O JEDNÉ DOSTI OBECNÉ A VELMI ÚČINNÉ METODĚ STATISTICKÉHO DŮKAZU (ČEHOKOLI)

Jan Klaschka, Výzkumný ústav balneologický, Mariánské Lázně (1986),
Psychiatrické centrum, Praha (1991)

> It might be argued that most researchers have been trained in the ideals of science, and that they try hard to evaluate their results from the point of view of the field as a whole. I am sure that they do, but I am also doubtful about their ability to ignore completely their personal costs in favor of the field. I have seen too many cases of researchers trying first one method of analysis and then another, searching to find the one that will pronounce their results significant.
>
> T. A. Ryan [2]

> Dříve byla situace klinika neutěšená. Hromadil materiál a vyšetření, některá z nich často zbytečně... . Počínal si, jsa sám produktem přírody, nehospodárně jako příroda sama. Nyní pomocí computeru je možno vhodně a správně připravený materiál překvapivě zpracovat.
>
> V. Vondráček [3]

Statistická věda produkuje statistické metody a má vyhraněné metodologické představy o způsobu jejich aplikace. Výzkumník potýkající se s úkolem interpretovat chování empirických dat si vybírá mezi statistickými metodami vhodné nástroje. Pokud je pro jeho styl výzkumné práce metodologický rámec vymezený statistickou vědou příliš těsný, nenalezne pravděpodobně mezi nabízenými metodami přesně to, co potřebuje; protože se bez statistiky neobejde, pracuje pak s tím, co je k dispozici, a mnohdy si ne dokonale vyhovující nástroje amatérsky vylepšuje.

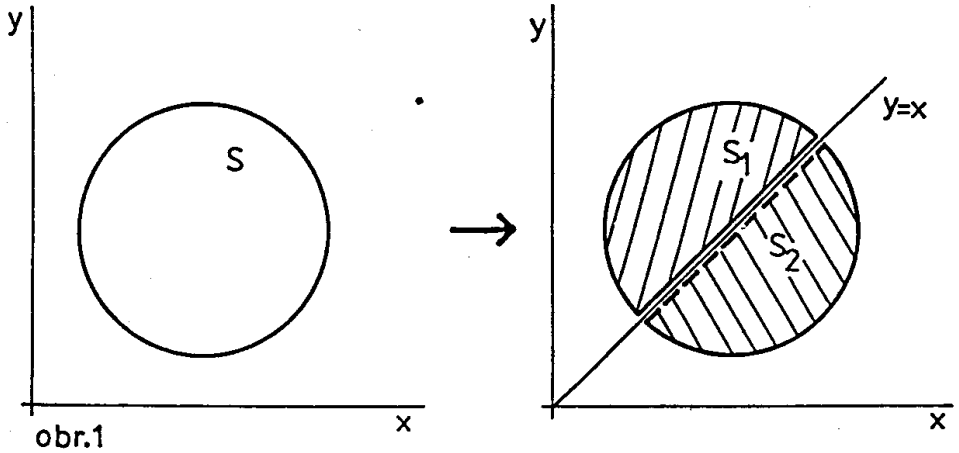
Jedno z takových vylepšení, odpozorované v praxi, inspirovalo autora k vytvoření metody cílené stratifikace, jež je předmětem této práce. Pojednání je názornou ukázkou toho, jak za pomoci matematického myšlení můžeme, všímáme-li si skutečných potřeb výzkumné praxe, věnovat nástroje šité na míru i stylové kategorii výzkumníků statistickou vědou tradičně opomíjené.

Stylovou kategorii, o niž se nám jedná, můžeme vymezit následovně: Vyjádříme hodnotami od 0 do 1 na spojité škále, v jaké míře výzkumník respektuje výsledky statistické analýzy. Maximálně možná hodnota 1 přísluší výzkumníkovi na statistické analýze tak závislému, že teprve nad výsledky výpočtů poznává, co a proč zkoumal. Druhá extrémní hodnota - 0 - patří naopak výzkumníkovi, který *jakékoli* výsledky statistické analýzy pokládá za potvrzení svých pracovních hypotéz. Pro styl, který je předmětem našeho zájmu, jsou pak charakteristické hodnoty uvedeného ukazatele cca 0,1. Výzkumník ohodnocený tímto číslem uznává, že závěry výzkumu se musejí se statistikou shodovat; protože však statistiku jakožto pouhý nástroj cele podřizuje hlavním cílům výzkumu, nepřizpůsobuje zásadně závěry práce výsledkům statistické analýzy, ale jedině naopak.

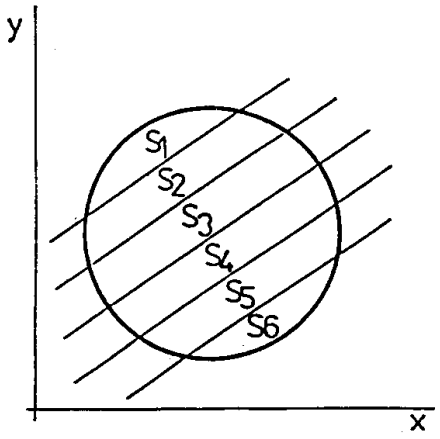
Výzkumník patřící do právě popsané stylové kategorie - říkejme mu v zájmu stručnosti "náš výzkumník" - má při aplikaci statistiky specifické problémy. Dokáže většinou, co potřebuje, ne vždy však snadno. Shromáždění statistických argumentů vyžaduje často úmorné opakování různých variant analýzy, v horších případech je nutno vzpírající se datový soubor pracně "jednotit". Slabinou běžných statistických metod z hlediska našeho výzkumníka je, obrazně řečeno, že tyto metody data především citlivě měří a dávají jim příliš velkou šanci prosadit své sklony. Náš výzkumník jde za svým cílem, o datech potřebuje vědět jen tolik, aby byl schopen překonat jejich případný odpor. Měření mu slouží jen k míření. Ideální nástroje v jeho pojetí nemají povahu měřicí aparatury, ale zbraně.

Tolik o povaze výzkumníka, jehož statistický arzenál chceme obohatit. Nyní konečně k věci.

Myšlenka cílené stratifikace vyrašila ze zeleného stromu života: Cílem jistého experimentu bylo prokázat kladnou korelaci mezi dvěma veličinami - první zpracování dat však kýžený výsledek nedalo. Řešitelé se nevzdali a rozdělili datový soubor S na dva podsoubory S_1 a S_2 způsobem znázorněným na obr. 1. V každém z podsouborů S_1, S_2 pak byl žádoucí vztah prokázán.

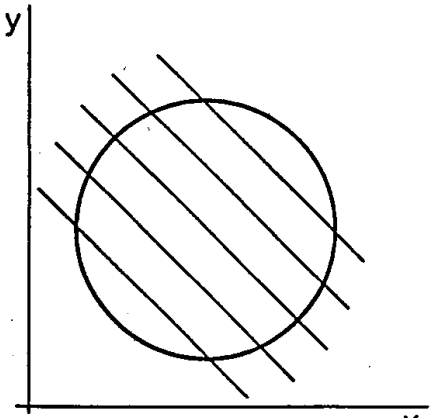


obr.1



obr.2

Statistika může jistě nad tímto příkladem napadnout ledacos - možná také to, že data byla zušlechtěna vynalézavě, ale poloviČatě. Proč nerozdělit soubor rovnou dle obr. 2?



obr.3

Odtud je jen krůček k obecné podobě metody cílené stratifikace: Budiž našim cílem verifikovat nějakou hypotézu (ať už ve smyslu zamítnutí protiřečící nulové hypotézy nebo jinak) o vztazích mezi reálněhodnotovými veličinami $x_1, \ldots, x_N$ a mějme k dispozici měření $\mathbf{x}(s) = (x_1(s), \ldots, x_N(s))$ pořízená na objektech s ze souboru S. Potřebujeme (to je prakticky nejsilnější předpoklad), aby data $\{\mathbf{x}(s); s \in S\}$ tvořila v N-rozměrném euklidovském prostoru R^N "oblak dat".

Základní idea metody cílené stratifikace je: ROZDĚL A PANUJ! Totiž - rozděl oblak dat na části, jejichž tvar a poloha vesměs vyhovují dokazované

hypotéze; význam slova "panuj" je pak jasný.

Formálněji: Budiž $C \subseteq R^N$ množina obsahující všechny přípustné hodnoty dat. Budiž $f: C \to R$ taková spojitá funkce (čtenář nechť laskavě promine případné nedostatky v matematických jemnostech - ve společnosti toho, co je jádrem práce, by snad neměly příliš vynikat), že pro libovolné $K \in R$ oblak dat těsně soustředěný kolem nadplochy $\{\mathbf{x} \in R^N;\ f(\mathbf{x}) = K\}$ výrazně podporuje danou hypotézu. Sestrojíme dostatečně jemný rozklad množiny C na části $C_i = \{\mathbf{x} \in R^N;\ f(\mathbf{x}) \in I_i\}$, $i=1, \ldots, M$, kde $I_1, \ldots, I_M$ je rozklad reálné přímky R na intervaly. Soubor S pak rozdělíme na podsoubory $S_i = \{s \in S;\ \mathbf{x}(s) \in C_i\}$, $i=1, \ldots, M$ a na každý z těchto podsouborů zvlášť aplikujeme zcela běžnou statistickou analýzu.

Na několika příkladech ukážeme, jak popsanou metodou lze i "nevýrazná" data (v R^2) přimět promluvit o různých vztazích - jednoduchých, složitějších i téměř exotických.

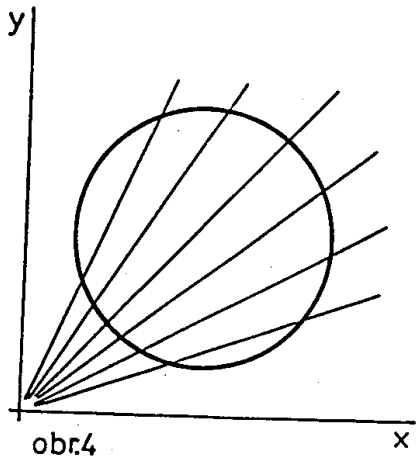
a) Volbou $f(x, y) = x - y$ dojdeme ke způsobu dělení zachycenému na obr. 2.

b) Při $f(x, y) = x + y$ vzniká stratifikace dle obr. 3.

c) Obr. 4 znárorňuje situaci při $f(x, y) = x / y$ (touto volbou usilujeme o shodu s modelem regrese procházející počátkem).

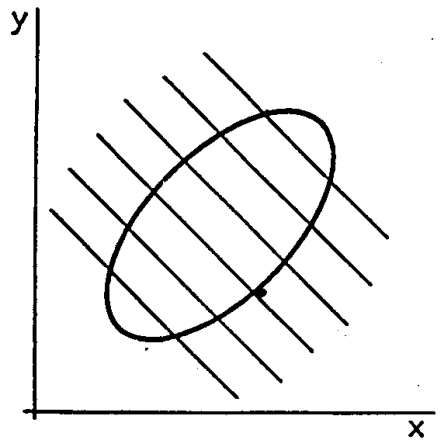
d) Pro $f(x, y) = x \,.\, y$ získáme "krájení" hyperbolami (s odpovídajícími možnostmi interpretace chování dat v podsouborech).

e) K velice zajímavým výsledkům povede volba $f(x, y) = y - A \,.\, \sin(Bx + C)$.



obr.4

Poddají se ovšem nejen data "nerozhodná" ale i "vzpurná" - viz obr. 5.



obr.5

Uživatel metody cílené stratifikace musí vyřešit při každé aplikaci problém, jak zdůvodnit stratifikaci. Uvážíme-li, o jakého uživatele půjde, vidíme, že problém nebude neřešitelný - našemu výzkumníkovi nebude jistě při argumentaci chybět zběhlost ani zaujetí. Přesto, abychom eliminovali nebezpečí, že tato fáze výzkumu pohltí většinu práce ušetřené při výpočtech, vyslovíme užitečné doporučení: Stratifikaci lze prohlásit za rozdělení nehomogenního souboru na několik homogenních. Kritéria homogenity by neměla být, jak známo [1], jen statistická, ale také (a především) všeobecně vědecká - badatel s dostatečnou vědeckou autoritou může tudíž prosadit svá vlastní; v naší situaci je třeba kritérium založit na rozpětí hodnot funkce f.

Pro ilustraci uveďme příklad: Je zkoumán vliv počtu organismů typu X na počet organismů typu Y v určitém prostředí. Předpokládá se vztah "čím více X, tím méně Y", z něhož by měla plynout strategie potírání škodlivých Y-ů nasazováním neškodných X-ů. Člověk míní, příroda mění - experimenty ukáží "čím více X, tím více Y". Mezitím jsou dlouhodobě naplánovány další etapy výzkumu, přiděleny finanční prostředky, rozestavěna nová zařízení pro chov X-ů, navázány patronátní kontakty se šlechtiteli, schváleny zahraniční stáže. Všechmo volá po potvrzení původního předpokladu. V této situaci přicházíme se spásnou myšlenkou, že je třeba analyzovat homogenní podsoubory (soubor jako celek homogenní nebyl, proto je původní analýza chybná). Za homogenní prohlásíme soubor vzorků odebraných v místech s přibližně stejnou hustotou osídlení oběma typy organismů - hustota je vyjádřena součtem počtu X-ů a Y-ů. Pak smíme volit $f(x, y) = x + y$ a jsme přesně v situaci znázorněné na obr. 5, a tím i u cíle.

Pokud si uživatel poradí s argumentací ve prospěch zvoleného dělení, může se už jen těšit z řady příjemných vlastností metody cílené stratifikace:

- Při cílené stratifikaci dělíme, formálně vzato, prostor R^N nezávisle na datech. Na základě alespoň přibližné představy, jak mohou data vypadat, lze skutečně stratifikaci zaručující požadovaný výsledek navrhnout předem. To je výhodné nejméně ze dvou hledisek:

  - Z hlediska ekonomie práce: není nutno data vícekrát analyzovat a zkoušet různé metody; žádoucí výsledek získáme napoprvé a libovolnou v úvahu přicházející (standardní) metodou.

  - Z hlediska metodologické čistoty: Je možno dodržet zásadu "na jedna data použít jen jednou jeden předem určený test". Totéž pravidlo dokonce zakazuje potenciálním kritikům analyzovat data znovu v celku.

- Soubor dat se rozdělí na "vyhovující" podsoubory téměř beze zbytku (nepoddajné okrajové podskupiny lze prohlásit za malé nebo atypické). To je jistě podstatné zlepšení proti podezření mnohdy vyvolávající technice dělení souboru na část vhodnou k analýze a odpad. Navíc jsou naše tvrzení dokazována ne jednou, ale hned mnohonásobně - shodnými výsledky v podskupinách.

- Potřebné manipulace s daty jsou podporovány standardními statistickými programovými balíky (viz např. parametr USE v řídicím jazyce BMDP).

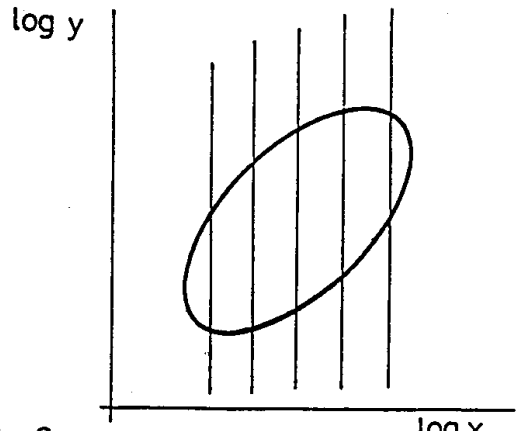
- Metoda je vysoce robustní vůči téměř jakémukoli chování dat.

Závěrem jedno varování: Metoda cílené stratifikace je zbraň. Je jí celkem lhostejné, jakému účelu poslouží. Kdo neví, že má co činit se zbraní (má-li), popř. nezná její možnosti nebo s ní neumí zacházet, obrátí ji snadno proti sobě. Že tuto výstrahu není radno brát na lehkou váhu, dokazuje následující tragický příběh, který se skutečně stal.

Byla jednou snaha prokázat - dodejme, že celkem právem - závislost mezi dvěma veličinami. Byla také nechuť k logaritmické transformaci spolu s nejasným povědomím, že výběrový Pearsonův korelační koeficient pro velice šikmá data není to pravé. Ze všeho toho (a neznámo z čeho ještě) vzešla myšlenka analyzovat data bez transformace a standardně (tj. vypočítat korelační koeficient plus regresní rovnici metodou nejmenších čtverců), ale v "homogenních" podsouborech, jak znázorněno (v logaritmické stupnici) na obr. 6.

Takový postup nemůžeme nazvat jinak než stratifikací cílenou k vědecké sebevraždě. Naštěstí bylo ve vylíčené historii dovršeno tím, že hodnoty jen přibližně měřené veličiny X byly ve většině pásů konstantní...

Poučení: Metodu cílené stratifikace si nemůže dovolit podceňovat ani ten, kdo se s její myšlenkou neztotožňuje.



obr.6

P. S. 1991

1.) Úkaz, jehož aplikací jsem se zde zabýval, se jmenuje Simpsonův paradox. Jednu z jeho podob na ROBUSTu už dříve popsal Dan Pokorný [4] (viz Paradox marginální).

2.) Když tento příspěvek vznikal, bylo mi docela jasné, čí šejdířské prsty je třeba hlídat a kdo je k tomu povolán, kdo by tudíž měl být vyzbrojen a před kým by zbraně měly být utajeny. Několik týdnů po ROBUSTu 86 mi hlásila jedna pražská kolegyně: "Tak už jsme to použili!" Zajímal jsem se, co odhalili, čemu zabránili. Samá voda. "Měli jsme jedny data, trápili jsme se s nima čtrnáct dní a pořád nic nevycházelo. Tak jsme to řízli..."

LITERATURA:

[1] Feinstein, A. R.: Clinical biostatistics XII. On exorcising the ghost of Gauss and the curse of Kelvin. Clinical Pharmacology and Therapeutics *12* (1971), 1003 - 1016.

[2] Ryan, T. A.: "Ensemble-Adjusted p Values": How Are They to Be Weighted? Psychological Bulletin *97* (1985), 521 - 526.

[3] Vondráček, V.: Úvahy psychologicko-psychiatrické. Avicenum, Praha 1981, kap.: Klinická věda a klinická věda psychiatrická, ss. 18 - 34.

[4] Pokorný, D.: Kontingenční paradoxy. In: ROBUST 84, JČSMF, Praha 1984, 92 - 99.