Metodické studie

ANALÝZA RELIABILITY KLASIFIKAČNÍCH PROCEDUR – PRAVDĚPODOBNOSTNÍ MODEL CHYBY V KLASIFIKACI

PETR BOSCHEK
Filozofická fakulta UK, katedra psychologie

ABSTRACT

Analyses of classification procedures reliability – probabilistic model of classification error

P. Boschek

It is known that there is a very powerful and detailed methodology for a reliability analysis of tests that are used for the measurement in psychological diagnostics. However, there are also various diagnostic classification procedures in which observed variable is measured only on a nominal scale. For such classifications, the information about their reliability is available very rarely. Of course, one reason for this is the fact that there is no sufficiently developed methodology for the nominal classification, unlike for the test measurements. The analysis of reliability in these cases is usually based on the Cohen-Fleiss kappa conception, i.e. the concordance analysis of two or more replications of classification. The aim of this paper is to present new, innovative methods of the analysis of reliability of classification procedures, based on a probabilistic model of error in classification. This model is based on analogous principles as the „true-error" model of classical test theory.

key words:

reliability of classification,
error of classification,
reliability of nominal scales,
probabilistic model of error in classification

klíčová slova:

spolehlivost klasifikace,
chyba klasifikace,
reliabilita nominální škály,
pravděpodobnostní model chyby v klasifikaci

ÚVOD

Nezbytnou součástí metodologie psychologického výzkumu zejména v oblasti psychologické a samozřejmě i lékařské diagnostiky je konstrukce měřících, škálovacích či klasifikačních procedur s vysokým stupněm reliability. Jedině tak lze spolehlivě analyzovat sledované psychické jevy pro účely formulace psychologických zákonitostí, resp. stanovit s dostatečnou spolehlivostí diagnostickou kategorii testovaných jedinců.

Základy systematického přístupu řešení této problematiky jsou spojeny se jménem Charlese Spearmana, který již v roce 1910 postuluje rozklad pozorovaného (testového) skóru X na dvě stochasticky nezávislé složky, skutečný skór T (= true score) a chybu měření E (= error of measurement). Tento postulát ($X = T + E$) a potažmo odvozený rozklad variance observovaného skóru představuje princip klasického testového modelu, který poskytuje metodiku analýzy reliability testových procedur.

Došlo: 10. 10. 2011; P. B., Filozofická fakulta UK, katedra psychologie, nám. J. Palacha 2, 116 38 Praha 1; e-mail: boschek@centrum.cz
Studie vznikla za podpory grantového projektu GAČR č. 406/09/0424.

Předpokladem pro aplikaci klasického testového modelu je samozřejmě intervalový charakter nejen pozorovaného skóru, ale i jeho latentních aditivních složek. Tento předpoklad zaručuje smysluplnost rozkladu variance pozorovaného skóru na dílčí složky odpovídající skutečnému skóru a chybě měření a umožňuje tak definovat míru reliability (koeficient reliability, index reliability). Bodové a intervalové odhady těchto měr reliability získáváme buď na základě Pearsonova koeficientu korelace při strategii paralelních observací (dependabilita, ekvivalence, objektivita) nebo na základě modelu analýzy variance (odhad vnitřní konzistence) v případě, kdy analyzovaný testový skór je lineární kombinací skórů položek testu. K tomu, abychom mohli získat validní induktivní soudy o příslušných populačních parametrech, přistupuje navíc požadavek normálního rozdělení složek *X, E*.

V psychologické diagnostice se však vyskytuje poměrně veliké množství testových a škálovacích procedur, které nesplňují předpoklad intervalové škály. V případě, že pozorovaný skór má pouze ordinální charakter (známkové klasifikace, Likertovy a rozličné jiné posuzovací škály atp.), nelze charakteristiky (míry) reliability vůbec definovat. Pořadové statistiky, které se zde pro účely kvantifikace stupně reliability používají, mají charakter pouhého indikátoru bez adekvátního empirického relačního systému. Navíc induktivní metody se v těchto případech zužují na pouhý test hypotézy pořadové nezávislosti, který se navíc často komplikuje velkým počtem shodných pořadí u škál s malým rozsahem. V případě dvojnásobné klasifikace se v těchto datových situacích používá převážně Kendallův, případně Spearmanův korelační koeficient. V případě více replikací klasifikační procedury je možno (se stejnými výhradami) užít Kendallův koeficient shody (konkordance), který je modifikací Friedmanovy „pořadové analýzy rozptylu".

Zcela odlišnou kategorii představují klasifikační procedury, kdy pozorovaný statistický znak má pouze nominální charakter (např. stanovení diagnózy, pozitivita-negativita testu, dichotomické klasifikace v rozličných výběrových řízeních atp.), tedy zařazení osob do disjunktních diagnostických kategorií. I v těchto případech je potřebné, stejně jako u testových měření, mít k dispozici informaci o reliabilitě takové klasifikace, neboť je známo, že stupeň reliability ovlivňuje potažmo i výsledky statistické analýzy. Skutečnost je bohužel taková, že i u běžně používaných klasifikačních procedur se informace o jejich reliabilitě vyskytuje zcela vzácně. V těchto výjimečných případech je analýza reliability většinou založena na známém Fleiss-Cohenově kappa modelu, tedy odhadu příslušné varianty kappa-koeficientu konkordance na základě dvou či více nezávislých replikací klasifikace. Replikaci může představovat model „test-retest", avšak asi nejčastější případ, který se zde vyskytuje, představují nezávislé klasifikace dvou, nebo více hodnotitelů, tedy analýza objektivity klasifikace, která je v těchto případech zřejmě nejpodstatnější komponentou reliability.

Cílem této stati je prezentace netradiční metody analýzy reliability klasifikačních procedur založené na pravděpodobnostním modelu chyby v klasifikaci. Tato metoda poskytuje oproti klasické koncepci Fleiss-Cohenovy kappa vydatnější informaci o stupni reliability analyzované klasifikační procedury, a co je asi nejpodstatnější, získané statistiky lze také lépe interpretovat.

Při prezentaci modelu chyby v klasifikaci budeme z didaktických důvodů postupovat systematicky. Princip metody a potřebný nástin teorie vysvětlíme úvodem na nejjednodušším modelu dichotomické klasifikace, a to nejprve pro případ dvou nezávislých klasifikací. Posléze tento model zobecníme na obecný případ $R \geq 2$ klasifikací. Již na tomto místě je užitečné připomenout, že princip obecného modelu, tzn. případ multinomické klasifikace s obecným počtem replikací je zcela analogický. Samozřejmě se v takové datové situaci nesmírně komplikuje indexové označení

používané v popisu jednotlivých komponent modelu, což je nevýhodné pro úvodní prezentaci principu. Kromě této skutečnosti je účelné již úvodem se zmínit o dalším důležitém aspektu, který ovlivňuje formální způsob prezentace metody v této stati. Jak bývá obvyklé u vícerozměrných statistických metod, je výpočetní postup inferenčních metod velice komplikovaný. Kromě úvodního nejjednoduššího modelu představuje odhad parametrů řešení soustavy nelineárních rovnic. To je možné pouze s použitím vhodných numerických metod (tzn., že výpočet odhadů nelze vyjádřit přímo analyticky). Pro tyto účely byl vytvořen program, který kompletně řeší výpočet všech inferenčních statistik a jehož výstupy u prezentovaných příkladů využijeme. V závěrečné části této stati je uveden nástin nejdůležitějších algoritmů, které tento program obsahuje. Jednotlivé specifikované modely budou bohatě ilustrované konkrétními příklady. Současně budou tyto datové situace analyzované také pomocí kappa-metodologie, abychom mohli porovnat informační přínos obou metod.

Pravděpodobnostní model chyby v dichotomické klasifikaci

Úvodem je účelné zde krátce (i zjednodušeně) připomenout známý princip stochastického modelu klasické testové teorie, abychom si uvědomili silnou ideovou analogii obou modelů. Klasický testový model je založen na postulátu $X = T + E$, tedy vyjádření měřeného skóru X jako součtu dvou nezávislých náhodných veličin, pravého (True) skóru T a chybové složky (Error) E. Předpoklad nezávislosti těchto komponent umožnuje definovat míru reliability a na základě nezávislých opakování získat její výběrový odhad, směrodatnou odchylku chybové komponenty, a tím potažmo i konfidenční interval pro pravý testový skór.

Pravděpodobnostní model chyby v klasifikaci, zcela analogicky jako klasický testový model, postuluje distribuci observované kategorie X nominální klasifikace jako funkci pravé (skutečné) kategorie T a podmíněné chybové složky $X \mid T$, která kvantifikuje stupeň reliability této klasifikace. Tento princip je v našem případě dichotomické klasifikace, tedy klasifikace do dvou kategorií označených (1, 2), patrný ze vztahu distribuce těchto náhodných proměnných, což není nic jiného, než známý vzorec pro úplnou pravděpodobnost.

$$P(X = 1) = P(T = 1)P(X = 1 \mid T = 1) + P(T = 2)P(X = 1 \mid T = 2)$$
$$P(X = 2) = P(T = 1)P(X = 2 \mid T = 1) + P(T = 2)P(X = 2 \mid T = 2)$$

resp.

$$P(X = i) = \sum_{k=1}^{2} P(\mathrm{T} = k)P(X = i) \mid T = k),$$

kde jednotlivé složky představují (pro $i,k = 1,2$):

$P(X = i)$ rozdělení pravděpodobnosti pozorované (klasifikované) kategorie.
$P(T = k)$ rozdělení pravděpodobnosti skutečné („true“) resp. pravé kategorie.
$P(X = i \mid T = k)$ podmíněné rozdělení pravděpodobnosti pozorované kategorie podmíněno skutečnou kategorií (vyjádření chybové složky klasifikace).

Zde např. první z prezentovaných vztahů vyjadřuje pravděpodobnost klasifikace objektu do první kategorie. To může nastat dvěma způsoby: bezchybnou klasifikací, pokud první kategorie je „pravou“ kategorií, nebo chybnou klasifikací do „nepravé“ první kategorie. Vzhledem k tomu, že tyto možnosti představují disjunktní jevy, lze pravděpodobnost výsledku klasifikace vyjádřit v uvedené součtové podobě (analogicky druhý vztah).

Je zřejmé, že vysoká reliabilita takové klasifikace je vyjádřená nízkými hodnotami podmíněných pravděpodobností chyby klasifikace $P(X=i \mid T = k)$, pro rozdílné kategorie $i \neq k$, a naopak. V extrémním případě nulových pravděpodobností by se jednalo o totální reliabilitu, což znamená ekvivalentní distribuci observovaných a pravých kategorií.

Opět analogicky, jako v případě klasické teorie testů, je zřejmé, že přímý odhad pravděpodobnostní funkce můžeme získat pouze u pozorovaných kategorií, a pro získání výběrových odhadů distribuce T a $X \mid T$ je třeba mít k dispozici data z nezávislých replikací těchto klasifikací. Jak již bylo řečeno, budeme se pro snadnější orientaci v dané problematice úvodem zabývat nejjednodušším případem dvou nezávislých klasifikací (v praxi asi nejčastěji nezávislá hodnocení resp. klasifikace dvou hodnotitelů). V textu později tento předpoklad rozšíříme na obecný počet klasifikací, který samozřejmě poskytuje podstatně vyšší míru informace pro vlastní analýzu reliability.

Vstupní data v případě dvou klasifikací (dvojnásobná replikace) představují klasické McNemarovské schema, tedy četnosti n_{ij}, $i,j = 1,2$ sdružené klasifikace např. dvou hodnotitelů (X_1, X_2) ve čtyřpolní kontingenční tabulce. Tyto výběrové četnosti pocházejí z populačního dvourozměrného rozdělení pravděpodobností $P(X_1 = i, X_2 = j)$ pro $i,j = 1,2$.

Za předpokladu, že „podmíněné" klasifikace $X_1 \mid T$ a $X_2 \mid T$ jsou stejně rozdělené a nezávislé (analogie klasicky paralelních testů), tedy $P(X_1 \mid T) = P(X_2 \mid T)$ a

$P(X_1 \cap X_2 \mid T) = P(X_1 \mid T).\ P(X_2 \mid T)$, lze pravděpodobnostní funkci tohoto dvojrozměrného (dichotomického) rozdělení vyjádřit jako funkci rozdělení pro analýzu reliability podstatných komponent, pravé kategorie T a podmíněné klasifikace $X \mid T$ takto:

$$(1.1) \qquad P(X_1 = i, X_2 = j) = \sum_{k=1}^{2} P(T = k) P(X_1 = i \mid T = k) P(X_2 = j \mid T = k) \qquad i, j = 1,2$$

Uvedený vztah představuje základ pravděpodobnostního modelu pro analýzu reliability dichotomické klasifikační procedury. Samozřejmě pro teoretické účely (odvození odhadů parametrů), avšak nikoliv pro pochopení principu modelu, je užitečné prezentovaný vztah dále modifikovat, resp. algebraicky upravit. Důvodem pro tuto úpravu je fakt, že v diskrétním rozdělení pravděpodobností je vždy jedna pravděpodobnost vázaná (součet pravděpodobností je vždy roven = 1).

Pozn: v dalším textu budeme z důvodu zjednodušení zápisu používat následující označení. Vzhledem k tomu, že se zde jedná o populační charakteristiky (parametry modelu), použijeme standardně řeckou abecedu:

$\pi_{ij} = P(X_1 = i, X_2 = j)$, $i,j = 1,2$; *sdružené rozdělení observovaných kategorií*
$\tau_k = P(T = k)$, $k = 1,2$; *rozdělení pravé kategorie*
$\varepsilon_{ks} = P(X = s \mid T = k)$, $k,s = 1,2$; *podmíněné rozdělení $X \mid T$ chyby ($k \neq s$) v klasifikaci)*

Vzhledem k tomu, že inferenční statistiky (bodové odhady, intervalové odhady a testy hypotéz o shodě dat s modelem) jsou závislé na případných dalších předpokladech týkajících se parametrů tohoto modelu, budeme se jednotlivým specifikacím modelu věnovat postupně.

Model A: Předpoklad shodné pravděpodobnosti obou chyb v klasifikaci

V úvodu se budeme zabývat numericky nejjednodušším modelem dichotomické klasifikace, který předpokládá shodnou pravděpodobnost chybné klasifikace do obou uvažovaných kategorií $\varepsilon_{12} = \varepsilon_{21}$, tedy $P(X = 2 \mid T = 1) = P(X = 1 \mid T = 2)$. Tento předpoklad nepředstavuje nic jiného, než analogii předpokladu nezávislosti složek E, T v klasickém testovém modelu, tedy shodu distribuce chybové složky E na všech úrov-

ních pravého skóru *T*. Je zřejmé, že pro dosti širokou škálu rozličných klasifikačních procedur není uvedený předpoklad příliš omezující, avšak jeho přijetí je samozřejmě nutné pečlivě uvážit v závislosti na konkrétní datové situaci.

Jestliže budeme uvedený předpoklad akceptovat, strategie analýzy se výrazně zjednoduší, neboť se v modelu sníží počet neznámých (volných) parametrů. V tomto případě lze získat maximálně věrohodné odhady obou neznámých populačních parametrů $\varepsilon = \varepsilon_{12} = \varepsilon_{21}$

a $\tau = P(T = 1)$ přímo, bez použití složitých numerických metod (podrobněji viz Anděl, 2002). Příslušné odhady jsou standardně označeny pomocí stříšky

$$(2.1) \qquad \hat{\varepsilon} = \frac{1 - \sqrt{1 - \frac{2(n_{12} + n_{21})}{n}}}{2}, \qquad \hat{\tau} = \frac{n_{11}(1-\hat{\varepsilon})^2 - n_{22}\hat{\varepsilon}^2}{(n_{11} + n_{22})(1 - 2\hat{\varepsilon})}$$

Rozptyl odhadu pravděpodobnosti chyby klasifikace $\hat{\varepsilon}$ se potažmo získá na základě vztahu

$$(2.2) \qquad s^2(\hat{\varepsilon}) = \left(n\frac{4(\hat{\varepsilon} - \hat{\tau})^2}{\hat{\tau} - 2\hat{\tau}\hat{\varepsilon} + \hat{\varepsilon}^2} + 2n\frac{(1-2\hat{\varepsilon})^2}{\hat{\varepsilon} - \hat{\varepsilon}^2} + n\frac{4(\hat{\varepsilon} + \hat{\tau} + 1)^2}{(1-\hat{\varepsilon})^2 - \hat{\tau} + 2\hat{\tau}\hat{\varepsilon}} \right)^{-1}$$

Z teorie ML-odhadu plyne, že rozdělení odhadu pravděpodobnosti chyby klasifikace konverguje v distribuci k normálnímu rozdělení $N[\, \varepsilon\, ,\, s^2(\varepsilon)\,]$ (důkaz viz Anděl, 2002). To umožňuje konstrukci intervalu spolehlivosti pro neznámou populační pravděpodobnost chyby klasifikace:

$$CI(95\%): \quad \hat{\varepsilon} \pm 1{,}96 s(\hat{\varepsilon})$$

Postup analýzy budeme nyní prezentovat na následujícím reálném příkladu.

Příklad 1: Ve dvoustupňovém přijímacím řízení na jisté vysoké škole v ČR s uměleckým zaměřením představuje první část přijímací procedury talentovou zkoušku, ve které jsou uchazeči klasifikováni dichotomicky (1 = postupující, 2 = nepostupující). Pro účely analýzy reliability (objektivity) této klasifikace hodnotila dvojice pedagogů nezávisle výsledek talentové zkoušky (X_1, X_2) u souboru n = 155 uchazečů. Empirické četnosti sdružených klasifikací prezentuje čtyřpolní tabulka (tab. 1).

Tab. 1 Empirické četnosti klasifikace dvou pedagogů

KLAS	$X_2=1$	$X_2=2$
$X_1=1$	*85*	*21*
$X_1=2$	*13*	*36*

Dosazením těchto výběrových četností do uvedených vztahů získáme tyto výsledné hodnoty statistik:

$$\hat{\varepsilon} = 0{,}125 \qquad \hat{\tau} = 0{,}711 \qquad s(\hat{\varepsilon}) = 0{,}0095 \quad \mathrm{CI}(95\%): \; P[\, 0{,}107 < \varepsilon < 0{,}144\,] = 0.95$$

Získané statistiky je nutné doplnit testem symetrie vstupní kontingenční tabulky, abychom ověřili předpoklad shodné podmíněné distribuce obou hodnotitelů. V našem případě dvou replikací dichotomické klasifikace se jedná o McNemarův test. Na základě hodnoty testové charakteristiky: $\chi^2 = 1{,}88$; $p = 0{,}170$ lze hypotézu stejné distribuce přijmout.

Model B: Rozdílné pravděpodobnosti chyby v klasifikaci

Je samozřejmé, že existují dichotomické klasifikační procedury, kdy předpoklad shodné pravděpodobnosti obou chyb klasifikace je buď nereálný, nebo nastupuje z diagnostických důvodů přirozený požadavek rozdílných přípustných pravděpodobností obou chyb klasifikace. Jako příklad takové situace může být požadavek, aby se chyba falešné negativity diagnózy vyskytovala s nižší pravděpodobností, než případná chyba falešné pozitivity. V takovém případě se strategie analýzy z důvodu vazby mezi parametry náhodných proměnných, které model saturují, pochopitelně komplikuje. Z tohoto důvodu je nutné volit zcela odlišný induktivní postup. Strategie analýzy reliability dichotomické klasifikační procedury na základě dat z dvojnásobné klasifikace (např. nezávislé klasifikace dvou hodnotitelů) je v tomto případě založena na testu hypotézy, který je analogický jako klasický Pearsonův chí-kvadrát test dobré shody a má tedy následující kroky:

1. Určíme pevně hodnoty podmíněných pravděpodobností $\varepsilon_{ks} = P(X = s \mid T = k)$, $k,s=1,2$, tedy přípustné pravděpodobnosti chybné a potažmo i správné klasifikace. Tím určíme formou (nulové) hypotézy požadovaný stupeň spolehlivosti klasifikační procedury.
2. Stanovení těchto hypotetických pravděpodobností umožňuje na základě získaných empirických četností pomocí modifikované metody minimálního χ^2 odhadnout neznámé pravděpodobnosti pravých kategorií $\tau_k = P(T = k)$, *pro* $k = 1,2$.[1)] (podrobněji viz Anděl, 2002).
3. Dosazením těchto odhadů spolu s přípustnými pravděpodobnostmi chyby klasifikace do základního vztahu modelu získáme teoretické (očekávané) pravděpodobnosti

$$\hat{\pi}_{ij} = \sum_{k=1}^{2} \hat{\tau}_k \varepsilon_{ik} \varepsilon_{jk} \quad , \quad i,j = 1,2 \tag{2.3}$$

pro χ^2 test dobré shody empirických četností četnosti n_{ij} s modelem specifikovaným hypotézou o hodnotách přípustných pravděpodobností $\varepsilon_{ks} = P(X = s \mid T = k)$, $k,s=1,2$.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_j - n\hat{\pi})^2}{n\hat{\pi}_{ij}} \quad (df = 2) \tag{2.4}$$

Jinak řečeno, test rozhoduje, zda empirické rozdělení četností sdružené klasifikace odporuje, či neodporuje požadovanému stupni reliability klasifikace, který je specifikovaný přípustnými pravděpodobnostmi chybné klasifikace.

Uvedený algoritmus analýzy reliability si ukážeme podrobně na již uvedeném příkladu talentové zkoušky, přičemž přípustné pravděpodobnosti chybné klasifikace tentokrát volíme nesymetricky. Důvodem této volby může být požadavek, aby se chybné vyloučení talentovaného uchazeče vyskytlo s menší pravděpodobností, než chybný postup netalentovaného. Z tohoto důvodu byly přípustné pravděpodobnosti zvoleny následovně:

$P(X = 2 \mid T = 1) = 0,05$ a $P(X = 1 \mid T = 2) = 0,15$. Tyto hypotetické pravděpodobnosti poskytují odhad rozdělení pravděpodobností pravých kategorií metodou minimálního chí-kvadrátu: $\hat{\tau}_1 = 0,63$, $\hat{\tau}_2 = 0,37$. Teoretické resp. očekávané četnosti specifikované přípustnými pravděpodobnostmi chybné klasifikace uvádí tab. 2.

[1)] Odhady parametrů (pravděpodobnostní funkce pravých kategorií) τ_1 a potažmo $\tau_2 = 1 - \tau_1$ modifikovanou metodou minimálního χ^2 představuje řešení nelineární rovnice (v případě multinomické klasifikace potom soustavy nelineárních rovnic)

$\sum_{i=1}^{2} \sum_{j=1}^{2} \frac{n_{ij}}{\pi_{ij}(\tau_1, \hat{a})} \frac{\partial \pi_{ij}(\tau_1, \hat{a})}{\partial \tau_1} = 0$,

kde ε značí matici přípustných pravděpodobností ε_{ks}.

Tab. 2 Odhady teoretických (očekávaných) četností

KLAS	$X_2=1$	$X_2=2$
$X_1=1$	*88,9*	*12,0*
$X_1=2$	*12,0*	*42,1*

Tab. 3 Standardizované reziduály

KLAS	$X_2=1$	$X_2=2$
$X_1=1$	*-0,41*	*2,60*
$X_1=2$	*0,29*	*-0,94*

Shodu empirických dat s modelem zvolených přípustných pravděpodobností testuje χ^2 test dobré shody: χ^2 ($df = 2$) = 7,89*;* p = 0,019. Výsledek testu (zde zamítáme nulovou hypotézu) ukazuje neshodu empirických četností s hypotetickým modelem specifikovaným přípustnými pravděpodobnostmi chyby klasifikace. Podrobnější „post hoc" informaci k tomu poskytují připojené standardizované reziduály (viz tab. 3), ze kterých je patrné, že v kategorii ($X_1 = 1$, $X_2 = 2$) je signifikantně vyšší četnost výskytu této klasifikační kombinace, než by odpovídalo stanoveným přípustným pravděpodobnostem chybné klasifikace.

Pro ilustraci zde připojíme výsledek téhož testu, když zvětšíme přípustné pravděpodobnosti chybné klasifikace takto: $P(X = 2 \mid T = 1) = 0{,}1$ a $P(X = 1 \mid T = 2) = 0{,}15$. V tomto případě bychom na základě hodnoty testové statistiky χ^2 (df = 2) = 2,25; p = 0,325 hypotézu o přípustných hodnotách pravděpodobností nezamítli.

Jak již bylo řečeno v úvodu, ukážeme na datech příkladu talentové zkoušky také strategii analýzy reliability pomocí Cohenova κ–koeficientu. Pro výpočet použijeme příslušnou proceduru ze statistického programu SPSS. Ta hodnotu výběrového κ–koeficientu doplňuje dvěma inferenčními statistikami, testem hypotézy nulové hodnoty populačního κ–koeficientu a asymptotickou směrodatnou odchylkou odhadu koeficientu metodou maximální věrohodnosti, která umožňuje vyčíslení (asymptotického) intervalu spolehlivosti pro populační κ–koeficient. Tento (95%) interval je k výstupní tab. 4 dodatečně připojen, neboť není součástí výstupu programu SPSS.

Tab. 4 κ – koeficient (výstup SPSS)

		Value	Std. Error Asympt.[a]	Approx. T[b]	Approx. Sig.
Agreement	Phi	,517			,000
	Kappa	,514	,072	6,442	,000
N of Valid CASE		155			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypoth.

CI(95%): (0,514 ±1,96.0,072) : [0,37 ; 0,66]

Obecný model: Multinomická klasifikace a obecný počet replikací

Zabývejme se nyní nejobecnějším vícerozměrným modelem analýzy reliability multinomické klasifikační procedury. Ten představuje klasifikaci do $K \geq 2$ kategorií (označené 1, 2, … , K) na základě obecného počtu $R \geq 2$ nezávislých replikací. Vstupními daty pro analýzu jsou v tomto případě empirické četnosti K^R klasifikačních kombinací $n_{i_j \ldots i_R}$, $i_1, \ldots, i_R = 1, \ldots, K$ v R-rozměrné kontingenční tabulce. Tyto výběrové četnosti pocházejí z (populačního) rozdělení pravděpodobností $\pi_{i_1 \ldots i_R} = P(X_1 = i_1, \ldots, X_R = i_R)$, $i_1, \ldots, i_R = 1, \ldots, K$.

Pro snazší představu si již na tomto místě uvedeme datovou situaci, na které budeme později celý postup analýzy ilustrovat.

Příklad 2: V rámci volejbalové statistiky byla sledována kvalita příjmu podání. Tři školení posuzovatelé klasifikovali 468 příjmů (tři ligová utkání) do dvou kategorií (1 = „výborný“, tzn. možnost kombinace v následujícím útoku, 2 = „dobrý“, tedy bez možnosti kombinace). Z důvodu zjednodušení jsme zde vynechali další možné kategorie, tedy bez možnosti útoku a přímé body z podání, kdy možnost chybné klasifikace nepředpokládáme. Empirické četnosti klasifikace získaných od tří nezávislých posuzovatelů jsou uvedeny v trojrozměrné kontingenční tabulce (viz tab. 5.)

Tab. 5 Empirické četnosti klasifikace tří hodnotitelů

Klasifikace (X_1, X_2, X_3)	*111*	*112*	*121*	*122*	*211*	*212*	*221*	*222*	SUMA
Četnosti	262	19	23	12	16	9	11	116	468

Stejně jako v případě dvou posuzovatelů, i zde předpokládáme, že posuzovatelé klasifikují zcela nezávisle. Pro účely našeho modelu je samozřejmě nutné tento předpoklad formálně upřesnit. Budeme tudíž předpokládat, že podmíněné distribuce klasifikace $X_j \mid T$, *(j = 1, ... , R)* jsou u všech hodnotitelů stejně rozdělené a vzájemně nezávislé. Předpoklad shodné distribuce lze ověřit testem symetrie vícerozměrné kontingenční tabulky. Pro tyto účely byl zobecněn algoritmus Bowkerova testu symetrie dvojrozměrné kontingenční tabulky a jeho numerické řešení je rovněž součástí vytvořeného programu pro analýzu reliability (viz tab. 5).

V případě splnění těchto poměrně racionálních předpokladů lze vyjádřit pravděpodobnosti sdružené klasifikace pro obecný případ *R* posuzovatelů opět jako funkci „true“ a „error“ komponent:

$$\pi_{i_1 \ldots i_R} = \sum_{k=1}^{K} P(T = k) \prod_{j=1}^{R} P(X_j = i_j \mid T = k) \tag{3.1}$$

Vzhledem k tomu, že v tomto mnohorozměrném modelu je základní vztah z důvodu nutnosti použít komplikované indexy modelu na první pohled poněkud nepřehledný, považuji proto za užitečné prezentovat i jeho konkrétní úpravu pro náš příklad tří nezávislých dichotomických hodnocení. Potom pravděpodobnost, např. klasifikace objektu po řadě do kategorií (2, 1, 2), je vyjádřena následovně:

$$\pi_{212} = \sum_{k=1}^{2} P(T = k) P(X_1 = 2 \mid T = k) P(X_2 = 1 \mid T = k) P(X_3 = 2 \mid T = k) \tag{3.2}$$

Algoritmus řešení, tj. odhad parametrů modelu a potažmo test hypotézy, že empirické četnosti *R*-násobné multinomické klasifikace odpovídají pravděpodobnostnímu modelu, který je specifikovaný přípustnými pravděpodobnostmi chybné klasifikace, je zcela identický jako v předcházejícím příkladu, kdy jsme analyzovali pouze dvě nezávislé klasifikace. Přípustné hodnoty pravděpodobnosti chybné klasifikace byly v tomto případě stanoveny symetricky, tedy $P(X = 2 \mid T = 1) = P(X = 1 \mid T = 2) = 0{,}05$. Výsledky kompletní statistické analýzy dat příkladu 2 (textový výstup programu) prezentuje tab. 6.

Tabulka obsahuje (po řadě)

- Základní informaci o analyzovaných proměnných tj. počet případů, analyzované proměnné, počet kategorií a replikací (nezávislých klasifikací), distribuci kategorií v jednotlivých replikacích.
- Uživatelem stanovenou matici přípustných pravděpodobností $P(X \mid T)$ a následné odhady pravděpodobností „true kategorií“ $P(T)$.

Tab. 6 Výsledky analýzy reliability – příklad č. 2 (textový výstup programu)

```
NUMBER OF CASES :                    468
NUMBER OF REPETITIONS :               3
NUMBER OF CATEGORIES :                2

VARIABLES :    VAR 1  : R1
               VAR 2  : R2
               VAR 3  : R3
CASES WEIGHT          : FREQ

DISTRIBUTION CAT / REP:
CAT \ REP    1          2          3
  1        316        306        312
           67,5%      65,4%      66,7%

  2        152        162        156
           32,5%      34,6%      33,3%

TOTAL   100,0% 100,0% 100,0%

ACCEPTABLE PROBABILITIES  (T → X)
 (T → X)    1       2
 1        0,950   0,050
 2        0,050   0,950

ESTIMATED PROBABILITIES          P(j)
P(1) = 0,688
P(2) = 0,312

OBS      EXPECT    STAND     FREEMAN   INDEX
FREQ     FREQ      RES       TUKEY     1 2 3
262      276,10     -0,84      -0,84    1 1 1
 19       14,88      1,07       1,05    1 1 2
 23       14,88      2,11       1,92    1 2 1
 12        7,35      1,71       1,55    1 2 2
 16       14,88      0,29       0,34    2 1 1
  9        7,35      0,61       0,65    2 1 2
 11        7,35      1,34       1,27    2 2 1
116      125,22     -0,82      -0,82    2 2 2

GOODNESS OF FIT TEST / ACCEPT PROB
CHI-SQ      = 12,17
DF          =  6
SIGN        =  0,058

SYMETRY OF CONTINGENCY TABLE
CHI-SQ   =  1,71
DF       =  4
SIGN
```

– Pro jednotlivé (indexem označené) kombinační kategorie jsou uvedeny observované četnosti, očekávané četnosti (specifikované přípustnými pravděpodobnostmi $P(X|\ T)$ a pro post hoc analýzu standardizované a Freeman-Tukeyovy reziduály.
– χ^2 test dobré shody empirických dat s modelem specifikovaným hypotézou o přípustných pravděpodobnostech chybné klasifikace.
– Test symetrie vstupní kontingenční tabulky (testuje splnění nutného předpokladu modelu o shodné distribuci $P(X \mid T)$ u všech replikací).

Z inspekce výstupní tabulky zjistíme tyto stěžejní statistické nálezy:

– Na základě výsledku χ^2 testu symetrie kontingenční tabulky $\{\chi^2\ (df = 4) = 1{,}71\ , p = 0{,}948\}$ lze hypotézu symetrie přijmout, což značí splnění předpokladu o shodném rozdělení podmíněné klasifikace u všech tří hodnotitelů.
– Výsledek χ^2 testu dobré shody empirických četností trojnásobné klasifikace s modelem (specifikovaným zvolenými přípustnými pravděpodobnostmi) $\{\chi^2\ (df = 6) = 12{,}17, p = 0{,}058\}$ znamená přijetí hypotézy o přípustných hodnotách pravděpodobnosti chyby klasifikace $P(X = 2 \mid T = 1) = P(X = 1 \mid T = 2) = 0{,}05$.
– Jak je patrné, obsahuje výstupní tabulka, kromě těchto dvou stěžejních induktivních soudů, další užitečné statistiky, především procentuální vyjádření vstupních frekvencí, bodové odhady populační pravděpodobnosti „pravých" klasifikačních kategorií, očekávané četnosti sdružené klasifikace za podmínky přípustných pravděpodobností chyby klasifikace. Pro porovnání empirických a očekávaných četností sdružené klasifikace jsou k dispozici standardizované reziduály a Freeman-Tukeyovy odchylky.

Stejně jako u datové situace dvou nezávislých replikací klasifikace (příklad č. 1) i zde ukážeme paralelní analýzu reliability na základě Cohenova κ – koeficientu. Vzhledem k tomu, že pro obecný počet replikací (více než dva hodnotitelé) není k dispozici žádný standardní statistický software, využijeme zde příslušnou proceduru z našeho již zmíněného programu, s těmito výstupními statistikami (podrobné vzorce pro použité inferenční statistiky uvádí např.: Bortz, Lienert, Boehnke, 1990).

Výběrový odhad hodnoty κ – koeficientu:	$\kappa = 0{,}712$
Výběrová směrodatná odchylka (standardní chyba):	$s(\kappa) = 0{,}033$
95% konfidenční interval pro populační κ	[0,678 ; 0,776]

Pravděpodobnostní model chyby v klasifikaci vs. kappa model

Pokusíme se nyní o slíbené porovnání praktického využití obou strategií analýzy reliability nominálních klasifikací. Všeobecně známý Fleiss-Cohenův koeficient kappa je poměrně sofistikovaná míra konkordance v opakovaných klasifikacích. Její princip spočívá v relativním vyjádření nadnáhodných frekvencí shodných observací. Při totální konkordanci je hodnota koeficientu rovna jedné a v případě nezávislosti výsledných klasifikací, kdy empirická frekvence shodných klasifikací je rovna očekávané četnosti při platnosti hypotézy nezávislosti, je jeho hodnota nulová. Bohužel má tento koeficient charakter pouhého indikátoru bez empirického vyjádření. Tím máme na mysli fakt, že interpretace, resp. představa konkrétní hodnoty jeho bodového odhadu je pro uživatele jistě méně čitelná než např. obyčejná relativní četnost konkordantních klasifikací. To nutně vede k nešťastnému pseudoinformativnímu intervalovému třídění hodnot koeficientu s nic neříkajícím označením intervalů: výborný, uspokojivý, dobrý,.., atd., které lze v některých publikacích najít.

Prezentovaný pravděpodobnostní model chyby v klasifikaci poskytuje, na rozdíl od kappa - strategie, určitě lépe interpretovatelnou informaci. Inferenční soudy, intervalové odhady a testy hypotéz se zde týkají populačních pravděpodobností chybné, a tím samozřejmě potažmo i bezchybné klasifikace, tedy přímého vyjádření stupně rizika, s jakým může uživatel očekávat chybu klasifikace, např. falešnou pozitivitu či falešnou negativitu diagnózy.

Stručný nástin programu pro analýzu reliability

Hlavní procedura programu představuje kompletní numerické řešení obecného pravděpodobnostního modelu chyby v klasifikaci pro obecný počet klasifikačních kategorií (multinomická klasifikace) a obecný počet replikací. Vstupem jsou empirická data

z replikací klasifikace, a to buď ve formě *R*-rozměrné kontingenční tabulky (váhová proměnná), nebo ve formě primárních dat, tedy observovaného vektoru po statistických jednotkách. Uživatel si volí matici přípustných pravděpodobností chybné resp. správné klasifikace *P(X | T)*. Na základě těchto vstupních dat se vypočtou (metoda minimálního chí-kvadrátu) odhady pravděpodobnostní funkce *P(T)* pravých kategorií modelu. Tyto odhady představují řešení soustavy nelineárních rovnic (Newton-Raphsonova metoda). Získané odhady umožňují výpočet odhadů očekávaných četností pro vstupní *R*-rozměrnou kontingenční tabulku a potažmo výpočet hodnoty testové charakteristiky chí-kvadrát testu dobré shody s modelem. Tento chí-kvadrát test je pro „post-hoc" analýzu doplněn standardizovanými reziduály a Freeman-Tukeyovými odchylkami. Součástí tohoto algoritmu je i odhad pravděpodobnosti chybné klasifikace (metoda maximální věrohodnosti) pro případ dichotomické klasifikace s předpokladem, že obě chyby klasifikace mají shodnou pravděpodobnost. Pro ověření předpokladu shodné distribuce replikací klasifikace slouží zobecněný Bowkerův test symetrie *R*-rozměrné kontingeční tabulky.

Další procedura programu řeší problematiku reliability na základě Fleiss-Cohenovy koncepce při obecném počtu replikací. Výstupními statistikami jsou vážený i klasický kappa - koeficient, jeho asymptotický intervalový odhad, test hypotézy o nulové hodnotě populačního kappa a dílčí koeficienty kappa pro „post hoc" analýzu konkordance v jednotlivých klasifikačních kategoriích.

ZÁVĚR

Prezentovaný pravděpodobnostní model chyby v klasifikaci poskytuje novou, netradiční metodu analýzy spolehlivosti nominálních klasifikací. Ta poskytuje oproti klasické „kappa" metodologii inferenční soudy, které mají mnohem přirozenější, a tím samozřejmě i praktičtější interpretaci. Vzhledem k tomu, že u této metody je numerické řešení výstupních statistik poměrně složité, byl souběžně vytvořen program, jehož jádrem jsou iterativní metody výpočtu všech těchto statistik.

LITERATURA

Agresti, A. (2002): Categorical data analysis. New Jersey, John Wiley & Sons.

Anděl, J. (2002): Základy matematické statistiky. Praha, Matfyzpress.

Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D. (1999): Beyond kappa: A review of interrater agreement measures. Canadian Journal of Statistics, 27, 3-23.

Bortz, J., Lienert, G. A., Boehnke, K. (1990): Verteilungsfreie Methoden in der Biostatistik. Berlin, Springer.

Boushka, W. M., Marinez, Y. N., Prihoda, T. J., Dunford, R., Barnwell, G. M. (1990): A computer program for calculating kappa: application to interexaminer agreement in periodontal research. Computer Methods and Programs in Biomedecine, 33, 35-41.

Cohen, J. (1960): A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1968): Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220.

Conger, A. J. (1980): Integration and generalization of kappas for multiple raters. Psychological Bulletin, 88, 322-328.

Fleiss, J. L., Cohen, J., Everitt, B. S. (1969): Large sample standard errors of kappa and weighted kappa, Psychological Bulletin, 1969, 72, 323-327.

Fleiss, J. L. (1971): Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-392.

Fleiss, J. L. (1978): Inference about weighted Kappa in the non-null case. Applied Psychological Measurement, 1, 113-117.

Fleiss, J. L., Cohen, J. (1983): The equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33, 613-619.

Reed, J. F., Reed, J. J. (1992): Cohen's weighted kappa with Turbo Pascal (FORTRAN). Computer Methods and Programs in Biomedicine, 38, 153-165.

Walter, S. D., Irwig, L. M. (1988): Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. Journal of Clinical Epidemiology, 41, 923-937.
Wirtz, M., Caspar, F. (2002): Beurteilerübereinstimmung und Beurteilerreliabilität. München, Hogrefe.
Zvára, K., Štěpán, J. (2002): Pravděpodobnost a matematická statistika. Praha, Matfyzpress.

SOUHRN

Je známo, že pro analýzu reliability testů, které se užívají pro měření v psychologické diagnostice, je k dispozici velice silná a podrobně rozpracovaná metodologie. V diagnostice se však vyskytují i rozličné klasifikační procedury, kdy observovaná proměnná má pouze nominální charakter. U takových klasifikací se informace o jejich reliabilitě vyskytuje velice vzácně. Jedním z důvodů je samozřejmě fakt, že pro nominální klasifikace není k dispozici tak vypracovaná metodologie, jako u testového měření. Analýza spolehlivosti je v těchto případech většinou založena na Cohen-Fleissově kappa koncepci, tedy analýze konkordance při dvou, nebo více replikacích klasifikace. Cílem této stati je prezentace nové, netradiční metody analýzy reliability klasifikačních procedur, jejímž základem je pravděpodobnostní model chyby v klasifikaci. Tento model je založen na analogickém principu jako „true-error“ model klasické teorie testů.