Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie
Metody analizy danych

916

ISSN 1898-6447
Zesz. Nauk. UEK, 2013; 916: 5–12

Małgorzata Rószkiewicz
Instytut Statystyki i Demografii
Szkoła Główna Handlowa w Warszawie

Problemy analityczne metaanalizy – efekt procesu badawczego*

Streszczenie

Wraz z poszerzaniem się problematyki badań społecznych poszerzają się również obszary realizowanych badań empirycznych. W związku z licznymi badaniami często odnoszącymi się do tych samych zagadnień pojawia się pytanie, jak prowadzić analizę sumaryczną, pozwalającą ustalić, co w zasadzie wiadomo na temat przebadanych obszarów. Poszczególne badania są realizowane na próbach losowych, co sprawia, że uzyskiwane wyniki nie są jednakowe. Ich zakres zmienności opisuje wariancja losowa estymatora. W artykule rozważono problemy związane z analizą fluktuacji wyników badań, których rozstrzygnięcie wyznacza paradygmat statystycznej metaanalizy.

Słowa kluczowe: metaanaliza, efekt badania, testy wiązane, bezpieczna liczba badań.

1. Wprowadzenie

Wśród najczęściej przytaczanych powodów sięgania do metaanalizy wskazuje się wzrost liczby badań i ich wyników w ramach poszczególnych domen oraz oczekiwanie, że procedura ta spowoduje wzrost mocy statystycznej [Becker i Cohen 2003]. Obok opinii bardzo pozytywnych wyrażano jednak również poglądy sceptyczne, podważające walory poznawcze tego podejścia badawczego [Hadges i Olkin 1985]. W literaturze tematu brak jest jednej, powszechnie uznanej definicji metaanalizy. Na potrzeby niniejszego opracowania przyjęto, że metaanaliza w ujęciu

* Opracowanie przygotowano w ramach grantu NCN nr UMO-2011/01/B/HS4/00970.

statystycznym polega na wypracowaniu ogólnej konkluzji z wyników pochodzących z wielu badań. Kluczową jej częścią jest ocena spójności analizowanej serii wyników. Jeśli bowiem nie wiadomo, na ile są one spójne, to nie wiadomo, w jaki sposób dokonywać ich uogólnienia. Z tego względu ważne jest rozstrzygnięcie, czy odmienności obserwowane między poszczególnymi badaniami są powodowane efektami systematycznymi, czy też stanowią jedynie artefakty wynikające ze zmienności losowej poszczególnych badań. Celem artykułu jest omówienie ścieżek postępowania w diagnozowaniu efektu procesu badawczego oraz ocena ich skuteczności.

2. Metody oceny efektu procesu badawczego

Diagnozowanie homogeniczności wyników badań jest na ogół dokonywane za pomocą testu wykorzystującego statystykę zaproponowaną przez W.G. Cochrana, opartą na rozkładzie chi-kwadrat [Higgins i Thompson 2002, Rószkiewicz 2009]. Test ten zwiększa swą moc, wykazując nadmierną skłonność do wskazywania na heterogeniczność wyników, gdy liczba badań jest znaczna, a szczególnie gdy badania są realizowane na dużych próbach. Moc ta jest zależna również od wariancji wewnątrzgrupowych wyników, czyli poziomów błędu średniokwadratowego każdego wyniku. Może się zatem zdarzyć, że mimo iż po przeprowadzeniu testu o homogeniczności serii wyników nie będzie podstaw do podważania ich spójności, wystąpi widoczna (znaczna) wartość ich wariancji międzygrupowej. Problem diagnozowania homogeniczności wyników z rozważanej serii badań jest zatem konsekwencją niedoskonałości procedur weryfikujących założenie o ich spójności.

Najstarszymi podejściami rekomendowanymi w ocenie istotności różnic między wynikami niezależnych badań są procedury ważone, takie jak *t*-test i *z*-test. Procedury te określa się mianem testów wiązanych (*combined tests*). Poniżej wybrano testy prezentowane przez R.A. Fishera [1932, 1948], S.A. Stouffera *et al*. [1949] oraz B.J. Winera [1971].

R.A. Fisher [1948] zaproponował procedurę wykorzystująca istotność krytyczną dla wyników poszczególnych badań w następującej postaci:

$$\chi^2 = -2\sum_{i=1}^{k} \ln p_i ,$$

gdzie:

p_i – istotność krytyczna dla wyniku i-tego badania,
$i = 1, 2, \ldots, k$,

w której statystyka testująca posiada asymptotyczny rozkład chi-kwadrat o liczbie stopni swobody $\nu = 2k$. Procedura ta jest przedstawiana jako asymptotycznie optymalna w porównaniu do innych metod wiązanych [Kozil i Perlman 1978, Littell i Folks 1973]. Jest z nią związanych jednak wiele ograniczeń [Rosenthal 1984]. Może ona dostarczyć wyników niezgodnych z prostą procedurą testu istotności w sytuacji, gdy większość wyników badań będzie wskazywać na rezultaty zbieżne co do kierunku, ale nieistotne. Z kolei S.A. Stouffer zaproponował, by dla łączenia niezależnych testów realizowanych z uwzględnieniem statystyki *t*-Studenta sumować wartości statystyki testującej *t* i przekształcać do statystyki *Z* według wzoru:

$$Z = \frac{\sum t_i}{\sqrt{\sum n_i}},$$

gdzie:

t_i – wynik testu istotności w badaniu *i*,

n_i – rozmiar próby w badaniu *i*.

Procedura ta opiera się na własności addytywności rozkładu normalnego z wariancją równą liczbie obserwowanych wyników prób. Z racji zbieżności rozkładu *t*-Studenta z rozkładem normalnym procedura ta nie jest odpowiednia dla testów pochodzących z małych próbek. Niemniej procedura Stouffera jest łatwiejsza od procedury Fishera, w której konieczne są przekształcenia logarytmiczne. Procedura zaproponowana przez B.J. Winera jest identyczna z procedurą Stouffera i przyjmuje postać:

$$Z = \frac{\sum t_i}{\sqrt{\sum \frac{df_i}{df_i - 2}}},$$

gdzie df_i oznacza liczbę stopni swobody statystyki testującej wyniki badania *i*.

Oba podejścia są zbieżne wraz ze wzrostem liczby próbek.

Charakterystyczną cechą najstarszych rozwiązań rekomendowanych do oceny efektu procesu badawczego jest traktowanie równorzędnie wyników wszystkich prób niezależnie od ich rozmiarów. Próby mniej liczne z oczywistych względów mogą nie spełniać wymogu reprezentatywności lub spełniać go w ograniczonym zakresie w porównaniu do prób o znacznych rozmiarach. W takich sytuacjach nadanie wszystkim wynikom jednakowych wag nie wydaje się poprawne, gdyż może prowadzić do nadawania nadmiernego znaczenia wynikom o niskiej reprezentatywności. By przełamać ograniczenia wynikające z różnej reprezentatywności prób o różnych rozmiarach F.M. Mosteller i B.R. Bush [1954] zaproponowali

nadanie wag odpowiadających odchyleniom standardowym wynikom agregowanym według procedury Stouffera, czyli:

$$Z = \frac{\sum df_i S_i}{\sqrt{\sum df_i^2}},$$

gdzie S_i jest odchyleniem standardowym wyników w badaniu i.

Zupełnie odmienne podejście do oceny homogeniczności serii wyników badań zaproponowali L.V. Hadges i I. Olkin [1985]. Wykazali oni, że testująca homogeniczność statystyka Q-Cochrana może być podzielona na dwa składniki, analogicznie do dekompozycji odchyleń od średniej, w przypadku danych klasyfikowanych ze względu na wybraną cechę, tj. na składnik między- i wewnątrzgrupowy, gdzie grupy są definiowane przez wartości lub kategorie cech potencjalnie wpływających na wyniki badań. W myśl tego rozwiązania jeśli:

$$Q = \sum_{i=1}^{k} \frac{\left(T_i - \overline{T}_.\right)^2}{\upsilon_i},$$

to:

$$\left(T_{ij} - \overline{T}_.\right) = \left(T_{ij} - \overline{T}_j\right) + \left(\overline{T}_j - T_.\right),$$

gdzie:

T_{ij} – wynik badania i ze względu na j-tą kategorię lub wartość cechy charakteryzującej proces badawczy,

$\overline{T}$ – wartość uogólniona wszystkich wyników badań,

$\overline{T}_j$ – wartość uogólniona z tych wyników badań, które odnoszą się do j-tej wartości lub kategorii cechy charakteryzującej proces badawczy.

Zaproponowana dekompozycja wartości statystyki testującej pozwala zidentyfikować, które cechy badań potencjalnie wpływają na ich wyniki. Każdy z wyodrębnionych w wyniku dekompozycji składników może być testowany oddzielnie. Jeśli cecha badań, ze względu na którą dokonano dekompozycji, byłaby istotnym czynnikiem zakłócającym homogeniczność wyników, to zarówno wynik testu Q-Cochrana powinien wskazywać na niejednorodność analizowanej serii wyników badań, jak i istotny powinien okazać się składnik międzygrupowy statystyki Q. Składnik wewnątrzgrupowy statystyki Q dotyczący zmienności losowej poszczególnych wyników prób powinien być nieistotny. Gdyby składnik wewnątrzgrupowy okazał się istotny, przeprowadzoną procedurę dla wyróżnionego czynnika zakłócającego należałoby uznać za niewystarczającą. Oznaczać to może, że występują jeszcze inne, nieuwzględnione czynniki zakłócające porów-

nywalność wyników. W takiej sytuacji należałoby rozważyć włączenie do analizy dodatkowych zmiennych opisujących potencjalne różnice między badaniami.

Przedstawiona dekompozycja zakłada, że czynniki zakłócające mogą być opisywane przez zmienne nominalne lub dyskretne. Jeśli czynnikami zakłócającymi wyniki badań i odnoszącymi się do procesu badawczego są zmienne o charakterze ciągłym, to ich oddziaływanie może być opisane modelem regresji w ujęciu modelu mieszanego [Konstantopoulos 2006].

Modele mieszane wykorzystywane w metaanalizie należą do grupy ogólnych liniowych modeli mieszanych. Jeśli założyć, że efekty procesu badawczego są opisane za pomocą p predyktorów $X_1, X_2, \ldots, X_p$, które są powiązane liniowo z wynikiem badania, to przykładowy model drugiego poziomu dla wyniku pochodzącego z i-tego badania przyjmuje postać:

$$T_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \eta_i + \varepsilon_i,$$

gdzie:

x_{ij} – poziom j-tego predyktora opisującego działanie zmiennej zakłócającej w i-tym badaniu,

ε_i – składnik losowy próby w badaniu i,

η_i – efekt losowy specyficzny dla i-tego badania, o rozkładzie $N(0, \tau^2)$.

Jeśli wariancja międzygrupowa wyników badań (τ^2) byłaby znana, to współczynniki regresji mogłyby być szacowane ważoną metodą najmniejszych kwadratów, a ich istotność rozstrzygałaby o efektach systematycznych związanych ze specyfiką poszczególnych badań. Na ogół jednak τ^2 nie jest znane – wówczas możliwe są cztery drogi postępowania:

1) oszacowanie τ^2 na podstawie dostępnych wyników badań i wykorzystanie oceny tej wariancji do estymacji współczynników regresji ważoną metodą najmniejszych kwadratów, co umożliwi weryfikację istotności efektów systematycznych poszczególnych badań,

2) łączne oszacowanie wektora współczynników i wariancji τ^2 za pomocą nierestykcyjnej metody największej wiarygodności (REML),

3) przyjęcie, że wariancja międzygrupowa wyników badań wynosi zero (tak jak to ma miejsce w analizie efektów ustalonych), co – jak już wspomniano – może powodować obciążenie wyników, ale ponieważ jest ono tym mniejsze, im liczniejszy jest zbiór dostępnych badań, rozwiązanie takie można rekomendować w przypadku, gdy zbiór wyników badań jest dostatecznie liczny,

4) oszacowanie wektora współczynników regresji dla każdej wartości τ^2 z przedziału prawdopodobnych wartości i użycie średniej ważonej dla tych wyników, przyjmując za wagi wartości odpowiednie do prawdopodobieństw, tak jak ma to miejsce w podejściu bayesowskim.

Wprowadzając metodologię ogólnych modeli mieszanych do procedur metaanalizy, należy odnotować dwie podstawowe różnice między ogólnymi modelami mieszanymi a modelami wykorzystywanymi w metaanalizie. Po pierwsze, w metaanalizie wariancje błędu próby ε_i, co do którego zakłada się, że ma rozkład normalny z wartością oczekiwaną 0 i wariancją υ_i, nie są identyczne w poszczególnych badaniach. W związku z tym założenie, że są sobie równe, jest nierealne. Jak wiadomo, wariancja błędu próby zależy od wielu czynników wynikających z konstrukcji próby, tj. rozmiaru próby i metody losowania. Po drugie, wariancje błędu poszczególnych badań w metaanalizie nie są identyczne, ale są znane. Z tego względu model wykorzystywany w metaanalizie rozważa specjalny przypadek ogólnego liniowego modelu hierarchicznego, w którym wariancje pierwszego poziomu są różne, lecz znane. Szacowanie wariancji międzygrupowej przy ustalonych cechach badań opisanych zmiennymi X_i przebiega wówczas według formuły:

$$\tau^2 = \frac{\chi^2 - k + p}{c},$$

$$c = tr\left(\mathbf{V}^{-1}\right) - tr\left[\left(\mathbf{X}^T\mathbf{V}^{-1}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-2}\mathbf{X}\right],$$

$$V = diag\left(\upsilon_1, \upsilon_2, ..., \upsilon_k\right).$$

Wnioskowanie o istotności wariancji międzygrupowej przy ustalonych cechach badań opisanych zmiennymi X_i opiera się na wartości statystyki χ^2 o rozkładzie chi-kwadrat z liczbą stopni swobody $(k - p)$, którą definiuje formuła:

$$\chi^2 = \mathbf{T}^T\left[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\right]\mathbf{T}.$$

Dokonując metaanalizy, należy pamiętać, że wiele sytuacji badawczych może prowadzić do obciążonych lub błędnych wyników. Należą do nich:

– wykorzystanie tylko pozytywnych wyników, a pominięcie wyników negatywnych,

– traktowanie równoważnie wyników wszystkich badań, które odnoszą się do tych samych pytań badawczych, nawet wówczas, gdy występują między nimi różnice jakościowe,

– wielokrotne wykorzystanie wyników tego samego badania,

– brak gwarancji wysokiego stopnia zgodności lub rzetelności między rangami przy kodowaniu charakterystyk badań, kiedy są podstawowe lub metodologiczne.

Można wskazać kilka sposobów przełamania tych trudności. Problem pomijania w publikacjach wyników, które okazały się nieistotne, i występowanie nadreprezentatywnych wyników pozytywnych w literaturze z danej domeny badań R. Rosenthal [1984] nazwał problemem doboru badań, proponując określenie,

jaka liczba badań jest niezbędna, by potwierdzić hipotezę zerową o nieistotności wyniku uogólnienia. H. Cooper [1979] określił tę liczebność mianem bezpiecznej liczby badań z liczby N (N_{fs} – *fail-safe N*). Jest to dodatkowa liczba badań, niezbędna by przekształcić całkowitą istotność statystyki testu połączonych prób w wartość wyższą niż krytyczna istotność – na ogół 0,05 lub 0,01. Dla $p = 0{,}05$ jest to:

$$N_{fs.0,05} = \left(\frac{\sum Z}{1{,}645}\right)^2 - N,$$

zaś dla $p = 0{,}01$ jest to:

$$N_{fs.0,01} = \left(\frac{\sum Z}{2{,}33}\right)^2 - N,$$

gdzie:

ΣZ – suma wartości statystyki Z dla prób łączonych według procedury istotności Stouffera,

N – liczba prób łączonych.

Tak wyznaczone liczebności określają zatem, ile dodatkowych badań – każde o wyniku nieistotnym lub w sumie dające $\Sigma Z = 0$ – powinno być włączonych, by nie było podstaw do odrzucenia hipotezy zerowej o nieistotności wyniku.

3. Podsumowanie

Warto zwrócić uwagę na korzyści analityczne, jakich dostarcza metaanaliza, a które można uznać za szczególne, gdyż są nieosiągalne w incydentalnych badaniach przekrojowych. Po pierwsze, metaanaliza stwarza możliwość wykorzystania pełnego zakresu informacji wynikowych ze zrealizowanych badań w ramach danej dziedziny, niezależnie od tego, czy wyniki te okazały się istotne statystycznie, czy też nie. Po drugie, umożliwia prowadzenia bardziej szczegółowych podziałów i klasyfikacji poprzez integrację cząstkowych zbiorów danych w jeden zbiór o znacznym rozmiarze. Ponadto wskazane ścieżki postępowania umożliwiające identyfikację czynników zakłócających procesy badawcze i powodujące ich niejednorodność dają szansę modyfikacji postępowania w przyszłych badaniach. Należy podkreślić, że ocena kontekstu badawczego jest możliwa jedynie w metaanalizie, a jej znaczenie dla właściwej interpretacji wyników poszczególnych badań jest nie do przecenienia.

Literatura

Becker B.J., Cohen L.D. [2003], *How Meta-analysis Increases Statistical Power*, „Psychological Methods", vol. 8, nr 3.
Cooper H. [1979], *Statistically Combining Independent Studies: A Meta-analysis of Sex Differences in Conformity Research*, „Journal of Personality and Social Psychology", vol. 37, nr 1.
Fisher R.A. [1932], *Statistical Method for Research Workers*, wyd. 4, Oliver and Boyd, London.
Fisher R.A [1948], *Combining Independent Tests of Significance*, „American Statistician", vol. 2, nr 5.
Hadges L.V., Olkin I. [1985], *Statistical Methods for Meta-analysis*, Academic Press, New York.
Higgins J.P.T., Thompson S.G. [2002], *Quantifying Heterogeneity in a Meta-analysis*, „Statistics in Medicine", vol. 21, nr 11.
Konstantopoulos S. [2006], *Fixed and Mixed Effects Models in Meta-analysis*, Discussion Paper Series, IZA DP, nr 2198, IZA, Bonn.
Kozil J.A., Perlman M.D. [1978], *Combining Independent Chi-squared Tests*, „Journal of American Statistical Association", nr 73.
Littell R.C., Folks J.L. [1973], *Asymptotic Optimality of Fisher's Method of Combining Independent Tests II*, „Journal of American Statistical Association", nr 68.
Mosteller F.M., Bush R.R. [1954], *Selected Quantitative Techniques* [w:] *Handbook of Social Psychology*, red. G. Lindzey, vol. 1, Addison-Wesley, Cambridge.
Rosenthal R. [1984], *Meta-analysis Procedure for Social Research*, Sage, Beverly Hills.
Rószkiewicz M. [2009], *Możliwości i ograniczenia metaanalizy w obszarze ilościowych badań marketingowych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 96, Wrocław.
Stouffer S.A. *et al*. [1949], *The American Soldier: Adjustment during Army Life*, vol. 1, Princeton University Press, Princeton.
Winer B.J. [1971], *Statistical Principals in Experimental Design*, wyd. 2, McGraw-Hill, New York.

Analytical Problems of Meta-analysis – the Effect of Heterogeneous Studies

The growing number of issues in social science research has led to a correspondingly large body of research studies. The sheer volume of research related to the same topics poses a question as to how to organise and summarise these findings to identify and use what is known and focus research on promising areas. Effect size estimates from the studies are not identical. This is to be expected because the estimates are based on data from samples, and random variations due to sampling should introduce fluctuations into estimates. This paper introduces methods for exploring these fluctuations as a paradigm for explorations in meta-analyses generally.

Keywords: meta-analysis, survey effect, combined tests, fail-safe *N*.