Emil Schoenbaum

O matematické statistice

Časopis pro pěstování matematiky a fysiky, Vol. 64 (1935), No. 5, 124--131

Persistent URL: http://dml.cz/dmlcz/121277

O matematické statistice.

Dr. *Emil Schoenbaum,* Praha.

Úkol, předvésti hlavní pokroky v matematické statistice docílené v posledních několika letech, vyžadoval by doby několika hodin a nebyl by řešitelný ani s použitím znamenitých přehledů, které v poslední době vydává I. O. Irwin v Journal of the Royal Statistical Society a Darmois v časopise Econometrica. Neboť obor matematické statistiky rozšiřuje se každým rokem jednak co do rozsahu, jednak co do hloubky tak intensivně, že není možno v krátké přednášce ani sledovati tento rozvoj v celku, natož analysovati nebo kritisovati. Tak vztahuje se referát Irwinův v letošním ročníku Journal of the Royal Statistical Society na více než 110 prací, z nichž je celá řada velmi obsáhlých a po matematické stránce značně nepřístupných. Statistikovi starší školy je novinkou obzvláště prohloubení po matematické stránce, které potvrzuje námi dávno hájené mínění, že studium matematické statistiky vyžaduje v prvé řadě pronikavého a důkladného školení matematického.

Usnadnil jsem si proto svůj referát tím, že omezuji se na podání obsahu prací, jež považuji za nejdůležitější pro další vývoj matematické statistiky, protože udávají nové směry nebo přinášejí zpřesnění nebo zobecnění dosavadních výsledků. Nepopírám, že pro tento subjektivní ráz bude můj referát jednostranný a neúplný. Ale tomu nelze se vyhnouti.

I. Každý statistik, stojící před úkolem najíti pro statistický kolektiv adekvátní vyjádření, učiní aspoň v myšlenkách pokus, docíliti analytického vyjádření pomocí zákona Gaussova. Důvod, že tento pokus se často a s velikým úspěchem daří, jest známý a je tento: Sečteme-li veliký počet nahodilých proměnných sledujících libovolné zákony rozdělení, splňuje součet zákon blízký zákonu Gaussovu za podmínek, jež lze populárně, třeba ne úplně přesně vyjádřiti tak, že každá ze složek součtu má býti malá v poměru k střední odchylce součtu. Nepřehledný počet prací, z nichž uvádím pouze jako nejdůležitější Čebyševa, Liapunova, Markova a z novějších práci Lindbergovu, Pólyovu, P. Lévyho, Cantelliho, Khintčina, Kolmogorova atd. a v letošním ročníku časopisu Giornale dell'Istituto Italiano degli Attuari důležitou práci H. Cramér a o stejnoměrném zákoně velkých čísel, zpřesňuje nebo zevšeobecňuje tyto podmínky nebo je rozšiřuje na případ více pro-

měnných. Ale podstatou všech těchto prací zůstává, že nahodilé veličiny skládající veličinu výslednou jsou neodvislé.

Velikým obohacením matematické statistiky je proto, že se podařilo v počtu pravděpodobnosti již Markovovi rozšířiti platnost zákona Gaussova na případ proměnných závisejících zvláštním způsobem v řetězu. Generalisaci těchto prací Markovových, jimiž zabývá se také z různých hledisek Fréchet a Hostinský, důležitou pro matematickou statistiku obsahuje velmi závažná práce S. Bernsteina v 97. svazku Mathematische Annalen. Bernstein vychází ze zvláštního případu závislosti, jímž zabýval se již dříve Bruns a zvláště Markov, totiž takového, že veličiny dostatečně od sebe vzdálené jsou úplně nezávislé a rozšiřuje vzdálenost závislých veličin tak, aby Gaussův zákon t. zv. Grenzwertsatz ještě platil. Jeho výsledky lze vyjádřiti tak, že jestliže závislost se dostatečně zeslabuje se vzdáleností veličin, součet proměnných veličin řídí se přece Gaussovým zákonem. Tento případ vyskytá se v praxi statistické velmi často, totiž, že za sebou jdoucí veličiny nahodilé jsou sice závislé, ale jejich závislost mizí po několika členech. Bernsteinova práce obsahuje jako speciální případ Markovovy výsledky a dochází mimo to také rozšířením na 2 součty závislých veličin k přesné formulaci obecných podmínek pro teorii normální korelace, dovoluje tedy značné rozšíření platnosti Gaussova zákona na zjevy, které nebyly dosud v teorii uvažovány, ale které jsou v praxi právě obvyklé; zdá se mi proto, že právě matematická statistika mohla by vytěžiti z této Bernsteinovy práce, která znovu prokazuje důležitost teorie Markovových řetězů, velmi mnoho.

Privilegované postavení Gaussova zákona vysvětluje se, jak známo tím, že zkušenost nám ukazuje, že v nejrůznějších oborech lidského vědění lze různá rozdělení statistická dostatečně přesně vyjádřiti Gaussovou křivkou. Mimo to však učí nás teorie, že sčítáním velikého počtu malých nahodilých a nezávislých účinků vzniká za velmi obecných podmínek Gaussova křivka. Tak vysvětluje tedy teorie empirický zjev, že Gaussova křivka se často vyskytuje. Ale ovšem víme, že existuje mnoho kolektivů statistických, jež nelze Gaussovou křivkou přibližně vyjádřiti a vzniká tak důležitý úkol najíti jiné teoretické křivky, jimiž by bylo lze vyjádřiti křivky empirické. A tu je nutno upozorniti na důležitou práci Pólyovu v Annales de l'Institut H. Poincaré 1931, která osvětluje výsledky dokázané Markovem a Bernsteinem a odvozuje systematicky zákony platné pro zjevy četné i řídké se závislostí slabou, silnou a obzvláště silnou. Zajímavo je, že jako limitní případ schematu nahodilých zjevů nakažlivých, tedy schematu daleko obecnějšího než jest schema urny s koulí nevrácenou, lze odvoditi

pouze dvě křivky Pearsonovy. Také tato Pólyova práce zaslouží pozornosti statistiků.

II. V souvislosti s pokusy vyjádřiti daný kolektiv křivkou Gaussovou je jeden z důležitých problémů matematické statistiky, jehož řešení působilo i po teoretické stránce, ale hlavně po praktické stránce veliké potíže, a jenž vykazuje v nejnovější době veliký pokrok po stránce počtářské. Jde o problém rozkladu dané frekvenční křivky ve dva dané zákony rozdělení. Jest známo, že před 40 lety podařilo se již Pearsonovi rozřešiti teoreticky problém rozkladu dané frekvenční křivky ve dvě složky sledující Gaussův zákon. Prakticky aplikoval Pearson tuto teorii na soubor lebek nalezených v předhistorickém bavorském pohřebišti a prokázal, že se jedná o soubor vzniklý smíšením dvou čistých ras. Po počtářské stránce jde o velmi obtížný úkol, který vede k řešení velmi komplikované algebraické rovnice 9. stupně, o úkol podle Charliera heroický. Charlier a Wicksell v Arkiv f. Mat. och Physik 1923 zjednodušili řešení Pearsonovo, ale přes to prohlašují úkol stále ještě za počtářsky nesnadný. Je zásluhou 2 prací uveřejněných v tomto roce v Skandinavisk Aktuarietidskrift 1934 (Burrau: „The half invariants of the sum of two typical laws of errors with an application to the problem of dissecting a frequency curve into components" a Strömgren: „Tables and diagrams for dissecting a frequency curve into components by the half invariant method"), že činí problém přístupný právě po počtářské stránce, poskytujíce předpis pro jednoduchý postup početní. Této výhody je docíleno použitím Thieleových semiinvariantů. Jde o problém, vyjádřiti danou frekvenční funkci $\varphi(x)$ jako součet dvou Gaussových křivek

$$\varphi(x) = h_1 e^{-\frac{(x-m_1)^2}{2n_1^2}} + h_2 e^{-\frac{(x-m_2)^2}{2n_2^2}}.$$

Použijeme-li Thieleových invariantů, máme ihned rovnici

$$e^{\frac{\lambda_1}{1!}t + \frac{\lambda_2}{2!}t^2 + \ldots + \frac{\lambda_n}{n!}t^n + \ldots} = \frac{n_1 h_1}{n_1 h_1 + n_2 h_2} \cdot e^{m_1 t + \frac{1}{2}n_1^2 t^2} +$$

$$+ \frac{m_2 h_2}{n_1 h_1 + n_2 h_2} \cdot e^{m_2 t + \frac{1}{2}n_2^2 t^2},$$

$$\frac{n_1 h_1}{n_1 h_1 + n_2 h_2} = k_1, \qquad \frac{n_2 h_2}{n_1 h_1 + n_2 h_2} = k_2,$$

kde λ_i jsou poloinvarianty i-tého stupně dané křivky, m_1, m_2 aritmetické průměry, n_1, n_2 střední chyby obou Gaussových zákonů, h_1, h_2 jsou počty individuí patřících do složkových souborů. Postupným diferencováním rovnice, klademe-li potom $t = 0$, obdržíme relace pro proměnné $k_1, k_2, m_1, m_2, n_1, n_2$, při čemž lze nalézti dosti jednoduchý aritmetický zákon pro n-tý semiinvariant.

Prvých 6 relací slouží k určení 6 konstant. Odtud vycházejíc odvozuje druhá cit. práce 2 početní předpisy, jednoduchý s menší přesností a složitý s větší přesností. Zjednodušením je ovšem, jsou-li směrodatné odchylky tytéž; tento úkol řešil Steffensen ve svém počtu pravděpodobnosti. Početní postup je ilustrován na 5 příkladech a doložen důležitými tabulkami usnadňujícími výpočet. Daleko těžší problém, rozložiti danou frekvenční křivku ve dvě křivky sledující jiné zákony, na př. Poissonův nebo Pearsonovy typy, čeká ještě na řešení.

III. Určení matematické formy rozdělení četnosti čili řešení t. zv. problémů specifikace je usnadněno bohatým materiálem tabulkovým a nesčetnými pracemi K. Pearsona a jeho žáků, obzvláště tabulkovým dílem Tables for statisticians and biometricians (I. a II. díl 1931). Teorie specifikace jest obsahem prací R. A. Fishera a jeho školy a obzvláště jeho práce „Theory of statistical Estimation" (Proceedings of the Cambridge Phil. Society 1925), jež vyjádřily přesně důležitost problému odhadu. Práce R. A. Fishera a jeho školy, přístupné i širšímu kruhu interesentů jeho učebnicí „Statistical methods for research workers 1930" přinesly veliký pokrok teorii representativní metody, která stává se středem usilovného badání matematických statistiků obzvláště anglosaských, a to z obou hledisek náhodného nebo záměrného výběru. Metoda momentů, jíž se zde všeobecně užívá, doznala velikého obohacení důležitou prací St. Georgescu (Biometrika 1932), která úplně novou metodou dociluje výsledků obdržených již dříve C. C. Craigem, R. A. Fisherem, ale nadto dociluje přesných formulací pro momenty a semiinvarianty malých náhodných výběrů a přibližných pro rozsáhlé výběry.

Tohoto nejvýše pozoruhodného výsledku je docíleno zavedením pojmu t. zv. asociované funkce. Budiž

$$y_1, y_2, \ldots, y_t$$

posloupnost proměnných a

$$\{y_{i_1}, y_{i_2}, \ldots, y_{i_p}\}$$

střední hodnota součinu p proměnných, pak se nazývá výraz

$$\alpha_p(u_1, u_2, \ldots, u_p) = \sum_{i_1=1}^{t} \sum_{i_2=1}^{t} \cdots \sum_{i_p=1}^{t} \{y_{i_1} y_{i_2} \ldots y_{i_p}\} \frac{u_1^{i_1}}{i_1!} \cdot \frac{u_2^{i_2}}{i_2!} \cdots \frac{u_p^{i_p}}{i_p!}$$

funkcí přidruženou (asociovanou) k součinovým momentům p-tého řádu. Obdobně definována je funkce asociovaná k semiinvariantům p-tého řádu. Pomocí asociované funkce lze definovati funkci charakteristickou

$$\varphi(t_1, t_2, \ldots, t_n, \ldots) = 1 + \frac{\lambda}{1!}\sum_i \alpha_1(t_i) + \frac{\lambda^2}{2!}\sum_{i,j} \alpha_2(t_i, t_j) + \ldots$$

Pomocí vytvořující funkce $G(t)$ definované

$$G(t) = y_0 + y_1 \frac{t}{1!} + y_2 \frac{t^2}{2!} + \ldots + y_p \frac{t^p}{p!} + \ldots \qquad (2)$$

a je-li ještě $F \,.\, dv$ elementární pravděpodobnost výsledku $(y_0, y_1, \ldots, y_p, \ldots)$, lze ukázati, že

$$\alpha_p(t_1, t_2, \ldots, t_p) = \int_D F \, G(t_1) \, G(t_2) \ldots G(t) \, dv, \qquad (3)$$

při čemž se integruje přes všechny možné hodnoty y. Je potom

$$\psi(t_1, t_2, \ldots, t_n, \ldots) = \int_D F \, e^{\lambda[G(t_1) + G(t_2) + \ldots + G(t_n) + \ldots]} \, dv. \qquad (4)$$

Srovnáním (3) a (4) obdrží se asociovaná funkce rozdělení momentů kolem libovolného počátku a řada přiblížení libovolného stupně pro semiinvarianty a produkční momenty semiinvariantů náhodného výběru. Velmi významná práce končí dvěmi tabulkami. Jedna dává semiinvarianty druhého řádu semiinvariantů výběru až do váhy 12 a až po člen $1/N^2$, druhá semiinvarianty produkčních součinů náhodného výběru pro střed aritmetický až do váhy 12. Obecné výsledky dosažené Georgescu a jeho metoda zasluhují co největší pozornosti.

IV. Ježto práce Georgescu-ova týká se výpočtu momentů náhodného výběru, bude snad na místě upozorniti na nejnovější důležité práce Guldbergovy a R. Frische, které ukazují, jak z diferenčních rovnic, jimž hoví frekvenční funkce, lze odvoditi jednoduché rekursní formule pro úplné a neúplné momenty a z těchto zase odvoditi kriteria, jež dovolují rozhodnouti, zda daný kolektiv statistický lze vyjádřiti teoretickou frekvenční formulí. Kdežto R. Frisch v práci v Metronu 1932 zabývá se obecnou analysou problému, odvozuje Guldberg pro zákon binomický, Poissonův, Pascalův a hypergeometrický momenty i kriteria t. zv. místní i úhrnné (globální) a v nejnovější práci ve Skandinavisk Aktuarietidskrift 1934 rozšiřuje tyto výsledky na kolektivy o 2 a více proměnných. V komunikaci předložené tomuto kongresu aplikuje dr. Fischer Guldbergovu metodu s úspěchem též na zákon Pólya-Eggenbergův. Práce Guldbergovy mají též značný význam pedagogický.

V. Důležitou otázku, jak posouditi, zdali hypotésy zvolené za podklad odhadu jsou ve shodě s úhrnem empirických dat, řeší známé práce K. Pearsona, jejichž základem je používání testu x^2. Položíme-li si otázku o zákonu pravděpodobnosti veličiny χ^2, stojíme před novým problémem, který byl podrobně zkoumán ve veliké práci E. S. Pearsona a Neymana „On the problem of the Most Efficient Tests of statistical Hypotheses" ve Philosophical Transactions 1933. Jedná se v principu o úctyhodně starý problém

teorému Bayesova, zajímavý po stránce filosofické. Autoři vycházejí z pojmu nejlepšího kritického oboru pro danou hypotésu. Je to obor, jenž poskytuje určitou pravděpodobnost pro zamítnutí hypotésy a činí maximem pravděpodobnost pro přijetí hypotésy druhé. Jest zřejmo, že jedná se o problém variačního počtu. Existují-li ovšem více nežli 2 hypotésy, máme problém daleko obtížnější. Teorii illustrují autoři zajímavými příklady.

K problému „maximální likelihood" vztahuje se polemika mezi Jeffreyem, R. A. Fisherem a Bartletem, o níž nemohu se blíže zmíniti, týkající se pojednání „An Alternative to the Repetition of Observations".

Zajímavých problémů statistických se dotýkají nejnovější práce Lotkovy, Misesovy, Meidellovy, Zwinggiho, Tricomiho a jiných, které navazují na použití integrálních rovnic v analyse populace a v jiných populačních problémech, dále řada prací o matematické epidemiologii, které však leží již na rozhraní statistiky a specielních věd.

VI. Ve statistice vyskytuje se často úkol, najíti závislost mezi 2 znaky, jež nelze kvantitativně určiti, ale u nichž známe pořadí (rank). Teorii takových kolektivů je věnována řada prací, z nichž uvádím práce Spearmanovu a Esscherovu. Ve veliké práci „On the Mean character and variance of a Ranked individual and on the Mean and Variance of the Intervals between Ranked Intervals" (Biometrica 1932) zabývá se Pearson s hlediska této teorie studiem exponenciální populace a ukazuje, že momenty závisí zde na diferenciálních koeficientech logaritmu beta funkce.

$$\mathrm{B}(q, n-q+1)$$

$$\frac{\mathrm{d}}{\mathrm{d}n}\log \mathrm{B} \qquad = \text{dibeta funkce}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}n^2}\log \mathrm{B} \qquad = \text{tribeta funkce atd.}$$

Odvozuje vztahy mezi těmito funkcemi a funkcemi gama, digama, atd. a ukazuje, že lze je vyjádřiti pomocí součtu mocnin reciprokých čísel $\Sigma\, 1/n^k$, pro něž uveřejňuje bohaté tabulky na 12 desetinných míst. Své po matematické stránce zajímavé úvahy illustruje Pearson četnými příklady.

VII. Rostoucí význam řad ortogonálních polynomů pro vyjadřování kolektivů vede k četným pracím, z nichž obzvláště upozorňuji na novou práci Aitkenovu, která navazuje na známou práci Charlierovu, vyjadřující řadou Hermiteových polynomů. Totéž provádí Aitken v Proceedings of the Royal Society Edinburg 1933 „On the Orthogonal Polynomials in Frequencies of Type B" s Charlierovou řadou B.

$$\log f(x) = \log \psi(x) + k_2 K_2(x) + k_3 K_3(x) + \ldots$$

$$K_\nu(x) = (-1)^\nu . \nabla^\nu \psi(x), \quad \psi(x) = \frac{e^{-m} . m^x}{x!}, \nabla \psi(x) = \psi(x) - \psi(x-1).$$

Odvození orthogonálních polynomů ze zákona Pólya-Eggenbergova o pravděpodobnosti nakažlivých zjevů obdobných polynomům odvozeným Romanovským obsahuje komunikace p. dra Fischera, předložená tomuto sjezdu.

VIII. Z prací, jež řeší jednotlivé problémy, chci zmíniti se o práci H. Wolda „Sulla correzione di Sheppard" v Giornale dell'Istituto Italiano degli Attuari červenec 1934, která znamená jistý pokrok v řešení tohoto problému dosud neuspokojivém a přebíraném z učebnice do učebnice. Autor odvozuje opravu Sheppardovu, aniž musil předpokládati dotyk řádu vyššího, a vyjadřuje stupeň aproximace tím, že odvozuje zbytek přibližného výrazu. Jedinou podmínkou, je

$$x^{n+1} . \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x+t)\, dt \to 0 \qquad \text{pro } t \to \pm\infty;$$

autor prokazuje, že tento předpoklad neznamená podstatnou restrikci proti obvyklé restrikci

$$x^m . f^{(2m)}(x) = 0 \qquad \text{pro } x = \pm\infty \text{ a pro všechny hodnoty } m,$$

ale průkaz nezdá se mi bezvadný. Analogické vzorce jako Sheppardův jsou odvozeny pro momenty faktorielní, pro něž byla odvozena Sheppardova oprava již dříve Langdonem a Orem 1930 v Annales of mathematics.

IX. Okolnost, že mocninné momenty jsou symetrickými funkcemi argumentů, vedla v nové době k důležitým pracím matematicko-statistickým o symetrických funkcích; uvádím pouze práci Birgera Meidella v Skandinavisk Aktuarietidskrift 1928, v níž dochází k nerovnostem majícím roli v matematické statistice a práci Toolovu 1932 v Annales of Math. Statistics „On Symetric functions of more than One Variable and of Frequency Functions", kterou znám pouze z referátu Irwinova.

Svůj referát o matematické statistice končím poukazem na důležitou diskussi na 22. schůzi mezinárodního ústavu pro statistiku v Londýně tohoto roku o t. zv. koeficientu korelace, o níž podává výstižnou zprávu prof. M. Fréchet. Nesprávné užívání tohoto koeficientu k měření stupně závislosti statistických veličin bez potřebných výhrad vedlo k podání návrhu zvláštní komise, který byl podložen anketou, jíž se zúčastnili: Bowley, Wilson, Darmois, Huntyngton, Steffensen, P. Lévy, Rietz, Gini, Guldberg, Gumbel, Hotelling, Mises, R. Frish, Risser, Jordan. Návrh zní:

„L'institut International de Statistique, dans sa Session d'avril 1934,

I. considérant que de nombreux statisticiens emploient le coefficient de corrélation, dit de Bravais-Galton, en croyant ou comme s'ils croyaient — mathématiquement démontré que ce coefficient est, en tout état de cause, une bonne mesure de la rigueur de la dépendance fonctionnelle entre deux variables statistiques;

attire l'attention sur les résultats d'une enquête mené auprès de quelques-uns des statisticiens les plus connus par leur compétence mathématique;

et, observant que parmi les réponses à cette enquête il n'en est aucune, soutenant qu'en l'absence de tout autre renseignement, il suffit de savoir que le coefficient de deux variables statistiques X et Y est voisin de zéro pour conclure à l'inexistence d'une relation fonctionnelle entre X et Y, qu'au contraire la plupart s'élèvent contre cette assertion;

constate qu'on ne saurait considérer cette assertion comme une vérité mathématiquement démontrée, ou, à tout le moins, reconnue universellement comme telle;

(que le champ de validité de l'emploi du coefficient de corrélation est singulièrement plus étroit que beaucoup ne l'avaient d'abord supposé;)

(qu'aucune valeur du coefficient de corrélation, fut-elle égale à $\pm$ 1 ne saurait, à elle seule, garantir la causalité, laquelle doit être prouvée par des raisons d'un ordre différent;)

(et rappele que, pour éviter aux débutants toute erreur sur ce point, plusieurs auteurs ont proposé de substituer à l'expression „coefficient de corrélation" celle de coefficient de linéarité;)

II. en ce qui concerne le cas plus délicat où l'usage du coefficient de corrélation serait limité, en ce qui concerne le degré de la dépendance, aux observations où la courbe des moyennes est presque linéaire ou au cas plus restreint encore où l'on suppose que les variables X, Y vérifient presque exactement la loi normale;

constatant que certains auteurs élèvent des objections, même dans ce cas particulier, contre l'assertion ci-dessus mentionnée décide que le mandat de la commission sera prorogé pour étudier ces cas particuliers;

III. constatant d'autre part que le coefficient de correlation possède une signification statistique qu'il serait intéressant de dégager aux points de vue de la concordance, de la linéarité, de l'interpolation, etc.;

décide que la commission devra aussi étudier la signification du coefficient de corrélation et préciser dans quels cas et pour quels buts, son emploi peut être recommandé."