Siniša Lakić[1]
Univerzitet u Banjoj Luci
Filozofski fakultet
Odsjek za psihologiju
Banja Luka, Republika Srpska, Bosna i Hercegovina

UDK 159.9.072:311
Pregledni rad
DOI: https://doi.org/10.46630/gpsi.18.2019.03

BAYESOV FAKTOR: OPIS I RAZLOZI ZA UPOTREBU U PSIHOLOŠKIM ISTRAŽIVANJIMA

Apstrakt

Cilj ovog rada je da našu širu psihološku publiku bolje upoznam sa Bayesovim faktorom (u notaciji BF ili B), odskora izuzetno popularnim statističkim metodom testiranja hipoteza u psihologiji koji čak pretenduje da zamijeni funkciju P-vrijednosti. Rastuća popularnost se najočitije vidi iz rezultata pretrage Google Scholar bibliografske baze kada se kao ključne riječi postave "Bayes factor" i "psychology". Za 2006. godinu se registruje tek 76 pogodaka, 2010. je to 176, 2014. već 436, dok je broj radova objavljenih samo u 2018. godini već 1570.[2] Uprkos bujajućem trendu, na našim jezicima nisam pronašao tekstove koji bi opisali BF i pojasnili njegove prednosti u odnosu na P-vrijednosti. Studenti psihologije, praktičari koji žele da prate naučne trendove, kao i iskusni istraživači, su tako osuđeni na relevantne radove na engleskom jeziku koji su često saturirani naprednijom statističkom terminologijom i statističkom notacijom što svakako otežava razumijevanje teksta i demotiviše čitaoce. Iz tog razloga, nastojao sam da ovaj tekst napišem jezikom razumljivim svima onima koji posjeduju fundamentalna statistička znanja. Rad počinjem opisom motiva za upotrebu BF, nakon čega prikazujem teorijsku podlogu BF kroz veoma jednostavne primjere, a rad završavam navodeći prednosti i ograničenja BF, uz sugerisanje softverskih rješenja pomoću kojih zainteresovani mogu sami da izračunavaju BF za niz različitih kvantitativnih nacrta i uklope novu paradigmu u svoj istraživački repertoar.

Ključne riječi: Bayesov faktor, P-vrijednosti, statističko testiranje hipoteza.

Uvod

Odavno je poznato da interpretiranje P-vrijednosti nije jednostavno, usljed čega se donose pogrešni zaključci (vidjeti npr. Cohen, 1995; Meehl, 1967; ili Gigerenzer 2004). O tome je pisano i na našim jezicima (Tenjović & Smederevac, 2011), a istraživači su podsticani da što više koriste standardizovane veličine efekta i intervalne procjene parametra, kao i da se oslone na replikacije i metaanalitički pristup u integrisanju nalaza različitih istraživanja (Cumming, 2014). Međutim, svjedoci smo da su P-vrijednosti i dalje sveprisutne (npr. svi empirijski radovi iz 2018. go-

[1] Adresa autora: sinisa.lakic@ff.unibl.org

[2] Da rast nije proizvod sve većeg ukupnog broja objavljenih radova pokazuje analiza pretrage za ključne riječi "psychology" i "regression", gdje se nakon rasta dešava pad u 2018. godini: 66400 (2006), 106000 (2010), 117000 (2014), 50600 (2018). U oba slučaja iz pretrage su izuzeti patenti i citiranja.

dine u ovom časopisu koriste P-vrijednosti, a situacija je veoma slična i u drugim časopisima na našem jezičkom prostoru). Osim što nose oreal objektivnosti, razlog zbog kojeg P-vrijednosti privlače istraživače je što na osnovu njih istraživači donose sudove o hipotezama od interesa. Za razliku od preporučivanih intervalnih procjena, metode testiranja hipoteza se neposrednije obraćaju živom korisniku koji treba da donese binarnu odluku na osnovu empirijskih podataka; na primjer, da li je uz psihoterapiju anksioznosti poželjno uzimati i farmakoterapiju, ima li inkluzija pozitivne efekte, koji od dva dizajnerska rješenja privlači veću pažnju potencijalnih kupaca, te predviđa li teorija A podatke bolje od teorije B.

Problemi korištenja P-vrijednosti i testiranja značajnosti nulte hipoteze

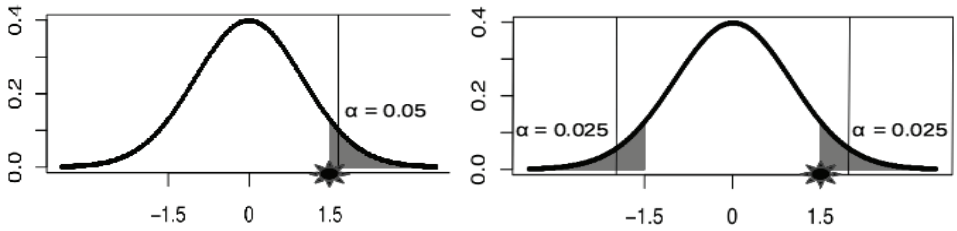
Ogromna većina psihologa poznaje P-vrijednosti isključivo u kontekstu testiranja nulte hipoteze (eng. Null Hypothesis Significance Testing) koju ovdje moram opisati kako bih bio siguran da čitalac pravilno razumije i Bayesov faktor (BF). Dakle, procedura testiranja započinje statističkom operacionalizacijom nulte hipoteze tako što unaprijed definišemo jedinstvenu, precizno određenu vrijednost populacionog parametra. Obično je taj parametar jednak 0 (eng. *nill hypothesis*), pa zavisno od statističkih mjera koje koristimo u nacrtu mislimo na vrijednost korelacija, razlika ili logaritmovanih omjera koji su jednaki 0. Vrijednost 0 služi za to da „predstavi idealizovanu poziciju skeptika koji vjeruje da se dobijeni podaci mogu objasniti isključivo dejstvom slučajnosti“ (Van Doorn, Matzke, & Wagenmakers, 2019, str. 4). Rjeđe specifikujemo neku drugu vrijednost, kao kada su u pitanju hipoteze prevalencije gdje, na primjer, želimo da testiramo hipotezu da je prevalencija anksioznosti među studentima različita od pretpostavljene vrijednosti (npr. 5%).

Nakon što smo precizirali vrijednost populacionog parametra od interesa, potrebno je da opišemo očekivanu distribuciju statistika (npr. *r*, *d*, *log(OR)*) na slučajno odabranim uzorcima iz populacije. Drugim riječima, zadatak distribucije statistika je da nam konkretizuje vjerovatnoću da dobijemo tačno određenu mjeru od interesa. Pri definisanju distribucije se najčešće[3] služimo dokazanim principima vjerovatnoće (npr. centralna granična teorema) i odabiramo neku od poznatih matematičkih distribucija (npr. *t, z*, *F*) koje nam čarobno prikazuju rezultate beskonačno dugog niza ponovljenih istraživanja na populaciji u kojoj je nulta hipoteza istinita. Kakav će tačno biti oblik distribucije ovisi ne samo o vrijednosti parametra od interesa za H_0, nego i o standardnoj grešci, koja predstavlja statistički očekivanu razliku između parametra i statistika usljed greške uzorkovanja. Standardna greška definiše širinu obuhvata distribucije i obično se procjenjuje na osnovu veličine uzorka (na većim uzorcima očekujemo manju grešku procjene) i količine varijabilnosti unutar uzorka (ukoliko

[3] Distribuciju statistika možemo konstruisati i egzaktnim proračunima (primjenom kombinatorne vjerovatnoće gdje je unaprijed poznata učestalost svakog ishoda), kao i simuliranim uzorkovanjima (npr. bootstrapping, Monte-Carlo simulacije) kada umjesto pretpostavljenog beskonačnog niza ponavljanja kalkulacija na uzorcima, koristimo veoma veliki broj ponavljanja (npr. 10000 ili više, ovisno o konkretnom postupku) koji se smatra razumnim da se predstavi ono što bi se desilo da je broj ponavljanja beskonačan.

su podaci međusobno sličniji, parametar od interesa bolje predstavlja grupu, pa je i mogućnost greške manja). I dolazimo napokon do P-vrijednosti.

Grafički objašnjeno, P-vrijednost nam govori kolika je površina od mjere koju smo dobili na uzorku (npr. datog t-skora) do kraja distribucije koju smo koristili da predstavimo H_0 (npr. distribucije t-skorova). U slučaju dvosmjernih testova koji se u psihologiji koriste kao zadane opcije, razmatramo apsolutnu vrijednost mjere koju smo dobili na uzorku, odnosno u obzir uzimamo ne jedan, nego oba kraja pretpostavljene distribucije (Slika 1 prikazuje takvu situaciju za klasični t-test). P-vrijednosti se izražavaju kao proporcije što znači da su kao matematički iskazi vjerovatnoće ograničene na skali od 0 do 1, a ukoliko ih pomnožimo sa 100 dobijamo procenat površine pod distribucionom krivom. Iz navedenog slijedi ispravna definicija koja kaže da je P-vrijednost vjerovatnoća dobijanja datog ili ekstremnijeg statistika ukoliko je nulta hipoteza u populaciji istinita.



Slika 1. Prikaz površine pod t-distribucijom i kritične vrijednosti testa značajnosti za slučaj kada je $t(100) = 1.50$ (označeno zvjezdicom). Osjenčene površine predstavljaju P-vrijednosti, a uspravne linije kritične vrijednosti za nivo značajnosti $\alpha = .05$ (za dvosmjerni test se sabiraju negativni i pozitivni efekti). Na lijevoj strani je prikazan jednosmjerni test gdje je $P = .068$, a kritična t-vrijednost 1.660, a na desnoj dvosmjerni test gdje je $P = .136$, a kritična t-vrijednost = 1.984.

Kada je dobijena P-vrijednost niža od nekog, u istraživanju definisanog nivoa statističke značajnosti (koji se notacijom označava kao α) – a obično, niža od .05 – dogovor zajednice ovlaštava istraživača da formalno odbaci nultu hipotezu i proglasi nalaze statistički značajnim, odnosno predstavi alternativnu hipotezu kao da je ona istinita. Ukoliko je ipak P-vrijednost jednaka ili veća od definisanog nivoa značajnosti, istraživač zadržava nultu hipotezu, ali ne može da zaključi ništa o njenoj istinitosti. Time se statističko promišljanje završava i proces se mehanički ponavlja iz istraživanja u istraživanje ili čak unutar istog istraživanja (vidi Gigerenzer, 2004 o „ritualu nule“ ili Bakan, 1966. o automatizovanom zaključivanju). Istraživaču je ostavljeno na volju da u sekciji diskusija eventualno upozori na neke moguće metodološke propuste koji su mogli dovesti ili do Greške Tip 1 (odbacio je H_0 kada je ona zapravo istinita u populaciji) ili do Greške Tip 2 (kada je H_0 zadržao, a u populaciji je zapravo istinita neka alternativna hipoteza, H_A). Kada se radi o samo jednom istraživanju, istraživač ne može biti siguran da li je na osnovu P-vrijednosti donio pravo odluku ili napravio neku od navedenih grešaka.

Kritičari testiranja nulte hipoteze (Bakan, 1966; Gigerenzer, 2004; Haig, 2016; Perezgonzalez, 2015) opisuju navedenu proceduru kao prijesnu mješavinu origi-

nalnog doprinosa Roland Fishera, tvorca P-vrijednosti, i Jerzy Neymana and Egon Pearsona, autora teorije testiranja hipoteza čiji je osnovni postulat kontrolisati nivo grešaka u zaključivanju u dugom nizu istraživanja. Na primjer, Gigerenzer navodi kako je Fisher u poznijoj istraživačkoj karijeri (1) zagovarao upotrebu P-vrijednosti kao kontinuirane mjere dokaza protiv nulte hipoteze, (2) jedinstvenu nultu hipotezu sa parametrom nula smatrao kao prikladnu samo u kontekstima kada postoji izuzetno malo relevantnih informacija, te (3) insistirao na tome da se P-vrijednosti ispod kritične tačke .05 moraju više puta replikovati kako bi se nalaz uvrstio u domen naučnog. Na drugoj strani, Neyman i Pearson su smatrali nužnim da se pri procesu testiranja neke hipoteze H_0 (koja može imati parametar 0 ili neki drugi), u obzir mora uzeti i dodatna alternativna hipoteza H_A. Logično bi bilo da alternativna hipoteza operacionalizuje najmanju veličinu efekta u populaciji koja bi predstavljala praktično osjetnu razliku. Na primjer, ako je za osnivanje univerzitetskog psihološkog savjetovališta dovoljan argument da je 10% ili više studenata anksiozno, onda bi H_A mogla glasiti da je broj anksioznih studenata u populaciji 10% (a ne nultom hipotezom pretpostavljenih 5%). Uvodeći pojam Greške Tip 2, Neyman i E. Pearson su predvidjeli da istraživači treba unaprijed da pretpostave stopu vjerovatnoće pogrešnog zaključka (β) kada je H_A istinita, odnosno da treba da znaju u koliko ponovljenih istraživanja bi greška uzorkovanja mogla biti uzrok statistički neznačajnog rezultata. Međutim, realnost je takva da se podaci o statističkoj moći testa i dalje relativno rijetko spominju u radovima, a i kada se spomenu najčešće su tu kao usputna informacija zasnovana na maglovitim predstavama o snazi efekta na osnovu nominalnih statističkih pragova koje je predložio Cohen (1992). Situaciju će možda promijeniti konkretnije preporuke o tome kako da definišemo smislene minimalno važne veličine efekta (vidjeti npr. Anvari & Lakens, 2019; i King, 2011), ali zasad se naučni časopisi u oblasti psihologije na našim prostorima, a i šire, dominatno oslanjaju na P-vrijednost i nivoe značajnosti kojim se razmatra isključivo nulta hipoteza.

Može se zaključiti da je pojednostavljena procedura testiranja značajnosti nulte hipoteze krajnje zavodljiva usljed sljedećeg. Kao prvo, P-vrijednosti su univerzalno prisutne u svim zamislivim vrstama kvantitativnih nacrta – bilo da su oni kategorički, numerički ili mješoviti, bivarijatni ili multivarijatni, deskriptivni ili eksperimentalni – a princip zaključivanja je banalno jednostavan: ako je $P < \alpha$ onda imamo dokaz o efektu, a ako je $P \geq \alpha$ praktično nemamo dokaz[4]. Usput, tradicija se potrudila da se mahom eliminiše i problem definisanja kritičnih vrijednosti, pa je nivo .05 postao sveprisutan[5]. Drugo, nameće se slika istraživača čija je uloga nakon planiranja nacrta i prikupljanja podataka svedena na minimum. Naime, objektivne principe frekvencionističke vjerovatnoće će na prikupljene podatke primijeniti statistički algoritam, a ne istraživač. Zadatak istraživača je samo da prepiše P-vrijednosti dobijene u statističkom ispisu i racionalno ih komenta-

[4] Ipak, nerijetko se nalazi diskutuju u pravcu istinitosti alternativne hipoteze ukoliko je P-vrijednost i dalje relativno niska (npr. < .10).

[5] Mada su posljednjih godina glasne masovne inicijative da se nivo značajnosti automatski pooštri na .005 (Benjamin et al. 2018) ili da se konkretna kritična vrijednost argumentovano definiše u svakom pojedinačnom istraživanju (Lakens et al., 2018). Pogledati i masovni kritički odgovor na te inicijative (Trafimow et al. 2018).

riše u skladu sa teorijskim postavkama. Ovo ostavlja moćan privid o neprikosnovenoj objektivnosti (za konceptualna objašnjenja izvora iluzije vidjeti Berger & Berry, 1988; i Gelman & Hennig, 2017).

Nažalost, prikupljeno je sijaset empirijskih dokaza da je u stvarnosti situacija drugačija: svaki istraživač mora da napravi čitav niz subjektivnih odluka kada prikuplja podatke i statistički ih analizira, a u izvještaju se samo dobijenim P-vrijednostima posvećuje adekvatna pažnja (Gelman & Loken, 2013; Leek & Peng, 2015; Simmons, Nelson & Simonsohn, 2011; Wicherts et al., 2016). Globalni potres koji je ukazao na neodrživost postojeće strukture izazvao je projekat 100 nezavisnih istraživačkih timova širom svijeta (Open Science Collaboration, 2015). Izvedeno je ukupno 100 direktnih replikativnih istraživanja, čiji je cilj bio da provjeri održivost zaključaka iz članaka objavljenih u tri vrlo respektibilna psihološka časopisa. U samo 36.1% istraživanja su nalazi dobijeni na osnovu P-vrijednosti bile saglasni sa nalazima originalnih izvještaja gdje je P-vrijednost bila niža od .05. Uz to, veličina efekta u replikacionim istraživanjima je bila prepolovljena u odnosu na ono što je dobijeno u originalnim istraživanjima. Uslijedila je široka kampanja samopreispitivanja koja je nadišla ekspertsku zajednicu i prelila se u druge naučne discipline i opšti javni diskurs, gdje se naziva *krizom reproducibilnosti* ili *krizom replikabilnosti*.

Osim netransparentnosti istraživačkog postupanja i izvještavanja, na listi krivaca za to stanje se našlo i pretjerano oslanjanje na testiranje značajnosti nulte hipoteze. Naime, više nije bilo sumnje da se P-vrijednostima daje prekomjeran značaj koji direktno uslovljava odluku hoće li uredništvo htjeti objaviti neki članak ili ne. Unutar visoko kompetitivne discipline, opterećene metrijskim pristupom produktivnosti (vidjeti Lindner, Torralba & Khan, 2018; Nosek, Spies & Motyl, 2012), takva praksa podstiče površne i preslobodne istraživačke poteze pri statističkoj analizi. Naime, kako će biti jasnije iz ostatka teksta, u mnogim slučajevima kada je P-vrijednost nešto niža od .05 alternativna hipoteza može biti podjednako slabo vjerovatna kao i nulta. Isto tako, P-vrijednosti veće od .05 mogu da govore snažnije u prilog alternativne hipoteze nego nulte. I ovim napokon dolazimo do uloge BF, jer nam BF omogućava da kvantifikovano, na kontinuiranoj mjernoj skali, saznamo ono zbog čega smo i započeli istraživanje: Uzimajući u obzir ranija znanja koja imamo o pojavi i podatke koje smo prikupili svojim istraživanjem, u kojoj mjeri je jedna hipoteza vjerovatnija od druge?

Šta je to Bayesov faktor (BF)?

Najsažetija definicija, koju ćemo naknadno dopuniti, kaže da je BF omjer vjerovatnoća dvije suprotstavljene hipoteze. Najprije ću podsjetiti da omjer predstavlja količnik vjerovatnoće da se desi jedan ishod u odnosu na neki drugi ishod: omjer = p / (1-p). U statističkoj komunikaciji se omjer iskazuje ili sa dva broja (npr. 2:1, 3:1) ili se drugi dio (:1) podrazumijeva i izostavlja. Na primjer, vjerovatnoća da bacajući igraću kockicu sa šest strana dobijete paran broj jeste 1 (ili 1:1) s obzirom da postoje 3 mogućnosti za paran broj i 3 mogućnosti za neparan broj (3:3 = 1:1). S druge strane, omjer vjerovatnoća da dobijete broj 6 pri jednom bacanju je 1:5 (ili 0.20), a ne 1:6 što bi zapravo predstavljalo vjerovatnoću p = 0.167 ili 16.7% da se

dobije šestica. Iz naprijed navedenog slijedi da omjer 1 predstavlja potpuno jednaku šansu dva upoređena ishoda. Omjeri veći od 1 opisuju da je dati ishod vjerovatniji i to onoliko puta koliko iznosi konkretan omjer. Na primjer, dva puta je veća šansa da se na kockici dobije broj 3 ili veći u odnosu na ishode 1 ili 2. S druge strane, omjeri niži od 1 označavaju manju vjerovatnoću te hipoteze. Omjere niže od 1 slabije intuitivno razumijemo, te bi se oni po pravilu trebali izražavati u inverznom obliku[6] (npr. ukoliko je omjer jednak 0.5, onda je suprotni ishod dva puta vjerovatniji: 1/0.5 = 2; u primjeru sa kockicom i dobijanjem šestice, 5 puta je veća vjerovatnoća da nećete dobiti šesticu: 1/0.2 = 5).

Obje suprotstavljene hipoteze koje se uzimaju u obzir prilikom poređenja moraju biti statistički definisane. Mada to nije nužno, obično je jedna od njih klasična, jedinstvena nulta hipoteza, dok se druga (ili obje) definiše kao čitav interval vrijednosti duž kojeg vjerovatnoća potencijalnog parametra može biti promjenjiva. Radi lakšeg razumijevanja koncepta BF i njegovog ključnog aspekta koji se odnosi na definisanje alternativnih hipoteza, krećem sa opisom najjednostavnijeg mogućeg slučaja kada bi obje hipoteze bile definisane kao jedinstvene tačke na mjernoj skali (a analogno primjeru koji prikazuje Morey, 2014).

Ilustrativni primjer izračunavanja BF

Uzmimo za primjer situaciju u kojoj istraživač pretpostavlja da na uzrastu od 16 mjeseci više od 50% djece (H_0: $\pi = .50$) može njima nepoznate predmete koji su po nekoj karakteristici vizuelno slični svrstati u istu grupu na osnovu zajedničkog imena (primjer iz istraživanja Tutnjević & Lakić, 2018). Štaviše, recimo da istraživač pretpostavlja da je taj postotak tačno 60% (H_A: $\pi = .60$). Pretpostavimo dalje da je u istraživanju bilo 100 djece ($n = 100$) od kojih je 58 uspješno riješilo zadatak, što bi ukazivalo na to da je dijete ovladalo pomenutom sposobnošću. Ukoliko bismo testirali značajnost nulte hipoteze uobičajenim putem, koristili bismo binomni test sa specifikovanom H_0. Ako istraživač ima dovoljno razloga da vjeruje da je procenat djece veći od 50% moglo bi se argumentovati da se koristi jednosmjerni test. U takvom proračunu bismo dobili P-vrijednost .067, te kao posljedicu zadržali nultu hipotezu[7].

Ono što se još može izračunati osim P-vrijednosti, za koju već znamo da predstavlja površinu distribucije za rezultat 58 ili veći, jeste tačna vjerovatnoća (p) da se dobije statistik 58 i ona iznosi: $p(n_+ = 58 \mid \pi = .50) = 0.0223$. Koristeći binomnu distribuciju očekivanih statistika, konkretna vjerovatnoća se može izračunati i za alternativnu hipotezu i ona iznosi: $p(n_+=58 \mid \pi = .60) = 0.0742$. Dijeljenjem ove dvije vjerovatnoće dobijamo da je – uzimajući u obzir podatke na uzorku – alternativna hipoteza 3.33 puta vjerovatnija od nulte (.0742/.0223 = 3.327). Dobijeni rezultat

[6] Druga mogućnost je da se koriste logaritmovani omjeri. Tada je umjesto vrijednosti 1, neutralna vrijednost jednaka 0, a skala negativnih i pozitivnih vrijednosti je simetrična i ukazuje na stepen dokaza za jednu ili drugu hipotezu.

[7] Za uobičajeni dvosmjerni test bismo dobili još veću P-vrijednost (.133), a 95% interval povjerenja bi obuhvatao vrijednosti od .49 do 1.00.

predstavlja omjer vjerodostojnosti dvije hipoteze/modela (eng. likelihood ratio, *LR*) što je u datom slučaju identično Bayesovom faktoru, pa je $BF_{10} = 3.33$. Shodno tome, inverzna vrijednost Bayesovog faktora koja govori o snazi dokaza u prilog nulte hipoteze iznosi 0.30 ($BF_{01} = 1/3.33 = 0.30$).

Međutim, teško je opravdati jedinstvenu alternativnu hipotezu i situacije su kompleksnije u realnosti. Kako bih to pokazao proširujem navedeni primjer najprije time što ćemo sada pretpostaviti da su oba istraživača imala različite i tvrdoglave vizije alternativnih hipoteza. Dok je prvi istraživač (A) pretpostavljao da je populaciona vrijednost 60%, drugi istraživač (B) je pretpostavljao da je ta vrijednost 70%. Ako nemamo jake argumente da prije istraživanja preferiramo jednu od te dvije hipoteze, odnosno ako su one apriorno pojednako vjerovatne, možemo sada izračunati BF tako što ćemo uprosječiti dobijene omjere vjerodostojnosti[8]: $BF_{10} = \{[p(n_+=58 \mid \pi = .60) / p(n_+=58 \mid \pi = .50)] + [p(n_+=58 \mid \pi = .70)/ p(n_+=58 \mid \pi = .50)]\} / 2 = 1.74$. Dakle, nakon uzimanja u obzir podataka sa uzorka alternativna hipoteza da je populaciona vrijednost ili 60% ili 70% je i dalje nešto vjerovatnija nego nulta hipoteza da je to tačno 50%. BF je niži nego u slučaju kada smo imali samo hipotezu da je prevalencija jednaka 60%, budući da je vjerovatnoća $p(n_+=58 \mid \pi = .70)$ tek 0.003.

Dalje se nadovezujući, ukoliko bi prije istraživanja pomenuti istraživači pristali na neki kompromis, logično bi bilo da se postavi niz realističnih alternativa u pogledu parametra koje bi na primjer sezale od 55% do 75%. Vrijednosti unutar tog opsega bi mogle biti ili podjednako vjerovatne ili bi se neke tačke na skali mogle smatrati vjerovatnijim (npr. 65%). U svakom slučaju, naprijed opisani proces određivanja vjerovatnoće parametra prije nego što su podaci uopšte prikupljeni jeste najbitnija definišuća karakteristika Bayesianske statistike, kao i procedura neophodna za kalkulaciju Bayesovog faktora. Prije nego što stignemo do kalkulacija vezanih za ono što se naziva subjektivnim apriornim vjerovatnoćama, pomoći će nam da prvo na istom istraživačkom primjeru razmotrimo kako bi to izgledao potpuno objektivni Bayesov faktor.

Ako o tome promislimo, brzo ćemo se složiti da u teoriji postoji određena početna distribucija vjerovatnoća koja bi bila potpuno objektivna. Naime, sve moguće apriorne vrijednosti parametra bi tada morale biti jednako vjerovatne, odnosno sve vrijednosti od 0% do 100% bi morale imati jednako kvantifikovanu početnu vjerovatnoću. Tu pretpostavku oslikava uniformna distribucija, poznata i kao Beta distribucija sa parametrima 1 i 1 (Slika 2). S obzirom na to da se radi o kontinuiranoj dimenziji, gdje su tačke na skali beskonačno male (npr. postoji mogućnost da je parameter jednak 33.33333333%), da bismo izračunali vjerovatnoće za sve moguće alternativne hipoteze neophodno je da koristimo integralni račun. Ali, sasvim dobru procjenu možemo dobiti i korištenjem aproksimativne mreže (eng. approximative grid), što ću iskoristiti u didaktičke svrhe. Konkretno, u tom slučaju bi kalkulacija mogla podrazumijevati da se za svaki cijeli broj od 0 do 100 zasebno izračuna omjer vjerodostojnosti modela, te da se ta suma – analogno primjeru sa samo dvije

[8] Ukoliko bismo hipotezu od 60% smatrali dva puta vjerovatnijom prije početka istraživanja, onda bismo konačan BF mogli dobiti proporcionalnim ponderisanjem omjera za A i B slučaj.

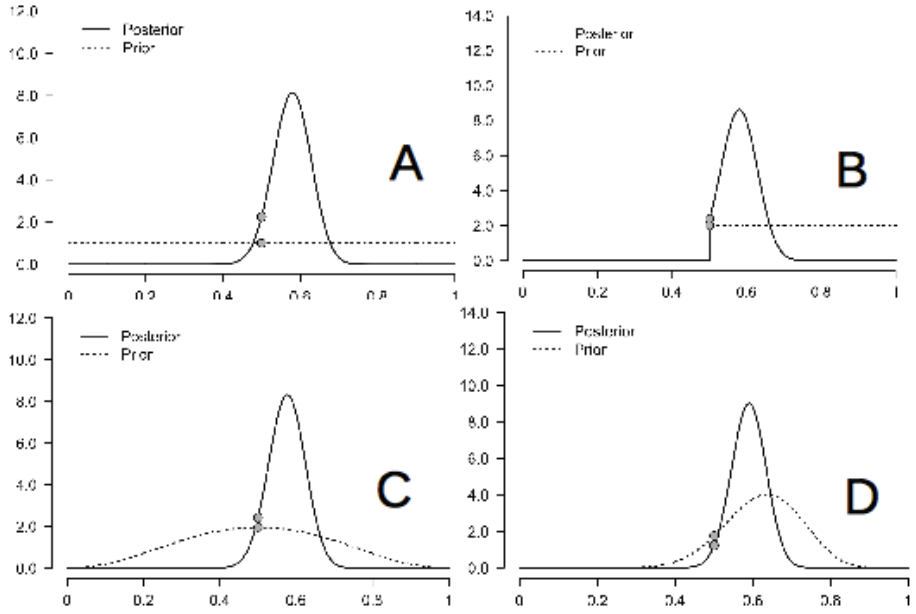
alternativne istraživačke hipoteze – podijeli sa ukupnim brojem modela koji je 101. Rezultat koji se dobija takvom približnom procjenom (BF_{01} = 2.27) praktično je jednak onome što se dobija integralnim računom (BF_{01} = 2.25) ili kada se mjerna skala umjesto na 101 tačku podijeli na 1 000 001 tačku (BF_{01} = 2.25). Dakle, nulta hipoteza je tada nešto više od dva puta vjerovatnija od potpuno objektive ili neinformativne alternativne hipoteze (podsjećanja radi, za dvosmjerni binomni test – koji bi bio frekvencionistički pandan objektivnom BF – je dobijena P-vrijednost .133).

Razmotrimo još neke razumne mogućnosti za koje bi se istraživači mogli odlučiti pri postavljanju apriornih vjerovatnoća. Jedna od njih je analogna upotrebi jednosmjernog binomnog testa. Nultu hipotezu bismo ostavili definisanu kao jedinstvenu vrijednost parametra na 50%, ali bismo alternativnu specifikovali kao niz vrijednosti parametra većih od 50%, sve do 100%. U tom slučaju kalkulacijom dobijamo da je BF_{01} = 1.19 (BF_{10} = 0.84) što bi značilo da su nulta i alternativna hipoteza, praktično gledano, podjednako vjerovatne. Međutim, u slučaju kada istraživači žele da ustanove da li je većina djece datog uzrasta ovladala određenom sposobnošću čini se da bi još racionalnije bilo direktno uporediti dvije kombinovane hipoteze: jednu, da je taj postotak veći od 50%, naspram druge, da je postotak niži od 50%. Tada podaci sa uzorka relativno jasno govore u prilog hipoteze da je većina djece ovladala datom sposobnošću: BF_{+-} = 17.12. Inače, BF se može izračunati u takvoj situaciji zahvaljujući osobini tranzitivnosti, odnosno mogućnosti da se izračuna omjer omjera, kada postoji zajednički nazivnik, tj. $BF_{+-} = BF_{+0} \ / \ BF_{-0}$ = 0.839/0.049 = 17.12.

Još dva tipa apriornih vjerovatnoća su veoma prisutna u istraživačkoj literaturi: kada se alternativna hipoteza definiše kao difuzna hipoteza koja svoj mod ima na tački nultog parametra (tzv. slabo informativna alternativna hipoteza), te kada se na osnovu postojećih saznanja definišu informativne apriorne distribucije vjerovatnoće parametra. Za ilustraciju prvog tipa, pretpostavimo da su istraživači već uzeli u obzir da je praktično nemoguće da vrijednosti budu 0% ili 100%, pa su tim vrijednostima dodijelili krajnje niske vjerovatnoće, dok su vrijednosti oko 50% smatrane vjerovatnijim (npr. Beta(3.257,3.257)). Slika 2 pokazuje tu pretpostavku. Ovakvu distribuciju je čak moguće kombinovati sa komparacijama kompozitnih hipoteza predstavljenim u prethodnom pasusu i upravo takvu strategiju smo i koristili kao autori istraživanja u ranije objavljenom članku (Tutnjević & Lakić, 2018). Treba napomenuti da je BF_{10} u tom slučaju nešto niži nego kada se kao apriorne vjerovatnoće koristi uniformna distribucija.

Drugi tip, koji podrazumijeva definisanje potpuno informativnih hipoteza, zasniva se na pretpostavci da postoji validan korpus znanja o vjerovatnoći parametara prije nego što su podaci za dato istraživanje i prikupljeni. Na primjer, bilo bi opravdano da rezultati istraživanja na našem stvarnom istraživanju sa malim brojem djece (n_+ = 15, n_- = 9, odnosno 62.5% tačno riješenih zadataka) budu uzeti u obzir kao apriorne vjerovatnoće za novo, replikativno istraživanje (na primjer, sa fiktivnim podacima sa 100 djece od kojih je 58 tačno uradilo zadatak; vidjeti Sliku 2). S obzirom da je i u ranijem istraživanju natpolovičan broj ispitanika riješio zadatak, BF_{10} bi tad bio veći od 1 (BF_{10} = 1.43). Informativne apriorne vjerovatnoće bismo

mogli definisati i ispitujući eksperte koji su tokom decenija radili u praksi sa djecom na pitanju razvoja jezičke kategorizacije. U svakom slučaju, statistički mehanizam kombinuje apriorne vjerovatnoće sa izgledošću modela nakon što su podaci uzeti u obzir, a kao rezultat se dobijaju ažurirane, posteriorne vjerovatnoće parametra. To je procedura kojom i puni Bayesianski pristup integriše racionalnost i kumulativnost, kao osnovna načela nauke, u statistički model.



Slika 2. Prikaz postavljanja različitih apriornih distribucija vjerovatnoće parametra. Isprekidane linije prikazuju apriorne distribucije – vjerovatnoće prije nego što su podaci prikupljeni. Pune linije prikazuju posteriorne vjerovatnoće koje kombinuju apriorne vjerovatnoće i rezultate na uzorku (58 od 100 djece tačno riješili zadatak). Grafikoni su izvedeni u softveru JASP. A) uniformna distribucija Beta(1, 1) gdje su sve vrijednosti parametra π od 0 do 1 podjednako vjerovatne; B) poluuniformna distribucija gdje su podjednako vjerovatne vrijednosti π od 0.5 do 1, dok se vrijednosti manje od 0.5 smatraju nemogućim; C) slabo informativna distribucija Beta(3.257, 3.257); D) informativna distribucija Beta(15, 9) postavljena na osnovu podataka iz postojećeg istraživanja.

Kategorizacija apriornih vjerovatnoća pri definisanju BF

Usljed prostornih ograničenja, načela definisanja apriornih vjerovatnoća i izračunavanje BF sam ilustrovao tek na jednom vidu statističkog nacrta, ali su oni identični kada su u pitanju drugi nacrti (tj. korelacije, razlike u prosječnim vrijednostima, komparacije modela). U svim kontekstima BF će predstavljati omjer vjerovatnoća dvije hipoteze u svjetlu prikupljenih podataka. Kako smo vidjeli, u zavisnosti od količine znanja koju imaju prije istraživanja, istraživači učitavaju različitu količinu

informacija u statistički model. Kvalitativno se ta količina može vrlo grubo podijeliti na tri pristupa (Depaoli, & van de Schoot, 2017).

Prvi, potpuno neinformativni i apsolutno objektivni pristup zadovoljava istraživače koji strepe od bilo kakve subjektivnosti u nauci; ali pritom, ukoliko ga koriste, istraživači moraju biti svjesni da eksplicitno priznaju da ne znaju apsolutno ništa sigurno u vezi sa pojavom od interesa. Analogno testiranju nulte hipoteze, u takvim situacijama se obično definiše jedinstvena nulta hipoteza, te alternativna hipoteza koja ukazuje na jednaku vjerovatnoću parametra duž svih empirijski mogućih vrijednosti (npr. korelacije od -1.0 do +1.0).

Drugi pristup se naziva slabo informativnim[9] i podrazumijeva da je istraživač svjestan da vrijednosti parametra ne mogu da dosegnu krajnje granice teorijske skale. Ipak, u skladu sa temeljnim naučničkim skepticizmom obično se jedinstvena nulta hipoteza poredi sa alternativnom hipotezom koja je u slučaju jednostavnih nacrta (tj. prevalencija, korelacija, razlika) simetrična i centrirana na nultom parametru. Istraživač dodjeljuje veću apriornu vjerovatnoću vrijednostima bliskim nuli, a konkretan tip (npr. normalna, Cauchy, Beta, t) i širina takve distribucije se preciziraju ili na osnovu istraživačke argumentacije (npr. maksimalna praktično zamisliva vrijednost efekta kao granična vrijednost distribucije, veličina efekta iz prethodnog istraživanja kao mjera varijabilnosti distribucije, vidjeti Dienes (2019)) ili na osnovu nominalnih statističkih principa (npr. referentni pragovi standardizovanih veličina efekata).

Konačno, subjektivistički informativni pristup podrazumijeva da će istraživač odrediti distribuciju prethodnih vjerovatnoća alternativnoj hipotezi na osnovu dosadašnjih znanja koja proističu ili iz rezultata ranijih empirijskih istraživanja ili namjenski provedenih ekspertskih procjena. Takve distribucije su često asimetrične i modalna vrijednost odgovara najvjerovatnijoj vrijednosti parametra koja obično nije nula. Postupak konstruisanja distribucije/a na osnovu ispitivanja eksperata se naziva elicitiranje (O'Hagan et al., 2006) i ono se implementira uz pomoć softverskih rješenja (npr. MATCH http://optics.eee.nottingham.ac.uk/ match/uncertainty.php, SHELF 4 http://www.tonyohagan.co.uk/shelf/SHELF4.html ili EXCALIBUR http://www.lighttwist.net/wp/excalibur). Očito, prednost informativnog pristupa je to što će procjene dobijene tim putem dati najefikasnije, kao i najvalidnije zaključke, naravno pod uslovom da su apriorne vjerovatnoće razumno dobre. Istovremeno, iz pozicije objektivistički nastrojenog istraživača informativni pristup je kulminacija onoga što usljed namjerne zloupotrebe ili subjektivnog neznanja može ugroziti naučnu procjenu, pa sljedeći segment posvećujem diskusiji o potencijalnom uticaju apriornih vjerovatnoća na rezultate.

Senzitivnost BF na apriorne vjerovatnoće i procjena snage dokaza BF

Koliko povjerenja možemo imati u BF ako njegove vrijednosti direktno variraju u zavisnosti od odluka istraživača? Zar to nije suprotno idealima objektivne nauke? Pitanje postaje pogotovo relevantno kada se BF uporedi sa testiranjem značaj-

[9] Kada je raspon efekta i teorijski neograničen - kao u slučaju testiranja razlika između grupa, gdje one mogu da iznose od $-\infty$ do $+\infty$ – govori zapravo o zadanom neinformativnom pristupu.

nosti nulte hipoteze gdje P-vrijednosti ne variraju sem u slučaju kada se istraživači odluče za jednosmjerne umjesto dvosmjernih testova. Postoje teorijski i empirijski odgovori na pitanje senzitivnosti BF usljed postavljanja različitih apriornih vjerovatnoća, a prvo ću prikazati teorijske.

Kako u širem kontekstu navodi Vanpaemel (2010), definisanje različitih apriornih vjerovatnoća je *sine qua non* zrele nauke. Nužno je moći specifikovati model koji oslikava mehanizam generisanja podataka, a ne samo oslanjati se na probabilističko dokazivanje da je jedan netačan model (nulta hipoteza) zaista netačan. Drugim riječima, potrebno je moći prikupiti dokaze za ono u šta se vjeruje da je istina, a ne samo pokazati da nešto nije istina. Kada se to shvati, postaje očito da je osjetljivost BF na različite apriorne vjerovatnoće zapravo poželjna karakteristika, a ne smetnja, budući da podaci moraju u različitoj mjeri podržavati različite teorijske postavke, odnosno različite istraživačke hipoteze.

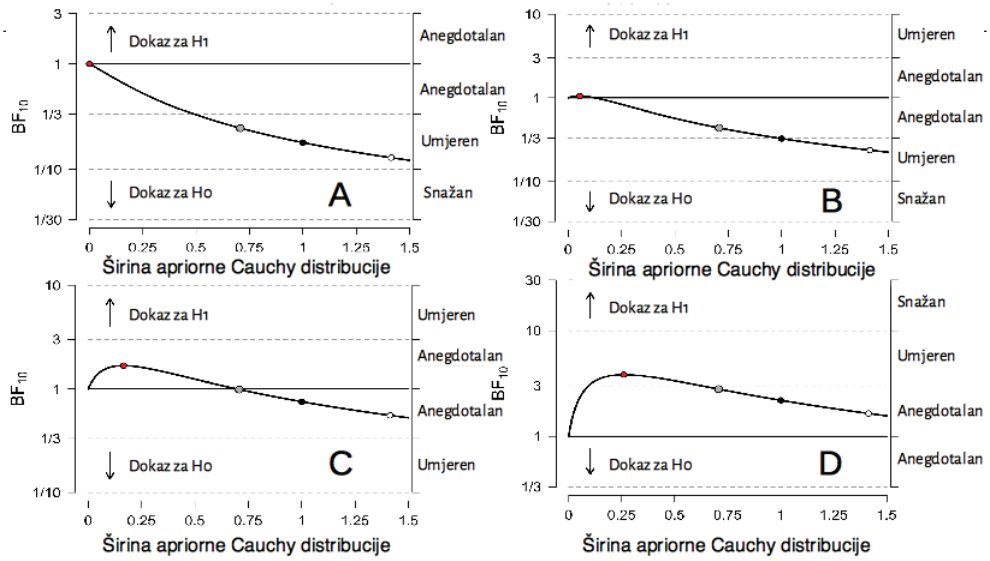
Naravno, dominantna bojazan skeptičnih je da se kao apriorne vjerovatnoće mogu postaviti potpuno arbitrarne vrijednosti koje dopuštaju da ih istraživači zapravo podese *post-hoc*, nakon što su podatke prikupili i analizirali. Na taj način bi nesavjesni istraživači mogli da zloupotrijebe fleksibilnost BF i da svoje istraživanje predstave značajnijim nego što jeste. Ali, čini se da je upravo suprotno tačno: ako bi neko potpuno arbitrarno definisao apriorne distribucije, to bi zapravo prije privuklo kritički pogled recenzenata i naučne zajednice nego što je to slučaj za trenutno važeći standard gdje se nekritički i automatizovano koriste jedinstvene nulte hipoteze, a da se alternativna hipoteza teorijski, odnosno statistički, ni ne definiše. Štaviše, korištenje potpuno neinformativnih hipoteza ili klasičnog *tabula rasa* stava dovodi do neefikasnih procjena koje zanemaruju postojeća istraživačka znanja ili čak i zdravu logiku. Koliko tumačenja rezultata usljed toga mogu biti različita može se vidjeti iz debate o statističkim dokazima u pogledu realnosti ekstrasenzorne percepcije (Bem, Utts & Johnson, 2011; Rouder & Morey, 2011; Wagenmakers et al., 2011, 2012).

Kada je u pitanju empirijska strana senzitivnosti apriornih vjerovatnoća evidentna je jedna snažna zakonitost: sa povećanjem broja podataka u istraživanju, smanjuje se uloga apriornih vjerovatnoća u kvalitativnoj interpretaciji rezultata. Drugim riječima, što su uzorci veći, apriorne vjerovatnoće postaju manje bitne kada je u pitanju odluka da li podaci govore u prilog jedne od hipoteza i zaključci su sve manje inkonkluzivni. Nadalje, u velikoj većini slučajeva slabo informativne apriorne vjerovatnoće igraju relativno slabu ulogu pri donošenju konačne odluke. Na primjer, u slučaju slabo informativnog pristupa se često upotrebljavaju ranije pominjane Beta distribucije. Za njih postoji nešto što se naziva efektivna veličina uzorka, odnosno može se izračunati koliko podataka zapravo same apriorne distribucije predstavljaju, tako što se jednostavno saberu a i b parameter koji definišu Beta distribuciju. Konkretno, uniformna distribucija Beta (1,1) podrazumijeva da smo imali uzorak od samo 2 ispitanika, dok bi Beta (10,10) podrazumijeva da smo imali uvjerenje koje bi bilo analogno istraživanju sa 20 ispitanika prije nego što smo započeli naše istraživanje. Ipak, istina je da vrijednosti BF mogu da variraju tako da promijene smjer dokaza. Kako bi bilo jasno u kojoj mjeri se to dešava, neophodno je najprije da definišemo značenje veličine BF.

Već sam naveo da vrijednost *BF* = 1 podrazumijeva da su obje razmatrane hipoteze podjednako vjerovatne. Veća odstupanja od 1 podrazumijevaju sve veću vjerovatnoću jedne od hipoteza, odnosno vrijednosti manje od 1 idu u pravcu dokaza za suprotnu hipotezu. Kao psiholozi smo se navikli da nam različite ekspertske preporuke o visini statistika (npr. u pogledu graničnih vrijednosti za indekse fita u strukturalnom modeliranju ili Cohenove sugestije o veličinama standardizovanih efekata) olakšaju odlučivanje i zbace sa istraživača odgovornost promišljanja. Ne iznenađuje zato činjenica da su neki autori preveli određene granične vrijednosti BF u kvalititativnu procjenu snage dokaza za jednu od hipoteza. Te generičke vrijednosti izgledaju smisleno, međutim, prije nego što ih navedem, ukazao bih na još jednu prednost upotrebe BF. Naime, s obzirom na to da je BF relativno intuitivna mjera, različiti korisnici informacija iz istraživanja (npr. istraživači, recenzenti, praktičari) mogu samostalno procijeniti snagu dokaza za hipotezu. Na primjer, moguće je držati se sljedećeg načela: „Postavi sebi takvu granicu BF nakon koje bi bio dovoljno uvjeren da ono što hipoteza implicira upotrijebiš sam u praksi ili da impliciranoj intervenciji podvrgneš bliske osobe (npr. partnera, dijete, roditelja)". Iako će ovaj princip nekima zazvučati zastrašujuće subjektivno – i teško primjenjivo u nekim fundamentalnim istraživanjima – njegovom upotrebom se može kalibrirati uvjerenje o praktičnom značaju koje nalazi imaju, što je veliki nedostatak apstraktnosti čvrsto fiksiranih P-vrijednosti, pa i standardizovanih veličina efekta.

Naravno, u akademskoj komunikaciji je preporučljivije držati se generičkih standarda, mada sa razumnim oprezom. Nekoliko radova (npr. Dienes & Mclatchie, 2017; Jarosz & Wiley, 2014; Kass & Raftery, 1995) poredi različite sugestije u vezi graničnih skorova. Ukratko, BF vrijednosti ispod 3 se generalno smatraju inkonkluzivnim i nedovoljno informativnim, odnosno tek anegdotalnim dokazom za neku od hipoteza. Vrijednost 3 često odgovara P-vrijednosti od .05 (Dienes, 2014) i obično se uzima kao prva indikativna granica za hipotezu, ali daleko od toga da predstavlja snažan dokaz. Empirijskim putem je utvrđeno da je razumna granica koju bi istraživači trebalo da dostignu u eksplorativnoj fazi barem 5, a poželjno 6, dok se tek za vrijednosti veće od 10 već može očekivati da predstavljaju snažan dokaz u prilog hipoteze (Schönbrodt & Wagenmakers, 2018). Vrijednosti veće od 100 se obično smatraju krucijalnim, odlučujućim dokazom u prilog hipotezi.

Sad kada su pojašnjeni interpretativni okviri, moguće je vratiti se na empirijsko razmatranje senzitivnosti BF. Jedna od preporuka je da istraživači sami provjere u svojim analizama koliko su vrijednosti BF robusne, tako što će kao apriorne vjerovatnoće varirati niz plauzibilnih parametara. Dati proces je olakšan time što je u najjednostavnijem softverskom rješenju, JASP-u (jasp-stats.org), dostupna i odgovarajuća opcija koja trenutno funkcioniše u slučaju neinformativnih i slabo informativnih apriornih distribucija parametara. Kao dobra ilustracija može poslužiti primjer koji prikazuje situaciju gdje su na uzorcima od po 100 ispitanika dobijene različite t-vrijednosti sa pratećim P-vrijednostima i Cohenovim d-vrijednostima.



Slika 3. Prikazane su četiri različite situacije gdje je korištena generička kalkulacija za dvosmjerni t-test za nezavisne uzorke. U svim situacijama obje grupe su imale po 100 ispitanika. Grafikon prikazuje kako smjer i snaga dokaza variraju u zavisnosti od apriorne širine distribucije vjerovatnoće parametra. Korištena je Cauchy distribucija za koju mjera širine pokazuje medijanu apsolutne veličine efekta. Širina sa vrijednošću 1 odgovara distribuciji u kojoj je 50% mogućih efekata između -1d i +1d, a 50% mogućih efekata ima veću apsolutnu vrijednost od 1d. Grafikoni su napravljeni u softveru JASP. A) $t(198) = 1.0$, $p = .319$, $d = 0.14$; B) $t(198) = 1.5$, $p = .135$, $d = 0.21$; C) $t(198) = 2.0$, $p = .047$, $d = 0.28$; D) $t(198) = 2.5$, $p = .013$, $d = 0.35$.

Uzmimo za primjer situaciju C (Slika 3) gdje je $t(198)=2.0$. Rezultat bi bio proglašen statistički značajnim i alternativna hipoteza bi bila prihvaćena sa malom veličinom efekta od $d = 0.28$. I zaista, BF mijenja svoj smjer i prelazi iz podrške alternativnoj hipotezi sa užim apriornim distribucijama u podršku nultoj hipotezi kada su apriorno predviđene i mnogo veće veličine efekta. Međutim, ono što je bitno jeste da vidimo da veličina BF ne izlazi iz okvira anegdotalnih dokaza. Štaviše, kada se informativna vjerovatnoća namjerno podesi na standardizovanu lokaciju .28 (vrijednost efekta na uzorku) i to sa vrlo malom varijansom distribucije (normalna distribucija, $M = 0.28$, $SD = 0.10$), BF ima tek vrijednost 5.95. Maksimalni BF bi bio dobijen za jedinstvenu alternativnu hipotezu na lokaciji 0.28d i iznosio bi 7.27, što znači da ni tada ne bi dostigao nivoe snažnog dokaza. Dakle, u jednostavnim nacrtima čak ni otvorene manipulacije, koje ne bi mogle promaći recenzentima, ne dovode nužno do drastično različite interpretacije BF (vidjeti i Etz, 2016, i Grange, 2015). Ipak, razborita je preporuka da autori istraživanja saopšte raspon BF vrijednosti za niz realističnih apriornih vrijednosti (Depaoli & van de Schoot, 2017).

Prednosti i ograničenja BF

Neke prednosti BF u odnosu na P-vrijednosti bi trebalo da budu očite iz dosad navedenog. Radi se o mjeri koja je intuitivnija za interpretaciju i koja se odnosi na

stepen dokaza za hipotezu u koju smo kao istraživači uvjereni. S obzirom na to da je BF u teoriji simetrična mjera sa centralnom vrijednosti 1 (u praksi potpuno simetrična kada koristimo logaritmovane omjere), za razliku od P-vrijednosti ona može da govori i u prilog nulte hipoteze, a takođe može da nam dâ do znanja da su rezultati istraživanja nedovoljno konkluzivni, i sve se to saopštava na kontinuiranoj skali. Vidjeli smo i da je BF konzistentniji indikator kada je u pitanju supstantivno znanje, u smislu da statistička operacionalizacija hipoteza može i treba da varira u zavisnosti od opšte prihvaćenog fundusa znanja o određenoj pojavi. Podaci iz pojedinačnog istraživanja moraju služiti kumulativnom ažuriranju uvjerenja koje već imamo o pojavi, a ne da pojedinačno istraživanje bude samo sebi svrha (Wagenmakers et al., 2017, 2018). Osvještavanje apriornih distribucija parametra uozbiljuje istraživački proces i u današnje vrijeme je to jednostavno provesti u intuitivnim grafičkim okruženjima.

Postoje i dodatne prednosti na koje ukazuje objavljena literatura. Na primjer, pokazano je da čak i upotreba veoma liberalnih (npr. BF > 3), a pogotovo razumnijih graničnih vrijednosti (npr. BF > 10) smanjuje vjerovatnoću lažno pozitivnih nalaza koja je dovela do krize reproducibilnosti (npr. Etz & Vandekerckhove, 2016). Kao posljedica upotrebe BF rezultati pojedinačnih istraživanja se opreznije tumače, što je pogotovo bitno za istraživanja za koja se zna da imaju eksplorativan karakter. S tim u vezi treba navesti još jednu prednost BF i Bayesianskog pristupa uopšte: za razliku od P-vrijednosti, BF je načelno imun na povećanje šanse da se načini greška u zaključivanju koje nastaje usljed višestrukog testiranja. Naime, stepen dokaza za hipotezu ostaje isti bez obzira na broj testiranja i zbog toga se BF posebno preporučuje za sekvencionо prikupljanje podataka, kada istraživač zaustavlja uzorkovanje onda kada dostigne određeni prag efekta izražen putem BF (npr. Rouder, 2014, Schönbrodt et al., 2017). Iz istog razloga je u istraživanjima u kojima se očekuje veliki broj testova moguće umjesto korekcija P-vrijednosti, ili uz njih, koristiti BF pragove (vidjeti npr. korištenje BF > 10, Lakić, Damjenić i Grahovac, 2018. u kontekstu obuhvatnog pristupa analizi multivarijatnih konstrukata ili Dušanić, Lakić i Turjačanin, 2019. za selekciju varijabli u kontekstu mrežne analize). Konačno, posebnu prednost predstavlja tranzitivnost BF što omogućava i naprednije opcije poređenja modela kao što su ordinalna poređenja unutar analize varijanse (Hoijtink et al., 2018) i direktne komparacije vjerovatnosti niza modela koji ne moraju biti ugnježđeni (npr. u sklopu opšteg linearnog modela, Rouder & Morey, 2012). Ove opcije su već uvrštene u softver JASP.

Vidjeli smo da se subjektivnost i senzitivnost apriornih vjerovatnoća mogu tumačiti i kao prednost i kao ograničenje BF što svakako treba uzeti u obzir pri evaluaciji upotrebe BF. Takođe, BF tehnički otežava potvrdu malih veličina efekata, ali je pokazano da je to zapravo logički konzistentna i poželjna osobina BF, kao i svakog naučnog metoda (za objašnjenje vidjeti Morey, 2015). Međutim, upućuju su i neke druge zamjerke BF, a zanimljivo je da one dolaze od Bayesianski orijentisanih statističara (npr. Gelman & Rubin, 1995; Kruschke, 2011, str. 311–312; Kruschke & Liddel, 2018a i 2018b; Kiers & Tenderio, 2019; van der Linden & Chryst, 2017; Tendeiro & Kiers, 2019). Jedna od ključnih je da se upotrebom BF ne koristi čitav arsenal koji stoji na raspolaganju u statističkoj procjeni snage efekata i neizvjesnosti

modela. Na primjer, dok je za BF dovoljno definisati samo distribuciju alternativne hipoteze, puni Bayesianski pristup obavezuje da se jasno definišu apriorne distribucije svih parametara iz kojih potiču podaci (npr. oblik izvorne populacije što može da adresira i problem stršećih mjera, varijabilnost podataka i mjera veličine efekta). Iz toga slijedi da i pretpostavke o valjanosti zaključaka na osnovu BF ovise o tome da li su zadovoljeni bazičniji preduslovi; naravno, validnost nacrta i validnost i pouzdanost mjerenja, ali i statistički preduslovi kao što su razlike u varijabilitetima i postojanje stršećih mjera. Prateća zamjerka je i da BF ima onoliko smisla koliko ga imaju i obje hipoteze, a kritičari prvenstveno naglašavaju da je u većini slučajeva potpuno nepotrebna upotreba jedinstvene nulte hipoteze (doduše, po ovom pitanju postoji napredak, vidjeti Morey & Rouder, 2011). Nadalje, puna Bayesianska analiza insistira na intervalnoj procjeni parametara. Kombinovanjem apriornih distribucija i podataka sa uzorka dobija se posteriorna distribucija parametara od interesa na kojoj je moguće odrediti uvjerljive intervale koji su supstantivno kompletniji od odluke o hipotezi.

Zaključak

Navedene kritike jesu ozbiljne i može se zaključiti da BF nije lijek za sve. Ipak, nepobitno je da u trenutnom spletu okolnosti BF smisleno proširuje statistički repertoar koji se koristi u psihologiji. Ako je nešto naučeno iz krize reproducibilnosti to je da statistička analiza ne smije počivati na samo jednom, automatizovanom kriterijumu kao što je to unazad nekoliko decenija bio slučaj. U vezi s tim predlaže se da se BF koristi zajedno sa P-vrijednostima (Dienes & Mclatchie, 2018; primjeri na našem prostoru Lepir, Lakić i Takšić, 2018, ili Mirković & Lakić, 2019) ili zajedno sa testom ekvivalentnosti koji i zasebno predstavlja smisleniji modus od P-vrijednosti (Lakens et al. 2018). Realno gledajući, ranije jesu postojale tehničke prepreke za korištenje BF u smislu da su softverska rješenja bila nedostupna, ali danas takvih prepreka nema. JASP (JASP Team, 2019) – kao i blizak srodnik Jamovi (The jamovi project, 2019) sa dopunskim paketom jsq – veoma je jednostavan za upotrebu, potpuno besplatan (open source) softver koji omogućava izračunavanje BF za velik broj procedura. Uz to, korisnicima R okruženja stoje na raspolaganju paketi BayesFactor (Morey & Rouder, 2018) i bain (Gu et al., 2019) koji takođe omogućavaju laku kalkulaciju BF u velikom broju nacrta. Dakle, sve je spremno da i vi počnete upotrebljavati BF u svojim istraživanjima.

Reference

Anvari, F., & Lakens, D. (2019, February 1). *Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest.* https://doi.org/10.31234/osf.io/syp5a

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423.

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*, 716–719.

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. (2017). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10. doi:10.1038/s41562-017-0189-z

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*(2), 159–165.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi: 10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p< .05). *American Psychologist, 49,* 997–1003. doi: 10.1037/0003-066X.50.12.1103

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240.

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

Dienes, Z. (2019, February 28). How do I know what my theory predicts?. https://doi.org/10.31234/osf.io/yqaj4

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

Dušanić, S., Lakić, S. i Turjačanin, V. (2019). *Građansko i političko učešće mladih: Psihološki pristup* (2. izdanje). Banja Luka: Friedrich Ebert Stiftung.

Etz, A. (2016, 19. jun). Understanding Bayes: How to cheat to get the maximum Bayes factor for a given p value [Blog post]. Preuzeto sa https://alexanderetz.com/2016/06/19/understanding-bayes-how-to-cheat-to-get-the-maximum-bayes-factor-for-a-given-p-value/

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PloS one*, *11*(2), e0149794.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gelman. A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology, 25,* 165–173.

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 967–1033.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can bea problem, even when there is no "fishing expedition" or "p- hacking" and the research hypothesis was posited ahead of time.* Preuzeto sa http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Grange, J. (2015, 27. novembar). Animating Robustness-Check of Bayes Factor Priors [Blog post]. Preuzeto sa jimgrange.wordpress.com/2015/11/27/animating-robustness-check-of-bayes-factor-priors/

Gu, X., Hoijtink, H., Mulder, J. & van Lissa, C. J. (2019). bain: Bayes Factors for Informative Hypotheses. R package version 0.2.1. https://CRAN.R-project.org/package=bain

Haig, B. D. (2016). Tests of Statistical Significance Made Sound. *Educational and Psychological Measurement*, 1–18. https://doi.org/10.1177/0013164416667981

Hoijtink, H., Mulder, J., van Lissa, C., and Gu, X. (2018). A tutorial on testing hypotheses using the Bayes factor. Psychological Methods. DOI: 10.1037/met0000201

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2. http://dx.doi.org/10.7771/1932-6246.1167

JASP Team (2019). JASP (Version 0.10)[Computer software].

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kiers, H., & Tendeiro, J. (2019, April 5). With Bayesian Estimation One Can Get All That Bayes Factors Offer, and More. https://doi.org/10.31234/osf.io/zbpmy

King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(2), 171–184.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS.* Burlington, MA.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Buchanan, E. M. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168.

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests, *The Journals of Gerontology: Series B*, , gby065, https://doi.org/10.1093/geronb/gby065

Lakić, S., Damjenić, M. i Grahovac, S. (2018). HEXACO crte ličnosti kao korelati korištenja pojedinačnih strategija učenja srednjoškolskih učenika. *Banjalučki novembarski susreti 2017 – Zbornik radova.* Banja Luka: Filozofski fakultet.

Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature News*, *520*(7549), 612.

Lepir, D., Lakić, S. i Takšić, V. (2018). Relacije bavljenja sportom i emocionalne inteligencije na srednjoškolskom uzrastu. *Primenjena psihologija, 11,* 285–300.

Lindner, M. D., Torralba, K. D., & Khan, N. A. (2018). Scientific productivity: An exploratory study of metrics and incentives. *PloS one*, *13*(4), e0195321.

Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of science*, *34*(2), 103–115.

Mirković, B. & Lakić, S. (2019). The Effects of Brand Popularity and the Big Five on Perceived Quality of Refreshment Products: An Exploratory Study. *Empirical Studies in Psychology – Proceedings of the XXV Scientific Conference.* Faculty of Philosophy, University of Belgrade.

Morey, R. D. (2014, 9. februar). What Is a Bayes Factor? [Blog post]. Preuzeto sa richarddmorey.org/2014/02/what-is-a-bayes-factor/

Morey, R. D. (2015, 10. april). All about that "bias, bias, bias" (it's no trouble). [Blog post]. Preuzeto sa http://bayesfactor.blogspot.com/2015/04/all-about-that-bias-bias-bias-its-no.html

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419. http://dx.doi.org/10.1037/a0024377

Morey R. D. & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... & Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

Open Science Collaboration. (2015, August 28). Estimating the reproducibility of psychological science. *Science, 349*(6251): aac4716. http://dx.doi.org/10.1126/science.aac4716

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology, 6(MAR), 1–11. https://doi.org/10.3389/fpsyg.2015.00223

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*(4), 682–689.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308.

Rouder, J. N. & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research, 47*, 877–903.

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322.

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods.*Advance online publication. http://dx.doi.org/10.1037/met0000221

Tenjović, L., & Smederevac, S. (2011). Mala reforma u statističkoj analizi podataka u psihologiji: malo p nije dovoljno, potrebna je i veličina efekta. *Primenjena Psihologija, 4*, 317–333.

Tutnjević, S., & Lakić, S. (2018). Language-mediated object categorization: A longitudinal study with 16-to 20-month-old Serbian-speaking children. *European Journal of Developmental Psychology*, *15*(5), 608–622.

The jamovi project (2019). *jamovi* (Version 1.0.1) [Computer Software]. Retrieved from https://www.jamovi.org

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C. J., Beh, E. J., Bilgiç, Y. K., ... & Chaigneau, S. E. (2018). Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology*, *9*.

van der Linden, S., & Chryst, B. (2017). No need for Bayes Factors: A fully Bayesian evidence synthesis. *Frontiers in Applied Mathematics and Statistics*, *3*, 12.

Van Doorn, J., Matzke, D., & Wagenmakers, E.-J. (2019). An In-Class Demonstration of Bayesian Inference. https://doi.org/10.1177/1475725719848574

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011).

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. *Psychological science under scrutiny: Recent challenges and proposed solutions*, 123–138.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25,* 35–57.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. http://dx.doi.org/10.3389/fpsyg.2016.01832

.

Siniša Lakić[10]
University of Banja Luka
Faculty of Philosophy
Department of Psychology
Banja Luka, Republic of Srpska, Bosnia and Herzegovina

BAYES FACTOR: WHAT IT IS AND WHY IT SHOULD BE USED IN PSYCHOLOGICAL RESEARCH

Abstract

The aim of this paper is to get our wider psychological audience acquainted with the Bayes factor (commonly denoted as BF or B). In recent times, BF became a highly po-

[10] Corresponding author email: sinisa.lakic@ff.unibl.org

pular statistical method for hypothesis testing in psychology, with claims that it could replace the role of P-values. Its increasing popularity is evident from the results of a Google Scholar search using the terms “Bayes factor” and “psychology”: there were only 76 hits in 2006, 176 in 2010, 436 in 2014, while the number of papers published only in 2018 reached 1570.[11] Despite this thriving trend, I could not find texts written in our languages (BCS) describing BF and clarifying its claimed advantages over P-values. Thus, psychology students, practitioners who want to keep up with research trends, as well as experienced researchers, have to read relevant papers in English. Unfortunately, those texts are often saturated with advanced statistical terminology and notation, which certainly impedes understanding and demotivates the readers. For that reason, I tried to write this paper using language which should be understood by all who possess fundamental statistical knowledge. I begin by describing motives for using BF, after which I present its theoretical background through some straightforward examples. I finish by presenting the advantages and limitations of BF, and suggesting which software the interested readers should use in order to calculate BF for various quantitative designs and thus, incorporate a new paradigm in their research repertoire.

Keywords: Bayes factor, P-values, statistical hypothesis testing

Primljeno: 07. 07. 2019.
Primljena korekcija: 26. 09. 2019.
Prihvaćeno za objavljivanje: 06. 09. 2019.

[11] That this increase is not a result of an overall increase in the number of published papers can be seen from the analogous search using the terms “psychology” and “regression”. After the increase, one observes a decrease in 2018: 66400 (2006), 106000(2010), 117000(2014), 50600(2018). In both cases, patents and citations were excluded from the search.