SROVNÁNÍ KLASICKÉ A BAYESOVSKÉ PRAVDĚPODOBNOСТИ A STATISTIKY (3.)

Petr Hebák*

(Dokončení z čísla 2/2012 AOP.)

5. Intervalové odhady

Vyjděme opět z toho, že v *klasickém pojetí* parametr θ je neznámá konstanta. Proti tomu krajní body intervalu *spolehlivosti* (D, H) jsou *před provedením výběru* náhodné veličiny, zatímco *po provedení výběru* jsou (d, h) vypočtené hodnoty těchto náhodných veličin. Avšak jakmile konkrétní výběr byl proveden, už nic náhodného není a vypočítaný interval buď obsahuje neznámou hodnotu parametru či nikoli, jenže my nevíme, který z těchto dvou případů nastal. Na tento interval se už tedy dále nemůžeme dívat jako na náhodný. Správná interpretace klasického intervalu spolehlivosti proto je, že $(1 - \alpha)100\%$ náhodných intervalů ze všech možných, vypočítaných tímto způsobem, *bude* obsahovat skutečnou hodnotu neznámého parametru π . V tomto smyslu máme tedy $(1 - \alpha)100\%$ stupeň důvěry, že zrovna ten jeden náhodně pořízený náhodný výběr vedl k vytvoření intervalu, který obsahuje hodnotu θ . Po provedení výběru pojmy pravděpodobnost či spolehlivost v klasickém pojetí nepřinášejí nic nového. Bayesovci proto říkají, že činit pravděpodobnostní úsudky na základě takto pojímaných intervalů spolehlivosti představuje chybnou a zavádějící interpretaci. Konstrukce klasického intervalu spolehlivosti je založená na *výběrovém rozdělení* použité statistiky, ze kterého je teoreticky možné vidět, jak se hodnoty této statistiky mění (liší) ve všech možných výběrech. Z toho však vyplývá, že pravděpodobnosti s tím souvisící jsou určeny *dříve*, než máme data, a tedy nezávisí na konkrétním posuzovaném výběru.

Proti tomu bayesovské pravděpodobnosti vycházejí z posteriorního rozdělení a vedou k bayesovskému intervalu spolehlivosti, který je bayesovci označován za *credible* interval. Bayesovský stupeň důvěry v tento interval má přímou pravděpodobnostní interpretaci (ve smyslu bayesovského *degree of belief*), který je přímo podmíněný napozorovanými daty, což je pro výzkumníka velmi užitečné a přímočaré. Při tomto pojetí se nemusíme zajímat o skutečnosti, které mohly nastat, ale nenastaly. Bayesovský přístup sumarizuje informaci získanou z konkrétních dat, zatímco klasický přístup je do jisté míry hypotetickou úvahou, vycházející z výběrového rozdělení, neboli z pravděpodobnostního chování zvolené *statistiky* ve všech možných (neprovedených) výběrech.

* Univerzita Hradec Králové, Fakulta informatiky a managementu; Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky (hebak@centrum.cz).

Klasická (četnostní) představa nebo teorie bývá často (nesprávně) shrnuta do výroku o existenci $(1 - \alpha)100\%$ spolehlivosti, že *skutečná* hodnota parametru se nachází v intervalu od d do h . V této souvislosti Samaniego (2010, str. 59) připomíná článek Neymana (1938), ve kterém uvedl, že *interpretace klasického intervalu spolehlivosti neobsahuje žádný pravděpodobnostní úsudek o neznámém parametru*. Podle Neymana interpretace intervalu spolehlivosti by měla vycházet z pojmů souvisejících s relativní četností výsledků v *opakovaných výběrech*, ale k tomu Samaniego poznamenává, že je neudržitelná a neobhajitelná interpretace, která se vůbec neopírá o data z pořízeného výběru a vychází výhradně ze situace před jeho provedením.

Pro bayesovce je výklad teorie intervalového odhadu mnohem jednodušší, protože se vyhýbá možným konfliktům s napozorovanými daty. Posteriorní rozdělení parametru θ je základem všech prováděných úsudků o θ , a tedy i jediným potřebným nástrojem pro konstrukci bayesovského intervalu spolehlivosti. Bayesovskému protějšku klasického intervalu spolehlivosti pro parametr θ se též říká *credibility interval* pro θ . Jednoznačně přijímaný překlad tohoto pojmu do češtiny asi zatím neexistuje. Pojem *credibility* je ve slovnících spojován s uvěřitelností, důvěryhodností, ale i se spolehlivostí, takže není úplně jasné, který z nich (či jiný) nejlépe vyjadřuje obsahový význam *bayesovského intervalu spolehlivosti*. Každopádně však výše uvedená chybně používaná interpretace klasického intervalu spolehlivosti je naopak správnou interpretací bayesovského kredibilního intervalu, jehož konstrukce výhradně vychází z posteriorního rozdělení parametru θ . Samaniego (2010, str. 60) shrnuje, že každý interval $(d, h) = (\theta_d, \theta_h)$, pro který

$$P(\theta_d \leq \theta \leq \theta_h) = \int_{\theta_d}^{\theta_h} g(\theta|y) d\theta = 1 - \alpha, \quad (29)$$

je $(1 - \alpha)100\%$ bayesovským intervalem spolehlivosti odhadovaného parametru θ . Problém je v tom, že určení mezi intervalu θ_d a θ_h není jednoznačné. Interval s vyhovujícími vlastnostmi je znám jako HPD (*highest posteriori density*) interval, který pro jednovrcholová rozdělení je nejkratším, neboli má nejmenší rozdíl $\theta_h - \theta_d$. Nejčastěji používaným je proto *nejkratší* bayesovský interval, pro který meze θ_d a θ_h jsou určeny tak, aby vyhovovaly vztahu

$$\int_{-\infty}^{\theta_d} g(\theta|y) d\theta = \alpha / 2 = \int_{\theta_h}^{\infty} g(\theta|y) d\theta. \quad (30)$$

Bayesovský interval spolehlivosti reprezentuje posteriorní úsudek o intervalu, který se zvolenou pravděpodobností obsahuje neznámý parametr θ .

Výpočet kvantilů posteriorního rozdělení může být proveden numericky řešením integrálu, což specializované pakety (např. *Minitab*) nabízejí, anebo v některých případech (stejně jako při klasickém přístupu) při *dostatečně* velkém rozsahu výběru pomocí kvantilů normovaného normálního rozdělení. Většinou jsou bayesovské intervaly spolehlivosti *kratší* než klasické intervaly, a to i při vágní apriorní informaci, přičemž rozdíly jsou tím větší, čím lepší je apriorní informace. Bolstad (2004, str. 142 a 183) nabízí velmi zajímavé a přesvědčivé příklady, které při odhadech parametrů

binomického a normálního rozdělení dokumentují přednosti bayesovských odhadů ve srovnání s klasickými.

6. Testování hypotéz

Tradiční způsob testování statistických hypotéz je ze všech statistických úsudků ten nejvíce kritizovaný, a to nikoli jen bayesovci. Než si ukážeme na sporná až chybná místa klasického přístupu, znovu se nejdříve trochu podívejme do historie. *Press* (2003) krátce shrnuje vývoj *teorie* testování hypotéz od roku 1892 do roku 1959 následujícím způsobem.

Karl Pearson (1892) se snažil přispět k vytvoření formální teorie pro testování hypotéz a líbila se mu možnost zapojit do svého snažení i Bayesův vzorec. Neuměl si však vytvořit potřebná apriorní rozdělení a netušil, že uvažování aspoň jedné alternativní hypotézy by výrazně zjednodušilo situaci. Podobně uvažoval i *Student* (*Gosset*) (1908), když se zabýval jím vytvořeným *t*-testem o střední hodnotě normálního rozdělení a testem o korelačním koeficientu. Byl by dal přednost využití Bayesova vzorce, ale stejně jako *K. Pearson* a někteří další statistici tohoto období nevěděl, jak určit vhodné apriorní rozdělení. *Fisher* (1925) sice navrhl formální postup testování hypotéz, ale rovněž ještě neuvažoval existenci alternativní hypotézy. Výhradně se zaměřil na možnost zamítnout hypotézu v případě přesvědčení o její neplatnosti.

Zásadní rozšíření a změnu postoje v oblasti testování hypotéz způsobily články *Neymana* a *Egona Pearsona* z let 1928 až 1933. Ucelená a dodnes velmi známá a často citovaná je jejich teorie (1933), protože obsahuje do té doby řadu zcela neznámých pojmů, jako je hladina významnosti, síla testu, chyby prvního a druhého druhu a další. Avšak (rozumím-li dobře poznámkám v literatuře) *Fisher* v té době zastával stejné názory, jaké obhajoval známý a vlivný filozof *Karl Popper* (1935 a 1959), a přistupoval proto k této problematice jinak než *Neyman* s *Pearsonem*. *Popper* i *Fisher* vycházeli z přesvědčení, že je třeba najít postup, který by umožňoval na základě dat prokázat, že určitá hypotéza (teorie) je chybná nebo dokonce (podle jejich slov) *neobhajitelná až podvodná*. Považovali testované hypotézy za *uměle nastrčené*; jak jinak rozumět tomu, že jim říkali *strawman hypothesis*? Jejich cílem bylo na základě dat prokázat, že takové hypotézy jsou neplatné, a tudíž teorie, které s nimi souvisí, jsou chybné.

Všechny v té době postupně vznikající teorie testování statistických hypotéz naprosto vyloučily možnost přiřadit pravděpodobnosti ověřovaným hypotézám, protože *Fisher* (nesporně největší autorita a nejvlivnější statistik dvacátého století) byl zásadním odpůrcem čehokoli subjektivního a bayesovský přístup zcela zamíтал. Není tedy divné, že odmítl považovat testované hypotézy za náhodné veličiny a přiřazovat jim pravděpodobnosti. *Wald* (1939) navrhl rozhodovací teorii, do které zahrnul i testování hypotéz, čímž se sice značně přiblížil k bayesovskému pojetí, ale jeho četnostní vnímání pravděpodobností mu asi znemožnilo hlubší propojení klasické a bayesovské školy. *Lehmann* (1959) shrnul klasický přístup *Fisherův*, *Neymanův* – *Pearsonův* i *Waldův* ve své knize věnované testování statistických hypotéz, ale stejně jako jeho předchůdci se držel četnostního pojetí pravděpodobnosti a přístupu založeného na opakovaných výběrech a výběrovém rozdělení. *Fisher* a s ním i mnozí jeho následovníci považovali test významnosti za příležitost soustředit datové *důkazy*

proti testované hypotéze, zatímco Neyman s E. Pearsonem navrhli metodu, jejímž cílem je rozhodnutí ve prospěch jedné ze soupeřících hypotéz na základě pravděpodobností jejich chybného přijetí či odmítnutí. Zajímavostí je, že dnes přístupy Fishera a Neymana – Pearsona při hodnocení téměř splývají, navzdory jejich odlišným cílům a nestejně interpretaci.

Principy a podstata klasického teorie i doporučeného postupu testování statistických hypotéz je zcela běžnou součástí většiny statisticky orientovaných statistických učebních textů. Rovněž v *Hebák et al.* (2000) může případný zájemce o výklad a klasifikaci nabízených typů testů najít potřebné informace. Tradiční přístup a postupy z něj vyplývající jsou většinově pro uživatele jedinou dostupnou informací, takže není překvapením, že jej nejspíše považují za jediný možný a správný. *Mechanické* interpretace méně zasvěcených jsme zmínili hned v úvodu tohoto příspěvku, ale běžné (svými důsledky horší) jsou i *optimistické* interpretace o *prokázání kauzálního efektu na základě výsledků provedených testů*.

Kritika teorie testování tradičního pojetí nepřichází jen od bayesovců a zvláště v posledních letech velice sílí. *Wang* (1993) v už zmíněné publikaci (*Sense and Nonsense of Statistical Inference*) na řadě příkladů a odkazů skvěle dokresluje přednosti a slabiny (tehdy diskutovaných) statistických úsudků, jakož i s tím souvisejících statistických metod. Nikoli náhodou hned první kapitulu pojmenoval *Fads and Fallacies in Hypothesis Testing* a jmenuje řadu důvodů, proč právě v této oblasti výzkumníci (zvláště ve společenských vědách) nejčastěji chybují. Cituje (kromě desítek dalších) ostrou kritiku *Breimana* (1985), debatu *DeGroota* s *Lehmannem* ve *Statistical Science* (1986) a další kriticky zaměřené příspěvky, aby ukázal, že situace je opravdu vážná a vyžaduje **zásadní zvýšení statistických znalostí osob používajících statistiku především ve společenskovědních oborech**. Zmiňuje tragické omyly vyplývající z používání (jen pro ilustraci zmíněných chí-kvadrát testů pro data velkého rozsahu), kdy doslovně říká ... *This leads to the rejection of any reasonable null hypothesis*. Označuje jako zcela zavádějící (a vysvětluje proč) zvyklost používat v této souvislosti pojem *vědecká metoda* a na mnoha příkladech dokresluje chybná použití či nesprávné interpretace testovacího postupu. Připomíná knihu *Lehmann* (1985): *The Neyman – Pearson Theory after 50 Years* a zdůrazňuje, že *teorie Neymana – Pearsona je důležitá, ale naneštěstí je to jen teorie. Není to metoda a přinejmenším to není metoda, která by mohla být nahrazena lepší metodou*. Mnozí autoři, např. *Morrison – Henkel* (1970), ale i mnoho dalších rovněž upozorňují na rostoucí počet připomínek a debat ke sporné užitečnosti statistických testů v oblasti společenských věd.

Přes všechny oprávněné kritiky a námitky považuji za přehnaný přístup ekonomky prof. *D. McCloskey*, která v roce 2010 měla na VŠE veřejnou přednášku pod názvem *Statistical significance is worse than useless*. Ve své argumentaci často zmiňovala knihu *Ziliak, S. T. – McCloskey, D. N.*, která pod názvem *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives* vyšla v roce 2007. Některé její argumenty proti testování statistických hypotéz byly sice oprávněné, ale pro méně informované o této problematice přednáška působila spíše jen *destruktivním* dojmem (proti jakékoli analýze dat), což doufám ani přednášející nezamýšlela. Samotnou knihu jsem nečetl, takže se možná mýlím, ale každopádně mne její argumenty a *důkazy*

připadaly povrchní. Správně sice odmítala způsob, kterým někteří ekonomové *prokazují* kvalitu svých teorií a modelů, ale v tomto konkrétním případě se její kritika neměla týkat jen testů významnosti, ale též *nedostatečné ekonomické teorie, málo přesvědčivých a nezdůvodněných modelů, nízké kvality a snížené vypovídací schopnosti některých agregovaných dat*, jakož i *způsobu výběru vysvětlujících proměnných v ekonometricky zaměřených úlohách*. Přesvědčivě nepůsobila ani (částečně oprávněná, ale *přílišná*) chvála *Studenta (W. S. Gosseta)* ve vztahu k *přehnané* (někdy asi neoprávněné a nezdůvodněné) kritice *R. A. Fishera*.

Námítky bayesovců vůči úsudkům prováděným na základě výběrových rozdělení už jsem popsal v souvislosti s bodovými a intervalovými odhady. Zmínit je však třeba ještě několik dalších důvodů zásadně odlišného přístupu obou porovnávaných statistických škol k problematice testování statistických hypotéz. Bayesovci upozorňují, že se o testování hypotéz zajímají méně než o jiné statistické úsudky. Důvodem je skutečnost, že posteriorní rozdělení a jeho charakteristiky obsahují o neznámém parametru (neznámých parametrech) vše potřebné, takže ve většině úloh není proč a kam pokračovat. *Press (2003)* však připouští, že třeba pro rozhodnutí, zda data potvrzují určitou teorii, pro posouzení rozdílů mezi několika teoriemi anebo obecně v experimentálních úlohách může být testování hypotéz užitečným pomocníkem. Bayesovský přístup má nespornou výhodu, že apriorní a posteriorní postoj k posuzovaným hypotézám má charakter rozhodovací pravděpodobnostní úlohy, což usnadňuje srovnání alternativních tvrzení či domněnek. Bayesovci kritizují i *libovůli* při volbě hladiny významnosti, která ve svém důsledku je jedinou rozhodovací *dělicí čarou*.

Při bayesovském přístupu je zdánlivě neřešitelným problémem skutečnost, že nejčastější hypotéza $H_0: \theta = c$, kde c je předpokládaná hodnota parametru, je pro spojitý parametr θ *netestovatelná*, jelikož její pravděpodobnost **pro všechny možné hodnoty tohoto parametru** je nulová. Zapišeme-li však testovanou hypotézu jako $H_0: \theta < c$ a **jednostrannou** alternativní hypotézu jako $H_1: \theta \geq c$, pak při znalosti $g(\theta|y)$ (v souladu s *Jackmanem (2010, str. 33)* můžeme posteriorní pravděpodobnosti hypotéz definovat následujícím způsobem:

$$P(H_0|y) = P(\theta < c|y) = \int_{-\infty}^c g(\theta|y)d\theta \text{ a } P(H_1|y) = P(\theta \geq c|y) = \int_c^{\infty} g(\theta|y)d\theta. \quad (31)$$

Je-li posteriorní pravděpodobnost $P(H_0|y)$ menší než zvolená hladina významnosti α , rozhodneme ve prospěch H_0 , a naopak. Opakují se, ale opět s konjugovaným apriorním rozdělením není žádný větší výpočetní problém, zatímco v jiných případech je vhodné využít moderní výpočetní postupy využívající Markovské řetězce a metodu Monte Carlo. *Jackman* dále zdůrazňuje, že klasická statistika nemá podobný nástroj jako je posteriorní rozdělení, i když takto (bayesovsky) leckdy bývají p -hodnoty chybně interpretovány. Ve skutečnosti p -hodnoty říkají, jak často můžeme při provedení všech možných výběrů a při platnosti hypotézy H_0 *získat vypočítanou hodnotu testového kritéria anebo hodnoty, které jsou pro testovanou hypotézu ještě více nepříznivé*. Taková interpretace působí velice neohrabaně až podezřele, takže je zcela pochopitelná snaha p -hodnoty jednoduše označit za hladinu významnosti, se kterou jsou porovnávány, což také (označením *sig* ve zkráceném smyslu *významnost*) mnohé výstupy statistických

paketů obsahují. *Berger – Selke* (1987) a *Casella – Berger* (1987) upozorňují ještě na skutečnost, že p -hodnoty vedou k častějšímu zamítání testované hypotézy než posteriorní rozdělení, čehož důvodem je skutečnost, že p -hodnoty (stejně jako hodnoty testových kritérií) závisí na rozsahu výběru n .

Bayesovci musí při stejné hypotéze $H_0: \theta = c$, ale dvoustranné alternativní hypotéze $H_1: \theta \neq c$, postupovat jinak. Bylo už řečeno, že při spojitém testovaném parametru θ nedává smysl pravděpodobnost $P(H_0)$, tedy pravděpodobnost rovnosti $\theta = c$, jelikož pravděpodobnost konkrétní hodnoty θ je nulová a nelze tedy určit posteriorní pravděpodobnost $P(H_0|y)$. Poměrně jednoduchým východiskem je v minulém oddíle popsáný $(1 - \alpha)100\%$ bayesovský dvoustranný interval spolehlivosti (*credible* nebo *credibility interval*). Při použití tohoto intervalu se testovaná hypotéza H_0 zamítá v případě, že hodnota c se v tomto intervalu nenachází, a naopak nezamítá, jestliže c se v tomto intervalu nachází. Pro ilustraci rozdílů mezi klasickým a bayesovským přístupem k pravděpodobnostním a statistickým úsudkům jsem zvolil jednoduchý příklad dvanácti hodů napínáčkem v relaci ke 12 hodům mincí podle už zmíněného vynikajícího výkladu *D. W. Lindleyho* z roku 1976. Převzaté zadání a řešení úlohy jsem pro potřeby tohoto příspěvku trochu rozšířil a doplnil o některé výpočty.

Příklad 4: Úsudky o π při dvanácti nezaújatých hodech napínáčkem a mincí

Máme k dispozici výsledky $n = 12$ hodů běžným napínáčkem. Při těchto hodech napínáček padl špičkou nahoru (U) při $y = 9$ hodech a špičkou dolů (D) při $n - y = 3$ hodech. Ve výkladu k příkladu Lindley vysvětlil, že házel pokud možno stále stejně a uvedl, že pořadí devíti U a tří D neposkytuje žádnou zajímavou informaci ($U, U, U, D, U, D, U, U, U, U, U, D$). Zdůraznil však, že předem nebyl určen rozsah výběru ($n = 12$) ani zastavovací pravidlo, takže se neví, zda počet hodů předem určil, anebo zda se rozhodl házet dokud napínáček nepadne celkem 3x špičkou dolů. Pro klasického statistika má odpověď na tuto otázku zcela zásadní význam, zatímco pro bayesovce je to lhostejné. Porovnejme klasické a bayesovské odpovědi na tyto otázky:

1. Jaká je pravděpodobnost, že v následujícím 13. hodu padne napínáček špičkou nahoru?
2. Jaký je bodový odhad $\pi = P(U)$ v každém hodu?
3. Jaký je 95% dvoustranný interval spolehlivosti pro π ?
4. Jaký je výsledek testu $H_0: \pi = 0,5$ proti $H_1: \pi \neq 0,5$ při zvolené hladině významnosti $\alpha = 0,05$?
5. Jak by se výsledky změnily, kdyby se místo napínáčku házelo mincí opět s devíti H a třemi O ?

Řešení

Ad 1 a 2) Při klasickém přístupu nelze jednoznačně odpovědět, protože je-li počet hodů pevně předem zvolen, je nejlepším odhadem π výběrová relativní četnost $p = 9/12 = 0,75$, zatímco házelo-li se až do tří D , je (podle Haldana) nezkresleným odhadem π podíl $(y - 1)/(n - 1) = 8/11 = 0,727$. Při vágní informaci o parametrech

a, b beta rozdělení zvolíme $a = b = 1$, takže posteriorní střední hodnota beta rozdělení s parametry $a = 9 + 1, b = 3 + 1$ má střední hodnotu $10/14 = 0,714$. Tyto hodnoty (při klasickém pojetí jsou zároveň odhady pravděpodobnosti, že v následujícím hodu padne napínáček špičkou nahoru, ale zároveň i odhady parametru π pro každý hod tímto napínáčkem po 12 hodech s uvedeným výsledkem $9 \times U$ a $3 \times D$. Pro klasické úvahy se 12 provedených hodů považuje za 12 nezávislých pokusů, zatímco pro bayesovský přístup stačí, že tyto pokusy jsou zaměnitelné, neboli každé pořadí s devíti U a třemi D má stejnou pravděpodobnost. Požadavek zaměnitelnosti pokusů je sám o sobě slabší než Bernoulliho předpoklad nezávislosti.

Ad 3) Pro velké n lze využít aproximaci binomického i beta rozdělení normálním rozdělením, ale pro $n = 12$ to vhodné není. Použijeme vzorce, které uvádí Kahounová (2000), ve kterých se využívají kvantily rozdělení F s parametry $2(y + 1)$ a $2(n - y)$ pro horní mez a s parametry $2(n - y + 1)$ a $2y$ pro dolní mez a pro $\alpha = 0,05$. Pak po dosazení dostáváme

$$\pi_h = \frac{f_{1-\alpha/2}(y+1)}{n - y + f_{1-\alpha/2}(y+1)} = \frac{5,17 \cdot 10}{3 + 5,17 \cdot 10} = \mathbf{0,945}, \quad (32)$$

$$\pi_d = \frac{y}{f_{1-\alpha/2}(n - y + 1) + y} = \frac{9}{3,01 \cdot (12 - 9 + 1) + 9} = \mathbf{0,428}, \quad (33)$$

takže podle klasického přístupu je 95% interval spolehlivosti pro π od **0,428** do **0,945**. Bayesovský *credible* interval spolehlivosti vyžaduje numerické řešení rovnice pro hledané meze intervalu

$$\int_{\pi_{0,025}}^{\pi_{0,975}} \frac{\Gamma(14)}{\Gamma(10)\Gamma(4)} \pi^{10}(1 - \pi)^4 d\pi = 0,95. \quad (34)$$

Statistický paket MINITAB nabízí bayesovská makra, pomocí kterých dostáváme následující výstup porovnávající beta rozdělení s parametry $a = b = 1$ a posteriorní rozdělení po dosazení hodnot $y = 9$ a $n = 12$.

Výstup 1

Posteriorní parametry beta rozdělení

a1 = 10

b1 = 4

Posteriorní střední hodnota (m1), rozptyl (v1) a směrodatná odchylka (s1)

m1 = 0,714

v1 = 0,014

s1 = 0,117

Kvantily posteriorního rozdělení

y1	kvantil
0,005	0,379
0,010	0,412
0,025	0,462
0,050	0,505
0,500	0,725
0,950	0,887
0,975	0,909
0,990	0,931
0,995	0,943

Hledaný 95% interval spolehlivosti pro π je od **0,462** do **0,909** a srovnání apriorního a posteriorního rozdělení je vidět na obrázku 2.

Obrázek 2**Apriorní a posteriorní rozdělení π k příkladu s napínáčky**

ad 4) Klasický test hypotézy $H_0: \pi = 0,5$ proti $H_1: \pi \neq 0,5$ na 5% hladině významnosti vyžaduje výpočet p -hodnoty. Výsledky, které ještě výrazněji zpochybňují představu o symetrickém napínáčku, jsou 10 : 2; 11 : 1 a 12 : 0 za předpokladu, že počet hodů $n = 12$ je předem daný. Včetně získaného 9 : 3 je výsledek $299/4\ 096 = \mathbf{0,073}$ v tomto případě statisticky nevýznamný a nedovoluje zamítnout H_0 . Při Haldanově experimentu jsou horší jen výsledky 10 : 3; 11 : 3; 12 : 3 atd. Pravděpodobnost těchto výsledků (včetně získaného poměru 9 : 3) je $134/4\ 096 = 0,033$, což na 5% hladině významnosti dovoluje zamítnout hypotézu $H_0: \pi = 0,5$. Pro bayesovský test stačilo zjistit, že hodnota $\pi = 0,5$ se nachází ve vypočteném bayesovském *credible* intervalu od **0,462** do **0,909**, takže hypotéza H_0 se nezamítá.

Ad 5) **Klasické výpočty pro mince jsou naprosto stejné jako pro napínáčky.** Zcela jinak je tomu však při bayesovském přístupu, při kterém využijeme beta rozdělení s parametry $a = b = 50$, protože naše přesvědčení (*degree of belief*) o symetrické minci je daleko silnější než při neurčitém postoji k napínáčku. Potom pravděpodobnost, že při 13. hodu padne mince *hlavou vzhůru* je dána posteriorní střední hodnotou $59/112 = 0,527$, protože u mince poměr 9 : 3 příliš nezměnil výchozí přesvědčení o symetrické minci. Jistě není nerozumné, že se náš závěr o napínáčcích výrazně liší od závěru o mincích. Posteriorní rozdělení je v tomto případě špičatější a i interval spolehlivosti od **0,434** do **0,618** a test hypotézy (též nezamítající testovanou hypotézu $\pi = 0,5$) tuto skutečnost mnohem více odráží než v případě napínáčků. Srovnání je velice zřejmé při pohledu na obrázek 3 a na následující výstup.

Výstup 2

Posteriorní parametry beta rozdělení

a1 = 59

b1 = 53

Posteriorní střední hodnota, rozptyl a směrodatná odchylka

m1 = 0,527

v1 = 0,002

s1 = 0,047

Kvantily posteriorního rozdělení

y1	kvantil
0,025	0,434
0,975	0,618

Obrázek 3

Apriorní a posteriorní rozdělení π k příkladu s mincemi

Závěr

Předkládám veřejnosti výkladový, ale částečně i polemicky zaměřený příspěvek, kterým reaguji na požadavek, abych porovnal klasickou statistickou školu, opírající se o pravděpodobnosti vnímané jako *dlouhodobé* relativní četnosti, se subjektivisticky orientovanou bayesovskou statistickou školou. Zkusil jsem v tomto (zdánlivě dlouhém, ale vzhledem k rozsahu problematiky velice stručném) příspěvku aspoň orientačně seznámit širší odbornou veřejnost s některými myšlenkami bayesovského přístupu, vycházejícími z možnosti kombinace získaných dat s apriorními znalostmi či informacemi o studovaných skutečnostech. V současné době (téměř výhradně v zahraničí) existují desítky bayesovsky orientovaných monografií, jakož i velké množství článků, které pravidelně vycházejí v odborných časopisech. S tím kontrastuje skutečnost, že (pokud je mi známo) v České republice zatím nebyla v knižní podobě uveřejněna ani jedna bayesovská monografie zaměřená na pedagogické a částečně i filozofické aspekty statistických úsudků (oblast informatiky úmyslně nezmiňuji) a neznám ani žádný článek nebo odbornou debatu zaměřenou na důsledky odlišností klasické a bayesovské statistické školy. Navíc u nás, na rozdíl od jiných zemí či univerzit, stále jednoznačně ve výzkumu, ve výuce i v aplikacích dominuje tradiční statistický přístup, patřící svojí výhradností (bez přehánění) spíše do třicátých až sedmdesátých let minulého století. Přestože se snažím ve svých volitelných předmětech vyučovaných na VŠE (*Rozhodování za rizika a nejistoty; Marketingová rozhodnutí a data; Subjektivní pravděpodobnost a bayesovská statistika*) aspoň trochu se dotknout bayesovského způsobu uvažování a vysvětlit studentům jeho přednosti, je to strašně málo. Nepodařilo se mi prosadit do výuky ani jeden povinný bayesovsky orientovaný předmět aspoň do doktorského studia na oboru statistika. Přednosti bayesovských pravděpodobnostních i statistických úsudků ve srovnání s klasickými jsou jasné a v zahraničí všeobecně uznávané, a to nejen na specializovaných oborech nebo univerzitách.

Předložený článek se proto pokouší o srovnání bayesovského a klasického (četnostního) vnímání pravděpodobnosti, dále o posouzení rozdílné kvality bodových a intervalových odhadů těchto dvou statistických škol a konečně i o představení rozdílného přístupu k testování statistických hypotéz. Už tak překročený původně vymezený prostor nebyl dostatečný pro naznačení aspoň podstaty moderních bayesovských výpočetních metod pomocí Markovských řetězců kombinovaných s metodou Monte Carlo. Přitom jejich prudký rozvoj na přelomu dvacátého a jednadvacátého století způsobil až revoluční změnu v zájmu teoretiků i uživatelů o bayesovské metody, používané téměř ve všech oblastech (jednorozměrné i vícerozměrné) statistiky a v aplikacích z různých vědních oborů. Článek neobsahuje teorii rizika ani teorii rozhodování, i když právě v těchto oblastech podíl bayesovského přístupu převládá nad jinými. Uvědomuji si, že i samotné srovnání statistických úsudků mohlo být provedeno mnohem důkladněji s hlubší statistickou argumentací.

Na druhé straně je velice zajímavé a pro klasicky orientované statistiky inspirující, že v posledních letech se objevují bayesovci *oprášené* starší směry (jako je *Empirical Bayes* nebo *Objective Bayes*) a různé jiné (jakoby bayesovsky orientované) přístupy. Částečně změněný postoj význačných bayesovských i klasických představitelů způsobil značné přiblížení aspoň v oblastech, ve kterých je to možné (za všechny

bayesovce jmenujme aspoň *Bradleyho Efrona*, jehož přínos pro statistiku je nezane-dbatelný). Bayesovská i klasická statistická škola mají svoje přednosti a nedostatky, takže jedině propojení všeho *lepšího* z každé z nich by mohlo být užitečné a pro budoucí rozvoj statistiky prospěšné. Význačný v tomto směru je dříve zmíněný článek, jehož autoři jsou *Bayarri – Berger*, který pod názvem *The Interplay of Bayesian and Frequentist Analysis* vyšel ve *Statistical Science* v roce 2004. Uznávaný bayesovec *J. O. Berger* k tomu v uvedeném článku na str. 59 říká (překlad PH), že *v souvislosti s očekávanou užitečností propojení klasické a bayesovské statistiky je třeba nechat stranou otázku, zda bayesovský či četnostní přístup je filozoficky (ne)korektní (i když je to potřebná otázka zásadní důležitosti), a rovněž je třeba (ted') neřešit velice závažnou otázku pedagogické korektnosti klasického či bayesovského způsobu výuky. Je třeba se nejdříve soustředit na méně kontroverzní otázky čistě metodologického rázu, protože ve dvou výše uvedených otázkách by zatím určitě ke shodě nedošlo a možná to dnes není ani nutné*. Druhá část tohoto článku, pojmenovaná jako *Inherently Joint Bayesian-Frequentist Situations*, naznačuje, co mají autoři na mysli a kam směřují.

Na velké datové soubory a na přístup označený jako *Empirical Bayes* je zaměřena aktuální monografie *Bradleyho Efrona* (2010), která i naznačuje Efronovu představu o budoucím vývoji statistiky. Možnosti užitečného propojení klasického a bayesovského přístupu nabízí (mnou několikrát v tomto příspěvku citovaná) kniha *Francisca J. Samaniego: A Comparison of the Bayesian and Frequentist Approaches to Estimation*, která vyšla v prestižním nakladatelství *Springer* rovněž v roce 2010. Celkově je možné shrnout, že souboje mezi klasickým a bayesovským přístupem přestávají být nesmiřitelné, přičemž se předpokládá, že všichni statistici aspoň částečně znají přednosti i slabiny obou škol. Hledají se styčné body, které by ve svém důsledku znamenaly posun v oblasti statistické teorie a poskytly lepší metody a postupy pro výzkumníky využívající statistické nástroje při analýze dat.

Literatura k 3. části

- BAYARRI, M. J.; BERGER, J. O. The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*. 2004, vol. 19, s. 58–80.
- BERGER, J. O.; SELKE, T. Testing a Point Null Hypothesis: the Irreconcilability of p-Values and Evidence. *Journal of the American Statistical Association*. 1987, vol. 82, s. 112–122.
- BLACKWELL, D. *Basic Statistics*. Los Angeles : Mc-Graw Hill, 1969.
- BOLSTAD, W. M. *Introduction to Bayesian Statistics*. Hoboken : John Wiley and Sons, 2004.
- BREIMAN, L. Nail Finders. Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer, Wadsworth, Belmont, California, 1985. Vol. 1, s. 201–212.
- CASELLA, G.; BERGER, L. R. Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association*, 1987, vol. 82, s. 106–135.
- EFRON, B. *Large Scale Inference – Empirical Bayes Methods for Estimation, Testing and Prediction*. Cambridge; New York : Cambridge University Press, 2010.
- FISHBURN, P. C. The Axioms of Subjective Probability (with Discussion). *Statistical Science*. 1986, vol. 1, s. 335–358.
- FISHER, R. A. *Statistical Methods for Research Workers*. 14. ed. Edinburgh; New York : Hafner, Oliver and Boyd, 1925.
- HÁTLE, J.; LIKEŠ, J. *Základy počtu pravděpodobnosti a matematické statistiky*. Praha : SNTL, Alfa, 1972.
- HAWKING, S. W. *A Brief History of Time*. New York : Bantam Books, 1988.
- HEBÁK, P. et al. *Praktikum k výuce matematické statistiky II: Testování hypotéz*. Praha : Vysoká škola ekonomická v Praze, 2000.
- JACKMAN, S. *Bayesian Analysis for the Social Sciences*. Chichester : John Wiley and Sons, 2010.
- KAHOUNOVÁ, J. *Praktikum k výuce matematické statistiky I: Odhady*. Praha : Vysoká škola ekonomická v Praze, 2000.
- LEHMANN, E. L. *Testing Statistical Hypotheses*. New York : John Wiley and Sons, 1959.
- LEHMANN, E. L. The Neyman Pearson Theory after 50 years. Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer, Wadsworth, Belmont, California, 1985, vol. 1, s. 1–14.
- LINDLEY, D. V.; PHILLIPS, L. D. Inference for a Bernoulli Process (a Bayesian View). *The American Statistician*. 1976, vol. 30, s. 112–119.
- MORRISON, D. E.; HENKEL, R. E. *The Significance Test Controversy*. Chicago : Aldine, 1970.
- NEYMAN, J.; PEARSON, E. S. On the Testing of Statistical Hypotheses in Relation to Probability A Priori. *Proc. of the Cambridge Phil. Soc.* 1933, vol. 29, s. 492–510.
- NEYMAN, J. Contribution to the Theory of Sampling Human Populations. *Journal of the American Statistical Association*. 1938, vol. 33, s. 101–116.
- PEARSON, K. *The Grammar of Science*. London : Adam and Charles Black, 1892.
- POPPER, K. *The Logic of Scientific Discovery*. London; New York : Hutchinson; Harper Torchbooks, 1935; 1959; 1968.
- PRESS, S. J. *Subjective and Objective Bayesian Statistics*. New York : John Wiley and Sons, 2003.
- STUDENT (William Sealy Gosset). Paper on the Student t- distribution. *Biometrika*. 1908, vol. 6, s. 1–25.
- WALD, A. Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *Ann. Math. Statistics*. 1939, vol. 10, s. 299–326.
- WANG, C. *Sense and Nonsense of Statistical Inference*. New York : Marcel Dekker, 1993.
- ZILIAK, S. T.; MCCLOSKEY, D. N. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Michigan : University Press, 2007.

A COMPARISON OF CLASSICAL AND BAYESIAN PROBABILITY AND STATISTICS (3)

Abstract: Statistics has been developing for almost 250 years – since the publication of an essay which included one theorem called Bayes' after the author. This whole period (since 1763 to this day) has been accompanied by a duel between the supporters of a subjective concept of probability and those who refuse everything but a purely objective concept of probability as well as statistics. While the 18th and 19th centuries accepted the importance of the subjective (let us say Bayesian) way of thinking for the development of probability and statistics without a problem, in the 20th century the classic (frequentist) way took over and has been dominant in teaching and textbooks to this day. Only in the second half of the 20th century did the situation begin to change slowly. Reasons for that are partly described in the present article, but arguments and simple examples supporting the Bayesian way in comparison with the classic one are clear and generally respected worldwide. Unsuspected new computing possibilities have caused an explosive development of Bayesian statistics, which has infiltrated almost all the areas of statistics and a number of other scientific fields. It is not possible to expect a retreat of the different philosophical or pedagogical positions of the fighting schools of thought (even though it is really needed), but the use of advantages of both the approaches is methodologically not only possible, but even expected. Part of the teaching of statistics must be prepared for these changes, but it has not been the case in the Czech Republic at all so far.

Keywords: subjective probability; frequentist statistics; classical and Bayesian approach and thinking; Bayes' theorem; point estimation; prior and posterior distribution; Bayesian Credible interval; hypothesis testing

JEL Classification: E21, C82