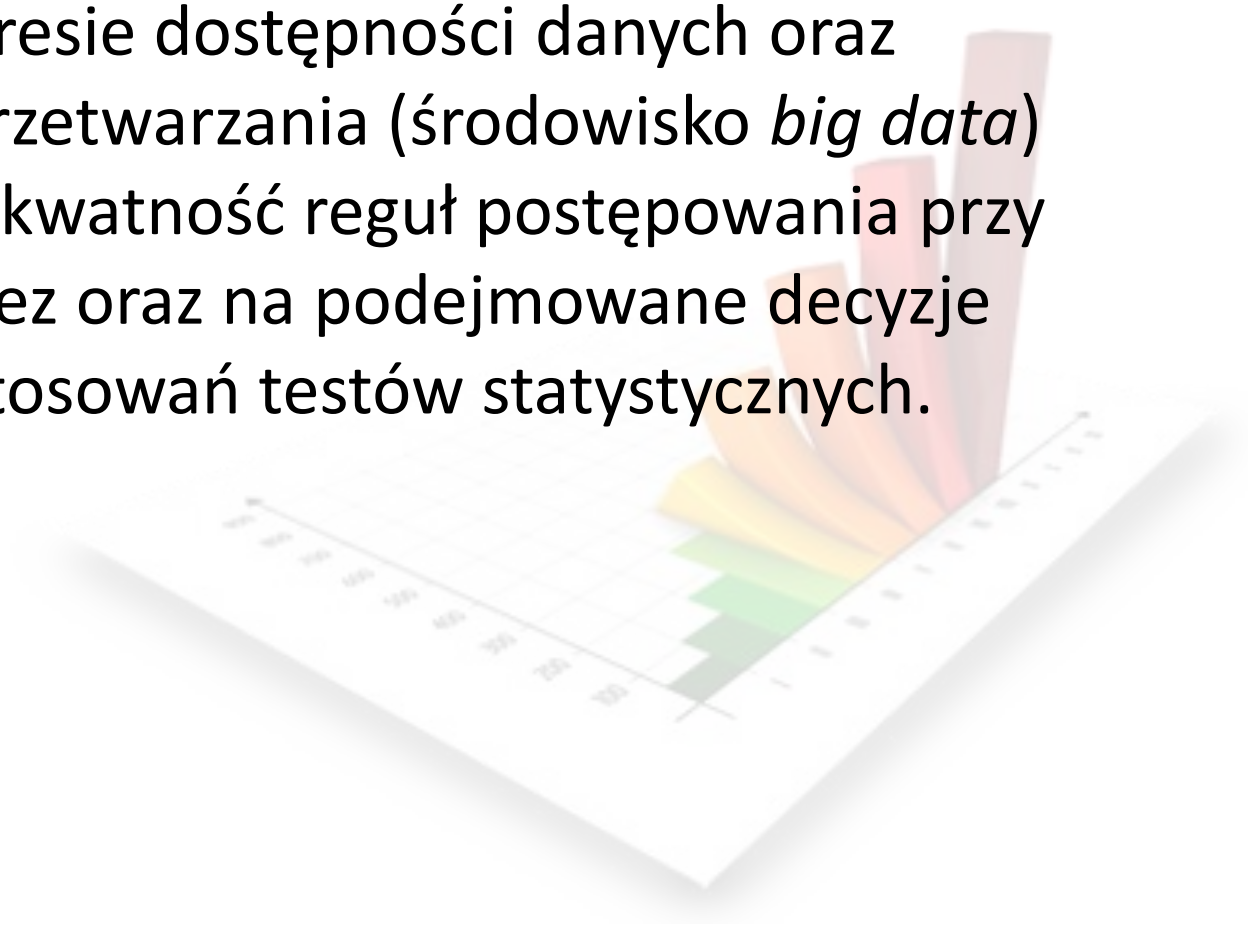
Mirosław Szreder
Uniwersytet Gdański

Istotność statystyczna w czasach *big data*

Konferencja naukowa
MET2019
Metodologia Badań Statystycznych
3-5 lipca 2019 r., Warszawa

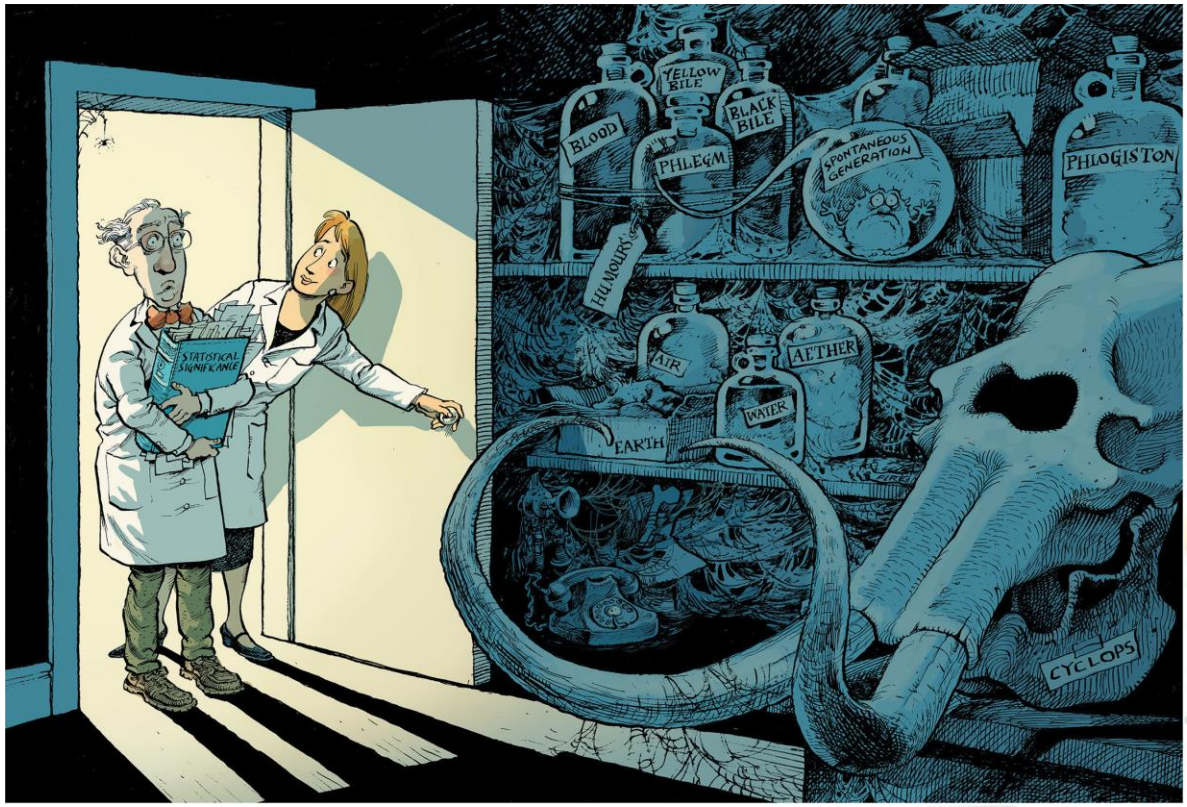
Cel badań: przedyskutowanie, na ile współczesne przemiany w zakresie dostępności danych oraz możliwości ich przetwarzania (środowisko *big data*) wpływają na adekwatność reguł postępowania przy weryfikacji hipotez oraz na podejmowane decyzje wynikające z zastosowań testów statystycznych.



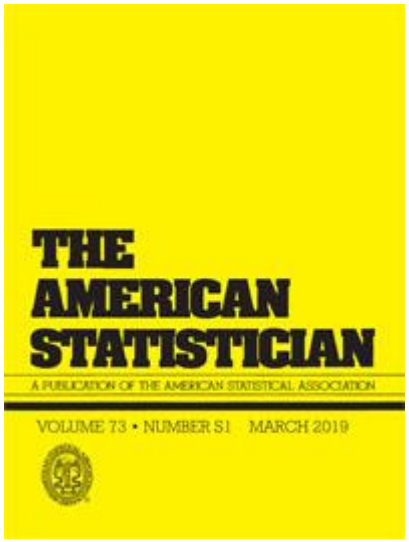
Motywacja:

1) Powrót dyskusji o istotności statystycznej i wykorzystaniu wskaźnika *p-value* w prestiżowych czasopismach z marca 2019 r.

„Nature”: Retire statistical significance



Valentin Amrhein, **Sander Greenland**, **Blake McShane** and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.



Moving to a World Beyond “p < 0.05”

Cytat z “Nature” na temat zbioru artykułów z powyższego voluminu „The American Statistician”:

It presents more than 40 papers on ‘Statistical inference in the 21st century: a world beyond P < 0.05’.

The editors introduce the collection with the caution “don’t say ‘statistically significant’.

2) Niewystarczająca reakcja środowiska statystyków, w tym redaktorów czasopism, na oświadczenie Amerykańskiego Towarzystwa Statystycznego – *The American Statistical Association* (ASA) z 2016 roku:

ASA Statement on Statistical Significance and *P-Values*

I. Wartości prawdopodobieństwa krytycznego (*p-value*) mogą wskazywać na to, jak nieprzystające do określonego modelu statystycznego są zaobserwowane dane.

II. *P-value* nie jest miarą prawdopodobieństwa tego, że analizowana hipoteza jest prawdziwa, ani tego że dane zostały uzyskane wyłącznie w drodze losowania (zostały wygenerowane przez proces losowy).

III. Konkluzje badawcze oraz decyzje ekonomiczne lub związane z określoną polityką działania nie powinny być oparte wyłącznie na tym, czy *p-value* przekroczyło określony próg.

IV. Poprawne wnioskowanie wymaga od badacza ujawnienia pełnej informacji oraz przejrzystości.

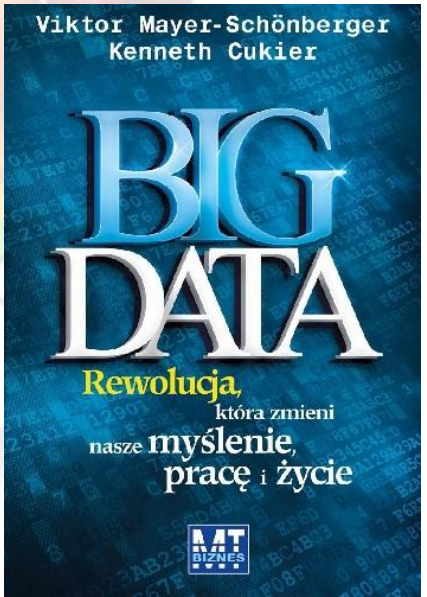
V. Ani *p-value*, ani statystyczna istotność nie mierzą wielkości efektu oraz nie mierzą ważności uzyskanego wyniku.

VI. Sam w sobie wskaźnik *p-value* nie stanowi dobrej miary wartości przesłanek dotyczących modelu lub hipotezy.

3) Wskazanie potencjalnych zagrożeń związanych z rosnącą rolą błędów nielosowych we wnioskowaniu statystycznym, zwłaszcza wśród osób przekonanych o tym, że większa liczba obserwacji może rekompensować brak dbałości o spełnienie założeń modelu wnioskowania.

„Jesteśmy gotowi do poświęcenia odrobiny dokładności w zamian za poznanie ogólnego trendu”.

(Mayer- Schönberger V., Cukier K., BIG DATA. Rewolucja, która zmieni nasze myślenie, pracę i życie. Wyd. MT Biznes, Warszawa 2014)



Statystyczna istotność i wskaźnik *p-value* – współczesna krytyka

Wynik statystycznie istotny nie zawsze oznacza ważny

significant ≠ important" , "influential"

Nowy lek obniżający ciśnienie krwi średnio o 0,10 jedn., z błędem standardowym 0,03 jedn. ~ **wynik statystycznie istotny**

Inny lek obniżający ciśnienie krwi o 10 jedn. z błędem standardowym 10 jedn. ~ **wynik statystycznie nieistotny**

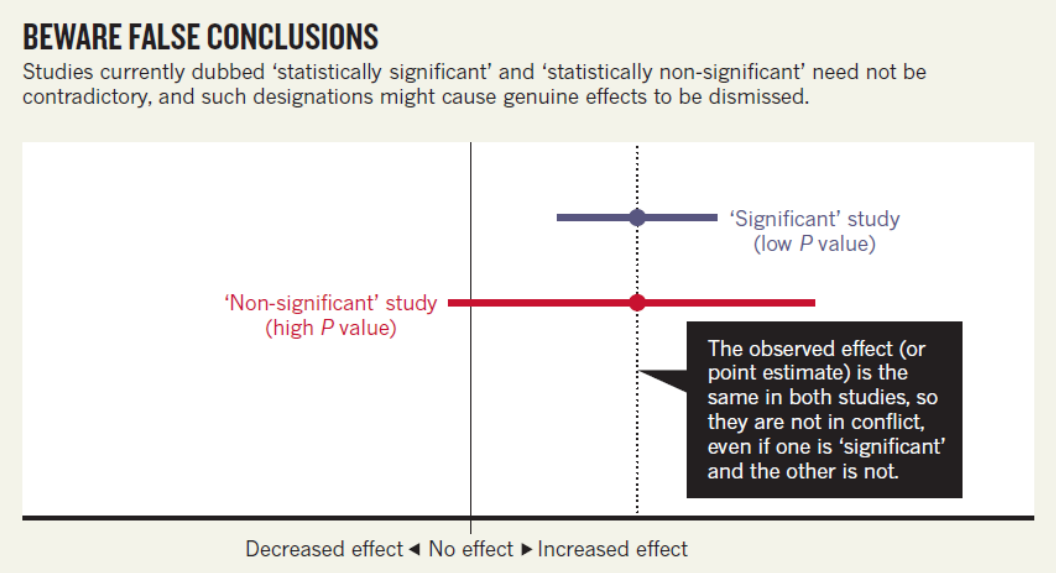
Gelman A., Stern H. (2006), The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, (4), 328-331.

Istota krytyki

Nie ma ostrej i jednoznacznej linii podziału pomiędzy wyrażeniami „statystycznie istotny” i „statystycznie nieistotny”.

Istotność w statystyce zmienia się w sposób ciągły, tak jak ciągłą zmienną losową jest wartość *p-value*.

Wyniki „statystycznie istotne” i „statystycznie nieistotne” nie muszą być sobie przeciwstawiane.



Różnica między statystyczną istotnością a nieistotnością sama w sobie nie jest statystycznie istotna

Przykład (z Gelman A., Stern H. (2006))

Załóżmy, że dla przetestowania istotności efektu działania pewnego czynnika wykonano dwa badania reprezentacyjne:

- w badaniu A otrzymano: średnią wielkość efektu równą 25 i odchylenie standardowe równe 10 → **statystycznie istotny**;
- w badaniu B otrzymano: średnią wielkość efektu równą 10 i odchylenie standardowe równe 10 → **statystycznie nieistotny**.

W rzeczywistości różnica ta – obliczona jako wartość oczekiwana różnicy między średnimi próbkowymi – jest **statystycznie nieistotna**:
różnica ta wynosi **15**, a odchylenie standardowe **14** (pierwiastek z sumy kwadratów odchyleń standardowych z obu badań, wynoszącej 200)

Nowe badanie C dla większej próby: średnia wielkość efektu = 2,5 i odchylenie standardowe = 1 → **statystycznie istotny**, tak jak w badaniu A, mimo że różnice między efektami w A i C są duże i statystycznie istotne
(wartość oczekiwana różnicy średnich wynosi 22,5 z odchyleniem standardowym 10,05).

Wskaźnik *p-value* powinien być traktowany jedynie jako **jedno ze źródeł** dowodzenia nieprawdziwości hipotezy zerowej.

Tym bardziej, że *p-value* odnosi się bezpośrednio nie tylko do hipotezy zerowej, jak przyjęło się uważać.
Odnosi się także do całego modelu wnioskowania i jego założeń, a więc także do wszystkich tych okoliczności i zakłóceń (błędów nielosowych), które miały wpływ na przyjęcie takiej, a nie innej wartości statystyki testowej.

"*The p-value was never meant to be used the way it's used today*"
(Goodman, na podst. Nuzzo (2014))

Skłonność do błędnego interpretowania *p-value*

Gdy *p-value* okazuje się wyższe niż przyjęty próg (np. 0,05), błędne są stwierdzenia:

„nie występuje różnica”,
„nie występuje współzależność” .

Na podstawie *p-value* ocenia się błędnie siłę dowodów prawdziwości hipotezy alternatywnej.

p = 0,05 → 5% szansa wystąpienia danych zaobserwowanych w próbie, przy założeniu, że prawdziwa jest H_0;
ale nie oznacza to, że
1 – p = 0,95 → 95% szansa wystąpienia tych danych, jeżeli prawdziwa jest H_1.

Przy takiej błędnej interpretacji iloraz szans na rzecz hipotezy alternatywnej wynosiłby (95 : 5) ≈ 19 : 1.
Benjamin i Berger (2019) dowodzą (odwołując się do czynnika bayesowskiego), że iloraz ten jest znacznie niższy i wynosi ≈ 2,5 : 1

Wskaźnikowi *p-value* nadano w ostatnich kilkunastu latach zbyt duże znaczenie, sugerujące, iż jego wartość jest w stanie wyrazić wszystkie najważniejsze elementy niepewności w procesie weryfikacji hipotez.

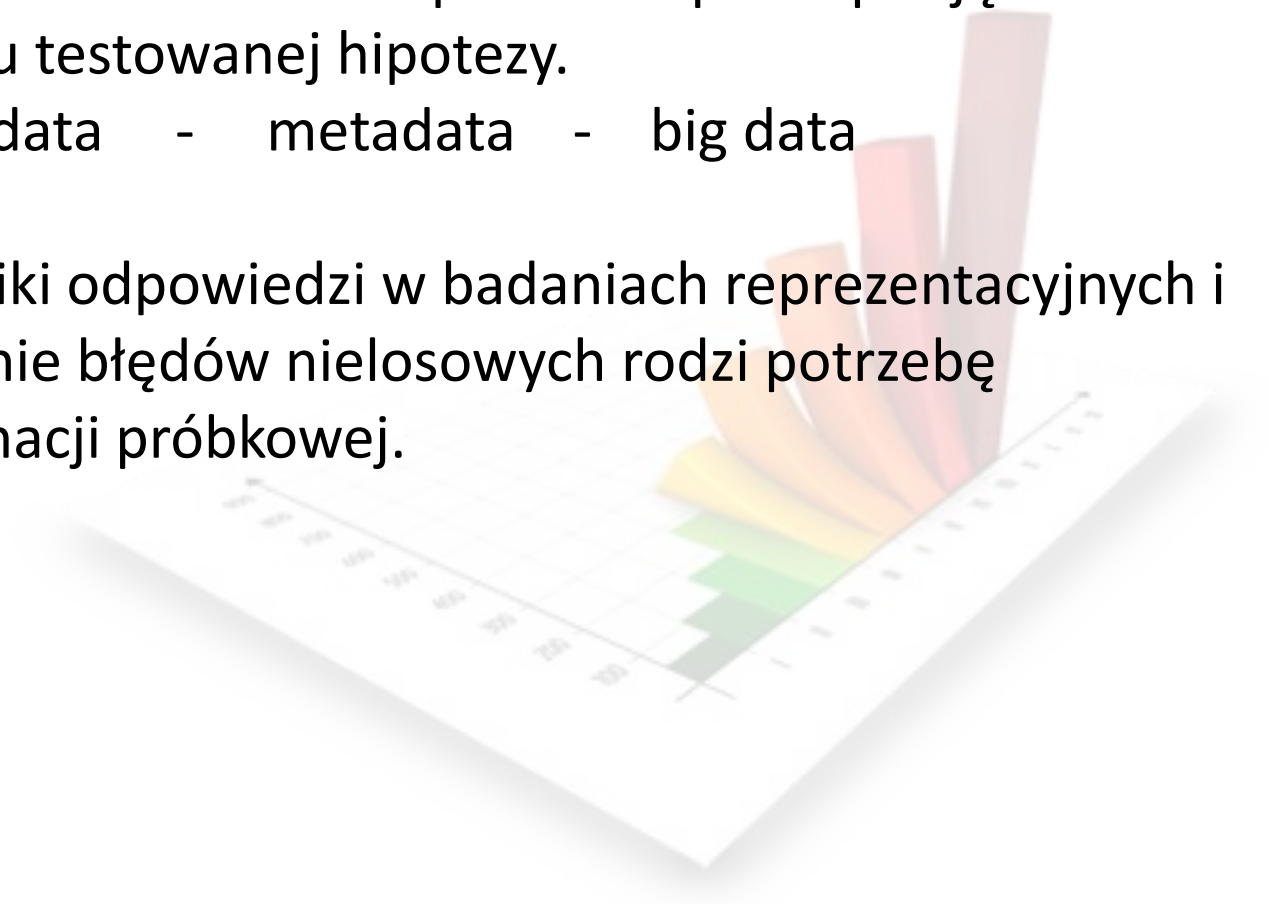
Zdaniem Goodmana, (2019), jest to już pewne zjawisko socjologiczne, nie tylko naukowe:
Tak jak wierzymy w wartość pieniądza, za którym w przekonaniu konsumentów kryje się prawo do nabycia określonych dóbr lub usług, tak wierzymy, że za określonymi wartościami *p-value* kryje się prawo do naukowego uznania określonych hipotez, a często także prawo do opublikowania uzyskanych wyników.

Big data szansą i wyzwaniem

1. Wykorzystanie innych źródeł danych dobrze wpisuje się w główny postulat płynący z omówionej dyskusji – nie rezygnacja z *p-value*, lecz głębsza analiza źródeł niepewności przed podjęciem decyzji o odrzuceniu testowanej hipotezy.

paradata - metadata - big data

2. Malejące wskaźniki odpowiedzi w badaniach reprezentacyjnych i wzrastające znaczenie błędów nielosowych rodzi potrzebę uzupełnienia informacji próbkowej.



3. Zagrożenie: pokusa, aby zwiększając liczebność próby doprowadzić do tego, by zaobserwowany, nawet niewielki efekt (ang. *size effect*) stanowił wartość statystycznie istotną (odrzucenie H_0) – zjawisko nazywane *p-hacking*.

Z oświadczenia ASA:
Ani p-value, ani statystyczna istotność nie mierzą wielkości efektu oraz nie mierzą ważności uzyskanego wyniku.

4. Wyzwanie: brak uznanego podejścia metodologicznego do wykorzystania big data we wnioskowaniu statystycznym.

5. Ostateczna konkluzja: uzyskanie *p-value* wskazującego na statystyczną istotność nie może być rozumiane jako dowód nieprawdziwości hipotezy, lecz wskazanie, iż warto ją dalej badać.

"*The numbers are where the scientific discussion should start, not end*."

(Goodman, na podst. Nuzzo (2014))

Literatura przywołana

Amrhein V., Greenland S., McShane B. (2019). *Retire statistical significance.* "Nature" (567), 305-307.

Benjamin D.J., Berger J.O. (2019). *Three Recommendations for Improving the Use of p-Values.* "The American Statistician", (73):sup1, 186-191.

Gelman A., Stern H. (2006), *The difference between "significant" and "not significant" is not itself statistically significant.* "The American Statistician", (4), 328-331.

Mayer-Schönberger V., Cukier K. (2013). BIG DATA. *Rewolucja, która zmieni nasze myślenie, pracę i życie.* Warszawa: MT Biznes.

Nuzzo R. (2014). *Statistical errors.* "Nature" (506), 150-152

Dziękuję za uwagę!