Poznámky k analýze sociologických dat

JAN ŘEHÁK
Ústav pro filosofii a sociologii ČSAV, Praha

Poznámky shrnuté v této stati nepředstavují úplnou inventarizaci úloh a problémů, s kterými se v běžné praxi analýzy sociologických dat setkáváme; týkají se několika vybraných aspektů statistické analýzy a některých aspektů, které se statistickou analýzou a statistickou inferencí úzce souvisí.

Jde hlavně o to, ukázat na omezení aplikačních možností statistických metod v sociologické praxi. Tato omezení vyplývají jednak z principů metod samých, jednak z kvality dat, s nimiž pracujeme. Statistika je vědní disciplína, která má svou vlastní teorii a metodologii a z těch vyplývají její procedury a přístupy. Teorii samu lze studovat pouze ze speciálních učebnic a se značnými matematickými znalostmi. Určitý pohled na přístupy ke statistické teorii a alespoň letmé a velmi stručné seznámení se s principy může však snad přispět k fundovanějšímu aplikačnímu postřehu.

Statistika stojí v postavení metody analýzy dat; kvalifikovaná aplikace nutně vyžaduje znalosti teorie statistiky na jedné straně a teoretické znalosti aplikační vědy na straně druhé a proto je taková aplikace většinou obtížným problémem. Kromě toho jsou tu stálé vtíravé otázky o kvalitě dat, s nimiž pracujeme. Proto závěry z dat se vytvářejí součinností inference metodologické, sociologické a statistické. Už tato jednota ukazuje, že analýza dat není vůbec jen rutinní záležitostí, ale součástí tvořivé inference vědecké.

1. Typy statistických závěrů

Statistické závěry činíme o *základním souboru, populaci*, tj. o definované množině statistických jednotek (statistických objektů), která je předmětem našeho zájmu. Základní soubor může být dvojího typu:

a) jednak je to přesně vymezený *konečný* soubor jednotek, který je jednoznačně určitelný například tím, že pořídíme jeho seznam (teoreticky je možno vytvořit seznam všech jednotek konečného souboru vždycky, prakticky to ale nebývá možné u větších souborů);

b) jednak je to soubor určený svými vlastnostmi, ale neomezený na daný konečný počet jednotek; je to soubor, jehož velikost je neznáma a někdy ani známa být nemůže — takový základní soubor můžeme nazvat *otevřeným*, respektive hypotetickým nebo *neurčitým*.

Příkladem souboru a) může být soubor všech školních dětí v určitém obvodu Prahy, příkladem b) může být soubor žáků experimentální školy v Bratislavě a všech žáků, kteří se v podobných podmínkách kdy ocitali, ocitají nebo budou ocitat v budoucnu; soubor je tu definován teoreticky a není možno jej určit jednoznačně seznamem, protože nevíme kolik dětí a které děti se budou učit za stejných podmínek v budoucnu.

Statistické údaje však zřídka získáváme od celého základního souboru. Většinou jsou data k dispozici pouze od jeho vybrané části: každou takovou část základního souboru nazveme *výběrovým souborem*, respektive *výběrem*. Tak výběrovým souborem v případě a) je například soubor dětí ze tří vybraných pražských tříd; v případě b) je výběrem ona bratislavská experimentální třída.

Data, která získáváme, mohou být klasifikována také z hlediska jejich proměnlivosti u statistické jednotky. Některé znaky jsou *konstantní*, nemění se podle nálady, okamžitého stavu respondenta, nepodléhají chybám při zjišťování, nemění se v čase. Jiné znaky (a těch je většina) jsou *proměnlivé*. Můžeme je považovat za náhodné proměnné, neboť jejich zjištění je podmíněno spoustou nejrůznějších faktorů — subjektivní pocity, chyba při počítání, nepozornost, špatná interpretace otázky či výběr nežádoucího obsahového prvku otázky. Tato data měříme s chybou. Zatímco pohlaví je neproměnné, věk v průběhu

krátkého časového období se nemění, zaměstnání se nemění apod. Otázka „Jak se cítíte spokojen v zaměstnání?" bude zodpovězena v závislosti na mnoha okolnostech a odpověď se může měnit den ze dne či někdy z hodiny na hodinu, obzvláště tam, kde odpovědi jsou sémanticky neurčité: velmi spokojen, spokojen atd.

Při typologii situací statistických závěrů můžeme vzít v úvahu obě hlediska současně a naznačit je v tabulce 1.

Tabulka 1.

	Data mají konstantní charakter	Data jsou realizacemi náhodných proměnných — měření s chybou
Data se získávají z celého základního souboru (ZS)	1. není problém zobecnění, úlohy se redukují na vhodné vytváření přehledných měr vlastností souborů	2. eliminujeme chyby měření a náhodnost při získávání dat
Data jsou získávána pouze z výběrového souboru (VS)	3. úlohou je tu zobecnit data z VS na ZS a vyjádřit přesnost zobecnění	4. současně 1. a 3.: tato situace je nejobtížnější a většinou se redukuje na 2. nebo 3.

Otevřený základní soubor je věcí modelu a přijaté konvence. Výběr z takového základního souboru (například experimentální třídu) můžeme považovat za základní soubor s tím, že nás nezajímá jen současný stav, ale proces, resp. výsledek procesu, který vede k současnému stavu. Dostáváme se tím z pole 3 do pole 2. Podle toho, co bylo řečeno o základních souborech, musíme rozlišovat v poli 3 dvě situace — zobecnění na konečný (3a) a na neurčitý (3b) základní soubor. V každém případě pak musíme k úloze přistupovat jinak, tj. tvořit jiné statistické, resp. pravděpodobnostní modely.

Většinou se při analýze dat dostáváme do pole 2 resp. do pole 4 a po redukci do pole 2. To je proto, že na složité otázky sociologického výzkumu odpovídá respondent většinou s možností chyby a subjektivního hodnocení obsahu kategorií odpovědi. Nebude třeba snad argumentovat pro to, že poznání možného zdroje nepřesných odpovědí a informací je velice důležité pro konečnou interpretaci. Některé pomocné „klasifikace" znaků uvádím v dalším paragrafu. (Slovo klasifikace je v uvozovkách, protože jde jen o pomocná třídění, která jsou spíše výhodná než nutná.)

2. Některé pohledy na znaky z hlediska chyb zjišťování

Většina znaků tedy podléhá různým chybám při zjišťování jejich hodnot pro jednotlivce nebo jiné statistické jednotky. Pomíjím zde problémy validity a reliability jako takové a uvádím typy znaků z hlediska chyb, které lze a priori předpokládat a tím i vhodnou metodou měření (či zjišťování) potlačovat na nejmenší míru. Přitom nemám na mysli chyby náhodné, tj. náhodné odchylky od skutečné hodnoty znaku; ty jsou relativně málo nebezpečné, neboť jsou identifikovatelné. (Příklad takové chyby je zjišťování skóre I.Q. testu: při každém zjišťování I.Q. skóre se dá předpokládat, že respondent bude mít poněkud jiný výsledek, přičemž nelze identifikovat žádný faktor, který tuto variabilitu způsobuje.)

Kritéria typů:

a) *Formální vztahy hodnot znaku:* zde rozeznáváme znaky kardinální, ordinální a nominální. Informace čerpaná z těchto vztahů mezi hodnotami znaku je podstatná pro vytváření nejrůznějších populačních měr. Obzvlášť obtížné je zpracování ordinálních znaků, tj. znaků, jejichž hodnoty jsou uspořádané kategorie (obvyklé „kvantifikace" pořadovým číslem odpovědi v uspořádání nejsou považovány vždy za vyhovující); to plyne jednak z faktu, že není známa vzdálenost kategorií na předpokládané škále a dále, že odpovědi bývají často sémanticky neurčité a subjektivně interpretovatelné respondentem.

b) *Přímé zjišťování* × *konstrukce znaků* (složené znaky, škály). Je třeba vyjádřit

podiv nad tím, že se málo využívá kumulativních škál, které vznikají jako syntéza odpovědí na řadu jednoduchých, většinou dichotomických otázek (sémanticky daleko jednoznačnějších a pro respondenta jednodušších). Výsledný kardinální znak pak měří (po pečlivé přípravě a rozboru škály) daný aspekt lépe a umožňuje vyhnout se řadě chyb plynoucích z aspektů dále uvedených.

c) *Znak vztahový* × *znak faktický.* Znaky vztahové odrážejí vztah respondenta k objektu, znaky faktické vypovídají o objektivně zjistitelných faktech, událostech a jevech.

d) *Časový rozměr znaků.* Vzhledem k době realizace předmětu znaku můžeme znaky dělit na minulé, přítomné, budoucí. (Více o c) a d) viz Hübschmanová–Řehák [15].)

e) *Vzdálenost subjektu a objektu:* výpověď o sobě, o nejbližším okolí, o širším okolí, o vzdálených objektech.

f) *Předmět znaku* naznačuje dělení podle typu objektu: osoby, předmětu, událostí, abstraktní entity.

V různých kategoriích a kombinacích kategorií vznikají chyby a zkreslení při odpovědích zcela jiným způsobem, který, podle mého soudu, může být předvídán; tím se může zamezit nevhodné formulaci otázky nebo způsobu zjišťování. Toto rozdělení vede také jako pomocné hledisko k lepší a opatrnější interpretaci dat. Je zřejmé, že je tu mnoho dalších faktorů, které musíme brát v úvahu při eliminaci zkreslených výpovědí nebo chyb měření – je to chování tazatele, časové a prostorové vlivy, prostředí, předmět výzkumu apod. Zdrojů chyb je tedy mnoho. Úkolem metodologie by měla být snaha modelově formulovat jednotlivé situace a hledat cesty, jak řešit problém výskytu a eliminace těchto chyb. V některých případech lze takové pokusy nalézt v literatuře.

3. Některé úlohy statistiky

V učebnicích statistiky se většinou dovídáme o různých úlohách, které jsou na statistiku kladeny. Uveďme zde jejich seznam:

1. Sběr a sumarizace dat
2. Plánování statistických experimentů a výběrových šetření
3. Měření variability dat; vysvětlování této variability
4. Odhad charakteristik konečných i otevřených populací – odhad parametrů pravděpodobnostních modelů
5. Konstrukce a hledání měr přesnosti odhadových charakteristik – vyjádření přesnosti měření
6. Testování hypotéz o parametrech modelů
7. Měření síly vztahů a závislosti mezi dvěma či více proměnnými kauzálního modelu
8. Problémy klasifikace (diskriminace, identifikace, diagnostika)
9. Úlohy predikce – odhad těch charakteristik modelu, které se týkají budoucnosti
10. Obecná teorie rozhodování za neurčitosti.

Body 2.–10. jsou předmětem matematické statistiky; k řešení těchto úloh potřebujeme vždy sestrojit nějaký pravděpodobnostní model, který je adekvátní situaci a ve kterém už je zahrnut výsledek apriorních teoretických konstrukcí. Tady vstupuje do statistiky teorie nebo alespoň aplikační intuice.

Vzniká tu otázka, kdy vlastně statistika vstupuje do výzkumu. Optimálně by to mělo být již na samém počátku sociologického výzkumu, tj. při práci nad teoretickými východisky a koncepcemi. Z nich totiž vycházejí všechny statistické úlohy, od plánování výběrového šetření až do posledních rozhodnutí nad statistickými daty. Bohužel tomu tak vždy není. Někdy se statistiky používá dokonce až po sběru dat, tj. v době, kdy jsou formulovány úlohy pro analýzu. Nejzazší okamžik vstupu statistiky do výzkumu by měl být okamžik plánování výběrového šetření, resp. statistického experimentu. Už v tomto stadiu lze zahrnout apriorní vědomosti a informaci do formalizovaného modelu a získat tím kvalitnější data.

1. Bodový odhad

Chtěl bych se zastavit u některých nejčastěji se vyskytujících úloh matematické statistiky. První je úloha *bodového odhadu.* Matematický (respektive pravděpodobnostní) model skutečnosti vzniká odvozením z formálních axionů, které jsou překladem teoretických sociologických východisek do

jazyka matematiky (resp. počtu pravděpodobnosti). V modelu, který je odvozován obecně, vystupují neznámé parametry. Model totiž většinou odpovídá více stavům reality – každá hodnota parametru odpovídá určitému stavu reality.

V úloze bodového odhadu jde o rozhodnutí, které hodnoty neznámých parametrů lze přijmout na základě empirických dat jako nejlepší přiblížení se realitě, který stav modelu přijmeme jako vhodně reprezentující skutečnost.

Uveďme velmi jednoduchý příklad: je dán konečný soubor statí jednoho tisíce článků v novinách, který má být předmětem analýzy. Nemůžeme sledovat všechny stati, ale vezmeme jich pouze 200. Máme tedy relaci „výběrový soubor – konečný základní soubor". Chceme vědět, kolik statí se zajímá o danou tematiku. To jest: „klademe každé stati otázku", na niž je možná dichotomická odpověď.

Zjišťujeme hodnotu dichotomického znaku u vybraných statí. Jaké procento statí v celém základním souboru má danou tematiku? Můžeme toto číslo odhadnout na základě pozorovaných empirických dat? Předpokládejme, že jsme provedli náhodný výběr a předpokládejme, že stejnou úlohu řeší tři osoby, A, B, C, z nichž každá si vybere svůj vlastní soubor o 200 článcích. Předpokládejme, že skutečné neznámé procento p v základním souboru je 65 %. Máme dohromady $\binom{1000}{200}$ různých výběrových souborů. Naši tři výzkumníci získali ve výběrových souborech postupně 62,5 %, 78,0 %, 56,3 %. Každý z nich má tedy jinou evidenci z výzkumných dat. První z nich je neznámé hodnotě nejblíže, jenže to neví. Každá ze tří osob přijme jiné rozhodnutí o parametru p. To je zákonité, ale současně to ukazuje na omezení interpretační možnosti celého postupu. Je ovšem přirozené intuitivně očekávat, že při dostatečném množství pozorovaných jednotek výběrového souboru se naše hodnoty budou skutečné hodnotě parametru blížit. Jelikož velká většina výběrů ukazuje na % velmi blízké oněm 65 % v populaci, doufáme, že jsme měli dostatečné štěstí při hře náhody, která vybírala soubor, že *náš* soubor patří k těm „blízkým" a nikoli k té menšině vzdálených.

V případě prostého náhodného výběru statí dostáváme jako nejlepší rozhodnutí (z určitých hledisek) výběrové procento. Ve třech uvedených případech tedy máme tři různá konečná rozhodnutí: 62,5; 78,0; 56,3.

Takový odhad má tu vlastnost, že je *nejlepší nevychýlený*, tj.:

a) Kdybychom provedli všechny výběry o velikosti 200 a spočítali u nich výběrová procenta, pak průměr všech těchto čísel by byl stejný jako skutečná hodnota. Tato vlastnost je důležitá při opakování pokusů. Můžeme očekávat, že průměr více šetření je blízký skutečné hodnotě s daleko větší spolehlivostí než jednotlivá procenta z dílčích šetření. Čtvrtá osoba, znalá všech tří výše uvedených výsledků, může odhadnout procento jako:

$$\frac{62{,}5 + 78{,}0 + 56{,}3}{3} = 65{,}6.$$

Této vlastnosti říkáme vlastnost nevychýlení.

b) Mezi všemi možnými rozhodovacími pravidly, která mají tuto vlastnost, je nejpřesnější, to znamená, že odhadová charakteristika má *nejmenší výběrovou chybu* (koncentruje své hodnoty co nejblíže kolem skutečných hodnot).

Příkladů na takové nejlepší nevychýlené odhady je mnoho (stejně by např. výběrový průměr kardinálního znaku byl odhadem populačního průměru). Poznamenejme ještě, že tato vlastnost je vázána na způsob vybírání, tj. zde na prostý náhodný výběr.

Výhodou bodového odhadu je to, že se s ním dobře pracuje; je to jediné číslo, které může být použito při konstrukci nějakých dalších měr pro srovnání apod. Nevýhodou je, že tu není charakterizována přesnost, s jakou měříme.

5. Intervalový odhad

V postupu *intervalového odhadu* se snažíme nikoli o nalezení optimální bodové odhadové charakteristiky, ale snažíme se vytvářet intervaly, které by se se zadanou spolehlivostí (např. 95 %) pokrývaly neznámou hodnotu parametru. Tím už zavádíme do úvah pojem přesnosti; jestliže interval je (při zadané spolehlivosti) úzký, pak předpokládáme přesnost závěrů, je-li široký, pak je naše informace neurčitější. V našem případě budeme takový interval

vytvářet tak, že jej rozprostřeme symetricky kolem bodového odhadu $\bar{p}$:

$$\bar{p} - d \qquad \bar{p} \qquad \bar{p} + d$$

Šířka d bude zřejmě závislá na přesnosti měření a na spolehlivosti, s jakou chceme náš *konfidenční interval* (interval spolehlivosti) konstruovat. Předpokládejme, že spolehlivost má být 95%; pak

$$d = 1{,}96 \cdot s_{\bar{p}},$$

kde $S_{\bar{p}}$ je výběrová chyba. (Kdybychom chtěli spolehlivost např. 99 %, pak by se koeficient změnil z 1,96 na 2,58 apod.)

Výběrová chyba se vyjádří vzorcem:

$$S_{\bar{p}} = \sqrt{\left(1 - \frac{n}{N}\right) \frac{\bar{p}\,(100 - \bar{p})}{n - 1}}$$

kde N je velikost populace (N = 1000)
n je velikost výběru (n = 200)
p je procento ve výběru (postupně 62,5; 78.0; 56.3).

Pro naše tři výzkumníky dostaneme postupně výpočtem (a po zaokrouhlení) tabulku 2.

Tabulka 2

$\bar{p}$	62,5%	78,0%	56,3%
$1 - \frac{n}{N}$	0,8	0,8	0,8
$\frac{\bar{p}(100-\bar{p})}{n-1}$	11,78	8,62	12,93
s_p^2	9,42	6,90	10,34
s_p^2	3,07	2,63	3,22
$1{,}96\ s_p$	6,02	5,15	6,30
interval spolehlivosti	(56,48;68,52)	(72,85;83,15)	(50,00;62,60)

Každý výzkumník dostává zcela odlišné výsledky, pouze první (aniž o tom ví) pokrývá skutečnou hodnotu. Všichni však doufají, že jejich výsledky jsou správné a mají za sebou spolehlivost 95%.

Jak interpretovat slovo *spolehlivost?* Jestliže vybereme náhodně jeden ze souborů o 200 statích, máme 95% šanci, že neuděláme tímto statistickým postupem chybu; v 95 % všech možných výběrů o velikosti 200 vede tento postup k pokrytí skutečného neznámého parametru. A to je též klíč k volbě tohoto čísla. Ve společenských vědách se většinou rozhodujeme pro 95 %, někdy pro 99 %. Podle problému však ve statistických úlohách může být hladina spolehlivosti volena až na 99,9 %, což je případ některých medicínských úloh nebo situací, kdy se na základě měření rozhodujeme pro velké investice.

Máme přirozeně zájem na tom, aby interval spolehlivosti byl co nejkratší. Platí ovšem, že čím větší spolehlivost požadujeme, tím širší interval dostaneme a naopak. Je tu tedy konflikt mezi oběma hledisky. Jednou krajností je bodový odhad, který má spolehlivost 0 % a na druhé straně je tu 100% spolehlivost, té však odpovídá celá škála možných hodnot, takže takový výsledek je prakticky bezcenný.

Výhodou postupu je to, že je v něm už přímo zabudována přesnost – čím užší interval při dané spolehlivosti, tím méně entropičnosti v rozhodování a tím přesnější informace. Nevýhodou však je právě to, že jde o interval, s nímž nemůžeme snadno pracovat při vytváření dalších měr a při srovnávání různých výsledků. Kromě toho je tu stále základní fakt statistických závěrů – jsou to závěry za neurčitosti. Pouze doufáme, že nám je náhoda při měření příznivá. V dlouhé řadě konstrukcí intervalů spolehlivosti očekáváme 5 % případů nepokrytí. Je-li zvolená spolehlivost dostatečně vysoká, musí každý výzkumník rozhodnout sám.

6. Testování hypotéz

Nejčastější úlohy formulované ve společenskovědním výzkumu jsou úlohy testování hypotéz. To jsou případy, kdy nás nezajímá odhadování hodnoty parametru přímo, ale rozhodnutí, zda lze přijmout ten či jiný výrok o parametrech nebo o rozložení veličin vůbec.

Existuje celá řada typů hypotéz. Tak např. výzkumník uvedený ad 4. si mohl postavit hypotézu, že skutečné procento p = 60 %. Test hypotézy pak mohl být ekvivalentní s konstrukcí intervalů spolehlivosti; jestliže interval spolehlivosti po-

kryje hodnotu 60 %, pak není důvodů hypotézu odmítnout, je-li pak 60 % vně intervalu, považujeme hypotézu za nesprávnou (málo pravděpodobnou). I zde je vidět omezení statistických aplikací. Předpokládejme, že naše tři osoby, každá nezávisle, vyslovily tři různé hypotézy: $p = 60\,\%$, $p = 65\,\%$, $p = 55\,\%$. První a třetí osoba přijímají své hypotézy, které jsou ale odlišné pro tentýž soubor, tj. pro tutéž realitu; druhá osoba odmítá hypotézu (ač hypotetická hodnota je tatáž jako skutečná v souboru).

Další příklady hypotéz:

a) průměrný výsledek bodovaného testu z matematiky je větší u studentů sociologie než u studentů psychologie;

b) existuje průkazná závislost mezi znakem „informovanost o daném objektu" a znakem „postoj k objektu";

c) na základě minulých měření předpokládáme, že se v průměru zlepší počasí a budou létat letadlové linky mezi Prahou a Bratislavou;

d) na základě zjištěných symptomů přijímáme diagnózu D.

Mluvíme tu o statistických hypotézách, tj. o hypotézách týkajících se pravděpodobnostních vlastností náhodných veličin, které měříme a zjišťujeme. Mluvíme-li o statistice, předpokládáme, že překlad z jazyka sociologie do jazyka statistiky, tj. statistická operacionalizace, je už provedena. Konečný výsledek rozhodnutí o veličinách musí být opět převeden zpět do jazyka sociologie.

Mnohdy při testování hypotéz nás výsledky překvapují svou neočekávatelností a zdají se zcela nemožnými — odmítáme hypotézy, které by měly být podle teoretického východiska jednoznačně potvrzeny. Interpretace takovýchto výsledků musí být krajně zdrženlivá a opatrná. Je pochopitelně nutné takový výsledek vysvětlit a na bázi teorie se s ním vyrovnat. Nesmíme však zapomínat, že to může být způsobeno nejrůznějšími faktory a ukvapené závěry o revizi východisek nelze přijmout bez prověření všech kroků ve výzkumu. Uveďme některé možné příčiny:

1. chyby v operacionalizaci teoretických pojmů;
2. špatná konstrukce modelů — tj. špatný překlad ze sociologie do matematiky;
3. chybný plán a realizace sběru dat;
4. špatná volba znaků vypovídajících o objektech; špatná volba *referenčního systému znaků;*
5. nevhodná formulace otázek;
6. sémantická neurčitost otázek — respondent je chápe jinak než výzkumník;
7. špatný zpětný překlad statistického rozhodnutí do pojmového rámce výchozí teorie;
8. je k dispozici příliš málo dat, aby mohly být prokázány formulované hypotézy;
9. data nejsou analyzována podle modelu, který byl aplikován;
10. mohl nastat případ, že právě u této konkrétní hypotézy se projevilo pravděpodobnostní riziko, se kterým pracujeme ve statistice vždy.

Není možné tedy zamítnout škrtem pera východiska, aniž bychom nesledovali celou dlouhou řadu vlivů (z nichž některé zde byly jmenovány), které mohou ovlivňovat celkové závěry a celý proces vědecké inference — nejen její statistickou část.

Statistické testování hypotéz, jako všechny úlohy řešené matematikou, vycházejí z modelu, který je abstrakcí skutečnosti. Z toho plynou další omezení aplikace metody — vždy je zachycen jen některý z aspektů reality.

Shrňme jen stručně některé z možných přístupů k testování hypotéz.

A) Formulujeme hypotézu H a za předpokladu, že platí, odvodíme teoretické pravděpodobnostní chování sledovaných veličin a jevů. Jestliže data, která získáme jsou za této hypotézy málo pravděpodobná, pak hypotézu zamítáme. Tento přístup nazveme přístupem *R. A. Fishera*. Bylo by možno jej charakterizovat jako zeslabenou analogii logického postupu

$$(H \Rightarrow D) \Leftrightarrow (\text{non } D \Rightarrow \text{non } H)$$

(z hypotézy (H) plyne chování dat (D); nechovají-li se data podle předpokladu, hypotéza neplatí).

Uvažujeme vždy riziko α (odpovídající riziku nepokrytí skutečné hodnoty při intervalovém odhadu, např. 5% resp. v jednotkách pravděpodobnosti 0,05), které je stanoveno jako horní pravděpodobnostní hranice proti hypotéze, jestliže je správná. Nastane-li málo pravděpodobný jev, pak buď hypotéza neplatí, nebo nastal zázrak (R. A. Fisher).

B) Přístup *Neyman-Pearsonův* lze charakterizovat dvojicí hypotéz, které jsou postaveny proti sobě. Základní nulová hypotéza (H_0) je testována proti nějaké jiné možnosti — alterativní hypotéze (H_1). Předpokládáme teoreticky dva možné stavy skutečnosti — zde tedy vstupuje teorie ještě o krok dále do statistické procedury — vymezuje možnou alternativu. To umožňuje lepší výběr mezi rozhodovacími pravidly. Je tu možné formulovat hledisko optimálnosti pro takový výběr.

Rozhodovací situace může být naznačena tabulkou

Tabulka 3

Ve skutečnosti platí	Rozhodnutí pro H_0	Rozhodnutí pro H_1
H_0	správné rozhodnutí	chyba 1. druhu
H_1	chyba 2. druhu	správné rozhodnutí

Rozhodovací pravidla jsou volena tak, aby pravděpodobnost chyby 1. druhu nepřevýšila dané číslo (např. 0.05; 0,01; 0.001; značíme ji obvykle α) a přitom aby pravděpodobnost chyby 2. druhu byla co nejmenší (značíme obvykle β). Rozhodnutí je tu závislé na schopnosti přesně formulovat oba modely odpovídající oběma hypotézám, na počtu pozorování a (jako vždy) na náhodě.

C) *Bayesovský přístup* je založen na Bayesově větě (publikované v roce 1763 Thomasem Bayesem). Tento přístup formuluje hypotézy $H_1 \ldots H_k$ a předpokládá, že jsou a priori (před sběrem dat) známy pravděpodobnosti těchto hypotéz; tyto apriorní pravděpodobnosti mívají nejrůznější interpretace a celý tento přístup má mnoho různých škol podle názoru na apriorní pravděpodobnosti. Můžeme snad souhrnně říci, že apriorní pravděpodobnosti odrážejí stupeň našich znalostí a zkušeností o možných jevech, ať už je vyjadřujeme empiricky a na základě dřívějších zkoumání, nebo jako subjektivní názor; mohou reprezentovat také subjektivní důvěru v platnost hypotézy.

Pomocí Bayesovy formule pak opravujeme apriorní pravděpodobnosti na základě evidence ze sebraných dat. Takto získané aposteriorní pravděpodobnosti jsou základem pro statistické rozhodování o hypotézách. Hypotézu, která má aposteriori dostatečně vysokou pravděpodobnost, můžeme přijmout za správnou. Bayesovská analýza dat je technicky značně náročná. Závěry závisí na apriorních pravděpodobnostech a na modelu, který byl zvolen.

D) V případě, že nejsme s to, či nechceme, nebo se neodvažujeme konstruovat matematický model, který umožňuje inferenci o parametrech, můžeme volit *neparametrické techniky*, v nichž jsou předpoklady daleko volnější a nezavádí se parametry, kromě velice obecných (jako je posunutí počátku, změna měřítka apod.). Výhodou těchto technik je, že se obejdou bez modelu; jedna technika zahrnuje daleko více aplikačních situací. Jsou většinou velice jednoduché, snadno pochopitelné a jejich heuristické odvození je jasné. Přesto že jsou „do šíře" aplikabilnější, nelze doporučit jejich použití tam, kde lze sestrojit model. Do modelu totiž vkládáme už apriorní informaci a ta nám umožňuje analyzovat data efektivněji; pro stejnou přesnost závěru je třeba méně pozorování. Chyby druhého druhu jsou menší. Přínos informace skrze model umožňuje snížit entropičnost rozhodovací situace. Také formální podoba neparametrických technik tento fakt potvrzuje. Jsou to většinou metody založené na pořadí nebo na konstatování výskytu nějakých jevů. Kardinální znaky pak pro použití takových testů ordinalizujeme, což je pochopitelně jistou ztrátou informace. Formální stránka, která vede k využití pořadí pozorování však předurčuje neparametrickým technikám velice široké použití ve společenských vědách — mnohdy totiž nejsme schopni získávat kardinální znaky a naše měření jsou pouze pořadového charakteru. (Např. nejsme schopni měřit schopnost žáků, ale jsme schopni je podle schopnosti uspořádat.) Formulace úlohy je obdobná jako u Neyman-Pearsonova přístupu; zavádíme též nulovou a alternativní hypotézu a chyby 1. a 2. druhu. Rozdíl spočívá v tom, že zde nezavádíme parametrické modely.

7. Typy statistických hypotéz

Pro lepší orientaci v různých úlohách testování hypotéz je možné rozčlenit je do

tří nejčastěji se vyskytujících typů. Toto členění je pomocné a má úlohu pouze orientační (z hlediska formálního i terminologického mu lze leccos vytknout). Velká většina hypotéz je v současné době založena na Neyman-Pearsonově přístupu a neparametrických technikách, které jsou jeho rozšířením.

Ve výzkumné praxi se vyskytují nejčastěji tři typy statistických hypotéz:

1. *Hypotézy o stavu populace — testy dobré shody*

H_0 je hypotéza, která říká, že pro danou populaci platí určitý model, resp. určitý výrok o neznámých parametrech. Termín testy dobré shody je tu chápán šířeji než ve statistické literatuře. Jde obecně o tvrzení, že populace je v nějakém daném hypotetickém stavu. Jako příklady mohou sloužit:

a) veličina má normální rozložení;

b) víme, či předpokládáme, že veličina má normální rozložení a H_0 specifikuje, že má průměr 38,5;

c) průměrné skóre I.Q. testu v dané konečné populaci je $\bar{X} = 70$ bodů (ze 100 možných).

V případě c) např. můžeme formulovat řadu alterativních hypotéz:

$H_1 : \bar{X} \neq 70$ $\quad$ $H_4 : \bar{X} > 70$

$H_2 : \bar{X} = 80$ $\quad$ $H_5 : \bar{X} < 70$

$H_3 : \bar{X} = 62$ $\quad$ $H_6 : \bar{X} < 50$

a mnoho dalších; výběr z nich záleží na situaci a teorii, kterou vkládáme do procesu inference.

Nulová hypotéza může být specifikována na rozložení četností, průměr, procento, rozptyl, medián, symetrii apod.

2. *Hypotézy o srovnání dvou nebo více populací — testy homogenity*

V těchto testech se promítá do statistiky srovnávací metoda. Těmito testy se srovnávají dvě nebo více populací, resp. některé rysy populací. Tak např. je možno porovnávat rozložení četností, aritmetické průměry apod.

Nulová hypotéza bývá formulována obvykle jako hypotéza homogenity, tj. hypotéza, že soubory jsou z daného hlediska statisticky podobné (je možno je smíchat a výsledné rozložení, či zkoumaná vlastnost je stejná pro celý soubor právě tak, jako pro jeho původní části). Alternativní hypotézy specifikují rozdílnost: například rozdílnost průměrů nebo jejich uspořádání atd.

3. *Testy o struktuře vztahu mezi proměnnými — testy závislosti*

V tomto případě (který by bylo možno obecně zahrnout do první kategorie) je předmětem rozhodovacího procesu pouze jedna populace, ale více proměnných, jejichž vztah nás zajímá.

Nulová hypotéza je většinou formulována jako nezávislost znaků. Alternativní hypotéza pak je specifikována kauzálním modelem, který nás zajímá, tj. strukturou vztahů, které chceme prokázat statisticky.

Je přirozené, že existuje řada dalších typů hypotéz a testů pro ně a že toto dělení je jen zcela hrubé a pomocné. Přesto může být snad užitečné při formulaci statistických úloh a při následné interpretaci.

8. Závěr

Tyto poznámky neměly být receptem jak analyzovat data, ale měly čtenáře seznámit s tím, jaké úlohy před námi ve statistice stojí, jaké požadavky můžeme formulovat a hlavně naznačit, s jakými omezeními při použití musíme počítat, jaká omezení máme. Jsou do jisté míry reakcí na současný stav aplikace statistických metod v sociologickém výzkumu . Musím tu varovat před přespřílišným a hlavně nekvalifikovaným používáním testů a jiných procedur matematické statistiky. Nekvalifikovaná aplikace udělá obvykle více škod než užitku. Domnívám se, že aplikace matematické statistiky a matematiky vůbec má velký význam pro analýzu dat i pro modelové účely a výstavbu dílčích teorií, že se bez nich společenská věda, která chce kvalifikovaně hodnotit jednotlivé měřené aspekty skutečnosti, neobejde; soudím však, že je třeba postupovat opatrně a odborně. Chtěl bych varovat před magií čísel a před absolutizováním četnosti jako důkazového materiálu. Doufám, že jsem uvedl dost různých okolností, které takovýto přístup zpochybňují.

Chceme-li docílit dobré aplikace statistických metod a zasvěcené analýzy dat, je nutné statistiku studovat delší dobu a pravidelně, a to nejen její metody, ale také a především její metodologii, principy, východiska, její filosofii a také její limity. Jen tak je možno se vyhnout nepříjemným chybám.

Literatura

[1] Bernstein A. L.: *Spravočnik statističeskich rešenij*, Statistika, Moskva 1968
[2] Cyhelský V., Novák I.: *Statistika I*, SNTL 1967
[3] Cyhelský V.: *Statistika v příkladech*, SNTL 1967
[4] Čermák V.: *Statistika II*, SNTL, Praha 1968
[5] Diamond S.: *Mir věrojatnostěj*, Statistika, Moskva 1970
[6] Edwards A. L.: *Statistical Methods for the Behavioral Sciences*, Holt, Rinehart and Winston, New York 1966
[7] Ezekiel M., Fox K. A.: *Metody analiza korreljacij i regressij*, Statistika, Moskva 1966
[8] Fisher R. A.: *Statistical Methods and Scientific Inference*, Oliver and Boyd, London 1966
[9] Freeman L. C.: *Elementary Applied Statistics*, Wiley, New York 1965
[10] Guilford J. P.: *Fundamental Statistics in Psychology and Education*, Mc Graw-Hill, New York 1956
též: *Podstatowe metody Statystyczne w psychologii i pedagogice*, Warszawa, PWN 1960
[11] Hacking I.: *Logic of Statistical Inference*, Cambridge University Press 1965
[12] Hájek J.: *Teorie pravděpodobnostního výběru s aplikacemi na výběrová šetření*, NČSAV, Praha 1962
[13] Hillebrandt H. F.: *Elementárna statistika pre psychológov, sociológov a pedagógov*, SPN, Bratislava 1968
[14] Hübschmannová M., Řehák J.: *Etnikum a komunikace*, Sociologický časopis č. 6/1970, s. 548—559
[15] Josífko M.: *Pravděpodobnost a matematická statistika pro biology* — (skriptum pro přírodovědeckou fakultu), SPN 1969
[16] Kyburg H. E., Jr., Smokler H. E., eds.: *Studies in Subjective Probability*, Wiley, New York 1964
[17] Malý V.: *Statistické metody ve společenských vědách*, (skriptum pro pedagogickou fakultu), SPN, Praha 1970
[18] Maxwell A. E.: *Analysis Qualitative Data*, Methuen, London 1961
[19] Mittenecker E.: *Plánování a statistické hodnocení experimentů*, SPN, Praha 1968
[20] Moroney M. S.: *Facts from Figures*, Pelican Book 1969
[21] Mosteller R., Rurke R., Thomas G.: *Věrojatnost*, Mir, Moskva 1969
[22] Mosteller F., Tukey J. W.: *Data Analysis, Including Statistics*, in G. Lindley, E. Aronsons (eds): *Revised Handbook of Social Psychology*, 1969 (p. 180—203)
[23] Neyman J.: *First Course in Probability and Statistics*, Holt, Rinehart and Winston, New York 1950
též ruský překlad: *Vvodnyj kurs teorii věrojatnostěj i matematičeskoj statistiky*, Nauka, Moskva 1968
[24] Ostle B.: *Statistics in Research*, The Iowa State University Press 1966
[25] Reichman W. J.: *Use and Abuse of Statistics*, Pelican Book 1968
[26] Řehák J.: *Počet pravděpodobnosti a teoretické rozdělení četnosti*, v: Lamser V., Růžička L.: Základy statistiky pro sociology, Svoboda 1970
[27] Savage I. R.: *Statistics: Uncertainty and Behavior*, [*Hougton Miffin Comp.*,] Boston 1968
[28] Urbach V. J.: *Biometričeskoje metody*, Nauka, Moskva 1964
[29] Veněckij J. G.: *Variacionnyje rjady i ich charakteristiki*, Statistika, Moskva 1970
[30] Yattes F.: *Vyboročnyj metod v perepisjach i obsledovanijach*, Statistika, Moskva 1965
[31] Yule U. G., Kendall M. G.: *Wstep do teorii statistiky*, PWN, Warszawa 1966

Резюме

Ржегак Я.: Заметки к анализу социологических данных

Статья занимается некоторыми аспектами анализа данных и кратким списком некоторых задач, возникающих при анализе данных. Речь идет прежде всего об аспектах методологических и статистических, каким образом их рассматривает социологическая методология.

В первой очереди обсуждаются типы статистических, нас интересующих выводов, во-первых по типу данных, которые есть в распоряжении — измерение с ошибкой и без ошибки, данные всей основной совокупности или ее части — во-вторых по типу основной совокупности — конечной и открытой. Ошибки при сборе данных возможно элиминировать при помощи подходящей методологической подготовки исследования и посредством анализа измерительного аппарата, которым является часто вопрос анкеты. Для этой цели намечены некоторые классификации знаков, оказавшихся на практике автора выгодными для априорной оценки анкеты и для самого анализа данных.

В следующей части дается список главных статистических задач и более подробно обсуждаются три основные задачи статистических выводов: точечная оценка, интервальная оценка и проверка гипотез. В части о проверке гипотез кратко намечены основные методологические приемы к проблеме: Р. А. Фишера, Нейман-Пирсона, байесовский и непараметрический. В конце даются три ситуации проверки гипотез, пожалуй, чаще всех появляющихся в социологической практике; их указание должно также служить вспомогательным и ориентировочным средством в заданиях исследовательской деятельности.

В статье подчеркиваются те аспекты отдельных приемов и методов, которые определяют их ограниченную применяемость и необходимую осторожность при их применении.

Summary

Řehák J.: Statistical Inference. Comments on the Analysis of Sociological Data

The paper discusses some aspects of data analysis and presents a brief survey of some problems arising in the course of such an analysis. This concerns, first of all, the methodological and the statistical aspects, as viewed by sociological methodology. First of all, the types of statistical conclusions which we are interested in are discussed 1. according to the types of available data (measuring with and without error; data from the entire target population or from its part), 2. according to the type of the target population (finite and open). Errors in data collection can be eliminated by an adequate methodological preparation of the research, as well as by an analysis of the measuring apparatus — which frequently assumes the form of a question included in the questionnaire. For this purpose, some classification possibilities of variables are indicated; in the author's practice, these have proved to be advantageous for the a priori evaluation of the questionnaire, as well as for the analysis of data taking place after their collection.

In the following part, a list of the main statistical problems is presented and three fundamental areas of statistical inference — the point estimation, the interval estimation and the hypotheses testing — are discussed in more detail. In the part dealing with the testing of hypotheses, the basic methodological approaches to the problem (R. A. Fisher's, Neyman-Pearson's, the Bayesian and the non-parametrical) are briefly mentioned. Finally, three situations of hypotheses testing which, in all probability, occur most frequently in sociological practice are introduced. Their description should also serve as an auxiliary and orientation instrument in the tasks formulated during the research activity.

In the paper, stress is laid upon those aspects of the individual approaches and methods which determine their limited applicability and the caution necessary in their application.