Petr Emanovský

Statistické testování hypotéz třikrát jinak

Učitel matematiky, Vol. 30 (2022), No. 1, 3–14

Persistent URL: `http://dml.cz/dmlcz/150380`

STATISTICKÉ TESTOVÁNÍ HYPOTÉZ TŘIKRÁT JINAK

Petr Emanovský[1]

Úvod

Statistické testování hypotéz představuje metodu, která je široce využívána v kvantitativním výzkumu v rámci empirických věd včetně pedagogiky a sociologie. Tato všeobecně známá procedura je běžně využívána nejen profesionálními výzkumníky, ale i studenty při zpracovávání kvalifikačních prací. Statistické testování hypotéz představuje jednu z typických forem statistického usuzování, při němž se uplatňuje statistická indukce. Tento princip hraje klíčovou roli při zkoumání hromadných jevů a jeho podstatou je srovnání vlastností vzorku „náhodně“ vybraného z populace s teoretickým pravděpodobnostním modelem. Na základě tohoto srovnání se usuzuje o vlastnostech celé populace, přičemž lze kvantifikovat riziko omylu. Při tomto přístupu je pravděpodobnost definována jako relativní četnost výskytu náhodného jevu. S touto metodou se lze podrobně seznámit v každé učebnici matematické statistiky, viz např. (Hendl, 2004; Chráska, 2007; Klementa et al., 1984; Komenda, 1994). Méně známá je historie vzniku a vývoje této metody. Myšlenky, na nichž je tato metoda založena, se objevují již počátkem 18. století v práci skotského matematika Johna Arbuthnota (Arbuthnot, 1710; Emanovský, 2021b). Další vývoj v této oblasti je spojen s řadou významných matematických myslitelů, jako byli např. Abraham de Moivre (1667–1754), Nicolaus

[1]Článek vznikl za podpory projektu Univerzity Palackého Olomouc „Algebraické a geometrické struktury“ IGA PrF 2021 030.

Bernoulli (1687–1759), Daniel Bernoulli (1700–1782), Pierre Simon Laplace (1749–1827), Karl Fridrich Gauss (1777–1855) a další (Fienberg, 1992). Významnou měrou k rozvoji této metody přispěl rovněž William Sealy Gosset (1876–1937), autor významného článku *The probable error of a mean* z roku 1908, v němž je popsána nová metoda vhodná pro statistické testování málo početných vzorků, dnes známá jako Studentův t-test (Student, 1908; Fisher, 1934; Box, 1981; Emanovský, 2021a). Článek zaujal zejména statistika Ronalda A. Fishera, který tuto metodu dále rozvinul.

Dnešní podoba statistického testování hypotéz je syntézou dvou fundamentálně odlišných statistických přístupů, které se objevily v první polovině 20. století. Na jedné straně zde stál významný anglický statistik R. A. Fisher, na druhé polský matematik J. Neyman spolu s anglickým statistikem E. Pearsonem.

Fisherovské testování

Ronald Aylmer Fisher (1890–1962) byl geniální anglický statistik, který je považován za hlavního zakladatele moderní matematické statistiky. Navrhl např. metodu maximální věrohodnosti, analýzu rozptylu, plánování experimentů a přispěl i k teorii testování hypotéz a teorii odhadu. Ve své významné práci *Statistical Methods for Research Workers* z roku 1934 shrnul své originální statistické metody a postupně se zasloužil o jejich široké uplatnění v biometrice, ekonomii, medicíně a dalších vědních oborech. Za své vědecké výsledky byl dokonce jmenován do šlechtického stavu (Kalina, 2012). Statistické metody aplikoval v biologii, zejména v genetice, evoluční biologii a eugenice (nauka o zušlechťování lidského genofondu). Po ukončení studia matematiky na University of Cambridge v roce 1912 pracoval v Londýně jako učitel matematiky a fyziky a od roku 1919 působil 14 let v Rothamsted Experimental Station. Zde se zabýval zejména experimentálním zemědělským výzkumem. Uvádí se, že díky jeho nové metodě analýzy rozptylu byly miliony lidí zachráněny před hladem.

Hlavním specifickým rysem testování hypotéz podle R. A. Fishera je fakt, že se neuvažuje žádná alternativní hypotéza. Pracuje se pouze s jednou (nulovou) hypotézou a testování lze chápat jako proces ověřování jediného modelu, který odpovídá této hypotéze. Pravděpodobnostní rozdělení tohoto „nulového modelu“ je zpravidla známé. Logiku fisherovského testování lze přirovnat ke způsobu uvažování při provádění důkazu sporem. Zajímají nás především data, která jsou v rozporu s uvažovaným modelem.

Ukažme si podstatu tohoto testování na následujícím jednoduchém příkladu:

Příklad 1. Uvažujme diskrétní náhodnou veličinu r, která může nabývat hodnot 1, 2, 3 a 4, přičemž pravděpodobnostní rozdělení hodnot této veličiny závisí na parametru Θ, který nabývá hodnot 0, 1 a 2 (viz tab. 1).

Tab. 1: Pravděpodobnostní rozdělení náhodné veličiny r v závislosti na parametru Θ (Christensen, 2005)

r	1	2	3	4
$f(r/0)$	0,980	0,005	0,005	0,010
$f(r/1)$	0,100	0,200	0,200	0,500
$f(r/2)$	0,098	0,001	0,001	0,900

Předpokládejme, že chceme fisherovsky testovat nulovou hypotézu $H_0\colon \Theta = 0$ na hladině významnosti $\alpha = 0{,}01$. Tomuto předpokladu odpovídá rozdělení pravděpodobnosti $f(r/0)$ v tabulce 1. Hladinu významnosti zde chápeme jako pravděpodobnost představující maximální riziko chyby 1. druhu, které jsme ještě ochotni akceptovat. Vidíme, že pravděpodobnost výskytu hodnot 2 a 3 je v tomto případě menší než 0,01. Pro tyto pozorované hodnoty bychom tedy nulovou hypotézu zamítli na hladině významnosti $\alpha = 0{,}01$. Po zvýšení hladiny významnosti např. na 0,02 bychom nulovou hypotézu zamítli i pro $r = 4$.

V rámci fisherovského testování hypotéz hraje pojem p-hodnoty (pozorované hladiny významnosti) významnější roli než pojem hladiny významnosti. Fisher chápal p-hodnotu jako celkovou pravděpodobnost výskytu pozorovaných dat nebo dat méně příznivých nulové hypotéze. Tabulka 2 udává p-hodnoty pro náš příklad:

Tab. 2: Tabulka p-hodnot pro rozdělení $f(r/0)$ (Christensen, 2005)

r	1	2	3	4
$f(r/0)$	0,980	0,005	0,005	0,010
p-hodnota	1,00	0,01	0,01	0,02

Hladina významnosti α představuje pouze námi stanovenou hraniční hodnotu rizika chyby 1. druhu. Tato hodnota určuje, která pozorovaná data povedou k zamítnutí nulové hypotézy a která nikoliv. Tím je určen tzv. kritický obor, který je chápán jako množina všech pozorovaných hodnot, pro něž je příslušná p-hodnota menší nebo rovna α, tj. hodnot, které vedou k zamítnutí nulové hypotézy. Všimněme si, že pro náš příklad je výsledek testu pro $\alpha = 0{,}01$ stejný jako pro $\alpha = 0{,}0125$. V obou případech zamítáme nulovou hypotézu při pozorovaných hodnotách $r = 2$ nebo $r = 3$. Uvědomme si také, že pro $\alpha = 0{,}0125$ je pravděpodobnost chyby 1. druhu 0,01, nikoliv 0,0125.

Přestože fisherovské testování explicitně nepracuje s alternativní hypotézou, nic nám nebrání provést dva nezávislé fisherovské testy pro dvě různé nulové hypotézy. Předpokládejme, že při prvním testu je nulová hypotéza $H_{01}\colon \Theta = 0$ a při druhém testu je nulová hypotéza $H_{02}\colon \Theta = 2$. Data potřebná pro testování jsou uvedena v tabulce 3. Z hodnot tabulky 2 je zřejmé, že pro hypotézu H_{01} jsou malé p-hodnoty pro $r = 2, 3, 4$. Pro tyto pozorované hodnoty bychom nulovou hypotézu H_{01} zamítli. Podobně nulovou hypotézu H_{02} zamítáme pro $r = 2, 3$ (tab. 4). Pro $r = 1$ nezamítáme H_{01} a pro $r = 4$ nezamítáme H_{02}. Pozorujeme-li hodnoty $r = 2, 3$, zamítneme obě hypotézy. Na základě fisherovského testování tedy nejsme schopni vybrat odpovídající pravděpodobnostní rozdělení.

Tab. 3: Pravděpodobnostní rozdělení pro $f(r/0)$ a $f(r/2)$ (Christensen, 2005)

r	1	2	3	4
$f(r/0)$	0,980	0,005	0,005	0,010
$f(r/2)$	0,098	0,001	0,001	0,900

Tab. 4: Tabulka p-hodnot pro rozdělení $f(r/2)$ (vlastní výpočet)

r	1	2	3	4
$f(r/2)$	0,098	0,001	0,001	0,900
p-hodnota	0,100	0,002	0,002	1,000

Obecně testování podle Fishera probíhá zpravidla v následujících krocích (Gill, 1999; Gigerenzer, 2004):

1. Formulujeme nulovou hypotézu, která nemusí odpovídat „nulovému stavu", tj. stavu odpovídajícímu náhodě.
2. Zvolíme vhodný statistický test s příslušným rozdělením, které odpovídá předpokladu, že nulová hypotéza platí.
3. Vypočítáme testové kritérium ze získaných dat.
4. Určíme p-hodnotu (pozorovanou hladinu významnosti), která koresponduje s vypočteným testovým kritériem na základě známého rozdělení nulového modelu.
5. Zamítneme H_0, pokud je p-hodnota dostatečně malá. V opačném případě neděláme žádný závěr.

Vzniká přirozená otázka, co se rozumí dostatečně malou p-hodnotou. Fisher zpočátku nebyl zastáncem stanovení pevné hranice pro tuto hodnotu a domníval se, že by měla vyplynout z kontextu konkrétního výzkumného problému. Později však v rámci svých agronomických a biologických experimentů stanovil hladiny významnosti 0,01 (higher standard) a 0,05 (lower standard), které se všeobecně ujaly (Gill, 1999).

Testování podle Neymana a Pearsona (NP)

Jerzy Neyman (1894–1981) byl přední polský matematik a statistik, který se významně zasloužil o rozvoj moderní teoretické statistiky. Jeho jméno je spojeno zejména s novým přístupem ke statistickému testování hypotéz, který navrhl spolu s E. Pearsonem. Tento nový přístup záhy našel uplatnění v lékařské diagnostice, genetice, meteorologii a astronomii. Od roku 1912 studoval Neyman na Charkovské univerzitě, kde ho velmi ovlivnil ruský matematik S. N. Bernštejn. V roce 1924 získal titul doktor filozofie na Varšavské univerzitě za dizertační práci s názvem *O použití teorie pravděpodobnosti v zemědělství.* Několik dalších let působil v Londýně (University College London) a Paříži (Université de Paris), kde spolupracoval s Karlem Pearsonem. Po návratu do Polska založil Biometrickou laboratoř v Nenckiho ústavu experimentální biologie ve Varšavě. Zabýval se zde především statistickými metodami testování léků. Od roku 1938 do konce života žil v USA, kde pracoval na kalifornské univerzitě v Berkeley (Neyman, 1982).

Egon Sharpe Pearson (1895–1980) byl podobně jako jeho otec Karl Pearson významný anglický statistik. Po studiu na Winchester College a Trinity College v Cambridge působil jako profesor statistiky na University College London a jako redaktor časopisu Biometrika. Známým se stal hlavně v souvislosti s novým přístupem ke statistickému testování hypotéz, který zveřejnil společně s Jerzym Neymanem.

Při NP testování se pracuje se dvěma hypotézami zcela odlišným způsobem. Toto testování chápeme jako rozhodovací proces mezi dvěma komplementárními alternativami H_A a H_B. Původně tyto hypotézy nebyly označovány jako nulová a alternativní, byly chápány jako dvě komplementární alternativy. Pro lepší orientaci je však takto můžeme nazývat. Uvažujme (nulovou) hypotézu $H_A\colon \Theta = 0$ a (alternativní) hypotézu $H_B\colon \Theta = 2$ (poznamenejme, že test $H_A\colon \Theta = 2$ versus $H_B\colon \Theta = 0$ je odlišný). Testování je v tomto případě založeno na určení optimálního testu pro předem stanovenou hodnotu hladiny významnosti α, kterou zde chápeme jako pravděpodobnost neoprávněného zamítnutí hypotézy H_A (chyba 1. druhu). Kritickým oborem pak rozumíme množinu pozorovaných hodnot, které vedou k zamítnutí hypotézy H_A. Při takto definované hladině významnosti tedy α odpovídá pravděpodobnosti padnutí pozorované hodnoty do kritického oboru. Podobně zavádíme parametr β jako pravděpodobnost neoprávněného přijetí hypotézy H_B (chyba 2. druhu). Klíčovou roli při NP rozhodování však hraje hodnota $1-\beta$ (síla testu). Na rozdíl od fisherovského přístupu, kde je hodnota parametru α odvozena dodatečně (aposteriori) od pozorovaných dat, při NP testování je tato hodnota stanovena předem (apriori) podle použitého testu.

Chceme-li uvažovat určitou hodnotu α, můžeme toho dosáhnout pomocí tzv. randomizovaných testů. Randomizovaný test je založen na náhodném určení kritického oboru. Ukažme si například, jak lze pro naše data realizovat randomizovaný test s hladinou významnosti $\alpha = 0{,}0125$. Této hodnoty α můžeme dosáhnout pomocí tří různých testů:

1. Pozorujeme-li $r = 4$, pak H_A zamítáme. Padne-li při hodu mincí panna, zamítáme H_A i pro $r = 2$. Kritický obor tedy závisí na výsledku hodu mincí (padne-li panna, je $K = \{4, 2\}$, padne-li orel, $K = \{4\}$). Pravděpodobnost neoprávněného zamítnutí H_A (tedy hodnota α) je $\alpha = 0{,}01 + 0{,}5 \cdot 0{,}005 = 0{,}0125$.
2. Pozorujeme-li $r = 4$, pak H_A zamítáme. Padne-li při hodu mincí panna, zamítáme H_A i pro $r = 3$. Kritický obor tedy závisí na výsledku hodu mincí (padne-li panna, je $K = \{4, 3\}$, padne-li orel, $K = \{4\}$). Pravděpodobnost neoprávněného zamítnutí H_A je $\alpha = 0{,}01 + 0{,}5 \cdot 0{,}005 = 0{,}0125$.
3. Pozorujeme-li $r = 2$ nebo $r = 3$, pak H_A zamítáme. Padne-li při dvojím hodu mincí dvakrát panna, zamítáme H_A i pro $r = 4$. Kritický obor tedy závisí na výsledku hodu mincí (padne-li dvakrát panna, je $K = \{2, 3, 4\}$, nepadne-li dvakrát panna, $K = \{2, 3\}$). Pravděpodobnost neoprávněného zamítnutí H_A je $\alpha = 0{,}005 + 0{,}005 + 0{,}5 \cdot 0{,}5 \cdot 0{,}01 = 0{,}0125$.

Je zřejmé, že právě popsaná procedura slouží k ilustraci randomizovaných testů, ale pro praktické testování bude stěží použitelná. V praxi zpravidla za nejlepší test považujeme test s největší silou, tj. s největší pravděpodobností zamítnutí hypotézy H_A v případě, že hypotéza H_B je pravdivá. Podle Neyman-Pearsonova lemmatu (Lehmann, 1997) odpovídá optimální NP test maximálnímu pravděpodobnostnímu poměru $f(r/2)/f(r/0) = 90$ (viz tab. 5). Tento test zamítá hypotézu H_A při pozorované hodnotě $r = 4$ na hladině významnosti $\alpha = 0{,}01$. Tento výsledek je zcela odlišný od fisherovského testu na stejné hladině významnosti, který zamítá H_A pro $r = 2$ nebo $r = 3$. Naproti tomu na hladině významnosti $\alpha = 0{,}02$ dávají oba testy stejné výsledky, tj. zamítají H_A pro $r = 2$, 3 a 4. Síla NP testu je na hladině významnosti $\alpha = 0{,}01$ rovna 0,9, zatímco test podle Fishera má na této hladině významnosti sílu pouze $0{,}001 + 0{,}001 = 0{,}002$. Vidíme, že test fisherovského typu není vhodný pro rozhodování mezi těmito alternativami. Slouží spíše ke zkoumání vlastností modelu odpovídajícího nulové hypotéze.

Tab. 5: Pravděpodobnostní poměry pro nulovou a alternativní hypotézu (Christensen, 2005)

r	1	2	3	4
$f(r/0)$	0,980	0,005	0,005	0,010
$f(r/2)$	0,098	0,001	0,001	0,900
$f(r/2)/f(r/0)$	0,1	0,2	0,2	90

Přestože význam parametru α je u obou typů testu odlišný, ukázali jsme na příkladech $\alpha = 0{,}01$ a $\alpha = 0{,}02$, že kritický obor fisherovského testu může definovat NP test se stejnou numerickou hodnotou α. Pro jinou hodnotu α, např. $\alpha = 0{,}0125$, to tak být nemusí.

NP testování zpravidla probíhá v následujících krocích (Gill, 1999):

1. Formulace zkoumané hypotézy H_B a komplementární hypotézy H_A.
2. Volba vhodného statistického testu s příslušným rozdělením odpovídajícím platnosti H_A.
3. Určení hladiny významnosti α a jí příslušného kritického oboru za předpokladu platnosti H_A.
4. Výpočet testového kritéria ze získaných dat.
5. Zamítnutí H_A a přijetí H_B v případě, že vypočtené testové kritérium padne do kritického oboru. Přijetí H_A v opačném případě.

Vidíme, že NP testování je na rozdíl od fisherovského založeno na pevném stanovení hladiny významnosti a následném určení příslušného kritického oboru. Nepracuje s pojmem p-hodnota, naopak významnou roli hraje pojem síla testu.

Testování významnosti nulové hypotézy (NHST)

V současnosti nejvíce uplatňovaným přístupem ke statistickému testování hypotéz v rámci vědeckého výzkumu je postup představující syntézu fisherovského a NP testování. Jak již bylo uvedeno, fisherovské testování je založeno na ověřování jedné hypotézy pomocí pojmu p-hodnota, přičemž hladina významnosti se určuje až na základě pozorovaných dat. Naproti tomu NP testování chápeme jako rozhodovací proces mezi dvěma komplementárními hypotézami založený na apriorním určení hladiny významnosti a s ním spojeného kritického oboru. Syntézou obou přístupů dostáváme tzv. testování významnosti nulové hypotézy (NHST). V rámci tohoto testování pracujeme se dvěma hypotézami, z nichž jednu explicitně označujeme jako nulovou (podobně jako Fisher) a druhou jako alternativní. Alternativní hypotéza zpravidla koresponduje s výzkumnou (věcnou) hypotézou. Podobně jako u NP testování se zde jedná o rozhodovací proces, jehož výsledkem je však pouze zamítnutí či nezamítnutí nulové hypotézy. Bohužel, k této proceduře se často přistupuje zcela mechanicky a bez hlubšího statistického uvažování. Některé kritické články rovněž hovoří o tzv. nulovém rituálu (null ritual) (Gigerenzer, 2004). Testovací proces probíhá v zásadě vždy v následujících krocích:

1. Formulování statistické nulové hypotézy odpovídající „nulovému rozdílu“, „nulové korelaci“ apod. Formulování alternativní hypotézy korespondující s výzkumnou hypotézou.
2. Stanovení konvenční pětiprocentní hladiny významnosti α pro zamítnutí nulové hypotézy.
3. Porovnání pozorované p-hodnoty se stanovenou hladinou významnosti.
4. Zdůraznění statisticky významných výsledků, pro něž je $\alpha < p$.

Metoda NHST tedy převzala z fisherovského přístupu hlavně interpretaci pojmu p-hodnota a z NP testování práci s parametry α a β.

Zajímavostí je, že i přes trvalou rivalitu mezi Fisherem a Neymanem s Pearsonem nikdo nevystoupil s nárokem na přiznání autorství NHST metody (Gill, 1999). Poprvé se pravděpodobně tento přístup objevil v knize E. F. Lindquista *Statistical Analysis in Educational Research* z roku 1940 (Perezgonzalez, 2015). Poznamenejme, že tato kniha vyšla rovněž v českém překladu v roce 1967.

Závěr

Výše popsané přístupy ke statistickému testování hypotéz nelze chápat jako konkurenční neslučitelné cesty, neboť každý z nich má ve statistice své specifické uplatnění při řešení různých problémů statistické analýzy. Pro úplnost poznamenejme, že kromě uvedených tří přístupů ke statistickému testování hypotéz existují některé další. Jmenujme alespoň alternativní bayesovský přístup založený na Bayesově větě, která byla publikována již v 18. století (Bayes & Price, 1763; Bolstad, 2007). Renesance bayesovského přístupu ke statistické inferenci začíná však až koncem 20. století díky novým možnostem výpočetní techniky a vhodného softwaru. Bayesovské testování je vhodnější než NP testování v případě rozhodování mezi několika alternativními hypotézami.

Literatura

[1] Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society, 27*, 186–190.

[2] Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418.

[3] Bolstad, W. M. (2007). *Itroduction to Bayesian statistics.* N. J.: John Wiley.

[4] Box, J. F. (1981). Gosset, Fisher, and t Distribution. *The American Statistician, 35*(2), 61–66.

[5] Emanovský, P. (2021a). První statistické testování hypotézy podle Johna Arbuthnota. *Učitel matematiky, 29*(1), 26–36.

[6] Emanovský, P. (2021b). Jak pivovarský sládek způsobil revoluci ve statistice. *Učitel matematiky, 29*(2), 96–110.

[7] Fienberg, S. E. (1992). A brief history of statistics in three and one-half chapters: a review essay. *Statistical Science, 7*(2), 208–225.

[8] Fisher, R. A. (1934). *Statistical methods for research workers.* Paternoster Row.

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

[10] Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Qurterly, 52*(3), 647–674.

[11] Hendl, J. (2004). *Přehled statistických metod – zpracování dat.* Portál.

[12] Chráska, M. (2007). *Metody pedagogického výzkumu.* Portál.

[13] Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*(2), 121–126. DOI: 10.1198/000313005X20871.

[14] Kalina, J. (2012). Ronald Fisher, otec biostatistiky. *Pokroky matematiky, fyziky a astronomie, 57*(3), 186–190.

[15] Klementa, J., Komenda, S., & Kunert, E. (1984). *Statistické metody v pedagogickém výzkumu.* VUP.

[16] Komenda, S. (1994). *Biometrie.* VUP.

[17] Lehmann, E. L. (1997). *Testing statistical hypotheses* (2nd ed.). Springer.

[18] Lindquist, E. F. (1940). *Statistical analysis in educational research.* Houghton Mifflin.

[19] Lindquist, E. F. (1967). *Statistická analýza v pedagogickém výzkumu.* SPN.

[20] Neyman, J. (1982). Vznik matematické statistiky. *Pokroky matematiky, fyziky a astronomie, 27*(3), 136–147.

[21] Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*, 1–11.

[22] Student. (1908). The probable error of a mean. *Biometrika, 6*, 1–25.

Abstract

One of the biggest controversies in the history of statistics is considered the dispute over the correct conduct of statistical significance tests. The beginning of this dispute can be traced back to the first half of the 20th century. On the one hand, there is the important English statistician R. A. Fisher and against him Polish mathematician J. Neyman together with English statistician E. Pearson. The combination of these two approaches finally led to the method of null hypothesis statistical testing (NHST), which is commonly used in empirical research today. The article describes the essence and differences of these testing methods in mathematical and historical context.

Petr Emanovský
Přírodovědecká fakulta Univerzity Palackého v Olomouci
17. listopadu 1192/12
771 46 Olomouc
e-mail: petr.emanovsky@upol.cz