統計的検定における検定力と効果量

データ解析演習　2010年7月14日
M1　内海健太

本日の発表内容

- t検定と分散分析
- 「平均」の比較と「分散」分析
- 分散分析の復習
- 検定力
- 効果量

t検定と分散分析

- t検定

  2つの標本の母集団の母分散が未知の場合における平均値差の検定の際に用いられる。

- 分散分析

  3つ以上の平均値の比較の際に用いられる。

t検定と分散分析

○ なぜ分散分析が必要か？

- t検定で3つ以上の平均値を比較すると，
  危険率が高くなる。
  → 実際は差が無いが，たまたま有意差が出る確率
- A, B, C3つの平均値の比較の場合を考えてみる。
  この場合の危険率，すなわち「3回の比較のうち
  少なくとも1回は有意差がでる確率」は・・・

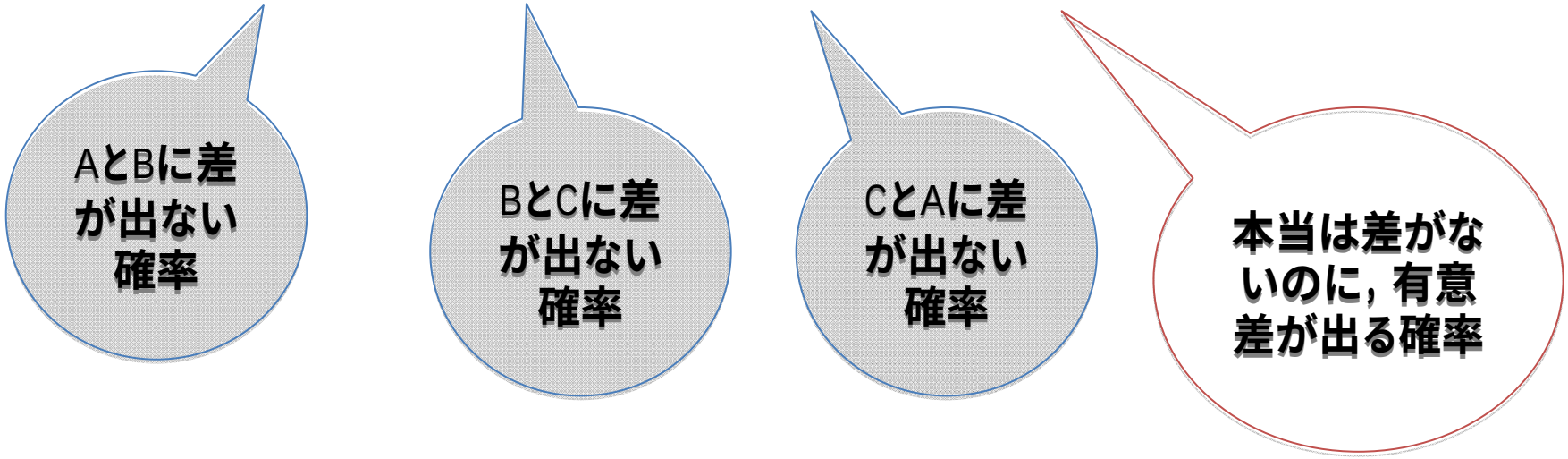
「1ー余事象（3回とも有意差が出ない確率）」

で，算出できる。

t検定と分散分析

- 「1－余事象」の算出

$$1-(\underline{1-0.05})\times(\underline{1-0.05})\times(\underline{1-0.05})=0.14$$



- t検定を繰り返すと有意水準を下げることになる。
  → “検定多重性の問題”
- t検定で比較ができるのは，2つの平均値差まで。

「平均」の比較と「分散」分析

○ 平均の比較になぜ分散分析が用いられるか？

- 分散とは偏差（値ー平均）の2乗和をデータ数で割ったもの。
- 分散分析では，各データの値の全体平均からのばらつき（全体平方和）は，実験要因で説明可能なばらつき（群間平方和）と，実験要因では説明できないばらつき（群内平方和）に分解できる，と考える。

値の分散　＝ 要因で説明可能なばらつき＋説明できないばらつき

全体平方和　＝　群間平方和　＋　群内（誤差）平方和

「平均」の比較と「分散」分析

・平方和はデータ数nで割れば分散になる。ゆえにこの分析は，群の平均値の比較を分散を用いて行うものと言える。

・つまり分散分析には，「ばらつきの大きさを比較しなければ，各群間に観察された平均値の差に意味があるかどうかはわからない」という前提がある。

・すなわち，平均値に差があるかどうかは，データの分散にかかっていると言える。

分散分析

○ 分散分析表 (2要因被験者内)

変動因	平方和	自由度	平均平方	F値	p	
被験者:S	1138614.30	23	49504.97			
要因A	104957.10	1	104957.10	4.94	0.04	*
誤差:A×S	488295.20	23	21230.23			
要因B	64346.20	2	32173.10	7.06	0.002	****
誤差:B×S	209672.39	46	4558.10			
交互作用:A×B	35134.88	2	17567.44	2.07	0.14	n.s.
誤差:A×B×S	390898.19	46	8497.79			
全体	2431918.26	143				

+ p<.10, * p<.05, ** p<.01, *** p<.005, **** p<.001

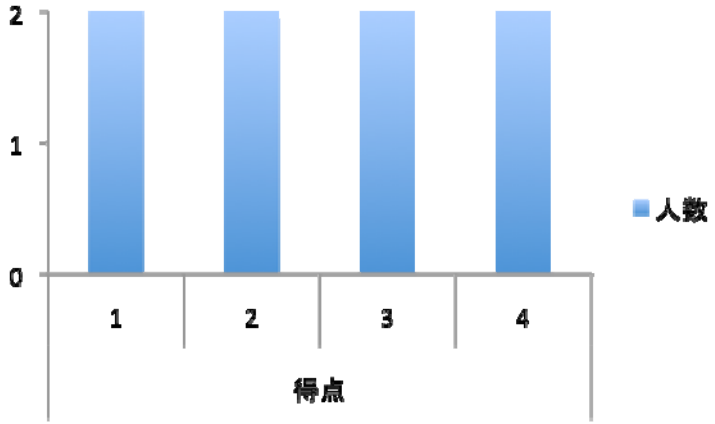
分散分析

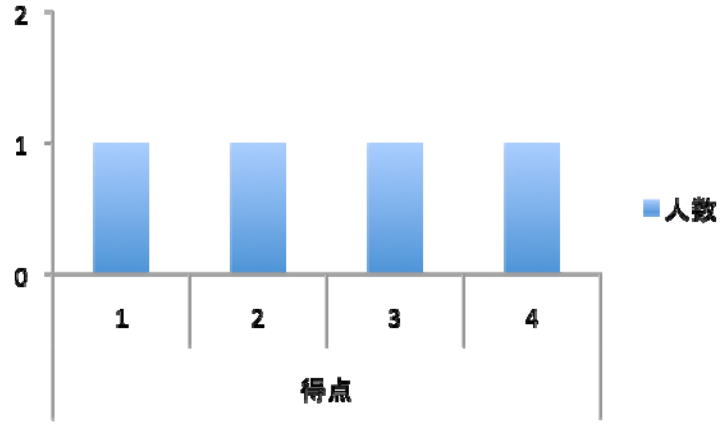
○ 平均平方

・「平方和÷自由度」で算出

・ 平方和の比較だけでは，平方和の大小関係が，ばらつきの大小と自由度の大小いずれに起因しているのかわからない。

(例)





・上記2つのデータは，平均・分散共に同じであるが，平方和が異なる。

(平方和；図1：10，図2：5)

→ 平方和は自由度が大きいほど大きい値をとるため，上記の公式にあてはめ，自由度1あたりの平方和の値(平均平方)を求めなければならない。

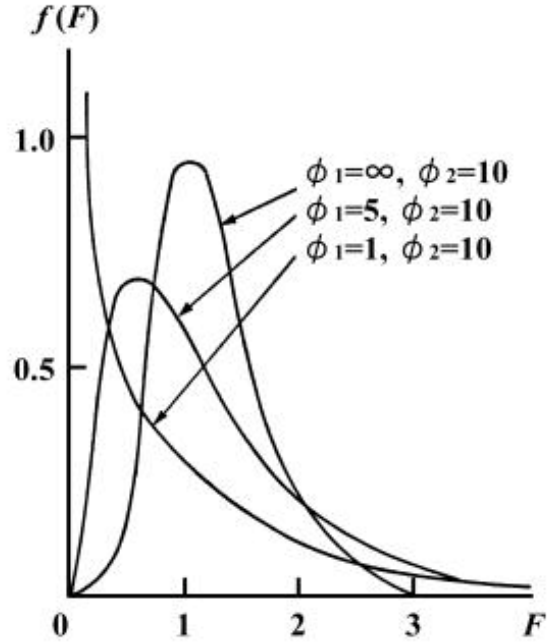
分散分析

○ F値

・分散分析の検定統計量として用いられる。

・F値は，「群間の平均平方(MS_A)／群内の平均平方(MSe)」で算出。 → <u>F値が大きいほど，群間の平均値に差がある。</u>

○ F分布



：F値

：確率密度

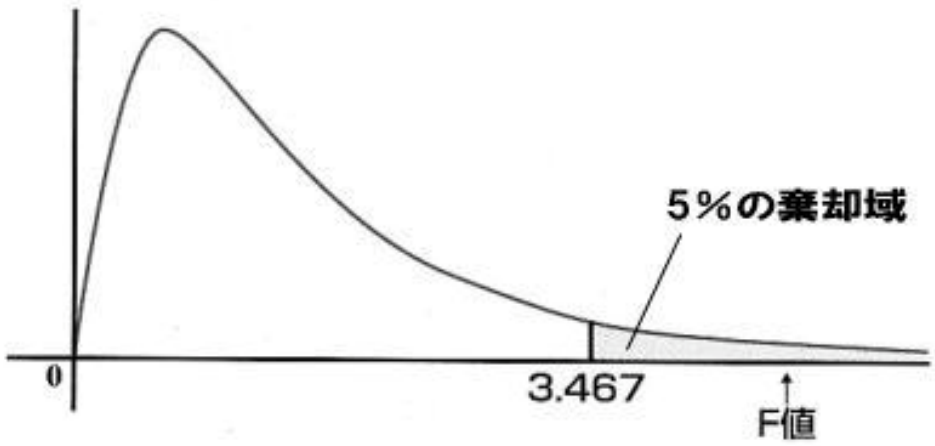
φ1：群間の自由度

：群内の自由度

分散分析

○ F分布

(例) ある要因 (水準数：3) の主効果に有意が認められるか
どうかの検定。分析の結果，F(2,21) = 7.058。



– 上の図より，上記の検定統計量の実現値は棄却域に入る。
＝ 帰無仮説が正しいという前提のもとでは，確率的に
めったに得られない様な値が得られた，ということ。

分散分析

○ 検定

・実際のデータから求められた検定統計量Fの実現値も参考にして帰無仮説が棄却されるかいなかを判断。

※ A, B, Cの3群のデータを比較していた場合

H_0：μA=μB=μC，3群の母平均は等しい。

H_1：H_0ではない。

・帰無仮説が棄却されても，3群間のどこに差があるのかはわからない。

→ 下位検定が必要。

分散分析

○ 下位検定

・多重比較

- 3条件以上の平均値の比較においてANOVAの結果有意差が認められた場合，どの条件間に差があるのかを確かめるための検定。

注意点

・ANOVAはF統計量を用いる検定

- F統計量を用いる多重比較(Scheffe法，FisherのPLSD法など）は，ANOVA後に行わなければならない。
- 一方，F統計量を用いない多重比較(TukeyのHSD法，Bonferroni法など）では，ANOVAで有意差が認められなくても，有意となることがある。 → 前もって分散分析をする必要はない。

分散分析

○ 下位検定

・単純主効果検定

- 交互作用が出た場合にそのうちどれかの要因における水準毎に，その他の要因の効果を検定する手法

注意点

・“交互作用が有意”だからといって，“単純主効果”も有意になるとは限らない。

・上記と逆のことも同様に言える。

・「交互作用が有意」→「単純主効果の検定」という流れが一般的であるが，この2つの検定は包括関係では無いので，交互作用が有意でなくても単純主効果検定を行うことは間違いではない。

効果量とは

○ 「統計的に有意＝実質的に意味のある差」？

（例）正規分布に従うA, Bの２集団間からそれぞれ1000人ずつ被験者を抽出し，知能検査の結果を比較しました。

結果は・・・

集団A・・・平均：110，SD：10

集団B・・・平均：109，SD：10

＜分析＞

このデータについて，有意水準５％でt検定を行うと，

$t(1998) = 2.22, p<.05$ と，有意になる。

→ ~~２群の平均値差が大きいことを示す。~~

→ サンプルサイズが大きいため有意となってしまった。

効果量とは

○「統計的に有意＝実質的に意味のある差」？

・この例は，「統計的に有意な差＝実質的に意味のある差」ではないことを示している。

→ 検定統計量は，効果の大きさとサンプルサイズの両方から影響を受ける。

・先ほどの例では，サンプルサイズが，検定統計量が棄却域に入るほどの大きさであった，と考えられる。

・効果の大きさとは，この例の場合「グループの違い」が平均値差に及ぼす影響を指す。

この効果の大きさのことを，効果量(effect size)という。

効果量とは

○「統計的に有意＝実質的に意味のある差」？

- 先の例より…
  - p値の大小が“効果の大きさ”を表しているわけではない。

  - 統計的仮説検定で扱うのは「有意差があるか否か」の二者択一の判断であることを忘れてはならない。

↓

サンプルサイズに左右されるp値とは別に，
それに左右されない「効果の大きさ」を表す必要がある。

効果量とは

○「統計的に有意＝実質的に意味のある差」？

・つまり，p値が小さくなる状況は2つ考えられるということである
- 効果量が大きい時
- サンプル数が多い時

・ 統計的な有意が認められた場合，どちらかの影響によるものかを調べなければならない。

効果量とは

○「統計的に有意＝実質的に意味のある差」？

・先ほどの例は，「サンプルサイズが大きいため，検定結果が 有意になる」という例であった。
しかし，逆に「サンプルサイズが小さいため，検定結果が非有意になる」という場合もある。

↓

検定力が低いことが原因

・つまり，「統計的に非有意」な結果は，必ずしも「実質的に意味がない差」ということを指すわけではない。

検定力

○検定力

母集団に差があるとき，サンプルにおいて有意な結果が得られる確率。

	真実	
決定	H_0は正しい	H_0は間違い
H_0を棄却	第１種の誤り 確率はα	正しい決定 確率は$1-\beta$ （検定力）
H_0を棄却しない	正しい決定 確率は$1-\alpha$	第２種の誤り 確率はβ

※「1－β」が検定力

検定力

○検定力

α：実際には差が無いのに，有意差があるとしてしまう確率
（第１種の誤り）

β：実際には差があるのに，有意差無しとしてしまう確率
（第２種の誤り）

・当然ながら，α・β共に小さい方が望ましい。
しかし，標本抽出などの条件を固定すると，αとβの両方を
小さくすることは，通常できない。

検定力

○検定力

・すなわち，αとβはトレードオフの関係。
  – 第1種の誤りが発生しないよう危険率を低く設定すればするほど，第2種の誤りが増える（＝検定力が弱くなる）。

・しかし，<u>αとβの和は一定ではない</u>。
  - 帰無仮説が真で，正しく採択される確率：1－α
  - 帰無仮説が偽で，正しく棄却される確率：1－β
    → αとβが混在する状況が存在しない。

検定力

○検定力

・有意水準αは，5％や1％と定められることがほとんどであるが，βは定めることが難しい？

- 帰無仮説は「$\mu_{実験} = \mu_{統制}$」と，限定した状況を表しているため，αを定めることが可能。

- 一方，「帰無仮説が偽」であるのは様々な状況があるので，βを定めることは難しい？

→ Cohen(1992)では，「検定力（1－β）が0.80以下の場合には，第2種の誤りを犯す可能性が高くなる」としている。

検定力

○検定力

Q. 第2種の誤りは，ランダムに現れるのか？

A. そうではない。

・特定の危険率αに対して，標本数が一定であるならば，
効果量が大きいほどβは小さくなる。

・効果量が一定であるならば，標本数が大きくなるほどβは
小さくなる。

検定力

○ 検定力のまとめ

・有意水準α・効果量・検定力・標本数の4つは，**他の3つが決まれば残りの1つも決まってくる。**

＜他の2つの条件が同じである場合＞

・ αを大きくすると検定力も大きくなる。逆も同様。

・ 効果量が大きければ検定力も大きい。逆も同様

・ 標本数を大きくすると検定力も大きくなる。

→ 一方で，少しの差にも統計的有意が認められる可能性がある。

効果量

○ 効果量とは

・実験的操作の効果や変数間の関係の強さを表す指標。

(Field & Hole, 2003, p.152)

・要因の操作による影響が強い場合，および要因内のばらつき（誤差）からの影響が弱い場合に大きな値をとる。

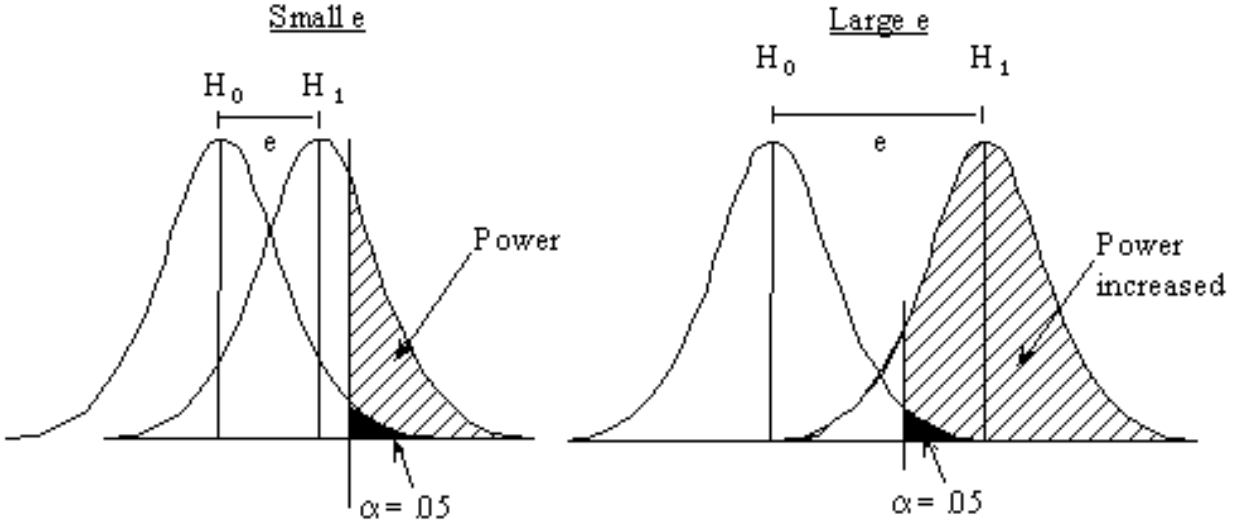
・サンプルサイズによって変化することのない，標準化された指標。

→ 測定単位の影響を受けない，等のメリット。

効果量

○ 効果量とは

両条件ともSDが等しく，かつ，正規分布に従う場合



効果量小　　**効果量大**

※ 効果量の大小によって，両分布の重なりの程度が異なってくる。

効果量

○ 効果量はなぜ必要か？

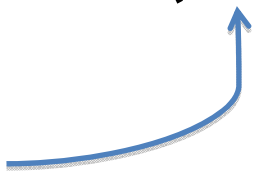
・F値との違い

- F値はデータ数が多い程増加する。

cf) F値＝「群間の平均平方(MS_A)／群内の平均平方(MSe)」

- MSe ＝ 平方和 (SSe) ／自由度(df)

→ データ数が多いほど小さくなる



F値は大

※ 効果量は影響を受けない。

効果量

○ 効果量を表す指標

(1) *d* family (Cohen's *d*, *f* など)

グループごとの平均値の差を標準化した効果量

(2) *r* family (r^2, η^2, partial η^2, ω^2, R^2 など)

変数間の関係の強さを示す効果量

効果量

(1) *d* family

・Cohen's *d*は，t検定のような2グループの平均値の差を比較するときに使用。

|2つの条件の平均値差|

$$d = \frac{(\text{実験群の平均} - \text{統制群の平均})}{\sqrt{\frac{\text{実験群の標準偏差}^2 + \text{統制群の標準偏差}^2}{2}}}$$

2つの条件の標準偏差の平均値

効果量

(1) *d* family

・ちなみに，先ほどの*d*は，2群間のデータ数が等しい時の計算方法である。

・以下に各群のデータ数が異なる場合の*d*の算出方法を記す。

$$d = \frac{(\text{実験群の平均} - \text{統制群の平均})}{\sqrt{\frac{(\text{実験群の人数} - 1) \times \text{実験群の標準偏差}^2 + (\text{統制群の人数} - 1) \times \text{統制群の標準偏差}^2}{(\text{実験群の人数} + \text{統制群の人数}) - 2}}}$$

効果量

(1) *d* family

- グループごとの平均値の差を標準化した値が得られる。
  → *SD*を単位として，平均値がどれだけ離れているか
  の指標。(例) $d = 1$は，1*SD*離れていることを指す。
- つまり、2つの条件の平均値差が、それらの標準偏差
  の何倍であるかを示す値。
- この値が大きいほど、2つの条件の分布の重なりが小さ
  いことを示す。
  → 一方の条件がもう一方の条件よりも値が大きい傾向が
  より顕著。
- 数式の理論上，上限と下限は無制限なので，d の値は1を
  超え得る。

効果量

(2) *r* family

・相関係数に基づいた効果量。

・t検定における効果量*r* は，t値と自由度を使って計算される。

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

・ ここで計算される*r* は，実験群と統制群のグループを表す名義尺度（例；実験群：0，統制群：1）と，データの間の<u>点双列相関係数</u>(point – biserial correlation coefficient)である。

効果量

(2) *r* family

・**点双列相関係数**(point – biserial correlation coefficient)

- **2つの変数のうち，一方の変数が2値しかとらず，もう一方の変数が連続変数の場合の，2変数間の相関係数。**

・**ピアソンの積率相関係数*r*とは異なり，算出される値は常にプラス。**

効果量

(2) *r* family

- 分散分析では，相関比率 η を用い， η^2 で効果量を表すことが多い。

$$\eta^2 = \frac{\text{ある要因の平方和}\ (SS_{effect})}{\text{全体平方和}\ (SS_{total})}$$

- η^2 では，全体における，ある要因の占める割合を計算。
- 分母が全体平方和なので，$0 \leqq \eta^2 \leqq 1$ となる。
- 「ある要因の平方和」が大きい場合，η^2 は大きい値をとる。

  → 群の違いによって説明できるばらつき
- 「全体平方和」が大きい場合， η^2 は小さい値をとる。

効果量

(2) *r* family

欠点

1. 標本数が小さい時は，算出値が大きくなる。

2. 他の独立変数の個数や，その独立変数の優位性の影響を受ける。

- **η^2 の合計は1.0であるため，独立変数の数が増えると1つの独立変数が占める割合も少なくなる。**
  ＝ 独立変数の数が増えると，η^2は小さくなる
- **つまり1つの独立変数の効果がどれほどのものであるかの判断がしにくい。**

↓

偏相関比(partial η^2)

効果量

(2) *r* family

○偏相関比(partial η^2)

他の要因の影響を統制した上で，1つの独立変数の影響の効果量を計算。

$$\text{Partial } \eta^2 = \frac{SS_{\text{treatment}}}{SS_{\text{treatment}} + SS_{\text{error}}}$$

ある要因の平方和 (SS effect)

- SSerror (誤差) が大きいと，partial η^2 は小さくなる。
- partial η^2の効果の大きさの明確な基準は無い。

効果量

(2) r family

○偏相関比(partial η^2)

<u>欠点</u>

・partial η^2 の合計値が1.0を超えるため，換算を行っても従属変数に占める各独立変数の割合が解釈困難。

・つまりη^2は母集団における推定値が不正確であり，一般化ができない。

→ <u>より正確な母集団推定値ω^2</u>

$$\hat{\omega}^2 = \frac{SS_{\text{treatments}} - df_{\text{treatments}} MS_{\text{error}}}{SS_{\text{total}} + MS_{\text{error}}}$$

効果量

(2) *r* family

○ ω^2 (Omega Squared)

・独立変数によって説明される母集団における不偏分散の推定値を提供

・前ページで紹介したpartial η^2 の欠点を解消

<u>注意点</u>

・ω は各群の人数が等しい時しか使えない

効果量の指標と目安

cf) **水本篤, 竹内理**(2008)

検定(分析)の種類ごとに見る代表的な効果量の指標と大きさの目安

使用される検定（分析）	対象と注意	効果量の指標	効果量の目安 小 (Small)	中 (Medium)	大 (Large)
(1) 相関分析		r	.10	.30	.50
(2) 重回帰分析		R^2	.02	.13	.26
		f^2	.02	.15	.35
(3) t 検定（t-test）	対応あり・なしともに同じ	r	.10	.30	.50
		d	.20	.50	.80
(4) 一元配置分散分析 (One-way ANOVA)	全体の差の検定	η^2	.01	.06	.14
		partial η^2	-	-	-
		ω^2	.01	.09	.25
		f	.10	.25	.40
	多重比較	r	.10	.30	.50
(5) 二元配置分散分析 (Two-way ANOVA) 多元配置分散分析* (Multi-way ANOVA) *三元配置以上の分散分析	主効果	η^2	.01	.06	.14
		partial η^2	-	-	-
		ω^2	.01	.09	.25
	交互作用	η^2	.01	.06	.14
		partial η^2	-	-	-
		ω^2	.01	.09	.25
	多重比較	r	.10	.30	.50

効果量の指標と目安

(6) 共分散分析（ANCOVA）	共変量の影響を取り除いて分析し，主効果，交互作用，多重比較の効果量は (4) や (5) と 同じ				
(7) 多変量分散分析 (MANOVA)	多変量検定	multivariate η^2 (multivariate R^2)	-	-	-
		multivariate partial η^2	-	-	-
多変量共分散分析 (MANCOVA)	従属変数ごとの分散分析	主効果，交互作用，多重比較の効果量は (4) や (5) と 同じ			
(8) カイ 2 乗検定 (χ^2 test)	2×2 の分割表	φ (= W)	.10	.30	.50
	2×2 以外	Cramer's V	.10	.30	.50
(9) マン・ホイットニーの U検定 ウィルコクスンの符号順位和検定 クラスカル・ウォリスの順位和検定 フリードマン検定	検定統計量を Z に変換して r を求める	r	.10	.30	.50

Note. Cohen (1998; 1992), Field (2005), Tabachnick and Fidell (2006) などを基に作成。 効果量の大きさはあくまで目安であるので研究分野によって変わる。(3) d, (4) f, (8) W についての詳細は，Cohen (1988) を参照のこと。 η^2 の大きさの目安は文献によっては，r を 2 乗した r^2 に合わせて，η^2 = .01（効果量小），η^2 = .09（効果量中），η^2 =.25（効果量大）としているものもある。また，partial η^2 の効果の大きさの基準は明確なものがない。multivariate η^2 と multivariate partial η^2 の値は従属変数（dependent variable）の数によって変わるため，効果量の目安は Cohen (1998) を参照。

効果量

○ 効果量の用途

（1）p値とともに結果に示す。

本日も紹介した様な例から・・・

- 有意差が認められたが，実際はあまり効果がない。
- 有意差は認められなかったが，実質的効果が期待される。

の様なケースが実際に起こりうる。

そのため，有意差が認められようが認められまいが，いずれにせよ効果量は報告すべきであろう。

効果量

○ 効果量の用途

(2) メタ分析

- p値を基準とした比較は，サンプルサイズの影響を受けるため，メタ分析にはふさわしくない。ゆえにメタ分析にはp値以外の基準が必要。
- そこでメタ分析では，類似した複数の実験結果から，それぞれ効果量を算出したものを総合し，全体としてどのくらいの効果があるかを評価する。
- メタ分析を行う時に使用する効果量は同じものに統一。

効果量

○ 効果量の用途

(3) 検定力分析

・被験者の人数を決めるために事前に行うもの。

・実験して有意差が得られなかった場合に「第2種の誤り」を事後的に計算するもの。

最後に

統計的仮説検定において，有意差があった場合に（＊）とp値だけを書く論文が多い。しかし，今日挙げた理由などから，この傾向は改められるべきである。

効果量を示す習慣を身につければ，（1）有意水準の高さを要因の操作による効果の高さとする誤解，（2）統計的に有意は認められていないが意味のある平均値の差の見落とし，等の危険性が少なくなるだろう。

参考文献

山田剛史･村井潤一郎(2004) よくわかる心理統計　ミネルヴァ書房

吉田寿夫(1998) 本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本　北大路書房

森敏昭･吉田寿夫(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

水本篤･竹内理(2008) 研究論文における効果量の報告のために－基礎的概念と周囲点「英語教育研究」31, 57-66

印南洋･Hauser, E(2004) The statistical power of language acquisition research: A review.

http://www7b.biglobe.ne.jp/~koizumi/Innami/041028Hauser.doc

中山真孝(2009) 分散分析と効果量 (2009年度心理データ解析レジュメ)