Učitel matematiky

Petr Emanovský

První statistické testování hypotézy podle Johna Arbuthnota

Učitel matematiky, Vol. 29 (2021), No. 1, 26–36

Persistent URL: http://dml.cz/dmlcz/148843

PRVNÍ STATISTICKÉ TESTOVÁNÍ HYPOTÉZY PODLE JOHNA ARBUTHNOTA

Petr Emanovský[1]

Úvod

Je „dobrou“ tradicí vysokoškolské přípravy budoucích učitelů zakončit studium obhájením diplomové práce, která často obsahuje tzv. výzkumnou neboli empirickou část. Podobně je tomu u disertačních prací studentů doktorských studijních programů zaměřených na pedagogické disciplíny. Pokud se jedná o kvantitativní výzkum, jsou tito studenti nuceni zvládnout základy metodologie pedagogického výzkumu včetně potřebných metod statistického zpracování dat. Jedním z důležitých a hojně užívaných nástrojů pedagogického výzkumu je statistické testování hypotéz neboli testy statistické významnosti. Bohužel, tento nástroj je často používán nesprávně a statistická významnost (pedagogických) jevů bývá špatně interpretována, viz např. články (Blahuš, 2000; Soukup, 2010). Málokterý uživatel této metody asi také tuší, že vznikala již počátkem 18. století zásluhou skotského matematika Johna Arbuthnota. Do dnešní podoby byla tato metoda rozvinuta až ve 20. a 30. letech 20. století, zejména matematiky Ronaldem Fisherem (1890–1962), Jerzym Neymanem (1894–1981) a Egonem Pearsonem (1895–1980).

Statistické testování hypotéz

Statistické testování hypotéz je dnes podrobně popsáno v každé učebnici statistiky obsahující kromě základů popisné statistiky

[1]The work presented in this paper has been supported by the Palacky University project „Mathematical Structures“ IGA PrF 2020 012.

rovněž kapitolu o statistice indukční neboli inferenční neboli matematické, viz např. (Hendl, 2004; Chráska, 2007; Klementa et al., 1984; Komenda, 1994). Podstatu a význam této metody výstižně shrnuje profesor Komenda ve svém výroku: „Statistická indukce, budující cesty úsudkům z informace obsažené ve výběru na poměry v populaci jako celku, bere na sebe nejčastěji podobu testu statistické hypotézy.“ (Komenda, 1994, s. 80). Předmětem statistického testování může být pouze statistická hypotéza týkající se tzv. hromadného jevu, jehož výskyt či nepřítomnost lze sledovat opakovaně v mnoha situacích. Podstatou tohoto testování je porovnání pozorovaného výsledku, který jsme zjistili u náhodně vybraného vzorku, s teoretickým matematickým modelem, který předpokládá platnost tzv. nulové hypotézy. Na základě tohoto srovnání nulovou hypotézu buď zamítneme, nebo nezamítneme. Rozhodujícím kritériem tohoto rozhodovacího procesu je přitom výpočet tzv. p-hodnoty, tj. pravděpodobnosti, že nesprávně zamítneme nulovou hypotézu (chyba 1. druhu). Vypočtenou p-hodnotu srovnáváme s tzv. hladinou významnosti, tedy s rizikem chyby 1. druhu, která je pro nás ještě přípustná. Teoretickým srovnávacím modelem bývá vzorec (testové kritérium, statistika), do kterého dosazujeme naměřené hodnoty a který určuje náhodnou veličinu, jejíž pravděpodobnostní rozdělení je při platnosti nulové hypotézy známo. V dnešní době existuje řada testů statistické významnosti využívajících různé matematické modely podle typu dat a konkrétní situace statistického usuzování (Hendl, 2004). K nejčastěji používaným testům patří t-test, χ^2-test a F-test. Všimněme si blíže tzv. znaménkového testu, který poprvé použil John Arbuthnot.

Znaménkový test

Znaménkový test je jedním z nejjednodušších statistických testů významnosti, který patří mezi neparametrické testy. Je založen na redukci metrických nebo ordinálních dat na alternativní (pouze znaménko + nebo −) (Klementa et al., 1984). Jedna z variant tohoto testu umožňuje testovat statistickou významnost rozdílu mezi hodnotami dvou závislých spárovaných vzorků. Tento test

se nazývá znaménkový test pro dva párové výběry (někdy je pokládáme za jeden párový výběr). Máme-li n nezávislých párů dat (x_i, y_i), přiřadíme každému páru znaménko + v případě, že $x_i > y_i$, a znaménko − v případě, že $x_i < y_i$. Případné dvojice (x_i, y_i), pro něž platí $x_i = y_i$, do testu nezahrnujeme. Test je založen na představě, že při platnosti nulové hypotézy by se měly ve výběrech vyskytovat páry s oběma znaménky se stejnou pravděpodobností 0,5. Budou-li ve výběru převažovat páry se znaménkem + nebo naopak páry se znaménkem −, svědčí to v neprospěch nulové hypotézy. Pokud je alternativní hypotézou tvrzení, že rozdíly mezi hodnotami x_i a y_i jsou statisticky významné, hovoříme o dvoustranném znaménkovém testu. V případě, že alternativní hypotézou je tvrzení, že hodnoty x_i jsou statisticky významně větší (nebo menší) než hodnoty y_i, jedná se o test jednostranný. Abychom byli schopni rozhodnout, zda nulovou hypotézu zamítneme, či nikoliv, potřebujeme vypočítat p-hodnotu, tj. pravděpodobnost výskytu pozorovaného počtu znamének (řekněme +) nebo počtu ještě nepříznivějšího pro nulovou hypotézu. V případě jednostranného testu s alternativní hypotézou $x_i > y_i$ budou pro nulovou hypotézu nepříznivé vysoké hodnoty výskytu znamének +. Pro alternativní hypotézu $x_i < y_i$ budou naopak pro nulovou hypotézu kritické nízké hodnoty výskytu znamének +. Jedná-li se o dvoustranný test, musíme zvážit významně malý i významně velký výskyt znamének +. Z pravděpodobnostního hlediska se jedná o stejnou situaci, jako kdybychom n-krát házeli mincí a zajímala nás pravděpodobnost, že orel (nebo panna) padne právě k-krát (anebo méněkrát, resp. vícekrát). Je známo, že rozdělení pravděpodobností této náhodné veličiny odpovídá teoretickému modelu tzv. binomického rozdělení. Náhodná veličina s tímto rozdělením nabývá hodnot $k = 0, 1, 2, \ldots, n$ s pravděpodobnostmi

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

V našem případě $p = 0{,}5$, tedy

$$P(X = k) = \binom{n}{k} 0{,}5^n.$$

Jelikož musíme zvážit nejenom pravděpodobnost výskytu hodnoty k, ale i všech hodnot ještě méně příznivých pro nulovou hypotézu, budou nás spíš zajímat pravděpodobnosti

$$P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} 0{,}5^n \quad \text{a} \quad P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} 0{,}5^n.$$

Poznamenejme, že binomické rozdělení poprvé přesně popsal Jacob Bernoulli v roce 1713 a Abraham de Moivre dokázal v roce 1733, že se toto rozdělení blíží normálnímu (Hendl, 2004).

Příklad. Deset náhodně vybraných žáků 8. ročníku školy vykazovalo výsledky vědomostních testů z matematiky a fyziky uvedené v tabulce 1. Oba testy byly přibližně stejně náročné a maximální počet bodů v každém testu byl 100. Zajímá nás, jestli je rozdíl ve výsledcích obou testů pro žáky 8. ročníku této školy statisticky významný.

Tab. 1: Výsledky testů z matematiky a fyziky

i (žák)	x_i (Test M)	y_i (Test F)	znaménko (x_i, y_i)
1	47	44	+
2	58	56	+
3	69	65	+
4	59	74	−
5	82	80	+
6	93	91	+
7	96	94	+
8	58	54	+
9	76	85	−
10	88	86	+

Nulovou hypotézou je zde tvrzení, že mezi výsledky obou testů není statisticky významný rozdíl a alternativní hypotéza naopak předpokládá, že tento rozdíl statisticky významný je. Jedná se

tedy o dvoustranný test. O zamítnutí či nezamítnutí nulové hypotézy rozhoduje vypočtená p-hodnota. Ta je rovna celkové pravděpodobnosti, s níž dostaneme námi pozorovanou hodnotu (8krát znaménko + z 10 případů) nebo jakýkoliv jiný počet znamének +, který je stejně nebo méně příznivý pro nulovou hypotézu. V našem případě jsou méně příznivé případy $k = 9$ a $k = 10$. Protože $\binom{n}{k} = \binom{n}{n-k}$, máme $P(X = k) = P(X = n - k)$ a musíme do p-hodnoty zahrnout ještě případy $k = 0$, 1 a 2. Pro naše data bude tedy p-hodnota dána vztahem

$$\begin{aligned} p &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 8) + \\ &\quad + P(X = 9) + P(X = 10) \\ &= \binom{10}{0} 0{,}5^{10} + \binom{10}{1} 0{,}5^{10} + \binom{10}{2} 0{,}5^{10} + \binom{10}{8} 0{,}5^{10} + \\ &\quad + \binom{10}{9} 0{,}5^{10} + \binom{10}{10} 0{,}5^{10} \\ &= 0{,}0010 + 0{,}0098 + 0{,}0439 + 0{,}0439 + 0{,}0098 + 0{,}0010 \\ &= 0{,}1094. \end{aligned}$$

To znamená, že riziko nesprávného zamítnutí nulové hypotézy je větší než 10 %. Abychom mohli nulovou hypotézu zamítnout, musela by být vypočtená p-hodnota (tedy riziko chyby 1. druhu) menší než námi zvolená hladina významnosti (tj. riziko, které jsme ještě ochotni akceptovat). Přijmeme-li doporučenou hladinu významnosti 5 %, nulovou hypotézu nezamítáme a rozdíly mezi výsledky obou testů považujeme za statisticky nevýznamné. Vidíme, že síla tohoto testu není příliš velká, jelikož ani osm stejných znamének z deseti nevede k zamítnutí nulové hypotézy. Statisticky významný rozdíl bychom dostali až při devíti stejných znaménccích. V tomto případě by pro p-hodnotu platilo

$$\begin{aligned} p &= P(X = 0) + P(X = 1) + P(X = 9) + P(X = 10) \\ &= \binom{10}{0} 0{,}5^{10} + \binom{10}{1} 0{,}5^{10} + \binom{10}{9} 0{,}5^{10} + \binom{10}{10} 0{,}5^{10} \\ &= 0{,}0010 + 0{,}0098 + 0{,}0098 + 0{,}0010 = 0{,}0216 < 0{,}05. \end{aligned}$$

Z tabulky 1 vidíme, že počet znamének + je větší než počet znamének −. Mohli bychom tedy alternativní hypotézu formulovat ve tvaru: „Bodové hodnocení žáků v testu z matematiky je statisticky významně vyšší než v testu z fyziky.“ Pak by se jednalo o test jednostranný a p-hodnota by byla vzhledem k symetrii binomického rozdělení poloviční oproti dvoustrannému testu, tj.

$$\begin{aligned} p &= P(X = 8) + P(X = 9) + P(X = 10) \\ &= \binom{10}{8} 0{,}5^{10} + \binom{10}{9} 0{,}5^{10} + \binom{10}{10} 0{,}5^{10} \\ &= 0{,}0439 + 0{,}0098 + 0{,}0010 = 0{,}0547. \end{aligned}$$

Vypočtená p-hodnota je nyní jen mírně nad hranicí 5 %, což by však k zamítnutí nulové hypotézy nestačilo.

Poznamenejme, že vzhledem k této malé síle znaménkového testu, by bylo vhodnější použít jiný test, např. párový t-test nebo Wilcoxonův test (Hendl, 2004). Příklad použití znaménkového testu je zde uveden záměrně, abychom přiblížili způsob „statistických“ úvah Johna Arbuthnota v 18. století.

John Arbuthnot a „důkaz“ existence Boží prozřetelnosti

Dr. John Arbuthnot (1667–1735) byl skotský lékař a satirik žijící převážnou část svého života v Londýně. Časem se vypracoval mezi londýnskou smetánku, byl členem Královské společnosti (Royal Society) a Královské lékařské akademie (Royal College of Physicians). Od roku 1705 působil dokonce jako osobní lékař královny Anny. K jeho mnoha zájmům patřila matematika, zejména problémy související s pravděpodobností. V roce 1692 přeložil z latiny Huygensovo dílo o pravděpodobnosti *De ludo alease*. Zabýval se také výukou matematiky, o čemž svědčí jeho práce *Essay on the usefulness of mathematical learning* z roku 1700 (Shoesmith, 1978). V roce 1710 publikoval v časopise *Philosophical Transactions of the Royal Society* článek *An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes* (Arbuthnot, 1710). V tomto článku popisuje Arbuthnot

zajímavé výsledky svého pozorování založeného na studiu londýnských matričních záznamů za 82 let (1629–1710) (viz tab. 2). Podle těchto záznamů se každý rok během tohoto období narodilo více chlapců než dívek při zhruba konstantním poměru počtu obou pohlaví. Tento jev analyzuje již v roce 1662 John Graunt ve své práci *Bills of Mortality*, v níž si mimo jiné všímá překvapující stability poměru počtu narozených chlapců a dívek za 60 let v Londýně a Romsey (Fienberg, 1992). Zatímco Graunt zůstává na úrovni popisné statistické analýzy a spekulací o příčinách tohoto jevu, u Arbuthnota se objevuje první pokus o testování statistické významnosti. Pokud by pravděpodobnost narození chlapce a dívky byla stejná, tj. 0,5, pak by pravděpodobnost pozorovaných výsledků byla $0{,}5^{82} = \dfrac{1}{4{,}836 \cdot 10^{24}}$. Tato velmi malá hodnota utvrdila Arbuthnota v přesvědčení, že nejde o dílo náhody, ale Boží prozřetelnosti. Z hlediska dnešní statistické terminologie bychom řekli, že byla zamítnuta nulová hypotéza o stejném počtu rodících se chlapců a dívek (neboli o poměru počtů narozených pohlaví rovném 1) na základě velmi malé vypočtené p-hodnoty $0{,}5^{82}$. Jinak řečeno, rozdíl mezi počtem narozených chlapců a počtem narozených dívek byl shledán statisticky významným ve prospěch chlapců. Později podpořil Arbuthnot tyto „statistické“ úvahy dokonalejším matematickým modelem, který již odpovídá dnešnímu statistickému testování pomocí jednostranného párového znaménkového testu. Data z londýnské matriky (viz tab. 2) vlastně představují 82 párů se znaménkem +. Odtud dostáváme již zmiňovanou p-hodnotu $p = P(X = 82) = \binom{82}{82} 0{,}5^{82} = 0{,}5^{82}$. Na základě svých výpočtů a svého pozorování Arbuthnot usoudil, že tímto zásahem moudrý Stvořitel vyrovnává větší úmrtnost mužů a udržuje rovnováhu mezi počty obou pohlaví, a to nejenom ve zkoumaném období a nejenom v Londýně. Jedině tak je možné podle Arbuthnota vysvětlit, že každý muž na světě si může najít ve své zemi ženu vhodného věku. Bez Božího zásahu by bylo na světě více žen než mužů, což by vedlo k nežádoucí polygamii. V Arbuthnotově práci lze tedy nalézt první snahu o zobecnění výsledků pozorování vlastností vzorku na celou populaci, tedy první náznak statistické inference.

Tab. 2: Londýnské matriční záznamy 1629–1710 (zdroj: Arbuthnot, 1710)

Christened.			Christened.		
Anno.	*Males.*	*Females.*	*Anno.*	*Males.*	*Females.*
1629	5218	4683	1648	3363	3181
30	4858	4457	49	3079	2746
31	4422	4102	50	2890	2722
32	4994	4590	51	3231	2840
33	5158	4839	52	3220	2908
34	5035	4820	53	3196	2959
35	5106	4928	54	3441	3179
36	4917	4605	55	3655	3349
37	4703	4457	56	3668	3382
38	5359	4952	57	3396	3289
39	5366	4784	58	3157	3013
40	5518	5332	59	3209	2781
41	5470	5200	60	3724	3247
42	5460	4910	61	4748	4107
43	4793	4617	62	5216	4803
44	4107	3997	63	5411	4881
45	4047	3919	64	6041	5681
46	3768	3536	65	5114	4858
47	3796	3536	66	4678	4319

Christened.			Christened.		
Anno.	*Males.*	*Females.*	*Anno.*	*Males.*	*Females.*
1667	5616	5322	1689	7604	7267
68	6073	5560	90	7909	7302
69	6506	5829	91	7662	7392
70	6278	5719	92	7602	7316
71	6449	6061	93	7676	7483
72	6443	6120	94	6985	6647
73	6073	5822	95	7263	6713
74	6113	5738	96	7632	7229
75	6058	5717	97	8062	7767
76	6552	5847	98	8426	7626
77	6423	6203	99	7911	7452
78	6568	6033	1700	7578	7061
79	6247	6041	1701	8102	7514
80	6548	6299	1702	8031	7656
81	6822	6533	1703	7765	7683
82	6909	6744	1704	6113	5738
83	7577	7158	1705	8366	7779
84	7575	7127	1706	7952	7417
85	7484	7246	1707	8239	7623
86	7575	7119	1708	8239	7623
87	7737	7214	1709	7840	7380
88	7487	7101	1710	7640	7288

Arbuthnotovy výpočty

Arbuthnot srovnává porod dítěte s hodem mincí se stranami M (male) a F (female) (Arbuthnot, 1710) a využívá početní pravidlo, které by měl v dnešní době znát každý středoškolák a kterému se říká binomická věta. Při hodu jednou mincí padne buď strana M, nebo strana F, tedy každé pohlaví má jeden příznivý výsledek tohoto náhodného pokusu, což odpovídá koeficientům dvojčlenu $M+F$. Budeme-li házet dvěma mincemi, můžeme získat přehled o všech možných výsledcích pomocí druhé mocniny dvojčlenu $M+F$. Platí totiž $(M+F)^2 = M^2+2MF+F^2$. Koeficient členu M^2 je 1, což odpovídá jediné možnosti výsledku MM. Podobně máme jedinou možnost pro výsledek FF. Koeficient 2 prostředního členu $2MF$ nám říká, že máme 2 možnosti pro výsledek, při němž padne jednou M a jednou F (totiž MF a FM). Pokud nás zajímá pravděpodobnost, že při pokusu padne stejný počet M a F, bude hrát klíčovou roli právě koeficient tohoto prostředního členu. Tato pravděpodobnost je zřejmě $P = \frac{2}{4} = \frac{1}{2}$. Provedeme-li podobnou úvahu pro čtyři mince, dostaneme $(M+F)^4 = M^4 + {} + 4M^3F + 6M^2F^2 + 4MF^3 + F^4$. Pouze prostřední člen polynomu odpovídá výsledkům se stejným počtem M a F a koeficient tohoto členu udává počet takových výsledků ($MMFF$, $MFMF$, $MFFM$, $FFMM$, $FMFM$, $FMMF$). Pravděpodobnost stejného výskytu M a F je v tomto případě $P = \frac{6}{16} = \frac{3}{8}$. Podobně pro šest mincí by tato pravděpodobnost vyšla $\frac{20}{64} = \frac{5}{16}$, pro osm mincí $\frac{70}{256} = \frac{35}{128}$ a obecně pro n mincí (n je sudé přirozené číslo) dostaneme $P = \dfrac{\binom{n}{\frac{n}{2}}}{2^n} = \binom{n}{\frac{n}{2}} 0{,}5^n$. Je evidentní, že námi sledovaná pravděpodobnost je již od $n = 4$ menší než 0,5 a s rostoucím počtem mincí klesá. Na základě svých výpočtů, při nichž zřejmě pro větší n použil i logaritmy, Arbuthnot usoudil, že pravděpodobnost, že se ve sledovaném období narodí stejně chlapců jako dívek je menší než 0,5. Pro svůj „důkaz“ však předpokládal, že pravděpodobnost narození chlapce je stejná jako pravděpodobnost narození dívky (nulová hypotéza) a za tohoto předpokladu vypočetl pravděpodobnost $0{,}5^{82}$ (p-hodnotu), že „mužský rok“ nastane 82krát za sebou. Na základě této velmi malé vypočtené p-hodnoty „za-

mítl nulovou hypotézu“, tj. usoudil, že jev není náhodný, nýbrž je dílem Božím.

Závěr

Ačkoliv původ klasické teorie pravděpodobnosti je úzce spjat s hazardními hrami a počátky statistiky souvisí s potřebami státu evidovat informace o obyvatelstvu, je třeba si také uvědomit, že matematici, kteří se podíleli na vzniku a rozvoji těchto disciplín, byli vesměs hluboce věřící lidé. Jejich práce byla tedy často motivována teologicky, totiž snahou dokázat Boží existenci. Přesto lze říci, že Arbuthnotova práce byla ve své době novátorská v tom, že podala matematický (statistický) „důkaz“ svého tvrzení založený na kvantitativně chápaném pojmu pravděpodobnosti s numericky vyjádřenými argumenty. Tím se Arbuthnot zapsal do historie statistiky a je považován za prvního člověka, který použil statistický test významnosti.

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. Dostupné z `http://www.jstor.org/stable/103111`

[2] Blahuš, P. (2000). Statistická významnost proti vědecké průkaznosti výsledků výzkumu. *Česká kinantropologie, 4*(2), 53–72.

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

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

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

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

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

[8] Shoesmith, E. (1978). The continental controversy over Arbthnot's argument for divine providence. *Historia Mathematica, 14*, 133–146.

[9] Soukup, P. (2010). Nesprávná užívání statistické významnosti a jejich možná řešení. *Data a výzkum – SDA Info, 4*(2), 77–104.

Abstract

Statistical testing of hypotheses is one of the important tools of quantitative (not only educational) research. The idea of verifying the hypothesis using a "mathematical model" probably first appeared in the 18th century in John Arbuthnot's work. The article describes this first testing in historical and mathematical context.

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