



Jaká je skutečná chybovost hladiny významnosti $p < 0,05$ a proč nepoužívat označení statisticky významný/nevýznamný rozdíl

Kritika hodnoty p jakožto kritéria statisticky významných rozdílů

Jan F. Vojáček

I. interní kardioangiologická klinika, LF UK a FN Hradec Králové
Kardiologie na Bulovce, s. r. o., Praha

Hledání statisticky významných vztahů hodnocených hladinou významnosti pomocí výpočtu hodnoty p bylo donedávna základem vědeckého výzkumu, klinických studií a následně i zavádění nových postupů v rámci medicíny založené na důkazech. V poslední době jsme však svědky celkem zásadní revize tohoto postupu, dobře sumarizované například v American Statistical Association statement on p -values (1).

Skupina předních statistiků publikovala v roce 2018 rozsáhlý přehled o nedostatku reprodukovatelnosti vědeckých studií, který vyvolává rostoucí obavy z důvěryhodnosti závěrů studií založených na „statisticky významných“ výsledcích. Jako příčiny špatné reprodukovatelnosti výsledků studií byly označeny nízké statistické standardy důkazů pro tvrzení o nových objevch v mnoha oblastech vědy. Automatická akceptace „statisticky významného“ nálezu pouze na základě výpočtu s $p < 0,05$ vede k extrémně vysoké míře falešně pozitivních výsledků i při absenci jiných experimentálních, procedurálních či publikačních problémů (2).

Recentně byla publikována řada článků o nesprávném pochopení a nesprávném výkladu hodnoty p a mnoho špičkových časopisů je kritických k použití hodnoty p , jakožto průkazu statisticky významného výsledku výzkumu.

Hodnota p není a nesmí být zaměňována za chybovost neboli pravděpodobnost faleš-

ného odmítnutí správné nulové hypotézy – čili průkazu účinnosti metody, která ve skutečnosti účinná není (chyba typu I). Tuto chybovost lze chápat jako falešnou pozitivitu pro test hypotézy. Falešná pozitivita znamená, že ačkoliv ve skutečnosti je **zkoumaný vzorek odlišný od kontrolního pouze kvůli náhodě**, přesto vytváří falešně nízkou hodnotu p . Navzdory nízké hodnotě p však alternativní hypotéza není pravdivá. Na úrovni populace neexistuje žádný rozdíl například mezi starým a novým léčebným postupem.

Tlak na nalezení nízkých hodnot p v kombinaci s nepochopením toho, jak správně interpretovat hodnoty p , může fatálně zkreslit interpretaci výsledků.

Hodnota p poskytuje pouze omezené informace s vysokou chybovostí v **závislosti na předchozí pravděpodobnosti hypotézy**. Například hodnota p mezi 0,03 až 0,05 sama o sobě nabízí pouze slabý důkaz proti nulové hypotéze. Podobně ale ani relativně vysoká hodnota p nemusí znamenat důkaz ve prospěch nulové hypotézy; z těchto důvodů by analýza dat neměla končit výpočtem p -hodnoty, pokud jsou vhodné a proveditelné jiné přístupy (1).

Nulová hypotéza předpokládá, že v množině daných proměnných neexistuje mezi dvěma soubory žádný statistický vztah a významnost, že hodnocené náhodné výběry pocházejí z jednoho základního souboru. Nulová hypotéza je

označena jako H_0 a vyjadřuje předpoklad, že není rozdíl v průměrech hodnocených proměnných mezi dvěma populacemi (i s přihlédnutím k ukazatelům variability dat v souboru):

$$H_0: \mu_1 = \mu_2$$

kde H_0 = nulová hypotéza, že neexistuje rozdíl, μ_1 = průměr populace 1 a μ_2 = průměr populace 2.

Úroveň významnosti, označovaná také jako α , je pravděpodobnost nesprávného odmítnutí nulové hypotézy i přesto, že je pravdivá. Například hladina významnosti 0,05 označuje 5% riziko nesprávného závěru, že rozdíl existuje, i když ve skutečnosti není žádný rozdíl. Na obrázku 1 je znázorněna úroveň významnosti 0,05 ve formě červeně zbarvené plochy v hodnotě 2,5 % (0,025) hustoty pravděpodobnosti rozdělení na ose y , která je na obou koncích křivky, každá oblast má pravděpodobnost 0,025, celkem tedy 0,05. Jako praktický a zjednodušený příklad pro grafické znázornění jsme náhodně zvolili jako průměr populace hodnotu 854 a průměr zkoumaného vzorku 892. Průměr zkoumaného vzorku zasahuje do červeně zbarvené plochy, proto můžeme odmítnout nulovou hypotézu a konstatovat, že vzorek nepochází ze základního souboru a významně se liší na hladině významnosti 5 % (0,05). Tyto oblasti označují kritickou hodnotou pro dvoustranný test. Kritická oblast definuje, jak daleko

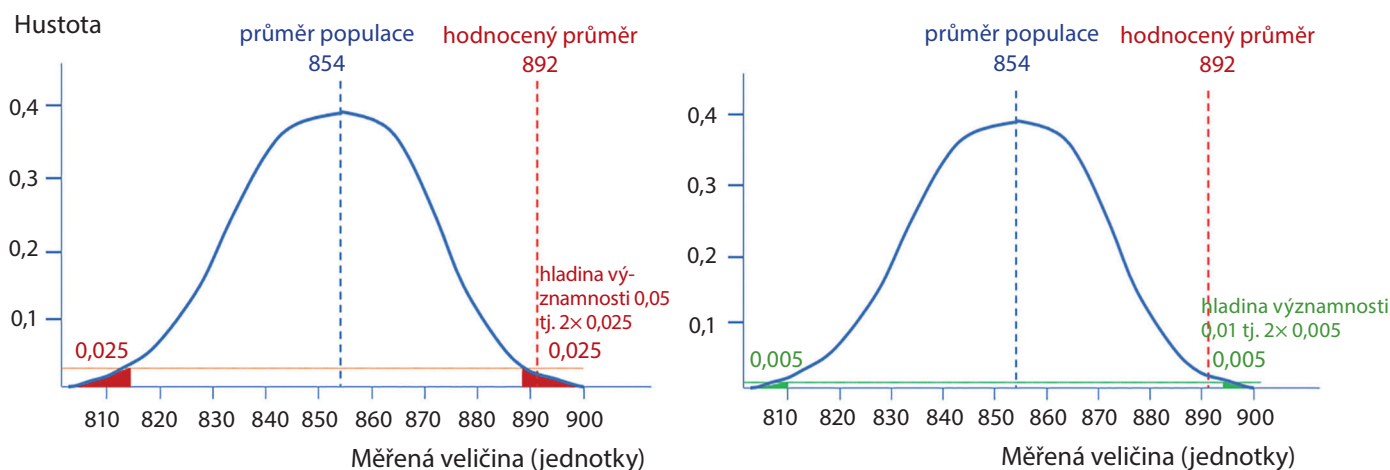
KORESPONDENČNÍ ADRESA AUTORA:

prof. MUDr. Jan F. Vojáček, DrSc., janvojacek@seznam.cz

I. interní kardioangiologická klinika LF UK a FN Hradec Králové, Sokolská 581, 500 05 Hradec Králové

Cit. zkr: Interv Akut Kardiolog. 2022;21(2):64-70

Obr. 1. Na obrázku je znázorněna úroveň významnosti 0,05 ve formě červeně zbarvené plochy v hodnotě 2,5% (0,025) hustoty pravděpodobnosti rozdělení na ose y, která je na obou koncích křivky, každá oblast má pravděpodobnost 0,025, celkem tedy 0,05. Jako praktický a zjednodušený příklad pro grafické znázornění jsme náhodně zvolili jako průměr populace hodnotu 854 a průměr zkoumaného vzorku 892. Průměr zkoumaného vzorku zasahuje do červeně zbarvené plochy, proto můžeme odmítnout nulovou hypotézu a konstatovat, že vzorek nepochází ze základního souboru a významně se liší na hladině významnosti 5 % (0,05). Tyto oblasti označují kritickou hodnotou pro dvoustranný test. Kritická oblast definuje, jak daleko musí být naše výběrová statistika od hodnoty nulové hypotézy, než můžeme říci, že je dostatečně neobvyklá a odmítnout nulovou hypotézu. Průměr porovnávaného vzorku 892 spadá do kritické oblasti, což naznačuje, že se liší na úrovni spolehlivosti 0,05. Můžeme také zkoumat, zda je statisticky významný pomocí jiné běžné hladiny významnosti 0,01. To je znázorněno na pravé části obrázku, kde každá ze dvou zeleně zbarvených oblastí má pravděpodobnost 0,005, což dává dohromady celkovou pravděpodobnost 0,01. Tentokrát průměr porovnávaného vzorku 892 nespadá do kritické oblasti, a proto nelze odmítnout nulovou hypotézu (15)

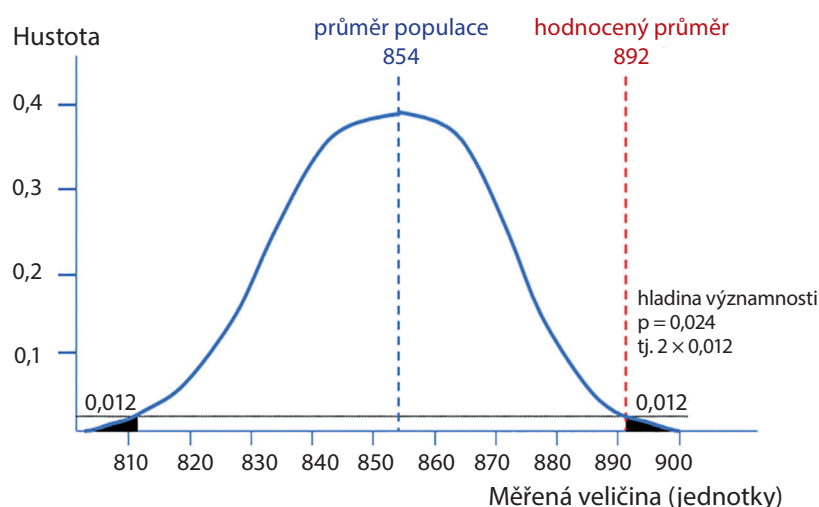


musí být naše výběrová statistika od hodnoty nulové hypotézy, než můžeme říci, že je dostatečně neobvyklá a odmítnout nulovou hypotézu. Průměr porovnávaného vzorku 892 spadá do kritické oblasti, což naznačuje, že se liší na úrovni spolehlivosti 0,05. Můžeme také zkoumat, zda je statisticky významný pomocí jiné běžné hladiny významnosti 0,01. To je znázorněno na pravé části obrázku, kde každá ze dvou zeleně zbarvených oblastí má pravděpodobnost 0,005, což dává dohromady celkovou pravděpodobnost 0,01. Tentokrát průměr porovnávaného vzorku 892 nespadá do kritické oblasti, a proto nelze odmítnout nulovou hypotézu. Toto srovnání ukazuje, proč je zásadně důležité vybrat úroveň významnosti před zahájením výzkumu. Dodatečný (ex post) výběr úrovně významnosti by mohl poskytovat falešně významné výsledky.

Hodnota p

Ronald Fisher zavedl hodnoty p v roce 1920, protože chtěl objektivní metodu pro porovnávání dat s nulovou hypotézou, spíše než neformální odhad: „Moje data vypadají jinak než nulová hypotéza.“ Výpočty hodnoty p zahrnují velikost efektu, velikost vzorku a variabilitu dat do jediného čísla, které objektivně říká, jak konzistentní jsou získaná data s nulovou hypotézou. Na obrázku 2 mají černě vybarvené koncové oblasti vpravo s hodnotou

Obr. 2. Na obrázku mají černě vybarvené koncové oblasti vpravo s hodnotou 892 pravděpodobnost výskytu 0,012 s celkovou pravděpodobností 0,024. Tento údaj představuje pravděpodobnost získání střední hodnoty vzorku, která je přinejmenším stejně extrémní/limitní jako průměr porovnávaného vzorku na obou koncích distribuce, pokud je průměr základního souboru 854 a hodnota 0,024 odpovídá hodnotě p (15)

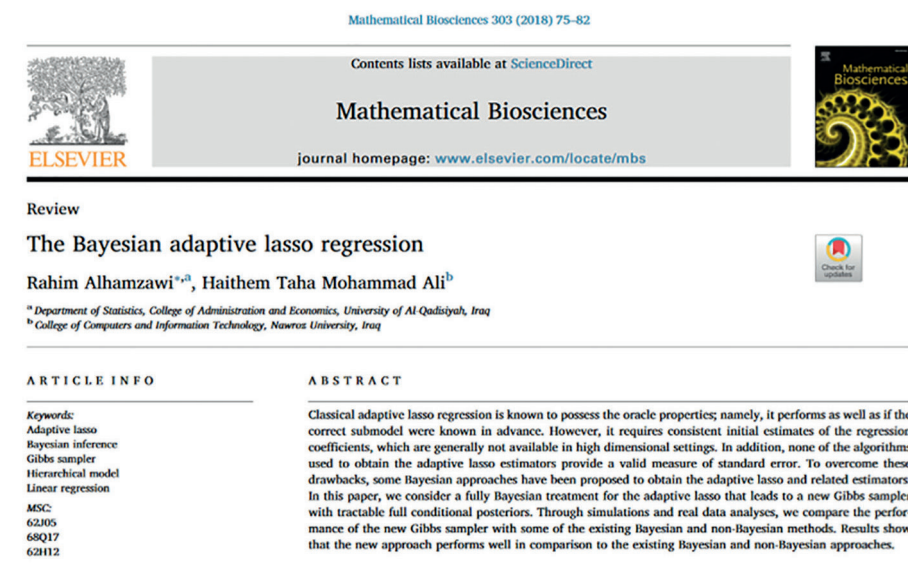


892 pravděpodobnost výskytu 0,012 s celkovou pravděpodobností 0,024. Tento údaj představuje pravděpodobnost získání střední hodnoty vzorku, která je přinejmenším stejně extrémní/limitní jako průměr porovnávaného vzorku na obou koncích distribuce, pokud je průměr základního souboru 854 a hodnota 0,024 odpovídá hodnotě p . Pokud je hodnota p menší nebo rovna nastavené hladině významnosti, odmítneme nulovou hypotézu. Pokud vezme hodnotu p pro náš příklad a porovnáme ji s běžnými úrovněmi významnosti, odpovídá předchozím grafickým výsledkům. Hodnota

$p = 0,024$ se liší na úrovni α (alfa) 0,05, ale ne na úrovni 0,01. **Častou chybou je interpretace hodnoty p jako stupně pravděpodobnosti, že nulová hypotéza je pravdivá.**

Při studii s postupem, o kterém víme, že je zcela neúčinný, tedy navzdory tomu, že nulová hypotéza je pravdivá, je možné, že bude rozdíl v datech v důsledku náhodné chyby měření. Ve skutečnosti je nepravděpodobné, že by hodnoty v experimentální skupině byly přesně stejné jako u kontrolní skupiny. V důsledku toho pozorovaný rozdíl ve vzorku nemusí odrážet skutečný rozdíl mezi populacemi. Hodnota p

Obr. 3. Publikace, které používají bayesovský přístup, získávají postupně uplatnění i v jiných oborech (matematika, astronomie, kvantová mechanika). Vlevo dole stoupající počet publikací podle PubMed.gov v posledních letech



Chybou je interpretace hodnoty p , jakožto pravděpodobnosti chyby odmítnutím skutečné nulové hypotézy (chyba typu I).

Je velmi důležité si uvědomit, že hladina $p < 0,05$ neznamená < 5% chybovost neboli pravděpodobnost falešného odmítnutí správné nulové hypotézy – čili průkaz účinnosti metody, která ve skutečnosti účinná není (chyba typu I).

Bayesovská versus četnostní statistika (Bayesian vs. frequentists statistics)

Navzdory tomu, že bayesiánská statistika je starší než frekventistická (neboli klasický četnostní přístup), počty publikací, které používají bayesovský přístup, v poslední době výrazně stoupají a tento přístup získává postupně stále širší uplatnění, především v medicíně, ale i v jiných oborech (obrázek 3).

Základem bayesovské teorie pravděpodobnosti je jediná písemná vědecká práce Reverenda Thomase Bayese „Esej o řešení problému doktriny pravděpodobnosti“. Ta je dodnes používána ve statistice jako Bayesův teorém a byla publikována ve Philosophical Transactions až po smrti autora 23. prosince 1763. Pozoruhodné je, že práci našel v pozůstalosti zesnulého Thomase Bayese jeho přítel – jakýsi Mr. Price – a toto pojednání považoval za tak unikátní, že by nemělo upadnout v zapomnění a s tímto zdůvodněním ji poslal dopisem členovi Královské společnosti Johnu Cantonovi. Ten měl právo práci přečíst na zasedání Royal Society of London a jak tehdy bylo zvykem, všechny práce tam přednesené byly následně otištěny v jednom z nejstarších tehdejších časopisů Philosophical Transactions of the Royal Society, aby mohly být rozeslány všem členům této společnosti (3).

V bayesiánské statistice jsou všechny relevantní informace pro vytvoření inference obsaženy v pozorovaných datech a validita je udržována správnou specifikací probablistického modelu. Zatímco četnostní statistika přijímá nebo odmítá nulovou hypotézu, bayesovský probablistický model nepoužívá nulovou ani alternativní hypotézu a závisí na správném stanovení **pravděpodobnosti před testem (prior probability)** dále **věrohodnosti (likelihood)** hypotézy a na jejich základě je odhadnuta **pravděpodobnost po testu**

1. Introduction

The standard multiple linear regression model assumes that a vector of centered responses $y = (y_1, \dots, y_n)^T$ can be written as

$$y = X\beta + \epsilon, \quad (1)$$

where λ is a tuning parameter. Lasso regression reduces to an ordinary least square (OLS) regression when the tuning parameter $\lambda = 0$. Model (2) continuously shrinks the unknown regression parameter vector β toward 0 as the tuning parameter λ increases, and some regression parameters are shrunk to exact 0 if λ is sufficiently large. The predictors associated with zero estimated regression parameters are considered

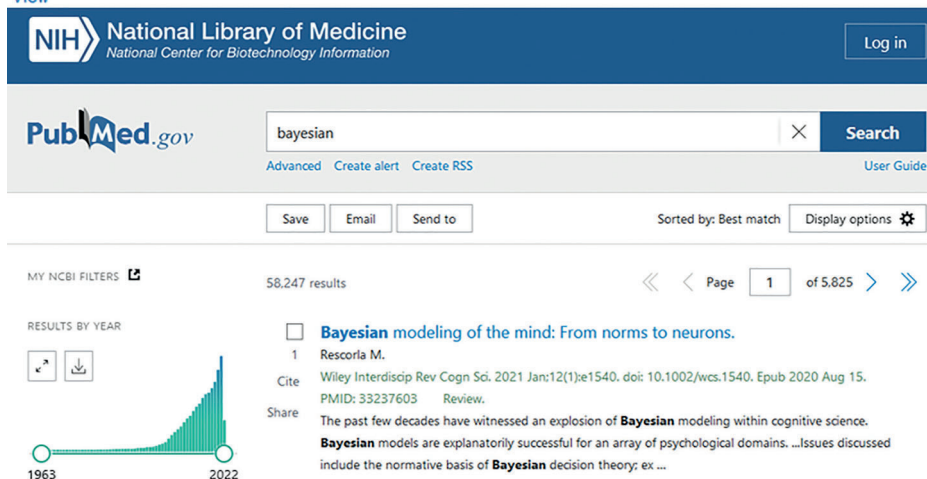
Nearly model independent constraints on dense matter equation of state in a Bayesian approach

Preprint Full-text available Mar 2022

N. K. Patra · Sk Md Adil Imam · B. K. Agrawal · Arunava Mukherjee · Tuhin Malik

We apply Bayesian approach to construct a large number of minimally constrained equations of state (EoSs) and study their correlations with a few selected properties of neutron star (NS). Our set of minimal constraints includes a few basic properties of saturated nuclear matter and low density pure neutron matter EoS which...

View



ukazuje stupeň pravděpodobnosti, že nulová hypotéza je pravdivá. Nízká hodnota p ukazuje, že zkoumaný vzorek poskytuje dostatek důkazů k odmítnutí nulové hypotézy pro celou základní skupinu.

Z technického hlediska vyjadřuje hodnota p pravděpodobnost získání hodnot dat ve zkoumaném vzorku za předpokladu pravdivosti nulové hypotézy. Předpokládejme ve studii

hodnotu $p = 0,04$. Tato hodnota p znamená, že pokud by mezi zkoumanými postupy nebyl žádný rozdíl, získali bychom významný rozdíl ve 4 % studií právě kvůli náhodné chybě při odběru vzorků. Hodnota p řeší pouze otázku, jak pravděpodobná jsou data za předpokladu platnosti nulové hypotézy. Neměří podporu alternativní hypotézy. Toto omezení je příčinou časté dezinterpretace hodnot p .

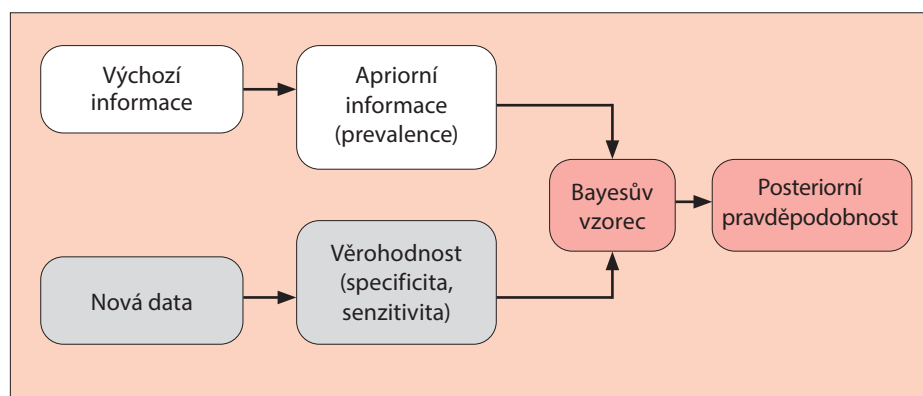
(**posterior probability**). Bayesovská statistika odpovídá na otázku: jaká je pravděpodobnost platnosti hypotézy vzhledem k naměřeným datům, odhaduje poměr pravděpodobností hodnocených hypotéz (4, 5) (obrázky 4 a 5).

Jedním z prvních praktických využití bayesovské statistiky v medicíně bylo vyhodnocení prediktivní hodnoty zátěžového testu k diagnostice ischemie myokardu u populace s nízkou prevalencí onemocnění a s necharakteristickými příznaky (6). **Pravděpodobnosti před testem** odpovídá známá **prevalence onemocnění** ve zkoumané populaci, dále **věrohodnost (likelihood)** hypotézy je v tomto příkladu správné stanovení **klinické specificity a senzitivity** testu a na jejich základě je odhadnuta **pravděpodobnost po testu (posterior probability = v uvedeném příkladě prediktivní hodnota pozitivního nebo negativního testu)**. Bayesovská statistika odpovídá na otázku: jaká je pravděpodobnost platnosti hypotézy vzhledem k naměřeným datům, odhaduje poměr pravděpodobností hodnocených hypotéz. V uvedeném příkladu se ukázalo, že i při provedení zátěžového testu se specificitou pro diagnostiku ischemie myokardu 95 % je při prevalenci onemocnění pouhých 5 % pouze 18 % z pozitivních testů správně pozitivních, ostatní jsou falešně pozitivní. Metoda proto nebyla doporučena k diagnostice ischemie myokardu v populaci (obrázek 6).

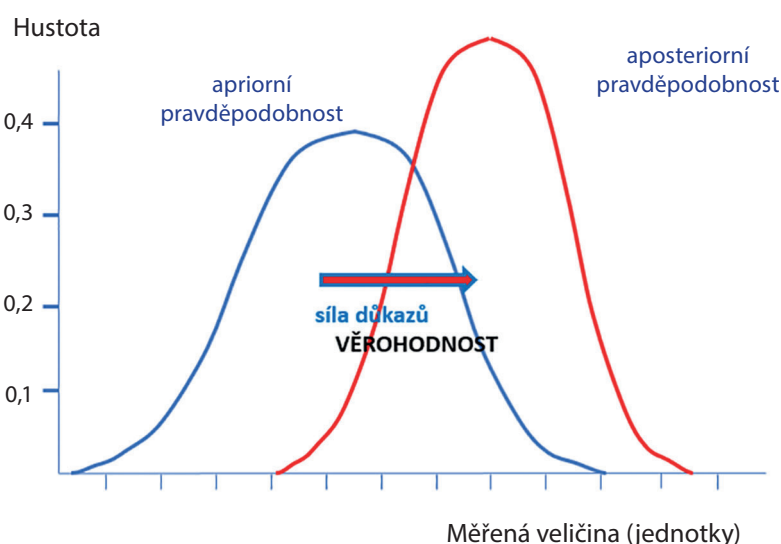
Využití v medicíně předpokládá znalost prevalence zkoumaného jevu (apriorní pravděpodobnost) a znalost spolehlivosti provedeného měření – zde například specificity a senzitivity provedeného testu, věrohodnosti hypotéz (v Bayesově teorii likelihood – věrohodnost), z těchto údajů můžeme podle Bayesova vzorce odvodit skutečnou pravděpodobnostní distribuci po provedení testu, tj. skutečnou prediktivní pravděpodobnost při pozitivním výsledku. Výsledná pravděpodobnost tedy podle bayesovské teorie závisí nejen na spolehlivosti a výsledku provedeného testu, ale i na apriorní pravděpodobnosti.

Hlavním problémem některých teoretiků v bayesovském přístupu je nedostatek objektivních informací o apriorních pravděpodobnostech (to však naštěstí neplatí v medicínském použití – zde většinou máme data o prevalenci nebo statistiky z předchozích

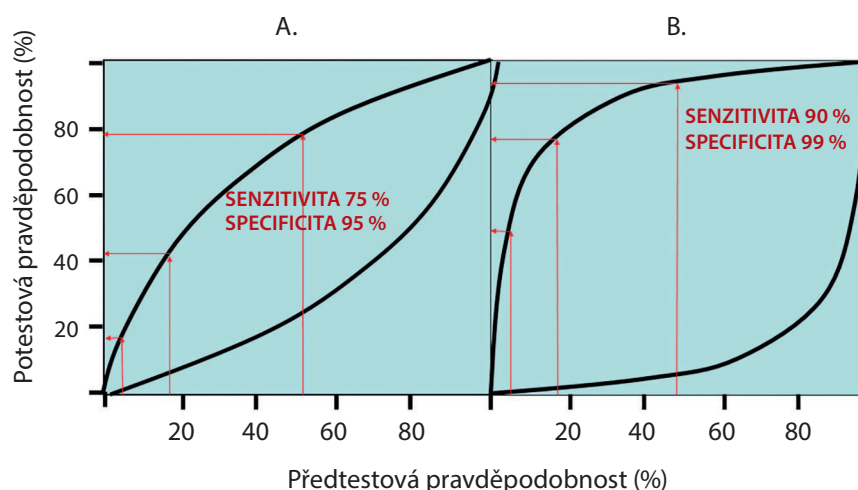
Obr. 4. Bayesovský probabilistický model nepoužívá nulovou ani alternativní hypotézu a závisí na správném stanovení pravděpodobnosti před testem (prior probability), dále věrohodnosti (likelihood) hypotézy a na jejich základě je odhadnuta pravděpodobnost po testu (posterior probability)

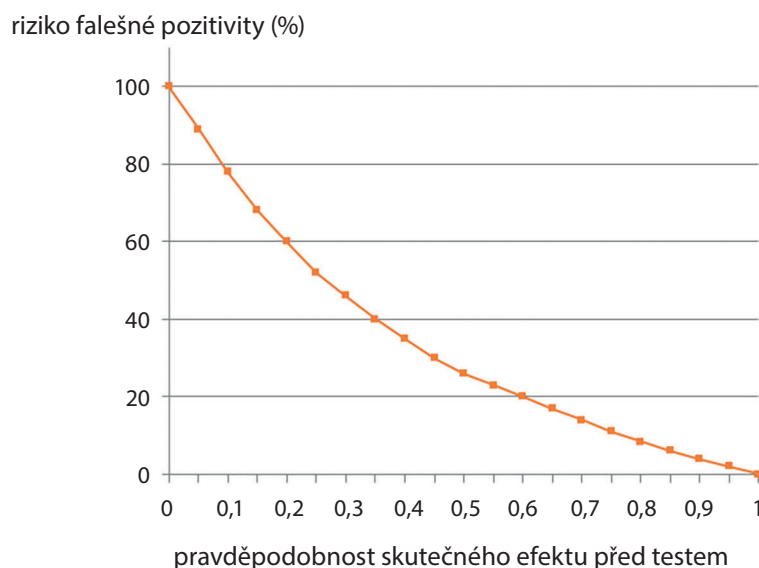


Obr. 5. Bayesovská statistika je založena na apriorní a aposteriorní distribuční funkci ovlivněné získanými důkazy – věrohodností (likelihood dle Bayese) a odpovídá na otázku: jaká je pravděpodobnost platnosti hypotézy vzhledem k naměřeným datům, odhaduje poměr pravděpodobností hodnocených hypotéz (4, 5)



Obr. 6. Bayesovská statistika v uvedeném příkladu ukazuje, že při provedení zátěžového testu se specificitou pro diagnostiku ischemie myokardu 95 % je při prevalenci onemocnění pouhých 5 % pouze 18 % z pozitivních testů správně pozitivních, ostatní jsou falešně pozitivní. Při stoupající prevalenci stoupá počet správně pozitivních testů. Prevalence v populaci závisí mimo jiné na věku, pohlaví a typu příznaků



Obr. 7. Závislost falešně pozitivního výsledku hodnoty p těsně pod 0,05 (0,045–0,05) na předchozí pravděpodobnosti reálného efektu (8)**Tab. 1.** Okolnosti určující míru falešné positivity hodnoty p

1	Prevalence reálných účinků	čím vyšší, tím méně falešně pozitivních výsledků
2	Síla studie – pravděpodobnost zjištění skutečného účinku zkoumaného postupu	čím vyšší, tím méně falešně pozitivních výsledků
3	Přesnost použitých metod zkoumání	čím vyšší, tím méně falešně pozitivních výsledků
4	Vypočtená hodnota P	čím nižší, tím méně falešně pozitivních výsledků

Tab. 2. Bayesův teorém

$$\text{Posteriorní pravděpodobnost} = \frac{\text{Pravděpodobnost} \times \text{Apriorní pravděpodobnost}}{\text{Důkaz}}$$

$$P(SP/PT) = \frac{P(/ PT/ SP) \times P(SP)}{P(PT)}$$

SP = skutečná (správná) pozitivita, PT = pozitivní test, P = pravděpodobnost

studii). Proto byl zaveden teoretický **princip lhostejnosti (the principle of indifference, principle of insufficient reason)**, který v bayesovské teorii požaduje, aby, pokud neexistují žádná předchozí relevantní data, všechny možné zvažované vstupy měly rovnoměrně rozloženou pravděpodobnost.

Z tohoto hlediska teoreticky existují následující tři možnosti:

Informativní prior (An informative prior) vyjadřuje konkrétní, definitivní informace o proměnné.

Posteriór z jednoho problému se stává priorem pro jiný problém; již existující důkazy se stávají součástí dalšího hodnocení, a jak se hromadí více důkazů, apriorní předpoklady jsou určeny převážně důkazy spíše než jakýmkoli původním teoretickým subjektivním odhadem. Pojmy „apriorní“ a „posteriorní“ jsou obecně relativní. Jako apriorní musí být

hodnoceny pozitivní i negativní výsledky, jinak bude uvedena chybná apriorní pravděpodobnost.

Slabě informativní prior (A weakly informative prior) vyjadřuje částečnou předchozí informaci o proměnné.

Neinformativní prior (An uninformative prior) vyjadřuje vágní nebo obecné předchozí informace o proměnné. Nejjednodušším a nejstarším pravidlem pro určení neinformativního prioru je princip lhostejnosti (**The principle of indifference**), který přiřazuje stejnou pravděpodobnost všem možnostem. V problémech s odhadem parametrů použití neinformativního prioru obvykle přináší výsledky, které se příliš neliší od konvenční statistické analýzy, protože pravděpodobnostní funkce často poskytuje více informací než neinformativní prior.

Jiným praktickým využitím bayesovského přístupu nejen v moderní kardiologii, ale i

v medicíně všeobecně, je hodnocení tzv. významného nebo nevýznamného rozdílu mezi dvěma či více hodnocenými soubory v klinickém výzkumu a v klinických studiích a stanovení skutečné **prediktivní pravděpodobnosti** a posouzení potestové **chybovosti** statistické **hodnoty „p“**.

Tento přístup vychází z posouzení **předtestové pravděpodobnosti** očekávaného rozdílu mezi hodnocenými postupy (**apriorní pravděpodobnosti – prior probability**), zhodnocení **věrohodnosti zkoumaných** hypotéz, výsledkem je **posteriorní pravděpodobnost** hypotéz pro daná data a skutečná **prediktivní pravděpodobnost pozitivního výsledku výzkumu nebo studie**.

Sellke a spolupracovníci spočítali, že hodnota $p = 0,05$ může mít míru **falešné positivity**, která v závislosti na **prevalenci skutečných účinků** dosahuje 20–70 % (7). Pokud má studie nižší pravděpodobnost zjištění skutečného účinku, vyšší podíl pozitivních výsledků bude falešně pozitivní (obrázek 7). Prevalence skutečných účinků má dopad na chybovost výsledku. Čím nižší prevalence skutečných účinků, tím vyšší chybovost výsledku.

Zatímco Sellke a spolupracovníci (7) použili bayesovský přístup, Colquhoun et al. (8) došel k podobným odhadům nebayesovským výpočtem.

To, co Colquhoun (8) nazývá **prevalencí reálných účinků** (označovaných jako **P(real) – reálné p**), bayesovský přístup nazývá předchozí (**apriorní**) **pravděpodobnost**. Je to podíl testů hypotéz, ve kterých je alternativní hypotéza pravdivá ještě před hodnocením výzkumem, a to buď subjektivní teoretickou úvahou, nebo na základě znalostí předchozích výzkumných dat. Lze si to představit jako dlouhodobou pravděpodobnost nebo záznam podobných typů studií. Je to věrohodnost alternativní hypotézy. Pokud je alternativní hypotéza nereálná nebo má špatné předchozí výsledky, $P(\text{real})$ je nízké. Například prevalence 0,1 naznačuje, že 10 % podobných alternativních hypotéz se ukázalo jako pravdivé, zatímco v 90 % byla pravdivá nulová hypotéza. Alternativní hypotéza může být neobvyklá, netestovaná nebo jinak nepravděpodobná. Pokud alternativní hypotéza odpovídá současné teorii, má identifikovaný patofyziologický mechanismus účinku a před-

chozí studie již ukázaly významné výsledky, $P(\text{real})$ je vyšší. Například prevalence 0,90 znamená, že alternativa je pravděpodobná na 90 %, kdežto nulová pouze na 10 %. Pokud je prevalence 0,5, existuje šance 50/50, že na začátku studie je pravdivá buď nulová, nebo alternativní hypotéza. Určitým problémem může být, že nejsou k dispozici apriorní data, ale vodítkem mohou být teoretické úvahy a výsledky předchozích pilotních studií.

Míru falešné positivity ovlivňuje prevalence výskytu skutečně správného rozdílu mezi srovnávanými postupy ve zkoumané populaci (apriorní pravděpodobnost podle Bayese – **prior probability**), síla studie – pravděpodobnost zjištění skutečného účinku zkoumaného postupu, přesnost použitých metod zkoumání a vypočtená hodnota p (tabulka 1). Síla použitých metod souvisí s falešně pozitivními výsledky, protože když má studie nižší pravděpodobnost zjištění skutečného účinku, vyšší podíl pozitivních výsledků bude falešně pozitivní (9).

Špatná interpretace významu hodnoty p spočívá v tom, že při $p = 0,05$, není poměr ve prospěch skutečného efektu 19 : 1 (95 % správně pozitivní : 5 % falešně pozitivní), jak by se na první pohled mohlo zdát, ale v závislosti na předchozí pravděpodobnosti reálného efektu a spolehlivosti použitých metod v nejlepším případě jen asi 3 : 1 (75 % správně pozitivní : 25 % falešně pozitivní). To je mnohem slabší důkaz než 19 : 1, který by mohl být nesprávně odvozen z hodnoty p . Pokud by mělo být omezeno riziko falešně pozitivního výsledku na 5 %, musel by být apriorní předpoklad 87 %, že existuje skutečný účinek zkoumaného postupu již před provedením experimentu. Pokud by vyšlo $p = 0,001$ v dobře provedené studii, je poměr pravděpodobnosti 100 : 1, že existuje skutečný efekt zkoumaného postupu. Riziko falešně pozitivního výsledku by bylo 8 %, pokud by předchozí pravděpodobnost skutečného účinku byla pouze 0,1. V tomto případě pro dosažení rizika falešné positivity 5 % by se muselo $p = 0,00045$.

Z tohoto hlediska je doporučováno nepoužívat výrazy «významný» a «nevýznamný». Spíše by hodnoty p měly být doplněny o specifikaci

předchozí pravděpodobnosti, která by byla potřebná k získání přijatelného např. 5% rizika falešně pozitivního výsledku. Může být také užitečné specifikovat minimální falešně pozitivní riziko spojené s pozorovanou hodnotou p . Praxe označení výsledku $p < 0,05$ jako „statisticky významného“ přispívá k nedostatečné reprodukovatelnosti v některých oblastech vědy. Je velmi pochybné vyvozovat závěry nebo dokonce uplatňovat praktické výstupy ze studií s nálezem „významného rozdílu“ v důsledku zamítnutí nulové hypotézy při $p < 0,05$. Řešením je replikace stejných výsledků v dalším výzkumu (8).

Na druhé straně je nutné uvést i názory, že nelze používat čisté subjektivní bayesovské apriorní pravděpodobnosti správného výsledku (10, 11), v tomto směru je však nutné ve statistické analýze vycházet ze všech dostupných (i negativních) předchozích relevantních výsledků (pilotních studií, replikovaných výzkumů).

Jak se tedy máme se znalostí výše uvedených omezení dívat na statistické využití hodnoty p ? Řada významných statistiků nesehlasí s požadavky některých časopisů, že hodnoty p nemají být vůbec používány a uváděny, ale doporučila pro zamítnutí nulové hypotézy v závislosti na dalších okolnostech rovnou stanovit limit p na 0,01 a nižší (12), jako celkem jednoznačný požadavek pro zavádění výsledků výzkumu do praxe je dále doporučována dříve podceňovaná replikace výsledků, která zvyšuje jejich validitu.

Závěrem

1. **Hodnota p je důležitá.** Čím nižší je hodnota p , tím nižší je chybovost, hodnota p blízko 0,05 má chybovost 25–50 %. Hodnota $p = 0,0027$ odpovídá chybovosti 4,5 %, což se blíží míře, kterou mnozí mylně připisují hodnotě $p = 0,05$. Nižší hodnota p znamená silnější důkaz pro odmítnutí nulové hypotézy. Hodnota p poblíž 0,05 naznačuje jen to, že výsledek stojí za další pohled, ale teprve, když se dostane k 0,001, bude šance na přijatelnou míru falešné positivity.

2. **Replikace výsledků.** Původní Fischerova myšlenka byla, že hodnoty p jsou jen jednou částí procesu, který zahrnuje experimentování, statistickou analýzu a replikaci pro validní vědecké závěry. Je nepravděpodobné, že by jediná studie, zejména pokud se hodnota p blíží 0,05, snížila míru falešné positivity na přijatelnou úroveň. Opakované studie vedou k dostatečně nízké chybovosti. Při hodnocení replikačních studií je však nezbytné zahrnout jak významné, tak nevýznamné studie, a ne vybrat pouze významné studie. Přesvědčivé prokázání hypotézy jedinou studií je velmi nepravděpodobné.

3. **Na velikosti efektu záleží.** Je třeba se zaměřit nejen na hodnotu p , ale pozornost musí být věnována i velikosti efektu. Jen proto, že je efekt statisticky významný, nemusí být smysluplný v reálném světě. Hodnota p ani neukazuje přesnost odhadované velikosti efektu.

4. **Na alternativní hypotéze záleží.** Máme tendenci si myslet, že ekvivalentní hodnoty p z různých studií poskytují stejnou podporu pro alternativní hypotézu. Výzkum ukazuje, že věrohodnost alternativní hypotézy výrazně ovlivňuje míru falešně pozitivních výsledků. Například vysoce věrohodná alternativní hypotéza a hodnota $p = 0,05$ jsou spojeny s chybovostí p kolem 12 %, zatímco nepravděpodobná alternativa je spojena s mírou chybovosti kolem 76 %.

5. **Odborná znalost je zásadní.** Vědecký úsudek o věrohodnosti hypotéz, výsledcích podobných studií, navrhovaných mechanismech, správném experimentálním designu jsou nepostradatelné pro transformaci statistiky z čísel na smysluplná a důvěryhodná zjištění (13, 14, 15).

*Podpořeno projektem Univerzity Karlovy
COOPERATIO.*

LITERATURA

1. Wasserstein RL, Lazar NA. The American Statistical Association statement on p-values: context, process, and purpose. *Am Stat.* 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108 [Epub ahead of print]. Available from: 10.1080/00031305.2016.1154108.

2. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine Statistical Significance. *Nat Hum Behav.* 2018;2(1):6-10. doi: 10.1038/s41562-017-0189-z.

3. Bayes T. An essay towards solving a problem in the doctrine

of chances. *Philosophical Transactions of the Royal Society of London.* 1763;53:370-418.

4. Branimir K. Hackenberger. Bayes or not Bayes, is this the question? *Croat Med J.* 2019;60:50-2. https://doi.org/10.3325/cmj.

2019. 60. 50.

5. Hebak P. Srovnání klasické a bayesovské pravděpodobnosti a statistiky (I.) Acta Oeconomica Pragensia. 2012;20(1):69-87. ISSN 0572-3043.

6. Vojáček J. Statistické hodnocení významnosti testů používaných k detekci ischemie myokardu. In: Vojáček J, Bultas J. Němá ischemie myokardu. Grada Praha. 1994, 71 s.

7. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypothesis. The American Statistician. 2001;55:62-71.

8. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. R. Soc. open sci. 2017;4:171085. [http://](http://dx.doi.org/10.1098/rsos.171085)

dx.doi.org/10.1098/rsos.171085.

9. Ruiz-Ruano García AM, López Puga J. Deciding on Null Hypotheses using P-values or Bayesian alternatives: A simulation study. Psicothema. 2018;30(1):110-115. doi: 10.7334/psicothema2017.308. PMID: 29363479.

10. Johnson VE. Uniformly most powerful Bayesian tests. Ann. Stat. 2013;41:1716-1741. (doi:10.1214/13-AOS1123)

11. Arandjelović O. A more principled use of the p-value? Not so fast: a critique of Colquhoun's argument. R Soc Open Sci. 2019;6(5):181519. doi: 10.1098/rsos.181519. eCollection 2019 May.

12. Johnson VE. 2013 Revised standards for statistical eviden-

ce. Proc. Natl Acad. Sci. USA 2013;110(48):19313-7. doi: 10.1073/pnas.1313476110. Epub 2013 Nov 11., PMID: 24218581 PMCID: PMC3845140.

13. <https://blog.minitab.com/en/author/minitab-blog-editor>, <https://blog.minitab.com/en/adventures-in-statistics-2/five-guidelines-for-using-p-values>.

14. Goodman SN. p-Values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am. J. Epidemiol. 1993;137:485-496. doi:10.1093/oxford-journals.aje.a116700.

15. Základní pojmy z lékařské statistiky. In: Vojáček J, Kettner J. Klinická kardiologie 5. vydání Maxdorf Praha 2022. 1000 s.