統計的に有意？
── 帰無仮説検定でわかること・わからないこと

専修大学人間科学部 教授
大久保街亜（おおくぼ　まちあ）

Profile―大久保街亜
2002年，東京大学大学院人文社会系研究科博士課程修了。博士（心理学）。日本学術振興会特別研究員，同海外特別研究員，専修大学文学部講師，准教授を経て，2014年より現職。専門は認知心理学。著訳書は『伝えるための心理統計：効果量・信頼区間・検定力』（共著，勁草書房），『新版　認知心理学：知のアーキテクチャを探る』（共著，有斐閣）など。

Facebookによる感情伝染実験

すこし前の話である。Facebookが，689,003人のユーザーを対象に大規模実験を行った。彼らは投稿記事の内容表示をユーザーに気づかれないよう操作した。あるユーザーにはポジティブな内容が多く表示され，別のユーザーにはネガティブな内容が多く表示された。実験の結果，図1のとおり，表示される投稿記事の内容に従って感情が変化する，つまり感情伝染が生ずることが示された（Kramer, Guillory & Hankock, 2014）。

この大規模実験の結果は，世界中の新聞，テレビ，さらにインターネットなどで大きく取り上げられた。たとえば，2014年7月2日の日本経済新聞によると「肯定的（ポジティブ）な印象を与える投稿を減らしたところ，利用者自身の投稿も否定的（ネガティブ）な内容が増えた。逆にネガティブな投稿を減らすと，ユーザーの投稿ではポジティブな内容が増えたという」と報道された。New York Timesでも2014年6月29日付けで「The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.」（Goel, 2014）という記事が配信された。

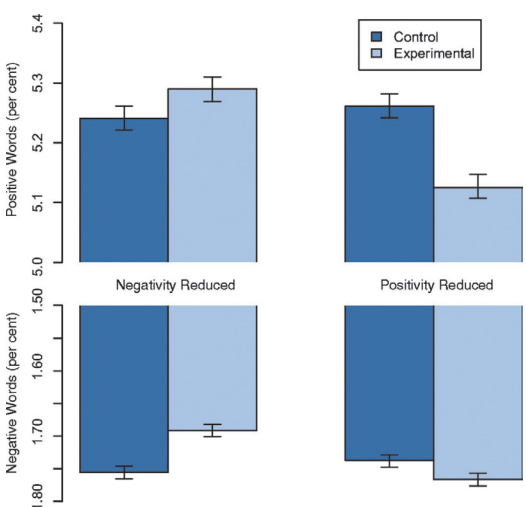

図1　Kramerらの Facebook による感情伝染実験における結果　投稿の操作によって変化したポジティブ語（上）とネガティブ語（下）の割合を示した（Kramer, Guillory, & Hankock, 2014, *Proceedings of the National Academy of Sciences of the United States of America, 111*, Figure 1, p.8789）

この実験はいくつかの点でセンセーショナルであった。Facebookがユーザーを実験動物扱いしたという批判もあった。ユーザーに対して事前にはっきりとした説明がなかったことも問題視された。Facebookが世論誘導を行うのではないかという疑問も呈された。これらの批判を受けFacebookは公式に謝罪した。

感情伝染実験で用いられた統計的手法

この研究では，感情伝染の効果を確認するためにいくつかの統計的手法が用いられた。わかりやすいものは，下位検定として行われた*t*検定の結果だろう。ここで一番効果の大きかったポジティブな内容表示を減らした実験群とそれを行わなかった統制群の比較を取り上げよう。これはメディアでも中心的に取り上げられ

た結果である。この実験群において投稿記事内にポジティブ単語が占める比率は統制群に比べて0.1％少なくなった。この差は統計的に有意だった。論文に書かれた統計量は以下の通りである。t（310,044）= -5.63, $P < 0.001$, Cohen's $d = 0.02$。

t検定の結果に基づくと，この差が偶然で生ずることは，1000回に1回より少ないことになる（$p < .001$）。なるほどたいしたものだ。だが，気をつけなければならない。実験群と統制群のポジティブ単語における比率の違いは0.1％しかない。要するに1000単語あったら1単語少ない程度の差である。Facebookでは10～20語程度の投稿記事が最も多い。そしてほぼすべての投稿記事が100語以下である。そう考えると取るに足らない小さな差と言ってよい。

論文で記載されたCohen's dという効果量も，この差が取るに足らないことを示している。Cohen's dは，二つの平均値の差をプールした標準偏差で割ったものである。つまり，検討対象となる差が標準偏差のいくつ分かを表す指標である。標準偏差を基準とするので，測定方法や単位に依存せず差の大きさを記述できる。さて，上述の通り，この条件での感情伝染の効果はCohen's $d = 0.02$であった。つまり，標準偏差の50分の1の差しかないことになる。

Facebook実験で観察された感情伝染の効果は実に小さなものだった。だが，この実験を報告した論文はもちろんテレビ，ラジオ，インターネットなどの各種報道も感情伝染の効果があったことのみを喧伝した。その効果がとても小さなものであったことを報じることは，ほぼなかった。

統計的手法というブラックボックス

なぜ，効果があったことのみが伝えられるのだろう。そして，その大きさについては，なぜふれられないのだろう。その理由は，統計的手法と関わりがある。統計的手法は，データから結論を導くための一連の手続きである。Facebookによる大規模実験において，感情伝染が生じたという結論が広く流布したのは，彼らが取ったデータが一連の手続きを経て（具体的には帰無仮説検定という統計的手法を経て），効果があるという結論が導かれたためである。

統計的手法によって，私たちはさまざまな現象を捉えることができ，研究の俎上にあげることができる。たとえば，Facebookの投稿をただ眺めているだけで感情伝染が生ずるか判断することは難しい。現象をつかむため，そしてそこから結論を導くために統計的手法が必要になる。統計的手法は，大まかにいうと①問題設定，②記述統計，③推測統計の順に進む。Facebookの実験を例にとると，まず，感情感染が起こるかどうかという問題をたてる（問題設定）。そして，それに適する実験条件を設定し，データをまとめる（記述統計），それらに対して検定や推定を行うことで差，効果，関係性を推測する（推測統計）。この一連の手続きを経ることで，現象を適切に理解できるようになるし，何がしかの結論を導くこともできる。

結論が出るのは良いことだ。しかし，しばしばこの結論だけがひとり歩きをする。当事者やごく一部の専門家や好事家をのぞき，圧倒的大多数は統計的手法に基づく結論だけを知ることになる。一般の人々はもちろん，専門分野の研究者であっても，論文などの一次ソースをすべて精査するわけではない。そして，論文中で紹介された（引用された）情報や，アブストラクトだけを眺めることも多い。しかしそれだけでは多くの場合結論しか知ることができない。大元の論文を丁寧に読まなければ，その研究で用いられた統計手法の詳細はわからないし，その統計手法に基づいて導かれた結論が妥当か正しく判断することは難しい。

また，統計的手法に対する知識が不十分な場合，論文など研究の一次ソースにふれたとしても，実質的には結論しか理解できない。残念ながら統計的手法の理解はしばしば手続き的になりがちだ。手続きを機械的に当てはめるため，統計的手法はブラックボックス化しているという批判が繰り返しなされてきた。ブラックボックス化した統計手法は，データから結論を導く魔法の箱になってしまう。実際の研究において

も，いわゆる「ハウ・ツー」のノウハウだけが流布しているため，手法を機械的に当てはめることもしばしばだ。2条件の平均値を取ったらt検定，3条件を超えたら分散分析。カテゴリカル・データを取ったら，とにかくχ^2検定。こんな具合である。このような浅はかな理解しかなければ，結論をうのみにするしかない。

帰無仮説検定の論理とその問題点

現在，心理学で使用される統計的手法において，帰無仮説検定は支配的なものである。ほぼすべての実証的な心理学研究で帰無仮説検定が用いられる。心理学における統計教育も帰無仮説検定が中心である。Facebookによる感情伝染の実験も帰無仮説検定が用いられていた。

帰無仮説検定とは，検定対象となる効果や差が全くないか，ないとは言えないのか調べるものだ。帰無仮説検定では，効果や差がないという前提のもと，観察された差や効果が生ずる確率を計算する。その確率がとても小さければ効果がないとは言えない。ここで背理法の論理を用い，ないとは言えないのだから，差や効果があるという結論が導かれる。帰無仮説検定で得られる結論は，その論理を考えると「ないとは言えない」という間接的な証拠に基づくことがわかる。

ここで気をつけなければならないことがある。帰無仮説検定で判断できるのは効果や差の「あり・なし」だけだ。その大きさについて全く情報は得られない。帰無仮説検定では，計算された確率が小さければ，検定対象の差や効果を有意であると判断する。この確率をp値と呼ぶ。また，その基準を有意水準と呼ぶ。心理学では慣習的に有意水準は.05（つまり5%）に設定されている。つまりp値が.05を下回るか否かで効果や差の「あり・なし」が決定される。

これは単純でわかりやすい。しかし，差や効果は「あり・なし」の二つにきれいに分類できるとは限らない。差や効果はどちらも全くない状態から非常に大きいあるいは強い状態まで連続し，さまざまな値をとりうる。そして，判断の基準となる確率も連続的である。.05を境目に，確率の意味そのものが劇的に変化するわけではない。それにもかかわらず，有意性の判断で，「あり・なし」の二つに分けてしまう。p値が確率であり，連続的なものであることを考えると，「あり・なし」の単純な二分法が現実を正確に反映するとは言いがたい。

なお，この.05という基準には明確な根拠がない。現代の推測統計を確立させたロナルド・フィッシャーがなんとなくその辺がよいと述べたことがきっかけとなり，慣習的に使い続けられているだけにすぎない。しかも，フィッシャー自身は.05を絶対的な基準とするアイデアには徹頭徹尾批判的だった。

さらに，p値がいくら小さくとも，その差や効果の大きさとは全く関係がない。p値は，いわば観察された差や効果がゼロとは言えない確率である。標本サイズが増えると，大数の法則に従い，差や効果の推定は正確になる。そのため巨大な標本サイズで行う帰無仮説検定では，精度が高くなりすぎて，ほんのわずかな差や効果ですらp値は有意水準を下回ってしまう（これは帰無仮説〔すなわち効果ゼロ〕というきわめて特殊な条件との比較を行うために生ずる）。個人差があり，ノイズが大きい心理学の測定では，比較条件間で測定値が完全に等しいことは実質的にあり得ない。小数点以下のどこかに必ず差はある。この性質のため，心理学関連の領域では，何万人という大規模標本サイズで帰無仮説検定を行うとほぼ必ず有意という結果がでてしまう。ただし，このようなときに有意な効果や差があっても，効果がゼロではないことを示すにすぎない。小数点以下のどこかに差はあるのだから，あまり意味のある情報ではないかもしれない。689,003人という大量の標本サイズで行ったFacebookの感情伝染の実験でも同様のことが起こったと考えられる。

まとめると，帰無仮説検定の結果として得られた結論は，「ないとは言えない」いう間接的な証拠に基づき，差や効果の大きさの情報は含まず，標本サイズに影響され，根拠が明確でない慣習的な基準によって得られたものということになる。列挙してみるとずいぶん心許ない。

基準と怠惰な精神

帰無仮説検定において，有意水準.05という明確な基準があることは，わかりやすさの点で利点がある。しかし，アメリカの疫学者ウイリアム・セジウィックが述べたように「基準とは，考えることを回避させ，怠惰な精神を作る格好の道具である」。しかも，一度基準ができ上がってしまうと，なかなか変えることは難しい。ブラックボックス化して，ハウ・ツーだけが浸透した帰無仮説検定は，データから結論を導く過程において，しっかりと自分の頭で考えない怠惰なデータ分析を生み出しているかもしれない。

ただし，帰無仮説検定を使用しても，基準を妄信せず丁寧にデータを吟味することでこれらの問題をある程度避けることができる。しかし，手続きを機械的に当てはめ帰無仮説検定を行うだけなら避けることは難しいだろう。あるいは，このような性質を逆手にとって，取るに足らない差や効果をあたかも重要なものとして見せることも可能だ。

統計改革とベイズ推定
―― 帰無仮説検定だけに頼らない

帰無仮説検定には多くの問題がある。したがって過度な依存は避けるべきだ。1990年代以降，心理学において用いられる統計的手法には変化が出てきている。具体的には，帰無仮説検定に対する過度な依存を脱して，さまざまな面からデータを評価することが明示的に求められるようになってきた（詳しくは，大久保・岡田〔2012〕を参照）。具体的には，帰無仮説検定におけるp値のみを判断の材料とせず，効果量や信頼区間，検定力などさまざまな指標から包括的にデータを検討することが強く推奨されるようになってきた。効果量の効果の大きさを表す指標である。たとえば，先述したCohen's dが代表的なものだ。一方，信頼区間は区間推定の一種である。平均値のような点だけではなく，区間という範囲をあわせて見ることで，データを多面的に評価できる。すなわち，効果量と信頼区間を考慮することで，効果の大きさとそのばらつきについて情報が得られる。また，検定力を見ることで標本サイズの問題についても留意することができる。つまり，統計改革で推奨される手法を用いることで，帰無仮説検定でわからないこともわかるようになる。

帰無仮説検定の代わりにベイズ推定を行うことも推奨されている。ベイズ推定では，帰無仮説のような特殊な仮説が必要でないし，調べたい仮説にかかわる分布を直接求め，確率を計算できる。また，データを更新することで推定は正確になるため，標本サイズの増大が致命的な問題にならない（ベイズ推定の利点については佐伯・松原〔2000〕が簡便にまとめている）。

もっとも統計改革で推奨される手法やベイズ推定を利用しても，その手続きだけを機械的に取り入れるだけでは十分でない。それだけでは帰無仮説検定に対する過度の依存と同質の問題がいずれ生ずる。データから結論を得る手続きについて，その内容や論理をきちんと理解する必要がある。それができて，データから適切な結論を導くことができる。また，データに基づいて他者が出した結論を正しいか判断できるようになる。

文　献

Goen, V.（2014）Facebook tinkers with user's emotions in news feed experiment, stirring outcry. The New York Times. <http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html>

Kramer, A.D.I, Guillory, J.E. & Hancock, J.T.（2014）Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 8788-8790.

日本経済新聞（2014）米フェイスブックが謝罪, 投稿操作し心理実験. 日本経済新聞 2014年7月2日 <http://www.nikkei.com/article/DGXNASGM0100I_S4A700C1000000/>

大久保街亜・岡田謙介（2012）『伝えるための心理統計：効果量・信頼区間・検定力』勁草書房

佐伯胖・松原望（2000）『実践としての統計学』東京大学出版会