

[教材研究]

Rによる効果量分析の実際

How to calculate effect sizes with R

盛崎俊浩[†]Toshihiro Morisaki[†]

†久留米大学 心理学研究科

† Graduate school of Psychology Kurume University.

1. はじめに

近年、学術論文において統計的検定を行なう際に、当該の統計量とともに効果量や検定力¹の報告を義務付けられることが多くなってきた。APA（アメリカ心理学会）は既に効果量の報告を義務付けており、日本の各心理系学会においても効果量と検定力は重要視され始めている。

効果量分析を行なえるソフトウェアとしてG*Powerがある。これはフリーソフトであり、しかもGUI(Graphical User Interface)で操作ができるため使い勝手がよい。しかしながら、GUIであるがゆえに操作の記録を残しにくく、また自分で式を入力するわけではないため、欲しい出力がいつも得られるわけではないという難点がある。これに対して、統計解析全般を行なえるソフトウェアとして普及し始めているGNU R（以下、Rと表記）は、式や関数を自分で入力する手間と引き換えに、上記2つの難点がない。特に、どのような操作をしたのかという記録が残せることは、分析のやりなおしや追試を行う上で非常に役立つ。しかしRはCUI(Character User Interface)がデフォルトであり、使うことに困難を感じる向きもあると思われる。そこで本稿では、Rを用いて初歩的な効果量分析のチュートリアルを行い、応用段階への足掛りとなることを目指した。

2. 効果量・検出力・サンプルサイズ・有意水準の関係

統計検定は一般に、1)帰無仮説と対立仮説を立て、2)帰無仮説の上で手持ちのデータがどれほどの確率で出現するかを推し測り、3)それがある有意水準よりも小さければ帰無仮説を捨て、対立仮説を採用するという手順を踏む。

例えば、ある病気に対して新しい治療法が開発されたとする。統計検定では、「新治療法は今までの治療法と同じくらいの効果しかない」という帰無仮説を立て、その帰無仮説が正しいとすると実際に調査で得られたデータはどれくらいの確率で起こるかを見つめるわけである。そして、その確率が一定値以下の場合、帰無仮説は間違っていたと判断することにし、一定値よりも大きい場合は新治療法と今までの治療法に意味のある差があるかどうかは判断を保留する。この一定値のことを有意水準と呼ぶ。

有意水準は通常0.05や0.01などに設定され、算出された統計量が帰無仮説の上で正しい（真である）確率がこれよりも小さい場合は帰無仮説が正しくないだろうと考えるのである。帰無仮説が棄却されるかどうかは効果量と検定力の大きさに依存する。

効果量とは、独立変数が従属変数に対して影響を及ぼしている程度や、変数間の関係の大きさといった、効果の大きさをあらわす統計的な指標のことであり、言い換えれば、帰無仮説が正しくない程度を量的に表す指標のことであり[1]。先程の新治療法と今までの治療法の例で言うならば、旧治療法での回復までの日数が平均50日、新治療法ではそれが平均35日で標準

偏差はどちらも5日であったとすると、平均の差を共通な標準偏差で割った $|(50 - 35)| / 5 = 3$ が効果量である。これは、旧治療法と新治療法での回復までの平均日数の差は、共通な標準偏差3つ分あるということを表わしている。本稿では、2群の差の効果量を表わす指標として Hedges の g と d_b を、分散分析における F 統計量に対する効果量として Cohen の f および f^2 を扱う。

検定力とは、上で述べた影響の程度や、関係の大きさを検出する度合いのことであり、 $1 - \beta$ という記号で表わされる。この β は第2種の過誤を犯す確率であり、真実は対立仮説が正しいのに帰無仮説を棄却しない確率のことを指す。先程の例で言えば、新治療法と旧治療法の回復までの平均日数に差があるのに、差があるとは言えないと結論してしまう確率である。この β が小さいほど、言い換えれば検定力が大きいほど、小さな効果でも有意であるとして検出しやすくなる。検定力は、サンプルサイズが大きい程大きくなる。

以上のように、効果量・検定力・サンプルサイズ・有意水準は互いに影響を与えあっており、検定力の分析はこれを利用する。

3. R 上での分析の実際

3.1. 準備

データの準備

データを R の画面上で入力することもできるが、図1のように手元にあるデータはあらかじめ表計算ソフトなどで作成しておき、R で読み込める形式で保存しておく方がよい。カンマ区切りテキスト(csv)形式が無難であるⁱⁱ⁾。

	A	B	C
1	name	pre	post
2	aker	7	-10
3	micky	11	1
4	mercy	0	3
5	kitty	-4	-5
6	manbow	4	3
7	fummy	10	1
8	tommy	9	-9
9	mary	6	-2
10	summy	6	6
11	satcher	5	-22
12	noa	9	16

図1 表計算ソフトへの入力例

R への入力

R で分析をするには、R の入力画面の「>」のマークに続いて式や関数を入力する（以降、この文章の中で特に断りのない限り「>」が最初にあるタイプライタ書体の部分はユーザーが入力する箇所を示す）。式や関数を全て入力した後に Enter (Return) キーを押すと結果が出力される。関数の間違いがあるとエラーメッセージが出されるが、閉じ括弧の数が足りないだけだと行頭に「+」が表示されて残りの入力ができる。ただ、R の画面上でする入力するより、

一旦他のテキストエディタに書いておき、それをRの入力画面にコピー&ペーストするのがよい。入力間違いを訂正しやすいからである。

カレントディレクトリの変更

Rを起動したら、分析をする前にRが参照するディレクトリを作業用の場所に変えておくとよい。これは`setwd()`関数（以下、文中では関数を“**関数名()**”の形式で表記する）やRcmdrの「フォルダの移動」などで行えるが、ここでは関数での指定方法を述べておく。例えば、図の内容をcsvで保存したデータがC:¥hogehoge¥fugafuga¥kokoにあり、そこを作業用の場所（カレントディレクトリ）として使いたい場合は、

```
> setwd("C:/hogehoge/fugafuga/koko")
```

と入力する。

データファイルの読み込み

カレントディレクトリにある`sampledata.csv`というファイルをRに読み込んで`data`という変数名で使えるようにするⁱⁱⁱには、

```
> data <- read.csv("sampledata.csv")
```

とする。`read.csv()`を使うと、Rのなかではデータフレームという形式になり、一番最初の行は列名ラベルとして読み込まれる。読み込んだラベル名で列を指定したいときは、基本的に`data$name`や`data$pre`という風に**変数名\$ラベル名**と入力する。

パッケージの読み込み

また、本稿ではpwrパッケージを使う。パッケージを使うには、Rにあらかじめ読み込んでおく必要がある。pwrパッケージを読み込むには、

```
> library(pwr)
```

と入力する。入力したすぐ次の行に「>」が再度表示されたら、pwrパッケージの読み込みは成功である。もしエラーメッセージが表示されたら、pwrパッケージがインストールされていないか、または何か他の問題がある。pwrパッケージがダウンロードされていない場合は、

```
> install.packages("pwr")
```

としてpwrパッケージをインストールし、あらためて`library(pwr)`と入力する^{iv}。

3.2. 2群の平均値の差の分析でのサンプルサイズ・効果量・検定力

3.2.1. サンプルサイズの決定

調査や実験の前に、どれほどサンプルを集めるべきかを検討する。まず、対応のない2群の平均値の差を検討することを念頭にサンプルサイズを考える。Cohen [2]は母集団における差の効果量の値が0.2を小さな効果量、0.5を中程度の効果量、0.8を大きな効果量と言っており、ひとまずこの基準に沿って検定力を0.8としてサンプルサイズを決めるのがよいだろう^v。

そこで中程度の効果量 (0.5) を見積もる場合, R に

```
> power.t.test(n=NULL, delta=0.5, sig.level=0.05, power=0.8)
```

と入力する。power.t.test()は, t 検定の効果量, 検定力, サンプルサイズ, 有意水準について分析してくれる関数である。ここで, n がサンプルサイズ, delta が効果量 (0.5), sig.level が有意水準 (0.05 : 5%), power が検定力 (0.8) の指定であり, サンプルサイズをなし (NULL) に指定している^{vi}。すると,

Two-sample t test power calculation

```
n = 63.76576
```

```
delta = 0.5
```

```
sd = 1
```

```
sig.level = 0.05
```

```
power = 0.8
```

```
alternative = two.sided
```

NOTE: n is number in *each* group

と表示される筈である^{vii}。ここで n が 63.76576 と計算されており, 先の条件で有意な差が観察されるためには少くとも 1 群あたりのサンプルサイズが 64 必要なことがわかる。

上で指定した効果量は Cohen [2]の基準で言うところの中程度を想定している。大きな効果があると予想する場合は power.t.test()関数の delta=0.5 の箇所を delta=0.8 と変えて実行すると, n=25.52463 と出力されて 1 群あたり 26 人で済むことがわかる。なお, 人数だけを出力させたい場合は

```
> power.t.test(n=NULL, delta=0.5, sig.level=0.05, power=0.8)$n
```

とする。

対応のある 2 群の平均値を比較することを考える場合は,

```
> power.t.test(n=NULL, delta=0.5, sig.level=0.05, power=0.8, type=paired)
```

である。type=paired の箇所が新しく加わっていることに注意されたい。すると, n=33.3672 と計算されてサンプルサイズは 34 必要なことがわかる。こちらも, delta と power の値を色々に変えるとそれに応じて出力される n が変化する。

3.2.2. t 検定の実行

前の節のようにしてサンプルサイズを決定した後, 実際に調査・実験を行い, 得られたサンプルのデータから有意性検定を行う。

Level 0 と Level 1 の 2 群で以下のような 22 人ずつのデータが得られたとする。

Level 0: -2, 1, 4, -7, 1, 5, 5, 4, -3, 5, 9, 7, 7, 11, 0, -4, 4, 10, 9, 6, 6, 5

Level 1: 9, 10, -9, -4, -7, 7, 3, 3, -6, 1, 16, -11, -10, 1, 3, -5, 3, 1, -9, -2, 6, -22

このデータを表計算ソフトで図2のように入力し、`betweendata.csv` という名前で csv 形式に保存した。

	A	B	C
1	level0	level1	
2	-2	9	
3	1	10	
4	4	-9	
5	-7	-4	
6	1	-7	
7	5	7	
8	5	3	
9	4	3	
10	2	6	

図 2 2群のデータの表計算への入力 (一部)

これを

```
> data <- read.csv("betweendata.csv")
```

として R に読み込み `data` という名前で使えるようにした。

Level 0 の平均は 3.8, 標準偏差は 4.72 であり, Level 1 の平均は -1.0, 標準偏差は 8.55 であった。R で対応のない平均値の差の分析を行なう関数は `t.test()` である。`t.test(一つ目の変数名, 二つ目の変数名)` という風に入力する。

```
> t.test(data$level0, data$level1)
```

と実行すると以下のように出力される。

Welch Two Sample t-test

```
data: data$level0 and data$level1
```

```
t = 2.2917, df = 32.707, p-value = 0.0285
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
 0.5341326 9.0113220
```

```
sample estimates:
```

```
mean of x mean of y
```

```
3.772727 -1.000000
```

R ではデフォルトで Welch の検定が行なわれ, 検定統計量 (`t`), 自由度 (`df`), 有意確率

(p-value), 95%信頼区間(confidence Interval), 平均値 (mean) が出力されている。等分散を仮定した t 検定の結果を出力させたい場合は,

```
> t.test(data$level0, data$level1, var.equal=TRUE)
```

のように「,」で区切って var.equal=TRUE を挿入する。この結果は,

```
> resultbetween <- t.test(data$level0, data$level1, var.equal=TRUE)
```

のように変数名を付けて使えるようにしておくと効果量の算出に役立つ。

この data を対応のあるデータとして分析し, resultwithin と変数名を付けて使えるようにするには,

```
> resultwithin <- t.test(data$level0, data$level1, paired=TRUE)
```

とする。

3.2.3. 効果量の計算

対応のない 2 群の平均値の差の効果量は, 母集団において 2 群の平均値の差を共通な標準偏差で割ったものとして定義される。これを推測する指標の一つが Hedges の g と呼ばれるものであり, t 値を対応のない t 検定での統計量, n_1 , n_2 をそれぞれ第 1 群, 第 2 群のサンプルサイズとして

$$g = |t\text{値}| \cdot \sqrt{\frac{(n_1 + n_2)}{(n_1 \cdot n_2)}}$$

という式で表現される[4]。この式の t 値は等分散を仮定したものだから, 先程の R の出力を保存した変数 resultbetween を使って,

```
> abs(resultbetween$statistic)*sqrt((n1+n2)/(n1*n2))
```

と入力すると

```
t
```

```
0.6909668
```

として g の値が出力される^{viii} (t と出力されているのは, 計算時に使った変数の出力様式がそのままになっているからである)。

一方, 対応のある 2 群の平均値の差の効果量は, 2 群で対応するサンプルの差得点を求め, その差得点の平均を差得点の標準偏差で割ったものとして求めることができる[1]。この効果量 (d_D と呼ばれる) は, t 値を対応のある t 検定での統計量, n を 1 群あたりのサンプルサイズとして

$$d_D = |t\text{値}| \cdot \sqrt{\frac{1}{n}}$$

という式で表現できる[4]。これも R の出力をを保存した変数 resultwithin を使って,

```
> abs(resultwithin$statistic)*sqrt((22+22)/(22*22))
```

と入力すると

```
t
```

```
0.4955609
```

と出力される。

3.2.4. 検定力の計算

検定力は、対応がなく 2 群のサンプルサイズが異なるときは `pwr` パッケージの `pwr.t2n.test()` を使い、それ以外の場合はサンプルサイズの算出で使ったのと同じ `power.t.test()` を使う。

`pwr.t2n.test()` の使い方は基本的に `power.t.test()` と同じであるが、

```
> pwr.t2n.test(n1=22, n2=22, d=0.6909668, sig.level=0.05, power=NULL)
```

のように、サンプルサイズを `n1` と `n2` の 2 つに分けて入力すること、効果量の指定を `delta` ではなく `d` とすることに注意されたい。対応のない例では検定力 (`power` の欄) が 0.6099282 と出力される。

対応のあるデータの場合、

```
> power.t.test(n=22, delta=0.6909668, sig.level=0.05, power=NULL,
type=paired)
```

という風に入力する。すると出力のなかに `power=0.8706844` とあり、対応のあるデータとして見た場合は十分な検定力を持っていることがわかる。

3.2.5. 明日への分析

効果量が得られると、その効果量で目指す検定力にはどれくらいのサンプルサイズが必要か計算することができる。豊田はこれを「明日への分析」を呼んでおり、次の調査や実験を企画するためのステップであると位置付けている。対応のない場合のみを例示すると、

```
> power.t.test(n=NULL, delta=0.6909668, sig.level=0.05, power=0.8)
```

という風に再度 `n` に `NULL` を、`delta` に得られた効果量を、`power` に `0.8` に指定してやるとよい。すると `n=33.86696` と出力され、有意水準 5% 水準で検定力 0.8 を目指すには 1 群あたり 34 人必要なことがわかる。

3.3. 1 要因分散分析でのサンプルサイズ・効果量・検定力

3.3.1. サンプルサイズの決定

被験者間計画の 1 要因分散分析では、効果量として Cohen の f が使われる。これは母集団において、

$$f = \frac{\text{各群の平均値の標準偏差}}{\text{各群内における標準偏差}}$$

として定義されており、母効果量 f について Cohen [2] は、0.1, 0.25, 0.4 をそれぞれ小さな効果量, 中程度の効果量, 大きな効果量と設定している。

サンプルサイズの決定には、pwr パッケージの `pwr.anova.test()` を使う。3 群の 1 要因被験者間分散分析で、有意水準 5% で効果量を 0.25 (中程度)、検出力を 0.8 とし、5% 水準で有意となることを目指す場合のサンプルサイズは

```
> pwr.anova.test(n=NULL, k=3, f=0.25, sig.level=0.05, power=0.8)
```

として求められる。t 検定の場合と異なるのは、群の数を k で指定すること、効果量を f で指定することである。

入力すると t 検定の場合と同様の出力が得られ、このなかに $n=21.10362$ とあることから、上の条件に必要な 1 群あたりのサンプルサイズは 22 であることがわかる。

一方、被験者内計画の 1 要因分散分析では、効果量として f ではなく f^2 が使われる。この母効果量 f^2 について Cohen [2] は、0.02, 0.15, 0.35 をそれぞれ小さな効果量, 中程度の効果量, 大きな効果量と設定しており、 f をそのまま 2 乗したものよりも厳しくなっている。

これに伴って、サンプルサイズの決定には pwr パッケージの `pwr.f2.test()` を使う。3 群の 1 要因被験者内分散分析で、有意水準 5% で効果量を 0.15 (中程度)、検出力を 0.8 とし、5% 水準で有意となることを目指す場合、

```
> pwr.f2.test(u=2, v=NULL, f2=0.15, sig.level=0.05, power=0.8)
```

と入力して出力のなかの v を u で割って 1 を足すとサンプルサイズが得られる。この場合は $v=64.319$ なので、 $64.319/2 + 1 = 33.1595$ となり、サンプルサイズは 34 と求められる。被験者間の場合と異なるのは、 u で要因の自由度を、 v で誤差の自由度を、 f^2 で効果量 f^2 を指定することである。

3.3.2. 効果量の計算

分散分析は R でも実行可能であるが、使い慣れたソフトウェアを使うことをお勧めする。ここでは、既に以下のような各群サンプルサイズ 15 のデータから表 1 (被験者間計画) あるいは表 2 (被験者内計画) の分散分析表が与えられているものとして話を進める。

A 群 : 12, 14, 9, 8, 10, 19, 7, 15, 9, 13, 15, 12, 10, 14, 14

平均 : 12.1, SD : 3.13

B 群 : 2, 13, 11, 16, 7, 11, 10, 14, 11, 8, 13, 10, 15, 16, 10

平均 : 11.1, SD : 3.58

C 群 : 14, 13, 14, 15, 12, 17, 13, 14, 16, 17, 14, 12, 12, 14, 13

平均 : 14.0, SD : 1.59

表 1 被験者間計画として見た場合の分散分析表

変動要因	平方和	自由度	平均平方	F値	確率
要因A	64.133	2	32.0667	3.5756	0.0368
誤差	376.667	42	8.9683		
全体	440.8	44			

表 2 被験者内計画として見た場合の分散分析表

変動要因	平方和	自由度	平均平方	F値	確率
要因A	64.133	2	32.0667	3.8835	0.0325
個人差	145.467	14	10.3905	1.2584	0.2919
誤差	231.2	28	8.2571		
全体	440.8	44			

先に述べた母効果量 f の標本における推測値 \hat{f} は、

$$\hat{f} = \sqrt{\frac{\text{要因の平方和}}{\text{誤差の平方和}}} = \sqrt{F\text{値} \cdot \frac{\text{要因の自由度}}{\text{誤差の自由度}}}$$

という式で表わされる。この式の最右辺を用いると、 F 値と自由度から標本での効果量が求められるので便利である。これを表1の F 値と要因の自由度および誤差の自由度から

> sqrt(3.5756*2/42)

として計算すると 0.41263 が得られ、大きな効果量に相当することがわかる。

被験者内計画における f^2 も同様に分散分析表の F 値と要因の自由度および誤差の自由度から求めることができる。表2の要因Aの F 値と要因の自由度および誤差の自由度から

> 3.8835*2/28

として計算すると 0.27739 が得られ、やや大きな効果量があることがわかる。

3.3.3. 検定力の計算

効果量 f と f^2 が求まったので、これを使って検定力を計算することができる。被験者間計画の場合、再びpwr.anova.test()を使って

> pwr.anova.test(n=15, k=3, f=0.41263, sig.level=0.05, power=NULL)

と入力すると、出力のなかに power=0.66405 とあり、66%ほどの検定力であったことがわかる。もし、欠損があってそれぞれの群でのサンプルサイズが等しくない場合、 n にはサンプルサイズの平均値を入れる。例えば、A群・B群・C群それぞれのサンプルサイズが12, 15, 9で f が上と同じであったとき、

```
> pwr.anova.test(n=mean(c(12, 15, 9)), k=3, f=0.41263, sig.level=0.05,  
power=NULL)
```

とするか、直接平均値（今回は12）を入れて

```
> pwr.anova.test(n=13, k=3, f=0.41263, sig.level=0.05, power=NULL)
```

とすると、 $\text{power}=0.55245$ となり、ほぼ2回に1回は有意差が得られないほど検定力が減少していることが確認できる。

被験者内計画の場合は、`pwr.f2.test()` を使って、

```
> pwr.f2.test(u=2, v=28, f2=0.27739, sig.level=0.05, power=NULL)
```

とすると、 $\text{power}=0.70079$ となり、検定力は少々低めであったことがわかる。

3.3.4. 明日への分析

得られた効果量と `pwr.anova.test()` を使うと被験者間分散分析での明日への分析ができる。今回の効果量で有意水準5%のときに検出力0.8を目指すとした場合、

```
> pwr.anova.test(k=3, n=NULL, f=0.41263, sig.level=0.05, power=0.8)
```

とした結果、 $n=19.896$ であり、各群にあと5つずつデータが必要であることがわかる。

被験者内分散分析での明日への分析は再び `pwr.f2.test()` を使う。被験者間要因と同様に有意水準5%で検出力0.8を目指すとした場合、まず得られた効果量から

```
> pwr.f2.test(u=2, v=NULL, f2=0.27739, sig.level=0.05, power=0.8)
```

として誤差の自由度 $v=34.896$ を算出する。これを $u=2$ で割って1足すと18.448が得られ、あと4人分被験者必要であることがわかる。

4. 終わりに

以上、t検定と1要因の分散分析を扱う場面で、サンプルサイズを決定し、データを取って検定を行ない、効果量と検出力を吟味し、次の研究に備えて明日の分析をするという一連の手順を示した。ここで挙げた以外の統計検定における効果量の分析は参考文献を参照されたい。

参考文献

- [1] 大久保街亜・岡田謙介, 伝えるための心理統計 効果量・信頼区間・検定力, 勁草書房, 2012
- [2] Cohen, J. Statistical power analysis for behavioral scientists (2nd Ed.) Hillsdale, 1988
- [3] 永田靖, サンプルサイズの決め方, 朝倉書店, 2003
- [4] 豊田秀樹 編著, 検定力分析入門 —Rで学ぶ最新データ解析—, 東京図書, 2009

註

ⁱ Statistical power のこと。検出力とも。

ⁱⁱ csv 形式で保存しても、たまに文字コードと改行コードの関係で読み込めないことがある。そのときは近くにいる詳しい人に尋ねるのがよい。

ⁱⁱⁱ 変数 `data` に「束縛する」、「代入する」、「付値する」などと言う。

^{iv} それでも解決しない場合は、近くにいる詳しい人に尋ねるのがよい。

^v 母数の区間推定の幅からサンプルサイズを設定する方法もある。[3]

^{vi} この場合は、

```
> power.t.test(delta=0.5,sig.level=0.05,power=0.8)
```

としても同様の出力を得られる。

^{vii} 字下げなど見た目が多少変わることがある。

^{viii} `abs()` は絶対値を返す関数である。