Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 304 · 2016

Informatyka i Ekonometria 7

Grażyna Trzpiot

Uniwersytet Ekonomiczny w Katowicach
Wydział Informatyki i Komunikacji
Katedra Demografii i Statystyki Ekonomicznej
grazyna.trzpiot@ue.katowice.pl

ROZWAŻANIA O *P-VALUE*

Streszczenie: Celem artykułu jest przybliżenie problemów związanych z powszechnym wykorzystywaniem oprogramowania statystycznego i interpretowaniem wyników badań. Badania prowadzone w naukach ekonomicznych różnią się założeniami i celem od analiz prowadzonych w naukach medycznych czy w naukach społecznych. Wykorzystujemy podobne metody badawcze, choć nie zawsze widzimy zasadnicze różne w podejściu do wykorzystywanych modeli statystycznych, zwłaszcza w warstwie spełnienia założeń o rozkładach zmiennych losowych. Poszukujemy szybkich i pewnych interpretacji wyników, co rodzi zagrożenia dla całości prowadzonych badań.

Słowa kluczowe: błąd I rodzaju, błąd II rodzaju, *p-value*.

Wprowadzenie

Celem artykułu jest zwrócenie uwagi na wykorzystywane powszechnie w badaniach naukowych różnorodne narzędzia informatyczne, które zastępują tradycyjne postępowanie analityczne. Powstają nowe środowiska badawcze, które generują wyniki badań w określony sposób. Wielokrotne postępowania, brak sprawdzania założeń stosowanych metod, a w tym wykorzystanie wartości *p-value* jako narzędzia testowania hipotez statystycznych prowadzi do nowej generacji popełnianych błędów w badaniach statystycznych. Wiele osób prowadzących analizy statystyczne (analizę danych) nie zawsze poprawnie interpretuje wyniki lub nadużywa pojęcia *p- value* do weryfikacji określonych hipotez.

W artykule pisma „ScienceNews" [Siegfried, 2014], opublikowanym 7 lutego 2014 r., wskazuje się, że: „metody statystyczne testowania hipotez […] mają więcej wad niż polityka prywatności Facebook'a". Tydzień później statystyk i blogger Jeff Leek (*Simply Statistics*) odpowiedział: „Problemem nie jest to, że

ludzie niepoprawnie wykorzystują wartości *p-value*. […] chodzi o to, że większość analiz statystycznych przeprowadzana jest przez ludzi nieodpowiednio przygotowanych do prowadzenia takich analiz" [Leek, 2014]. W tym samym tygodniu statystyk i pisarz naukowy Regina Nuzzo opublikowała artykuł w czasopiśmie „Nature" zatytułowany *Metoda naukowa: błędy statystyczne* [Nuzzo, 2014].

1. Podejście definicyjne

Prowadząc akademicki wykład, głosimy, że statystyka to nauka o metodach badania prawidłowości występujących w zjawiskach masowych. Obserwujemy w badaniach dwa poziomy oddziaływania badanych zmiennych (czynników). Pierwszy poziom wynika z własności badanego zjawiska. Przyjmujemy, że oddziaływanie jest jednakowe na wszystkie elementy badanej zbiorowości statystycznej – stanowią badaną prawidłowość[1]. Drugi poziom obserwacji wskazuje na zakłócenia, odchylenia od przyjętych założeń[2]. Dodatkowe czynniki losowe mają wpływ, który nie może być oceniony jednoznacznie. Wpływy nie są trwałe ani co do siły, ani co do kierunku oddziaływania. Stosowanie metod statystycznych jest powszechne w badaniach w różnych dyscyplinach naukowych. Poszukujemy dzięki nim ważnych czynników w zjawiskach masowych. Powinniśmy zatem dobrze określić badaną zbiorowość statystyczną (populację). Ze zbiorowości statystycznej losujemy do badania podzbiór – próbę statystyczną[3], a następnie na podstawie próby wnioskujemy o **populacji**. Zarówno populacja, jak i próba powinny być **jednorodne**. Przyjmuje się, że populacja jest jednorodna wtedy, gdy wszystkie jej elementy pozostają pod wpływem działania tych samych głównych czynników[4].

Przy weryfikacji hipotez z wykorzystaniem testów istotności rozważa się dwa błędy. Błąd pierwszego rodzaju polega na odrzuceniu hipotezy prawdziwej (*false positive*). Zakładamy prawdopodobieństwo popełnienia tego błędu – to prawdopodobieństwo jest nazywane *poziomem istotności* i oznaczane przez α[5]. Błąd drugiego rodzaju polega na przyjęciu hipotezy fałszywej (*false negative*) i oznaczany jest β. Wymienione prawdopodobieństwa dotyczą błędów w procesie decyzyjnym, nie oznaczają prawdopodobieństw prawdziwości hipotez zerowej i alternatywnej. Przy stosowaniu testów istotności można podjąć jedną z dwóch decyzji: odrzucić hipote-

[1] Zapisując model zjawiska, mówimy o składniku systematycznym.
[2] Zapisując model zjawiska, mówimy o składniku losowym.
[3] O metodach losowania próby traktuje metoda reprezentacyjna.
[4] Metody taksonomiczne obejmują procedury kontroli jednorodności populacji.
[5] Najczęściej przyjmowaną wartością jest 0,05.

zę zerową, a przyjąć hipotezę alternatywną albo przyjąć, że nie ma podstaw do odrzucenia hipotezy zerowej. Decyzję o odrzuceniu hipotezy zerowej podejmujemy na podstawie wyniku porównania empirycznej wartości statystyki testowej z wartością krytyczną odczytaną z tablic rozkładu statystyki testowej.

W opisanym zwięźle testowaniu hipotez pochodzącym od Neymana i Pearsona znajdujemy również podejście zaproponowane przez Fishera, łatwe do obliczenia *p-value*, powszechnie stosowane w oprogramowaniu statystycznym, nie zawsze jednak dobrze interpretowane. Decyzję o odrzuceniu hipotezy zerowej możemy podjąć, porównując wartość p (*p-value*) z poziomem istotności α.

Za A. Sokołowskim [2004] podamy trzy definicje *p-value*:

1. Pole pod funkcją gęstości rozkładu prawdopodobieństwa statystyki testowej obliczone od empirycznej wartości tej statystyki w kierunku wskazanym przez hipotezę alternatywną. Pole to może być jednoczęściowe (przy jednostronnej hipotezie alternatywnej) lub dwuczęściowe (przy hipotezie dwustronnej).
2. Prawdopodobieństwo uzyskania wyniku bardziej przeczącego hipotezie zerowej niż ten wynik, który właśnie uzyskaliśmy.
3. Najostrzejszy poziom istotności, przy którym możemy odrzucić testowaną hipotezę na podstawie danych empirycznych, które posiadamy.

Ponieważ obecnie przy stosowaniu oprogramowania obliczenia wykonywane są automatycznie i ten wynik mamy najczęściej jako wynik testu istotności, możemy podjąć decyzję o odrzuceniu hipotezy zerowej. Niewłaściwa interpretacja *wartości p* (*p-value*) to uznawanie jej za prawdopodobieństwo prawdziwości hipotezy zerowej.

2. Meta-analiza

Meta-analiza jest procedurą statystyczną służącą do łączenia danych z wielu badań. Gane Glass wprowadził ten termin w 1976 r. w odniesieniu do statystycznej analizy zbioru wyników analiz z indywidualnych badań w celu integracji wniosków [Glass 1976, s. 3-8]. Meta-analiza ma mocne strony oraz ograniczenia. Niemniej jednak jest ona teraz standardowym narzędziem do dostarczania powtarzalnych podsumowań w naukach społecznych, medycynie, edukacji i innych dziedzinach wiedzy.

Jeżeli wielkość wpływu pewnego badanego efektu[6] (np. skuteczność leczenia) jest zgodna w wielu badaniach, meta-analizy mogą być używane do identy-

[6] Efektu lub też czynnika lub dodatkowej zmiennej.

fikowania tego wspólnego efektu. Gdy oddziaływanie zmienia się, meta-analizy mogą być stosowane do określenia przyczyny różnic. Decyzje o ważności hipotezy nie mogą opierać się na wynikach jednego badania, ponieważ wyniki te zazwyczaj nie są zgodne, wahają się pomiędzy kolejnymi badaniami. Potrzebny jest mechanizm do syntezy danych pomiędzy badaniami. Wykorzystane były do tego celu opinie opisowe, ale taki przegląd opinii jest w dużej mierze subiektywny (różni eksperci mogą dojść do różnych wniosków). Meta-analiza pozwala na zastosowanie obiektywnych modeli, podobnie jak w pojedynczym badaniu, ponadto może być wykorzystywana do dowolnej liczby badań.

Zatem tak jak w zwykłych badaniach empirycznych muszą być dopracowane kryteria włączania i wyłączania obserwowanych czynników. Wyznaczamy reguły, w jaki sposób wyniki badań będą uogólniane na populację. Jednym z celów każdej meta-analizy jest zebranie reprezentatywnej próbypodstawowych badań, które zapewniają zdefiniowane kryteria ich doboru. Meta-analizy można prowadzić, wykorzystując jeden ze statystycznych modeli: model z efektem stałym[7] lub model z efektem losowym[8]. W podejściu przyjmującym model z efektem stałym wyniki wszystkich badań opisują tę samą rzeczywistą wielkość wpływu efektu, a różnice wynikają z błędu doboru próby. Efekt łączny jest wówczas wyznaczany jako średnia ważona wpływu efektów, przy czym wagi są wyznaczane jako odwrotności wariancji odpowiednich badań.

W modelu z efektami zmiennymi rozważamy problem taki, że nieznane rzeczywiste efekty różnią się pomiędzy badaniami. Zakładamy zazwyczaj, że mają one rozkład normalny. Na obserwowalny wynik wpływu badanego efektu składa się wówczas ogólny średni efekt, odchylenie rzeczywistego efektu w badaniu od średniego efektu i odchylenie wartości obserwowanej od rzeczywistego efektu w badaniu, związane z błędem losowym próby. Efekt łączny jest wyznaczany na podstawie średniej ważonej efektów, a wagi są wyznaczane jako odwrotności wariancji wewnątrz poszczególnych badań, powiększonej o wariancje pomiędzy badaniami.

Pojawia się dodatkowo problem testowania wielokrotnego. Tę kwestię podnosi w swoim artykule Nuzzo [2015]. Jeżeli zakładamy prawdopodobieństwo błędnego odrzucenia hipotezy zerowej równe 0,05, to zgadzamy się, że przeciętnie raz na 20 decyzji popełniamy błąd. Ten poziom istotności dotyczy jednokrotnego wykonania testu istotności. Jeżeli stosujemy test niezależnie 20 razy, to prawdopodobieństwo, że przynajmniej raz popełnimy błąd, wynosi ponad 2/3.

---

[7] *Fixed-effect model* lub *commone-effect model*.
[8] *Random effect model*.

Zagadnienia, w których mamy do czynienia z tego typu problemami, występują również w klasycznych metodach, takich jak porównywanie średnich parami (testy post-hoc w ANOVA), testowanie istotności elementów macierzy korelacji, budowa modeli regresji o dużej liczbie zmiennych objaśniających. W tych sytuacjach trzeba ocenić rozmiar „niebezpieczeństwa" powodowanego przez testowanie wielokrotnie (wprost wynikające z liczby jednocześnie rozpatrywanych zmiennych lub grup), a następnie zastosować metody umożliwiające korektę poziomu istotności bądź wartości *p* (*p-value*).

3. Big Data

Obecnie w otoczeniu badawczym pojawiają się określenia takie, jak: *Data Mining*, *Data Analysis*, *Data Science* i *Big Data*. Wszystkie wymienione terminy mają swoją historię i obejmują określony zbiór danych i metod badawczych. *Big Data* to zbiór danych, który ma swoiste charakterystyki. To zbiór dużych, zmiennych i różnorodnych danych, których przetwarzanie i analiza są trudne, ale jednocześnie wartościowe, ponieważ może to prowadzić do zdobycia nowej wiedzy. *Big Data* ma zastosowanie wszędzie tam, gdzie dużej ilości danych cyfrowych towarzyszy potrzeba zdobywania nowych informacji lub wiedzy. Dostępność Internetu oraz usług świadczonych drogą elektroniczną, które w naturalny sposób są przystosowane do wykorzystywania baz danych, poszerzają zapotrzebowanie na statystyczne analizy danych. W 2001 r. grupa badawcza META Group (obecnie Gartner) opublikowała raport, który opisuje *Big Data* w modelu 3V: duża ilość danych (*Volume*); duża zmienność danych (*Velocity*); duża różnorodność danych (*Variety*). Następnie dodano kolejną charakterystykę o ocenę (weryfikację, *Value*) posiadanych danych – dochodząc w ten sposób do modelu 4V. Opis i wykorzystanie tego modelu w polskiej wersji 4W przedstawia się następująco:

1) wykorzystanie – wykorzystaj wewnętrzne (własne) zasoby danych;
2) wnioskowanie – umiejętnie stosuj techniki analityczne, użyj ekspertów;
3) wzbogacanie – wzbogacaj własne dane o informacje z rynku, używaj słowników i baz referencyjnych;
4) weryfikacja – koniecznie weryfikuj hipotezy i wnioski.

W roku 2012 Gartner uzupełnił podaną wcześniej definicję, wskazując, iż „*Big Data* to zbiory informacji o dużej objętości, dużej zmienności lub dużej różnorodności, które wymagają nowych form przetwarzania w celu wspomagania podejmowania decyzji, odkrywania nowych zjawisk oraz optymalizacji procesów" [Douglas, 2012].

4. Wybrane zasady stosowania *p-value*

Rosnąca liczba badań naukowych oraz dużych, złożonych zbiorów danych w ostatnich latach sprawiły, że rozszerzył się zakres zastosowań metod statystycznych. Otworzyło to nowe możliwości dla rozwoju naukowego wielu dyscyplin, ale także dostarczyło wielu obaw dotyczących wniosków płynących z prowadzonych badań. Wiarygodność naukowych wniosków, włączając ich powtarzalność, zależy nie tylko od samych metod statystycznych. Odpowiednio dobrane techniki, poprawnie prowadzone analizy oraz właściwa interpretacja statystycznych wyników odgrywają kluczową rolę w zapewnieniu solidności wniosków oraz warunkują, że niepewność, która je otacza, jest odpowiednio reprezentowana.

U podstaw wielu opublikowanych wniosków naukowych pojawia się koncepcja „istotności statystycznej", zwyczajowo ocenianej za pomocą wskaźnika *p-value*. Podczas gdy *p-value* może być traktowane jako użyteczna statystyczna miara, jednocześnie powszechnie nadużywa się go i błędnie interpretuje. Doprowadziło to do sytuacji, że wiele czasopism naukowych zachęca do ograniczania używania *p-value*, natomiast niektórzy naukowcy i statystycy nawet rekomendują jego wykluczenie.

W opublikowanym przez Amerykańskie Towarzystwo Statystyczne (ASA) komunikacie znajdujemy formalne stanowisko wyjaśniające kilka powszechnie uzgodnionych zasad leżących u podstaw prawidłowego zastosowania oraz interpretacji *p-value* [Wasserstein, Lazar, 2016]. Nieformalnie *p-value* jest prawdopodobieństwem wyznaczonym na podstawie określonego modelu statystycznego, takim, że statystyczne podsumowanie danych (np. różnica średnich z prób pomiędzy dwiema porównywanymi grupami) będzie większe lub równe od wartości obserwowanej. W artykule zapisano następujące zasady stosowania *p-value*:

1. *Wartości p-value mogą wskazywać, jak duża jest niezgodność danych w obrębie określonego modelu statystycznego.*

P-value stanowi jedno podejście do podsumowania niezgodności pomiędzy konkretnym zbiorem danych a proponowanym modelem statystycznym dla tego zbioru danych. Najbardziej powszechnym kontekstem jest model zbudowany przy odpowiednim zestawie założeń, wraz z hipotezą zerową. Często hipoteza zerowa zakłada nieobecność badanego efektu, takiego jak brak różnicy pomiędzy dwoma grupami lub brak zależności pomiędzy czynnikiem a wynikiem. Im mniejsza wartość *p-value*, tym większa statystyczna niezgodność danych z hipotezą zerową, jeśli utrzymane są założenia przyjęte do wyznaczenia *p-value*. Ta niezgodność może być interpretowana jako podważenie lub dostarczenie dowodów przeciw hipotezie zerowej lub poczynionym założeniom.

2. *P-value nie mierzy prawdopodobieństwa, że weryfikowana hipoteza jest prawdziwa, lub prawdopodobieństwa, że dane uzyskano w sposób losowy.*

Badacze często chcą traktować *p-value* jako stanowisko odnoszone do prawdziwości hipotezy zerowej lub do prawdopodobieństwa, że dane mają charakter losowy. *P-value* nie jest ani jednym, ani drugim. Jest stanowiskiem wobec danych w odniesieniu do określonego hipotetycznego wyjaśnienia, natomiast nie stanowiskiem dotyczącym samego wyjaśnienia.

3. *Wnioski naukowe, decyzje polityczne i biznesowe nie powinny być oparte tylko na sytuacji statystycznej, kiedy wartości p-value przekraczają określony punkt progowy.*

Praktyki, które ograniczają analizę danych lub wnioskowanie naukowe do pewnych mechanicznych reguł, takich jak $p < 0,05$ dla uzasadnienia twierdzeń naukowych lub wniosków, mogą prowadzić do błędnych przekonań i niewłaściwie podejmowanych decyzji. Wniosek nie od razu jest prawdziwy z jednej strony punktu podziału i fałszywy z drugiej. Naukowcy powinni uwzględnić więcej kontekstowych czynników w prowadzeniu wnioskowania naukowego, włączając projektowanie badania, jakość pomiarów, zewnętrzne dowody badanego zjawiska oraz zasadność założeń leżących u podstaw analizy danych. Pragmatyczne rozważania często wymagają decyzji binarnych: tak lub nie, ale to nie oznacza, że samo *p-value* może zapewnić, że decyzja jest właściwa lub nie. Powszechne zastosowanie „istotności statystycznej" (zazwyczaj interpretowanej jako $p \leq 0,05$) w charakterze pozwolenia na twierdzenie o odkryciu naukowym (lub domniemaniu prawdziwości) prowadzi do znacznego zakłócenia procesu naukowego.

4. *Właściwe wnioskowanie wymaga pełnego raportowania i przejrzystości.*

P-value oraz powiązane analizy nie powinny być raportowane w sposób wybiórczy. Prowadzenie wielokrotnych analiz danych oraz raportowanie tylko tych z określonym poziomem *p-value* (zazwyczaj tych przekraczających progową istotność) czyni uzyskane wartości *p-value* nieinterpretowanymi. Takie wyniki, znane również jako pogłębianie danych czy też selektywne wnioskowanie, prowadzą do fałszywego nadmiaru statystycznie istotnych wyników w publikowanej literaturze i powinny być stanowczo unikane. Nie jest konieczne prowadzenie wielu testów statystycznych dla zauważenia takiego problemu: jeżeli badacz sam wybiera to, co chce przedstawić na podstawie wyników statystycznych, prawidłowa interpretacja tych wyników jest poważnie zagrożona, zwłaszcza jeśli czytelnik nie jest poinformowany o wyborze oraz jego przesłankach. Naukowcy powinni ujawniać ilość hipotez branych pod uwagę w czasie całego badania, wszelkie decyzje o zbieraniu danych, o prowadzonych analizach statystycznych, o wyznaczonych wartościach *p-value*. Poprawne wnioski na-

ukowe, bazujące na wartościach *p-value* i powiązanych statystykach, nie mogą być wyciągnięte bez informacji, jak wiele i jakie analizy zostały przeprowadzone oraz jak te analizy (włączając *p-value*) zostały wybrane do raportowania.

5. *P-value lub statystyczna istotność nie mierzy rozmiaru efektu ani znaczenia wyniku.*

Statystyczna istotność nie jest równoważna istotności naukowej (czy w innym znaczeniu np. ludzkiej czy ekonomicznej). Mniejsze wartości *p-value* niekoniecznie muszą oznaczać wystąpienie większego lub ważniejszego efektu, natomiast większe wartości *p-value* nie muszą wskazywać mniejszego efektu lub jego braku. Dowolny efekt, niezależnie od tego, jak jest mały, może wygenerować niewielką wartość *p-value*, jeśli rozmiar próbki lub precyzja pomiaru jest wystarczająco wysoka. Przeciwnie, duże efekty mogą generować imponujące wartości *p-value* jeśli wielkość próbki jest niewielka lub precyzja pomiaru okazuje się niedostateczna. Podobnie jednakowo oszacowane efekty mogą mieć różne wartości *p-value*, jeśli precyzja oszacowań się różni.

6. *Sama wartość p-value nie jest dobrą miarą „dowodzenia" modelu lub hipotezy.*

Badacze powinni rozpoznać, że *p-value* bez kontekstu oraz bez dodatkowych dowodów dostarcza ograniczonej informacji. Dla przykładu, wartość *p-value* bliska 0,05 sama przez siebie prezentuje słaby dowód wobec hipotezy zerowej. Podobnie stosunkowo duże *p-value* nie oznacza dowodu na korzyść hipotezy zerowej. Wiele innych hipotez może być równych lub bardziej zgodnych w odniesieniu do obserwowanych danych. Z tych też powodów analiza danych nie powinna kończyć się tylko na wyznaczeniu *p-value*, gdy inne metody są zasadne i wykonalne.

Podsumowanie

Ze względu na powszechne nadużywanie i nieporozumienia związane z *p-value*, niektórzy statystycy wolą je uzupełniać lub nawet zastępować innymi możliwościami. Należą do nich metody podkreślające przewagę estymacji nad testowaniem, takie jak: przedziały ufności, przedziały wiarygodności czy przedziały predykcji; metody bayesowskie; alternatywne metody dowodzenia, m.in. wskaźniki prawdopodobieństwa lub czynniki Bayesa (*Bayes Factors*); jak również inne podejścia: modelowanie teoretyczno-decyzyjne (*decision-theoretic modeling*) lub wskaźniki wykrycia fałszu (*false discovery rates*). Wszystkie te miary i podejścia opierają się na dalszych założeniach, ale mogą bardziej bezpośrednio dotyczyć wielkości efektu (i związanej z nim niepewności) lub określenia, czy hipoteza jest poprawna.

Dobra praktyka statystyczna, jako istotny składnik dobrej praktyki naukowej, podkreśla zasady poprawnego projektowania i prowadzenia badań, różnorodność numerycznych i graficznych zestawień danych, zrozumienie zjawiska będącego przedmiotem badania, interpretację wyników, pełne ich raportowanie oraz właściwe logiczne i ilościowe zrozumienie, co oznaczają sumaryczne wyniki takich analiz. Żaden pojedynczy wskaźnik nie zastąpi dogłębnego rozumowania naukowego.

Literatura

American Statistical Association (2010), *ASA Statement on Risk-Limiting Post Election Audits*, http://www.amstat.org/policy/pdfs/Risk-Limiting_Endorsement.pdf (dostęp: 20.04.2016).

Douglas L. (2012), *The Importance of "Big Data": A Definition*, Gartner, https://www.gartner.com/doc/2057415/importance-big-data-definition (dostęp: 21.06.2012).

Gelman A., Loken E. (2014), *The Statistical Crisis in Science*, "American Scientist", No. 102, http://www.americanscientist.org/issues/feature/2014/6/thestatistical-crisis-in-science (dostęp: 20.04.2016).

Glass G.V. (1976), *Primary Secondary and Meta-Analysis of Research*, "Educational Researcher", No. 5, s. 3-8.

Johnson V.E. (2013), *Uniformly Most Powerful Bayesian Tests*, "Annals of Statistics", No. 41, s. 1716-1741.

Leek J. (2014), *On the Scalability of Statistical Procedures: Why the p-value Bashers Just Don't Get It*, "Simply Statistics Blog", http://simplystatistics.org/2014/02/14/on-thescalability-of-statistical-procedures-why-the-p-value-bashers-just-dont-get-it/ (dostęp: 20.04.2016).

Morganstein D., Wasserstein R. (2014), *ASA Statement on Value-Added Models*, "Statistics and Public Policy", No. 1, s. 108-110, http://amstat.tandfonline.com/doi/full/10.1080/2330443X.2014.956906 (dostęp: 20.04.2016).

Nuzzo R. (2014), *Scientific Method: Statistical Errors*, "Nature", No. 506, s. 150-152, http://www.nature.com/news/scientific-method-statistical-errors-1.14700 (dostęp: 20.04.2016).

Peng R. (2015), *The Reproducibility Crisis in Science: A Statistical Counterattack*, "Significance", No. 12(3), s. 30-32.

Phys.org – Science News Wire (2013), *The Problem with-values: How Significant Are They, Really?* http://phys.org/wire-news/145707973/the-problem-with-p-values-howsignificant-are-they-really.html (dostęp: 20.04.2016).

Siegfried T. (2010), *Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics*, "ScienceNews", Vol. 177, No. 26, https://www.sciencenews.org/article/odds-are-its-wrong (dostęp: 20.04.2016).

Siegfried T. (2014), *To Make Science Better, Watch Out for Statistical Flaws*, "ScienceNews", https://www.sciencenews.org/blog/context/make-science-better-watch-out-statisticalflaws (dostęp: 20.04.2016).

Sokołowski A. (2004), *O niewłaściwym stosowaniu metod statystycznych*, StatSoft Polska, Kraków.

Trafimow D., Marks M. (2015), *Editorial*, "Basic and Applied Social Psychology", Vol. 37, Issue 1, s. 1-2.

Wasserstein R.L., Lazar N.A. (2016), *The ASA's Statement on p-values: Context, Process and Purpose*, "The American Statistician" [online], http://dx.doi.org/10.1080/00031305.2016.1154108.

A NOTE OF THE *P-VALUE*

Summary: The aim of this article is to present the problems associated with the commone use of statistical software and interpreting test results. Research in economics vary the spirit and purpose of research in the medical sciences and the social sciences. We use similar research methods, though not always see different fundamental approach to the used statistical models, especially in a layer to meet the assumptions of distributions of random variables. We are looking for fast and reliable interpretation of the results which poses a threat to the entire research.

Keywords: type I error, type II error, *p-value*.