DOWODZENIE HIPOTEZ ZA POMOCĄ CZYNNIKA BAYESOWSKIEGO (BAYES FACTOR): PRZYKŁADY UŻYCIA W BADANIACH EMPIRYCZNYCH

Artur Domurat[1]
Akademia Leona Koźmińskiego

Michał Białek[2]
Akademia Leona Koźmińskiego

Streszczenie: *Testów statystycznych używa się w nauce po to, żeby wesprzeć zaproponowane hipotezy badawcze (teorie, modele itp.). Czynnik bayesowski (Bayes Factor, BF) jest metodą bezpośrednio wskazującą tę z dwóch hipotez, która lepiej wyjaśnia uzyskane dane. Jego wykorzystanie we wnioskowaniu statystycznym prowadzi do jednego z trzech wniosków: albo badanie bardziej wspiera hipotezę zerową, albo alternatywną, albo wyniki nie wspierają żadnej w sposób rozstrzygający i są niekonkluzywne. Symetria tych wniosków jest przewagą metody czynnika bayesowskiego nad testami istotności. W powszechnie używanych testach istotności nie formułuje się wniosków wprost, lecz albo się odrzuca hipotezę zerową, albo się jej nie odrzuca. Rozdźwięk między taką decyzją a potrzebami badacza często jest powodem nadinterpretacji wyników testów statystycznych. W szczególności wyniki nieistotne statystycznie są często nadinterpretowane jako dowód braku różnic międzygrupowych lub niezależności zmiennych.*

W naszej pracy omawiamy założenia teoretyczne metody BF, w tym różnice między bayesowskim a częstościowym rozumieniem prawdopodobieństwa. Przedstawiamy sposób weryfikacji hipotez i formułowania wniosków według podejścia bayesowskiego. Do jego zalet należy m.in. możliwość gromadzenia dowodów na rzecz hipotezy zerowej. Wykorzystanie metody w praktyce ilustrujemy przykładami bayesowskiej reinterpretacji wyników kilku opublikowanych badań empirycznych, w których wykonywano tradycyjne testy istotności. Do obliczeń wykorzystaliśmy darmowy program JASP 0.8, specjalnie dedykowany bayesowskiej weryfikacji hipotez statystycznych.

[1] Artur Domurat, Centrum Psychologii Ekonomicznej i Badań Decyzji, Akademia Leona Koźmińskiego, ul. Jagiellońska 57/59, 03-301 Warszawa, e-mail: artur.domurat@kozminski.edu.pl

[2] Michał Białek, Centrum Psychologii Ekonomicznej i Badań Decyzji, Akademia Leona Koźmińskiego, ul. Jagiellońska 57/59, 03-301 Warszawa, e-mail: mbialek@kozminski.pl

 DOI: 10.7206/DEC.1733-0092.79

Słowa kluczowe: *wnioskowanie statystyczne, testowanie hipotezy zerowej, reguła Bayesa, czynnik bayesowski, wartość p.*

WEIGHING EVIDENCE IN FAVOUR OF RESEARCH HYPOTHESES USING BAYES FACTOR: EXAMPLES OF APPLICATION IN EMPIRICAL STUDIES

Abstract: *Statistical tests are used in science in order to support research hypotheses (theory, model). The Bayes Factor (BF) is a method that weighs evidence and shows which out of two hypotheses is better supported. Adopting the BF in statistical inference, we can show whether data provided stronger support for the null hypothesis, the alternative hypothesis or whether it is inconclusive and more data needs to be collected to provide more decisive evidence. Such a symmetry in interpretation is an advantage of the Bayes Factor over classical null hypothesis significance testing (NHST). Using NHST, a researcher draws conclusions indirectly, by rejecting or not rejecting the null hypothesis. The discrepancy between these decisions and the researcher's needs, often leads to misinterpretation of significance test results, e.g. by concluding that non-significant p-values are evidence for the absence of differences between groups or that variables are independent.*

In this work we show the main differences between the Bayesian and the frequential approach to the understanding of probability and statistical inference. We demonstrate how to verify hypotheses using the BF in practice and provide concrete examples of how it modifies conclusions about empirical findings based on the NHST procedure and the interpretation of p-values. We discuss the advantages of the BF – particularly the validation of a null hypothesis. Additionally, we provide some guidelines on how to do Bayesian statistics using the freeware statistical program JASP 0.8.

Key words: *statistical inference, null hypothesis testing, Bayes Factor, p-value, Bayes' rule.*

1. Statystyczne testy istotności a potrzeby ich użytkowników

1.1. O nadinterpretacji testów istotności

W jednym z badań Baumeistera, Bratyslavsky'ego, Muravena i Tice (1998) proszono uczestników o rysowanie figur bez odrywania ręki. Badanie prowadzono w labora-

 DOI: 10.7206/DEC.1733-0092.79

torium, w którym unosił się kuszący zapach świeżo upieczonych ciasteczek czekoladowych. Przed wykonaniem zadania badanym podano do zjedzenia albo rzodkiewki (grupa 1), albo ciasteczka czekoladowe (grupa 2), albo nie podano nic (grupa 3). Spodziewano się, że badani, którzy oparli się pokusie zjedzenia słodyczy i zjedli rzodkiewki, będą wyczerpani poznawczo (ang. *ego depletion*) i wcześniej porzucą zadanie właściwe, które tak naprawdę nie miało rozwiązania. Uzyskano następujące średnie i odchylenia standardowe w tych trzech grupach (w minutach): $M_1 = 8{,}35$, $SD_1 = 4{,}67$, $n_1 = 25$, $M_2 = 18{,}9$, $SD_2 = 6{,}86$, $n_2 = 21$, $M_3 = 20{,}86$, $SD_3 = 7{,}30$, $n_3 = 21$. Do użytych metod analizy statystycznej należały między innymi dwa testy t-Studenta dla prób niezależnych, których wyniki opisano za pomocą następujących dwóch stwierdzeń:

(A) Uczestnicy badania jedzący rzodkiewki porzucają frustrujące zadania wcześniej, niż jedzący ciasteczka czekoladowe, $t(44) = 6{,}03$, $p < .001$.

(B) Osoby, które jadły ciasteczka czekoladowe, nie różniły się czasem porzucenia zadania od osób, którym nie podano żadnego jedzenia, $t < 1$, wynik nieistotny.

Przy stwierdzeniu (A) zaraportowano wskaźnik istotności statystycznej $p < 0{,}001$ (gdyby wartość p podać dokładniej, byłoby to $p = 0{,}0000001524$). Zastanówmy się, co oznacza uzyskana wartość p[3] w wykonanym teście statystycznym?

1) Zdecydowanie obalono hipotezę zerową H_0: $\mu_1 = \mu_2$, orzekającą, że jedzący rzodkiewki przeciętnie tak samo szybko rezygnują z wykonania zadania, jak jedzący ciasteczka.
2) Wartość $p < 0{,}001$ to prawdopodobieństwo tego, że nie ma różnicy między typowym czasem wykonywania zadania po zjedzeniu rzodkiewek μ_1 i po zjedzeniu ciasteczek μ_2.
3) Całkowicie udowodniono hipotezę badawczą, że jedzący rzodkiewki różnią się przeciętnym czasem porzucenia zadania od jedzących ciasteczka.
4) Z wartości $p < 0{,}001$ można wyprowadzić szanse prawdziwości hipotezy badawczej o tym, że jedzący rzodkiewki różnią się przeciętnym czasem porzucenia zadania od jedzących ciastka.
5) Dzięki wartości $p < 0{,}001$ znamy prawdopodobieństwo niesłusznego odrzucenia hipotezy, że przeciętny czas rezygnacji jest taki sam u jedzących rzodkiewki i u jedzących ciasteczka.
6) Uzyskano wiarygodny wynik eksperymentalny w tym sensie, że jeśli taki eksperyment powtarzano by wiele razy, to istotną różnicę między jedzący-

[3] W literaturze przedmiotu anglojęzyczny termin *p-value* tłumaczony jest dosłownie jako *p*-wartość, co podkreśla, że jest to rozważane pojęcie teoretyczne (zob. np. Jarmakowska-Kostrzanowska, 2016; Koronacki i Mielniczuk, 2001), a nie konkretna wartość jakiejś zmiennej *p*. W niniejszej pracy terminów: „istotność", „istotność statystyczna" i „wartość *p*" używamy zamiennie dla naturalnego brzmienia wywodu. Jeśli chodzi o konkretne wartości, przedstawiamy je wprost, pisząc np. o wartości $p = 0{,}035$.

mi rzodkiewki i ciasteczka uzyskano by w ponad 999 przypadków na 1000 takich badań.

Wskaźnik istotności *p* pokazuje szansę, z jaką można było zaobserwować różnicę między badanymi grupami taką jak uzyskana lub większą, nawet jeśli w rzeczywistości ludzie jedzący rzodkiewki i ciasteczka rezygnują z wykonania zadania po tym samym czasie. Ogólniej i precyzyjniej, wartość *p* jest prawdopodobieństwem warunkowym uzyskania statystyki empirycznej *T* takiej, jak zaobserwowana T_p lub jeszcze bardziej nietypowej, przy założeniu, że hipoteza zerowa jest prawdziwa, $p = P(T \geq T_p | H_0)$. Ponieważ istotność charakteryzuje statystykę wyliczoną z danych *D*, można ją prościej zapisać jako $p = P(D|H_0)$.

Stwierdzenia 1 i 3 są zatem fałszywe, ponieważ test statystyczny nie daje podstaw do stwierdzeń ogólnych (bezwarunkowych, pomijających założenie o prawdziwości H_0) i kategorycznych (pewnych, nieprobabilistycznych). Stwierdzenia 2 i 4 są także błędne, ponieważ odwracają interpretację prawdopodobieństwa warunkowego, przypisując szanse hipotezom, zerowej $P(H_0|D)$ i badawczej $P(H_b|D)$, a nie danym, $P(D|H_0)$. Z podobnego powodu niewłaściwa jest interpretacja piąta, sugerująca, że $p = 1 - P(H_0)$ zawsze, podczas gdy odrzucenie H_0 jest błędem tylko wtedy, gdy jest ona prawdziwa. Błędna jest również interpretacja szósta, gdyż utożsamia istotność z prawdopodobieństwem $1 - P(D)$, co byłoby prawdą tylko wtedy, gdyby z góry wiadomo było, że tylko H_0 jest prawdziwa (Haller i Krauss, 2002; por. Gigerenzer, 2004).

Powyższe interpretacje sformułowaliśmy, nadając przykładową treść zestawowi ogólnych stwierdzeń (np. „Otrzymano prawdopodobieństwo tego, że hipoteza zerowa jest prawdziwa"), wykorzystanych w badaniu Hallera i Kraussa (2002). W badaniu tym wzięło udział 44 studentów, 39 badaczy oraz 30 wykładowców statystyki i metodologii, pochodzących z różnych uczelni niemieckich. Uczestnicy badania, po zapoznaniu się z pewną istotną statystyką empiryczną, mieli za zadanie ocenić prawdziwość tych stwierdzeń. Spośród 113 uczestników ankiety, zaledwie 6 wykładowców i 4 badaczy udzieliło poprawnej odpowiedzi, zaznaczając, że wszystkie z powyższych sześciu interpretacji są błędne.

1.2. Źródła nadinterpretacji testów istotności

Testy statystyczne istotności hipotezy zerowej (NHST, od ang. *null hypothesis significance testing*) są procedurą powszechnie stosowaną w naukach empirycznych. Z czego zatem wynikają tak częste błędne interpretacje lub nadinterpretacje wyników statystycznych?

Badanie Hallera i Kraussa stanowiło replikację wcześniejszego badania Oakesa (1986), w którym 66 spośród 68 uczestników, pracowników naukowych, uzna-

 DOI: 10.7206/DEC.1733-0092.79

ło za poprawną przynajmniej jedną z sześciu błędnych interpretacji. Powtarzając to badanie kilkanaście lat później i uzyskując podobny wynik, Haller i Krauss chcieli pokazać środowisku akademickiemu, że znajomość metod wnioskowania statystycznego jest wciąż niska i konieczne jest efektywniejsze nauczanie metod statystycznych. Wydaje się jednak, że niewłaściwa interpretacja wyników testów statystycznych może mieć również inne przyczyny. Są to przynajmniej trzy następujące kwestie: charakter zwyczajowo przyjętej procedury NHST, specyficzna logika testowania hipotez statystycznych, a przede wszystkim – rozdźwięk między wnioskami, które badacz chce wyciągnąć, a wnioskami, na które pozwalają przeprowadzone testy. Przyjrzyjmy się tym kwestiom po kolei.

1.2.1. Problemy z wiedzą

Po pierwsze, niepoprawna interpretacja może rzeczywiście wynikać z niewystarczającej wiedzy. Wyniki badań empirycznych komunikuje się w nauce zwięźle, zazwyczaj w formie wyciągniętego wniosku wraz z uzyskaną statystyką zastosowanego testu i jej istotnością – tak jak w zdaniach A i B. Taki lakoniczny zapis stanowi pewien skrót myślowy stosowany z założeniem, że czytelnik wie, jak przeprowadza się i interpretuje testy statystyczne. Ogólnie rzecz biorąc, ludzie mają naturalną skłonność do myślenia deterministycznego zamiast probabilistycznego (Tyszka, 2001). Niedostateczna znajomość metod statystycznych, w połączeniu ze zwięzłością języka nauki i wspomnianą tendencją do myślenia w kategoriach deterministycznych, mogą sprawiać wrażenie, że w badaniach „udowodniono" prawdziwość pewnych hipotez. Wyrażały to omawiane wcześniej stwierdzenia 1 i 3 z badania Hallera i Kraussa o tym, że zdecydowanie obalono hipotezę zerową, albo że całkowicie udowodniono hipotezę badawczą.

1.2.2. Problemy teorii testów statystycznych a praktyka ich stosowania

Po drugie, testy statystyczne są powszechnie nauczane i stosowane bez uwzględnienia kontekstu ich powstania i niezgodności teorii leżących u ich podstaw. Jak zauważył Gigerenzer (2004), procedura NHST stanowi niespójną hybrydę dwóch podejść: teorii R. A. Fishera i teorii J. Neymana i E. Pearsona, stanowiącej próbę udoskonalenia tej pierwszej. Od strony obliczeniowej teorie te są takie same: odnoszą się do pewnej hipotezy zerowej, postulują użycie tych samych statystyk do oceny nietypowości danych, wykorzystują wartość p wyliczaną w tych samych rozkładach dla tych samych statystyk. Teorie te różnią się jednak procedurą testu statystycznego, interpretacją wyników i odpowiedzią na fundamentalne pytanie: po co w ogóle testować dane empiryczne.

 DOI: 10.7206/DEC.1733-0092.79

Niezgodność podejść na poziomie założeń pociągnęła za sobą wiele burzliwych dyskusji między uczonymi i ich zwolennikami. W podejściu Fishera wynik testu ma być argumentem za ewentualnym odrzuceniem hipotezy zerowej (gdyż jest ona fałszywa, Fisher, 1955). Z kolei w podejściu Neymana-Pearsona chodzi o to, by mylić się możliwie najrzadziej, podejmując jakieś decyzje w oparciu o test statystyczny; sama prawdziwość hipotezy zerowej ma natomiast znaczenie drugorzędne (zob. np. Neyman, 1957). Więcej o tym sporze piszemy w Załączniku 1.

Badacze-praktycy, zmęczeni niekończącymi się sporami między tymi dwoma nurtami i coraz mniej rozumiejący, o co w sporze między nimi chodzi, zaczęli we własnych analizach dobierać z nich te elementy, które wydawały się im wspólne, zrozumiałe i przydatne. Z czasem nie tylko w praktyce badawczej, lecz nawet w licznych podręcznikach akademickich testy statystyczne przybrały postać następującego – jak to określił Gigerenzer (2004; por. Gigerenzer, Krauss i Vitouch, 2004) – rytuału (cyt. za Gigerenzer, 2004, s. 588):

1) Sformułuj statystyczną hipotezę zerową o „braku różnic" lub „korelacji zerowej". Nie precyzuj tego, co przewiduje twoja hipoteza badawcza, nie wyszczególniaj, czego można oczekiwać według konkurencyjnych hipotez badawczych.
2) Zastosuj 5% jako konwencję odrzucenia hipotezy zerowej. Jeśli wyniki są istotne, zaakceptuj hipotezę badawczą. Zaraportuj $p < 0{,}05$, $p < 0{,}01$, $p < 0{,}001$ – zgodnie z tym, jakie p otrzymałeś.
3) Tej procedury używaj zawsze.

Pogląd taki wydaje się przesadą, jednakże powyższy „rytuał" stosowany jest powszechnie. Publikowane są raporty z badań empirycznych weryfikujących hipotezy badawcze w oparciu o istotność statystyk, natomiast do rzadkości należą rozważania nad mocą testu i błędami I i II rodzaju. Prowadzi to do błędnego wrażenia i nadinterpretacji wyrażonych w stwierdzeniach 5 i 6 z początkowego przykładu.

1.2.3. Problemy z odwracaniem prawdopodobieństw

Zapominając o tym, że wartość p jest prawdopodobieństwem warunkowym, określonym przez rozkład statystyki testu przy założeniu prawdziwości hipotezy zerowej, odbiorca komunikatu jest „prowokowany" do popełnienia błędu odwracania prawdopodobieństw (ang. *inverse fallacy* – zob. np. Białek, 2015; Domurat i in., 2015; Villejoubert i Mandel, 2002). Prawdopodobieństwo wystąpienia śmiertelnego wypadku drogowego w wyniku jazdy samochodem w stanie nietrzeźwym, P(śmierć|nietrzeźwość) jest czym innym, niż prawdopodobieństwo tego, że jakiś śmiertelny wypadek był spowodowany przez nietrzeźwego kierowcę, P(nietrzeźwość|śmierć). Błąd ten

 DOI: 10.7206/DEC.1733-0092.79

we wnioskowaniu statystycznym pojawia się wtedy, gdy w sądach probabilistycznych bezkrytycznie stosuje się zasadę logiczną *modus tollens*, $[(A \Rightarrow B) \wedge \sim B] => \sim A$, zgodnie z którą zaprzeczenie następnika implikacji każe odrzucić poprzednik (Krämer i Gigerenzer, 2005). W efekcie we wnioskowaniu statystycznym stosowane są błędne sylogizmy (ich przegląd – zob. Westover, Westover i Bianchi, 2011). Jednym z nich jest rozumowanie następujące (Westover i in., 2011, s. 13):

Jeśli H_0 jest prawdziwa, to najprawdopodobniej $p > \alpha$.
$p \leq \alpha$

———————

$\therefore$ H_0 jest najprawdopodobniej fałszywa.

Błędne jest na przykład następujące rozumowanie zgodne z tym sylogizmem. Typowy pomnik warszawski ma nieduży cokół lub kolumnę (H_0), a więc większość warszawskich pomników ma niskie cokoły i kolumny ($p > \alpha$). Obserwując pomnik Zygmunta III Wazy na Starówce, stwierdzamy, że ma kolumnę wysoką, co należy do rzadkości ($p < \alpha$). A zatem najprawdopodobniej typowy pomniki warszawski nie ma niskiej kolumny. Chyba że – zgodnie z dysjunkcją Fishera (zob. Załącznik 1) – kolumna Zygmunta nie należy do populacji warszawskich pomników i znajduje się poza Warszawą.

W teście statystycznym hipotezę zerową H_0 opisuje rozkład teoretyczny statystyki testu, charakteryzowanej przez $p = P(D|H_0)$. Obserwacja danych mało prawdopodobnych dla tej populacji, np. takich, że $p = P(D|H_0) < 0{,}05$, nie oznacza automatycznie, że są one typowe dla hipotezy alternatywnej (nieprawdą jest, że $p = 1 - P(H_1|D)$). Hipoteza alternatywna może być nawet jeszcze mniej prawdopodobna. Jak wspomnieliśmy wcześniej, test NHST zakłada prawdziwość H_0, nie sposób zatem orzekać o prawdziwości takiej hipotezy w świetle wartości p. Z perspektywy NHST omówione wcześniej interpretacje 2 i 4 badania Baumeistera o szansach prawdziwości hipotezy zerowej i badawczej są więc błędne.

1.3. (Niezaspokojone) potrzeby badacza-empiryka

Celem badań empirycznych jest dowodzenie hipotez badawczych. Dane zbiera się po to, aby mieć argumenty przemawiające za występowaniem pewnych zjawisk lub po to, by im zaprzeczyć. Przedstawione na początku artykułu przykładowe interpretacje statystyk pokazują, że badacz chce po prostu:

- wesprzeć swoją hipotezę badawczą (stwierdzenie A),
- pokazać, że pewnych różnic, korelacji, zależności itd. nie ma (stwierdzenie B).

 DOI: 10.7206/DEC.1733-0092.79

Stwierdzenia te stanowią interpretację wyników z perspektywy hipotezy badawczej, a nie z perspektywy hipotezy zerowej. W dotychczasowych rozważaniach chcieliśmy pokazać, że błędne interpretacje lub nadinterpretacje testów istotności niekoniecznie wynikają z braku wiedzy statystycznej, lecz raczej z tego, że testy statystyczne nie realizują powyższych celów badacza wprost. Ich specyficzna logika jest daleka od typowego myślenia „zwykłego człowieka". Ludzie na co dzień snują przypuszczenia dotyczące różnych rzeczy, szukając dla nich uzasadnienia. Mało kto najpierw zaprzecza tym domysłom, a następnie rozważa szanse wystąpienia jakichś zdarzeń, gdyby to zaprzeczenie (czyli założenie, że nic się nie dzieje) było prawdziwe. Brzmi to zawile, ale tak pokrótce można by opisać testowanie hipotezy zerowej zwykłym, pozastatystycznym językiem. Raczej jest tak, że ludzie – obserwując takie a nie inne fakty – argumentów i dowodów za tym, że mają rację lub że się mylą, szukają wprost (Stanovich, 2009). Opisane wcześniej nadinterpretacje testów statystycznych pokazały, że ich użytkownicy myślą podobnie: chcą po prostu wiedzieć, czy mogą uznać postawione hipotezy za wiarygodne i prawdopodobne, czy są one niesłuszne lub nieprzekonujące w świetle przeprowadzonych analiz. Tymczasem statystyczne testy istotności:

- mówią o odrzucaniu hipotez, zamiast o ich przyjmowaniu lub wspieraniu,
- bezpośrednio zajmują się hipotezą zerową H_0, a nie hipotezą badawczą H_b,
- opisują zachowanie się danych w kontekście H_0, $P(D|\ H_0)$, zamiast uwiarygodniać zaproponowane modele lub hipotezy w świetle zebranych danych, $P(H_b|\ D)$,
- a przede wszystkim – charakteryzują się asymetrią wnioskowania.

Ideę weryfikacji hipotez w procedurze NHST pokażemy na przykładzie zadania demonstrującego złudzenie gracza (ang. *gambler's fallacy*).

***Przykład. Złudzenie gracza*[4].** Rzucono monetą dziesięciokrotnie i dziesięć razy wypadł orzeł. Co wypadnie za jedenastym razem?

Poza zebranymi dowodami, czyli zaobserwowaną serią rzutów, znaczenie ma tu też nasze wcześniejsze doświadczenie (teorie, modele itd.), które mogą zasugerować odpowiadającemu następujące wyjaśnienia obserwacji dziesięciu orłów[5]:

[4] Błąd ten obserwuje się w zachowaniu graczy w kasynach (stąd nazwa), którzy grając na przykład w ruletkę, po dłuższej serii pól czarnych, bardziej skłonni są obstawiać pola czerwone w kolejnych zakładach (zob. np. Tyszka, 1999). Uleganie złudzeniu gracza wynika z kierowania się zasadą reprezentatywności i wyraża się przekonaniem, że po serii wyników jednego rodzaju, większą szansę mają wyniki takie, które zniwelują odstępstwo od normy tych wcześniejszych (Tversky i Kahneman, 1971).

[5] Oczekiwaną „podręcznikowo" odpowiedzią jest stwierdzenie, że szanse wyrzucenia orła i reszki w kolejnym rzucie są takie same, gdyż moneta „nie ma pamięci", a kolejne rzuty są zdarzeniami niezależnymi. O złudzeniu mówi się wtedy, gdy ktoś stwierdzi, że większą szansę ma reszka, ponieważ liczba reszek i orłów powinna się zacząć wreszcie wyrównywać. Czy przypisanie większych szans orłowi jest natomiast błędne? Wyobraźmy sobie sceptyka, który – zobaczywszy dziesięć orłów pod rząd – nabierze wątpliwości co

 DOI: 10.7206/DEC.1733-0092.79

- model M_1: eksperymentator użył monety symetrycznej,
- model M_2: wykonując trik iluzjonistyczny, eksperymentator użył monety fałszywej z orłami po obu jej stronach.

Testowana jest H_0: $\theta = 0{,}5$ wobec H_1: $\theta \neq 0{,}5$. Przyjęto poziom istotności $\alpha = 0{,}01$. W doświadczeniu uzyskano $f = 10$ orłów w $n = 10$ rzutach. Statystyką testu jest częstość, w = f/n = 10/10 = 1, dla której dokładną dwustronną wartość p wyznaczamy z rozkładu dwumianowego: $p = 2 \times P(10; 10; 0{,}5) = 2 \times 1/1024 \cong 0{,}001954$. Ponieważ $p < \alpha$, należy odrzucić H_0 na rzecz H_1. Odrzuciwszy H_0, wynik interpretuje się zgodnie z hipotezą badawczą – że moneta ma dwa orły.

Zgodnie z „rytuałem" opisanym przez Gigerenzera wynik można opisać następująco: Badanie „wsparło" hipotezę badawczą o tym, że rzucano monetą z dwoma orłami (w = 1, $p < 0{,}001$). Zauważmy, że to rozwiązanie:

- nigdy nie potwierdzi M_1 o symetryczności monety, gdyż konstrukcja testu z góry zakłada prawdziwość hipotezy H_0 i – co najwyżej – test nie da podstaw do jej odrzucenia,
- jest pośrednie: NHST zajmuje się H_0 a nie H_1, chociaż to właśnie z interpretacji hipotezy alternatywnej H_1 wyprowadzony zostaje wniosek badawczy,
- zazwyczaj nie rozróżnia potencjalnych modeli objętych hipotezą alternatywną[6], takich jak na przykład następujące dwa (oprócz M_2):
  - M_3: moneta była fałszywa, częściej wyrzuca orły,
  - M_4: moneta może być równie dobrze symetryczna, jak przeciążona, ale nie wiemy zupełnie nic o kierunku jej skrzywienia,
- całkowicie pomija wiedzę o tym, jak często w podobnych sytuacjach używa się monet symetrycznych i przeciążonych lub fałszywych, czyli o prawdopodobieństwie modeli *a priori* (a przecież, ponieważ zdecydowanie częściej w życiu spotykamy monety symetryczne, to użycia takiej monety powinniśmy bardziej się spodziewać, jeszcze zanim poznamy rezultat dziesięciu rzutów monetą).

Podsumowując, NHST nie daje rozstrzygających wskazówek dla hipotezy badawczej mimo tego, że polega na podjęciu dwóch rozłącznych decyzji: albo odrzucić,

do symetryczności monety. Być może eksperymentator jest iluzjonistą manipulującym rzucanymi monetami i używającym monety z dwoma orłami (o takiej możliwości dowiedzieliśmy się właśnie od znajomego iluzjonisty – zastosowałby taką monetę właśnie dlatego, że mało kto by się tego spodziewał).

[6] Piszemy tu o praktyce badawczej np. z wykorzystaniem popularnych pakietów statystycznych. W podręcznikach wyróżnia się hipotezy proste i złożone, często na tych pierwszych ilustrując pojęcie mocy testu. Zauważmy na przykład, że jeśliby przyjęto $\alpha = 0{,}05$ i postawiono hipotezę alternatywną prostą H_1: $\theta = 1$ (rozważanie wyłącznie modelu M_2 jako alternatywy dla M_1), to $\beta = 0$ już dla $n \geq 5$ rzutów. W praktyce badacze unikają jednak zbyt dokładnej specyfikacji hipotezy alternatywnej, a nawet postawienia hipotezy kierunkowej, gdyż nie chcą być posądzeni o „*p*-hacking", czyli manipulacje analizami tak, by obniżyć wartość p – zwykle poniżej poziomu istotności $\alpha = 0{,}05$ (zob. Head, Holman, Lanfear, Kahn i Jennions, 2015).

 DOI: 10.7206/DEC.1733-0092.79

albo nie odrzucać H_0. Wysoka wartość *p* nigdy nie sankcjonuje przyjęcia hipotezy zerowej, ponieważ sama konstrukcja testu opiera się na założeniu jej prawdziwości. Odrzucenie hipotezy zerowej nie oznacza natomiast, że automatycznie hipoteza alternatywna jest prawdziwa. Asymetrię wniosków wyciąganych w NHST na podstawie wartości *p* można opisać następująco:

- jeżeli $p < \alpha$, to należy odrzucić H_0 na rzecz hipotezy alternatywnej H_1; przyjęcie H_1 wiąże się z ryzykiem popełnienia błędu I rodzaju;
- jeżeli $p \geq \alpha$, to nie ma podstaw do odrzucenia H_0, ale nie wiadomo, czy jest tak dlatego, że H_0 jest prawdziwa, czy może dlatego, że badanie nie jest w stanie tego rozstrzygnąć (bo np. zebrano za mało danych); przyjęcie H_0 wiąże się z ryzykiem popełnienia błędu II rodzaju.

Tymczasem przeprowadziwszy testy, badacz chciałby po prostu dowiedzieć się, czy uzyskane wyniki:

- świadczą na korzyść hipotezy zerowej H_0,
- świadczą na korzyść hipotezy alternatywnej H_1,
- są niekonkluzywne i wskazują na potrzebę zebrania większej ilości dowodów.

Metodą pozwalającą ocenić wiarygodność konkurencyjnych hipotez w tych trzech kategoriach jest czynnik bayesowski.

2. Czynnik bayesowski, BF (ang. *Bayes Factor*)

2.1. Prawdopodobieństwo jako miara siły przekonań i dowodów

Analiza czynnika bayesowskiego należy do ogólniejszej klasy metod bayesowskich (np. bayesowskie metody wnioskowania, uwzględniające funkcje wiarygodności i funkcje strat, przedstawiono w pracach: Jóźwiak i Podgórski, 2001, s. 321–330 oraz Domański i Pruska, 2000, s. 240–259). W metodach tych nowe informacje, np. w postaci wyników badań empirycznych, traktuje się jako potencjalne dowody na rzecz postawionych hipotez. Miarą wiarygodności hipotez jest przypisane im prawdopodobieństwo przed i po uzyskaniu informacji. Niezależnie od źródeł wiedzy o prawdopodobieństwie – czy są nimi subiektywne przekonania, czy uzasadnienia odwołujące się do obiektywnych danych statystycznych – wymaga się spełnienia zasad aksjomatycznej teorii prawdopodobieństwa (Ramsey, 1931; Dienes, 2011; Vallverdú, 2015). Jest to inny sposób pojmowania prawdopodobieństwa, niż ma to miejsce w testach NHST, w których prawdopodobieństwo jest przede wszystkim wskaźnikiem nietypowości danych w rozważanej populacji.

 DOI: 10.7206/DEC.1733-0092.79

Dla lepszego zrozumienia różnic w rozumieniu prawdopodobieństwa w podejściu częstościowym i w podejściu bayesowskim, punktem wyjścia niech będzie pytanie o to, w jakim celu badacz gromadzi dane empiryczne lub jakieś inne informacje. Otóż prowadzenie badań służy gromadzeniu dowodów przemawiających za lub przeciw pewnym hipotezom. O tych dowodach można powiedzieć, że są mniej lub bardziej przekonujące, mocne, niezbite, słabe; mawia się też, że w tych danych dowodów na poparcie jakiejś hipotezy brak. Jednak w podejściu częstościowym (frekwentystycznym), stosując testy istotności, nie waży się argumentów, nie mówi się o tym, że dane uprawdopodobniają jedną hipotezę bardziej niż drugą. Hipotezy statystyczne stanowią dowolne przypuszczenia o rozkładach zmiennych losowych w populacji ze stałymi, choć niekoniecznie znanymi, parametrami (zob. np. Pawłowski, 1976, s. 134; Jóźwiak i Podgórski, 2001, s. 235; Koronacki i Mielniczuk, 2001, s. 213). W testach NHST na podstawie istotności policzonych statystyk podejmuje się więc decyzje o tym, by hipotezy odrzucać, lub żeby ich nie odrzucać. Wartości *p* nie używa się, wbrew opisanym wcześniej błędnym interpretacjom, do oceny siły dowodu, lecz jedynie jako kryterium oceny położenia statystyki testu[7]. Żeby zaś mówić o typowości lub nietypowości tej statystyki, trzeba wpierw określić populację, w obrębie której wartości zmiennej losowej i różne statystyki przyjmują różne wartości. Podsumowując, w testach NHST prawdopodobieństwo opisuje dane i statystyki testu, a nie przekonania lub hipotezy.

We wprowadzonym tu podejściu bayesowskim jest odwrotnie: pewne informacje lub dane z próby traktuje się jako ustalony fakt i z perspektywy tych danych „odgaduje się", na ile wiarygodne i uzasadnione są sądy na temat różnych rzeczy (czyli właśnie hipotezy). W podejściu tym prawdopodobieństwo traktuje się jako naturalną miarę siły dowodu (Morey, Romein i Rouder, 2016). Łatwo się przekonać, że jest ono obecne w potocznym myśleniu ludzi, w którym i dowodom, i hipotezom przypisuje się różne stopnie wiarygodności (być może stąd się biorą tak częste nadinterpretacje NHST).

Weźmy na przykład prognozę pogody, w której synoptycy zapowiedzieli kilka słonecznych dni. Obudziwszy się nazajutrz, widzimy przez okno, że jest jednak ciemno i pochmurno. Możliwe, że będzie padać. Przechodnie na ulicy za oknem chodzą w kurtkach i z parasolami pod pachą, będzie lało jak nic. Jednak na horyzoncie przejaśnia się. Być może synoptycy mieli jednak rację? Ta prosta historyjka pokazuje nam trzy rzeczy. Po pierwsze, ludzie w naturalny sposób przypisują różne szanse swym przypuszczeniom, oceniając ich trafność lub wiarygodność. Po drugie, przekonanie zmienia się pod wpływem zastanych okoliczności i napływających infor-

[7] Po pierwsze, w podejściu N – P nieważna jest wartość *p*, lecz to, czy przekroczyła poziom istotności α, czy nie. Po drugie, trudno uznać wartość *p* za miarę siły dowodu, skoro ma jakościowo różne znaczenie w kontekście wielkości próby. Na przykład $p = 0{,}032$ w teście t-Studenta dla prób niezależnych ma inny wydźwięk przy próbach o liczebności $n_1 = 15$ i $n_2 = 15$, a inny – przy $n_1 = 150$ i $n_2 = 150$ (taka sama istotność wskazuje na silniejszy efekt eksperymentalny przy mniejszej próbie lub na większą moc testu przy większej).

 DOI: 10.7206/DEC.1733-0092.79

macji. Prawdopodobieństwa warunkowe, którymi wyrażono, choć w sposób wysoce nieprecyzyjny, niepewność co do tego, że spadnie deszcz – P(deszcz| prognoza), P(deszcz| pochmurno), P(deszcz| parasolki), P(deszcz| przejaśnia się) – są coraz to inne. Zwiększają się, gdy napływają informacje kojarzone z pogodą deszczową, zmniejszają – gdy z pogodą słoneczną. Po trzecie, że same dowody są mniej lub bardziej przekonujące, co też można wyrazić w kategoriach szans. Zachmurzenie wydaje się oznaką deszczu bardziej wiarygodną niż noszenie parasolek (P(zachmurzenie| zanosi się na deszcz) > P(parasolki| zanosi się na deszcz)).

Z perspektywy testów NHST takie probabilistyczne wypowiedzi nie mają sensu, ponieważ nie zdefiniowano wcześniej populacji, której dotyczą (dni, godziny, ludzie czy inne jednostki statystyczne). Prawdopodobieństwo przypisano tu samym hipotezom, a nie danym w rozkładach jakichś zmiennych losowych. Niemniej jednak, jak pokazały to badania cytowane na początku niniejszego tekstu, użytkownicy testów statystycznych mają naturalną skłonność do takiej właśnie interpretacji.

Podejście bayesowskie wychodzi z definicji prawdopodobieństwa jako miary niepewności sądów, przekonań, orzeczeń. Do jej oceny można użyć jakichś danych opisujących populacje, można jakichś innych ocen – ważne, by spełnić warunki narzucone przez aksjomatyczną teorię prawdopodobieństwa. Głównym zadaniem analizy bayesowskiej jest analiza zmienności przekonań w kontekście zaobserwowanych faktów. Jakieś dane (obserwacje, informacje) są tu więc ustalone i to z ich perspektywy prowadzone jest wnioskowanie.

2.2. Parametr jako zmienna losowa

W podejściu frekwentystycznym (częstościowym) przyjmuje się, że parametr θ rozkładu zmiennej losowej – lub ogólniej model M, którego dotyczy wnioskowanie statystyczne – jest ustalony. Jest on wprawdzie nieznany, lecz to względem niego określa się losowość uzyskiwanych wyników, czyli zebranych danych D (pomiarów i policzonych statystyk). Celem badania jest pokazanie, na ile dane D są prawdopodobne przy założeniu prawdziwości M lub θ (M lub θ określają sposób sformułowania hipotezy zerowej).

W przyjmowanym tu podejściu bayesowskim jest inaczej: parametr θ jest zmienną losową, zaś dane D – ustalonym faktem. Rozróżnia się dwa typy rozkładu parametru θ. Rozkład *a priori* $P(\theta)$ reprezentuje wiedzę początkową o szansach różnych wartości parametru θ przed zebraniem danych D. Rozkład *a posteriori* $P(\theta \mid D)$ stanowi natomiast modyfikację rozkładu *a priori* po uwzględnieniu danych D. Dane D, w postaci na przykład danych empirycznych, są ustalonym faktem i służą aktualizacji wiedzy o rozkładzie parametru.

 DOI: 10.7206/DEC.1733-0092.79

Przykład. Złudzenie gracza – cd.

Podejście częstościowe. Punktem odniesienia był rozkład częstości uzyskiwanych orłów w dziesięciokrotnych rzutach monetą symetryczną, charakteryzowany parametrem $\theta = 0{,}5$. Czy rzeczywiście moneta taka jest – tego nie wiemy. Odpowiedni wniosek wyciągniemy, testując istotność danych D w postaci wyrzuconych 10 orłów w rozkładzie dwumianowym B(10; 1/2) (jak wcześniej pokazano, $p = 1/1024 < \alpha = 0{,}001$, co kazało odrzucić hipotezę zerową o symetryczności monety).

Podejście bayesowskie. Odsetek orłów, jakie wyrzuca moneta, jest zmienną losową θ, gdyż nie wiemy tak naprawdę, z jaką monetą mamy do czynienia. Dane w postaci dziesięciu orłów w dziesięciu rzutach monetą można uzyskać nawet dla monet wyrzucających w większości reszki. Jest to możliwe na przykład dla monety, która wyrzuca zaledwie 10% orłów i 90% reszek (dla której $\theta = 0{,}1$), choć szansa takiego wyniku wynosi zaledwie jeden na dziesięć miliardów, $P(D|\theta = 0{,}1) = 0{,}1^{10}$. Szanse uzyskania dziesięciu orłów są oczywiście wyższe dla monet o wyższych wartościach parametru θ. Uzyskawszy dziesięć orłów, spodziewamy się więc raczej wysokiej wartości θ, gdyż to właśnie wtedy moneta będzie spadać przeważnie orłem do góry. Rozkład parametru θ można wyznaczyć, posługując się regułą Bayesa.

2.3. Reguła Bayesa

Badania i analizy danych „nie wiszą w próżni" i poprzedzone są zazwyczaj jakimiś przemyśleniami. Mniej lub bardziej sceptyczny stosunek do testowanej hipotezy może wynikać z analiz teoretycznych, obserwacji, argumentacji, wcześniejszych badań itd. Przekonanie badacza o prawdziwości pewnej hipotezy (modelu, teorii itp.) i to, jak się zmienia pod wpływem zebranych informacji, opisać można prawdopodobieństwem wyznaczanym według reguły Bayesa.

O wiarygodności hipotezy H badacz ma już jakieś przekonanie, zanim pozyska informację, np. w postaci wyników badania empirycznego. Wiedzę tę opisuje prawdopodobieństwo *a priori* P(H). Dane D weryfikują to przekonanie, za ich pomocą badacz aktualizuje wiedzę, zmieniając prawdopodobieństwo P(H) na prawdopodobieństwo *a posteriori* P(H|D). Żeby z danych wnioskować o hipotezie, trzeba wiedzieć, na ile wiarygodnie hipoteza H przewiduje wystąpienie tych danych – wyraża to prawdopodobieństwo warunkowe P(D|H) – oraz na ile możliwe w ogóle jest uzyskanie danych D w różnych okolicznościach, P(D). Aby zaktualizować przekonanie o wiarygodności hipotezy H, prawdopodobieństwo *a posteriori* wyznacza się według reguły Bayesa (Bayes i Price, 1763):

$$P(H|D) = \frac{\mathrm{P(D|H)P(H)}}{\mathrm{P(D)}}$$

 DOI: 10.7206/DEC.1733-0092.79

Prawdopodobieństwo całkowite P(D) pokazuje, na ile możliwe jest uzyskanie takich a nie innych danych przy uwzględnieniu wszystkich hipotez: $P(D) = \sum_{i=1}^{k} P(D|H_i)P(H_i)$. Przyjrzyjmy się zastosowaniu reguły Bayesa do analizy przykładowych danych.

Przykład 1. Taksówki *(na podstawie: Kahneman i Tversky, 1972 oraz Tyszka, 2010, s. 242)*

W pewnym mieście po zmroku przechodzień został potrącony przez kierowcę taksówki, który zbiegł z miejsca wypadku. Wiadomo też, że 85% taksówek jeżdżących po mieście należy do firmy „Zielone" a 15% do firmy „Niebieskie". Przesłuchany świadek twierdzi, że taksówka była niebieska. Okazuje się jednak, że jego zeznanie jest wiarygodne na 80%, gdyż w 20% takich warunków myli on kolor niebieski z zielonym. Jaka jest więc szansa, że była to taksówka niebieska? A jaka – że zielona?

Nazwijmy hipotezą zerową przypuszczenie, że taksówka była niebieska (H_0), alternatywną – że taksówka była zielona (H_1). Szanse prawdziwości tych hipotez *a priori* – czyli przed zebraniem danych *D* w postaci zeznania świadka – wynoszą $P(H_0) = 0{,}15$ i $P(H_1) = 0{,}85$. Szanse poprawnego rozpoznania taksówki wynoszą $P(D|H_0) = 0{,}8$, zaś błędnego – zielonej jako niebieskiej – $P(D|H_1) = 0{,}2$. Zaktualizowane prawdopodobieństwo *a posteriori* dla hipotezy zerowej wynosi:

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D)} = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0) + P(D|H_1)P(H_1)} =$$

$$= \frac{0{,}8 \times 0{,}15}{0{,}8 \times 0{,}15 + 0{,}2 \times 0{,}85} \cong 0{,}4138.$$

Szansa *a posteriori* dla hipotezy alternatywnej, obliczona według tego samego sposobu, wynosi $P(H_1|D) \cong 0{,}5862$. A zatem, mimo że dane uwiarygodniły hipotezę zerową o niebieskiej taksówce, nadal bardziej prawdopodobna jest hipoteza alternatywna, że taksówka jest zielona.

Przykład. Złudzenie gracza – cd. Załóżmy, że nasze wcześniejsze doświadczenia podpowiadają nam, że w grę wchodzi:

- albo użycie monety symetrycznej, M_1: $\theta = 1/2$,
- albo monety fałszywej z dwoma orłami, M_2: $\theta = 1$,
- albo monety przeciążonej, wyrzucającej częściej orły, M_3: $\theta \in (1/2;1)$.

Dodatkowo zakładamy – w oparciu o wcześniejsze doświadczenia z podobnymi zadaniami – że monet fałszywych używa się w takich sytuacjach stukrotnie rzadziej

 DOI: 10.7206/DEC.1733-0092.79

niż symetrycznych, a przeciążonych dziesięciokrotnie rzadziej niż symetrycznych. Prawdopodobieństwo *a priori* rozważanych modeli wynosi wtedy:

- $P(M_1) = 100/111$ czyli 90,09%,
- $P(M_2) = 1/111$, czyli 0,90% oraz
- $P(M_3) = 10/111$, czyli 9,01%.

Obliczamy szanse uzyskania 10 orłów dla każdego z tych modeli, kolejno: $P(D|M_1) = (1/2)^{10}$, $P(D|M_2) = 1^{10} = 1$, a $P(D|M_3) = \theta^{10}$ dla wszystkich $\theta \in (1/2;1)$ (M_3 opisuje zmienną ciągłą z funkcją gęstości $f(\theta) = 20/111$). Ustaliliśmy już prawdopodobieństwo *a priori* każdego modelu oraz szansę uzyskania dziesięciu orłów dla każdego z nich. Należy jeszcze ustalić prawdopodobieństwo całkowite uzyskania dziesięciu orłów dla wszystkich trzech modeli. Wynosi ono:

$$P(D) = P(M_1) \times P(D|M_1) + P(M_2) \times P(D|M_2) + P(M_2) \times P(D|M_2) =$$
$$= (100/111) \times (1/2)^{10} + (1/111) \times 1^{10} + (10/111) \times \int_{1/2}^{1} (1/(1-1/2)) \times \theta^{10} d\theta =$$
$$= 0{,}00088 + 0{,}00901 + 0{,}01637 = 0{,}02626.$$

Wykorzystując powyższe obliczenia, oceniamy teraz szanse *a posteriori* tego, że zastosowano rozważane typy monet. Według formuły Bayesa:

- $P(M_1|D) = P(M_1) \times P(D|M_1) / P(D) = 0{,}00088/0{,}02626 = 0{,}03350 \approx 3{,}35\%$
- $P(M_2|D) = P(M_2) \times P(D|M_2) / P(D) = 0{,}00901/0{,}02626 = 0{,}34298 \approx 34{,}31\%$
- $P(M_3|D) = P(M_3) \times P(D|M_3) / P(D) = 0{,}01637/0{,}02626 = 0{,}62352 \approx 62{,}34\%$.

W świetle danych *D* najbardziej przekonujące jest wyjaśnienie, że moneta jest po prostu przeciążona (M_3) – szanse na to wzrosły z przewidywanych 9,01% do prawie dwóch trzecich ($P(M_3|D) = 62{,}34\%$). Mniej prawdopodobne jest to, że moneta ma dwa orły (szanse tego przypuszczenia wzrosły z 0,90% do $P(M_2|D) = 34{,}31\%$). *A posteriori* najmniej prawdopodobne jest użycie monety symetrycznej. Wiarygodność tego przypuszczenia spadła z 90,09% do zaledwie $P(M_1|D) = 3{,}35\%$.

Czytelnik może samodzielnie przekonać się, na ile zmienią się oceny *a posteriori*, jeśli przyjmie się inne wartości prawdopodobieństwa *a priori* dla modeli M_1 – M_3 (uzna się na przykład, że monety z dwoma orłami są znacznie częstsze, niż tu założono). Różne wartości *a posteriori* mogą sprawiać poczucie subiektywizmu i tego, że badacz może uzyskiwać takie wyniki, jak chce, o ile przypisze swoim wejściowym przypuszczeniom odpowiednio duże szanse. Nie jest to jednak prawdą, gdyż zebranie odpowiedniej liczby dowodów (obserwacji) w dłuższym rozrachunku powinno każdego doprowadzić do tego samego wniosku[8].

[8] Dopasowywanie się przekonań pod wpływem napływających danych jest możliwe tylko wtedy, gdy spełniona jest tzw. zasada konwergencji (zob. np. Gaifman i Snir, 1982). Opisuje ona upodabnianie się funkcji

2.4. Czynnik bayesowski

W bayesowskim testowaniu hipotez kładzie się nacisk na pokazanie przewagi jednych hipotez (teorii, modeli itp.) nad drugimi w wyjaśnianiu lub przewidywaniu obserwowanych danych (uzyskanych informacji, wyników badań empirycznych itd.). Załóżmy, że zaobserwowano dane D i rozważane są dwie hipotezy, zerowa H_0 i alternatywna H_1. Zgodnie z regułą Bayesa prawdopodobieństwo *a posteriori* każdej z nich wynosi: $P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D)}$ i $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D)}$. Dzieląc te wyrażenia stronami, otrzymujemy:

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

Czynnikiem bayesowskim (ang. *Bayes Factor*) nazywamy – występujący w tym równaniu – iloraz, pokazujący, ile razy bardziej prawdopodobne jest uzyskanie danych D wtedy, gdy prawdziwa jest hipoteza H_1, niż wtedy, gdy prawdziwa jest hipoteza H_0:

$$BF_{10} = \frac{P(D|H_1)}{P(D|H_0)}$$

Czynnik bayesowski BF_{10} pokazuje, ile razy lepszym wyjaśnieniem danych D jest hipoteza alternatywna od hipotezy zerowej[9]. Jednocześnie możliwa jest interpretacja wskaźnika „w drugą stronę" jako znaczenia danych. Widać to, gdy przekształcimy powyższe wzory do postaci:

$$P(H_1|D) = S \times BF_{10} \times P(H_0|D)$$

Symbolem S oznaczono proporcję prawdopodobieństw *a priori*. Jak widać, im wyższe są wartości BF_{10}, tym wyższe jest prawdopodobieństwo *a posteriori* dla hipotezy H_1, $P(H_1|D)$, w porównaniu do prawdopodobieństwa *a posteriori* dla hipotezy H_0, $P(H_0|D)$. Wskaźnik BF_{10} wyraża więc stopień, w jakim dane D przechylają szalę dowodu na korzyść hipotezy H_1 względem hipotezy H_0.

Czynnik bayesowski sam w sobie nie zawiera ani szans *a priori*, ani szans *a posteriori*. Zarówno więc zwolennik testowanej hipotezy H_1 (wysoka wartość S),

prawdopodobieństwa pod wpływem aktualizacji dostateczną liczbą danych. O tych funkcjach mówi się, że powinny być kompatybilne, tzn. nie może być tak, że dla pewnych zdarzeń (a w kontekście rozważań w naszej pracy: dla jakichś wartości parametru), jedna z nich przyjmuje wartość zerową, a druga nie. Innymi słowy, zasada konwergencji nie jest spełniona, jeśli np. wartości parametru uznane za możliwe przez jednego badacza uznane będą za niemożliwe przez drugiego badacza.

[9] Na pierwszy rzut oka czynnik bayesowski może wydawać się innym określeniem tzw. ilorazu wiarogodności (ang. *likelihood ratio*, LR). Są to rzeczywiście pojęcia podobne i często tak samo zapisywane. W literaturze przedmiotu spotyka się stanowisko utożsamiające iloraz wiarogodności z czynnikiem bayesowskim (Nickerson, 2000). Nie są one ze sobą jednak tożsame. Jak zauważają Kass i Raftery (1995, s. 776), poza celem, w jakim się je zwykle oblicza, różni je sposób obliczania. Gdy dwie konkurencyjne hipotezy są proste (jak w przykładzie z taksówkami), czynnik bayesowski oblicza się tak samo jak iloraz wiarogodności. Gdy natomiast hipotezy są złożone, inny sposób ich obliczania wskazuje na odmienność tych pojęć.

 DOI: 10.7206/DEC.1733-0092.79

jak i jej przeciwnik (niska wartość *S*), powinien tak samo interpretować jego wartości jako uprawdopodobnienie hipotezy H_1 względem H_0. Natomiast to, jakie ktoś ostatecznie wyciągnie wnioski o prawdziwości tych dwóch hipotez, zależy od szans *a posteriori*. Poza wskaźnikiem *BF* kształtują je prawdopodobieństwa *a priori*, których proporcja *S* zależy od mniej lub bardziej sceptycznego lub entuzjastycznego nastawienia badacza do testowanej hipotezy.

Wskaźnik *BF* pozwala ocenić, ile razy lepiej jedna z hipotez (teorii, modeli itd.) przewiduje zaobserwowane fakty niż druga lub – co na jedno wychodzi – na korzyść której z tych hipotez dane bardziej przemawiają. Wartości czynnika równe 1 oznaczają, że uzyskanie danych było tak samo możliwe dla każdej z hipotez. Wartości $BF_{10} > 1$ (albo $BF_{01} < 1$ – jeśli we wzorze zamienimy licznik z mianownikiem miejscami) oznaczają przewagę hipotezy H_1, a poniżej 1 (albo $BF_{01} > 1$) – przewagę hipotezy H_0. Chociaż interpretacja czynnika bayesowskiego jest jednoznaczna i ciągła, dla wygody opisu wyników zaproponowano stopniowanie słownych określeń przewagi jednej hipotezy nad drugą (zob. np. Jeffreys, 1939/1961; Kass i Raftery, 1995; Wetzels i in., 2011). Jeden z najpopularniejszych sposobów opisu siły dowodu przedstawiono w tabeli 1. Gdy wskaźnik *BF* przyjmuje bardzo małe lub bardzo duże wartości, wygodnie jest się posłużyć logarytmem dziesiętnym ze wskaźnika, log *BF*.

Tabela 1
Interpretacja wielkości czynnika bayesowskiego BF jako siły dowodu

BF_{10}	log BF_{10}	BF_{01}	log BF_{01}	Poparcie dla H_1 względem H_0
1	0	1	0	takie samo
1-3	0 – 0,48	1/3 – 1	-0,48 – 0	niewystarczające (ang. *anecdotal)*
3-10	0,48 – 1	1/10 – 1/3	-1 – -0 ,48	znaczące (ang. *substantial*)
10 – 30	1 – 1,48	1/30 – 1/10	-1,48 – -1	silne
30 – 100	1,48 – 2	1/100 – 1/30	-2 – -1,48	bardzo silne
> 100	> 2	< 1/100	< -2	zdecydowane

Opracowanie na podstawie: Wetzels i in. (2011).

Przykład. Taksówki, cd.

Uzyskanie zeznania świadka, że taksówka była niebieska, jest w 80% możliwe, jeśli taksówka rzeczywiście taka była, $P(D|H_0) = 0{,}8$ i w 20% – jeśli była zielona, $P(D|H_1) = 0{,}2$. Czynnik bayesowski wynosi $BF_{10} = \frac{P(D|H_1)}{P(D|H_0)} = \frac{0{,}2}{0{,}8} = 0{,}25$. Przeprowadzone badanie – zeznanie świadka – jest więc znaczącym, czterokrotnie silniejszym, dowodem za hipotezą H_0, że taksówka była niebieska, niż za H_1 – że była zielona.

Mimo że hipoteza H_0 wyjaśnia uzyskane dane lepiej niż H_1, ostateczna interpretacja zeznania zależy od wiedzy początkowej o tym, że więcej jest taksówek zielo-

 DOI: 10.7206/DEC.1733-0092.79

nych (85%) niż niebieskich (15%). Choć więc ciężar dowodu przechylił się znacząco w stronę H_0 i szanse na to, że taksówka była niebieska, wzrosły z 15% do 41,3%, nadal bardziej prawdopodobne jest to, że taksówka była zielona, $P(H_1|D) = 59,7\%$.

Przykład. Złudzenie gracza, cd.

Trafność trzech przypuszczeń o tym, jaką monetą rzucano, ocenimy, licząc trzy czynniki bayesowskie według wzoru:

$$\frac{P(M_j|D)}{P(M_k|D)} = BF_{jk} \times \frac{P(H_j)}{P(H_k)}$$

Przypomnijmy, prawdopodobieństwa *a priori* dla tych trzech modeli wyniosły kolejno 90,09%, 0,90% i 9,01%, natomiast prawdopodobieństwa *a posteriori* – 3,35%, 34,31% i 62,34%.

Dla pierwszych dwóch modeli, M_1 – monety rzetelnej i M_2 – monety z dwoma orłami, mamy: $\frac{3,35\%}{34,31\%} = BF_{12} \times \frac{90,09\%}{0,90\%}$, skąd $BF_{12} = 0,000977 = 1/1024$. Otrzymana wartość oznacza, że prawdopodobieństwo uzyskania takich danych było ponad tysiąckrotnie wyższe dla monety z dwoma orłami, niż dla monety symetrycznej. Analogiczne obliczenia dla porównania modeli M_1 i M_2 z modelem M_3 – z hipotezą, że moneta była przeciążona – wynoszą: $BF_{13} = 0,005374 \approx 1/186$ i $BF_{23} = 5,503$. Pierwszy ze wskaźników mówi o tym, że uzyskane wyniki były dla modelu M_1 około 186-krotnie mniej prawdopodobne, niż dla M_3, a drugi – że około 5,5-krotnie bardziej prawdopodobne dla modelu M_2 niż M_3.

Interpretując policzone czynniki według stopni zaproponowanych w tabeli 1, można stwierdzić, że uzyskanie dziesięciu orłów w dziesięciu rzutach monetą zdecydowanie bardziej wspiera przypuszczenie, że moneta miała dwa orły, niż przypuszczenie, że była symetryczna i wskazuje na znaczącą przewagę tego modelu nad przypuszczeniem, że rzucano monetą przeciążoną. Należy jednak pamiętać, że policzone czynniki stanowią ocenę samego badania jako dowodu za poszczególnymi hipotezami, a nie wskazówka, którą z hipotez należy „przyjąć". Prawdopodobieństwa *a posteriori* wskazują, że chociaż użycie monety symetrycznej było mało prawdopodobne, to użycie monety przeciążonej było prawie dwukrotnie bardziej prawdopodobne niż monety z dwoma orłami.

Weryfikacja hipotez statystycznych za pomocą czynników bayesowskich nie jest pomysłem nowym. Pierwsze propozycje ich wykorzystania można znaleźć już w pracach sprzed kilkudziesięciu lat (zob. np. Jeffreys, 1939/1961; Edwards, Lindman i Savage, 1963; Hays, 1973). Wyznaczanie prawdopodobieństwa jest jednak bardzo pracochłonne i skomplikowane od strony algebraicznej, gdyż reguła Bayesa pociąga

 DOI: 10.7206/DEC.1733-0092.79

za sobą łączenie rozkładów prawdopodobieństwa różnych typów (i wymaga np. intensywnych obliczeń przy wyznaczaniu całek z opisujących je funkcji). Już nawet tak prosty przykład, jak ten ze złudzeniem gracza, zawierał analizę rozkładu mieszanego zmiennej losowej (prawdopodobieństwo było skokowo opisane dla wartości 1/2 i 1 i w sposób ciągły dla przedziału między tymi wartościami). Dawniej trudności analityczne próbowano pokonać, wykorzystując na przykład do opisu prawdopodobieństwa *a priori* P(H) i wiarygodności hipotez P(D|H) takie rozkłady, które gwarantowałyby uzyskanie rozkładów *a posteriori* należących do tej samej rodziny (zob. np. Hays, 1973, s. 820–821). Zabiegi takie były jednak kolejnym argumentem dla przeciwników stosowania metod bayesowskich, zarzucających im subiektywizm (Pawłowski, 1976). Współcześnie sytuacja zmieniła się dzięki wzrostowi mocy obliczeniowej komputerów i implementacji technik symulacyjnych, takich jak na przykład metody Monte Carlo oparte na łańcuchach Markowa (ang. *Markov chain Monte Carlo*, w skrócie MCMC). Umożliwiło to opracowanie programów komputerowych do bayesowskiej weryfikacji hipotez statystycznych (np. JASP; zob. np. Wagenmakers, Morey, Lee, 2016).

3. Przykłady wykorzystania czynnika Bayesa *BF* w praktyce

W tej części przedstawimy przykłady praktycznego zastosowania czynnika bayesowskiego do weryfikacji hipotez statystycznych. Porównamy wnioski płynące z zastosowania tej metody z wnioskami wynikającymi z zastosowania procedury NHST.

W naszych analizach wykorzystaliśmy darmowy program JASP 0.8 (https://jasp-stats.org/), wykorzystujący rozwijane równolegle przez jego autorów moduły pakietu statystycznego R. Jedną z zalet programu JASP jest to, że umożliwia wyznaczenie i analizę *BF* na podstawie statystyk zwyczajowo raportowanych w artykułach naukowych, bez dostępu do pełnego zbioru danych. Do wykonania bayesowskich odpowiedników testów istotności, takich jak test t-Studenta, test Chi-kwadrat, test korelacji, wystarcza znajomość liczebności grup, średnich grupowych, wartości statystyki t-Studenta, współczynnika *r* Pearsona itp.

Na początek ilustracją wykorzystania czynnika bayesowskiego z zastosowaniem JASP niech będzie reinterpretacja wyników badania Baumeistera i in. (1998), komentowanego na początku niniejszego artykułu.

Przykład. *Badanie nad wyczerpaniem ego (Baumeister i in., 1998)*

Przypomnijmy, w badaniu tym porównywano wyniki trzech grup: G1 – jedzący rzodkiewki, G2 – jedzący czekoladę, oraz G3 – grupa kontrolna. Zmienna zależną był

 DOI: 10.7206/DEC.1733-0092.79

czas wykonywania zadania, w minutach. Badacze stwierdzili, że osoby, które musiały powstrzymać się od zjedzenia czekolady i zjadły rzodkiewki, wyczerpały swoje ego, w związku z tym wcześniej porzucały zadanie niż osoby z pozostałych grup. Porównując grupy, postulowano zatem następującą zależność między typowymi czasami wykonywania zadania: $\mu_1 < \mu_2, \mu_1 < \mu_3$ i $\mu_2 = \mu_3$. Oszacowanie czynnika *BF* możliwe jest nie tylko na podstawie danych surowych, lecz także statystyk z próby: średnich grupowych, liczebności grup i wartości statystyki *t*. Badacze nie podali dokładnej wartości statystyki *t*, ale korzystając ze średnich i odchyleń standardowych, wyznaczyliśmy dokładną wartość $t(40) = 0{,}801$. (Czytelnik może samodzielnie użyć programu JASP, aby prześledzić ten przykład. Należy w tym celu użyć niedostępnej domyślnie zakładki „statystyki zbiorcze": Modules – Summary Stats). Badanie silnie wsparło przypuszczenie o różnicy między grupami pierwszą i drugą, $BF_{10} = 37{,}614$ i bardzo silnie, zdecydowanie – między grupami pierwszą i trzecią, $BF_{10} = 525{,}87$. Dla różnicy między grupami drugą i trzecią, uznanej przez badaczy za nieistotną, uzyskaliśmy czynnik $BF_{01} = 2{,}8$, sugerujący, że dane są niekonkluzywne: nie przemawiają ani za hipotezą o różnicy między tymi dwiema grupami, ani za jej brakiem, choć nieco przechylają szalę dowodu na rzecz braku różnic. Wnioski Baumeistera i współpracowników wydają się zatem uzasadnione, choć przypuszczenie, że osoby, które zjadły czekoladę, nie różnią się od tych, które nie jadły nic, ma słabe wsparcie i wymagałoby zebrania większej ilości danych.

W powyższym przykładzie wnioski oparte na błędnej interpretacji wartości *p* zostały potwierdzone przez wskaźnik *BF*. Nie jest to jednak regułą. W ostatnich latach coraz częściej badacze śledzą wyniki raportowane przez innych, i wykazują wady wnioskowania statystycznego. Na przykład Aczel, Palfi, Szaszi, Szollosi i Dienes (2015) pokazali, że podobny błąd popełnili autorzy artykułu opublikowanego w „Science" (Hu i in., 2015), którzy twierdzili, że stosując wymyśloną przez nich specjalną technikę, można podczas snu oduczyć się uprzedzeń wobec innych. Dla ich tezy ważne było pokazanie, że nie zachodzi interakcja między czynnikami w przeprowadzonej analizie wariancji. Brak istotności efektów interakcyjnych uznali więc za dowód przeciwko ich występowaniu. Aczel i in. (2015), wykorzystawszy czynnik bayesowski, podważyli ten wniosek i pokazali, że o ile siła uprzedzeń zaraz po obudzeniu była mniejsza (efekt główny), to brak jest dowodów, że efekt ten utrzymuje się dłużej w wyniku zastosowania techniki, niż bez jej stosowania.

W dalszej części pokażemy, jak weryfikować hipotezy statystyczne za pomocą czynnika bayesowskiego *BF*. Porównamy interpretację tego czynnika z wnioskami wyciąganymi na podstawie wartości *p* w analogicznych testach istotności. Przyjrzymy się powszechnie używanym metodom: analizie tabel krzyżowych (z testem niezależności Chi-kwadrat), testom różnic (test t-Studenta dla prób niezależnych oraz ANOVA) i analizie korelacji. Jako źródło danych wybraliśmy rzeczywiste wyniki opublikowanych badań empirycznych.

 DOI: 10.7206/DEC.1733-0092.79

3.1. Bayesowskie testy niezależności Chi-kwadrat

Tyszka, Cieślik, Domurat i Macko (2011) przeprowadzili badanie, w którym porównywali skłonność do ryzyka wśród osób pracujących na etacie (grupa 1, $n_1 = 120$), przedsiębiorców z wyboru (grupa druga, $n_2 = 64$) i przedsiębiorców z konieczności (grupa trzecia, $n_3 = 54$). Badani mogli wybrać, czy za udział w badaniu chcą dostać kwotę pewną, czy kwotę zależną od liczby poprawnych odpowiedzi na kilka pytań o różne sprawy społeczne i ekonomiczne. Otrzymano następujące proporcje wyborów udziału w kwizie: 27%, 39% i 36% i statystykę Chi-kwadrat $\chi^2(2) = 3{,}392$, $p = 0{,}183$. Wynik ten zinterpretowano jako brak różnic między przedsiębiorcami a nie-przedsiębiorcami pod względem skłonności do ryzyka. Jak pisaliśmy wcześniej, twierdzenia kategoryczne na temat hipotezy zerowej w oparciu o NHST są nadużyciem. Odtworzywszy bazę danych na podstawie powyższych statystyk i wykonawszy odpowiednie analizy w programie JASP (opcje: Common – Frequencies – Bayesian Contingency Tables) uzyskaliśmy czynnik bayesowski na poziomie $BF_{01} = 5{,}608$. Oznacza on, że hipoteza o braku różnic znalazła znaczące, prawie sześciokrotnie silniejsze wsparcie w danych, niż hipoteza o różnicach między odsetkami w tych trzech grupach.

W świetle uzyskanego czynnika bayesowskiego, autorzy wyciągnęli właściwy wniosek, choć dopuścili się nadinterpretacji wyniku testu istotności. Czynnik bayesowski pokazał, że najprawdopodobniej porównywane grupy nie różnią się skłonnością do ryzyka.

3.2. Bayesowskie testy korelacji

Barr, Pennycook, Stolz i Fugelsang (2015) postanowili sprawdzić, czy smartfony zastępują ludziom myślenie. Twierdzili oni, że im niższa jest skłonność ludzi do myślenia refleksyjnego (mierzona za pomocą skali myślenia refleksyjnego CRT, ang. *Cognitive Reflection Test*; zob. Frederick 2005), tym więcej czasu spędzają oni na wyszukiwaniu informacji przez telefon w porównaniu do osób bardziej refleksyjnych. Oczekiwano natomiast, że czas poświęcany na poszukiwanie informacji, mierzony w minutach spędzonych dziennie na googlowaniu, nie zależy od zdolności poznawczych, mierzonych jako sumaryczny wynik poprawnie rozwiązanych zadań: serii ośmiu sylogizmów (zob. De Neys i Franssens, 2009), czterech zadań na temat proporcji podstawowej (ang. *base rate*; zob. De Neys i Glumicic, 2008) i czternastu zadań heurystycznych (zaczerpniętych z pracy: Toplak, West i Stanovich, 2011). Korelacje[10] między refleksyj-

[10] Chcąc tutaj pokazać wykorzystanie wskaźnika *BF* w analizie korelacji na przykładzie z ciekawą treścią, posłużyliśmy się zbiorem danych surowych udostępnionym przez Pennycooka. W cytowanym artykule zamieszczono inne analizy niż wykonane przez nas. Podzielono tam badanych arbitralnie na trzy grupy (rzadko, umiarkowanie często i często szukających informacji) i sprawdzano istotność różnic pod względem refleksyjności i zdolności poznawczych. Różnice te okazały się istotne statystycznie tylko dla CRT, tak jak policzona przez nas korelacja. Istotność korelacji sugeruje więc identyczne wnioski, jak zaproponowane przez autorów.

 DOI: 10.7206/DEC.1733-0092.79

nością poznawczą i zdolnościami poznawczymi a czasem poświęcanym dziennie na poszukiwanie informacji przedstawiono w tabeli 2. Tabela zawiera wskaźniki istotności statystycznej i czynniki bayesowskie obliczone za pomocą programu JASP (opcje: Regresja – Korelacje z testami klasycznymi, Korelacje z wskaźnikiem *BF*).

Tabela 2
Korelacje między refleksyjnością poznawczą i zdolnościami poznawczymi a czasem spędzanym na poszukiwaniu informacji

	***r* Pearsona**	***p***	BF_{01}
Refleksyjność poznawcza	-0,138	0,038	1,410
Zdolności poznawcze	-0,119	0,072	2,453

Istotność korelacji podpowiada wnioski takie same, jak te wyciągnięte przez autorów: że czas poszukiwania informacji negatywnie zależy od refleksyjności poznawczej i że nie zależy od zdolności poznawczych. Tymczasem czynniki bayesowskie sugerują, że badanie jest niekonkluzywne i nie można mówić ani o istnieniu tych zależności, ani o ich braku. By to rozstrzygnąć, należałoby zebrać większą ilość danych. Uzyskane obecnie obie wartości *BF* są niskie, ale powyżej 1. Może się więc okazać, że ilość czasu poszukiwania informacji za pośrednictwem smartfona zależy i od refleksyjności, i od zdolności poznawczych.

3.3. Bayesowskie testy różnic

Testowanie hipotez dotyczących różnic międzygrupowych prześledzimy na przykładzie badania Tyszki i współpracowników (2016), w którym sprawdzano, od czego zależy skłonność ludzi do wiary w prawo serii na rynku giełdowym, czyli stosowania tzw. strategii prognostycznej momentum. Na przykład inwestorzy giełdowi stosujący tę strategię są przekonani, że rosnące ceny akcji będą nadal rosnąć, a spadające – spadać. Zadaniem uczestników dwóch z czterech eksperymentów było przewidywanie, jakie będzie kolejne zdarzenie po zaobserwowaniu serii zdarzeń wcześniejszych. Badani dokonywali przewidywań od momentu, gdy pokazano im serię przynajmniej trzech zdarzeń: strzałek w górę (zdarzenia losowe, eksperyment nr 2) lub skutecznych trafień koszykarza (zdarzenia nielosowe, eksperyment nr 4). Niezależnie od tego, czy badany przewidywał kontynuację serii (czyli stosował strategię momentum), czy przewidywał odwrócenie trendu (czyli stosował strategię kontrariańską), serię obserwowanych zdarzeń wydłużano aż do dziewięciu identycznych zdarzeń pod rząd. Uzyskiwano więc sześć przewidywań od każdego badanego, poprzedzonych serią identycznych zdarzeń. Liczba zdarzeń przewidywanych zgodnie z serią była wskaźnikiem wiary w kontynuowanie serii (czyli stosowania strategii momentum).

 DOI: 10.7206/DEC.1733-0092.79

3.3.1. Test t-Studenta dla prób niezależnych

Badacze oczekiwali, że ograniczenie możliwości przetwarzania informacji (konieczność słuchania historii o faunie podczas rozwiązywania zadania) będzie miało znaczenie dla przewidywania zdarzeń zależnych od losu (kierunku strzałek) i nie będzie miało znaczenia dla przewidywania zdarzeń zależnych od umiejętności (trafnych rzutów koszykarza). Od strony statystycznej oczekiwano więc istotności bądź nieistotności statystyk, odpowiadającej dwóm hipotezom w teście statystycznym t-Studenta dla prób niezależnych: hipotezie alternatywnej $\mu_1 \neq \mu_2$ dla zdarzeń zależnych od losu oraz hipotezie zerowej $\mu_1 = \mu_2$ dla zdarzeń zależnych od umiejętności. Statystyki opisowe, statystyki *t*, ich istotność *p* i policzone czynniki bayesowskie *BF* przedstawia tabela 3:

Tabela 3
Wyniki eksperymentów 2 oraz 4 z badania Tyszki i in. (2016) i ich weryfikacja za pomocą testów istotności i czynnika bayesowskiego

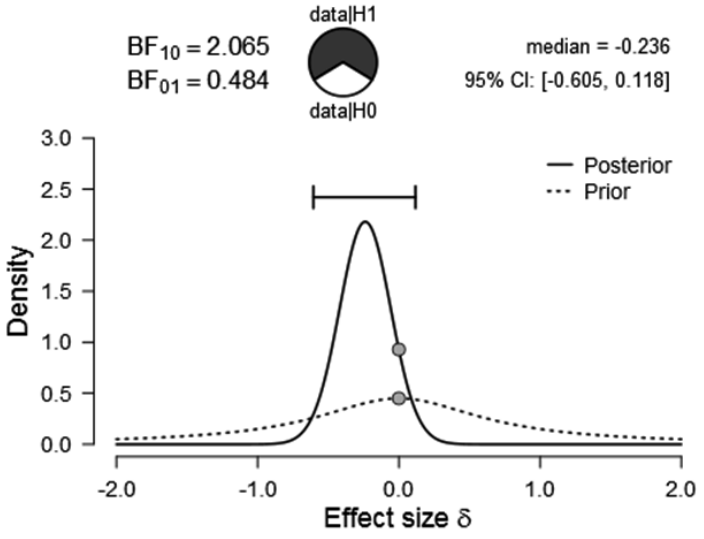
	Grupa kontrolna	Grupa z obciążeniem poznawczym	t	p	BF_{10}
Eksperyment 2: Zdarzenia losowe	3,05 (1,62), n=40	3,96 (1,40), n = 28	-2,422	,018	2,89
Eksperyment 4: Zdarzenia nielosowe	3,44 (1,58), n = 50	3,86 (1,54), n = 57	-1,389	,413	2,07

Zmienną zależną jest częstość stosowania strategii momentum.

Tak jak w analizie wyników badania Baumeistera i in. (1998), tak i tutaj czynniki bayesowskie oszacowaliśmy nie na podstawie danych surowych, lecz na podstawie opublikowanych statystyk zbiorczych (wykorzystaliśmy moduł „Statystyki zbiorcze" programu JASP). Rysunek 1, uzyskany w programie JASP, przedstawia oszacowaną siłę efektu („Effect size *δ*") dla hipotezy 2 o braku różnic, wartości wskaźników bayesowskiech BF_{10} i BF_{01}, medianę oraz 95% przedział wiarygodności (ang. *credible interval*) dla oszacowanego efektu (95% CI). BF_{10} mówi o sile dowodu na rzecz H_1 wobec H_0, zaś BF_{01} odwrotnie – o sile dowodu na rzecz H_0 względem H_1 (BF_{01} jest odwrotnością BF_{10}). Raportując wyniki analizy wygodnie jest użyć tego wskaźnika o wartości powyżej jedności. Dziewięćdziesięciopięcioprocentowy przedział wiarygodności CI dla szacowanej siły efektu należy rozumieć w ten sposób, że szansa na to, że szacowany parametr znajduje się w tym przedziale, wynosi właśnie 95%. Przedziału wiarygodności nie należy mylić z przedziałami ufności liczonymi przy testach istotności według klasycznej estymacji przedziałowej. Przedział wiarygodności zamieszczony jest dodatkowo na wykresie prezentującym rozkład siły efektu („Effect size") założony *a priori* („Prior", linia przerywana) oraz rozkład *a posteriori* („Posterior", linia ciągła). W testach wykonywanych przez program JASP jako rozkład *a priori* przyjmuje się dla hipotezy alternatywnej nieinformatywny rozkładu Cauchy'ego (linia przerywana). Rozkład ten nie ma wartości oczekiwanej, jest symetryczny,

 DOI: 10.7206/DEC.1733-0092.79

z medianą i modalną równą zero. Rozkład siły efektu *a posteriori* aktualizuje rozkład *a priori*, uwzględniając dane podlegające analizie. Rozkład *a posteriori* uzyskany w jednym badaniu może nadać kierunek kolejnym badaniom jako rozkład *a priori*.



Rysunek 1. **Bayesowska ocena hipotezy o braku różnic w eksperymencie 4 z artykułu Tyszka i in. (2016). Wykres otrzymany w programie JASP .0.8.**

Uzyskany w tym badaniu czynnik BF_{10} =2,065 sugeruje, że dwukrotnie bardziej prawdopodobne jest to, że grupy wykonujące zadanie pod obciążeniem poznawczym różnią się od grupy kontrolnej częstością stosowania strategii momentum. Innymi słowy szansa tego, że grupy się różnią, wynosi około 2/3, a tego, że się nie różnią – około 1/3. Drugi z czynników bayesowskich wskazuje prawie trzykrotnie większe wsparcie hipotezy o różnicach między grupami w przewidywaniach zdarzeń losowych.

Testy istotności wykonane przez autorów skłoniły ich do wniosku, że obciążenie poznawcze działa tylko dla przewidywania zdarzeń losowych, a nie działa dla zdarzeń nielosowych. Weryfikacja tych samych hipotez za pomocą czynnika bayesowskiego sugeruje jednak coś innego. Jeśli jako minimum rozstrzygnięcia przyjąć kryterium $BF > 3$ (zob. tab. 1), to wskaźniki *BF* przypisane obydwu statystykom t-Studenta pokazują, że wyniki obydwu eksperymentów są niekonkluzywne (BF_{10} między 2 a 3).

 DOI: 10.7206/DEC.1733-0092.79

3.3.2. (Dwuczynnikowa) analiza wariancji

Alternatywnie, do weryfikacji opisywanych tu hipotez można wykorzystać dwuczynnikową analizę wariancji. W tym celu należy połączyć dane z obydwóch eksperymentów i potraktować jako czynniki międzygrupowe dwie zmienne: obciążenie poznawcze (tak, nie) oraz rodzaj zdarzenia (losowe, nielosowe). Kierując się przypuszczeniem autorów badania, można oczekiwać, że jeśli obciążenie poznawcze działa tylko przy przewidywaniu zdarzeń losowych, to powinniśmy zaobserwować istotną interakcję między tymi dwoma czynnikami. Istotność czynnika opisującego rodzaj zdarzenia (losowe vs. nielosowe, eksperyment 2 vs. eksperyment 4) świadczyłaby natomiast o tym, że ludzie bardziej wierzą w kontynuację serii wtedy, gdy zdarzenia są nielosowe.

W tabeli 4 zamieściliśmy wyniki testów z dwuczynnikowej analizie wariancji, którą udało nam się powtórzyć za autorami badania na otrzymanej od nich bazie wyników surowych. W tabeli 5 przedstawiliśmy wyniki bayesowskiej analizy wariancji, obliczonej za pomocą programu JASP.

Tabela 4
Testy istotności w dwuczynnikowej analizie wariancji

Czynnik	Suma Kwadratów	df	F	p	η^2
obciążenie	18,109	1	7,552	0,007	0,042
typ zdarzenia	0,829	1	0,346	0,557	0,002
interakcja	2,490	1	1,038	0,310	0,006

Obciążenie poznawcze skłaniało badanych do częstszego stosowania strategii momentum, $F(1,171) = 7,552$, $p = 0,007$. Interakcja w opisywanym modelu dwuczynnikowym okazała się natomiast nieistotna statystycznie, $F(1,171) = 1,038$, $p = 0,310$. Podobnie nieistotny był wpływ rodzaju przewidywanych zdarzeń, $F(1,171) = 0,346$, $p = 0,557$. Analiza nie pozwala więc na stwierdzenie, że pod wpływem obciążenia poznawczego skłonność do stosowania strategii momentum zmienia się w innym stopniu wtedy, gdy przewidywane jest pojawanie się strzałek, a w innym – gdy przewidywane są rezultaty rzutu koszykarza. Ostatnie dwa testy istotności są tu niekonkluzywne: nieistotność interakcji nie uprawnia wniosku, że wpływ obciążenia poznawczego jest taki sam w zadaniach polegających na przewidywaniu zdarzeń sprawnościowych, jak w zadaniach polegających na przewidywaniu zdarzeń losowych.

W analizie bayesowskiej rozważa się jednocześnie kilka potencjalnych modeli, które można stworzyć w oparciu o czynniki i ich interakcję, i sprawdza się, który z tych modeli najlepiej przewiduje uzyskane dane. W przypadku analizy dwuczynnikowej porównuje się pięć następujących modeli, przewidujących: M_0 – brak jakichkolwiek efektów, M_1 – wpływ tylko pierwszego czynnika, obciążenia poznawczego,

DOI: 10.7206/DEC.1733-0092.79

M_2 – wpływ tylko drugiego czynnika, typu zdarzenia, M_3 – wpływ obydwu tych czynników, bez interakcji oraz M_4 – wpływ obydwu efektów głównych oraz ich interakcji. „Na wejściu" nie ma żadnych przesłanek, by uważać któryś z tych modeli za bardziej wiarygodny od innych. Dlatego ustala się dla nich jednakowe prawdopodobieństwo *a priori*. W tym przykładzie jest to P(M) = 0,2, ponieważ rozważa się pięć potencjalnych modeli. Ponieważ porównywane są więcej niż dwa modele, oprócz wskaźników BF_{10}, pokazujących, na ile dany model jest bardziej wiarygodny od modelu zerowego, wyznacza się prawdopodobieństwo *a posteriori* P(M|dane) dla każdego z modeli – zob. tabela 5. Liczy się także wskaźnik BF_M, który ocenia wiarygodność danego modelu w porównaniu do zbioru wszystkich pozostałych modeli. Ideą tego wskaźnika jest spostrzeżenie, że choć sam model może przewidywać dane lepiej niż model zerowy, to powinien on jeszcze mieć wyraźną przewagę nad pozostałymi modelami. Ostatnia kolumna pokazuje, z jaką dokładnością generowane są statystyki *BF* w symulacji MCMC wykorzystywanej w programie JASP.

Tabela 5
Analiza zbiorcza dla eksperymentów 2 oraz 4 w badaniu Tyszki i innych (2016) przeprowadzona przy zastosowaniu metod bayesowskich

Modele	P(M)	P(M|dane)	BF_M	BF_{10}	% błąd
M_1: model zerowy	0,200	0,144	0,673	1,000	
M_2: typ zdarzenia	0,200	0,037	0,155	0,259	0,000009
M_3: obciążenie poznawcze	0,200	0,639	7,080	4,436	0,0000008
M_4: obciążenie, typ zdarzenia	0,200	0,135	0,622	0,934	0,97
M_5: obciążenie, typ zdarzenia, interakcja	0,200	0,045	0,189	0,313	2,855

Porównując czynniki bayesowskie zawarte w tabeli 5, stwierdzamy, że BF_{10} ma wysoką wartość dla modelu jednoczynnikowego M_3 z czynnikiem „obciążenie poznawcze", BF_{10} = 4,436. Model ten jest ponad siedmiokrotnie bardziej wiarygodny, niż pozostałe modele razem wzięte, BF_M = 7,08. Jego prawdopodobieństwo a posteriori jest najwyższe, P(M_3|dane) = 0,639, pokazując jego ostateczną przewagę nad pozostałymi modelami. Z kolei model z interakcją jest ponad pięciokrotnie mniej prawdopodobny niż pozostałe modele, BF_M = 0,189 i ponad trzykrotnie mniej wiarygodny, niż model zerowy, $BF_{01} = 0{,}313^{-1} = 3{,}195$. Wbrew oczekiwaniom autorów badania, bayesowska analiza wariancji wskazuje więc, że nie ma interakcji między czynnikami. Gdy ograniczy się uczestnikom badania możliwość przetwarzania informacji, obciąży ich poznawczo, to są oni skłonni bardziej wierzyć w kontynuację trendu zarówno wtedy, gdy mają ocenić serię zdarzeń nielosowych (wyniki rzutów koszykarza), jak i wtedy – gdy losowych (pojawiające się strzałki).

 DOI: 10.7206/DEC.1733-0092.79

Podsumowując, czynniki bayesowskie sugerują inne wnioski, niż wyciągnięte przez autorów na podstawie istotności statystyk t-Studenta. Jednym z wniosków z abstraktu omawianego artykułu było stwierdzenie, że odwrócenie trendu jest spodziewane wtedy, gdy seria zdarzeń rozpoznawana jest jako seria zdarzeń losowych (niezależnych np. od wpływu czyichś umiejętności). Wskaźniki *BF*, opisujące zarówno wiarygodność hipotez stawianych w testach t-Studenta dla prób niezależnych, jak i wiarygodność hipotez formułowanych w obrębie modeli analizy wariancji, wskazały, że wniosek ten jest nieprawomocny, a uzyskane dane są co najwyżej niekonkluzywne i wymagają dalszych studiów. Najprawdopodobniej jednak obciążenie poznawcze skłania ludzi do stosowania strategii momentum niezależnie od typu przewidywanych zdarzeń.

4. Użycie czynnika bayesowskiego: korzyści i pułapki

Czynnik bayesowski ma szereg zalet w porównaniu do testów istotności opartych na analizie wartości *p* (Wagenmakers, 2007; Dienes, 2014). Po pierwsze, odpowiada na właściwe pytanie: wprost pokazuje badaczowi, którą z hipotez lepiej wspierają uzyskane dane, oraz która z hipotez zyskała (lub straciła) na wiarygodności w świetle uzyskanych danych. Po drugie, w przeciwieństwie do NHST, metoda czynnika bayesowskiego odnosi się do danych faktycznie otrzymanych (przypomnijmy, wartość *p* opisuje hipotetyczne wyniki, wynikające z założenia prawdziwości hipotezy zerowej). Po trzecie, w przeciwieństwie do wartości *p* metoda czynnika bayesowskiego, tak jak i inne metody bayesowskie, ma ugruntowane, spójne podstawy matematyczne, wynikające z aksjomatycznej teorii prawdopodobieństwa (Hays, 1973). W metodzie tej prawdopodobieństwo jest miarą siły dowodu, czego nie da się powiedzieć o koncepcji Neymana i Pearsona oraz testach NHST. W tych drugich „rządzi" kryterium minimalizacji błędów decyzyjnych, a wartość *p* jest ledwie miarą nietypowości danych, przez co jego znaczenie zależy od kontekstu (np. wielkości próby, poziomu istotności przyjętego przez badacza itd.). Kolejną korzystną własnością czynnika bayesowskiego, której nie ma wartość *p*, jest spełnienie zasady kumulatywności wiedzy i dowodów (Dienes, 2016). Kumulatywność wiedzy opiera się na założeniu, że nawet mało informatywne badanie, w którym $BF < 3$, może być źródłem informacji o prawdopodobieństwie *a priori* dla kolejnych badaczy zbierających kolejne dane. Ma to wielkie znaczenie, gdy bada się trudno dostępne próby, rzadko występujące zdarzenia, lub gdy gromadzenie danych jest niezwykle kosztowne.

Zasady kumulatywności nie spełnia procedura NHST, gdzie wręcz zaleca się nie badać zbyt dużych prób, gdyż wtedy nawet niewielkie efekty, o niskim znaczeniu

praktycznym, zyskują istotność statystyczną. Powiększanie próby do momentu, aż statystyka przekroczy próg istotności jest więc niedopuszczalną manipulacją. Jest bowiem tak, że jeśli hipoteza zerowa jest prawdziwa, to wzrost liczebności próby powoduje, że prędzej czy później wartość *p* spada poniżej dowolnego progu (zob. np. symulacje wykonane przez Jarmakowską-Kostrzanowską, 2016). Jeśli więc badacz zwiększa próbę, aż uzyska np. istotną korelację, słusznie jest posądzany o tzw. *p*-hacking (zob. np. Head i in., 2015). Dlatego niektóre z czasopism (np. „Psychological Science" czy „Science Translational Medicine") wymagają pokazania metody określenia wielkości próby przed przeprowadzeniem badania (por. Cumming, 2014).

Użycie czynnika bayesowskiego wiąże się również z pewnymi pułapkami. Głównym zarzutem wobec metod bayesowskich jest ich subiektywizm (zob. dyskusję w: Carlin i Louis, 1997 oraz Dienes, 2011). Rzeczywiście na pierwszy rzut oka można mieć wrażenie, że metody bayesowskie sankcjonują dowolne hipotezy i teorie, skoro różni ludzie mogą mieć rozmaite przesłanki „na wejściu" i przyjmować dowolne rozkłady *a priori*. Zauważmy jednak, że, po pierwsze, użycie czynnika bayesowskiego jest próbą ominięcia tej trudności – pokazuje on bowiem, w którym kierunku przekonania powinny się zmienić, a nie, którą hipotezę „wybrać", albo która z hipotez jest bardziej prawdopodobna *a posteriori*. Po drugie, to właśnie metody bayesowskie kładą silny nacisk na stopniową aktualizację wiedzy pod wpływem napływających informacji. Innymi słowy, zarówno sceptycy, jak i entuzjaści danych hipotez powinni wraz z rosnącą ilością danych mieć coraz bardziej zbliżone rozkłady *a posteriori*. Innymi słowy, mimo potencjalnie różnych subiektywnych założeń początkowych, wraz z napływającymi informacjami stanowiska badaczy powinny zbliżać się do siebie (por. przypis 8). Jak stwierdzili Edwards i in. (1963), przeciwnicy metod bayesowskich, odrzucający je i przyrównujący do budowania zamków na piasku, proponując pozostanie przy testach istotności, zalecają tak naprawdę budowanie w próżni.

Trudnością praktyczną jest konieczność założenia, że każdy, choćby mało prawdopodobny model, ma jakąś szansę przewidywania danych. W przykładzie z monetami uwzględniono nie tylko monety symetryczne i z dwoma orłami, lecz także monety przeciążone. Gdyby je wykluczyć *a priori* (przypisując im *a priori* prawdopodobieństwo równe zero), zaobserwowane 10 orłów sugerowałoby, że moneta najprawdopodobniej ma dwa orły, z nikłą szansą, że jest jednak symetryczna. Analizując ten przykład dalej, można mieć (być może całkiem słuszne) wątpliwości, czy zasadne było pominięcie monet przeciążonych w stronę reszki. Właśnie dlatego zaleca się, by szczególnie przy pierwszych badaniach unikać zbyt szczegółowej specyfikacji modeli i przyjmować nieinformatywne rozkłady *a priori* (np. rozkład jednostajny lub rozkład Cauchy'ego).

 DOI: 10.7206/DEC.1733-0092.79

Przyjmując perspektywę bayesowską, szukamy odpowiedzi na właściwe pytanie: która z hipotez jest bardziej wiarygodna w świetle nowych informacji. Nie dziwne więc, że czynnik bayesowski zdobywa stopniowo popularność wśród badaczy. Można mieć nawet obawy, czy jego rosnąca atrakcyjność nie sprawi kiedyś, że mechaniczny rytuał „$p < 0{,}05$" zostanie zastąpiony kolejnym – „$B > 3$". Mimo to zachęcamy czytelników do ponownego przyjrzenia się swoim starym danym, uznanym za niepublikowalne z uwagi na nieistotność statystyczną. Podejście bayesowskie zmienia stosunek badacza do własnej pracy między innymi dlatego, że nie narzuca ograniczeń na wielkość próby. Im więcej dowodów gromadzi badacz, tym zawsze jest lepiej, gdyż tym większe wsparcie dla którejś z weryfikowanych hipotez. Jak to ładnie wyraził Dienes (2011, s. 285), w podejściu bayesowskim nie ma miejsca na poczucie winy z powodu ilością gromadzonych danych, gdyż owoc z drzewa poznania zawsze smakuje dobrze.

Załącznik 1. Różnice między teorią A.F. Fishera a teorią J. Neymana i E. Pearsona

Stosowana w praktyce procedura NHST stanowi niespójną hybrydę dwóch podejść: teorii R.A. Fishera oraz teorii J. Neymana i E. Pearsona (w skrócie: N – P). W podejściu Fishera formułuje się wyłącznie hipotezę zerową (wbrew powszechnemu odczuciu, może ona opisywać różne wartości – niezerową różnicę, korelację liniową itd., np. H_0: $\rho=0{,}3$; H_0: $\mu_1 - \mu_2 = 13$ itp.). Wartość p jest miarą istotności statystyki testu, czyli jej nietypowości z perspektywy hipotezy zerowej: im bardziej nietypowe dane i statystyki empiryczne uzyskano, tym mniej prawdopodobne są one przy założeniu prawdziwości hipotezy zerowej. Fisherowska istotność jest stopniowalna, tzn. tym większa, im wartości p niższe. Kwestią wygody (związanej np. z korzystaniem z tablic) jest przyjęcie poziomu istotności, np. $\alpha = 5\%$, wyznaczającego obszar krytyczny statystyk skłaniających badacza do odrzucenia hipotezy zerowej. Jeśli badacz chce być bardziej restrykcyjny, może użyć np. $\alpha = 1\%$ (zob. np. rozważania Fishera: Fisher, 1925/1950, s. 78–83). Istotność statystyki pozwala na jedną z dwóch interpretacji: albo uzyskano nietypową próbę, albo hipotezę zerową należy odrzucić. Nie dopuszcza się tu wyjaśnień alternatywnych, gdyż testuje się wyłącznie model opisany hipotezą zerową.

Hipoteza alternatywna jest natomiast jednym z głównych elementów koncepcji N – P. Według tego podejścia celem testu statystycznego jest optymalizacja decyzji badacza. Minimalizuje się tu ryzyko popełnienia dwóch błędów: odrzucenia prawdziwej hipotezy zerowej (błąd I rodzaju, α) i przyjęcia hipotezy zerowej fałszywej

 DOI: 10.7206/DEC.1733-0092.79

(błąd II rodzaju, β). Wielkość $1 - \beta$ nazywana jest mocą testu i pokazuje, jaki odsetek badań takich jak prowadzone jest w stanie wykryć, że jest inaczej, niż postuluje hipoteza zerowa. Hipoteza alternatywna jest potrzebna, gdyż na jej podstawie określa się wielkość znaczącego efektu eksperymentalnego i moc testu. W podejściu N – P badacz określa wysokość błędu α przed analizą danych i na jego podstawie wyznacza obszar krytyczny w rozkładzie teoretycznym statystyki testu. Następnie wyznacza statystykę empiryczną. Chociaż wartość p określa położenie tej statystyki w obszarze krytycznym ($p < \alpha$) lub w obszarze ufności ($p \geq \alpha$), jej dokładna wielkość jest bez znaczenia. Test jest przecież optymalizowany względem błędu I rodzaju α, a nie względem wartości p: po ewentualnym odrzuceniu hipotezy zerowej podejmuje się działania odpowiadające hipotezie alternatywnej (zob. Neyman i Pearson, 1933; Aranowska i Rytel, 1997; Gigerenzer, 2004).

Jak zauważa Gigerenzer (2004), nie sposób przecenić przydatności podejścia N – P w podejmowaniu decyzji powtarzalnych lub dotyczących zjawisk masowych. Na przykład pakując zapałki do pudełek zawierających według normy 48 ± 2 sztuki, ważne jest, by wykryć problem rozregulowania się maszyny i nie uznawać nadal, że norma jest zachowana (błąd II rodzaju) lub – by niepotrzebnie tej maszyny nie zatrzymywać (błąd I rodzaju). Przyjęcie podejścia N – P ma jednak zaskakujące konsekwencje. W tym przykładzie nie chodzi o to, by udowadniać, że maszyna faktycznie się rozregulowała, lub że pracuje zgodnie z normą. Chodzi natomiast o to, by nie przeprowadzać niepotrzebnej korekty, ale i nie przegapić sytuacji, kiedy taka korekta jest niezbędna. Ogólniej nie chodzi więc o to, by dowodzić prawdziwości hipotezy zerowej lub alternatywnej, lecz o to, by podejmowane działania były trafne. Dlatego zdecydowanie odrzucał to podejście Fisher. Sprzeciwiał się on behawioralnym założeniom, ukierunkowanym na działania, decyzje i preferencje. Uważał, że metody wnioskowania statystycznego powinny tak samo przekonywać każdego racjonalnie myślącego człowieka, niezależnie od zastosowań tej wiedzy. Statystyka jego zdaniem jest narzędziem wnioskowania indukcyjnego, dostarczającym racjonalnego wsparcia teoriom i hipotezom w postaci wyników badań empirycznych. Daje szanse wprowadzenia nowych treści do teorii i jest potrzebna wtedy, gdy nie wystarcza wnioskowanie dedukcyjne. Kalkulowanie błędów decyzyjnych kłóci się z tym celem, ponieważ uzależnia interpretację testu od zamiarów i intencji badacza, a nie od testowanej teorii (Fisher, 1955). Więcej szczegółów o przejawach narastającego antagonizmu między Fisherem i Neymanem – nie tylko na gruncie naukowym – czytelnik znajdzie w pracach: Lehmann (2011), Gigerenzer (2004) i Jarmakowska-Kostrzanowska (2016). W pracach tych oraz w pracy Gigerenzera i in. (2004) opisane są również liczne kontrowersje związane z wykorzystaniem testów istotności do weryfikacji hipotez statystycznych.

 DOI: 10.7206/DEC.1733-0092.79

Bibliografia

Aczel, B., Palfi, B., Szaszi, B., Szollosi, A., & Dienes, Z. (2015). Commentary: Unlearning implicit social biases during sleep. *Frontiers in Psychology*, *6*, 1428.

Aranowska, E., & Rytel, J. (1997). Istotność statystyczna – co to naprawdę znaczy? *Przegląd Psychologiczny*, *40*, 249-260.

Barr, N., Pennycook, G., Stolz, J.A., & Fugelsang, J.A. (2015). The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Computers in Human Behavior, 48*, 473-480.

Baumeister, R.E., Bratslavsky, E., Muraven, M., & Tice, D.M. (1998). Ego Depletion: Is the Active Self a Limited Resource? *Journal of Personality and Social Psychology 74*, 1252-1265.

Bayes, M., & Price, M. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, FRS Communicated by Mr. Price, in a Letter to John Canton, AMFRS. *Philosophical Transactions of the Royal Society of London, 53*, 370-418.

Białek, M., (2015) Przegląd badań współczesnej kognitywistyki nad efektem przekonań. *Przegląd Filozoficzny*. *Nowa seria*, 95, 91-107.

Carlin, B.P., & Louis, T.A. (1997). Bayes and empirical Bayes methods for data analysis. *Statistics and Computing, 7*, 153-154.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29.

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition, 113*, 45-61.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*, 1284-1299.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274-290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*,78–89.

Domański, H., & Pruska, K. (2000). *Nieklasyczne metody statystyczne*. Warszawa: PWE.

Domurat, A., Kowalczuk, O., Idzikowska, K., Borzymowska, Z., & Nowak-Przygodzka, M. (2015). Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations. *Frontiers in Psychology*, *6*, 1194.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

Fisher, R.A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society*. *Series B (Methodological), 17*, 69-78.

Fisher, R.A. (1925/1950). *Statistical methods for research workers*. *Biological monographs and manuals*. *No. V. (11th ed.)*. Londyn: Oliver and Boyd.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25-42.

Gaifman, H., & Snir, M. (1982). Probabilities over rich languages, testing and randomness. *Journal of Symbolic Logic*, *47*, 495-548.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Gigerenzer G., Krauss S., Vitouch O. (2004). The null ritual. What you always wanted to know about significance testing but were afraid to ask. W: Kaplan D. (red.), *The Sage Handbook of Quantitative Methodology for the Social Sciences* (s. 391–408). Thousand Oaks, CA: Sage

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research, 7*, 1-20.

Hays, W.L. (1973). *Statistics for the Social Sciences. 2nd ed*. Nowy Jork: Holt Rinehart & Winston.

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., & Jennions, M.D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*, e1002106.

Hu, X., Antony, J.W., Creery, J.D., Vargas, I.M., Bodenhausen, G.V., & Paller, K.A. (2015). Unlearning implicit social biases during sleep. *Science, 348*, 1013-1015.

Jarmakowska-Kostrzanowska (2016). W statystycznym matriksie: kontrowersje wokół testowania istotności hipotezy zerowej oraz p-wartości. *Psychologia Społeczna*.

Jeffreys, H. (1939/1961). *Theory of Probability*. Oxford: Oxford University Press.

Jóźwiak, J., & Podgórski, J. (2001) *Statystyka od podstaw. Wyd. V zm.* Warszawa: PWE.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kass, R.E., & Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*, 773-795.

Koronacki, J., & Mielniczuk, J. (2001). *Statystyka dla studentów kierunków technicznych i przyrodniczych*. Warszawa: Wyd. Naukowo-Techniczne.

Krämer, W., & Gigerenzer, G. (2005). How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities. *Statistical Science, 20*, 223-230.

Lehmann, E.L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*. Nowy Jork: Springer Science & Business Media.

Morey, R.D., Romeijn, J.W., & Rouder, J.N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6-18.

Neyman, J. (1957). "Inductive Behavior" as a basic concept of philosophy of science. *Review of the International Statistical Institute, 25*, 7-22.

Neyman, J., & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, 231*, 289-337.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chichester: Wiley.

Pawłowski Z. (1976). *Statystyka matematyczna*. Warszawa: PWN.

Ramsey, F.P. (1931). Truth and probability. W: Trench P.K. (red.), *The foundations of mathematics and other logical essays*. Londyn: Truber.

Stanovich, K. E. (2009). Rational and irrational thought: The thinking that IQ tests miss. *Scientific American Mind*, 20, 34-39.

Toplak, M.V., West, R.F., & Stanovich, K.E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*, 1275–1289.

 DOI: 10.7206/DEC.1733-0092.79

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Tyszka, T. (1999). *Psychologiczne pułapki oceniania i podejmowania decyzji*. Gdańsk: GWP.

Tyszka, T. (2001). Kłopoty z myśleniem probabilistycznym. *Roczniki Psychologiczne, 4*, 179-191.

Tyszka, T. (2010). *Decyzje. Perspektywa psychologiczna i ekonomiczna*. Warszawa: Wydawnictwo Naukowe SCHOLAR.

Tyszka, T., Cieślik, J., Domurat, A., & Macko, A. (2011). Motivation, self-efficacy, and risk attitudes among entrepreneurs during transition to a market economy. *The Journal of Socio-Economics*, 40, 124-131.

Tyszka, T., Markiewicz, Ł., Kubińska, E., Gawryluk, K., & Zielonka, P. (2016). A belief in trend reversal requires access to cognitive resources. *Journal of Cognitive Psychology*.

Vallverdú, J. (2015). *Bayesians Versus Frequentists: A Philosophical Debate on Statistical Reasoning*. Nowy Jork: Springer.

Villejoubert, G., & Mandel, D.R. (2002). The inverse fallacy: An account of deviations from Bayes theorem and the additivity principle. *Memory & Cognition, 30*, 171-178.

Wagenmakers, E.-J., Morey, R.D., & Lee, M.D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25*, 169-176.

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779-804.

Westover, M.B., Westover, K.D., & Bianchi, M.T. (2011). Significance testing as perverse probabilistic reasoning. *BMC Medicine, 9*, 9-20.

Wetzels, R., Matzke, D., Lee, M.D., Rouder, J.N., Iverson, G.J., & Wagenmakers, E.J. (2011). Statistical evidence in experimental psychology an empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*, 291-298.

 DOI: 10.7206/DEC.1733-0092.79