František Fabian

K theorii regrese a vyrovnání rozdělení statistických souborů

Pokroky matematiky, fyziky a astronomie, Vol. 1 (1956), No. 5-6, 559--570

Persistent URL: http://dml.cz/dmlcz/137358

Problémy

1. Dokázat nebo zamítnout tvrzení, že když v pravé polorovině nejsou sing. body, množina singulárních bodů na ose y je řídká.

Poznámka. Když je splněn předpoklad, pak maximální a minimální řešení bodů na ose y vytínají na $x = 1$ disjunktní uzavřené intervaly. Přiřaďme každému styčnému intervalu diskontinua na $x = 1$, $0 \leqslant y \leqslant 1$ příslušný „diadický" bod na $x = 0$, $0 \leqslant y \leqslant 1$. Lze sestrojit diferenciální rovnici? (M. Neubauer).

2. Dokázat, že množina singulárních bodů v E^2 je F_σ. Dokázat, že nemusí být G_δ.

3. Je množina singulárních bodů řídká v E^2? Může protnout osu v intervalu?

4. Úplně popsat strukturu singulárních bodů když je jen „konečný počet singulárních řešení" (jak přesně formulovat?).

5. Jak rozšířit theorii na dynamické systémy v rovině? na soustavy rovnic?

Poznámka. Pokud se týká soustav, fundamentální větu 6 nelze rozšířit v plném znění — [2], a nové práce sovětských autorů. Na druhé straně v. také [9].

Literatura

[1] O. Perron, *Ein neuer Existenzbeweis für die Integrale der Differentialgleichung* $y' = F(x, y)$, Math. Ann., 76, 1915, str. 471.
[2] S. A. Čaplygin, *Sobranije sočinénij I*, Gostechizdat, 1948, str. 347.
[3] G. Peano, *Démonstration de l'integrabilité des équations différentielles ordinaires*, Math. Ann., 37, 1890, str. 182.
[4] E. Kamke, *Differentialgleichungen reeler Funktionen*, Leipzig, 1930.
[4'] E. Kamke, *Differentialgleichungen, Lösungsmethoden u. Lösungen*, Leipzig, 1951.
[5] P. S. Alexandrov, *Úvod do obecné teorie množin a funkcí* (překlad z ruštiny), Praha, 1954.
[6] K. Kuratowski, *Wstęp do teorii mnogości i topologii*, Warszawa, 1955.
[7] G. Birkhoff, *Těorija struktur* (překlad z angličtiny), IIL, 1952.
[8] G. Mie, *Beweis der Integrierbarkeit gewöhnlicher Differentialgleichungssysteme nach Peano*, Math. Ann., 43, 1893.
[9] Fukuhara, Proc. Imp. Acad. Jap., 4, 1928, str. 447.

F. FABIAN

K THEORII REGRESE A VYROVNÁNÍ ROZDĚLENÍ STATISTICKÝCH SOUBORŮ

1

Věda má poskytnout člověku poznání zákonitostí reálného světa, které jej vyzbrojují k tomu, aby prakticky měnil skutečnost. Pouze znalost objektivních zákonů umožňuje předvídat vývoj událostí a tudíž účelně jednat. Vynikající ruský chemik Mendělejev řekl o přírodovědě: „Vědecké zkoumání věci má dva základní, nebo konečné cíle: předvídání a užitek." Předvídání podle Mendělejevových slov „... má ten nejvyšší význam, že ukazuje lidem možnost proniknout v samu podstatu věci. Na druhé straně splnění vědeckých předpovědí by mělo pro lidi velmi malý význam, kdyby nakonec nevedlo k přímému všeobecnému užitku. Tento užitek vyplývá z toho, že vědecké předvídání, které spočívá na studiu, poskytuje lidem takovou jistotu, že mohou usměrňovat podstatu věcí žádaným směrem a dosahovat toho, aby žádané a chtěné se přiblížilo ke skutečnému a neviditelné k viditelnému."

Úkol vědy spočívá v tom, aby za zdánlivým chaosem nesčetných jevů vystupujících na povrch, našla vnitřní zákony, kterým je vývoj jevů podřízen, a aby pronikla do jejich podstaty.

Přírodní zákony jsou konkretním projevem všeobecné zákonitosti, všeobecné nutné vzájemné souvislosti všech věcí, všech jevů přírody. Poznat přírodní jevy — nebo, jak se někdy říká, vysvětlit je — znamená pochopit tyto jevy v jejich vnitřním vzájemném zákonitém vztahu, jinými slovy, vyložit je na základě jejich vlastní zákonitosti, odhalit jejich podstatu.

Existují tedy objektivní zákonitosti přírody a vědecké zákony, odrážející tyto zákonitosti objektivního světa a mající proto rovněž objektivní charakter.

Při řešení otázky zákona zápolí spolu dvě naprosto protichůdné filosofické koncepce: linie subjektivisticko-idealistická a linie materialisticko-vědecká.

Charles Pearson, jeden z největších představitelů machismu, mnohokrát citovaný a kritisovaný Leninem v knize „Materialismus a empiriokriticismus", byl nekompromisním zastáncem prvé filosofické linie. S hlediska této koncepce nemají přírodní a ani společenské zákony objektivní charakter, neexistují nezávisle na duševní činnost lidí, na jejich vědomí, ale jsou člověkem do přírody vnášeny, jsou lidmi tvořeny podle jejich libovůle. Idealističtí filosofové se snaží „dokázat", všemožným způsobem, že v přírodě panuje beznadějný chaos a že záleží na vůli lidí, jak tento chaos „uspořádají". „Rozum diktuje zákony přírodě", říkal Kant, a Pearson tvrdil, že „vědecké zákony jsou daleko více produkty lidského ducha než fakty vnějšího světa", že „dalekosáhlý ráz přírodního zákona děkuje za svou existenci vynalézavosti lidského ducha", že „člověk je tvůrcem přírodního zákona", a že konečně „má mnohem více smyslu tvrdit, že člověk dává zákony přírodě, než tvrdit opak, že příroda dává zákony člověku".

Stanovisko marxismu v této otázce je jiné. Marxismus pojímá zákony vědy — a je lhostejné, jsou-li to zákony přírodní či zákony politické ekonomie — jako odraz objektivních procesů, probíhajících nezávisle na vůli lidí. Lidé mohou odhalit tyto zákony, poznat je, prozkoumat je, upotřebit při své činnosti, využít jich v zájmu společnosti, nemohou je však změnit, nebo je zrušit. Tím méně mohou vytvářet nové zákony vědy.

K vymýšlení „zákonitostí" se dá velmi snadno zneužít matematického aparátu vůbec a matematické statistiky zvláště.

Matematika, zapomene-li se na její empirický původ, se stává aprioristickou a její závěry nabývají idealistického, od skutečnosti odtrženého charakteru; celý proces odtahování matematiky od skutečnosti potom skončí tím, že se matematika postaví proti vnějšímu světu jako něco naprosto samostatného, co z reálného světa nepochází, ale podle čeho se má vnější svět řídit. Následkem tohoto postupu vystupuje potom matematická logika jako všeobecná methoda pro theoretické zpracovávání přírodovědeckých a dokonce i společensko-ekonomických otázek.

Dialektický materialismus nepopírá úlohu a význam matematických method a matematického poznání vůbec v žádném vědním oboru. Naopak. Je však nutno pečlivě rozlišovat mezi vysvětlováním jevů pomocí matematiky, to jest mezi pomocnou úlohou matematiky, a mezi uplatňováním nároku na to, aby se stala všeobecnou vědní methodou. [1]

Je všeobecně známo, že je to především aparát matematické statistiky, který mezi ostatními odvětvími matematiky přichází v nejčastější kontakt s reálným světem. Matematická statistika odráží jistým způsobem objektivní zákony reálného světa. Zákonitosti, které odráží, jsou zákonitostmi pravděpodobnostními. Nelze proto, jak se většinou dělo dříve, a jak se mnohdy děje i dnes, vycházet z pouhého popisu experimentálně získaného materiálu, zůstat na tomto stupni a výsledek tohoto popisu položit roven zákonitosti, bez dalšího výkladu kvalitativní souvislosti. Na druhé straně nutno u matematické statistiky vyzvednout právě to, že má nepoměrně více než ostatní matematické disciplíny možnost upozornit na případnou existenci dosud neznámé zákonitosti na podkladě výsledků zpracování experimentálních dat.

Jedna z forem zpracování experimentálního materiálu je dána v theorii regrese; jde o methodu, pomocí které lze z experimentálních dat získat v jistém smyslu nejlepší formu kvantitativního vztahu mezi pozorovanými veličinami, avšak právě jen pro množinu napozorovaných dat. To nám ovšem bez další indukce neříká ještě nic o tom, je-li tento vztah zákonem nebo ne. Přijali-li bychom tuto formu bez dalšího kvalitativního rozboru za zákon, dostali bychom „zákon" na úrovni popisu získaných dat prostřednictvím smyslových orgánů; tím bychom na zákon povýšili vztahy mezi výsledky našich počitků; tím bychom pomáhali otvírat dveře machismu a subjektivismu. Nestačí proto a nesmí nikdy stačit sebe dokonalejší popis výsledků vykonaných experimentů. Zákonitosti, které vyšetřuje matematická statistika, jsou zákonitostmi hromadných náhodných jevů, které nelze nahradit prostým statistickým popisem. Tento popis nás může na případnou jejich existenci upozornit a je teprve na dalším zkoumání, abychom odhalili — na př. methodami theorie pravděpodobnosti — vnitřní souvislosti zkoumaných vztahů. Ustrnout na konstatování, že křivka, jejíž konstanty jsme určili z experimentálního materiálu, popisuje tento materiál v jistém smyslu lépe než křivka jiná, a je proto „lepším zákonem", je pro odhalování objektivního zákona naprosto nepostačující.

2

S jakými otázkami se setkáváme v theorii regrese? Je to především určování vztahů mezi dvěma (případně více) pozorovanými proměnnými.

a) Máme stanovit jistým, co nejlepším způsobem hodnoty konstant, které se objeví v určitých objektivních fysikálních či jiných zákonech. Tak na př. máme určit hodnoty konstant v rovnici

$$s = s_0 + vt + \frac{1}{2} gt^2,$$

pro dráhu vrženého tělesa, nebo v rovnici

$$y = a \sin (\omega t - \varepsilon),$$

která vystupuje ve velkém množství zákonů, jimž vždy odpovídají jisté hodnoty parametrů a, ω, ε, charakterisující ten který případ. Tak se dá na př. relace užít pro stanovení vztahu mezi časem t (v siderických dnech) a azimutálním úhlovým pohybem vzhledem k ekliptice; jiná skupina zákonů má tvar

$$y = a + \frac{b}{x^a}, \quad \text{nebo} \quad y = ab^{a^x},$$

atd.

b) V převážné většině případů tvar fysikálního zákona neznáme, ba nemáme ani mnohdy možnost učinit si představu o jeho formě. Přesto máme za úkol stanovit z experimentálních dat aspoň aproximační závislost mezi pozorovanými proměnnými. Tak přistupujeme k řešení vždy, když methody vědy, která problém formulovala, nejsou s to stanovit tvar vztahu z vnitřní zákonitosti sledovaného jevu.

V tomto případě máme možnost vyjádřit tvar zákona vždy polynomem, jehož koeficienty určíme známými methodami matematické statistiky. Nebudeme se zde zabývat methodami získávání těchto křivek, ale ukážeme na to, že, jak se zdá, možnost formalistického přístupu k chápání pojmu regrese, kterou získáváme z experimentálních dat, vyplývá již z matematických vět, které ukazují, že za předpokladu, že objektivní funkce vztahu mezi proměnnými (objektivnost musí být opodstatněna mateřskou vědou), na př. $f(x)$ je reálnou spojitou funkcí na $\langle a, b \rangle$, ji lze s libovolnou přesností předem danou aproximovat algebraickým polynomem.

V tomto smyslu možno pokládat problém za rozřešený, těmito matematickými větami [2]:

Věta S. N. Bernsteina: *Je-li $f(x)$ spojitá na $\langle 0,1\rangle$, pak stejnoměrně v x*

$$\lim_{n\to\infty} B_n(x) = f(x),$$

kde polynom

$$B_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} x^k (1-x)^{n-k}$$

nazýváme Bernsteinovým polynomem funkce $f(x)$.

Věta Weierstrassova: *Nechť $f(x)$ je reálná spojitá funkce, definovaná na $\langle a,b\rangle$. Pak pro každé $\varepsilon > 0$ existuje takový polynom $P_n(x)$, že pro všechna $x \in \langle a,b\rangle$*

$$|P_n(x) - f(x)| < \varepsilon.$$

Závěr, který jsme již shora uvedli, vyplývá rovněž jasně z této formulace poslední věty:

Každá reálná, na intervalu $\langle a,b\rangle$ spojitá funkce $f(x)$ může být pokládána za limitu některé stejnoměrně konvergující posloupnosti polynomů.

Je známo, že analogické věty platí na př. i pro trigonometrické vyrovnávací polynomy, v případě periodických jevů, vyjádřených funkcemi spojitými periodickými. Vedle těchto dá se ukázat celá další řada vyrovnávacích funkcí, vhodných pro ten který konkretní případ.

Již z tohoto čistě matematického aspektu plyne, že vyrovnání (pro jehož získání má aparát matematické statistiky velkou řadu method), které dostaneme, nemusí být — bez dalšího kvalitativního ověřování — výrazem zákona reálně existujícího, ale pouze aproximací, i když jakkoli dobrou. Nutno si také dále uvědomit, že z uvedeného vyplývá, že za uvedených předpokladů lze (methodami stanovenými matematickou statistikou) ke dvěma řadám jakýchkoli hodnot vždy nalézt křivku, která je libovolně přesně svazuje, ač ve skutečnosti tato data nemusí nijak vnitřně souviset.

3

Přejdeme k druhému, vcelku analogickému případu. Týká se analytického vyrovnávání rozdělení statistických souborů určitými typy křivek. Nejrozsáhlejší „křivkové schema" podal Ch. Pearson ve snaze sestrojit aparát, kterým by mohl v co nejvíce případech zavést v základních souborech pořádek, to jest nalézt systém křivek a kriterií pro užití každé z nich, z kterého by se k popsání sledovaného kolektivu vždy nějaká hodila. Odvození, které podal Pearson, je čistě formalistické a je na něm zřejmý Pearsonův machistický světový názor: popsat totiž výsledky, které získáme našimi počitky, bez ohledu na zhodnocení vnitřních zákonitostí — v tomto případě pravděpodobnostních — použitím vnějších analogií.

Ukažme stručně postup takového odvození:

Rovnici analytické funkce $y = f(x)$ určíme tak, že vymezíme podmínky, které vyhovují co největšímu počtu skutečně pozorovaných empirických souborů.

Ze „zkušenosti" víme, že „všechny" soubory mají své krajní body s nulovými četnostmi, a že existuje bod, ve kterém je četnost maximální (berouce v úvahu i případná vedlejší maxima). Vyjádřeno matematicky:

$$y(A) = 0, \qquad y'(A) = 0,$$
$$y(B) = 0, \qquad y'(B) = 0,$$
$$y(-a) = \max, \qquad y'(-a) = 0,$$

($-a$ má zřejmě význam modu).

Lze snadno ukázat, že tyto podmínky splňuje rovnice

$$y' = y(x + a). \tag{1}$$

Tato rovnice by však mnoho křivek nezahrnovala a bylo ji nutno zobecnit. Bylo nasnadě vynásobit pravou stranu relace (1) jistým faktorem $\psi(x)$, který by výchozí podmínky nerušil. Za $\psi(x)$ byla zvolena funkce

$$\psi(x) = \frac{1}{b_0 + b_1 x + b_2 x^2}.$$

Výsledná rovnice

$$y' = \frac{x + a}{b_0 + b_1 x + b_2 x^2} y \tag{2}$$

splňuje nejen okrajové podmínky, ale platí i v celém intervalu $\langle A, B \rangle$.

Volba funkce $\psi(x)$ byla podložena pouze analogií s diferenční rovnicí, které vyhovuje hypergeometrické rozdělení; na rovnici možno také hledět jako na zobecněnou diferenciální rovnici, definující Gaussovu funkci.

V závislosti na volbě parametrů, které vystupují ve výsledné diferenciální rovnici, získal Pearson při jejím řešení 12 typů křivek, pro které pak snadno dostal jistá kriteria pro použití každého typu.

Je tedy dán „soubor" (v každém případě vyrovnáváme však výběr, zpravidla náhodný), kde sledujeme empirické rozdělení určitého znaku. Jde o to vyrovnat toto empirické rozdělení (chybně: získat theoretické rozdělení). Je zřejmé, že vyjdeme-li z popisného hlediska, pak

a) musí křivka splňovat uvedenou diferenciální rovnici,
b) vlastnosti výběrových dat musí být zachyceny v konstantách (parametrech),
c) řešení diferenciální rovnice musí co nejlépe aproximovat zkoumaný výběr.

Pearson a celá jeho škola, která svou popisovací tendenci stavěla na nejrozmanitějších druzích všech možných charakteristik, v převážné většině bez pravděpodobnostního podkladu, vyšli z předpokladu, že empirické a „theoretické" rozdělení budou se tím více shodovat, čím více budou mít společných momentů.

Funkce, které získáme řešením diferenciální rovnice (2), závisí na čtyřech parametrech, které zároveň s koeficientem úměrnosti jsou definovány pěti momentovými rovnicemi $\left(\text{položí se } a = \frac{c}{\alpha}, b_0 = \frac{\beta}{\alpha}, b_1 = \frac{\gamma}{\alpha}, b_2 = \frac{\delta}{\alpha}\right)$. Je přirozené, že Pearsonovy křivky by mohly být nahrazeny libovolnou třídou funkcí, majících stejné momenty; jistě bychom získali ještě rozsáhlejší systém, a tím i funkce s větším množstvím parametrů, a tím i větší možnost lepších vyrovnání, jestliže bychom určovali „theoretickou" funkci $f(x)$ daného tvaru na základě (kromě koeficientu úměrnosti) k libovolných parametrů, určovaných z $(k + 1)$ momentových rovnic; při tom momenty jsou

$$m_h = \int_{-\infty}^{\infty} f(x)\, x^h \, \mathrm{d}x, \quad h = 0, 1, 2, \ldots, k. \tag{3}$$

Jelikož můžeme mít vždycky za to [3], že empirická funkce rozdělení $S(x)$ neklesá při rostoucím x pomaleji než geometrická řada, to jest

$$S(x) < B \cdot \mathrm{e}^{-\alpha |x|},$$

kde B a α jsou kladná čísla, potom určujíce „theoretickou" funkci rozdělení z dostatečného počtu momentových rovnic, můžeme s hlediska theorie

pravděpodobnosti dostat libovolně přesné přiblížení k libovolné statistické křivce za jediných předpokladů, že $f(x) \neq 0$ pro ta x, pro která $S(x) > 0$, a že $f(x)$ splňuje podmínku [3]

$$f(x) < B\,\mathrm{e}^{-\alpha|x|}.$$

Vidíme, že jsme získali v podstatě analogický výsledek jako v předchozím odstavci.

Rozeberme alespoň stručně některá další úskalí, která většině odborníků, kteří pak tyto methody aplikují, uniknou anebo kteří si tuto problematiku ani neuvědomí.

Momentová methoda je velmi rozšířený způsob pro zpracovávání napozorovaného materiálu. Je známo, že máme-li dvě identická rozdělení, pak momenty všech řádů jsou stejné. V opačném případě však tvrzení nemusí být splněno. Momentová methoda klade na funkci rozdělení podstatný požadavek, aby totiž její momenty libovolného řádu byly konečné. Přirozeně nelze očekávat, že funkce rozdělení, o kterých předpokládáme, že objektivně existují, budou a priori tento požadavek splňovat, a proto nelze hledanou funkci a priori takovým požadavkem svazovat.

Z toho všeho plyne, že obecně znalost všech momentů neurčuje funkci rozdělení jednoznačně.

Jelikož je možno připustit, že takový případ v praxi málokdy nastane, lze v mnoha případech (avšak velmi obezřetně a s dalším ověřením) rovnosti momentů využívat aspoň k přibližnému srovnávání rozdělení. Upozorněme dále na další možnost přestupků proti realitě při použití uvedené methody.

Z Pearsonovy theorie vyplývá, že momenty skutečného rozdělení jsou pokládány za rovné momentům určeným z výběru; na podkladě této rovnosti se hledá „theoretické" rozdělení. Jinými slovy: to co zjistíme smysly, pokládáme za totožné se zákonem objektivní skutečnosti; podle toho nám tedy má stačit to, co zjistíme počitky, bez hledání další vnitřní souvislosti, to jest bez dalšího zdůvodňování toho kterého nalezeného rozdělení. Rozdíl mezi „výběrem" a „základním souborem", asi stejně jako mezi „počitkem" a „objektivní skutečností" je ignorován.

Pearsonova methodika zkoumání, dodnes ještě často užívaná při aplikaci statistických method, nevyužívá celého tohoto indukčního aparátu matematické statistiky, který, jsa budován na principech pravděpodobnostních zákonitostí, umožňuje činit objektivní závěry z výběru na základní soubor.

Vztahy mezi výběrovými momenty a koeficienty dostaneme integrací rovnic

$$y'(b_0 + b_1x + b_2x^2)\,x^h = y(x + a)\,x^h,\ h = 0, 1, 2, 3,$$

předpokládajíce, že

a) $\int_{-\infty}^{\infty} y\,\mathrm{d}x = 1,$

b) počátek je v aritmetickém průměru.

Je patrno, že v tomto případě vystačíme s prvními čtyřmi momenty. Jelikož řešení diferenciální rovnice (2) má obecný tvar

$$y = c \cdot \mathrm{e}^{\int \frac{x+a}{b_0 + b_1x + b_2x^2}\,\mathrm{d}x}, \tag{4}$$

je typ křivky formálně matematicky závislý na vlastnostech jmenovatele $b_0 + b_1x + b_2x^2$. Snadno se dá ukázat, že všechny tyto křivky lze rozdělit na základě charakteristiky Pearsonovských křivek

$$K = \frac{b_1^2}{4\,b_0\,b_2},$$

což je výraz upravený z diskriminantu rovnice $b_0 + b_1 x + b_2 x^2 = 0$. Podle shora nastíněné methody jsou koeficienty a, b_0, b_1, b_2 funkcemi prvních čtyř momentů, a tedy i koeficient K lze pomocí výběrových momentů vyjádřit. [4]

Shoda „theoretické" křivky určitého typu, tímto způsobem získané, s empirickými daty, nám neříká nic více, než že data získaná výběrem (a výsledky měření nejsou také nic více než výběr) přibližně sledují uvedený typ křivky. O chování sledovaného znaku v základním souboru, který Pearson a všichni, kteří tímto způsobem „hledají zákony", nerozlišují od výběru, neříká nám takto provedený rozbor nic zásadního. Je zřejmé, že tímto způsobem se ke skutečnému zákonu nepropracujeme bez použití dalších úvah; užijeme-li pouze takovéhoto (machistického) způsobu rozboru, pak nám to přímo brání dopracovat se k zákonu, dopracovat se za počitkovou (popisnou) činnost.

Na druhé straně je ovšem pravda, že touto formalistickou popisnou stránkou (výběrových dat) se dospělo k velmi důležitému matematickému aparátu. Bylo proto oprávněné se domnívat, že za Pearsonovou methodou bude skryto hlubší pravděpodobnostní opodstatnění, které nám poskytne reálné kriterium pro aplikaci té které křivky v tom kterém konkretním případě. Bez odhalení této podstaty mohou být křivky nejvýše jen jakýmsi signálem, který upozorňuje na možnost existence objektivního zákona, což již nejednou přineslo velký užitek a usnadnilo další bádání; zároveň však nesmíme zapomínat, že tento postup může vést ke skreslování skutečnosti, ba i k záměrnému zneužívání za účelem „dokazování" předem vykonstruovaného, potřebného závěru.

Pravděpodobnostní zdůvodnění, které ukazuje vnitřní proces děje, na který pak možno tu kterou křivku aplikovat, dokázal pro případ, kdy kořeny jmenovatele v $\psi(x)$ jsou reálné, S. N. Bernstein. Ukázal, že existují reálné pravděpodobnostní procesy, kterým uvedené křivky odpovídají, a tak jejich použití jako matematického vyjádření zákona rozdělení je zdůvodněno tehdy, vyhovuje-li zkoumaný jev pravděpodobnostnímu procesu směřujícímu v matematickém rouše potom k té které křivce. Vedle toho dokázal A. N. Kolmogorov, že lze udat vždy pravděpodobnostní schema, které nás k té které konkretní křivce „Pearsonova systemu" přivede.

Ukažme stručně Bernsteinův postup.

S. N. Bernstein vyšel z obecného Pólyova urnového schematu: Mějme v urně x kuliček bílých a y kuliček černých. Táhněme z urny n-krát po jedné kuličce, při čemž po každém tahu kuličku vrátíme zpět a přidáme ještě Δ ($\Delta = 0, \pm 1, \pm 2, \ldots, \pm k, \pm \ldots$) kuliček téže barvy jako byla kulička vytažená. Ptejme se po pravděpodobnosti, že v těchto n tazích byla vytažena bílá kulička celkem m-krát.

Snadno zjistíme, že hledaná pravděpodobnost bude rovna

$$P_{m,n} = \binom{n}{m} \frac{x^{[m]}\, y^{[n-m]}}{(x+y)^{[n]}},$$

kde

$$r^{[s]} = r(r+\Delta)\ldots[r+(s-1)\Delta].$$

Utvoříme-li podíl $\frac{P_{m+1,n} - P_{m,n}}{P_{m,n}}$ a položíme $\frac{x}{x+y} = \alpha$, $\frac{y}{x+y} = \beta$, $\frac{\Delta}{x+y} = \delta$, můžeme psát

$$\frac{P_{m+1,n} - P_{m,n}}{P_{m,n}} = \frac{Am + B}{Cm^2 + Dm + E}, \tag{5}$$

kde konstanty A, B, C, D, E jsou závislé na konstantách α, β, δ a na celkovém počtu pozorování n.

Vhodným limitním přechodem v relaci (5) dostaneme Pearsonovu diferenciální rovnici. V závislosti na výrazech α, β, δ vede řešení diferenciální rovnice na ten který typ rozdělení. Z uvedeného schematu je zřejmé, že konstanty α, β, δ mají reálný smysl. Tak, jestliže na př. zaměníme vytažené bílé kuličky realisací určitých příznivých podmínek pro individuum jisté kategorie, takových, že je mu umožněno „rozmnožování", to jest přibývání podobných individuí v pozorovaném souboru, a vytažení černé kuličky realisací týchž podmínek pro individua (prvky) opačné třídy, potom počet individuí (prvků) $x + m\,\Delta$ po n-pokusech, při n dostatečně velikém, vyhovuje aproximativně jednomu z možných rozdělení Pearsonova systému.

Z tohoto postupu vyplývá, že existuje vždy souhrn příčin, který vede reálně k zcela určitému typu křivky ze souboru křivek definovaných rovnicí (4). Ze srovnání příčin vedoucích k zjištěnému empirickému rozdělení se souhrnem příčin vedoucích k té které křivce, můžeme usuzovat na rozdělení v základním souboru, odkud byl výběr, který nám poskytl empirické rozdělení, vzat. Pouhá shoda empirického rozdělení s „pearsonovsky" vyrovnanou křivkou nás může případně přivést na dobrou cestu, ale nezdůvodňuje užití té které křivky jako objektivního zákona.

4

Vezměme konkretní případ, na př. rozdělení náhodných chyb, a ukažme postup, jak určit zákon rozdělení, na podkladě vnitřních zákonitostí jevu samého pomocí theorie pravděpodobnosti a matematické statistiky.

Konstatujme nejprve tyto zásadní poznatky:

a) Křivka musí být odrazem chování sledovaného jevu v reálném světě, to jest odrazem souhrnu příčin, majících za následek to které rozdělení.

b) Křivku nutno vyvozovat buď

α) z určitého souhrnu postulátů charakterisujících obecně, pokud možno co nejlépe, chování příslušného jevu v reálném světle, z nichž potom dále pokračujeme deduktivními úvahami, nebo

β) z urnového (případně jiného) schematu, vystihujícího co možná nejvěrněji mechaniku sledovaného jevu při jeho realisaci.

c) Obě methody uvedené sub b) mají, byly-li výchozí předpoklady v obou případech správně vystiženy — vést ke stejnému závěru. Při tom není vyloučeno, že stále dokonalejším poznáváním mechanismu jevů bude možno v tom kterém konkretním případě předpoklady zpřesňovat a na podkladě toho potom získat rozdělení sledovaného znaku v stále dokonalejší formě.

Jak známo, formulují se postuláty pro vyvozování zákona rozdělení náhodných chyb asi takto:

Počet chyb, které padnou do intervalu délky Δx, je úměrný

1. počtu pozorování n,
2. délce intervalu Δx, kde x je velikost chyby,
3. jisté spojité funkci $f(x)$ rozdělení, kde x je opět velikost chyby.

Požaduje se, aby funkce $f(x)$ potom vyhovovala těmto požadavkům:

1. pravděpodobnost malých náhodných chyb je větší než pravděpodobnost větších náhodných chyb (v absolutní hodnotě), to jest hledané rozdělení je klesající funkcí absolutní velikosti chyb,
2. pravděpodobnost realisace náhodných chyb nezávisí na jejich znaménku, to jest náhodné chyby co do absolutní hodnoty stejné vyskytují se stejně často.

Konečně nutno poznamenat, že se předpokládá a skutečnost to potvrzuje, že náhodné

chyby spadají do oblasti hromadných náhodných jevů a vyhovují tedy všem podmínkám pro toto zařazení.

Z těchto předpokladů snadno získáme [5] funkcionální rovnici

$$w(u) \cdot w(v) = C\, w(u + v), C = \text{konst},$$

jejíž řešení je známo. Vezmeme-li v úvahu vyslovené předpoklady a uvážíme-li, že jde o rozdělení pravděpodobností, dostaneme celkem snadným postupem pro hledané rozdělení Gaussovu křivku.

Ukažme, že ke stejnému výsledku dospějeme, postulujeme-li výchozí předpoklady, odpovídající reálnému tvoření náhodných chyb pomocí pravděpodobnostního schematu. Vyjděme z těchto zřejmých předpokladů:

I. každá serie měření některé proměnné je nevyhnutelně doprovázena chybami,

II. každou chybu možno si představit jako výsledek velkého počtu velmi malých a co do absolutní hodnoty prakticky stejných elementárních chyb, jejichž vliv na celkovou chybu se s jejich růstem zmenšuje. Předpokládejme, že elementární chyby jsou nezávislé,

III. každá elementární chyba může být při každém měření realisována se stejnou pravděpodobností pro znaménko „plus" či „minus".

Mějme tedy posloupnost nezávislých náhodných proměnných $\{\xi_k\}$, z nichž každá může s pravděpodobností $\frac{1}{2}$ nabýt hodnotu $+\delta$ nebo $-\delta$; nechť m náhodných proměnných (elementárních chyb) nabude hodnoty $+\delta$, ostatní pak hodnoty $-\delta$. Nechť počet vykonaných měření je n. Velikost celkové chyby bude tedy

$$\eta_n = (2m - n)\,\delta.$$

Vidíme, že η_n nabude každé dané hodnoty, na př. k, pouze při zcela určité hodnotě m. Lze tedy aplikovat binomické rozdělení

$$P(\eta_n = k) = \binom{n}{m} p^m (1-p)^{n-m}, \text{ kde } m = \left(\frac{k}{\delta} + n\right) \cdot \frac{1}{2}, p = \frac{1}{2}.$$

Snadno zjistíme, že

$$P\left[\left(\frac{m}{n} - p\right) \sqrt{\frac{n}{p(1-p)}} < \beta\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta} e^{-\frac{1}{2}t^2}\, dt = \Phi(\beta), n \to \infty. \qquad (6)$$

Vhodnou volbou velikosti elementárních chyb (tak aby $\delta_n \to 0$, jestliže $n \to \infty$) a celkem jednoduchou úpravou výrazu (6) dostaneme

$$P(k < y) \approx \Phi\left(\frac{y}{c}\right)$$

pro n dostatečně velká, při čemž c je jistá konstanta.

Lze konečně ukázat nejobecnější předpoklady, za nichž pomocí theorie pravděpodobnosti zjistíme, že rozdělení náhodných chyb (za předpokladu, že uvedené požadavky odpovídají chování náhodných chyb v realitě) je vždycky aproximativně normální.

Tento problém je beze zbytku vyřešen v theorii limitních vět pro součty nezávislých náhodných proměnných. Reálné předpoklady pro chování náhodných chyb jsou v této formulaci dány tak zvanou Lindebergovou-Fellerovou podmínkou, případně podmínkou Ljapunovovou. [6]

Jestliže předpokládáme (což je celkem reálný předpoklad), že nezávislé náhodné

proměnné (v našem případě elementární chyby) jsou stejně rozděleny, a že mají konečné, od nuly různé rozptyly, pak shora uvedené podmínky jsou splněny, a tedy stejnoměrně v x při $n \to \infty$

$$P\left\{\frac{1}{B_k}\sum_{k=1}^{n}(\xi_k - a_k) < x\right\} \to \Phi(x), \text{ kde } a_k = E\,\xi_k, B_n^2 = \sum_{k=1}^{n} D\,\xi_k,$$

při čemž předpokládáme, že $\{\xi_k\}$ je posloupnost nezávislých náhodných proměnných.

Na základě uvedeného možno učinit tento závěr:

Je-li oprávněn předpoklad, že

A. náhodné chyby lze interpretovat jako náhodné proměnné,

B. že každá chyba je výsledkem součtu nezávislých elementárních chyb, majících neznámou, stejnou distribuci, potom rozdělení takto vzniklých chyb může vést pouze na rozdělení normální, při čemž máme zaručeno, že tento závěr vyplývá z analysy vnitřní podstaty sledovaného jevu.

Při tom můžeme poznamenat, že uvedený závěr platí, i když připustíme, že uvažované distribuce elementárních chyb nejsou stejné, a že místo toho náhodné proměnné splňují Lindenbergovu-Fellerovu podmínku (což lze opět očekávat i pro elementární chyby). Rovněž tak místo „nezávislosti" můžeme připustit i „malou závislost", jistým způsobem definovanou pro elementární chyby. Ani v tomto případě nepozbývá závěr platnosti. V případě „silné závislosti" nemá výzkum uvedeným směrem vedený reálný smysl.

Právě uvedený výklad myslím zřetelně ukazuje, proč a v jakém smyslu jsme oprávněni mluvit o tom, že náhodné chyby mají normání rozdělení; a pouze proto jsme oprávněni mluvit o tom, že normální rozdělení je skutečně zákonem rozdělení náhodných chyb.

5

Je naprosto přirozené, že i když máme zákon rozdělení opodstatněný a zdůvodněný, nesmíme čekat, že výběry, které z určitého základního souboru s tímto rozdělením vezmeme, budou tento zákon beze zbytku splňovat. Výběrové a skutečné rozdělení se budou vždycky lišit, a to ve smyslu theorie pravděpodobnosti. Jsou-li odchylky náhodné nebo významné, jinak řečeno, pochází-li výběr ze základního souboru s určenou distribuční funkcí nebo nikoli, to nám řekne statistická theorie testování hypothes. Náhodnost odchylek je způsobována velkou řadou nejrůznějších příčin. Významnost odchylek odhalená statistickým testem nám potom ukazuje, že předpoklady, za nichž byla funkce vyvozena, nejsou (s určitou pravděpodobností) splněny. Poznamenejme, že právě v důsledku těchto mnoha příčin se bude výběrové rozdělení (které je vlastně opět náhodnou proměnnou) jsa bráno ze základního souboru, na př. s rozdělením normálním, v každém konkretním případě lišit od rozdělení normálního, kteréžto odchylky (i když hypothesa o normálním rozdělení v základním souboru bude správná) použitím vhodné jiné funkce s větším počtem parametrů můžeme podstatně snížit. Přijetí nebo zamítnutí určité hypothesy je myšleno ve smyslu statistickém, to jest vždy vzhledem k určité hladině významnosti.

Ukáže-li test shodu, byť i na přísné hladině významnosti, nemusí to znamenat, že hypothesa, kterou jsme zvolili a kterou testujeme, odpovídá skutečné distribuční funkci. Testovaná hypothesa může být tak dobrou aproximací skutečného zákona, že test přijme tuto aproximaci za „zákon" (ukázali jsme, že tak dobrou aproximaci lze vždy nalézt). Odtud plyne, že test nám v podstatě určuje v jistém smyslu stupeň aproximace.

Jak si tedy počínat v případě, kdy nemáme žádnou možnost udat předpoklady, z kterých by bylo možno skutečnou distribuční funkci vyvodit. Lze postupovat na př. tímto způsobem: Určitou methodou získáme vyrovnání napozorovaných hodnot v jistém smyslu „theoretickou" distribucní funkci a otestujeme. Přijme-li test tuto hypothesu na určité

dosti přísné hladině významnosti, pak můžeme tuto aproximaci, získanou vyrovnáním, vzít za „theoretickou" (ne skutečnou) distribuční funkci ve smyslu použité methody vyrovnání a užitého testu. To platí i pro případ, kdy „theoretickou" funkci rozdělení určujeme použitím theorie odhadu pro příslušné parametry (na př. pomocí momentů atd.).

Touto methodou, tak zvanou parametrickou, není typ skutečné distribuční funkce vyvozen přímo z vnitřních zákonitostí sledovaného jevu a nedává nám podklad k tomu, považovat vyrovnanou („theoretickou") distribuční funkci za skutečnou, byť je sebelepší a k praktickým účelům sebevyhovující aproximací; tato methoda nám neříká nic o skutečných distribučních funkcích, o skutečných zákonech rozdělení; může v sobě zahrnovat značný prvek subjektivismu, chceme-li potom s aproximací přijatou na základě statistického testu zacházet jako se zákonem. Onen subjektivismus odráží se při „stanovení" typu funkce rozdělení, ve které potom pomocí v podstatě objektivních matematicko-statistických method určíme správně příslušné konstanty.

Methodou, která nás zbaví těchto prvků subjektivismu a která vede k objektivnímu závěru vzhledem k zákonu rozdělení v případě, kdy nemáme možnost postihnout reálné předpoklady, které by nám umožnily zákon rozdělení získat deduktivně, je tak zv. neparametrická methoda, nepředpokládající a priori znalost skutečné distribuční funkce.

Nechť tedy sledovaná náhodná proměnná ξ (na př. pozorované chyby) má neznámou funkci rozdělení, o které pouze předpokládejme, že je spojitá. Označme příslušnou kumulativní distribuční funkci $F(x)$. Mějme dále řadu nezávislých pozorování s empirickou komulativní distribuční funkcí $S_n(x)$, kde n je počet pozorování. Při hodnocení takto získaného rozdělení nesmíme, jak se často činí, nikdy zapomenout, že jde o rozdělení výběrové, o rozdělení charakterisující právě získanou množinu dat, které je s rozdělením skutečným spjato zákonitostmi pravděpodobnostními. Předpokládejme dále, že mechanismus jevu neznáme natolik, abychom mohli formulovat takové postuláty, z nichž bychom funkci $F(x)$ mohli vyvodit deduktivně, i když na druhé straně máme opodstatněn předpoklad, že tato objektivně, reálně existuje. Moderní matematická statistika nám dává prostřednictvím neparametrických method možnost konstruovat na základě výběrů (řad získaných měření) takové meze, v nichž se, na určité hladině α (to jest ve smyslu pravděpodobnosti) bude neznámá skutečná distribuční funkce nalézat. Pravděpodobnostní hladinu α dosti přísnou si předem zvolíme ($0 \ll \alpha < 1$).

Nechť potom získáme posloupnost výsledků nezávislých pozorování $x_1^*, x_2^*, \ldots, x_n^*$, které uspořádáme v tak zvanou „variační řadu": $x_1 \leq x_2 \leq \ldots \leq x_n$. Nechť této řadě přísluší empirická distribuční funkce

$$S_n(x) = \begin{cases} 0 & \text{pro} \quad x \leq x_1, \\ \dfrac{k}{n} & \text{pro} \quad x_k < x \leq x_{k+1}, \\ 1 & \text{pro} \quad x_n < x. \end{cases}$$

Principiální řešení naší úlohy nám podává Kolmogorovova věta, říkající že za shora uvedených předpokladů platí, že stejnoměrně v z

$$\lim_{n \to \infty} P\{\sqrt{n} \sup_x |S_n(x) - F(x)| < z\} = K(z),$$

kde

$$K(z) = \begin{cases} 0 & \text{pro} \quad z \leq 0, \\ \displaystyle\sum_{k=-\infty}^{\infty} (-1)^k \, e^{-2k^2 z^2} & \text{pro} \quad z > 0. \end{cases}$$

Tato věta, mající nejrozmanitější aplikace, je historickým východiskem (z r. 1933) při vývoji jedněch z nejprogresivnějších method matematicko-statistického zkoumání — method neparametrických. Velká přednost těchto method spočívá v tom, že kriteria, na základě nich konstruovaná, nepředpokládají znalost typu distribuční funkce v základním souboru, čímž odstraňují možnost subjektivního vlivu v tomto směru. V důsledku této objektivity jsou tyto methody středem pozornosti především v SSSR, odkud také pochází převážná část i dalších principiálních závěrů. Poznamenejme, že uvedené limitní formulace lze použít již celkem pro nevelká n.

Upravme shora uvedený výraz na tvar

$$\lim_{n\to\infty} P\left\{S_n(x) - \frac{z}{\sqrt{n}} < F(x) < S_n(x) + \frac{z}{\sqrt{n}}\right\} = K(z),$$

který opět platí pro všechna x, stejnoměrně v z. Zvolme nyní za $K(z)$ určitou pevnou hodnotu dostatečně blízkou 1. Označme ji α. Z tabulek funkce $K(z)$ pro tuto hodnotu α získáme určitou hodnotu $z = z_0$.

Získaný výsledek nám potom říká, že ve smyslu pravděpodobnostních zákonitostí bude skutečný zákon rozdělení ležet s pravděpodobností α v takto stanoveném pásu

$$S_n(x) - \frac{z_0}{\sqrt{n}} < F(x) < S_n(x) + \frac{z}{\sqrt{n}}$$

pro všechny hodnoty x.

Všechny distribuční funkce, které vyhovují této nerovnosti na hladině významnosti α, můžeme označit jako α — aproximace skutečné distribuce funkce. Kromě toho navíc víme, že skutečná distribuční funkce bude s pravděpodobností α ležet v takto stanoveném pásu. [7]

Poznámka: a) Neparametrické testy mají platnost v právě uvedeném smyslu — jak dokázal Kolmogorov — i po vypuštění předpokladu spojitosti pro distribuční funkci v základním souboru.

b) V současné době máme již možnost užívat místo limitních funkcí tvarů pro konečná n [8].

c) Velmi zajímavý článek v tomto směru napsal Wieslaw Sadowski: *O założeniu normalności w teorii błędow*, Przegląd statystyczny, 1956, 2, Warszawa, kde je podán rozbor jisté originální úvahy, která je v úplném souladu s koncepcí tohoto článku.

V tomtéž smyslu upozorňuji dále na článek B. Pardubského, *Některá rozdělení chyb měření*, Čs. čas. fys., 1955, 5, který rovněž podává řešení vycházející z analysy procesu tvoření chyb.

Citovaná literatura.

[1] F. Fabian, *Poznámka k pojmu „pravděpodobnost"*, Filos. časopis, III, č. 4, 1955.
[2] I. P. Natanson, *Konstruktivnaja těorija funkcij*, 1949, str. 19—26.
[3] S. N. Bernštejn, *Těorija verojatnostěj*, 1946, str. 339.
[4] Viz na př. W. P. Elderton, *Frequency Curves and Correlation*, 1938.
[5] Viz na př. A. N. Krylov, *Lekcii o priblížennych vyčislenijach*, str. 365—369.
[6] B. V. Gněděnko, *Kurs těorii verojatnostěj*, 2. vyd., 1954, str. 241—249.
[7] F. Fabian, *O neparametrických methodách*, SV—Matematika, fysika, astronomie, IV, č. 5, 1954, str. 671.
[8] B. V. Gněděnko, *Proverka něizmennosti rasp. ver. v dvuch nězavisimych vyborkach*, Mathematische Nachrichten, sv. 12, seš. 1/2, 1954.