

Jak neargumentovat při transformaci proměnné

ONDŘEJ ŠPAČEK

Fakulta humanitních studií, Univerzita Karlova, Praha

Diskuzní příspěvek k textu Paleček, A. 2017.
„Transformace kategorické proměnné na proměnnou
kontinuální na příkladu návštěvnosti bohoslužeb.“
Sociologický časopis / Czech Sociological Review 53 (4): 593–615.

V metodologicky zaměřeném článku „Transformace kategorické proměnné na proměnnou kontinuální na příkladu návštěvnosti bohoslužeb“ se Antonín Paleček věnuje možnostem transformace ordinální proměnné návštěvnosti bohoslužeb na kvantitativní škálu vyčíslením intenzity prostřednictvím počtu případů, kdy byl respondent na bohoslužbě za rok. Autor argumentuje výhody tohoto postupu a ukazuje je na příkladu dat z vybraných zemí, které se zúčastnily šetření ISSP 2008.

Mým záměrem v tomto diskuzním příspěvku není tvrdit, že navrhovaná transformace proměnné návštěvnosti bohoslužeb je metodologicky nevhodné řešení. Myslím, že je zcela obhajitelné, v zásadě až triviální, a kdyby se jednalo o empirický text z oblasti religiozity, mohla by se tato kvantifikace obejít bez zvláštní pozornosti, maximálně být okomentovaná pod čarou v poznámce technického typu. Autor ale přichází s primárně metodologickou ambicí a smyslem textu je prokázání obecných kvalit navrhovaného postupu oproti jiným alternativám. V tomto případě je nutné argumentaci kriticky rozebrat a položit si otázku, zda autor skutečně korektně vymezuje přínosy tohoto postupu.

Nejprve se zaměřuji na detailní rozbor některých argumentů, kterými chce autor doložit vhodnost a přínosnost navržené transformace. V druhé části pak za užití stejných dat, jaká byla v původním článku použita, empiricky srovnávám navržený kvantifikovaný postup s jeho alternativami. V poslední, třetí části se již pouze výčtově věnuji některým přešlapům či zavádějícím tvrzením, které mají obecnější metodologicko-statistickou povahu a v textu tohoto typu bychom se s nimi neměli setkat.

Vzhledem k deklarovanému metodologickému zaměření Palečkova článku (s. 595) nechávám zcela stranou nejen věcnou stránku řešeného problému, ale i další metodologické aspekty, které přesahují otázku statistické analýzy (např. nakolik navržené interpretace regresních modelů překračují jejich skutečné epistemologické možnosti apod.).

Proč kvantifikovat?

Napříč diskutovaným textem lze najít celou řadu argumentů, proč řešenou proměnnou kvantifikovat způsobem, který autor navrhuje. Ukazovány jsou zejména záporné postupy založené na kategorických datech a předpokládané přínosy využití kvantifikace. Problematické argumentační postupy lze přitom najít v obou rovinách.

Paleček zpochybňuje užití metod založených na práci s kategorickými daty (např. logistická regrese) na základě obtížnosti jejich aplikace či limitů spojených s nedostatečným zastoupením méně častých kategorií. První argument je pro metodologicky erudovanou diskuzi stěží přijatelný. V kontextu popularizace či srozumitelnosti pro širokou veřejnost by se jistě jednalo o validní kritérium, avšak pokud se máme bavit o metodologicky rigorózních přístupech, náročnost provedení analýzy může být až sekundárním kritériem.

Druhý argument je pak do značné míry nekonzistentní. Pokud jsou některé kategorie proměnné natolik málo zastoupené, že je nelze zařadit do analytického modelu, lze předpokládat, že v kvantitativním ohledu nebudou příliš podstatné, a neměli bychom tak mít obavu je v případě ordinální povahy proměnné sloučit se sousedními hodnotami. Autorův silně kritický postoj ke slučování kategorií ordinálních proměnných není dostatečně podložen a naznačuje ze strany autora spíše nepatřičnou objektivizaci do značné míry arbitrárních aspektů výzkumných nástrojů (jakým podoba kategorizace u ordinálních proměnných nepochybně je) než věcně podloženou kritiku.

Stejně tak lze argumentační oporu pro kvantifikaci rozporovat i v rovině domnělých pozitiv, která podle autora kvantifikace proměnné přináší. Leitmotivem tohoto omylu je domnělá přesnost transformované proměnné. Jádrem této zavádějící konstrukce je především to, že autor „zapomíná“ na přibližnost, která byla do kvantifikace nutně vložena při transformaci, a dále pracuje s kvantifikovanými údaji jako s exaktními hodnotami.

Vždy je určitý problém, pokud předpokládáme, že pouhou transformací proměnných data zásadním způsobem informačně obohatíme. Původní proměnné jsou ordinální. Je srozumitelné, že Paleček považuje za ztrátu to, že při práci s ordinálními daty nezohledňujeme kvantifikaci obsaženou v jednotlivých kategoriích, tedy např. to, že se jedná o frekvenci „1× za měsíc“ nebo „2–3× za měsíc“. Při kvantifikaci tento údaj zohledníme, ale při následném užití Paleček zapomíná na nepřesnost, která je v nově vytvořených kvantifikovaných proměnných obsažena, a pracuje s nimi, jako by se jednalo o přesné údaje. Srovná-li například kvantifikaci kategorie „asi jednou za měsíc“ a „několikrát za rok“, dochází k závěru, že první je dvakrát větší než druhá. Pokud bychom ale zohlednili nutnou nepřesnost kvantifikace, mohl by jejich vzájemný poměr být kdekoli v intervalu cca od 1 do 5,6.¹

¹ Hodnota „asi jednou za měsíc“ odpovídá kvantifikaci 12, avšak vzhledem k sousedním kategoriím ordinální proměnné se může jednat o jakékoliv číslo v intervalu (9; 21). Podobně

Z podobného důvodu je také problematické, když autor zdůrazňuje, že díky možnosti vypočítat průměr „je také přínosné, že na rozdíl od původní proměnné o devíti kategoriích (...) jsou stavy i změny návštěvnosti vyjádřeny pouze jednou hodnotou“ (s. 601). V první řadě při konstrukci tohoto „přesného“ průměru nedokáže zohlednit nepřesnost obsaženou v samotných hodnotách.² Zároveň pak autor zcela opomíjí zásadní reduktivní rozměr průměrování, který je obzvláště u proměnných, se kterými zde pracuje, zásadní a může vést ke špatné představě o podobě dat. Opakovaná deklarace přínosnosti toho, že pracujeme „pouze s jednou hodnotou“, svým způsobem obrací to, jakým způsobem by měly být vedeny metodologické úvahy. Výzkumník by měl zacházet s takovým počtem hodnot (parametrů), jaký je vyžadován empirickou realitou dat. A priori tento počet snižovat bez prokázání toho, že nedošlo k zásadní újmě na popisu sledovaného jevu, představuje podstatné riziko zkreslující redukce. Důsledky tohoto rizika jsou konkrétně demonstrovány v následující části.

Je kvantifikovaná proměnná lepší než původní?

Výše uvedenou argumentaci lze konkrétně demonstrovat na rozporování empirických příkladů, které Paleček v textu přináší. Zvláštní pozornost zde věnuji otázce aritmetického průměru kvantifikovaných hodnot, do kterých autor vkládá značné „naděje“ při získávání přehledu o datech.

Autor například referuje o tom, že díky kvantifikaci může ukázat, že průměrná roční návštěvnost bohoslužeb je v ČR 5,5, v SR 23,7 a v Polsku 29,8. Co to ale znamená? Podíváme-li se na rozložení celé proměnné, zjistíme, že hodnota 5,5 (odpovídající přibližně kategorii „několikrát za rok“) je velmi netypická. Údaj o průměru tak naprosto zakrývá fakt, že více než 60 % respondentů nenavštěvuje bohoslužby vůbec, dalších téměř 20 % méně než dvakrát za rok a do oné „průměrné“ kategorie spadá pouze 6,5 % respondentů. Tedy stěží se jedná o průměr jakožto vyjádření typičnosti.

Podobný problém je zřetelný i v polském a slovenském případě, kde průměrná hodnota odpovídající frekvenci „2–3× za měsíc“ představuje vlastně zcela netypické „lokální minimum“, mezi významně častějšími typy jednání odpovídajícími buďto navštěvování bohoslužeb každý týden, nebo naopak několikrát za rok (což je mimochodem zřetelně patrné i z grafu 1 na s. 604).

„několikrát za rok“ je kvantifikováno hodnotou 6, ale kategorie zahrnuje celý interval (3,75; 9). Výsledný poměr tak může být téměř sousedící (kolem hraniční hodnoty 9) nebo v poměru $21/3,75 = 5,6$.

² Bylo by možné uvažovat o tom, že skutečné hodnoty návštěvnosti bohoslužeb jsou v rámci jednotlivých měřených kategorií rozptýleny rovnoměrně a výpočet aritmetického průměru tak neovlivní. To ale není vůbec diskutováno (a nejspíše by ani nebylo hodnověrné vzhledem k povaze rozdělení dané proměnné).

Tabulka 1. Průměrné hodnoty a rozložení proměnné návštěvnosti bohoslužeb (ISSP 2008)

	Francie	Japonsko	Česká republika
Průměr (kvantifikováno)	5,2	5,4	5,5
Ordinální škála (sloučeno do 4 kategorií)			
Nikdy	57,9	21,0	61,5
Nejvýše dvakrát za rok	22,1	28,8	19,8
Několikrát za rok až jednou měsíčně	11,6	46,4	8,8
Dvakrát měsíčně a častěji	8,4	3,8	9,9
N	1 188	2 367	1 484

Nebezpečnost tvrzení o přínosech kvantifikace v podobě snadného výpočtu průměrných hodnot za celé populace lze velmi snadno demonstrovat i na širším srovnání ve stejném datovém souboru ISSP 2008. Analyzovanou situaci české populace můžeme doplnit o srovnání dvou dalších zemí, Francie a Japonska, které vykazují téměř shodnou průměrnou hodnotu návštěvnosti bohoslužeb vypočtenou prostřednictvím navrhované kvantifikace. Průměrná hodnota této proměnné se statisticky významně neliší od hodnoty v České republice, optikou Palečkových argumentů se tedy jedná o potvrzení shodné situace. Z rozepsaných hodnot jednotlivých kategorií původní ordinální proměnné (a to i v situaci, kdy jsem pro snadnější čtení původní dělení sloučil do čtyř kategorií) je však zřetelné, že srovnání průměrů nám poskytuje značně redukovanou informaci (viz Tabulka 1).

Zcela zavádějící je domnělá shoda mezi japonskou a českou společností. Zatímco v ČR více než 60 % respondentů bohoslužby vůbec nenavštěvuje, v Japonsku se jedná pouze o něco více než 20 %. Naopak japonská norma se odráží v návštěvě bohoslužeb několikrát za rok (původní kategorii „několikrát za rok“ zvolilo 42 % respondentů). Je tedy více než zřejmé, že pod zdánlivě stejným průměrem se skrývá zcela odlišná struktura jevu.

I v případě Francie ale lze popsat určité jemnější nuance, které se pod redukcí průměrné hodnoty ztrácí. Kupříkladu zatímco z průměrného údaje se zdá, že ve Francii je návštěva bohoslužeb méně obvyklá (byť statisticky tento rozdíl není potvrzen), při srovnání původní ordinální struktury se situace vyjevuje odlišně. V České republice najdeme významně více lidí nenavštěvujících bohoslužby vůbec. Naopak ve Francii jsou obvyklejší střední kategorie škály. Za domnělou vyšší intenzitou návštěvy bohoslužeb v ČR pak stojí pouze o 1,5 procentního bodu častější zastoupení kategorie s nejvyšší intenzitou. Tyto aspekty zcela zaniknou při spojení průměrů na nutně reduktivní průměrný údaj.

Srovnání průměrů tak v daném případě nelze popsat lépe než přesné srov-

nání nepřesných hodnot, respektive hodnot, které pouze málo vypovídají o relevantní komplexitě řešených dat. Tím rozhodně nemá být řečeno, že práce s průměry a jejich srovnávání je obecně špatně. Naopak, oprávněně se jedná o jednu z nejužitečnějších statistických reduktivních operací. Autor však v článku přichází se zcela neudržitelnými tvrzeními o superioritě kvantifikujícího postupu oproti jeho alternativám, které mohou svést jeho případné následovníky ke zcela nepodložené jistotě o podobě dat.

Druhý moment, kde lze zřetelně demonstrovat nedostatečnost argumentace v Palečkově textu, se týká přidané hodnoty kvantifikace na příkladu sledování intergenerační transmise prostřednictvím regresní analýzy. Ponechme v daný moment stranou teoretické argumenty o vhodnosti či nevhodnosti jednotlivých postupů a zaměříme se na to, zda v navrhované transformaci přináší regresní model lepší schopnost popsat situaci v datech.³ Pro zřetelnost a jednoduchost se v následující analýze nesnažím srovnat možnosti různých analytických postupů (např. variant regresních modelů pracujících s kategoriemi proměnnými), nýbrž různé predikční schopnosti lineárního regresního modelu, tak jak ho prezentuje Paleček, s alternativní kvantifikací.

Alternativním kvantifikujícím postupem může paradoxně být využití arbitrárně zvolených hodnot kódování původní ordinální proměnné v datovém souboru (např. 1 = „nikdy“, 2 = „méně než jednou ročně“, 3 = „asi 1× nebo 2× za rok“ atd.). Pouze zdánlivě se přitom jedná o svévolný postup. Lze totiž předpokládat, že u opakovaně používaných a prověřených otázek v dotazníku se použítá škála kategorií blíží optimální empiricky založené struktuře. Jinak řečeno, pokud jsou použité kategorie ordinální proměnné dobře metodologicky zvoleny, tak odrážejí jak kognitivní možnosti respondentů na danou otázku odpovědět, tak sociálně relevantní rozdíly v intenzitě daného jevu. Tato struktura přitom nemusí odpovídat lineární metrice, obzvláště v případě intenzitních proměnných, jako je právě frekvence návštěv bohoslužeb.⁴

Přímé srovnání přitom ukazuje, že pokud v regresních modelech představených na str. 608 použijeme proměnné kvantifikované prostřednictvím pořadového čísla kategorie (tj. 1, 2, 3... 9), tak ve všech případech jejich prediktivní schopnosti významně vzrostou (viz Tabulka 2). Jinak řečeno, za jinak stejných podmínek⁵ dosahuje lepších úspěchů v predikci regresní model, který není za-

³ Samozřejmě si jsem vědom, že vysvětlující schopnost statistických modelů sama o sobě nutně neprokazuje jejich sociologickou hodnotu. Pokud by kupříkladu perfektní statistický model pracoval se sociologicky nevalidním měřením proměnných, byl by v zásadě bezcenný. Zde však přistupuji na metodologický charakter a argumenty diskutovaného článku.

⁴ U míry intenzity mívají užívané škály spíše logaritmický charakter. Tento princip lze identifikovat i v proměnné návštěvnosti bohoslužeb, která je zde předmětem analýzy (obzvláště, pokud bychom použili původní verzi o osmi kategoriích).

⁵ Stejná metoda analýzy, stejný regresní model, stejný počet „kategorií“ (tj. různých hodnot) u nezávislé a závislé proměnných. Odlišnost spočívá pouze ve způsobu kvantifikace původních ordinálních kategorií.

Tabulka 2. Pearsonovo R lineárně regresních modelů podle použité metody kvantifikace

Metoda kvantifikace	Česká republika	Slovensko	Polsko
Paleček 2017	0,823	0,769	0,596
pořadové číslo kategorie	0,883	0,849	0,678

ložen na lineární kvantifikaci intenzity návštěv bohoslužeb, kterou navrhuje Paleček ve svém textu. K podobným závěrům bychom mohli dospět i při použití pokročilejších metod statistické analýzy, za diskuzi by zde mohla stát zejména kombinace regrese a optimálního škálování (metoda CATREG v SPSS).

Další omyly a zavádějící tvrzení

Kromě výše uvedených problémů spojených s cílem samotné studie je nutné rozporovat i řadu drobnějších tvrzení, která by vzhledem ke statisticko-metodologické povaze studie mohla svést čtenáře k chybným závěrům či představám o statistických metodách:

– Linearita měřeného vztahu v lineární regresní analýze je předpokladem, tedy něčím, co je potřeba prokazovat jinými prostředky než samotnou regresní analýzou (s. 602). Lineární regresní analýza nemůže sama o sobě z podstaty prokázat, zda měřený vztah je, či není lineární. Snad kromě zcela hypotetické situace, kdy by byla závisle proměnná v modelu (téměř) stoprocentně determinována nezávisle proměnnými.

– Normalita rozdělení transformovaných proměnných není v článku relevantně prokazována. Pozornost je věnována pouze symetrické centralizaci kolem průměru (s. 601), což je jeden ze znaků normálního rozdělení, nikoli však ten hlavní. Adekvátní analýza tvaru distribuční křivky v případě polských dat by záhy ukázala, že ani v tomto případě se rozdělení zdaleka neblíží Gaussově křivce (což je ostatně patrné i z grafu, ve kterém má křivka dva vrcholy).

– I samotný graf 1 (s. 604) zobrazuje data zcela chybným způsobem. Osa x neodpovídá deklarované kvantifikující transformaci. Je nanejvýš paradoxní, že autor zde zobrazuje proměnnou v duchu původní kategoriální proměnné. Jinak řečeno, zatímco představená kvantifikace má za cíl vyčíslit vzdálenosti mezi jednotlivými kategoriemi jako různé, v grafu jsou zobrazeny ve shodných jednotkových intervalech, ať je na dané škále jejich rozestup jedna polovina, jedna či 52. Na základě takto zkresleného grafu si čtenář může utvářet velmi zavádějící představu o reálné podobě rozložení nové proměnné.

– Nepřímé tvrzení, že v dřívějších dobách byly datové soubory malé a nyní jsou větší (s. 602), není pravdivé. Velikost datových souborů není ani tak parametrem

doby, jako spíše účelu a potřeb daného výzkumu. I v dávné minulosti kvantitativního výzkumu často najdeme na dnešní poměry velmi rozsáhlá výběrová šetření.

– Výklad nerozlišuje mezi procentními body a procenty, což je u metodologické statě alarmující. Pokud v ČR navštěvují bohoslužby jedenkrát měsíčně 2 % respondentů a na Slovensku 5 % respondentů, nelze říci, že v ČR je těchto respondentů o 3 % méně (s. 607). Správná formulace je o 60 % méně (nebo o 3 procentní body). V každém případě je ale i otázka, nakolik je korektní vzhledem k velikosti vzorku vůbec hovořit o přesné velikosti rozdílu v takto málo zastoupených kategoriích.

– P-hodnota (sig.) nikdy není rovna nule, jak je v textu opakovaně uváděno. Vždy se jedná o velmi malé, ale nenulové číslo.

Závěrem

Kvantifikovat proměnnou prostřednictvím navrženého postupu je samozřejmě možný a v řadě případů legitimní postup kvantitativní analýzy dat, který má své přínosy a relevantní zdůvodnění. Autor však ve svém textu předkládá tvrzení, která jsou zcela neadekvátní, v mnoha ohledech i zavádějící a z hlediska kvality metodologických postupů nebezpečná. Aritmetický průměr je pravděpodobně nejpoužívanější a nejefektivnější statistickou operací, považovat ho ale za řešení problému dat s komplexnějším rozdělením je nesmyslné. Stejně tak již sociální vědy řadu let pracují s kategorickými metodami regresní analýzy právě proto, že lineární regresní analýza v mnoha případech může vést ke zkresleným a interpretacně obtížně obhajitelným výsledkům. Autor zcela podceňuje specifikum škály intenzity – například to, že u dané proměnné je nula přirozené minimum, které v žádném případě nemůže být překročeno. Aplikovat na proměnnou této povahy lineární regresní analýzu je přešlapem nejen pro metodologické puristy, ale významným rizikem i pro pragmatického analytika.