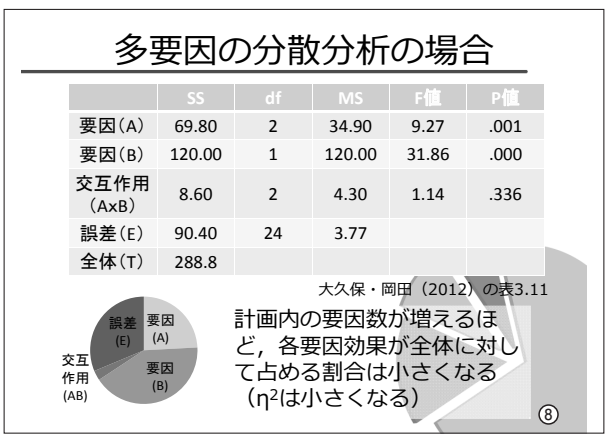
先ほども岡田先生からお話しいただきましたが，分散分析というのは基本的にはデータのばらつきを要因によるばらつきと誤差によるばらつきに分けるものです（スライド7)。それが基本的





な発想です。そのばらつきの度合いを示したものが平方和，SSと省略していますが，ここに挙げているような数値です。この平方和の数値をそれぞれの自由度で割ってあげる。割って出したものが平均平方です。要因の平均平方を誤差の平均平方で割って，F値を計算する。このF値を自由度を参照して有意かどうかを判定する。そうやって算出したものがp値である。およそこういった図式が分散分析の仕組みになっています。

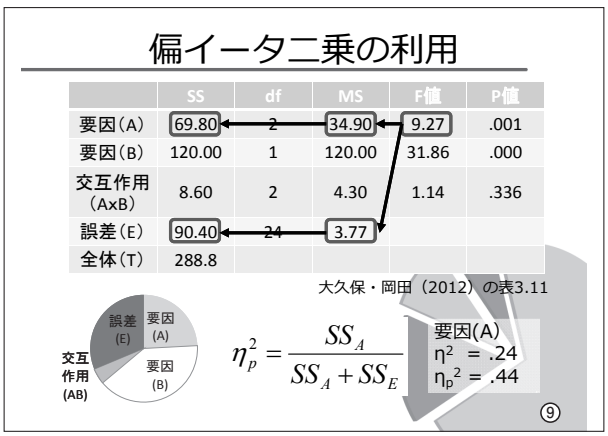
この関係性を図であらわしたものがこちらの円グラフです（スライド7）。これも先ほど岡田先生の発表にあったのでそれほど詳しい説明は要らないと思いますが，全体のばらつきをこのように要因と誤差のばらつきに分けていることを見やすく表現したものです。ここから考えると，ある要因の効果量というものは，要因によるばらつきが全体のばらつきの中でどのくらいの面積を占めるかで表すことができます。こういう発想で簡単に計算できるものがη^2と呼ばれる指標です。こういうふうに見ていくと，η^2というのはそんなにわかりづらいものではありません。そもそもの分散分析の想定からいくとかなりストレートに理解できる指標ではないかと思います。



しかし，皆さんご存じのようにη^2以外にも，分散分析の効果量にはたくさんの指標があります。その事情の1つとしてη^2だと困る場合があるわけです。これが先ほど挙がっていたような多要因

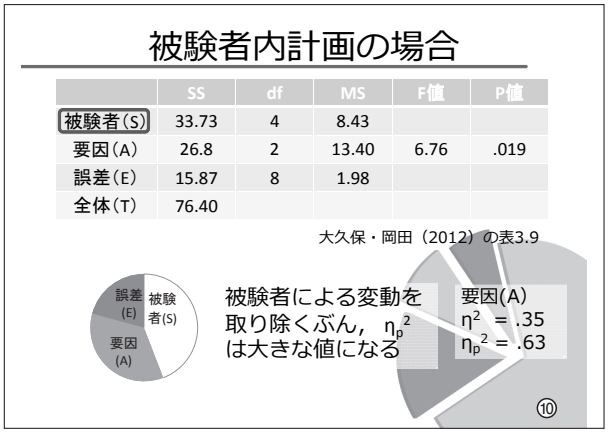
の分散分析の場合です。ここにあるのは2要因の被験者間計画の分散分析の表です（スライド8）。この場合も先ほどと同じように円グラフで平方和を表すとどうなるでしょうか。まずここに効果がたくさんあることから予想がつきますように，最初に要因Aの効果というものがある。それから要因Bの効果というものがあり，交互作用A×Bの効果というものがあって，残りが誤差ということになります。これで先ほどと同じようにη^2を計算するとどうなるでしょうか。計算すると今回は要因Aの効果というのは全体の中で占める割合がとても小さくなります。もちろん元々の数値が違うということもありますが，全体に占める割合というのは分析の中に入る要因の数が多ければ多いほど，さらに交互作用が多ければ多いほど小さくなります。仮にこの要因Aの効果が先ほどの1要因の分析の場合と同じだけの大きさだったとしても，計画の中にたくさんの要因があることによってη^2は小さくなってしまうわけです。そうすると，要因計画が違う研究どうしで効果量の大きさを比べるときにη^2だと都合が悪いということになります。

そこでこういう場合に参照できるのがη_p^2ということになります。これは先ほどとまったく同じ分散分析表ですが，ここで要因Aの効果の効果量について考えてみましょう（スライド9）。

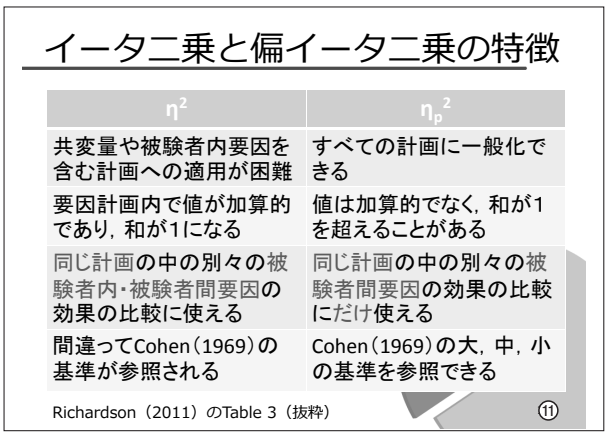


要因Aについて，この効果が有意かどうかを判定するときには，表右端のF値を使って判定していました。ここでさかのぼって，このF値がどこから出てきたかを思い出してみると，これは要因Aと誤差の平均平方を割って出した値でした。さらにさかのぼってこの2つがどこからきたかというと，これは当然それぞれの要因AとEの平方和からきているわけです。ということは，この効果について考えるときに重要なのは表の中でもこの2つの行だけではないか，そういうふうに考えることもできると思います。そうするとこの円グラフもそれに対応させて，要因Aと誤差のEのところだけ考えればいい。これを式にしたものをスライドの下の方に示しています。このようにしてあげると，同じ計画の中にたくさんの要因とか交互作用が入っていてもそれぞれの要因の効果を適切に判定できるのではないか。少なくとも，いま効果量を問題にしているのとは別の，ほかの効果の要因は除外して考えることができます。実際にこの分散分析表についてη^2を計算すると0.24ですが，η_p^2を計算すると0.44になり，この例ではこのくらい数値に違いが出てきます。

この関係は被験者内計画の分散分析の場合にも同じように適用できます。被験者内の分散分

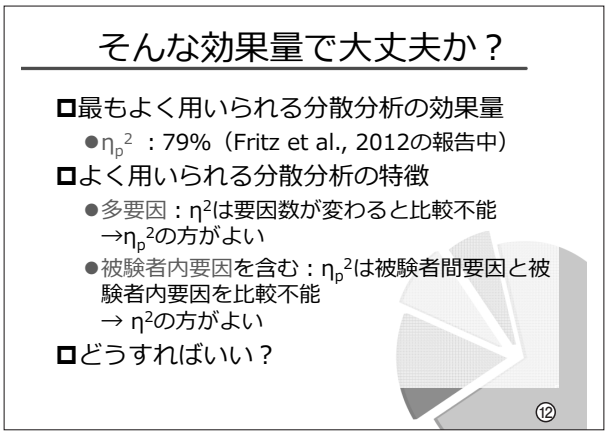


析の場合，こちらの分散分析表のようにデータのばらつきが分解されます（スライド10）。被験者による部分と，この場合1要因の分析ですから要因による部分と誤差による部分です。いま挙げた被験者による部分というのがこれまでの被験者間の分散分析にはなかった部分です。先ほど2要因の計画で行ったように，η_p^2を計算するときにはこの被験者によるばらつきの分も除外して考えます。そうすると被験者による変動を取り除く分，η_p^2はη^2より大きな値になります。実際に計算した値はこのとおりで，η^2よりη_p^2のほうがかなり大きな値になることがわかると思います。このことは，被験者内計画の方が検出力が高いという直感にも合致するでしょう。



ここまでに紹介したη^2とη_p^2の特徴をまとめてみましょう。これはRichardsonという人が書いた論文の表の一部分ですが，お手もとの資料にあると思いますので詳しくはゆっくり見てください（スライド11）。ただ，いまの文脈で重要なのは赤字になっているところです。η^2の特徴としては同じデザインであれば同じ計画の中の被験者間と被験者内の要因の効果を適切に比較できることがあります。これはなぜかというと，η^2は各要因の効果が円グラフの全体に対して占める割合を反映するので，単純に全体の中で占める割合を考えるという点で被験者間要因でも被験者内要因でも同じように比較できます。ところがη_p^2の場合にはこれができない。同じ計画の中の別々の被

験者間要因の効果の比較だけに使えて，被験者内の場合には使えないといいます。これはなぜかというと，先ほどη_p^2のときに被験者による部分を除いて，つまり完全な円ではなくて円の中の一部分だけを使ってその中での要因効果の割合を計算していました。そうすると，割合をとるときのもとになる分母の大きさがそもそも違っているわけです。だからη_p^2を使ったときには，デザインが違う場合，それから同じデザインであっても被験者内効果の場合は値を比較できないという問題が起こると言われています。



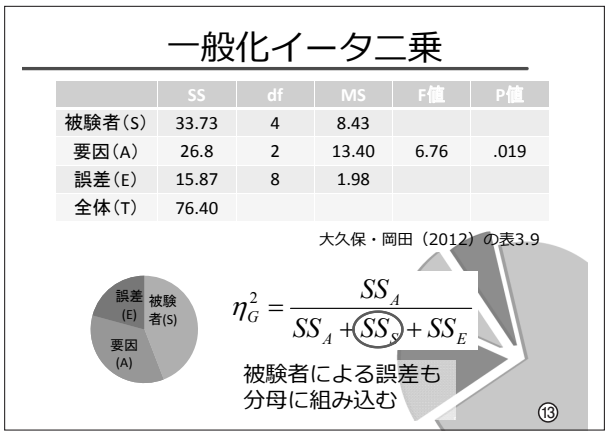
ここで最初に検討した分散分析に対する実験心理学者のニーズをふりかえってみましょう。

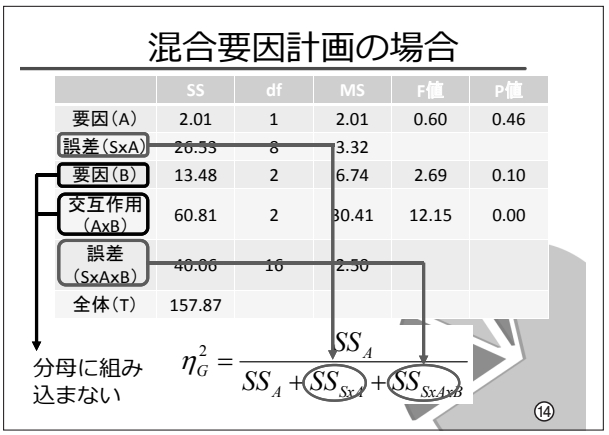
Fritzの研究の結果として先にも引用した箇所ですが，最もよく用いられている分散分析の効果量はη_p^2で，79%の論文が報告していました。一方で，よく用いられている分散分析の特徴を思い出してみましょう。1つは多要因の分析が多い，2要因以上の分析が多いということ。これと先ほど整理した効果量の特徴をあわせて考えてみますと，η^2のほうは要因の数が変わると比較するのが難しいという特徴がありました。多要因の分析が多いことを考えてみると，効果量として使用するのはη_p^2のほうがよさそうに思えます。

ところが，もう1つのニーズとして，被験者内要因を含む分析が多いということがありました。こちらを考えると，多くの場合η_p^2は被験者間要因と被験者内要因をお互いに比較できないので，この観点からするとη^2のほうがいいじゃないかということになってしまいます。そうすると，あちらを立てればこちらが立たずでどちらの効果量の指標を使えばいいかわからなくなるという問題が起こってきます。

ここで利用できそうなのが，2003年に提案された一般化イータ二乗（generalized eta squared）という指標です。このη_G^2（一般化イータ二乗）の基本的な考え方もそんなに難しくはありません。η_p^2のように円が欠けた状態を作って，分母の大きさがそろわなくなっているから比較できないのであれば，この取り払ってしまった部分を返してあげればいいではないかというのが基本的なアイデアだと思っていただければわかりやすいのではないでしょうか。ただ，この説明の仕方はやや正確ではないかもしれません。

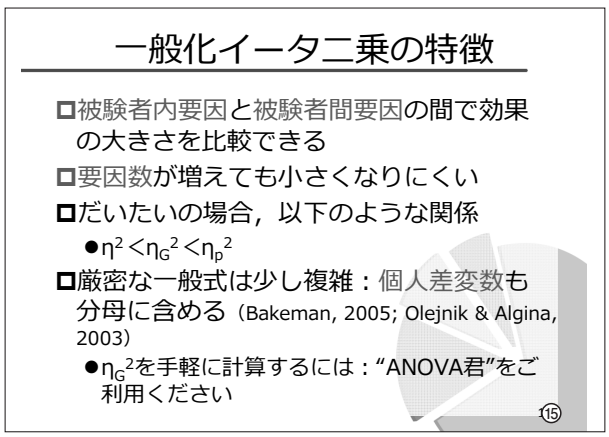
たとえば，η_G^2はこのようなグラフで表せるでしょう（スライド13）。η_p^2とどこが違うかというと，





被験者による誤差の部分を取り戻してあげていることです。この1要因の被験者内の式だとη_G^2と無印のη^2の式は同じになるので，どこが優れているのかまだわからない感じですが，もう少し複雑な計画の場合だとこの違いがわかってくるようになります。今度は混合要因計画の場合の分散分析表です（スライド14）。要因Aが被験者間要因，要因Bが被験者内要因になっています。この場合，要因効果Aの効果量のη_G^2の式はスライドのようになります。η_p^2の場合とどこが違うかといいますと，誤差の部分が2つありますが，両方ともが式に含まれている点です。η_p^2の場合ですと，要因Aの効果を出すときの誤差，分散分析表でいうと上のほうの誤差しか入りません。だからこのη_G^2の式はη_p^2の式と違うというのがわかると思います。

それから，これがもし無印のη^2の式だとすると，この2つの誤差に加えて要因Bの効果と交互作用A×Bの効果も含めた全体のばらつきというものを分母にとることになります。しかし，η_G^2の場合だとこれらの要因効果は計算に含めないわけです。このようなかたちで複雑な要因計画になるとはっきりするのですが，η_G^2の式というのは，場合によっては無印のη^2ともη_p^2とも違う式になることがわかると思います。

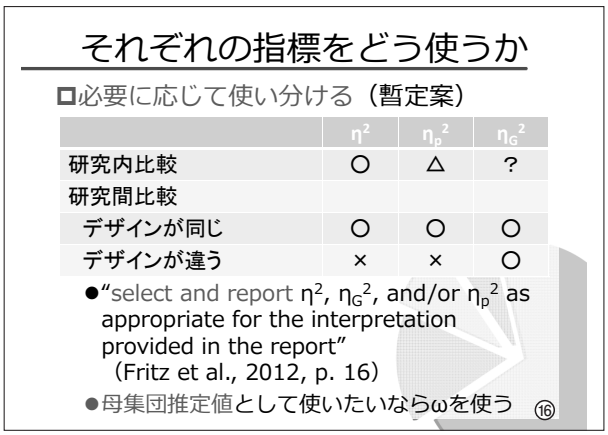


このようなη_G^2の特徴ですが，まず第一に指摘しておきたいのは，被験者内要因と被験者間要

因の間で効果の大きさを比較できることです。もともとこのことを目的につくられています。一方でη^2と違って，要因数が増えても値が小さくなりにくいということが言えると思います。これがなぜかというのは先ほど混合要因計画について説明したように，他の要因効果や交互作用による効果の部分を分母に含めないからです。η^2だとこれらの部分も分母に含めることになります。

このような特徴から，常にとはいえませんが，多くの場合は次のような関係が成り立ちます。つまり，η_G^2というのはη^2とη_p^2のだいたい中間くらいの値になることが多いと言えます。ただし，η_G^2の厳密な定義式，一般式はもう少し複雑なものになります。というのは，例えば男女の違いですとか，もともとの記憶能力の違いのような，個人差の変数も全部分母に入れないといけないということになっているので，これらも考慮しますと少々煩雑になるので今回は省略します。詳しくはOlejnikとAlginaの論文やその他の引用文献を見てください。また，このη_G^2を手軽に計算するには私がつくったものですが，「ANOVA君」というプログラムをご利用いただくのがよいと思います。インターネットで検索していただくと簡単に見つかると思います。

ここまでのところで無印のη^2とη_p^2，η_G^2と，様々な効果量の指標が出てきました。そうすると，いろいろな指標が出てきたけれども，ではどれを使ったらいいのかということが問題になってくると思います。これにつきましては先ほどから何度も参照しているFritzらはこのように書いています。「select and report η^2, η_G^2, and/or η_p^2 as appropriate for the interpretation provided in the report」と。つまり自分の論文に合ったものを選んで報告せよと言っているわけです。あともう1つのポイントとして，and/orと書いてあることに注目していただきたいと思います。and/orということは,1つだけ報告するのではなくて,自分の分析の目的に合ったものを複数併用してもよい，というよりは，場合によっては複数を併用して使ってほしいということです。η^2やη_p^2はあくまで効果量の「指標」であって，効果の大きさを絶対的に，一意に表しているわけではありません。厳密に言えば，真の効果量を推定するための指標，手がかりであるわけです。ですから，複数の指標を使って多角的に真の効果量を推定しようというのは適切なアプローチだと思います。



さらに今回は話が複雑になるので紹介しなかったのですが，もし母集団推定値として効果量を使いたいならω^2を使ったほうがいいと思います。これは岡田先生のお話にもあったと思いますが，

η（イータ）のほうは記述統計量に当たるような，サンプルについての効果量ですが，母集団における効果の大きさを推定したいのであればε（イプシロン）やω（オメガ）を使う必要があると思います。ωについても今回の議論とまったく同じように，何もついていない無印のω^2と，ω_P^2（偏オメガ二乗）と，ω_G^2（一般化オメガ二乗）の3種類があります。ηの場合とだいたい同じようなかたちで使い分けることができると思います。

さらに効果量の報告に関してもう1点だけ言っておきたいこととして，MSEの有用性について指摘しておきたいと思います。MSEというのは分散分析のときに出てくる誤差のほうの平均平方ですが，これを報告せよということがけっこう古い実験心理学の伝統としてあります。それをどうしていま持ち出すのかというと，この，MSEがあれば分散分析表の大部分が再現できるという利点があるからです。F値をどうやって計算したかというと，MS_AをMSEで割ったもので出している。だから，F値とMSEの2つの値があれば，MS_Aの値は逆算できますよね，割り算ですから。さらに，通常，分散分析の結果を報告するときには要因効果と誤差についての2つの自由度も報告します。そうすると，ある要因効果の検定に関わる，分散分析表の残りの部分もすべて再現できます。これはη_p^2ではできないことです。本当にできないかどうかはまたゆっくり計算してみてください。

このような関係性に注目すると，実はMSEのほうが効果量よりも情報量が多いのではないかという考えが浮かんできます。先ほど査読者も効果量についてよくわかっていないのではないかということを言いましたが，実は以前分散分析の結果にMSEをつけて投稿したところ，MSEは要らない，η_p^2などの効果量の指標をつけなさいというコメントがきたんです。そのときはこの人は何を言っているんだろうと思ったのですが，それはMSEを提示しておけばあとで読者のほうで好きな指標を計算できるという考えがあったからです。ただ，あるジャーナルや学問領域のスタンスとして，効果量のこの特定の指標を報告してくださいというのはありだとは思います。また，報告した結果をメタ分析に利用してほしいと考えるならMSEのほうがいいかもしれません。多数の研究の間で報告されている指標がそろっていないとせっかく集まったデータをメタ分析にかけられないということが起こります。あるいはメタ分析の分析者によっては分析のための指標として

何を使いたいかがあらかじめわからないということもあるかもしれません。こんなとき，特定の効果量を決め打ちで書いておくよりも，分析に関係する多くの統計量を再現できるMSEのほうを報告しておけば便利かもしれません。もしデータをメタ分析に使ってほしいという積極的な要望を持っているのであれば，効果量だけでなくMSEも書いたほうがいいのではないでしょうか。Fritzらも複雑なデザインを用いた分析の場合などにはMSEを記述することを勧めています。

効果量の“意味”

ここまでで効果量の指標がいろいろありますよということと，それをどう使うとよいかというガイドラインみたいなことをお話ししてきました。最後に，こうやって手に入れた効果量をどうやって使ったらいいかということについてお話ししたいと思います。

実験心理学の分野で研究をしていますと，よく言われるのはどのくらいの差であれば意味があるかということです。というのは，見方にもよりますが，実験心理学では比較的に抽象的な指標を使うことが多いからです。例えばよく使われる指標の1つに反応時間というものがあります。反応の速さといえば特に抽象的でもない，具体的な指標ではないかと思われるかもしれません。ところが実験心理学で使う場合，実際に人間の反応の速度自体を推定したいわけではないことが多いのです。人間がどのくらいの速度で反応するかということそのものを知りたいわけではなくて，ある条件と別の条件の間にどのくらい差があるのかを知ろうとしています。例えば「ドクター」，「ナース」という順に単語が出てくる場合と，「ベッド」，「ナース」と出てくる場合で「ナース」という単語への反応時間にどのくらい差があるか，といった違いに主な興味があります。このような実験を実際に行ってみたときにどのくらい差があるかというと，だいたい，せいぜい20 ～ 30ミリ秒の差なのですが，この数十ミリの差に何の意味があるのかというのはよく議論に上る点です。同じような議論はほかの指標にも当てはまると思います。例えば単純接触効果の実験などでは，AとBのどちらの対象が好きですかといったことを尋ねて，何度も接触した対象とはじめて接触した対象の間で何％くらい選好率が違うかということを調べます。このとき扱っているのも実際には接触条件と統制条件の間の数％の違いです。これにどんな意味があるかと改めて尋ねられると実はけっこう困るんです。人から聞いた例ですが，プライミング実験のような反応時間の実験で，条件間の平均値が9ミリ秒の差で統計的に有意であるという結果を出した人がいました。そこを査読者に突っ込まれたという話なんですね。統計的には有意なんだけど9ミリしか差がないなんて，何かこれはおかしくはないか。この疑問に答えるのはすごく難しいわけです。特定の課題のベースラインの反応時間にもよりますが，プライミングなどであれば，経験的には20 ～ 30ミリの差で有意になるというのが普通で，50ミリもあればかなり大きい差です。10ミリ秒台の差が有意になったとしたらかなりすごいなという感じですが，では9ミリ秒だったら意味があるのかないのかと聞かれるとすごく困ります。

こうした議論に対して効果量は何か答えることができるでしょうか。1つの答え方としては，先ほども出てきたCohenの基準を参照するということがあります。一見9ミリ秒で差がすごく小さいからあやしく見えるかもしれないけれども，効果量に直すとCohenの基準からみて十分な大きさ

効果量をどう解釈したらいいか

- どのくらいの差なら意味があるか
  - 反応時間：数10ミリ秒の違い
  - 選好判断：数%の選択率の違い
- Cohenのη_p^2の基準：
  - 小＝.0099（r = .10）
  - 中＝.0588（r = .24）
  - 大＝.1379（r = .37）
    - 絶対的な基準ではない
    - 研究内容によって解釈は変わる
  - 他の効果量の指標にも同様の基準が望まれる（t値やrとは変換可能）

⑱

実質科学的な知見に基づく解釈

- 効果そのものに実質的な意味がある場合（e.g., 応用的研究）
  - X%の改善，Y点の上昇・下降
    →どの程度なら意味があるかは，研究者・研究支援者のニーズが決める
- 効果そのものには意味を見出しにくい場合（e.g., 基礎的・理論的研究）
  - ほとんどの場合，効果の大きさについては問われない（範囲を定める基準もない）
    →専門家の直感と経験に頼るほかない？

⑲

の効果になっていますよという答え方がある。しかしこれで本当に査読者が納得するかというと，私はちょっとわからないなという感じがします。理想的にはこういう問題をどうやって解釈したらいいかというと，実質科学的な知見から，統計ではなくて研究テーマの中から導かれた論理でもって解釈するのがよいと思います。しかしこの解決策は効果そのものに実質的な意味がある場合とそうでない場合とで有効性が違ってくるように思います。実質的な意味がある場合として，例えばこの薬を投与すると血糖値が何%正常値に近づきますよとか，何かの得点，例えば抑うつ得点が2ポイント下がりますよと言われると，どの程度の意味があるかというのは，おそらく感覚として決められるのではないかと思います。ところが，基礎的，理論的な研究だとこうした合意の形成が難しいということが往々にしてあるのではないでしょうか。9ミリ秒だとあやしいという話をしましたが，ほとんどの場合，効果がどのくらいの大きさだったかという議論が実験心理学の論文の中で大きく取り上げられることはないのではないかと思います。ある条件と別の条件の間で平均値に差があるかないかということはよく議論しますが，その差の大きさがどのくらいかということは明示的にはほとんど問題にしないのです。それはなぜかといえば，どのくらいの大きさなら「よい」とか「わるい」といえるのかといった範囲を定める基準がないからではないでしょうか。結局，話は基準がないから決められないというところにもどってくるわけです。

先行研究に基づく判断

- 先行研究・先行実験の効果量に照らして判断する
  - 今回の実験の効果が相対的に大きい・小さいことがわかる（テクニカルな意味）
    →効果の実質的な意味が加えられるわけではない
    →どの程度の差なら先行研究・先行実験と違うといえるのか，基準は特にない

⑳

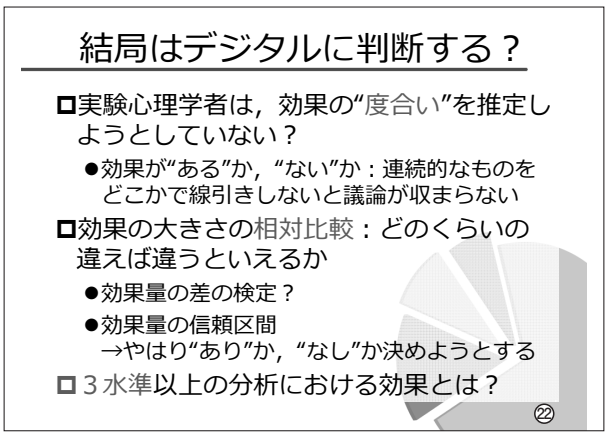
交互作用への興味

- 主な関心が交互作用の検出にある研究
  - 境界条件の画定
  - 理論的な要請
    →効果の大きさの評価基準を定めにくい？
- 主効果を重視する研究との関心の違い
  - × 特定の大きさへの関心
    （「方略Aで記憶成績を10%上げたい！」）
  - ○ 相対的な違いへの関心
    （「方略Aの効果は妨害課題を加えると消えるだろう」）

㉑

それからもう1つの基準となるのは先行研究です。あるいは自分の研究の中の先行実験で得られた効果量の値と比較する。これをやった場合に何がわかるかというと，今回の実験の効果がこれまでの実験と比べて相対的に大きいか小さいかがわかるでしょう。先ほど9ミリ秒の例を挙げましたが，効果量に変換すると実際にいくつだったのかはわかりませんが，仮に0.10くらいだったとしましょう。先行実験でも反応時間の差分はもっと大きかったものの，効果量は0.11とか0.12だったから今回の実験とそんなに変わらないとか，あるいは先行実験では0.20も効果量があったから今回の効果は小さいということがわかることでしょう。しかしここから，テクニカルに今回の実験と前の実験の効果量はどのくらい違うかということはわかりますが，これによって効果に実質的な意味が加わるわけではないんです。9ミリ秒が何をあらわしているのかという情報が新たに加わるわけではありません。先ほどの議論とも重なりますが，そもそも基準がないのでは先行研究と同じか違うかということも言えないと思います。例えば今回の実験は0.10，前の実験は0.11の効果量であったとしましょう。これなら違わないだろうと判断されるかもしれません。では0.12だったらどうだ，0.15は？とやっていくときりがないのではないでしょうか。

どうしてこんなふうに実験心理学で効果の大きさをあまり議論しないかというと，1つにはやはり交互作用への興味が強いということが一部の研究にあるからではないかと思っています。交互作用の検出に主な興味のある研究として，例えば境界条件を確定したいというようなタイプの研究があると思います。ある効果がこの環境では出てこの環境で出ない。例えばプライミング効果が自分の母語なら出るけれども外国語だと出ないというような仮説を検証したい場合を考えると，興味のあるのは交互作用ということになります。あるいは，理論的に交互作用があるはずだ，特にクロス型の交互作用があるはずだという研究の場合も基本的に交互作用にしか興味がないということになります。



これらの場合には効果の大きさがどのくらいという評価基準はたぶんすごく定めにくいのではないか。例えば特定の大きさへの関心がある研究として，方略Aを使わないときに比べて記憶成績を10%上げたいというのであれば，これはすごくわかりやすい目標で，どのくらいの差があれば効果ありと判断するのかはっきりとわかります。これに対して交互作用の場合，特に，方略A

の効果は妨害課題を与えると消えるだろうといった仮説を検証しようとする場合には，妨害課題を与えないときと与えるときの2つの状況の間にどのくらいの差があればいいのか，しかもどこにどういう差があればいいのかを予め想定するのはかなり難しくなるのではないでしょうか。

結局は実験心理学者というのは効果の度合いをあまり推定しようとしていないというところに問題があるのではないかと思います。実際，測定しているのは連続的な変量であることが多いので差分もたいていは連続量なのですが，どこかの時点で差があるかないかを線引きしたい，そうしないと議論がおさまらないという部分があって，仮説検定が隆盛を誇っているのではないかと思います。

このような文脈の中でそれでも実験心理学者が効果の大きさについて積極的に議論しようとする可能性としては，効果の大きさを相対比較したい場合があるのではないかと思います。これは実際に効果量を使って効果の大きさに違いがあることを主張したいというような研究を想定しています。例えば方略Aと方略Bでどのくらい学習効果が違うかということを言いたいとします。そのために効果量を使って主張したいという人がときどきいます。ところがこの場合もちょっと微妙な問題がありまして，効果量が方略Aだと0.20で，方略Bだと0.10だという結果が得られたとします。こう言われると差がありそうに思いますが，では方略Bの効果量が0.15だったらどうですか，あるいは0.17のときは？というふうになってくると，どこで差があるのかという問題にやっぱり戻ってきてしまうのではないか。だから効果量の差の検定はないのかと，そういう意見を耳にしたことがあります。

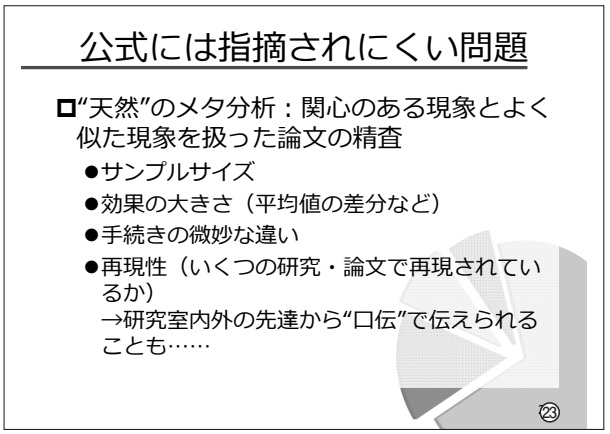
これについては実際に効果量の信頼区間を計算する方法がいま活発に開発されているので，信頼区間を利用して似たようなことができるようになるかもしれません。しかしこの根底には，2つの効果の間に差があるのかないのかを1か0かで決めようとするという，そういう発想があります。ですから，技術的には信頼区間を使うことによってうまく2つの効果量の違いを統計的に判定できるかもしれませんが，それをやってしまうと実は今回の統計革命が目指していたのとは違うところに着地してしまうのではないかということを感じます。今回の革命は，p値をみて機械的に効果があるかないかを判定するのはやめよう，もっときちんと効果の有無とか大きさといった問題に向き合おうという姿勢が根底にあったと思います。そのための効果量のはずなのに，この効果量の大きさや差を信頼区間や検定で機械的に決めてしまうのでは本末転倒ではないでしょうか。

それから，3水準以上の分析の場合の効果というものを考えにくい，想像しにくいということも実験心理学者が効果量に基づく判断がしづらい理由の1つかなと思います。いずれにしろ，実験心理学の中で効果というものについてのコンセプションというものがあまり考えられていないことが問題なのでしょう。あるいは統計学のほうでいう効果の概念とのすり合わせがうまくいっていない面があるのかもしれません。

最後に少しこれまでと矛盾するような話になるかもしれませんが，ここまで実験心理学者が効果量で評価しようとしているものについて，効果の度合いについて考えていないということをお話ししたのですが，実際はそれに近いことをやっているよという側面についても指摘しておきた

いと思います。実際，明示的には議論していなくても，実験心理学者は効果の大きさについての直感を持っていると思います。9ミリ秒の差は何かあやしいという査読者の指摘があったことはそうした直感があることを示しています。さらに，実はこのデータを出した研究者の方ご自身も「この結果についてはやっぱり何かおかしなことが起こっていた気がする」と話しておられました。やはり専門家の直感と経験は無視できないものがあるように思います。

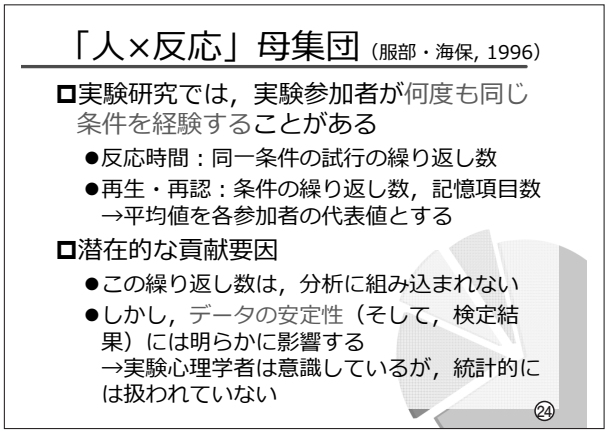
実験心理学者が実験を計画するときにまず何をするかというと，自分が関心のある現象とよく似た現象を扱った論文を詳しく読むということをすると思います。このとき何を詳しく見ているかというと，特に追試をしようというレベルまでくると，サンプルサイズだとか，効果の大きさ――この場合効果量というよりもだいたい平均値の差分ですが――，微妙な実験手続きの違い，それから実験結果の再現性のあたりに注目していると思います。論文やプレス記事でセンセーショナルに喧伝されている効果でも，原典をよく読んでみると一連の実験の中で1回しか検出できていない効果だったとか，その効果も実は有意傾向に留まったとか，そんなこともままあります。それから，同じ効果についてたくさん論文が出ているけれども全部同じ著者によるものだったとか，そういうこともあるわけです。



こういうような微妙な按配については研究室内外の先達から口伝で伝えられることもあります。これは，オフィシャルには言われていませんが，メタ分析に近い活動をやっていることになるのではないでしょうか。そのような事前の分析の一例として，あまりはっきり指摘されない要因のひとつとして，「人×反応」母集団について取り上げておきたいと思います。これをはっきり指摘している文献は，私の知る限りでは服部・海保（1996）だけです。「人×反応」母集団ということで何を言っているかというと，実験研究だと，一人の実験参加者が何度も同じ条件を経験するということがときどきあります。これは分散分析の反復測定という意味とは違います。どういうことかというと，例えば反応時間実験ですとまったく同一な条件に属する試行を何回も繰り返すということをふつうします。反応時間というのはかなり振れの大きい，敏感すぎる指標なので，20試行とか30試行の繰り返しをとるのが普通です。それから再生や再認でも同じようなことをします。試行ごとに再生や再認をする場合だと，反応時間実験の場合と同じように，同じ条件を繰り返し

経験することになります。あるいは，すべての項目を記憶したあとに項目を思い出してください，再認してくださいといったタイプの実験だと，記憶項目の数がくり返しの回数に相当することになります。この繰り返し数についても実は実験を追試しようとするときは研究者はけっこうよく見ているわけです。例えばこのタイプの実験をやるには全部で50個くらいの単語を覚えさせるのが普通だといったことを確認する。ところが実はこの試行数，繰り返し数というのは分析には直接は反映されていなくて，同じ条件に属するすべての試行の成績の平均値を各参加者の代表値として使用することが一般的なわけです。この繰り込む数というのは研究によって違うことがあります。しかし，反応時間の例で指摘したように，この繰り返し数は分析で考慮されていなくても，実際にはデータの安定性に明らかに影響しています。そしておそらく検定結果にも影響しています。これは私の知っているかなり狭い範囲の話ですが，例えば文章理解の分野の反応時間の実験だと，だいたい人×反応数が200 ～ 300はないと有意にならないなという感覚はあります。200くらいだと安定性が低くてやっぱりうまく出ていない＝統計的に有意でないとか，300を超えて400近いデータがあるとやっぱり安定して効果を検出できるということが実感として感じられたりします。だから繰り返し試行数が少なくても人数で補うことはできるんだなという感覚はあります。

実験データのこういう側面については実験心理学者は意識していますが，統計的には，明示的には扱われていないのではないかと思います。もしかしたら私が知らないだけできちんと問題として取り上げている研究なり手法なりというものもあるのかもしれませんが，そういうものがあったらまた教えていただきたいと思っています。



今回のメッセージとしては，効果というものについてもっと考えましょうということがいえるのではないかと思います。

実験心理学者は効果について，また効果の大きさというものについてより明示的に考えていく必要があります。効果についてのコンセプションが十分に練られていないからどんな評価をしてよいかわからない。結果として，ただ言われるままに自分でも意味のわからない統計量を書くだけということになりかねません。

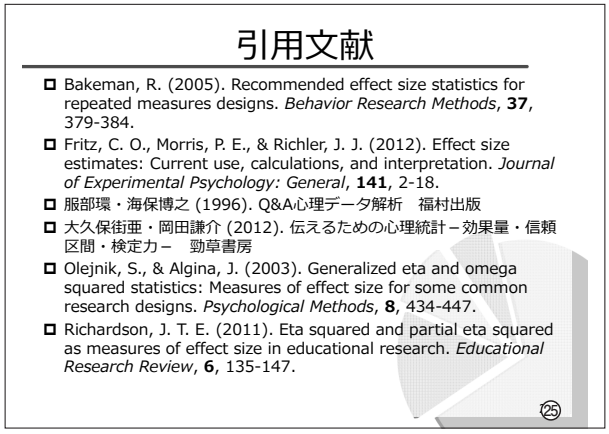
一方で，統計学の側に対しては，実験心理学者のニーズにより即した方法を提示していただき

たいと思っています。多くの統計量は反復測定デザインを前提としていませんが，実験心理学でよく用いられるのは反復測定要因を含むデザインです。また，一要因の計画だけに使える方法も実用的ではないということがいえます。

それから，交互作用や3水準以上を含む要因の効果や効果量というものをどう考えるか。これについては，実験心理学者も考える必要がありますし，統計学の側からもどんな意味合いがあるのか，どんな性質の統計量となるのかを説明する義務があるのではないでしょうか。そうしたやりとりの中で，もしかしたらこれまでの扱い方では十分でなかった，主効果とは区別する必要があるとか，2水準とそれ以上では意味が変わってくるとかいったことが判明するかもしれません。

効果量についての概説論文では，効果量を書くだけでなく，それについて考察すべきだということをよく言っています。Fritzらの論文にもそのように書かれています。しかし，そのような論文で効果量についての具体的な考察の例が書かれているのを私は見たことがありません。だいたいは直感的な観点から解釈していて，つまり，血糖値が何％下がったというような，もともと測定しているものの意味がわかるような例が出されているだけで，多くの実験心理学の研究には適用できないような議論になっています。この点については，統計学の側でも効果というものについて本当はよく考えていないことを反映しているのではないかと感じてしまいます。そうすると，統計学の側に対しても，効果というものをどう考えているのか，単に統計量という話ではなくて，それをどこにグラウンディングさせるのかについて問い直す余地があるのではないでしょうか。

実験心理学者も効果についてまったく考えてこなかったわけではありません。実験計画を聞いてだいたいこのくらいの効果が期待できそうだなと予測したり，データを見てこれは何だかおかしな結果だと判断できたりします。この専門家のもつ直感と経験をより多くの人々の間で共有できるようにする手段のひとつとして統計学は強力な役割を果たすことができます。この相互作用を実現するためには，各領域の専門家は実践の中で培ったものをもう一度ふりかえってとらえ直し，ことばにして伝える必要があります。統計学者の方には，できればそのニーズに答えうるものを提示していただきたいと思っています。また，逆に統計学の側から提示された数理的なモデルが実践家の悩んでいた問題を解決するヒントになることもあるでしょう。いずれにせよ，どちらかだけからの，一方通行の情報提示にとどめないようにすること，双方からのコミュニケーションをより密にしていくことが実験心理学と統計学の双方の研究をますます豊かにしていくのではないでしょうか。以上です。ご清聴ありがとうございました。



（岡田）　どうもありがとうございました。ご質問等ございましたらお願いします。

（質問者）　どのタイミングで質疑で聞けばよかったのか迷っているところなのですが，ぼくは動物の実験をしていまして，動物の実験心理学だと，nが全然稼げない，非常に小さいnしかとれないというような場合があったりするわけです。最初の大久保先生，岡田先生のお話とも絡むところだと思うのですが，極端にnが小さくしかとれないような場合なんかでは一体どういう工夫をするのがよいでしょうか。これはこれでニーズの1つかなと思うのですが，教えていただけるとありがたいです。

（井関）　ご質問ありがとうございます。極端にnが小さい場合にどうするかというお話ですね。私もそのあたりは専門というわけではなく，むしろ先生のほうがお詳しいと思いますが，やっぱり最後にお話ししたように試行数で稼ぐというのが1つの手ではないかと思います。試行数を，知覚実験なんかだとn=3で1,000試行とか1万試行とかやったりすることがあると思います。それから，その現象というのが，例えばそのn=3の研究で，3名の特殊な人でしか出せない効果ではないということをどこかで保証しておけば，科学的にほかの研究者から認められる成果になり得るのではないか。そのnの少ない個体たちがそれぞれ均質なものであるとか，ほかの根拠から，自分の想定している研究の範疇からずれたサンプルではないということを十分に論証できればそれはあやしいという話にならずにすむのではないかと思います。はっきりしたお答えではないかもしれませんが。

（質問者）　もちろん繰り返しがたくさんできればそれはそれでいいと思うのですが，非常に少ないnしかないときに有意性検定というのは非常に不利といいますか，難しい問題がおそらくあるだろうと。そこで効果量というものがそれを補うというか，あるいはリプレースするようなかたちで使えるとnが稼げないような事態であっても，科学的な議論がしやすくなるのではないのかなというふうな期待をもつのですが，そのあたりはいかがでしょうか。

（井関）　nが少ない場合効果量で何かものが言えないかということですが，それはおそらく実験心理学の分野でけっこう広くあるニーズのようで，ほかのところでも同じような議論を聞いたこと

があります。ちょうどこのセッションの前にその話をしていたのですが，それをたぶんいま査読でやると，nが少ないから検定で結果が出ていないだけで実はこんなに効果量があるのだというふうに主張すると，じゃあサンプルサイズを増やして本当に出るか見せてくださいということになるような気がしてならないんです。これをやれば大丈夫ですといった絶対の解決は私は見出せないと思います。その点は，サンプルサイズを増やせば信頼できる結果であることを確認できるといった発想についても同じです。ただ，まだいまのところは増やしなさいという話で終わってしまうのではないかという気がします。

（岡田） 効果量の信頼性に関しては標本サイズに依存しますので，そのへんはなかなか難しいところかと思います。

ほかにご質問いかがでしょうか。

（質問者）1つはコメントでもう1つは質問です。コメントですが，最後に井関先生ご指摘の，人×反応の数が明示的には分析に，統計に検定といったかたちで入ってこないという話があるのですが，ものすごく見当違いかもしれませんが，functional MRIの解析のときには使っているかなという気が直感ではします。個人レベルも，要するに脳の反応データですからすごくノイズが大きいのですが，簡単に言えば脳領域をものすごく小さく切って，個人レベル解析で最初にやってから，各個人で出てきた，有意になったらそれぞれの1個1個のデータ値を使ってもう1回集団レベル検定，2段階やっている，それに近いのかなと，全然違うかもしれませんけど。

もう1つは質問ですが，もう1つ同意できる点がすごく多いのですが，MSEの話です。あれがすたれたのは何でなのかなと，一昔前，私が大学院生だった90年代くらいはけっこうあれを書いていたと思うのですか，あれがスタンダードだと思っていたのですがいまはあまりないので，何ですたれちゃったんだろうなというのは，想像でお話ししていただきたいのと，私は学生らにレポートを書かせるときには必ず分散分析表を載っけてねと言うんです。というのは間違いがあったときにチェックできないので絶対書けと言うのですが，でも論文を書くときには要らないからねという矛盾した指導をしていて自分でもおかしいと思っているのですが，これは分散分析表を書かない理由は何だったかなと，スペースの問題だったかなという記憶がちょっとあるんですけど，あれは何で載せないとなっているのか，そのあたりの理由をご存じでしたらお話しお願いします。

（井関） ありがとうございます。MRIでのお話など，私も実験心理学の代表みたいな顔でここにいますが，すべての分野をフォローしているわけではないので，必ずしもそのあたりを詳しく知っているわけではありません。お話しいただいて参考になりました。おそらくマルチレベルとかそういうものを使って個人ごとのプロフィールみたいなものを分析してそれを上の段階で統合するみたいなことをすると，先ほどの人×反応数の問題を分析に反映させることはできるかと思います。ただ，現在はあまりやっていないということですね。1つの理由としては，実際の実験では欠損値が多く出るからだと思います。例えば反応時間の実験ではすべての試行のデータを同じように分析に使えるわけではありません。ある試行ではエラーになったりとか，ある試行のデータはどう見ても外れ値のようだなど，いろいろ起こる。しかも，これらのエラーも一律に同じ反応過程を原因としたものであればまだ何とかなりそうですが，判断の間違いであったり，よそに注意

を取られていたり，実験機器の問題であったりと多様なものでありえます。マルチレベルによる解決が普及するには，より優れた欠損値処理の方法が必要になるのではないでしょうか。

それから先ほどのMSEをどうして書かなくなったかというお話ですが，これは私にもなぜ書かないのかよくわからないというのが正直なところです。私が大学院生だったときは，書かなくてもいいけれども，ばらつきをあらわす指標として書くとよりいいよみたいなかたちになっていたと記憶しています。

分散分析表をそのまま載せないことについては，私も，いろいろ言うのなら分散分析表を載せたらいいのではないかと思っているのですけれども，載せない理由としては紙面の節約のためだということしかとりあえずは聞いたことはないです。ただ，いちいち論文に論述として細かく書くよりも分散分析表を載せてもらったほうが読むのに楽だと感じることはときどきあります。

（質問者） ありがとうございました。

（岡田） どうもありがとうございました。

（岡田） それでは時間となりましたので，残りのセッションを始めたいと思います。

続いての講師の先生は山形伸二先生です。山形先生は行動遺伝学，発達心理学，パーソナリティ心理学をご専門とされており，東京大学大学院での博士課程を修了後，日本学術振興会特別研究員を経て，現在は独立行政法人大学入試センターの特任助教でいらっしゃいます。

本日のご講演のタイトルは「行動遺伝学からみた効果量―遺伝子と環境はどのように個性を生み出すか―」ということですので，よろしくお願いいたします。