STATI

Pedagogika a edukometrie

STANISLAV KOMENDA

Anotace: Edukometrie je oblast pedagogiky, zabývající se možnostmi a metodami měření pedagogických jevů. Článek demonstruje použití prostředků statistického rozhodování jako nástroje popisu a analýzy klasifikace znalostí žáka v případě, kdy jsou tyto znalosti měřeny pomocí didaktického testu typu multiple-choice.

Klíčová slova: Učení, zpětná vazba, měření znalostí, statistika, edukometrie, didaktický test, operační charakteristika, entropie, distance, účinnost klasifikace

ÚVOD

Ve svém vývoji prochází každá empirická věda obdobím soustředěného úsilí uchopit alespoň některé své části kvantitativně, zařadit do svého instrumentária měření a zpracovat metodologii svého měření. Tento proces se nevyhnul ani vědám humanitním; názorným příkladem je psychologie.

Také pedagogika opírá některé ze svých disciplín o empirii a měření — třeba v oblasti zjišťování znalostí a jejich klasifikace. Je proto zcela přirozené klást si otázky o účinnosti měření a jeho optimizaci. Předkládaná studie je pokusem přiblížit tuto problematiku odborné pedagogické veřejnosti. V první části se zdůvodňuje místo induktivní statistiky v kontextu pedagogické zpětné vazby; druhá část je věnována aplikaci v případě klasifikace znalostí.

1. UČENÍ, TEST A GENERALIZACE

Základem rozumného počínání je stanovit, co je rozumné.

1.1 Učení

Genetický mechanismus, vtiskující potomkům vlastnosti rodičů, umožňuje zachovat pro zítřek, co bylo včera a co je dnes, aniž by potlačil možnost změn a odchylek, zaručujících plastičnost a dynamiku vývoje vlastností jedinců lidského rodu.

Stejně jako vlastnosti vrozené, také některé vlastnosti během života jedince získané (zejména poznání) je možno předávat z generace na generaci mechanismem řeči a písma. Teprve tato schopnost, v organizovaném procesu vzdělávání, v systému zvaném škola,

umožnila člověku překonat nezvratný fakt smrtelnosti jedince a zaručila nesmrtelnost lidského rodu; nic z toho, co bylo jednou individuem poznáno, nemusí být zapomenuto. Získávat poznatky bez toho, že by se mohly předávat z generace na generaci, bylo by počínáním beze smyslu.

1.2 Zpětná vazba

Učení je tak významná funkce v životě jedince, že je celá šestina až čtvrtina délky života věnována téměř výhradně jí — a téměř celý další život je jí více nebo méně intenzívně prostoupen.

Učení je procesem regulovaným, řízeným, případně procesem autoregulujícím. Nezbytnou součástí takového řízení je zpětná vazba, zajišťující tok informace o stavu vědění učícího se subjektu, případně soustavy subjektů. Tato informace se získává pozorováním a měřením. Co je škola školou, od chvíle, kdy stanul učitel před žáky a mezi žáky, kladl si zřejmě otázku, co žák umí, aby usměrnil svoje vlastní počínání, s cílem působit pokud možno účinně. Zkoušení všeho druhu, testování znalostí, je masově praktikovaným měřením ve službách zpětné vazby a řízení pedagogického procesu.

1.3 Induktivní usuzování

Měřit znalosti není možno vyčerpávajícím způsobem; už proto ne, že samotný proces měření znalosti subjektu ovlivňuje: rozšiřuje a modifikuje. Měření znalostí je měřením dílčího, omezeného vzorku — a je nutně provázeno úsilím zobecňovat dílčí poznatky na širší celek; z odpovědí na několik otázek, z řešení dílčí úlohy se usuzuje na celkové znalosti, na znalosti celku. Zkoušet — znamená generalizovat, zobecňovat; zkoušet znamená provádět induktivní úsudek — se všemi problémy, které takové induktivní usuzování nutně provázejí: totiž problémy rizika omylu, že vědění, případně nevědomost nebudou rozpoznány. Existuje vědní obor, zabývající se problémem induktivního usuzování výhradně a profesionálně. Tímto oborem je matematická statistika, založená na představách teorie pravděpodobnosti: na koncepci náhodného jevu, pravděpodobnosti jeho výskytu, náhodné veličiny a pravděpodobnostního rozdělení, na koncepci nezávislosti jevů a veličin. Tento vědní obor je tu už nějakých 100 let — a není důvod, proč ho nevyužít v pedagogice, stejně jako je už desítky let využíván v psychologii, biologii, medicíně a třeba i ekonomii. Buďme však přesnější — metody matematické statistiky nejsou pro pedagogiku neznámou pevninou. Spíše jde o to, že se dosud nestaly součástí učebnic pedagogiky tak, jako je tomu např. v psychologii. Spíše jde o frekvenci využívání a míru zdomácnění statistických metod v pedagogickém výzkumu; jde o to, aby se statistické koncepce stávaly východiskem chápání pedagogických jevů a součástí struktury pedagogických pojmů. Jde o to, aby se stejně jako psychometrie, antropometrie, ekonometrie a biometrie naplňoval také obsah oboru, který někdy označujeme názvem edumetrie nebo edukometrie.

1.4 Statistika, edukometrie a učitel

Na základě své dosavadní více než třicetileté zkušenosti pedagogické i výzkumné se cítím oprávněn soudit, že učitelům všech typů škol by nemělo být cizí statistické hledisko. Měli by si umět představit, že takové pojmy jako talent, schopnosti, inteligence a znalosti jsou měřitelné a uchopitelné jako veličiny s jistým frekvenčním rozdělením v referenční populaci.

Statistika, jak známo, se zabývá výlučně soubory; na jednotlivé případy se její závěry vztahují a dají aplikovat jenom v té míře, v jaké je jedinec prvkem nebo členem příslušné referenční populace. Učiteli je svěřována péče o kolektiv, soubor, o třídu. Pokud má učitel na mysli třídu jako celek, její úroveň a její vlastnosti, je na místě, aby k ní přistupoval jako statistik [2, 3].

Je tu ovšem jeden zásadní rozdíl: pro učitele není třída souborem anonymních prvků, ale souborem individualit, jejichž osobnost má být rozpoznána, pěstována a rozvíjena podle zásad optimálních pro každého jedince. Úkolem učitele je rozpoznat, diagnostikovat schopnosti a možnosti, stejně jako limity a omezení každého žáka, a dokázat mu účinně poradit při hledání jeho či její optimální životní strategie. Učitel nemá být statistik — statistika v rukou poučeného učitele je však schopna pomoci mu v kvalifikovaném plnění jeho životního poslání. Platí totiž, že právě tak jako se v některých krocích svého působení obrací učitel na žáka, na jedince, jsou i jiné kroky, v nichž se obrací a působí na kolektiv.

Chtěl bych se zastavit u jednoho z nejdůležitějších momentů práce učitele, jakým je nutnost předvídat, anticipovat budoucí vývoj, chování, nutnost odhadovat a usuzovat z minulého a přítomného na budoucí. Je samozřejmé, že individuální trajektorie jsou nesmírně proměnlivé, protože jsou ovlivňovány množstvím i subjektivních faktorů, jejichž váhu v dané chvíli je obtížné předvídat. Spolehlivost predikce je v takové situaci především funkcí nápaditosti a zkušenosti učitele. Naším úkolem je však také anticipovat vývoj kolektivů — vývoj názorů, postojů, motivace k práci, morálky, kriminality a jevů specifických pro školní prostředí: vztah žáka k učiteli a obráceně záškoláctví, vztah ke sportu, atmosféru v kolektivu třídy, úlohu přisuzovanou vůdčím osobnostem kolektivu. Řešit takový úkol mohou statistika a edukometrie rozhodně pomoci.

Každý z nás, ať se zabývá čímkoli, potřebuje uspokojení ze své činnosti, potřebuje mít satisfakci, vědomí, že své práci profesionálně rozumí a koná ji na úrovni, a že je tato úroveň objektivně ostatními kolem něho uznávána a respektována. Učitel má nárok mít ze své práce stejný pocit krásna jako třeba truhlář nad svým stolem, malíř nad plátnem nebo rolník nad svou úrodou. Je to jenom trochu méně bezprostřední — protože objekt jeho úsilí je nesrovnatelně složitější a křehčí než v případech výše uvedených.

Jsem zastáncem názoru, že k témuž cíli může vést řada cest. Stejně jako alternativní školství přestalo být zakázaným pojmem, neměly by ani možnosti kvantitativních metod edukometrie zůstat nevyužity — v pedagogické teorii, ve výzkumu, ale ani v pedagogické praxi. Pedagogika má nepochybně svůj specifický předmět zkoumání, z jistého hlediska nazírání jistě složitější než je předmět jiných vědních oborů. Člověk, jedinec, zůstane pro jiného člověka černou skříňkou i poté, co byly objektivními metodami moderní měřicí techniky prozkoumány jeho fyziologické funkce a změřeny jeho mentální výkony. Podstatné je, že i v pedagogické situaci jsou objektivní metody analýzy a řešení problémů využitelné.

1.5 Z historie měření školního výkonu

Systém měření školního výkonu a hodnocení znalostí žáků se v českých zemích ustálil koncem minulého století. Jeho „prehistorie" je ohraničena studijním řádem jezuitských škol na přelomu 16. a 17. století a zavedením maturit na gymnáziích podle pruského vzoru v polovině 19. století [2].

Historie tohoto vývoje je historií kritiky zkoušení, probíhající na hodnotící stupnici, která se počíná naprostou skepsí a odmítáním zkoušek — až po úsilí a racionální, objektivní měření, založené na přestavě, že bez poznání úrovně znalostí žáka se účinnost pedagogického procesu stává problematickou.

V české pedagogice 20. a 30. let je konfrontace obou přístupů k testování znalostí spojena se jmény Václava Příhody a Otakara Chlupa.

Příhoda se opírá o empiricky prokazované zjištění, že opakované hodnocení výkonu žáků (ať už týmž učitelem s časovým odstupem anebo skupinou učitelů) vykazuje nespornou variabilitu, což dokazuje přítomnost náhodné komponenty v hodnocení znalostí. Z toho vyvozuje, že (1) didaktický test by měl zdroj této nahodilosti kontrolovat a tím zkoušku objektivizovat, a (2) didaktický test je možno organizovat jako měření na vzorku ze souboru poznatků. Tento bod je možno považovat za parafrázi základního edukometrického para-

digmatu, na němž spočívá také autorův vlastní přístup, a který je východiskem jeho studií: každá zkouška je výběrem, vzorkem z rozsáhlejšího souboru, populace možných znalostí, které by bylo možno prověřovat; za jistých, pro pedagogickou situaci přijatelných okolností, splňuje tento výběr kritéria náhodného výběru — a lze proto na jeho výsledky aplikovat metody statistické indukce [3].

Příhoda byl, samozřejmě, ve svých přístupech k testování ovlivněn svou americkou zkušeností — předmluvu své knihy Psychologie a hygiena zkoušky vročuje do roku 1923 (české vydání je z roku 1924) v Madisonu, ve státě Wisconsin, USA [6].

Podstatou Chlupovy kritiky didaktického testování bylo tvrzení, že testy se soustřeďují na znalosti povrchní a jednotliviny, a že nejsou schopny postihovat hlubší znalosti tématu ani souvislostí, které jsou nepochybně podstatou poznání a vědění. Tato kritika je zásadní; aktuální zůstává dodnes, zejména v případě nestandardizovaných testů, konstruovaných ad hoc. Odpovědí na ni je taxonomie výukových cílů a různé systémy klasifikací úrovně znalostí (Bloom, Niemierko), soustavně rozpracované v praxi testování už desítky let. Profesionálně kontruované didaktické testy Chlupovu (a nejen jeho) námitku respektují [1].

Zhruba stejné důvody motivovaly českou učitelskou veřejnost k zaujetí dvojího stanoviska k didaktickým testům: progresivního a konzervativního. Prvý přístup odpovídal ve 30. letech zaměření a orientaci společenského a zejména hospodářského života na výkon (životní strategie u nás reprezentovaná především systémem Tomáše Bati) a na úspěch.

K výtkám didaktickému testování patřila také kritika, že podtrhuje standardní jednání, potlačuje individualitu a specifické rysy osobnosti.

Do vývoje názorů na didaktický test zasáhly i okolnosti irelevantní a iracionální, totiž politický vývoj.

V Sovětském svazu se od 20. let zkoušení ruší anebo alespoň jeho úloha výrazně snižuje (moje osobní zkušenost z vyprávění kolegyně z Lomonosovovy university: v poválečné generaci (myslí se 2. světová válka) akademické obce se užívalo rčení „toho člověka za sebe určitě ke zkoušce neposílali" o jedinci rozumu mdlého, byť akademicky vzdělaném — čímž mělo být řečeno, že si ho jeho studijní skupina na universitě nikdy nevybrala jako svého zástupce ke zkoušce; v kolektivistickém duchu oné doby bývalo zvykem, že studijní úspěch celé skupiny byl hodnocen podle úspěchu, kterého dosáhl vybraný zástupce; ostatní ke zkoušce nechodili). Ze statistického hlediska nazíráno, jde o extrémní případ uplatnění statistické indukce — tentokrát v roli očividně patologické, protože nejde o výběr náhodný, ale o úsudek z extrému na celek. Přestože bylo ve 30. letech od většiny pedagogických výstřelků, omezujících zkoušky, upuštěno, jakousi reminiscencí se tyto promítly do konce 40. a do 50. let českého školství, kdy bylo didaktické testování označeno za projev buržoasního elitářství a odsouzeno k odumření. Didaktické testy se tak ocitly na nějakou dobu v čestném společenství kybernetiky, populační mendeliánské genetiky, ekonometrie a jiných „buržoasních pavěd".

Je třeba konstatovat, že rezervovaný až odmítavý názor na smysl a možnosti didaktických testů přežívá v jisté části učitelské obce dodnes — kdy už irelevantní důvody jeho existence pominuly. K jeho recidivám přispívá ovšem často nekvalifikovaná aplikace didaktických testů, když se tyto stávají — podobně jako třeba statistika — nástrojem zkreslování, záměrné manipulace a dokazování předem určeného, misinterpretace anebo dokonce i přímé lži [1].

Vlastní práce autora je věnována především studiu testů rozlišujících (angl. norm-referenced tests, statistickonormativních podle Heluse), nikoli studiu testů ověřujících, kriteriálních (angl. criterion-referenced tests). Specifičnost autorova přístupu je určována především výchozím modelem: odpověď v testové položce se chápe jako nula-jedničková náhodná veličina (jde výhradně o testy typu multiple-choice), s týmž pravděpodobnostním rozdělením v souboru položek; odpovědi se přitom považují za vzájemně nezávislé [3]. Tato zjevně simplifikující představa umožňuje odvodit známé testové charakteristiky (např. koeficient

obtížnosti položky nebo index znalostí subjektu) jako parametry modelu, jejichž vlastnosti jsou statisticky popsatelné a pedagogicky smysluplně interpretovatelné. Se soudobými teoriemi didaktického testování (George Rasch, Frederic Lord) je v několika pracích model autora konfrontován a vzájemná souvislost je popsána [5, 8]. V tomto smyslu je autorský přístup jednou z modifikací „latent traits theory" – a představuje tak heuristickou proceduru, v níž je empiricky pozorovaná skutečnost (tj. testové skóre) vysvětlována pedagogicky interpretovatelnými proměnnými jako jsou obtížnost úlohy (položky testu) a úroveň znalostí subjektu. Podstatné přitom je, že se $N.n$ naměřených odpovědí (N = počet subjektů podstupujících test, n = počet položek testu) redukuje na $N + n$ latentních, teoretických, koncepčních veličin. Pro ilustraci stačí uvést, že např. pro $N = 30$ subjektů a $n = 20$ položek se $N.n = 600$ odpovědí vysvětluje pouhými $N + n = 50$ parametry.

O tom, že si pedagogika nemůže dovolit přehlížet možnosti kvantifikace a měření ve sféře svého zájmu, svědčí také citace ze základní monografie Educational Measurement, redigované Thorndikem: „Measure is one of the thousand most common words in printed English". Přitom, jak se uvádí, ve vzorku 2,5 miliónů slov se i slovo „measure" objevilo více než 400 krát a bylo použito ve více než 40 významech. To svědčí o tom, že i měření v pedagogice by mělo být chápáno dostatečně široce [7].

Nakonec ještě citát z výše zmíněné knihy Příhodovy [6]: „V nynějším stádiu vývoje (tj. v roce 1924 – pozn. S. K.) věd pedagogických nelze si již představit, aby pracovník vědecký ani školský nebyl obeznámen ani se základními pojmy statistickými. Není možno žádat ovšem, aby se každý učil operacím přesahujícím znalostí středoškolské matematiky. Není naštěstí v pomocné vědě statistické základních pojmů, které by šly za tyto hranice. Nutno zejména uměti vypočítati median a vážený aritmetický průměr z hodnot symbolisujících střední tendenci, nezbytno znáti metody tabulační a grafické, znázorňující formu distribuční, potřeba seznámiti se z jednotek variabilitních aspoň se čtvrtinovou odchylkou (Q), s výpočtem odchylky standardní (sigma) a se vzorci pro výpočet pravděpodobných chyb. Z metod vypočítávajících koeficient korelační (r, ro a R) postačí znalost metody Spearmanovy (pořadové) a vzorce Pearsonova-Bravaisova pro metodu násobkovou (product-moment formula)".

Tolik klasik české pedagogiky v roce 1924.

2. SPOLEHLIVOST KLASIFIKACE

Naše věda je zřejmě jenom jedním z možných způsobů výkladu a uchopení světa. Jestliže tedy připustíme alternativu ve způsobu pochopení, není důvodu váhat nad účelností kategorie přibližného.

2.1 Princip statistického rozhodování

Klasifikace znalostí žáků je počínání, v němž není obtížné rozpoznat prvky procesu rozhodování. Tato skutečnost může být podnětem k aplikaci principů statistického rozhodování, včetně vybudování modelové situace, v jejímž rámci je taková aplikace možná. Významným motivem je přitom úsilí ono rozhodování optimizovat, čímž je zpravidla míněna redukce „nespravedlivé" klasifikace, kdy jsou subjekty s touž úrovní znalostí klasifikovány více nebo méně rozdílně.

Předkládané úvahy se omezí na případ, kdy jsou znalosti hodnoceny na základě empirického měření založeného na testu typu multiple – choice [4].

2.2 Statistické pojetí klasifikace

Procedura klasifikace by měla uvažovat tři navzájem související struktury:

(a) Prostor Z možných úrovní z znalostí, které přicházejí v úvahu u subjektu, jehož

znalosti o daném tématu mají být klasifikovány. Předpokládá se, že tyto úrovně znalostí existují objektivně, přičemž měření examinátora jsou přístupné jenom nepřímo, přes výsledky testu, který zkoušený subjekt podstupuje.

Je respektována představa, že znalost tématu může mít rozličnou hloubku, například ve shodě s Bloomovou hierarchií úrovní znalostí: (1) zapamatování (2) porozumění, (3) analýza, (4) syntéza, (5) aplikace vědomostí o tématu, (6) řešení problémů s tématem souvisejících [1]. Diskutovaný model neuvažuje tyto úrovně explicitně; bere je v úvahu implicitně tím, jak promítá komplexnost a obtížnost do konstrukce testových položek.

(b) Skutečná úroveň znalostí, kterou má zkoušený subjekt o uvažovaném tématu, se projevuje v jeho odpovědích v testu; jinak řečeno, tyto odpovědi jsou na úrovni znalostí závislé.

Nechť S označuje prostor možných odpovědí zkoušeného subjektu, jimiž tento může v testu reagovat; s je prvek tohoto prostoru S. Je-li použito testu typu multiple – choice, je prostorem možných odpovědí zkoušeného subjektu množina hodnot testového skóre. Tyto odpovědi jsou jediným zdrojem informace o úrovni znalostí subjektu, přístupným examinátorovi. Zdrojem tím účinnějším, čím těsněji závisí odpověď na skutečné úrovni znalosti. A protože je tato závislost uchopitelná statisticky (ve formě podmíněných pravděpodobnostních rozdělení – jak uvidíme vzápětí), ji možno je také statistickými prostředky a metodami měřit.

Vyjádřeno formálněji, pravděpodobnostní rozdělení P(s|z) lze považovat za kanály, jimiž proudí informace z prostoru znalostí Z do prostoru odpovědí S. V nereálném, idealizovaném případě, kdy by úroveň znalostí byla přístupná přímému pozorování a evidenci, představovala by tato podmíněná pravděpodobnostní rozdělení vzájemně jednoznačná přiřazení odpovědí znalostem; šlo by tedy o případ kanálu bez šumu.

(c) Klasifikační procedura tak představuje soustavu rozhodovacích pravidel, přiřazujících každé odpovědi s z S právě jediný klasifikační stupeň (klasifikační kategorii) r klasifikační škály R. Znamená to, že klasifikační procedurou je zaváděn vztah mezi empirickou (examinační) evidencí S a klasifikací (rozhodnutím) R.

V pojmech experimentální psychologie to znamená, že prostor rozhodnutí R, stejně jako pravidla klasifikace jsou kontrolována examinátorem, zatímco prostor znalostí Z a prostor odpovědí S jsou kontrolovány zkoušeným subjektem.

Klasifikační procedura tak představuje exhaustivní (vyčerpávající) a jednoznačný rozklad prostoru odpovědí S do konečné množiny m vzájemně se nepřekrývajících podmnožin s_1, $s_2,\ldots$, s_m tak, aby byl zkoušený subjekt klasifikován stupněm r_i, jestliže jeho odpověď byla prvkem podmnožiny s_i, $i = 1,2,\ldots, m$.

Při hledání optimální klasifikační procedury lze aplikovat představy známé z Neymanovy a Pearsonovy koncepce testování statistických hypotéz [3].

Pro daná klasifikační pravidla se zavádí soustava tzv. operačních charakteristik klasifikace, což je m funkcí $P(r_i\,|z)$, $i = 1,2,\ldots,m$, každá z nich odpovídající právě jednomu z m klasifikačních stupňů r_i. Argumentem každé takové funkce je skutečná úroveň znalostí z;svých hodnot nabývá funkce v oboru $\langle 0, 1\rangle$ jako podmíněná pravděpodobnost, že zkoušený subjekt bude klasifikován stupněm r_i, tj., že jeho odpověď bude prvkem podmnožiny odpovědí s_i. Pro danou hodnotu z tvoří m odpovídajících hodnot $P(r_1|z)$ (podmíněné) pravděpodobnostní rozdělení, protože $P(r_1|z) + \ldots + P(r_m|z) = 1$.

S klasifikací je spojen problém její spolehlivosti. Zdrojem nespolehlivosti je přitom

(1) diskretizace, event. zhrubení spojité, případně detailněji vzorované škály odpovědí S do systému podmnožin $\{s_1, s_2,\ldots, s_m\}$ a

(2) neurčitost existující ve vztahu mezi úrovní znalostí z a odpovědí s, což způsobuje, že v množině subjektů s touž úrovní znalostí z nebudou všichni klasifikování stejně; tak se vynořuje otázka „klasifikační nespravedlnost“ co nejvíce redukovat.

2.3 Binomické rozdělení odpovědí

Takto chápaný model rozhodování může být hlouběji formalizován v případě, kdy se měření znalostí opírá o test typu multiple − choice a kdy se odpověď zkoušeného subjektu umisťuje na stupnici testového skóre.

Aktuální znalost subjektu o předmětu zkoušky představuje v modelu pravděpodobnost p jeho nesprávné odpovědi. Nízké hodnoty p (blízké nule, případně blízké „statistické nule", má-li v testu své místo možné dosažení správného řešení úlohy náhodným uhádnutím) odpovídají lepší znalosti, vyšší hodnoty p (blízké jedné) odpovídají špatné znalosti.

Předpokládá se, že test je tvořen souborem n vzájemně nezávislých položek (úloh) dané úrovně obtížnosti. S každou testovou úlohou se zkoušenému předkládá q alternativ odpovědí (různých řešení), z nichž právě jediná je správná a zbylých $q-1$ nabídek plní funkci distraktorů.

Odpovědí subjektu v testu jako souboru n položek je počet $k(0 \leqq k \leqq n)$ nesprávných odpovědí (k nesprávně řešených úloh) z n úloh předložených − tj. hrubé skóre nesprávných odpovědí. Znamená to, že podmíněné pravděpodobnosti odpovědí subjektu, kdy podmínkou je úroveň znalostí, nabývá formy binomického rozdělení

$$P(k|p) = \binom{n}{k} p^k (1-p)^{n-k}, (k = 0,1,\ldots, n), \quad 0 \leqq p \leqq 1 \tag{1}$$

Klasifikační pravidlo rozkládá množinu $\{0,1,\ldots, n\}$ možných odpovědí zkoušeného subjektu do m disjunktních podmnožin tak, aby každý z m klasifikačních stupňů byl přiřazován právě jediné z těchto podmnožin (vzájemně jednoznačné zobrazení). To pak umožňuje odvodit příslušný soubor m operačních charakteristik klasifikace.

2.4 Operační charakteristiky klasifikace

Soustava operačních charakteristik klasifikace je ovlivňována nejen rozsahem testu n, počtem používaných klasifikačních stupňů m, ale také tím, jakým způsobem byl proveden rozklad škály skóre chybných odpovědí do soustavy $\{s_1, s_2,\ldots, s_m\}$. Operační charakteristiky jsou tak definovány vzorci

$$P(r_i|p) = \sum_{k \in s_i} \binom{n}{k} p^k (1-p)^{n-k} \quad (i = 1,2,\ldots, m) \tag{2}$$

Testovanému výsledku s_1 (malý počet chyb) je zřejmě přiřazován nejlepší klasifikační stupeň, zatímco testový výsledek s_m (velký počet chyb) se oceňuje nejhorším klasifikačním stupněm m.

Obr. 1 − 6 zachycují soustavy operačních charakteristik v situacích, kdy jsou prověřovány znalosti pomocí testu s volbou z nabídnutých odpovědí (multiple − choice test) s $n = 10$ či 30 položkami a $m = 2,3$, případně 4 klasifikačními stupni. Klasifikační pravidla jsou definována v popisu každého grafu.

Pro každou hodnotu p (na vodorovné ose) udávají křivky operačních charakteristik, jaký podíl subjektů (jejichž znalostem odpovídá pravděpodobnost chybné odpovědi p) bude klasifikován tím kterým stupněm. To, že jedinci s touž úrovní znalostí mohou být klasifikováni rozdílně, je nepochybně nedostatkem procesu klasifikace − průvodním znakem jeho nedokonalosti, která může být interpretována jako projev nespravedlnosti hodnocení a tedy křivdy.

Vliv rozsahu testu n na nespravedlnost klasifikace je zřejmý: při daném počtu klasifikačních stupňů m jsou v případě rozsáhlejšího testu křivky operačních charakteristik strmější, díky čemuž se jednotlivé klasifikační stupně váží na znalosti specifičtěji.

Závislost operačních charakteristik a tím také nejednoznačnosti klasifikace na počtu použitých klasifikačních stupňů m není tak přímočaře interpretovatelná. Hodnocení a porovnání musí brát v úvahu, že klasifikační stupně jsou úrovněmi ordinálního znaku, a že „vzdálenost" mezi stupni „1" a „2" ve dvoubodové škále ($m = 2$) je podstatnější nežli „vzdálenost" mezi stupni „1" a „2" ve škále čtyřbodové ($m = 4$).

Z obr. 1 a 4 je zřejmé, že nejslabším místem klasifikace je ta úroveň znalostí p, pro niž nabývají obě operační charakteristiky téže hodnoty 0,5. Polovina subjektů s touto úrovní znalostí (odpovídajících v testu chybně s pravděpodobností přibližně $p = 0{,}2$ a ovládajících tedy asi 60 % zkoušené látky) bude klasifikována stupněm „1" a polovina stupněm „2". Tento nedostatek nebude zvětšením rozsahu testu z 10 na 30 úloh odstraněn — bude se však týkat stále menší podmnožiny subjektů.

V případě $m = 3$ klasifikačních stupňů je takovým „bolavým" místem klasifikace skupina subjektů s $p = 0{,}2$ (při aplikaci testu o $n = 10$ položkách), s $p = 0{,}2 - 0{,}4$ (při aplikaci testu o $n = 30$ položkách).

V případě $m = 4$ klasifikačních stupňů se „nespravedlnost" klasifikace zmírňuje a rozprostírá na širší oblast znalostí subjektů.

2.5 Entropie jako kritérium neurčitosti klasifikace

Pro každou úroveň znalostí p tvoří hodnoty operačních charakteristik $P(r_i|p)$, $i = 1,2,\ldots, m$, rozdělení pravděpodobností. Je žádoucí, aby toto rozdělení bylo maximálně nerovnoměrné, tj. aby co nejvíce jedinců bylo hodnoceno právě jediným z klasifikačních stupňů. Obráceně, čím rovnoměrnější je toto rozdělení, tím vyšší je neurčitost klasifikace v souboru jedinců s touž úrovní znalostí. Maxima se přitom dosahuje při $P(r_i|p) = 1/m$ pro každé i (je-li klasifikační škála chápána jako škála nominálního znaku), případně při $P(r_1|p) = P(r_m|p) = 1/2$ (je-li klasifikační škála chápána jako škála ordinální).

Neurčitost klasifikace je možno vyjádřit kondenzovanějším způsobem, tak, abychom místo souboru m operačních charakteristik vystačili s jedinou statistikou (nezávisle na počtu klasifikačních stupňů m).

Takovou statistikou (odpovídající nominálnímu pojetí klasifikační škály) je entropie, definovaná vztahem

$$H(p) = -\sum_{i=1}^{m} P(r_i|p) \log P(r_i|p), \tag{3}$$

a nabývající svých hodnot v intervalu $(0, \log m)$. Po provedení standardizace pak funkce z entropie odvozená, totiž

$$M(p) = 1 - \frac{H(p)}{\log m}. \tag{4}$$

Funkce (4) je schopna měřit spolehlivost (spravedlnost) klasifikace na stupnici (0,1) — přičemž hodnota 0 znamená minimální spolehlivost a hodnota 1 maximální spolehlivost klasifikace.

2.6 Distance jako kritérium neurčitosti klasifikace

Spíše než nominální má klasifikační škála povahu ordinální. Je proto vhodnější statistikou neurčitosti rozdělení $P(r_i|p)$, $i = 1,2,\ldots, m$, distance, schopná vzít v úvahu uspořádání stupňů klasifikační škály.

Distanci zavádíme vztahem

$$\begin{aligned} d(p) = {} & P(r_1|p)\,P(r_2|p) + \ldots + P(r_{m-1}|p)\,P(r_m|p) + \\ & + 2\{P(r_1|p)\,P(r_3|p) + \ldots + P(r_{m-2}|p)\,P(r_m|p)\} + \ldots \\ & + (m-1)\,P(r_1|p)\,P(r_m|p)\,. \end{aligned} \tag{5}$$

Funkce (5) nabývá svého minima $d(p) = 0$ (v případě, kdy $P(r_i|p) = 1$ pro některé i a nula pro i ostatní) a svého maxima $d(p) = (m-1)/4$ (v případě, kdy $P(r_1|p) = P(r_m|p) = 1/2$ a ostatní pravděpodobnosti jsou rovny nule).

Jako míra spolehlivosti klasifikace může proto sloužit statistika odvozená z (5) standardizací, totiž

$$D(p) = 1 - d(p)\,\frac{4}{m-1}\,, \tag{6}$$

jejíž minimální hodnta 0 znamená nejnižší a maximální hodnota 1 nejvyšší možnou spolehlivost klasifikace.

Křivky spolehlivosti kalsifikace (6) jsou zachyceny pro uvažované strategie na obr. 7 a 8.

Do průběhu křivek distance se promítají vlastnosti operačních charakteristik. Žádoucí je, aby křivky ležely co nejvýše, především v té oblasti znalostí p, kam patří největší počet zkoušených subjektů. Při použití pouze dvouhodnotové klasifikační škály ($m=2$) se zřejmě nevyhneme tomu, aby byla určitá skupina subjektů klasifikována výrazně nespravedlivě. Vyšší počet použitých klasifikačních stupňů tuto nespravedlnost zmírňuje – a rozprostírá ji na širší okruh subjektů. Testy většího rozsahu okruh takto postižených subjektů prokazatelně omezují.

ZÁVĚR

Statistická indukce, založená na paradigmatu populace a vzorku, z ní vybraného, který je studován a z jehož vlastností se usuzuje na vlastnosti populace, nabízí své metody také pedagogice. Metody takto používané a metodologie jejich aplikace bývají označovány názvem edukometrie.

Předkládaný text diskutuje některé klíčové momenty edukometrie. Podrobněji demonstruje její možnosti v případě klasifikace založené na testu volby z nabídnutých odpovědí, kde nabízí metodu kvantitativního hodnocení spolehlivosti procedury klasifikace, která je ve školní praxi pedagoga chlebem vezdejším.

LITERATURA

Byčkovský P.: *Základy měření výsledků výuky. Tvorba didaktického testu.* Praha, ČVUT 1982.

Hniličková J., Josífko M., Tuček A.: *Didaktické testy a jejich statistické zpracování.* Praha, SPN 1972.

Komenda S., Klementa J.: *Analýza náhodného v pedagogickém experimentu a praxi.* Praha, SPN 1981.

Komenda S., Mazuchová J.: *Reliability of the Test Supported Classification.* In: Referate des 21. Int. Symp. „Ingenieurpädagogik 92". Klagenfurt 1992.

Lord F. M.: *Application of Item Response Theory to Practical Testing Problems.* Hillsdale 1980.

Příhoda V.: *Psychologie a hygiena zkoušky.* Praha 1924.

Thorndike R. L. (ed.): *Educational Measurement.* Washington 1971.

Wright B. D., Stone M. H.: *Best Test Design.* Chicago 1979.

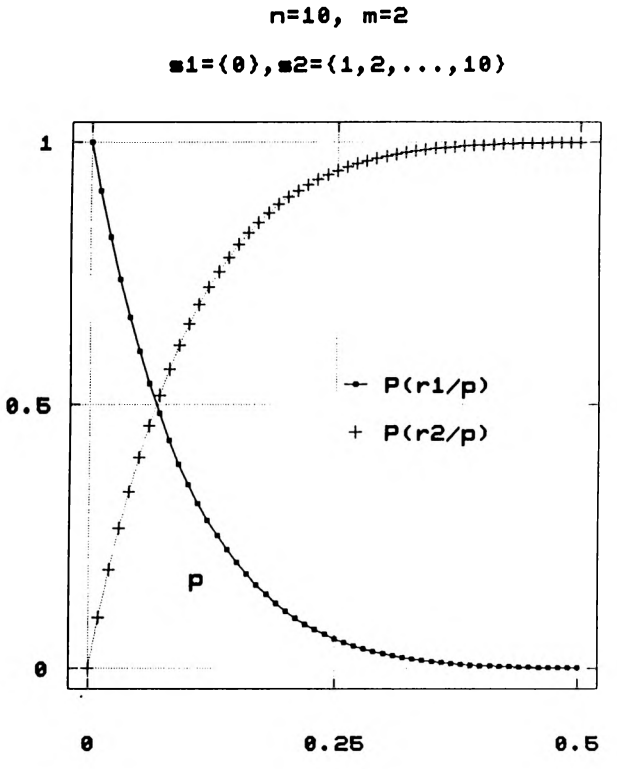
STANISLAV KOMENDA

EDUCATION AND EDUCOMETRICS

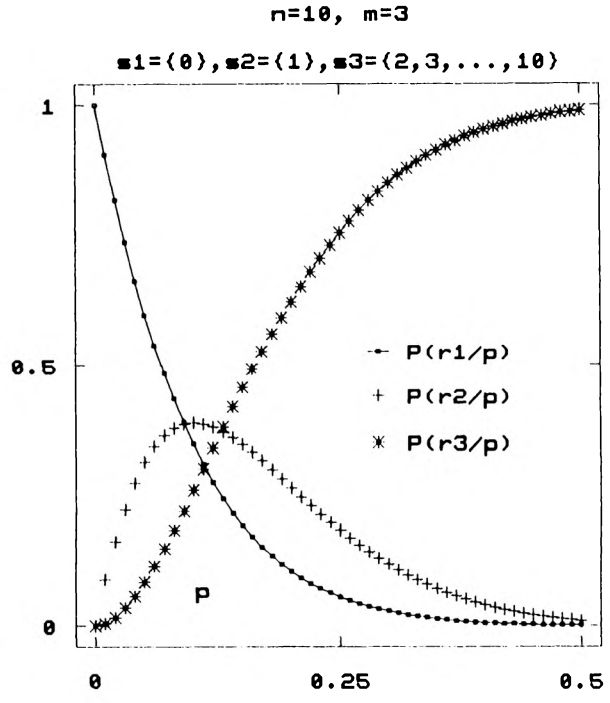
Learning and teaching as well are processes reasonable control of which is supported by the information feedback. Knowledge assessment can be incorporated as a part of this self-control or control process. In this paper the possibility is studied how to formalize this assessment in the case when the measurement of the knowledge of the subject to consider is supported by the school--achievement test of the multiple-choice type. The relation between the actual subject's knowledge and his/her response in the test is given by means of the binomial distribution and the assessment reliability is expressed by the set of operation characteristics. The possibility to apply entropy and distance of a frequency distribution as the means of the assessment efficiency evaluation is demostrated. Functioning of the method is proved in the cases of some multiple-choice tests of the size and number of distractors commonly used.

Došlo do redakce: 11. 6. 1993

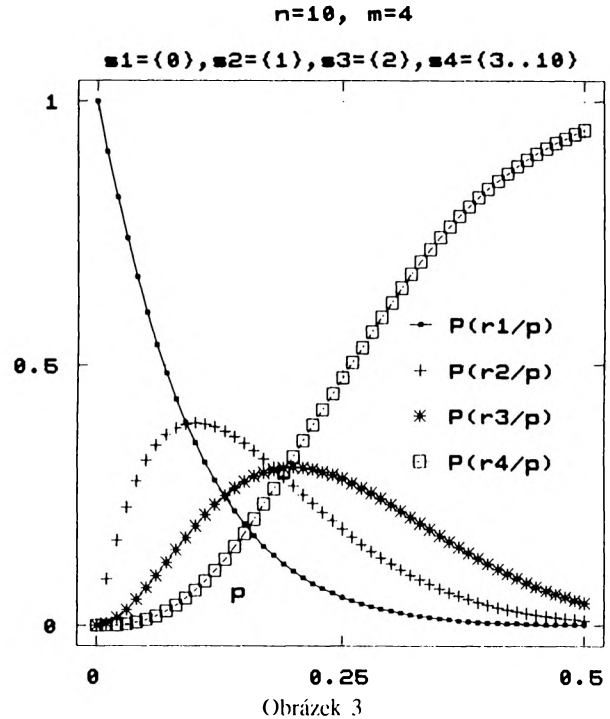
Autor: Doc. RNDr. STANISLAV KOMENDA, DrSc., Ústav lékařské biofyziky, biometrie a informatiky lékařské fakulty Univerzity Palackého, Hněvotínská 3, 775 15 Olomouc



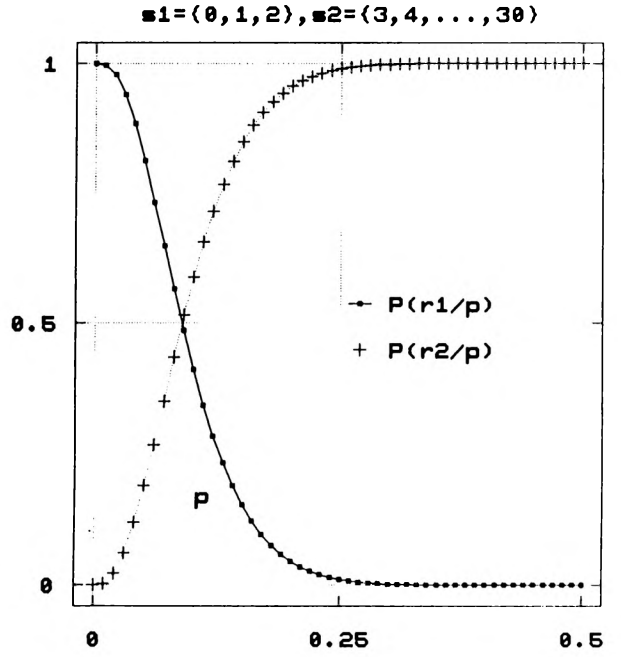
Obrázek 1



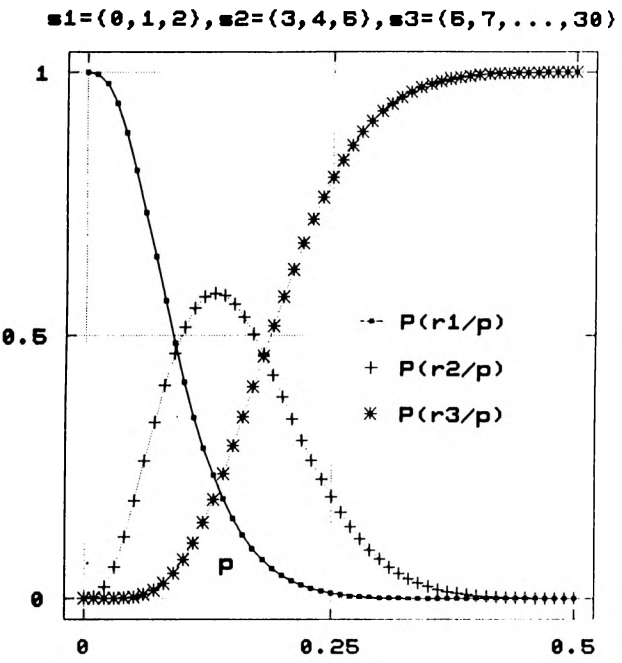
Obrázek 2



Obrázek 3



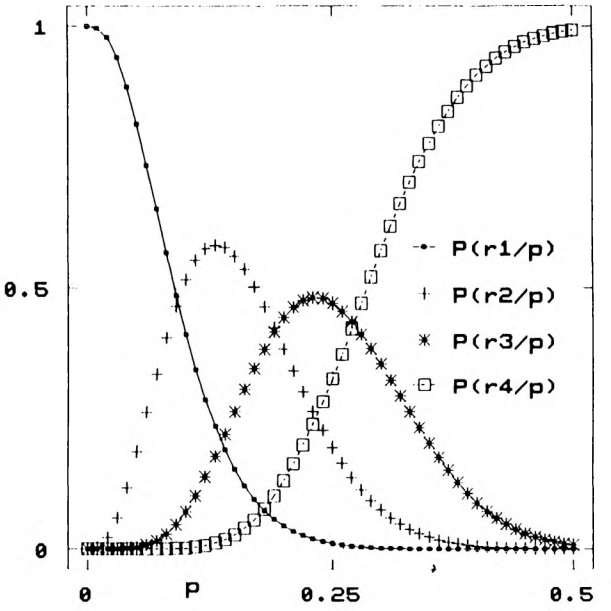
Obrázek 4



Obrázek 5

n=30, m=4

s1=(0,1,2),s2=(3,4,5),s3=(6,7,8),s4=(9,..,30)



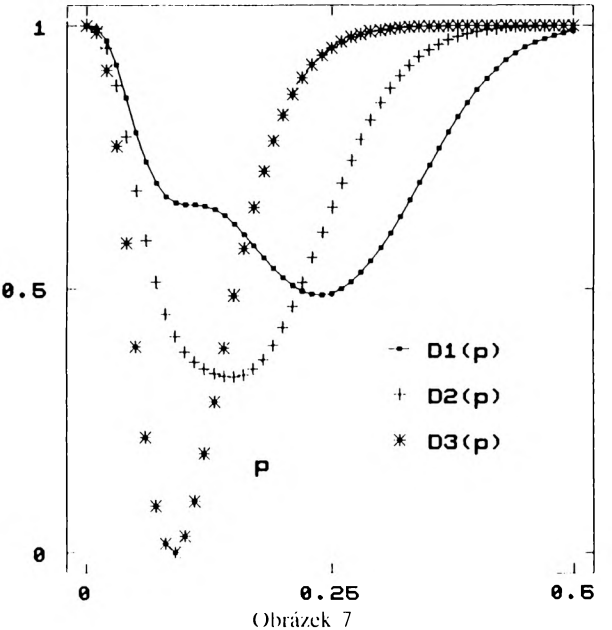
Obrázek 6

n=30

D1: s1=(0,1,2),s2=(3,4,5),s3=(6,7,8),s4=(9..30

D2: s1=(0,1,2),s2=(3,4,5),s3=(6,7,...,30)

D3: s1=(0,1,2),s2=(3,4,...,30)



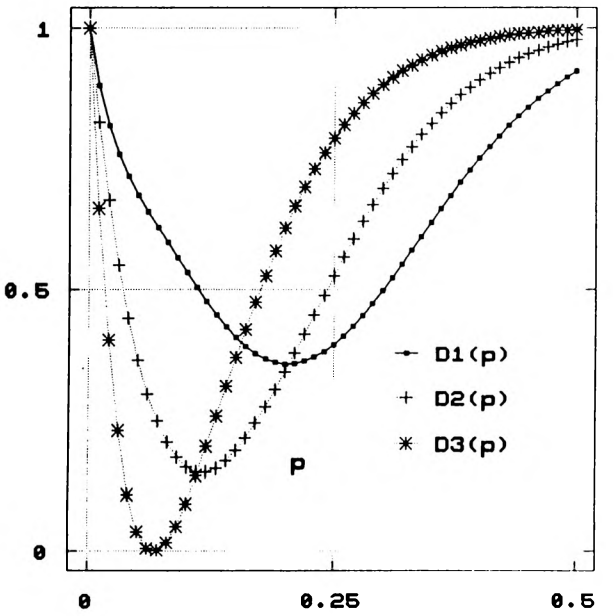
Obrázek 7

n=10

D1: s1={0},s2={1},s3={2},s4={3,4,...,10}

D2: s1={0},s2={1},s3={2,3,...,10}

D3: s1={0},s2={1,2,...,10}



Obrázek 8