Základní statistické testy pro rozložení diskrétních znaménkových dat

JAN ŘEHÁK
Ústav pro filozofii a sociologii ČSAV, Praha
BLANKA ŘEHÁKOVÁ
Ústav pro výzkum veřejného mínění FSÚ, Praha

V sociologickém výzkumu se často setkáváme s diskrétními (nespojitými) znaménkovými daty. Setkáváme se s nimi buď přímo, jako se znaky používanými v dotaznících, nebo nepřímo, jako se znaky odvozenými. Vyskytují se jednak v původních souborech výzkumných dat, ale jsou též používány pro komparativní cíle, resp. v sekundární analýze.

Diskrétní znaménková data jsou typem dat ordinálních: mají tři hodnoty, z nichž střední kategorie znamená neutrální hodnotu (přirozený počátek) a obě krajní kategorie reprezentují dva opačné póly na ordinální stupnici, rozložené okolo daného středu (−, 0, +).

Cílem stati je podat přehled různých statistických technik pro vyhodnocení takových dat. Nejprve uvádíme příklady diskrétních znaménkových dat (část 1), v druhé části nastiňujeme obecný postup testování statistických hypotéz. Dále pak probíráme jednotlivé techniky testování, a to: nejprve odhad parametrů, dále znaménkový test, McNemarův test a chí-kvadrát test pro symetrii.

Výběr metod pro tento přehled byl dán třemi hledisky: a) snahou pokrýt metody, které se běžně používají, a ukázat na jejich některé podobnosti, popřípadě i identičnost ve speciálních případech; b) podat co nejširší přehled těchto metod; c) upozornit na zajímavé hypotézy, které se vyskytly v naší praxi a které podle našeho názoru zasluhují rozšíření.

Nakonec uvádíme rozšíření pojmu diskrétních ordinálních dat se třemi hodnotami (−, 0, +) na případ, kdy je stupnice jak na kladné, tak na záporné straně rozšířena, kategorie jsou směrem ke kladnému i zápornému pólu odstupňovány. Pro tento případ uvádíme jednoduché rozšíření chí-kvadrátových testů. Stať má sloužit jako pracovní přehled a pracovní pomůcka. Proto zde neuvádíme žádné matematické odvozování (a to ani v případě rozšíření běžných testů na více než tříhodnotové znaky). Naproti tomu uvádíme tabulky kritických hodnot pro znaménkový test i pro běžně známý chí-kvadrát a pro normální rozložení (vybíráme ty hodnoty, které jsou pro uvedené úlohy užitečné a jejichž zařazení má usnadnit a zrychlit aplikaci).

Popsané metody patří do metodologie klasické statistické analýzy. Novou metodu, založenou na Bayesovském přístupu a odpovídající znaménkovému testu, popíšeme v jiné stati.

Znaménková data

Znaménková data vznikají realizací proměnné S, znaku, který má tyto vlastnosti:

a) má tři hodnoty,
b) má věcně daný počátek (neutrální kategorie),
c) kromě počátku má dvě pólové hodnoty, které určují opačné obsahové orientace na stupnici hodnot znaku.

částečně dokumentovat široké spektrum jejich výskytu. Stručný soupis ovšem nevystihuje metodologická úskalí a složitou metodologii konstrukce u některých uvedených typů.

Příklady diskrétních znaménkových proměnných:

a) Vliv zavádění nové techniky v závodech. Pro N dílen zkoumáme postoje k technice před změnou techniky a technologie a po ní, nebo krátce po zavedení a poté po dvou letech.
„+" = v dílně bylo zjištěno signifikantní zlepšení postoje,
„0" = rozdíl v postojích nebyl prokázán,
„—" = rozdíl statisticky prokázán, po zavedení nastalo zhoršení postojů.

b) V jedné dílně zkoumáme změnu postoje u N osob. Dotazujeme se před zavedením techniky a po zavedení techniky.
„+" = postoj se u respondenta zlepšil (kladný rozdíl),
„0" = postoj se u respondenta nezměnil,
„—" = postoj se u respondenta zhoršil (záporný rozdíl).
Určení hodnoty zde buď může plynout z přímého porovnání dvou odpovědí na tutéž otázku, nebo složitějším vyhodnocením celého profilu odpovědí, resp. dané baterie položek.

c) Zkoumáme očekávaný vliv zavádění další automatizace na různé aspekty pracovního života dělníků (např. na spokojenost s prací). Postoje jednotlivých respondentů jsou strukturovány takto:
„+" = automatizace způsobí zvýšení (zlepšení),
„0" = automatizace nebude mít vliv žádný,
„—" = automatizace způsobí snížení (zhoršení).

d) Zkoumání spokojenosti s prací může poskytnout tři hodnoty znaku:
„+" = respondent zcela nebo částečně spokojen,
„0" = respondent ani spokojen, ani nespokojen,
„—" = respondent zcela nebo spíše nespokojen.

e) Otázku stability zaměstnání lze hodnotit pomocí znaku:
„+" = respondent je rozhodnut změnit zaměstnání,
„0" = respondent neví, váhá, rozhoduje se,
„—" = respondent je rozhodnut neměnit zaměstnání.

f) Při porovnávání například automatizovaných transferlinek a dílen klasických jednoúčelových strojů v N závodech dostáváme komparací pro měřenou proměnnou X tři možnosti:[1 2])
„+" = v automatizované jednotce je $\overline{X}_a$ signifikantně vyšší než $\overline{X}_n$ v neautomatizované jednotce,
„0" = mezi průměry $\overline{X}_a$ a $\overline{X}_n$ není signifikantní rozdíl,
„—" = $\overline{X}_a$ je signifikantně nižší než $\overline{X}_n$.

g) Při hodnocení určitých aspektů života z hlediska stavu a z hlediska přání respondenta můžeme odpovědi členit takto:
„+" = stav je hodnocen výše než přání,
„0" = stav je stejný jako přání,
„—" = stav je hodnocen níže než přání.
Takovou řadu proměnných hodnotíme paralelně, splňují-li jisté metodologické předpoklady (odpovídají jedné určité dimenzi, nebo jsou to nezávislé komponenty zkoumaného a porovnávaného obsahu, které navíc jsou v obsahové dekompozici vyvážené a vyrovnané).

h) Dvě země (resp. dvě skupiny zemí), řekněme A a B, jsou postupně komparovány podle N různých proměnných:
„+" = A má signifikantně vyšší průměr než B u dané proměnné,
„0" = u dané proměnné se průměry pro A a B statisticky prokazatelně neliší,
„—" = A má signifikantně nižší průměr než B u dané proměnné.
O tomto případě platí stejné zásady jako u příkladu g). Jde tu o možnost vyhodnocování profilů pomocí znaménkové informace v datech.

---

[1]) Příklad je zvláště důležitý pro sekundární analýzu a zobecňování z různých výzkumů, v nichž bylo měření prováděno různými prostředky.

[2]) V tomto příkladě, stejně jako v a) a h), by bylo přesnější říci:
„+" = $\overline{X}_A$ je signifikantně rozdílný od $\overline{X}_N$ a $\overline{X}_A > \overline{X}_N$; „−" = $\overline{X}_A$ je signifikantně rozdílný od $\overline{X}_N$ a $\overline{X}_N > \overline{X}_A$. Takovéto pravidlo je v praxi běžně používáno, i když neodpovídá přesně dvoustranné alternativní hypotéze $H_0 : \overline{X}_A = \overline{X}_N$, $H_1 : \overline{X}_A \neq \overline{X}_N$. Jestliže zamítneme H_0, pak chyba ve špatném určení znaménka je sice nenulová, je však velmi malá, a proto je postup v praxi používán. Současné testování dvou jednostranných hypotéz též neodpovídá modelu.

„+" = dávám přednost sportu,
„0" = čtu rád o obojím asi stejně,
„—" = dávám přednost kultuře.
(Zde je přiřazeno „+" a „—" oběma obsahovým pólům libovolně)

Vznik těchto dat může být charakterizován z různých hledisek (určených vždy dvěma polárními alternativami):

a) vznikají jako diskrétní × vznikají diskretizací spojité škály (resp. sloučením podrobnějších kategorií);
b) trichotomizace je určena pevnými prahy × je určena jednoznačným pravidlem (např. výsledkem statistického testu);
c) data vznikají přímo ve znaménkovém tvaru × jsou do znaménkového tvaru převedena;
d) vznikají v experimentu (přirozeném, resp. řízeném) × vznik neodpovídá experimentální situaci;
e) vznikají porovnáním dvou různých časových okamžiků × neobsahují časovou dimenzi;
f) porovnávají zkoumané jednotky vzhledem k dané proměnné (jedinec, skupina, soubor, země apod.) × porovnávají proměnné na téže statistické jednotce;
g) znaménka polárních kategorií mají jednoznačnou obsahově orientovanou interpretaci × znaménka polárních kategorií jsou libovolně volitelná;
h) porovnání je vedeno na těchže statistických jednotkách (párově; statisticky jde o závislé výběry) × porovnáváme statisticky nezávisle vzniklá data (nezávislé výběry).

Každý z uvedených pohledů je nutné vzít v úvahu při zkoumání použitelnosti statistických technik. Podle geneze dat musíme určit, zda platí (v rámci přípustných tolerancí) výchozí statistický model.

Ani tento široký výčet hledisek není úplný. Například dichotomický znak s odpověďmi „ano", „ne" a velkým počtem odpovědí „nevím" můžeme považovat za znaménkový, v němž („+" = = „ano", „0" = „nevím", „—" = „ne").

Popis obecného modelu

Možné příklady lze v tomto případě obecně shrnout takto: Zkoumáme N objektů (podle situace jsou to osoby, skupiny, země, proměnné, obsahové okruhy apod.). Těchto N objektů bylo vybráno na základě prostého náhodného výběru. Určení jednotlivých objektů do souboru bylo provedeno tak, že každý z nich byl určen nezávisle na jiných, současně však u všech objektů populace je stejná pravděpodobnost, že se dostanou do výběru; totéž platí o všech dvojicích, trojicích, čtveřicích atd. objektů.

Tento předpoklad modelu je v praxi opravdu jen zřídka striktně splněn, přípustné odchylky od modelu záleží na jednotlivých konkrétních případech, a proto je zde nemůžeme obecně řešit.

Jiný předpoklad (který vede k témuž statistickému zpracování) je charakterizován takto: jednotlivé údaje, které vytvářejí soubor znaménkových dat, vznikají nezávisle jeden na druhém, tj. jeden údaj neovlivňuje hodnotu jiného.

Ověření těchto předpokladů není v běžných výzkumných situacích obtížné. V situacích netypických může znamenat ověření předpokladů složité meritorní metodologické i statistické úvahy. Proto v těchto případech doporučujeme konzultaci s matematickým statistikem.

V dalším kroku zařadíme N zkoumaných objektů do tří kategorií, odpovídajících třem hodnotám znaku. Výsledkem je velice jednoduchá jednorozměrná tabulka (tabulka 1. stupně třídění).

Tabulka 1. Rozložení četností diskrétního znaménkového znaku

Hodnota znaku	„−“	„0“	„+“	Celkem
absolutní četnost výskytu	N_-	N_0	N_+	N
relativní četnost výskytu	$f_- = \frac{N_-}{N}$	$f_0 = \frac{N_0}{N}$	$f_+ = \frac{N_+}{N}$	1
procento výskytu	$100\, f_-$	$100\, f_0$	$100\, f_+$	100

Předpokládáme, že relativní výběrové četnosti (f_-, f_0, f_+) odpovídají populačním výběrovým četnostem, resp. pravděpodobnostem výskytu (p_-, p_0, p_+). Úkolem statistické inference je využít informaci v absolutních, resp. relativních výběrových četnostech k přímé výpovědi o neznámých skutečných parametrech (p_-, p_0, p_+) či k výpovědi o jejich vybraných vlastnostech a jejich vzájemných relacích.

Všechny testy i metody odhadu parametrů jsou v této práci založeny na předpokladu multinomického rozložení (resp. jeho speciálního případu: binomického rozložení).

Při analýze dat řešíme dále tyto základní úlohy:

1. odhad parametrů p_-, p_0, p_+,
2. ověření hypotézy, že jedna z pravděpodobností je větší než jedna polovina (např. $p_+ > \frac{1}{2}$),
3. ověření hypotézy symetrie, tj. že $p_+ = p_-$.

Odhad parametrů rozložení diskrétních parametrických dat

Jednou ze základních úloh analýzy znaménkových dat je odhad neznámých parametrů výchozího statistického modelu, a to jak odhad bodový (číslo, které nejlépe charakterizuje neznámý parametr — je mu nejblíže), tak intervalový (interval, který charakterizuje navíc přesnost statistického měření a pás, v němž se s vysokou spolehlivostí neznámá hodnota parametru nachází). Pro větší četnosti se používá většinou klasický odhad pomocí relativních četností.

$$(1) \qquad \ddot{p}_- = \frac{N_-}{N}, \qquad \ddot{p}_0 = \frac{N_0}{N}, \qquad \ddot{p}_+ = \frac{N_+}{N}$$

Konfidenční intervaly pro parametr p_+ je pro malá N obtížné spočítat. Tyto intervaly lze však nalézt přímo v tabulkách [Janko 1958], [Bolšev, Smirnov 1965], [Owen 1966]. Pro větší N, je-li splněna podmínka

$$(2) \qquad \min\,(Np_+, N(1-p_+)) > 5$$

dává dostatečně přesnou aproximaci vzorec (viz Nayatani, Kurahara [1964]):

Hranice konfidenčního intervalu $(1-\alpha)\ 100\ \%$ pro p_+ jsou:

$$(3) \qquad p_+ = \frac{N_+}{N} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{N_+}{N^2}\left(1 - \frac{N_+}{N}\right)} = \frac{N_+}{N} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{N_+(N - N_+)}{N^3}}$$

$z_{\frac{\alpha}{2}}$ je $(1 - \frac{\alpha}{2})$ 100% kvantil normálního rozložení, pro který uvedeme tabulku nejpoužívanějších hodnot.

Tabulka A. Tabulka koeficientů pro dvoustranné konfidenční intervaly

$(1-\alpha)$ 100 %	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$	$z_{\frac{\alpha}{2}}$ s velkou přesností
90 %	0,05	1,645	1,6448 5363
95 %	0,025	1,960	1,9599 6398
99 %	0,005	2,576	2,5758 2930
99,9 %	0,0005	3,291	3,2905 2673

Pro ostatní parametry p_-, p_0 provedeme jednoduchou záměnu ve vzorci (2): místo N_+ zavedeme N_-, resp. N_0.

Příklad 1

25 respondentů z jedné pracovní skupiny, charakterizované určitým typem technologie odpovědělo na otázku, jaký vliv bude mít zavádění další automatizace na jejich spokojenost s prací takto:

pracovní spokojenost se po zavedení další automatizace zvýší = 12 (N_+)
pracovní spokojenost se po zavedení další automatizace nezmění = 7 (N_0)
pracovní spokojenost se po zavedení další automatizace sníží = 6 (N_-)

Zajímá nás, do jaké míry lze tento výsledek přijmout z hlediska spolehlivosti.

Pro p_+ dostáváme bodový odhad $p_+ = \frac{12}{25} = .48$. V tabulkách vyhledáme přesný 95% interval spolehlivosti: (.278, .687). Použijeme-li normální aproximace vzorce (3) dostaneme:

a) $1 - \alpha = .95$, $\alpha = .05$, $\frac{\alpha}{2} = .025$
v tabulce A: $z_{.025} = 1{,}960$

b) dosadíme do (3):

$$\text{hranice} = \frac{12}{25} \pm 1{,}96 \sqrt{\frac{12\,(25-12)}{25^3}}$$

$= .48 \pm .1958$
$= (.284, .676)$
šířka intervalu $= .392$

c) odhad dolní hranice (2):
$\min (Np_+, N(1-p_+)) = \min (12, 13) = 12 > 5$

Bereme-li v úvahu náhodná působení (chyby, které nelze nijak blíže specifikovat) při výpovědích respondentů, pak lze na základě našeho modelu provést závěr: *„skutečné procento kladného očekávání leží, se spolehlivostí 95 %, v hranicích (28 %, 68 %) s nejvěrohodnějším odhadem 48 %"*.

Stejný postup bychom volili, kdybychom skupinu 25 osob považovali za náhodný výběr z populace pracující ve stejných technologických podmínkách a chtěli — za předpokladu skutečné spolehlivých výpovědí — zobecnit výsledky na celou takovou populaci.

Příklad 2

Z téhož souboru odpovědělo 20 lidí, že očekává zlepšené perspektivy pro mladé lidi, 3 osoby neočekávaly žádné změny, 1 osoba očekávala zhoršení (1 respondent neodpověděl).

Máme zde:

$$N_- = 1, \quad N_0 = 3, \quad N_+ = 20, \quad N = 24$$

Odhad pro $p_+ = \frac{N_+}{N} = \frac{20}{24} = .83$, tj. 83 %,

přesný interval 95% spolehlivosti: (.626, .953)
normální aproximace intervalu spolehlivosti vzorcem (3): (.684, .982).
Normální aproximace se liší jak šířkou intervalu (interval je užší), tak posunutím intervalu. Nepřesnost znamená posun o 5 %, resp. 3 %, je tedy nepřípustná.

Neadekvátnost normální aproximace je indikována též odhadem podmínky (2): min (Np_+, $N(1-p_+)$) = min (20,4) = 4 < 5. Neshodu jsme tudíž mohli předvídat.

Závěry u tohoto příkladu jsou podobné jako u příkladu 1: „95% interval spolehlivosti pro skutečnou hodnotu parametru p_+ je (.63,.95), v procentech: (63%, 95%)."

Příklad jsme uvedli hlavně proto, abychom upozornili na nutnost dodržení podmínek statistických postupů. V tomto případě šlo o podmínku pro použití numerické aproximace konfidenčního intervalu.

Existence majoritní pravděpodobnosti v modelu

U tohoto modelu znaménkových dat nás může zajímat, zda některá z pravděpodobností p_+, p_0, p_- je větší než 1/2, tj. zda z hlediska daného znaku existuje majoritní názor, majoritní nadpoloviční skupina jednotek v jedné kategorii. (Poznamenejme, že postup této části neplatí jen pro znaménková data, ale pro jakýkoli nominální znak, u něhož nás zajímá, zda jedna z jeho kategorií odpovídá svou četností nadpoloviční většině v populaci).

Úlohu budeme formulovat pro parametr p_+, prostou záměnou indexů, avšak všechny výsledky platí i pro p_-, p_0 (resp. pro p_i u nominálního znaku).

A. Znaménkový test (binomický test)

Úlohu řešíme znaménkovým, resp. binomickým testem.

Postup:

1. *Formulace hypotézy:*

$$H_0 : p_+ = \frac{1}{2} \qquad H_1 : p_+ > \frac{1}{2}$$

2. *Data:*
Je známo rozložení četností: N_-, N_0, N_+ (celkem N). K postupu stačí znát: N_+, $N - N_+$

3. Intuitivní model: jestliže platí H_0, N_+ by mělo být přibližně rovno $\frac{N}{2}$, jestliže platí H_1, N_+ by mělo být mnohem větší než $\frac{N}{2}$.

4. *Testové kritérium pro malé výběry* může být provedeno ve dvou ekvivalentních tvarech. Postupy jsou aplikovatelné podle dostupnosti tabulek.
Postup A: Zjistíme hodnotu statistické významnosti:

$$(4) \quad P = P_B(X \geqq N_+ | p = \frac{1}{2}, N) = (\frac{1}{2})^N \sum_{i=N_+}^{N} \binom{N}{i}$$

kde P_B je binomická pravděpodobnost s uvedenými parametry a vyjadřuje výskytovost obdrženého N_+ nebo ještě extrémnější hodnoty (za předpokladu H_0). P můžeme zjistit buď výpočtem (např. pomocí Pascalova trojúhelníku), nebo přímo v tabulkách distribuční funkce binomického rozložení (např. [Janko 1958], [Bolšev, Smirnov 1965]). Je-li $P \leqq \alpha$, zamítáme H_0 ve prospěch H_1. Je-li $P > \alpha$, přijímáme H_0.

Postup B: Zjistíme *počet výskytů* v dané kategorii, tj. N_+, které porovnáme ve speciálních tabulkách s kritickou hodnotou k_N, danou pro každé N. Je-li $N_+ \geqq k_N$,

zamítáme H_0 ve prospěch H_1. Je-li $N_+ < k_N$, nemáme důvod H_0 zamítnout. Kritické hodnoty pro $N \leqq 50$ a pro vybraná N dále až do 1000 uvádíme v tabulce B, která je upravena tak, že hypotézu $H_1 : p_+ > \frac{1}{2}$ přijímáme tehdy, jestliže $N - N_+ \leqq$ hodnota v tabulce. (Postup je pochopitelně ekvivalentní. Čím větší je N_+, tím menší je $N - N_+$.)

5. *Testové kritérium pro výběry $N \geq 30$*

Tabulky pro postupy A i B existují pouze do relativně nízkého počtu pozorování. Je-li $N \geqq 30$, můžeme použít normální aproximaci binomického rozložení, a to jak pro numerické aproximace P, tak pro numerické aproximace k_N.

Postup A: Numerická aproximace hodnoty P může být nalezena dvěma způsoby:

a) *Gram-Charlierova aproximace* (s korekcí na spojitost)

$$Y = \frac{2N_+ + 1 - N}{\sqrt{N}} \tag{5}$$

$$P \doteq 1 - \Phi(Y) = \Phi(-Y)$$

b) *Camp-Paulsonova aproximace*

$$Y = \frac{X}{3\sqrt{Z}} \tag{6}$$

$$X = \left(\frac{N_+}{M+1}\right)^{\frac{1}{3}} \cdot \left(9 - \frac{1}{N_+}\right) - 9 + \frac{1}{M+1}$$

$$Z = \left(\frac{N_+}{M+1}\right)^{\frac{2}{3}} \frac{1}{N_+} + \frac{1}{M+1}$$

$$M = N - N_+$$

$$P \doteq 1 - \Phi(Y) = \Phi(-Y)$$

V obou případech je $\Phi(y)$ distribuční funkce standardizovaného normálního rozložení, která je tabelována v každých statistických tabulkách i v učebnicích. Proto můžeme k Y nalézt příslušnou pravděpodobnost. To není ovšem nutné, pokud nás nezajímá přímo hodnota P samotná. Pro zvolené α můžeme Y porovnat přímo s α-kvantilem normálního rozložení $N(0,1)$, který značíme z_α. Rozhodovací pravidlo je pak:

$$Y \geqq z_\alpha, \quad H_0 \text{ zamítáme ve prospěch } H_1 \tag{7}$$

$$Y < z_\alpha, \quad H_0 \text{ přijímáme}$$

Postup B: Pro $N > 100$ můžeme nalézt přibližně hranice pro odmítnutí hypotézy. H_0 zamítáme ve prospěch H_1 ($\equiv p_+ > \frac{1}{2}$), jestliže

$$N_+ \geqq k_N \tag{8}$$

kde $k_N \doteq \frac{N-1}{2} + r\sqrt{N+1}$, a je zaokrouhlené na celé číslo, a hodnoty r jsou dány pro různé hladiny α tabulkou D.

Tabulka B. Kritické hodnoty znaménkového testu

pro $H_0 : p_+ = \frac{1}{2}$, $H_1 : p_+ > \frac{1}{2}$

(H_0 zamítáme ve prospěch H_1, je-li $(N - N_+) \leqq$ hodnota v tabulce. „—" znamená, že testovat nelze.)

(Tabulku lze použít též pro testování hypotézy H_0 proti $H_1 : p_+ \neq \frac{1}{2}$, tedy pro dvoustranný znaménkový test. Chceme-li v tom případě testovat na hladině α, použijeme sloupec označený $\frac{\alpha}{2}$.)

n	0,005 (0,5%)	0,01 (1%)	0,025 (2,5%)	0,05 (5%)	0,10 (10%)
4	—	—	—	—	0
5	—	—	—	0	0
6	—	—	0	0	0
7	—	0	0	0	1
8	0	0	0	1	1
9	0	0	1	1	2
10	0	0	1	1	2
11	0	1	1	2	2
12	1	1	2	2	3
13	1	1	2	3	3
14	1	2	2	3	4
15	2	2	3	3	4
16	2	2	3	4	4
17	2	3	4	4	5
18	3	3	4	5	5
19	3	4	4	5	6
20	3	4	5	5	6
21	4	4	5	6	7
22	4	5	5	6	7
23	4	5	6	7	7
24	5	5	6	7	8
25	5	6	7	7	8
26	6	6	7	8	9
27	6	7	7	8	9
28	6	7	8	9	10
29	7	7	8	9	10
30	7	8	9	10	10
31	7	8	9	10	11
32	8	8	9	10	11
33	8	9	10	11	12
34	9	9	10	11	12
35	9	10	11	12	13
36	9	10	11	12	13
37	10	10	12	13	14
38	10	11	12	13	14
39	11	11	12	13	15
40	11	12	13	14	15
41	11	12	13	14	15
42	12	13	14	15	16
43	12	13	14	15	16
44	13	13	15	16	17
45	13	14	15	16	17

n	0,005 (0,5 %)	0,01 (1 %)	0,025 (2,5 %)	0,05 (5 %)	0,10 (10 %)
46	13	14	15	16	18
47	14	15	16	17	18
48	14	15	16	17	19
49	15	15	17	18	19
50	15	16	17	18	19
52	16	17	18	19	20
54	17	18	19	20	21
56	17	18	20	21	22
58	18	19	21	22	23
60	19	20	21	23	24
62	20	21	22	24	25
64	21	22	23	24	26
66	22	23	24	25	27
68	22	23	25	26	28
70	23	24	26	27	29
72	24	25	27	28	30
74	25	26	28	29	30
76	26	27	28	30	31
78	27	28	29	31	32
80	28	29	30	32	33
82	28	30	31	33	34
84	29	30	32	33	35
86	30	31	33	34	36
88	31	32	34	35	37
90	32	33	35	36	38
92	33	34	36	37	39
94	34	35	37	38	40
96	34	36	37	39	41
98	35	37	38	40	42
100	36	37	39	41	43
110	41	42	44	45	47
120	45	46	48	50	52
130	49	51	53	55	57
140	54	55	57	59	61
150	58	60	62	64	66
160	63	64	67	69	71
170	67	69	71	73	76
180	72	73	76	78	80
190	76	78	81	83	85
200	81	83	85	87	90
220	90	92	94	97	99
240	99	101	104	106	109
260	108	110	113	116	119
280	117	120	123	125	128
300	127	129	132	135	138

n	0,005 (0,5 %)	0,01 (1 %)	0,025 (2,5 %)	0,05 (5 %)	0,10 (10 %)
320	136	138	141	144	148
340	145	148	151	154	157
360	155	157	160	163	167
380	164	166	170	173	177
400	173	176	179	183	186
420	183	185	189	192	196
440	192	195	198	202	206
460	201	204	208	211	215
480	211	214	218	221	225
500	220	223	227	231	235
550	244	247	251	255	259
600	267	271	275	279	283
650	291	294	299	303	308
700	315	318	323	327	332
750	339	342	347	351	356
800	363	366	371	376	381
850	386	390	395	400	405
900	410	414	420	424	430
950	434	438	444	449	454
1000	458	462	468	473	479

Příklad 3

Ve 115 závodech byly porovnávány postoje k práci u pracovníků z dílen dvou různých technologických úrovní. V 71 případech byl postoj pracovníků z dílen vyšší technologické úrovně lepší, v 25 případech nebyl nalezen podstatný rozdíl, v 39 případech byl lepší postoj k práci v dílnách s nižší úrovní technologie. Otázkou je, zda můžeme zobecnit výrok, že v nadpoloviční většině závodů vůbec lze očekávat lepší postoj k práci u dělníků pracujících v dílnách vyšší technologické úrovně. Hladinu významnosti volíme $\alpha = 0{,}05$.

1. Hypotéza: $H_0 : p_+ = 1/2$, $H_1 : p_+ > 1/2$
2. Data: $N_+ = 71$, $N - N_+ = 44$, $N = 115$

Ukážeme oba dva postupy; v obou případech musíme volit aproximace:

Postup A

metoda a) (Gram-Charlierova aproximace):

$$y = \frac{2 \cdot 71 + 1 - 115}{\sqrt{115}} = \frac{28}{10 \cdot 72} = 2{,}6110$$

pro $\alpha = 0.05$ nalezneme v tabulce C hodnotu $z = 1{,}645$ a vzhledem k tomu, že $y = 2{,}6110 > 1{,}645$ můžeme přijmout H_1 jako dostatečně prokázanou.

Tabulka C. Hodnoty z_α pro vybrané hladiny významnosti

α	100α %	z_α	z_α s velkou přesností
0,1	10 %	1,282	1,2815 5157
0,05	5 %	1,645	1,6448 5363
0,01	1 %	2,326	2,3263 4787
0,005	0,5 %	2,576	2,5758 2930
0,001	0,1 %	3,090	3,0902 3231

Podle Janko [1958: 115] nalezneme:

$$P = 1 - \Phi(2{,}6110) = 1 - .995473 = .0045$$

(P je menší než $\alpha = 0{,}05$, H_1 přijímáme jako dostatečně statisticky prokázanou).
metoda b) (Camp-Paulsenova aproximace):

$$x = \left(\frac{71}{45}\right)^{\frac{1}{3}} \left(9 - \frac{1}{71}\right) - 9 + \frac{1}{45} = 1{,}4833$$

$$= \left(\frac{71}{45}\right)^{\frac{2}{3}} \frac{1}{71} + \frac{1}{45} = 0{,}0413$$

$$y = \frac{1{,}4833}{3\sqrt{.0413}} = 2{,}4327$$

Opět platí, že $y > 1{,}645$

$$P = 1 - \Phi(2{,}43) = \Phi(-2{,}43) = .007549, \quad P < 0{,}05$$

Postup B:

Aproximace kritické hodnoty se provádí pomocí vzorce (8) a tabulky D:

$$\alpha = 0{,}05 \Rightarrow r = .8224$$

$$k_{115} \doteq \frac{115 - 1}{2} + .8224\sqrt{115 + 1} = 65{,}86$$

Vzhledem k tomu, že $N_+ = 71 > k_{115} = 65{,}86$, můžeme hypotézu ($H_1$) přijmout.

Tabulka D. Koeficienty pro aproximaci kritických hodnot

α	r	r s velkou přesností
0,1	0,6408	0,6407 75785
0,05	0,8224	0,8224 26815
0,01	1,1632	1,1631 73935
0,005	1,2879	1,2879 14650
0,001	1,5451	1,5451 16155

Hypotéza symetrie pro znaménková data

Testy, které dále uvádíme, jsou testy pro symetrii rozložení znaménkových dat. Tedy bez ohledu na obsazení neutrální kategorie „0" nás zajímá vztah dvou pravděpodobností p_+ a p_-.

A) Znaménkový test

Za předpokladu, že nulová změna, resp. obsazení neutrální kategorie, má nulovou pravděpodobnost, či za předpokladu, že znaménková data vznikají jako dichotomizace spojité veličiny, byl odvozen znaménkový test pro detekci převahy „zlepšení" nad „zhoršením". Pro případ výskytu tzv. „spojení" (v našem případě existence neutrální kategorie s nenulovým výskytem N_0) byly navrženy různé modifikace metody. Nejvýhodnější je vynechat spojení, to znamená redukovat výběrový soubor o N_0 pozorování. Máme pak $N - N_0 = N^*$ pozorování a u nich N_+ zlepšení a N_- zhoršení (viz např. [Van der Warden 1960]).

Hypotéza symetrie může být pomocí znaménkového testu prověřována proti třem různým alternativám: $p_+ > p_-$, $p_+ < p_-$, $p_+ \neq p_-$. První dvě hypotézy jsou symetrické, obě prověřujeme stejným způsobem s pouhou záměnou „+" a „−". První dvě alternativy se nazývají *jednostranné*, zatímco třetí alternativa se nazývá *dvoustrannou*.

Jednostranná alternativa

1. *Formulace hypotézy:* $H_0 : p_+ = p_-$, $H_1 : p_+ > p_-$
2. *Data:* N_+, N_-, N_0, přičemž pro testování stačí znát N_+, N_-
3. *Postup:* Postupujeme přesně stejným postupem jako v případě testování hypotézy majoritní pravděpodobnosti. Po vynechání p_0 a redukci souboru o neutrální kategorie jsou hypotézy formulovány:

$$H_0 : p'_+ = \frac{1}{2} , \quad H_1 : p'_+ > \frac{1}{2}$$

p'_+, p'_- jsou nové pravděpodobnosti (po redukci), které se vzájemně doplňují ($p'_+ + p'_- = 1$).

Dvoustranná alternativa

1. *Formulace hypotézy:* $H_0 : p_+ = p_-$ $H_1 : p_+ \neq p_-$
2. *Data:* N_+, N_-, $N^* = N_- + N_+$ (stejně jako výše)
3. *Postup:* Redukujeme data o N_0 a aplikujeme znaménkový test jako u jednostranné alternativy, pouze s těmito dvěma změnami: Místo $N^* - N_+ = N_-$ používáme v tabulce B:

$$n = \min (N_-, N_+)$$

a místo sloupce pro α používáme sloupce pro $\frac{\alpha}{2}$. U aproximací též postupujeme stejně: Místo N_- použijeme n a místo N_+ použijeme $N^* - n$; místo α opět používáme $\frac{\alpha}{2}$. Příslušné hodnoty $z_{\frac{\alpha}{2}}$ najdeme v tabulce A.

Pokračování příkladu 1:

Prokazovat hypotezu majoritní pravděpodobnosti, tj. $p_+ > \frac{1}{2}$, zřejmě nemá vůbec význam. Je však zajímavá hypotéza symetrie proti alternativě: názor na zvýšení dominuje nad názorem na snížení pracovní spokojenosti.

Máme tedy: $H_0 : p_+ = p_-$, $H_1 : p_+ > p_-$
Data: $N_+ = 12$, $N_- = 6$, $N^* = 18$

Testování provedeme přímo z tabulky B pro $\alpha = 0{,}05$. Asymetrie je prokázána pouze jestliže $N - N_+ = N_-$ je menší než 5 nebo rovno 5. To není splněno: $N_- = 6 > k_{18} = 5$. A tedy ani hypotéza asymetrie směrem k názoru o zlepšení není statisticky prokázána.

B) McNemarův test

McNemar navrhl test vhodný ke sledování vlivu faktoru v čase a jeho efektu na zlepšení nebo zhoršení (resp. přijetí či ztrátu určitého názoru, vlastnosti ap.). Tento test odpovídá znaménkovým datům. McNemarovo uspořádání dat je dáno čtyřpolní tabulkou.

Z toho vzniká případ znaménkových dat takto:

$$N_+ = D , \quad N_- = A , \quad N_0 = B + C$$

McNemar navrhl test pro hypotézu, že ve skupině N osob se počet lidí majících danou vlastnost nezměnil.

1. *Formulace hypotézy* H_0 *a* H_1
   $H_0 : p_+ = p_-$, $H_1 : p_+ \neq p_-$ (dvoustranná alternativa)
2. *Data:* A, B, C, D uspořádaná jako v tabulce 2.

Tabulka 2. Uspořádání dat pro McNemarův test

		po skončení období		
		nemá vlastnost V	má vlastnost V	
na začátku období	má vlastnost V	A	B	A + B
	nemá vlastnost V	C	D	C + D
		A + C	B + D	N

3. *Intuitivní model:* Jestliže platí H_0, pak počet lidí, kteří změnili názor jedním směrem, musí být zhruba stejný jako počet těch, kteří změnili názor druhým směrem. Je-li rozdíl obou skupin velký (bez ohledu na znaménko), pak odmítáme H_0.

4. *Testové kritérium:*

$$X^2 = \frac{(A - D)^2}{A + D} = \frac{(N_+ - N_-)^2}{N_+ + N_-} \tag{9}$$

má rozložení chí-kvadrát s jedním stupněm volnosti. Je-li větší než kritická hodnota odpovídající zvolenému α, pak zamítáme hypotézu H_0 a přijímáme závěr o účinnosti vlivu zkoumaného faktoru.

Pro malé výběry se doporučuje modifikované testové kritérium s korekcí na spojitost:

$$X^2 = \frac{(|A - D| - \frac{1}{2})^2}{A + D} \cdot \tag{10}$$

Poznámka k testu: Testové kritérium (9) má rozložení chí-kvadrát pouze asymptoticky, tj. pro velký počet pozorování, přičemž *počet pozorování zde není N, ale* $A + D$. Tento test není nic jiného než úprava normální (Laplaceovské) aproximace znaménkového testu, resp. X^2 test pro symetrii. Aplikace tohoto testu je tedy možná pouze tehdy, je-li platná normální aproximace pro relativní četnosti $\frac{A}{A + D}, \frac{D}{A + D}$. I když se v literatuře jako dolní hranice pro aplikabilitu testu uvádí počet pozorování $A + D > 10$, doporučujeme používat tento test až nad 30. Pokud jsou k dispozici tabulky znaménkového (resp. přesného binomického) testu, doporučujeme používat tyto přesné testy, pokud to rozsah tabulek dovolí.

Test je ovšem možno použít i pro jiný typ dat, například pro závislé výběry (porovnání názoru muže a ženy v manželských dvojicích, porovnání názorů otce a syna, generálních ředitelů a výrobních ředitelů, řidiče a závozníka atp.).

C) Chí-kvadrát test symetrie znaménkových dat

Stejný test můžeme odvodit jako chí-kvadrát test dobré shody s modelem symetrie.

1. *Formulace H_0 a H_1*

$$H_0 : p_+ = p_- = a\,, \quad p_0 = 1 - 2a \quad (0 \leqq a \leqq \frac{1}{2}) \quad H_1 : p_+ \neq p_-$$

2. *Data:* N_+, N_0, N_-
3. *Intuitivní model:* N_+ musí být přibližně stejné jako N_-,

tj. $N_+ \doteq N_- \doteq \dfrac{N_+ + N_-}{2}$

4. *Testové kritérium*

(11) $$X^2 = \frac{(N_+ - N_-)^2}{N_+ + N_-}$$

má asymptoticky rozložení *chí-kvadrát s jedním stupněm volnosti*. H_0 zamítáme, jestliže hodnota testového kritéria je větší, nebo rovná kritické hodnotě chí-kvadrát pro dané α.

Příklad 4

Preference čtenářů, pokud jde o typ článků o sportu:
„+" = preference rozhovoru se sportovcem
„ 0 " = má rád obojí
„—" = preference reportáží ze zajímavého sportovního podniku
$N_+ = 326$, $N_0 = 423$, $N_- = 273$

O hypotéze majoritní pravděpodobnosti nemá význam statisticky uvažovat. Hypotéza asymetrie je však velmi zajímavá. Test pro takto velký soubor provedeme pomocí testu chí-kvadrát. Vzhledem k tomu, že nevíme předem nic o možném směru preferencí, vyhovuje nám dvoustranná alternativa $p_+ \neq p_-$.

Testové kritérium: $X^2 = \dfrac{(326 - 273)^2}{326 + 273} = \dfrac{2809}{599} = 4{,}689.$

Kritická hladina chí-kvadrátu pro $\alpha = 0{,}05$ je 3,84.

Vzhledem k tomu, že $X^2 = 4{,}689 > 3{,}84$, odmítáme hypotézu symetrie. Pro interpretaci zřejmě přijmeme výsledek převahy přání rozhovoru nad reportáží.

Rozšíření hypotézy symetrie pro znaky s odstupňovanými hodnotami

U některých ordinálních dat se vyskytuje polarita v tom smyslu, že jedny kategorie mají záporný a druhé kladný význam. Znak přitom buď má, nebo nemá neutrální kategorii. Kladné i záporné hodnoty jsou ordinálně odstupňovány. Jejich existence bude nejlépe patrná z příkladů.

Příklady

a) Ve výzkumu čtenářské obce časopisu klademe otázku: „Dáváte přednost sportu nebo kultuře?" Odpovědi jsou:
   1. dávám jednoznačně přednost sportu
   2. dávám spíše přednost sportu
   3. dávám spíše přednost kultuře
   4. dávám jednoznačně přednost kultuře

   Znak nemá neutrální kategorii. Kladnou orientaci můžeme volit libovolně. Pochopitelně je možné stupnici na obou stranách zjemnit.

b) Ve stejném výzkumu klademe otázku: „Dáváte přednost rozhovoru se sportovcem nebo popisu sportovní události?"
   Odpovědi:
   1. dávám jednoznačně přednost rozhovoru
   2. dávám spíše přednost rozhovoru
   3. nemám přednostní volbu
   4. dávám spíše přednost popisu sportovní události
   5. dávám jednoznačně přednost popisu sportovní události.

Znak má neutrální kategorii. Opět je možno volit jemnější škálu odpovědí.

Principy odvození testů

Dále uváděné testy vznikly aplikací obecné věty o chí-kvadrátových testech dobré shody [Rao 1965]. Odvození je vedeno takto:

1. specifikujeme hypotézu a její parametry,
2. parametry odhadneme pomocí metody maximální věrohodnosti,
3. spočítáme očekávané četnosti jakožto funkce parametrů,
4. aplikujeme vzorec

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{O_i} \tag{12}$$

O_i = očekávané četnosti v polích
E_i = empirické četnosti v polích

5. počet stupňů volnosti je

$$df = K - 1 - k\,, \tag{13}$$

kde K je počet kategorií znaku a k je počet nezávislých odhadnutých parametrů modelu. Důkazy těchto testů neuvádíme. Spočívají v uvedeném postupu, v nalezení rovnic pro maximálně věrohodné odhady a jejich řešení a v dosazení výsledných hodnot do výrazu pro X^2.

A) Hypotéza symetrie pro ordinální znak s neutrální kategorií

Znak má lichý počet kategorií, z nichž prostřední má neutrální význam.

1. *Formulace hypotéz H_0 a H_1*

   H_0: pro $K = 2k + 1$ kategorií platí

$$\begin{aligned} p_1 &= p_{2k+1} = a_1 \\ p_i &= p_{2k+2-i} = a_i \\ p_k &= p_{k+2} = a_k \end{aligned} \tag{14}$$

   Pro k parametrů platí $a_i \geqq 0$, $\Sigma a_i \leqq \frac{1}{2}$

   H_1: *omnibusová* alternativa, tj. libovolný stav odlišný od H_0.

2. *Data:* četnosti v jednotlivých kategoriích $N_1, N_2, \ldots, N_{2k+1}$, $N = \sum_{i=1}^{K} N_i$

3. *Intuitivní model:* za předpokladu H_0 by měly být *všechny* příslušné dvojice četností symetricky rozložené kolem $(k + 1)$-ní kategorie, tj. čísla N_i a N_{2k+2-i} by měla být (pro všechna i) přibližně stejná

$$N_i \doteq N_{2k+2-i} \doteq \frac{1}{2}\,(N_i + N_{2k+2-i})$$

4. *Testové kritérium:*

$$X^2 = \sum_i \frac{(N_i - N_{2k+2-i})^2}{N_i + N_{2k+2-i}} \tag{15}$$
$$df = k$$

   Testové kritérium je rozloženo asymptoticky jako chí-kvadrát rozložení s k stupni volnosti. Je-li

   $X^2 \geqq$ příslušná kritická hodnota rozdělení chí-kvadrát s k stupni volnosti, zamítneme H_0,

   $X^2 <$ příslušná kritická hodnota, nemáme důvod H_0 zamítnout.

   Pro velmi častý pětihodnotový znak je

$$X^2 = \frac{(N_1 - N_5)^2}{N_1 + N_5} + \frac{(N_2 - N_4)^2}{N_2 + N_4} \tag{16}$$
$$df = 2$$

5. *Rozklad* X^2

Jestliže zamítneme H_0, může nás zajímat, zda k tomu přispěly všechny symetricky rozložené dvojice kategorií, či pouze některá z nich. Statistická teorie [Rao 1965] umožňuje v tomto případě přímý rozklad k stupňů volnosti celkového chí-kvadrátu na k jednotlivých složek. Symbolicky naznačeno

$$X^2\ (df = k) = \Sigma\ X_i^2\ (df = 1) \tag{17}$$

V našem případě

$$X_i^2 = \frac{(N_i - N_{2k+2-i})^2}{N_i + N_{2k+2-i}} \tag{18}$$

je rozloženo asymptoticky jako chí-kvadrát rozložení s jedním stupněm volnosti. Zamítneme-li H_0, můžeme dále zkoumat jednotlivé sčítance a porovnávat je s kritickou hodnotou rozložení chí-kvadrát s jedním stupněm volnosti. Jestliže je i-tá složka významná, pak příslušná dvojice polí významně přispívá k zamítnutí H_0.

V tabulce *E* uvádíme pro snadnou dostupnost kritické hodnoty chí-kvadrátu, které potřebujeme pro naši úlohu.

Tabulka E. Kritické hodnoty chíkvadrátové distribuce pro praktické úkoly testování symetrie

Počet stupňů volnosti	Kritická hodnota pro α				
df	0,1 (10 %)	0,05 (5 %)	0,01 (1 %)	0,005 (0,5 %)	0,001 (0,1 %)
1	2,71	3,84	6,63	7,88	10,8
2	4,61	5,99	9,21	10,6	13,8
3	6,25	7,81	11,3	12,8	16,3
4	7,78	9,49	13,3	14,9	18,5
5	9,24	11,1	15,1	16,7	20,5
6	10,6	12,6	16,8	18,5	22,5
7	12,0	14,1	18,5	20,3	24,3
8	13,4	15,5	20,1	22,0	26,1
9	14,7	16,9	21,7	23,6	27,9
10	16,0	18,3	23,2	25,2	29,6

Příklad 5

Rozložení, které budeme analyzovat nyní, vzniklo z odpovědí na otázku: „Jakému typu článku dáváte přednost: portrétu našeho sportovce nebo portrétu zahraničního sportovce?" Rozložení četností odpovědí je dáno v tabulce 3:

Tabulka 3. Preference typů článků o sportu

Kategorie (skóre)	jednozn. přednost (— —)	spíše přednost (—)	těžko rozhodnu (0)	spíše přednost (+)	jednozn. přednost (+ +)	
Portrét našich sportovců	22 (5%)	55 (13%)	252 (58%)	72 (17%)	31 (7%)	Portrét zahraničních sportovců

Zajímá nás hypotéza symetrie: $p_{--} = p_{++}$ a $p_{-} = p_{+}$ pro otázku preference dvou pólových stimulů. Počet kategorií je lichý, znak má střední (neutrální) hodnotu, která ovšem k symetrii nepřispívá. Percentuální rozložení naznačuje mírnou preferenci prvého stimulu. Je však tento fakt statisticky prokazatelný?

Uvedeme nejprve heuristický postup:

1. zvolíme $\alpha = 0.05$
2. za předpokladu symetrie očekáváme
   obsazení 2. a 4. kategorie $= \frac{1}{2}(55 + 72) = 63.5$
   obsazení 1. a 5. kategorie $= \frac{1}{2}(22 + 31) = 26.5$

Obsazení střední kategorie je pro symetrii nezajímavé.
Tabulka očekávaných četností
26.5, 63.5, 252, 63.5, 26.5

3. podle obecného vzorce

$$X^2 = \frac{(22 - 26.5)^2}{26.5} + \frac{(55 - 63.5)^2}{63.5} + \frac{(252 - 252)^2}{252} + \frac{(72 - 63.5)^2}{63.5} + \frac{(31 - 26.5)^2}{26.5} = 0.764 + 1.138 + 0 + 1.138 + 0.764 = 3.804$$

Počet stupňů volnosti je $df = 5 - 1 - 2 = 2$, protože ze tří parametrů modelu symetrie (obsazení vnějších, vnitřních a neutrální kategorie) lze jeden odvodit z dalších dvou (proto 2 nezávislé parametry).

4. $X^2 = 3{,}804 < 5{,}99$, proto nemáme důvod zamítnout hypotézu symetrie.

Dosazení do vzorce (13) resp. (14) dává stejný výsledek rychleji

$$X^2 = \frac{(77 - 55)^2}{77 + 55} + \frac{(22 - 31)^2}{31 + 22} = 1{,}528 + 2{,}276 = 3{,}804$$

Příklad 6:

Tabulka 4. Preference článků se sportovní tematikou před jinými články

Články se sportovní tematikou	jednoznačná přednost (— —)	spíše přednost (—)	těžko rozhodnout (0)	spíše přednost (+)	jednoznačná přednost (++)	jiná než sportovní tematika
Absolutní četnosti	92	104	173	70	22	N = 461
(%)	(20%)	(23%)	(38%)	(15%)	(5%)	

$$X^2 = \frac{(104 - 70)^2}{104 + 70} + \frac{(92 - 22)^2}{92 + 22} = 6{,}644 + 42{,}982 = 49{,}626$$

$df = 2$

Číslo je zřejmě statisticky významné jak na hladině významnosti $\alpha = 5$ %, tak $\alpha = 1$ %.

Též obě složky jsou statisticky významné na obou hladinách významnosti: $6{,}644 > 6{,}63$, $42{,}982 > 6{,}63$ (na pravé straně nerovností jsou kritické hodnoty pro $\alpha = 1$ %; je-li významnost prokázaná pro 1 %, platí automaticky také pro 5 %). Ze srovnání obou složek vidíme, že druhá (odpovídající krajním kategoriím) je šestapůlkrát vyšší než první. Zřejmě tedy budou vnější kategorie přispívat k asymetrii podstatně více. *V daném souboru je tedy zájem o sport větší než o jiné věci, ale především je to způsobeno těmi respondenty, kteří čtou články o sportu především* (viz druhá složka X^2).

B) Hypotéza symetrie pro polarizovaný ordinální znak bez neutrální kategorie

Znak má sudý počet kategorií, z nichž polovina je odstupňována jedním směrem a polovina opačným směrem.

1. *Formulace hypotéz H_0 a H_1*

H_0: pro $K = 2k$ kategorií je

$$
\begin{aligned}
p_1 &= p_{2k} = a_1 \\
p_i &= p_{2k+1-i} = a_i \\
p_{k-1} &= p_{k+2} = a_{k-1} \\
p_k &= p_{k+1} = \frac{1}{2} - \sum_{i=1}^{k-1} a_i
\end{aligned} \tag{19}
$$

H_1: obecná *omnibusová* alternativa (jakýkoli jiný stav rozdílný od H_0)

2. *Data:* Pro $K = 2k$ kategorií máme četnosti $N_1, N_2, \ldots, N_{2k}$, $N = N_1 + \ldots + N_{2k}$

3. *Intuitivní model:* Všechny dvojice odpovídajících si četností by měly být přibližně stejné

$$N_i \doteq N_{2k+1-i} \doteq \frac{1}{2}(N_i + N_{2k+1-i})$$

4. *Testové kritérium*

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - N_{2k-i+1})^2}{N_i + N_{2k-i+1}} \tag{20}$$

má asymptoticky chí-kvadrát rozložení s k stupni volnosti. Je-li

$X^2 \geqq$ příslušná kritická hodnota rozložení chí-kvadrát s k stupni volnosti, pak zamítáme H_0.

V opačném případě přijímáme H_0.

Pro velmi častý případ 4 kategorií je

$$X^2 = \frac{(N_1 - N_4)^2}{N_1 + N_4} + \frac{(N_2 - N_3)^2}{N_2 + N_3} \tag{21}$$

$$\mathrm{d}f = 2$$

5. *Rozklad X^2 na složky* může být proveden zcela obdobně jako u testu části A)

$$X^2\,(\mathrm{d}f = k) = \Sigma\, X_i^2\,(\mathrm{d}f = 1) \tag{22}$$

$$X_i^2 = \frac{(N_i - N_{2k-i+1})^2}{N_i + N_{2k-i+1}} \tag{23}$$

X_i^2 jsou opět asymptoticky rozloženy jako chí-kvadrát rozložení s jedním stupněm volnosti.

Jestliže tedy zamítneme H_0, můžeme jednotlivé sčítance dále porovnat s kritickou hodnotou rozložení chí-kvadrát s jedním stupněm volnosti. Je-li i-tý sčítanec významný (překročil kritickou hodnotu), můžeme hledat u příslušné dvojice kategorií důvod asymetrie.

Příklad 7:

Opět uvedeme příklad preference čtenářů, tentokrát pokud jde o typy článků o kultuře: „rozhovor s umělcem" vs „článek o umělci". Výsledky jsou uvedeny v tabulce 5.

Tabulka 5. Preference typů článků o kultuře

Rozhovor s umělcem	jednoznačná přednost	spíše přednost	spíše přednost	jednoznačná přednost	Článek o umělci
Absolutní četnost	35	53	37	12	N = 137
(%)	(26)	(39)	(27)	(9)	

Nebudeme opakovat heuristický postup výpočtů, je stejný jako u příkladu 5.
Dosazením do vzorce (18), resp. (19) dostáváme:

$$X^2 = \frac{(53 - 37)^2}{53 + 37} + \frac{(35 - 12)^2}{35 + 12} = 2{,}8441 + 11{,}255 = 14{,}099$$

V tabulce najdeme kritickou hladinu pro $\alpha = 0{,}05$ a $df = 2$ jako 5,99.

$$X^2 = 14{,}10 > 5{,}99\,,$$

a tedy hypotézu symetrie odmítáme.

Omnibusová alternativa je však příliš široká, a proto se pokusíme získanou informaci specifikovat tím, že oba členy porovnáme s kritickou hodnotou na $df = 1$ (tj. s 3,84).

První člen je nevýznamný: $2{,}84 < 3{,}84$

Druhý člen je však významný: $11{,}26 > 3{,}84$

Vidíme (z významnosti druhého členu), že preference převažuje na stranu rozhovoru s umělcem, a to především u čtenářů s jednoznačně vyhraněným názorem (stejný závěr platí i pro $\alpha = 0{,}01$).

Závěr

V této stati jsme podali přehled základních metod pro analýzu diskrétních znaménkových dat. Ukázali jsme

a) metodu odhadu parametrů,
b) testování hypotézy majoritní pravděpodobnosti,
c) testování hypotézy symetrie,
d) rozšíření hypotézy symetrie na odstupňovaná znaménková data.

Diskrétní znaménková data jsou v analýze sociologických úloh velice četná, a proto jsme chtěli čtenáře v přehledu seznámit jednak s užitečnými běžnými metodami, jednak s rozšířením, které odpovídá zajímavým analytickým úlohám praxe sociologického výzkumu.

Uvedené metody mají tu výhodu, že jsou založeny na kvalitativních úvahách, tj. že nepředpokládají žádnou kvantifikaci, jsou proto neparametrické. Jsme toho názoru, že tam, kde není třeba kvantifikovat, je lépe se tomu vyhnout. Tím nechceme říci, že například běžné pseudokvantifikace očíslováním pořadí kategorií apod. jsou špatné a nevhodné. Takový výzkumný krok je však vždy relativně nebezpečný, neboť může skrýt některé zajímavé vlastnosti dat či jiné vlastnosti naopak nepřiměřeně zvýraznit.

Chtěli jsme tedy upozornit na jednoduché metody. Především jsme však chtěli upozornit na neprávem opomíjené úlohy a na samotný typ dat, jichž se týkají.

Uvedené úlohy používající diskrétních znaménkových dat nejsou pochopitelně vyčerpávající. Jde v nich především o možnost zobecnění hypotézy $p_+ = \frac{1}{2}$ vs. $p_+ > \frac{1}{2}$ na libovolnou hypotézu $p_+ = c (0 < c < 1)$ vs. alternativa $p_+ > c$ (resp. $p_+ < c$ resp. $p_+ \neq c$).

Nezmínili jsme se o hypotéze $H_0 : p_+ = p_- = p_0 = \frac{1}{3}$ proti omnibusové alternativě, vyjadřující libovolný odklon od H_0. Tato hypotéza se testuje jednoduše tak, že spočteme

$$X^2 = \frac{3\,(N_+^2 + N_0^2 + N_-^2)}{N} - N \tag{24}$$

a porovnáme s kritickou hladinou chí-kvadrátové distribuce na dvou stupních volnosti. (Jedná se o pouhou jednoduchou úpravu, která je výsledkem postupu formulí (12) a (13)).

Neuvedli jsme žádné další složitější hypotézy. Domníváme se, že takové úlohy budou postupně zpracovávány tak, jak si je bude vyžadovat výzkumná praxe.

Na závěr chceme upozornit znovu na širší aplikabilitu metod. Uvedené postupy jsou použitelné nejen na znaménková data, ale na libovolné nominální znaky s K kategoriemi, u nichž nás zajímá buď nadpoloviční obsazení jedné kategorie, nebo relace v obsazení dvou kategorií. Rozšíření pak může být formulováno jako párové srovnání v obsazení několika dvojic kategorií. Takové testy však je vhodné formulovat konkrétně k daným úlohám.

Literatura

Bolšev, L. N. — Smirnov, N. V.: *Tablicy matěmatičeskoj statistiky.* Moskva, Nauka 1965.
Janko, J.: *Statistické tabulky.* Praha, ČSAV 1958.
Nayatani, Y. — Kurahara, B.: *A condition for using the approximation by the normal and the Poisson distribution to compute the confidence intervals for the binomial parameter.* Reports of statistical application research, JUSE 11, s. 99—105.
Owen, D. B.: *Sbornik statističeskich tablic.* Moskva, Vyčislitelnyj centr AN SSSR. 1966.
Rao, C. R.: *Linear statistical inference and its applications.* New York, Wiley 1965.
Pirie, W. R. — Hamdan, M. A.: *Some revised continuity corrections for discrete distributions.* Biometrics 28, 1972, s. 693—701.
Van der Waerden, B. L.: *Matěmatičeskaja statistika.* IIL, Moskva 1960.

Резюме

Ржегак Я., Ржегакова Б.: Основные статистические критерии распределения дискретных знаковых данных

В статье приводятся дискретные знаковые данные, получаемые при реализации трихотомной переменной с естественно данной нейтральной категорией и с двумя противоположно ориентированными категориями («+», «—»). Приведены примеры таких данных и основные факторы, способствующие их получению.

Цель статьи — дать обзор статистических методов в качестве практически употребительной сводки. Приведены:

1) метод точечной оценки и метод доверительных интервалов;
2) критерии, порождающие существование преобладающего вероятия ($p_+ > \frac{1}{2}$) (критерий знаков и его нормальное приближение);
3) критерии, проверяющие гипотезу симметрии в отличие от односторонней или двусторонней гипотезы (критерий знаков, МакНемары критерий, критерий хи-квадрат и взаимосвязи между ними);
4) критерии симметрии для распространенных знаковых данных (построение категорий в обоих направлениях «+» и «—») в противовес к омнибусовой (общей) альтернативе (критерий хи-квадрат и его разложение для вклада отдельных пар полей по отношению к симметрии).

В заключение статьи анализируются возможности распространения статистических задач и использования приведенных методов при работе с номинальным типом переменных. В работе воспроизведены общеупотребительные статистические постоянные из таблиц, что должно способствовать использованию приведенных методов в социологических исследованиях.

Summary

Jan Řehák — Blanka Řeháková: Basic Statistical Tests for Discrete Sign Data Distributions

The paper introduces discrete sign data generated by a trichotomous variable with a naturally given neutral category and two differently oriented categories ("+", "—"). Examples of this type of data are presented, as well as basic factors influencing their genesis.

The purpose of the paper is to give a review of usual statistical methods and a working tool for analytic work. The following is discussed

1. Point and interval estimation method for category probabilities;
2. Testing existence of majority probabilities ($p_+ > \frac{1}{2}$): sign test and its normal approximations;
3. Testing symmetry of distribution ($p_+ = p_-$) vs one-sided and two-sided alternatives: sign test, McNemar test, Chi-square test and their interconnections;
4. Testing symmetry of distribution for extended sign data (having rank-ordered categories in both directions "+" and "—") against omnibus alternative: Chi-square test and its decomposition to single degrees of freedom indicating the contribution to asymmetry of each pair of respective categories).

A possible extension of the statistical hypotheses presented as well as a possibility of application of these techniques for nominal data are dealt with. The paper also gives useful statistical constants for an easy and quick application of the given methods.