

Rで学ぶ検定力

5/22

市村賢士郎

検定力分析を学ぶ意義

- 2013年に発表された*Nature*の新しいガイドラインによると...
「事前に設定した効果量をもとに、サンプル数をどのように決めたか」を明記しなければならない。
→**検定力分析を行うことを求めている!**
- Rを使えば簡単に検定力分析が可能。

本日のメニュー

- 初心にかえって仮説検定を考える
→ 検定の大事な基本を忘れてませんか?
- 検定力分析のいろはを学ぶ
→ 実験を理想に近づけるための事前・事後・明日への分析
- Rを使って検定力分析をしてみよう!
→ 関数を使えばとっても簡単

初心にかえって仮説検定を考える

なぜ仮説検定が必要か?

- 40人の生徒を2つのクラスA・Bに分け、Aでは新学習法(実験群)、Bでは旧学習法(対照群)で授業をしたところ、期末テストで、実験群の平均は60点、対照群の平均は50点であった。
- 実験群の方が対照群よりも平均点が高かったのは客観的事実。
- しかし、新旧の学習法で同じ効果しかないとしても、今回たまたま実験群の平均が高かっただけかもしれない...

そもそも検定とは?

- 平均にはデータから計算される標本平均と、確率変数の期待値である母平均(μ)がある。
- 実験で示したいのは、「実験群の母平均は、対照群の母平均より高い」
- しかし、現実の世界では、母平均はわからない。
- そこで、検定を行う。

「起きにくい」とは

- 検定の結果 $p = .0068$ であった。
→0.68%でしか「起きにくい」ことが起きた。
- 例えば、家から学校に行くまでの死亡率が0.68%だったら恐ろしい！
- 「起きにくい」という表現は状況や主観に影響される。
- そこで心理学では「起きにくい」という状態を表す確率をあらかじめ設定している。
- これが有意水準で心理学ではあくまで慣例的に5%, 1%, 0.1%が使われている。

有意水準

- 1%水準で有意差が見られた。
→「帰無仮説が正しい確率が1%だった。」
- 「帰無仮説が正しいと言う仮定の下で、1%の確率でしか起きないことが起きた」
- 0.1%水準で有意差の方が効果が高い？
- 低い確率の水準のほうが、**帰無仮説を安心して棄却できる。**

差の大きさを見積もる(信頼区間)

- 帰無仮説が棄却されたからと言って、標本平均の差=母平均とは限らない。(点推定)
- そこで信頼水準 $(1 - \alpha)$ で信頼区間を見積もり、結果の解釈に幅を持たせる。(区間推定)
 - 信頼水準が高くなると、信頼区間は広がる。
 - 信頼区間の幅は標本数の平方根に近似的に反比例する。

帰無仮説が採択されたら

- 帰無仮説の採択=採用ではない！
- 帰無仮説を採択後、信頼区間を観察したところ、 $[-7, 29]$ であった。
- 学習法の実験だとすると、29点の差はすごく大きな差。
→「差の有無に関する判断を保留」して標本数を増やし、信頼区間の幅を狭める。
- 信頼区間から**実質科学的な意味での差の有無を判断**することも大切！

差とは何か？

- 5cm差と1cm差はどちらが大きい？
- AさんとBさんの自宅から学校までの距離の差5cmと、0.1cmと1.1cmのダイヤの大きさの差1cmとでは？
- 差はデータの散らばりと見比べることで初めてその大きさを評価できる。
- 差の大きさを表現する指標が、標準化された平均値の差、すなわち**効果量**。

両側検定と片側検定

- 「新学習法のほうが旧学習法よりも効果が高い」という仮定のもと片側検定を行なった。
- 当然、片側検定の方が有意差が出やすい。
- しかし、「新学習法のほうが学習効果が高い」と言うのは実験者の願いにすぎない。
- 片側検定は「**一方が他方より母平均が高くなることは実質科学的にありえない**」場合のみ。
- そういうケースは稀。→検定は両側とする。

有意水準と効果量は別?

- 効果量の大きな実験は検定で有意差が出やすい。
- しかし、効果量が大きくても有意差が出ないこともある。
- なぜなら、**検定結果は効果量と標本数で決まるから**。
- 効果量が大きくて標本数が少ないときは、有意でなくとも信頼区間をみて、「差がない」ではなく「**態度保留**」。

実質的差があって初めて検定

- 100点満点のテスト平均の差が1点でも、各条件3000人ずつとると有意になる。
- しかし、このテストの1点に実質的差はない。
- 差には1.実質的差 2.標準化された差(効果量) 3.有意差(p値)があり、この順番に重要!
- 単純集計で**実質的差がなければ検定する必要なし!**
- 効果量を記すことで、有意水準の高さを効果の高さと誤解したり、有意でないけど重要な差を見逃す危険がなくなる。

頑健性

- 平均差の検定における3つの数理的な前提。
 - 正規性:観測対象の特性値は正規分布に従う
 - 等分散性:実験群と対象群の母分散は等しい
 - 独立性:測定値は確率分布から独立に観測されたもの
- 正規性、等分散性の前提は多少崩れても、分析結果に致命的な悪影響はない。(頑健性)
- 独立性に対しては頑健でないので注意。

独立性の前提

- 想定した母集団から単純無作為抽出でデータを集めれば独立性は保てるが、実質的にほぼ不可能。
- ある学校で実施した実験による学習法の効果は、生徒の学力レベルや地域の教育の特色などと独立とはいえないため、日本の高校生全体に結果を適用して解釈できない。
- 独立でないと考えられる**状況変数について出来るだけ詳しく、厚みのある記述**をつける。
 - 「〇〇の学力レベルの××の特色をもつ地域の生徒による実験です」

検定力分析のいろはを学ぶ

2種類の誤り

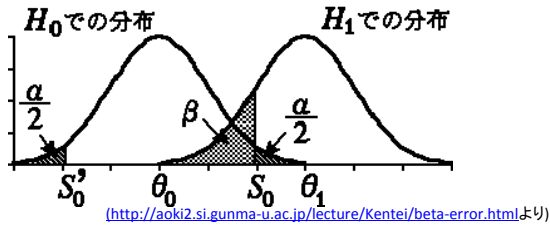
	帰無仮説を採択 (対立仮説を棄却) 帰無仮説を棄却 (対立仮説を採択)	
帰無仮説が真 (対立仮説が偽)	正しい判断 $1-\alpha$	第1種の誤り α
対立仮説が真 (帰無仮説が偽)	第2種の誤り β	正しい判断 $1-\beta$

豊田(2009)より作成

第1種の誤り：確率 α で帰無仮説が真なのに棄却してしまう誤り

第2種の誤り：確率 β で対立仮説が真なのに棄却してしまう誤り

2種類の誤り



α と β はトレードオフの関係にある
 →どちらを重視するかは帰無仮説がネガティブかポジティブかを考慮する
 ネガティブな例：新旧の学習法に差がない→ α を重視
 ポジティブな例：新旧の添加物の毒性に差がない→ β を重視

検定力とは?

- 帰無仮説が偽(対立仮説が真)であるときに有意差が検出される確率。 $1 - \beta$ で表わされる。
- $\alpha \cdot$ 効果量 \cdot 標本数 \cdot 検定力はカルテット。他の2つの条件が同じとき、以下の性質がある。
 - α を大きくすると検定力も大きくなる。
 - 効果量が大きければ検定力も大きい。
 - 標本数を大きくすると検定力も大きくなる。

なぜ検定力が必要か?

- 検定のカルテットのうち、 α は固定、効果量は未知。標本数のみ決められる。
- では、できるだけたくさんデータをとればよいか?
 - 20人でいいのに100人もとるのは研究費の無駄。
 - 統計的に有意でも実質科学的に無意味な差も検出してしまう。
- 第2種の誤りを十分に小さくした状態(検定力を大きくした状態)で、実質科学的に意味のある差を、できるだけ少ない標本数で実行することが検定の理想。
 →理想に近づけるために検定力分析!

どんな検定力分析があるか①

- 事前の分析
 - 実験の計画に先立って行なわれる
 - 効果量、検定力、 α から必要な標本数を決定する。
 - $\alpha = .05, .01, .001$ など
 - 検定力の目安として0.8が多い。標本があまりとれないときは0.7が最低ライン、第1種、第2種の誤りを同等に考える場合は0.95に指定。
 - 効果量はCohenの基準を参照する
 大 0.8 中 0.5 小 0.2 (Cohen's dの場合)

どんな検定力分析があるか②

- 事後の分析
 - 実施した実験を振り返る実験。
 - 実施した実験の検定力(標本検定力)を実際の標本数、標本効果量、 α を使って確認する。
- 明日への分析
 - 実験結果を踏まえて次の実験を企画する分析。
 - 標本効果量、検定力、 α から次の実験の標本数の目安をつける。

ずばり、標本数は?

標本の対応	es = 0.2	es = 0.5	es = 0.8
なし	394	64	26
あり	199	34	15

豊田(2009)より作成

※2群の平均値差の検定力分析
 ※もともと標準的な $\alpha = 0.5$,検定力0.8の場合
 ※対応なしの場合は表の人数 $\times 2$ の被験者が必要

<p style="text-align: center;">Rを使って検定力分析をしてみよう!</p>	<p style="text-align: center;">Rの特徴</p> <ul style="list-style-type: none"> ○ 無料! ○ プログラムコードが世界中に公開されているため、プログラムが洗練され、高い信頼性を持つ。 × GUIが弱く、SPSSのようにマウスでポチポチと言うわけにはいかない。Rコマンドー(Rcmdr)というパッケージを併用すれば多少は補える。
<p style="text-align: center;">Rの特徴</p> <p>△プログラミング言語であるため取っつきにくい。が、Rの場合は比較的易しい言語が用いられる。一度プログラムを書いてしまえば、同じ処理を正確に再現できる(再分析に便利!)。</p> <ul style="list-style-type: none"> ○ 多くの研究者によって、他の統計ソフトに導入されていない最新の統計手法まで、パッケージと言う形で公開され使用できる。 ○ Windows, Mac, Unix全てに対応しており、OS環境が違って同じように統計解析ができる。 	<p style="text-align: center;">Rの特徴</p> <ul style="list-style-type: none"> × SPSSのように結果の出力として、表などが自動的に出てこないため、分析結果に極力手を加えずに流用するという観点では弱い。 ○ 結果の出力をそのままでは利用しづらいものの、強力なグラフィック機能を備えており、非常にきれいなグラフが描け、様々なファイル形式で保存できる。
<p style="text-align: center;">Rをさわってみよう(四則演算編)</p> <ul style="list-style-type: none"> • Rの操作はプロンプト(>)にコマンドを入力して行なう。>の後にコマンドを入力してEnterで実行! • 四則演算は電卓同様にできる <ul style="list-style-type: none"> >3+5 >10-3 >2*3 >20/4 • 平方根や累乗の計算も <ul style="list-style-type: none"> >sqrt(16) #sqrt()は()内の平方根を求める関数 >2^2 	<p style="text-align: center;">Rをさわってみよう(t検定編)</p> <ul style="list-style-type: none"> • クラスAとBの統計学のテスト得点の差を検定してみましょう。 <pre>>クラスA <- c(54,55,52,48,50,38,41,40,53,52) >クラスB <- c(67,63,50,60,61,69,53,58,56,59) #クラスA,クラスBという変数を作り、そこに10人分のデータを格納している。 >t.test(クラスA,クラスB,var.equal=TRUE) #t検定の実行。var.equal=TRUEで等分散の仮定を満たしていることを指示。FALSE(規定値)にする等分散を仮定しないWelchの検定が実行される。</pre>

Rをさわってみよう(おまけ編)

```
>summary(テストA) #テストAの基本統計量の算出
>mean(テストA)    #テストAの平均の算出
>sd(テストA)      #テストAの標準偏差の算出
>var(テストA)     #テストAの標本分散の算出
>hist(テストA)    #テストAのヒストグラムを描く
>plot(テストA,テストB)
  #テストAとBの散布図を描く
などなど、簡単な関数でいろいろな計算ができる!
```

2群の平均値差の検定力分析

- power.t.test関数で一発!
- α 、効果量、標本数、検定力のうちの1つを指定せずに関数を実行すると、指定しなかった引数の値を返す。
- 2群の場合は等分散・等標本数を仮定
- 対応のある場合の標本数nはペアの数を表す。
- 具体的な関数の使い方はhelp関数で確認
help(power.t.test)と入力し、実行するとWebページが開きます。(ただし英語です)

事前の分析(検定力.8の場合)

- 対応なし、es = 0.2
– power.t.test(power=.8, delta=.2, sig.level=.05, type='two.sample')
- 対応なし、es = 0.5
– power.t.test(power=.8, delta=.5, sig.level=.05, type='two.sample')
- 対応なし、es = 0.8
– power.t.test(power=.8, delta=.8, sig.level=.05, type='two.sample')
- 対応あり、es = 0.2
– power.t.test(power=.8, delta=.2, sig.level=.05, type='paired')
- 対応あり、es = 0.5
– power.t.test(power=.8, delta=.5, sig.level=.05, type='paired')
- 対応あり、es = 0.8
– power.t.test(power=.8, delta=.8, sig.level=.05, type='paired')

事前の分析(オーダーメイドで)

- α と β を対等とみなし $\alpha = .05$ 、検定力を.95
– for(i in c(0.2,0.5,0.8)){print(ceiling(power.t.test
(power=.95,delta=i,sig.level=.05,
type='two.sample',strict=T)\$n))}
 - $\alpha = .01$ と厳しく、検定力は.7で我慢
– for(i in c(0.2,0.5,0.8)){print(ceiling(power.t.test
(power=.7,delta=i,sig.level=.01,
type='two.sample',strict=T)\$n))}
- 検定力や α 、効果量を指定してやることでそれに必要なサンプル数を知ることができる!

関数解説

- for文で変数iを作り、効果量にes =0.2,0.5,0.8を代入している。
 - \$nはpower.t.testの出力から標本数nだけ取り出す働きをしている。
 - ceilingでその標本数を整数に繰り上げている。
- ※これらの関数はいっぺんに分析し、かつ結果を見やすくするためのものです。Rに慣れないうちはpower.t.test()だけを使って、一つ一つ結果を出力していくことももちろん可能です。

事後の分析下準備

- 標本効果量を算出しておく必要がある。
- 残念ながら、Rで効果量を算出してくれる関数はない(ハズ)
- もちろん一度自分で関数を定義してスクリプトに保存しておけば、いつでも計算可能。
- ネットを探せば公開されているかも?
- それでもめんどくさい人はwww1.gifu-u.ac.jp/~gupsycho/miyamoto/effectsize.xlsがおススメ。

事後の分析

- 対応なしの標本数が20ずつで等しい実験の結果、標本効果量が0.91だった。
- 有意水準5%,1%,0.1%の検定力がどの程度だったかを推定する。
– `for(j in c(0.05,0.01,0.001)){print(round(power.t.test(n=20,delta=0.91,sig.level=j,type='two.sample',strict=T)$power,3))}`
- **実際のサンプル数や標本平均から自分の実験の標本検定力を知ることができる！**

関数解説

- for文で変数jを作り、危険率として `sig.level=0.05,0.01,0.001` を代入している。
- `$power` は関数 `power.t.test` の出力から検定力だけを取り出す働きをしている。
- `round` で検定力を小数代3位で四捨五入している。
- 実験の状況は `n=20,delta=0.91` で指定している。

明日への分析

- 明日への分析は豊田(2009)で提唱されているオリジナルな分析。
- そこで、まずは関数を定義する。
- 豊田(2009)参照。ソースコードは東京図書の「検定力分析入門」のWebページからダウンロードできる。
<http://www.tokyo-tosho.co.jp/books/ISBN978-4-489-02065-0.html>

明日への分析

- `power.t.tomorrow(効果量,両片=("two.sided","one.sided"),有意水準=c(.05,.01,.001),検定力=c(.7,.8,.9,.95))`
– 効果量、両側(規定値)か片側か、有意水準(規定値=5%,1%,0.1%)、検定力(規定値=0.7,0.8,0.9,0.95)を指定する。
- 実験を重ねた結果、次の実験の効果量の目安が0.9のとき
– `power.t.tomorrow(0.9)`
- **効果量の目安があることでより、妥当なサンプル数を知ることが出来る！**

pwrパッケージを使っていろいろな検定力分析をする。

- 「パッケージ」→「パッケージのインストール」からpwrをインストール
- `library(pwr)`とコマンドを入力して準備完了！

pwrパッケージに含まれる関数

- `pwr.2p.test()` : 2つの比率の差の検定(サンプル数が等しい場合)
- `pwr.2p2n.test()` : 2つの比率の差の検定(サンプル数が異なる場合)
- `pwr.anova.test()` : 一元配置分散分析
- `pwr.chisq.test()` : カイ二乗検定
- `pwr.f2.test()` : 一般線形モデル(分散分析・回帰分析など)

pwrパッケージに含まれる関数

- `pwr.norm.test()` : 正規分布の平均値の検定 (分散が既知の場合)
- `pwr.p.test()` : 比率の検定(1標本)
- `pwr.r.test()` : 相関係数の検定
- `pwr.t.test()` : 平均値に関するt検定(1群,2群, 対応あり)
- `pwr.t2n.test()` : サンプルサイズの異なる独立な2群の平均値に関するt検定

pwrパッケージで検定力分析をやってみよう!

- 基本的にはいずれも、効果の大きさ、サンプルサイズ、有意水準、検定力の4つのうち3つを指定してやることで残りの一つの値を求めることができる。
- 具体的な使い方は`help()`で。
- 明日への分析についてはpwrに含まれないので、豊田(2009)を参照。
- 自分のデータや計画について検定力分析をやってみよう!

参考文献

- 豊田秀樹 編著「検定力分析入門—Rで学ぶ最新データ解析—」2009 東京図書
- 山田剛史・杉澤武俊・村井潤一郎「Rによるやさしい統計学」

Rを学ぶためのおススメWebサイト

- R-Tips :
<http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html>