INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 30, číslo 2, červen 2019

VÝBĚROVÁ ŠETŘENÍ A STÉ VÝROČÍ ČESKÉHO STATISTICKÉHO ÚŘADU

SURVEY SAMPLING AND THE CENTENARY OF THE CZECH STATISTICAL OFFICE

Zuzana Prášková

Adresa: KPMS UK, Sokolovská 83, 186 75 Praha 8

E-mail: praskova@karlin.mff.cuni.cz

Abstrakt: Článek se zabývá historií výběrových šetření přibližně za posledních 100 let, tedy za období, kdy se formovala státní statistická služba v našich zemích. Všímá si rovněž rozvoje výuky matematické statistiky na českých vysokých školách a přínosu Jaroslava Hájka k obecné teorii výběrových šetření.

Klíčová slova: reprezentativnost, pravděpodobnostní výběr, výběry s nestejnými pravděpodobnostmi, statistická inference.

Abstract: The paper maps the history of survey sampling over the last hundred years covering the beginnings of the state statistical service in the Czech lands. The history of teaching mathematical statistics at Czech universities is also shortly mentioned and Jaroslav Hájek's contribution to general theory of survey sampling is discussed.

Keywords: representativness, probability sampling, sampling with varying probabilities, statistical inference.

1. Historická ohlédnutí a souvislosti

Záznamy o používání výběru pro získání informace o větším souboru vedou hluboko do historie. My začneme obdobím, které lze považovat za počátek moderní teorie výběrových šetření. Půjde o konec 19. a začátek 20. století, tedy o období, kdy vznikala i organizovaná statistika v českých zemích.

Připomeňme nejprve, že za počátek státem organizované statistiky v českých zemích je považován rok 1856, kdy byl ustaven Ústřední výbor pro statistiku polního a lesního hospodářství Čech při c. k. Vlasteneckohospodářské společnosti. Prvním předsedou byl zvolen hrabě Albert Nostitz. Výkonným orgánem tohoto Ústředního výboru byla Statistická kancelář, kterou vedl Eberhard Antonín Jonák, mj. autor jedněch z prvních učebnic statistiky[1] a národního hospodářství v českých zemích[2].

[1] Theorie der Statistik in Grundzügen, Wien 1856.

[2] Základové hospodářství, Praha 1871.

V roce 1870 byla ustavena Statistická komise královského hlavního města Prahy. Praha se tak zařadila mezi města jako Brémy, Berlín, Vídeň, Řím nebo New York. Výkonným orgánem komise byla statistická kancelář, jejímž ředitelem se stal docent Josef Erben z pražské polytechniky. Doc. Erben byl roku 1887 jako první Čech zvolen členem Mezinárodního statistického institutu (ISI), který byl založen v roce 1885 v Londýně 81 nejvýznamnějšími představiteli oficiální a akademické statistiky té doby. A konečně roku 1897 byl založen Zemský statistický úřad Království českého, první skutečně statistický úřad na území dnešní České republiky, jeho přednostou se stal Karel Kořistka.

V roce 1919 byl ustaven Státní úřad statistický (orgán výkonný) a Statistická rada státní (orgán poradní) jako orgány státní statistické služby nově vzniklého československého státu. Předseda statistického úřadu byl zároveň i předsedou statistické rady. Prvním předsedou se stal Dobroslav Krejčí, kterého brzy vystřídal František Weyr. O historii a činnosti této státní statistické instituce se lze více dočíst např. v článku [15] a na webu ČSÚ.

Než se začneme zabývat výběrovými šetřeními, uveďme rovněž několik poznámek k historii výuky matematické statistiky na vysokých školách a obecně k vývoji statistiky u nás. Před rokem 1890 se na univerzitě i na technice konaly různé přednášky z teorie pravděpodobnosti, hospodářské či průmyslové statistiky, většinou výběrovou formou. V roce 1904 byl zaveden dvouletý Učebný běh pojistné matematiky na české technice (od roku 1920 na Vysoké škole speciálních nauk ČVUT), kde se vedoucími osobnostmi stali Josef Beneš (pojistná matematika) a později Jaroslav Janko (pojistná matematika, matematická statistika). V roce 1921 byl schválen dvouletý cyklus přednášek o pojistné matematice a matematické statistice na přírodovědecké fakultě (ta vznikla roku 1920) University Karlovy, od roku 1923 zde působil Emil Schoenbaum (pojistná matematika), později Ladislav Truksa (matematická statistika, statistická dynamika) a Josef Bílý (hospodářská a finanční matematika, politická aritmetika).

V roce 1929 byla založena Československá statistická společnost[3], která vznikla jako vědecké fórum sdružující nejvýznamnější odborníky ze státní statistiky i z vysokých škol, předsedou byl zvolen Vilibald Mildschuh, profesor statistiky a národního hospodářství na Právnické fakultě UK, jednatelem Jaroslav Janko. Členové společnosti publikovali odborné články, účastnili se odborných statistických konferencí i mezinárodních kongresů organizovaných ISI. Vedle již existujících časopisů Pojistný obzor, Československý statistický

[3] Její činnost byla přerušena v roce 1939. Zanikla v roce 1948. Česká statistická společnost byla založena v roce 1990.

věstník (od roku 1931 přejmenovaný na Statistický obzor) a Časopis pro pěstování matematiky a fyziky, byl v roce 1930 založen časopis Aktuárské vědy, který měl sloužit k publikaci prací *ze všech oborů teoretického i praktického národohospodářství a sociálních věd, v nichž aplikují se metody matematické* a rovněž měl za cíl seznamovat s domácími výsledky i odbornou veřejnost zahraniční. Lze tak konstatovat, že činnost odborné statistické komunity se v novém státě všestranně rozvíjela. Tento vývoj byl násilně přerušen nacistickou okupací a uzavřením vysokých škol v roce 1939, zákazem vydávání odborných časopisů, činnost státního statistického úřadu byla omezena a podřízena válečnému hospodářství.

V roce 1945 byla opět zahájena výuka pojistné matematiky a matematické statistiky na přírodovědecké fakultě UK a na Vysoké škole speciálních nauk při ČVUT. Rostoucí význam aplikací matematických a statistických metod pro poválečnou obnovu a potřeba výchovy nových odborníků vedly k zákonnému rozšíření studia statistických oborů i k plnému obnovení a rozšíření státní statistické služby. Zákonem č. 122/1946 Sb., o úpravě statisticko-pojistného studia, bylo rozhodnuto o zavedení čtyřletého studia statisticko-pojistného inženýrství na Vysoké škole speciálních nauk a čtyřletého studia matematické statistiky, pojistné matematiky a ekonometriky na Univerzitě Karlově.

Společenské změny v roce 1948 se dotkly i státní a akademické statistiky. Studium statisticko-pojistného inženýrství na ČVUT bylo rozděleno na větev matematicko-statistickou (prof. Janko) a větev ekonomicko-statistickou (prof. František Egermayer). V roce 1952 byla Vysoká škola speciálních nauk zrušena, studium převedeno na nově vzniklou Vysokou školu ekonomickou (VŠE) a Matematicko-fyzikální fakultu Univerzity Karlovy (MFF UK). Katedra matematické statistiky[4] MFF UK byla založena v roce 1952, ve stejném roce jako fakulta. Externím vedoucím katedry se stal prof. Josef Novák (v letech 1953–1956). V letech 1956–1964 byl vedoucím katedry prof. Jaroslav Janko.

Prof. PhDr. Jaroslav Janko, DrSc. (1893–1965) byl mimořádnou osobností české statistiky prvních šedesáti let minulého století. Byl členem Státní rady statistické, trvale spolupracoval se Státním úřadem statistickým, zejména v odboru demografickém. Z této spolupráce vznikla řada demografických studií.[5] Publikoval v odborných časopisech domácích i zahraničních, účastnil se aktivně práce domácích i mezinárodních odborných institucí, psal učební

[4]Nyní katedra pravděpodobnosti a matematické statistiky.

[5]Např. Janko, J. (1931): Konstrukce úmrtnostních tabulek obyvatelstva na podkladě sčítání lidu. *Statistický obzor: revue pro statistickou teorii a praxi.* Praha: Státní úřad statistický **12**, No. 3–4, 129–140.

texty, podílel se na organizaci pedagogického a vědeckého života jako akademický funkcionář. Nákladem Státního úřadu statistického vyšla v roce 1937 Jankova kniha [9], považovaná za základní monografii o moderní matematické statistice, založené na teorii pravděpodobnosti a náhodném výběru. Později, díky iniciativě JČMF, vyšly jeho publikace [10], podávající přehled základních pojmů a moderních metod matematické statistiky formou přístupnou čtenáři se středoškolským vzděláním. Janko se zde zabývá i základními problémy výběrových šetření. Záslužným činem prof. Janko je zpracování a souborné vydání statistických tabulek a tabulek náhodných čísel [11], které se ve své době staly užitečnou pomůckou jak při výchově studentů matematické statistiky, tak odborníků aplikujících statistické metody v praxi.

2. Základy moderní teorie výběrových šetření

Na konci 19. století díky průmyslovému a hospodářskému rozvoji došlo i k rozvoji národních statistických institucí, které vyžadovaly úplné statistické šetření. Zároveň se však objevovaly i sociologické studie, ve kterých byly intenzivně vyšetřovány jen některé skupiny obyvatel, rodiny apod.

Anders Nikolai Kiaer, ředitel Norského Centrálního statistického úřadu (Norvegian Central Bureau of Statistics), již po roce 1876 prováděl neúplná statistická šetření, na základě kterých došel k závěru, že jsou postačující k získání užitečné a uspokojivé informace o celé populaci a není nutné provádět její úplné šetření. Se svými poznatky vystoupil v roce 1895 na zasedání ISI v Bernu, kde poprvé formuloval pojem reprezentativnosti výběru. Své výsledky publikoval v Bulletinu ISI o rok později.[6] Jeho práce se setkala s převážně nepříznivou reakcí zejména ze strany vedoucích představitelů národních statistických institucí, kteří tvořili jádro ISI a považovali census za jediný možný přístup ke statistickému šetření. Kiaer svou metodu representativního výběru, který vysvětloval jako miniaturní vzorek celé populace, upřesňoval a obhajoval i na dalších zasedáních ISI v roce 1897, v roce 1901 a v roce 1903.

Diskuse vedené na půdě ISI vedly v roce 1924 k ustavení komise pro studium reprezentativní metody. Pro nás může být zajímavé, že jejími členy byli i Lucien March z Francie a Franz Žižek z Německa, oba čestní členové Československé statistické společnosti. Závěry komise v roce 1925 vyústily do formální rezoluce, která akceptovala ideu reprezentativního výběru, a to buď jako náhodného výběru, do kterého může být zahrnuta každá jednotka se stejnou pravděpodobností, nebo jako záměrného výběru, neboli výběru skupin velkého rozsahu na základě úsudku a poté úplné šetření vybraných

[6] Observations et expériences concernant des dénombrements représentatifs, *Bull. Int. Statist. Inst.* **9** (1886), 176–183.

skupin. Volba metody byla ale zcela ponechána na rozhodnutí organizátora šetření.

V roce 1926 Arhur L. Bowley, profesor na London School of Economics, publikoval rozsáhlý článek[7], ve kterém shrnul současné poznatky o obou metodách výběru a navrhl jisté možnosti jejich dalšího vývoje. Snažil se zejména o přesnější formulaci a matematickou rigoróznost záměrného výběru. Reprezentativní metodou se zabýval i prof. Janko.[8]

V dalších letech mezi zastánci a oponenty jednotlivých variant reprezentativního výběru postupně začali převažovat stoupenci náhodného (pravděpodobnostního) výběru. Velkou měrou se o to zasloužil polský statistik Jerzy Neyman [14], který poukázal na to, že výhod záměrného výběru lze dosáhnout i na základě stratifikace a pravděpodobnostních principů, které poskytují nestranné odhady populačních charakteristik. Zdůrazňoval význam randomizace, díky které lze použít centrálních limitních vět a intervalů spolehlivosti pro statistické usuzování.

Během dalších 20 let došlo k velkému rozvoji teorie pravděpodobnostního výběru: byly definovány nové výběrové plány, nové varianty odhadu populačních charakteristik a variančních odhadů, navrženy vícestupňové výběry, výběry s nestejnými pravděpodobnostmi zahrnutí atd. Toto období je spojeno se jmény jako M. H. Hansen, W. N. Hurwitz, W. G. Cochran, H. O. Hartley, J. N. K. Rao, a také J. Hájek.

3. Jaroslav Hájek

Na tomto místě připomeňme práce prof. Jaroslava Hájka, který podstatně přispěl k teorii výběrových šetření v období od 50. do 70. let minulého století. Prof. Dr. Ing. Jaroslav Hájek, DrSc. (1926–1974) vystudoval statisticko-pojistné inženýrství na Vysoké škole speciálních nauk ČVUT (1945–1949). Poté nastoupil jako aspirant do Ústředního ústavu matematického (později Matematický ústav ČSAV), kde po obhájení disertační práce pracoval do roku 1966. V roce 1966 byl jmenován profesorem a vedoucím katedry matematické statistiky na MFF UK, externě vedl katedru od roku 1964.

Již v roce 1949 Hájek v článku [2] navrhl dvoufázovou proceduru pro nestranný odhad průměru základní populace rozdělené do nestejně velkých skupin, která ve srovnání s Neymanovým postupem z roku 1934 dává vý-

[7] Measurement of the precision obtained in sampling, *Bull. Int. Statist. Inst.* **22** (1926), Suppl. to Book 1, 1–62.

[8] K teorii representativní metody, *Časopis pro pěstování matematiky a fysiky* **57** (1928), No. 3–4, 286–290.

sledek s menším rozptylem. Stejný výsledek později publikovali O. B. Lahiri[9] a H. Midzuno[10].

V letech 1950–1960 J. Hájek publikoval nové výsledky týkající se poměrových odhadů, odhadů v oblastním výběru, zabýval se permutačním výběrem, optimálním a náhodným rozvržením výběru do oblastí. Zásadní význam má odvození limitních vět pro prostý náhodný výběr [5]. V roce 1960 rovněž vydal knihu [4], jejíž první část je v podstatě základní učebnicí, druhá část monografií, shrnující některé jeho původní výsledky. Vedle teoretických prací publikoval i výsledky některých výběrových šetření z lékařského výzkumu, na kterých aktivně spolupracoval.[11]

Zřejmě nejcitovanější Hájkovy práce se týkají poissonského a zamítacího výběru, viz [3], [6]. Nechť $\mathcal{S}$ je populace N jednotek, $s \subset \mathcal{S}$ výběr a $P = \{P(s), s \subset \mathcal{S}\}$ pravděpodobnostní rozdělení na množině všech možných výběrů z $\mathcal{S}$. Znak Y na populačních jednotkách nabývá hodnot $y_1, \ldots, y_N$. Pravděpodobnost zahrnutí jednotky i do výběru je $\pi_i = \sum_{s \ni i} P(s)$ a pravděpodobnost současného zahrnutí jednotek i a j je $\pi_{ij} = \sum_{s \ni i,j} P(s)$. Rozsah výběru je $K = \sum_{i=1}^{N} I_i$, kde I_i je indikátor zahrnutí jednotky i do výběru s, tj. náhodná veličina, která nabývá hodnoty 1, jestliže s obsahuje i a je 0 jinak. Prostý lineární odhad populačního úhrnu $y = \sum_{i=1}^{N} y_i$ (*Horvitz-Thompsonův odhad*) je

$$\widehat{y}_{\mathrm{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i}. \tag{1}$$

Je-li rozsah výběru pevný ($K = n$), používá se pro výpočet rozptylu $\widehat{y}_{\mathrm{HT}}$ Senova-Yatesova-Grundyho formule

$$V(\widehat{y}_{\mathrm{HT}}) = \mathrm{Var}(\widehat{y}_{\mathrm{HT}}) = \sum_{1 \le i < j \le N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \tag{2}$$

Potom nestranný odhad $V(\widehat{y}_{\mathrm{HT}})$ je

$$\widehat{V}(\widehat{y}_{\mathrm{HT}}) = \frac{1}{2} \sum_{i \in s} \sum_{j \in s} \left(\frac{\pi_i \pi_j}{\pi_{ij}} - 1 \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \tag{3}$$

[9] A method for selection providing unbiased estimates, *Int. Stat. Ass. Bull.* **33** (1951), 133–140.

[10] On the sampling system with probabilities proportionate to sum of sizes, *Ann. Inst. Stat. Math.* **3** (1951/52), 99–108.

[11] První výběrová šetření byla prováděna v letech 1956/57 a týkala se příjmů domácností, rodinných účtů, osevních ploch a dalších aplikací v zemědělské a hospodářské statistice (V. Čermák, J. Kozák, J. Walter a další.)

Poissonský výběr s parametry $0 < p_i < 1,\ i = 1, \ldots, N$, je definován pravděpodobnostmi

$$P(s) = \prod_{i \in s} p_i \prod_{i \in S\ -s} (1 - p_i). \tag{4}$$

Indikátory zahrnutí jsou nezávislé náhodné veličiny, $P(I_i = 1) = p_i = 1 - P(I_i = 0)$, a pro pravděpodobnosti zahrnutí platí $\pi_i = p_i$ pro $i = 1, \ldots, N$. Rozsah výběru je náhodná veličina, pro kterou $E(K) = \sum_{i=1}^{N} p_i$.

Zamítací výběr lze definovat jako podmíněný poissonský výběr, za podmínky, že rozsah výběru je pevný. Výhodou této definice je, že teoretické výsledky pro poissonský výběr, např. centrální limitní věty a věty o asymptotické normalitě odhadů, mohou být snadno využity pro zamítací výběr. Definice výběrového plánu jako podmíněný poissonský výběr však přináší problém jak spočítat pravděpodobnosti zahrnutí, které lze vyjádřit pomocí volitelných parametrů $p_1, \ldots, p_N$ pouze asymptoticky. Pro praktické účely se hodí alternativní definice zamítacího výběru jako podmíněný výběr s vracením, kdy se vybírají jednotky nezávisle s vracením, jakmile se však některá jednotka ve vzorku objeví znovu, celý vzorek se zamítne a procedura se opakuje. Pokračuje se tak dlouho, dokud ve vzorku nebudou jen vzájemně různé jednotky. Pravděpodobnostní rozdělení na množině všech možných výběrů potom je

$$\begin{aligned} P(s) &= c \prod_{i \in s} \alpha_i, \quad K(s) = n, \\ &= 0 \quad \text{jinak}, \end{aligned} \tag{5}$$

kde $0 \leq \alpha_1, \cdots, \alpha_N \leq 1$, $\sum_{i=1}^{N} \alpha_i = 1$, jsou pravděpodobnosti vytažení a c je konstanta, pro kterou $\sum_{s \subset S} P(s) = 1$. Hájek v [6] našel asymptotické vztahy mezi pravděpodobnostmi zamítnutí π_i a volitelnými pravděpodobnostmi vytažení α_i, dále vyjádřil pravděpodobnosti zahrnutí druhého řádu pomocí pravděpodobností prvního řádu, což mu umožnilo definovat nový varianční odhad úhrnu, a rovněž dokázal asymptotickou normalitu prostého lineárního odhadu. Kromě toho v [3] dokázal, že zamítací výběr má maximální entropii ve srovnání s jinými výběrovými plány se stejným rozsahem a stejnými pravděpodobnostmi zahrnutí. V [3] J. Hájek studoval v souvislosti s výběrovo-odhadovou strategií i obecný lineární odhad úhrnu pro obecný výběrový plán, jehož speciální varianta

$$\widehat{y}_H = N \frac{\sum_{i \in s} \frac{y_i}{\pi_i}}{\sum_{i \in s} \frac{1}{\pi_i}} \tag{6}$$

je často uváděna pod názvem Hájkův odhad. Tento odhad sice není nestranný, ale pro mnoho výběrových plánů má menší rozptyl než prostý lineární odhad.

Hájkovy práce z tohoto období jsou stále velmi citovány a byly zobecněny na celé třídy výběrových plánů s nestejnými pravděpodobnostmi zahrnutí. Jaroslav Hájek světově proslul rovněž svou teorií pořadových testů, kterou se zabýval v šedesátých letech, ale na začátku 70. let se znovu vrátil k teorii výběrových šetření, o kterých chtěl napsat novou monografii.[12] Ve spolupráci s Výzkumným ústavem sociálně ekonomických informací vydal první část zamýšlené monografie v češtině[13]. V akademickém roce 1972/73 byl tento text podrobně studován na semináři katedry matematické statistiky pod Hájkovým vedením. Na tomto semináři byly také formulovány otevřené problémy, některé z nich byly později vyřešeny a publikovány.[14] Hájek měl v úmyslu podobným způsobem pokračovat na dalších kapitolách, k tomu však již nedošlo. Když v roce 1974 zemřel, v jeho pozůstalosti se našel neúplný text a poznámky, které jeho kolegové pod vedením prof. Dupače uspořádali, takže nakonec se podařilo knihu s určitým zpožděním publikovat, viz [7].

4. Obecné teoretické principy výběrových šetření

V 60. letech minulého století se pravděpodobnostní výběr prosazovaný Neymanem stal standardní metodou výběrových šetření. V tomto přístupu je výběr pořízen náhodně podle pravděpodobnostního rozdělení P na množině všech možných výběrů z dané populace. Dále se předpokládá, že hodnoty $y_1, \ldots, y_N$ jsou konstanty. Základním problémem obvykle je na základě výběru odhadnout (nestranně) populační parametry (průměr či úhrn a rozptyl) a užít centrální limitní věty pro konstrukci intervalů spolehlivosti.

Statistické vlastnosti odhadů jsou tedy odvozeny z výběrového plánu; proto se tento přístup nazývá *design-based, založený na pravděpodobnostním plánu*. Volba pravděpodobnostního rozdělení je často motivována praktickými nebo administrativními důvody. Na jednotkách populace se však obvykle neuvažuje jediná proměnná y, ale často je k dispozici nějaká pomocná proměnná x známá např. z minulých šetření. Vztah mezi těmito proměnnými lze modelovat nějakým regresním modelem. Přístup, který se v současnosti označuje jako *model-assisted, s pomocí modelu*, předpokládá, že pomocná proměnná je známá buď na celé populaci nebo aspoň na vybraném vzorku a kromě

[12] Souvislost asymptotické teorie pořadových testů a výběrů z konečné populace je podrobně studována např. v [19].

[13] *Teorie výběru z konečných populací*, výzkumná zpráva 43, Výzkumný ústav sociálně ekonomických informací, Praha, 1972.

[14] Přehled lze nalézt např. v [8], kap. 40.

toho je znám její populační úhrn. Regresní parametry se potom odhadnou z hodnot (x, y) na daném vzorku, zatímco populační parametry se odhadují metodami výběrových šetření, tedy na základě výběrového plánu, nezávisle na předpokládaném modelu. Tato metoda je robustní vůči volbě modelu; při správné specifikaci modelu jsou pořízené odhady eficientnější. Tento přístup je uplatněn a podrobně vyložen např. v monografii [18].

Od 70. let[15] se užívá i odlišný přístup, který předpokládá, že hodnoty $y_1, \ldots, y_N$ jsou realizace náhodných veličin $Y_1, \ldots, Y_N$ (superpopulace) s rozdělením, které je funkcí nějakého parametru. Obvykle se předpokládá, že existuje hustota tohoto rozdělení. Populační charakteristiky (např. úhrn) jsou náhodné veličiny. Statistické usuzování se děje na základě výběru, který nemusí být nutně pořízen podle pravděpodobnostního schématu (může tedy být i záměrný). Parametry modelu se odhadují běžnými statistickými metodami, populační charakteristiky se odhadují pomocí predikce nebo bayesovsky. Tomuto postupu se říká *model-based, založený na modelu*.

Mezi zastánci a přívrženci postupů založených na pravděpodobnostním výběru a postupů založených na modelu (pro srovnání viz [16]) se vedly dlouhodobé metodologické spory. V současné době se uznávají oba přístupy a existují práce, které kombinují oba postupy (např. kalibrační postupy, [17]).

Počet prací, které se zabývají teorií i metodami výběrových šetření je enormní. Současný stav řešení problematiky výběrových šetření, včetně problematiky neodpovědí, malých výběrů, chybějících dat, bootstrapu, aplikací, softwaru, administrativních dat, metod výběru atd. je dobře zachycen v dvoudílném svazku [8], který v podstatě pokrývá všechny oblasti výběrových šetření, a je určen pro státní organizace, sociology, velké firmy, ale i pro studenty a vědecké pracovníky.

Budoucí vývoj v teorii výběrových šetření se bude muset orientovat zejména na problematiku velkých datových souborů (big data). Některé problémy a jejich řešení jsou naznačeny např. v článku [12], kde je rovněž možno nalézt další odkazy. Zde je prostor stále otevřený.

[15] Royall, R. M. (1970): On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377–387.

Literatura

[1] Fabian, F. (1964): K sedmdesátým narozeninám prof. Dr. Jaroslava Janko DrSc. *Časopis pro pěstování matematiky* **89**, 117–122.

[2] Hájek, J. (1949): Representativní výběr skupin metodou dvou fází. *Statistický obzor* **29**, 384–394.

[3] Hájek, J. (1959): Optimal strategy and other problems in probability sampling. *Časopis pro pěstování matematiky* **84**, 387–423.

[4] Hájek, J. (1960): *Teorie pravděpodobnostního výběru s aplikacemi na výběrová šetření.* NČSAV, Praha, 1960.

[5] Hájek, J. (1960): Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 361–374.

[6] Hájek, J. (1964): Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* **35**, 1491–1523.

[7] Hájek, J. (1981): *Sampling from a Finite Population.* Marcel Dekker, inc., New York and Basel, 1981.

[8] *Handbook of Statistics 29*, D. Pfeffermann and C. R. Rao, Eds, Elsevier-North-Holland, 2009.

[9] Janko, J. (1937): *Základy statistické indukce.* Státní úřad statistický, Praha, 1937.

[10] Janko, J. (1942 a 1944): *Jak vytváří statistika obrazy světa a života I, II.* JČMF, Praha, 1942, 1944.

[11] Janko, J. (1958): *Statistické tabulky.* NČSAV, Praha, 1958.

[12] Kim, J. K., Wang, Z. (2018): Sampling techniques for big data analysis. *Internat. Statist. Rev.*, DOI: `10.1111/insr.12290`

[13] Kruskal, W., Mosteller, F. (1980): Representative Sampling IV: The history of the concept in statistics, 1895–1939. *Internat. Statist. Rev.* **48**, 169–195.

[14] Neyman, J. (1934): On the two different aspects of the representative method: the method of stratified sampling, and the method of purposive selection. *J. Roy. Statist. Soc.* **97**, 558–606.

[15] Pištora, L. (2018): Hundred years of the Czech statistics. *Statistika* **98**, 385–389.

[16] Särndal, C.-E. (1978): Design-based and model-based inference in survey sampling. *Scand. J. Statist.* **5**, 27–52.

[17] Särndal, C.-E. (2011): Combined inference in survey sampling. *Pakist. J. Statist.* **27**, 359–370.

[18] Särndal, C.-E., Swensson, B., Wretman, J. H. (1992): *Model Assisted Survey Sampling*, Springer Verlag, New York, 1992.

[19] Sen, P. K. (1995): The Hájek asymptotics for finite population sampling and their ramifications. *Kybernetika* **31**, 251–268.

[20] Truksa, L. (1954): Šedesát let profesora Dr. Jaroslava Janko. *Časopis pro pěstování matematiky* **79**, 181–185.

[21] Závodský, P. (2010): Meziválečná československá statistická společnost. *Informační Bulletin České statistické společnosti* **22**, 4, 3–7.

[22] Zichová, J. (2009): Josef Erben a jeho přínos pro pražskou statistiku v 19. století. *Pokroky matematiky, fyziky a astronomie* **54**, 57–71.

[23] Český statistický úřad. URL: `https://www.czso.cz/`

LOGISTICKÁ REGRESE S KATEGORIÁLNÍ VYSVĚTLUJÍCÍ PROMĚNNOU: JAK TO, ŽE STATISTICA DÁVÁ JINÉ VÝSLEDKY NEŽ R?

LOGISTIC REGRESSION WITH CATEGORICAL EXPLANATORY VARIABLE: HOW IT COMES THAT RESULTS FROM STATISTICA AND R DIFFER?

Ondřej Vencálek[1], Jana Fürstová[2]

Adresa: [1] Katedra matematické analýzy a aplikací matematiky, PřF, Univerzita Palackého v Olomouci, 17. listopadu 12, 771 46 Olomouc
[2] Cyrilometodějská teologická fakulta, Univerzita Palackého v Olomouci, Univerzitní 244/22, 779 00 Olomouc

E-mail: ondrej.vencalek@upol.cz

Abstrakt: V příspěvku připomínáme možnost různě parametrizovat úlohu porovnání pravděpodobnosti výskytu určitého jevu (např. zlepšení zdravotního stavu) ve dvou či více skupinách (definovaných např. podávaným lékem). Nejednoznačnost parametrizace musíme mít na paměti, když analýzu provádíme v programu, na který nejsme zvyklí. Jiná parametrizace nás může zaskočit, v horším případě může vést ke špatným interpretacím odhadnutých parametrů.

Klíčová slova: Logistická regrese, parametrizace, R, Statistica.

Abstract: In the current contribution, we recall possibility of different parametrizations when describing model for comparison of chances (e.g. chances of recovery) in two or more groups (defined e.g. by different treatment). We must be aware of the ambiguity of parameterization mainly when performing analysis in a software that we are not used to. Another parameterization may be confusing, and may even lead to wrong interpretation of the estimated parameters.

Keywords: Logistic regression, parametrization, R, Statistica.

1. Úvod

Dostupnost různých programových nástrojů pro analýzu dat sice na jednu stranu otevírá možnost provedení statistických analýz širšímu okruhu zájemců, na straně druhé však vede k situacím, kdy se někteří uživatelé mohou cítit zmateně. Taková situace může nastat, pokud například chceme odhadnout parametry modelu logistické regrese při porovnání pravděpodobnosti

výskytu určitého jevu ve dvou či více skupinách a tento odhad provedeme jednak pomocí programu Statistica a poté pomocí programu R. Jenže ouha – každý z těchto dvou programů nám vrací jiný výsledek. Kde se stala chyba? Vysvětlení je poměrně jednoduché – model logistické regrese s kategoriální vysvětlující proměnnou (ale také jiné modely s kategoriálními vysvětlujícími proměnnými) může být různě parametrizován. Náš příspěvek si klade za cíl připomenout nejčastěji používané parametrizace. Zastavíme se u interpretace jednotlivých parametrů různě parametrizovaných modelů a ukážeme, jak změnit volbu parametrizace v programech R [2] a Statistica [3].

2. Případ srovnání šancí ve dvou skupinách

Představme si situaci, kdy máme pro léčbu určité nemoci k dispozici dva různé léky (označme je třeba A a B). Zajímá nás, zda jsou oba léky stejně účinné. Chceme proto porovnat pravděpodobnost zlepšení při podání léku A a pravděpodobnost zlepšení při podání léku B. Jednoduchý matematický popis této situace: veličina Y nabyde hodnoty 1, pokud lék u pacienta zabere a hodnoty 0 v opačném případě. Jelikož je výsledek dopředu nejistý, lze veličinu Y považovat za náhodnou. Rozdělení této veličiny (alternativní rozdělení) je popsáno parametrem p, který vyjadřuje pravděpodobnost, že Y nabyde hodnoty 1, tj. že u pacienta dojde k zlepšení:

$$p = \mathrm{P}(Y = 1).$$

Parametr p je číslo z intervalu $(0, 1)$, které záleží na tom, z které skupiny je pacient vybrán. Můžeme se totiž zabývat např. jen skupinou, jíž je podáván lék A (skupina A), nebo naopak jen skupinou, jíž je podáván lék B (skupina B). Místo pravděpodobnosti p můžeme pracovat také s šancí ω. Mezi těmito číselnými charakteristikami existuje jednoznačný vztah:

$$\omega = \frac{p}{1-p}.$$

Abychom mohli celou situaci dobře popsat řečí čísel, uvažujme (nenáhodnou) veličinu x, která nabyde určité hodnoty, pokud je pacient ze skupiny A a jiné hodnoty, pokud je pacient ze skupiny B. Existuje vícero způsobů, jak přiřadit hodnoty x. A právě v této nejednoznačnosti spočívá podstata problému – různé programy používají různé volby. My zde popíšeme standardní volby programu R a Statistica.

- **1. způsob (standard v programu R**[1]**)**: Jednu ze skupin zvolíme jako „referenční“ a přiřadíme jí hodnotu $x = 0$, druhé skupině přiřadíme hodnotu $x = 1$.
- **2. způsob (standard v programu Statistica)**: Jedné skupině přiřadíme hodnotu $x = 1$, druhé $x = -1$.

Model logistické regrese má v obou případech formálně stejnou podobu:

$$\ln \omega(x) = \alpha + \beta x, \tag{1}$$

jelikož však x kóduje skupiny různými způsoby, mají parametry při rozdílných kódováních rozdílnou interpretaci a nabývají rozdílných hodnot (jde o rozdílné parametrizace modelu). Abychom oba způsoby kódování od sebe odlišili, přidáme k parametrům (horní) index (R) pro kódování obvyklé v programu R a (S) pro kódování obvyklé ve Statistice.

2.1. Interpretace parametrů v kódování (R)

Věnujme se nyní interpretaci parametrů pro kódování obvyklé v R. Platí

$$\ln \omega(x^{(R)}) = \alpha^{(R)} + \beta^{(R)} x^{(R)}. \tag{2}$$

Pro skupinu A platí $x^{(R)} = 0$. Po dosazení této hodnoty do (2) dostáváme

$$\ln \omega(x^{(R)} = 0) = \alpha^{(R)}. \tag{3}$$

Pro skupinu B platí $x^{(R)} = 1$ a tedy

$$\ln \omega(x^{(R)} = 1) = \alpha^{(R)} + \beta^{(R)}. \tag{4}$$

V této parametrizaci tudíž

- $\mathrm{e}^{\alpha^{(R)}}$ vyjadřuje šanci (na zlepšení) ve skupině, které je podáván lék A (tj. v referenční skupině).
- $\mathrm{e}^{\beta^{(R)}}$ vyjadřuje poměr šancí na zlepšení skupiny B proti skupině A, jak plyne z rozdílu rovností (3) a (4).

2.2. Interpretace parametrů v kódování (S)

Rozmysleme si nyní interpretaci parametrů při druhém způsobu kódování, který je běžný ve Statistice. V rovnici modelu logistické regrese (1) nyní budeme psát index (S):

$$\ln \omega(x^{(S)}) = \alpha^{(S)} + \beta^{(S)} x^{(S)}. \tag{5}$$

[1] Dle sdělení jednoho z recenzentů je stejný standard také v programech Stata či SPSS.

Pro skupinu A platí $x^{(S)} = +1$ a tedy

$$\ln \omega(x^{(S)} = +1) = \alpha^{(S)} + \beta^{(S)}. \tag{6}$$

Pro skupinu B platí $x^{(S)} = -1$ a tudíž

$$\ln \omega(x^{(S)} = -1) = \alpha^{(S)} - \beta^{(S)}. \tag{7}$$

Sečtením respektive odečtením rovností (6) a (7) a jejich odlogaritmováním zjistíme, že

- $\mathrm{e}^{\alpha^{(S)}} = \sqrt{\omega(x^{(S)} = -1)\omega(x^{(S)} = +1)}$, jde tedy o geometrický průměr šancí v obou skupinách,
- $\mathrm{e}^{-2\beta^{(S)}}$ vyjadřuje poměr šancí na zlepšení skupiny B proti skupině A.

2.3. Vztahy mezi parametry v kódování (R) a (S)

Srovnáním rovnosti (3) s (6) a rovnosti (4) s (7) získáme vzájemný vztah mezi parametry v obou parametrizacích:

$$\begin{aligned} \alpha^{(R)} &= \alpha^{(S)} + \beta^{(S)}, \\ \alpha^{(R)} + \beta^{(R)} &= \alpha^{(S)} - \beta^{(S)}. \end{aligned}$$

Parametry kódování (R) můžeme tedy vyjádřit pomocí parametrů v kódování (S):

$$\begin{aligned} \alpha^{(R)} &= \alpha^{(S)} + \beta^{(S)}, \\ \beta^{(R)} &= -2\beta^{(S)} \end{aligned}$$

a naopak parametry modelu v kódování (S) lze vyjádřit pomocí parametrů modelu v kódování (R):

$$\begin{aligned} \alpha^{(S)} &= \alpha^{(R)} + \beta^{(R)}/2, \\ \beta^{(S)} &= -\beta^{(R)}/2. \end{aligned}$$

Parametry modelu používajícího jednu parametrizaci lze tedy přímo vypočítat z parametrů modelu využívajícího druhou parametrizaci. Neméně důležité je však také to, že díky Zehnaově principu invariance platí stejné vztahy také mezi odhady parametrů získanými metodou maximální věrohodnosti, viz [1], věta 7.87.

2.4. Příklad

Statistické výpočetní programy většinou vrací odhady parametrů $\alpha^{(R)}$ a $\beta^{(R)}$, případně $\alpha^{(S)}$ a $\beta^{(S)}$. Z nich potom dopočítáváme lépe interpretovatelné charakteristiky jako šance v jednotlivých skupinách či jejich poměry, které nezávisí na zvolené parametrizaci. My nyní postup obrátíme a z daných šancí vypočteme hodnoty parametrů $\alpha^{(R)}$, $\beta^{(R)}$, $\alpha^{(S)}$ a $\beta^{(S)}$. Poté na vhodně zvolených fiktivních datech necháme odhadnout parametry modelů v programu R a Statistica.

Nechť šance na zlepšení je 1:2 ve skupině A a 1:3 ve skupině B. Vzhledem k výše uvedeným interpretacím jednotlivých parametrů můžeme přímo psát

- $\mathrm{e}^{\alpha^{(R)}} = 1:2$, tudíž $\alpha^{(R)} = \ln(1/2) \doteq -0{,}693$,
- $\mathrm{e}^{\beta^{(R)}} = (1:3)/(1:2)$, tudíž $\beta^{(R)} = \ln(2/3) \doteq -0{,}405$,
- $\mathrm{e}^{\alpha^{(S)}} = \sqrt{\frac{1}{2}\frac{1}{3}}$, tudíž $\alpha^{(S)} = \frac{1}{2}\ln(1/6) \doteq -0{,}896$,
- $\mathrm{e}^{-2\beta^{(S)}} = (1:3)/(1:2)$, tudíž $\beta^{(S)} = \frac{1}{2}\ln(3/2) \doteq 0{,}203$.

Nyní uvažujme datovou sadu (pro jednoduchost co nejmenší), v níž empiricky pozorované šance jsou 1:2 a 1:3. Stačí mít 7 pozorování – 3 ze skupiny A, kde zlepšení nastalo v jediném případě, a 4 ze skupiny B, kde zlepšení nastalo rovněž v jediném případě. Odhad parametrů v R provedeme pomocí následujícího kódu:

```
> x  = factor(c(0,0,0, 1,1,1,1))
> y  =        c(0,0,1, 0,0,0,1)
> m0 = glm(y~x,family="binomial")
```

Odhady parametrů můžeme zobrazit příkazem `summary()`:

```
> summary(m0)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6931     1.2247  -0.566    0.571
x1           -0.4055     1.6832  -0.241    0.810
```

Jak vidíme, získali jsme odhady parametrů $\alpha^{(R)}$ a $\beta^{(R)}$.

Ve Statistice budeme postupovat následovně:

1. Načteme (či zadáme) data. Sloupec odpovídající zdravotnímu stavu značíme y, sloupec obsahující identifikátor skupiny (léčby) máme označen x.

2. V nabídce *Statistics* zvolíme *Advanced Models* a vybereme *Generalized Linear/Nonlinear* a poté *Logit model*.
3. V otevřeném okně *GLZ General custom design* specifikujeme proměnné – *Variables*: závisle proměnnou je y, nezávisle proměnnou x.
4. Upravíme *Response codes*: do prázdného řádku napíšeme s mezerou hodnoty 1 0, čímž určíme, že jednička kóduje událost, jejíž šance nás zajímá.
5. Po potvrzení předchozí volby se otevře okno *GLZ–Results*, v něm klikneme na *Estimates*.
6. Výstup má dvě části: odhady *Odds Ratios* (otevírá se automaticky) a *Parameter estimates* (na tuto část se můžeme „prokliknout“ v dolní či v levé části aktuálního okna).

2.5. Jak změnit parametrizaci v R a ve Statistice

Oba námi zmiňované programy umožňují svým uživatelům změnit standardně nastavenou parametrizaci. V R stačí změnit standardní nastavení argumentu `contrasts` v příkazu `glm`. Parametrizaci, která je standardem ve Statistice, můžeme v programu R nastavit příkazem

```
m = glm(y~x,family="binomial",contrasts = list(x = "contr.sum"))
```

Odhady parametrů získáme opět příkazem `summary()`:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8959     0.8416  -1.064    0.287
x1            0.2027     0.8416   0.241    0.810
```

Ve Statistice můžeme změnit parametrizaci tak, že mezi 3. a 4. krok výše popsaného postupu zadávání modelu vložíme ještě následující krok: v nabídce *GLZ General custom design* zvolíme možnost *Advanced* a z nabízených možností v oddíle *Parametrization* vybereme *Ref* (parametrizace používající referenční skupinu). Poté je ještě třeba nastavit, která úroveň má být považována za referenční (*Set reference level*).

2.6. Test významnosti rozdílu mezi skupinami

Chceme-li testovat, zda je pravděpodobnost studovaného jevu (zlepšení zdravotního stavu) stejná v obou skupinách, zajímá nás hypotéza $\mathrm{H}_0\colon \beta^{(R)} = 0$, respektive $\mathrm{H}_0\colon \beta^{(S)} = 0$. Připomeňme vztah $\beta^{(R)} = -2\beta^{(S)}$, viz kapitola 2.3.

Tvrzení o nulovosti obou parametrů jsou tedy ekvivalentní. Proto nás nepřekvapí, že p-hodnoty obou těchto testů jsou stejné, viz výstupy z programu R v kapitolách 2.4 a 2.5: p-hodnoty na řádku označeném `x1` jsou v obou případech 0,810.

3. Zobecnění pro víceúrovňový faktor

Ve výše uvedeném příkladě jsme srovnávali dvě skupiny – faktor „podávaný lék“ měl dvě úrovně. Co když bude ale úrovní více? Například kdybychom porovnávali tři nebo čtyři léčiva? Oba způsoby parametrizace úlohy mohou být použity i v tomto případě.

Uvažujme faktor, který má I různých úrovní (I je celé číslo větší než 1). Označme pro jednoduchost šanci v i-té skupině symbolem ω_i, kde $i = 1, \ldots, I$. Model můžeme zapsat opět jednotně (bez ohledu na parametrizaci), tentokrát pomocí $I - 1$ pomocných vysvětlujících proměnných, uspořádaných do vektoru $\boldsymbol{x} = (x_1, \ldots, x_{I-1})'$. Aby nedošlo ke zmatení, budeme opět hodnoty parametrů i vysvětlujících proměnných v jednotlivých parametrizacích označovat (horními) indexy (R) a (S).

- 1. způsob parametrizace (standardní v programu R): V parametrizaci běžné v programu R je pro $i = 1, \ldots, I - 1$

$$x_i^{(R)} = \begin{cases} 1, & \text{pokud je jedinec z } (i+1)\text{-ní skupiny,} \\ 0, & \text{jinak.} \end{cases}$$

  Jedinci z první (referenční) skupiny tak mají všechny hodnoty nulové: $x_1^{(R)} = \ldots = x_{I-1}^{(R)} = 0$. Šance v jednotlivých skupinách jsou tedy rovny

$$\begin{aligned} \ln \omega_1 &= \alpha^{(R)}, \\ \ln \omega_{i+1} &= \alpha^{(R)} + \beta_i^{(R)}, \quad i = 1, \ldots, I-1. \end{aligned}$$

  Odtud můžeme získat interpretaci jednotlivých parametrů:

  $\mathrm{e}^{\alpha^{(R)}} = \omega_1$ je šance v referenční skupině (skupině 1),

  $\mathrm{e}^{\beta_i^{(R)}} = \frac{\omega_{i+1}}{\omega_1}, i = 1, \ldots, I-1$, je poměr šance $(i+1)$-ní skupiny ku šanci referenční skupiny.

- 2. způsob parametrizace (standardní v programu Statistica): V parametrizaci běžné ve Statistice je pro $i = 1, \ldots, I-1$

$$x_i^{(S)} = \begin{cases} 1, & \text{pokud je jedinec z } i\text{-té skupiny,} \\ -1, & \text{pokud je jedinec z } I\text{-té skupiny,} \\ 0, & \text{jinak.} \end{cases}$$

Šance v jednotlivých skupinách jsou tedy rovny

$$\begin{aligned} \ln \omega_i &= \alpha^{(S)} + \beta_i^{(S)}, \quad i = 1, \ldots, I-1, \\ \ln \omega_I &= \alpha^{(S)} - \beta_1^{(S)} - \ldots - \beta_{I-1}^{(S)}. \end{aligned}$$

Sečtením všech I rovností, vydělením hodnotou I a následným odlogaritmováním získáme rovnost

$$\mathrm{e}^{\alpha^{(S)}} = \left(\prod_{i=1}^{I} \omega_i \right)^{\frac{1}{I}} =: g(\boldsymbol{\omega}),$$

kde $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_I)'$ a $g(\boldsymbol{\omega})$ je geometrický průměr šancí ve všech skupinách. Dosazením hodnoty $\ln(g(\boldsymbol{\omega}))$ za $\alpha^{(S)}$ v rovnicích pro $\ln \omega_i$ získáme interpretaci parametrů $\beta_i^{(S)}$:

$$\mathrm{e}^{\beta_i^{(S)}} = \frac{\omega_i}{g(\boldsymbol{\omega})}, \quad i = 1, \ldots, I-1.$$

Tedy

$\mathrm{e}^{\alpha^{(S)}}$ interpretujeme jako průměrnou šanci jednotlivých skupin (jde o geometrický průměr)

$\mathrm{e}^{\beta_i^{(S)}}$, $i = 1, \ldots, I-1$, je podíl šance i-té skupiny ku průměrné šanci.

Doporučujeme čtenáři, aby početně a poté za pomocí programu zkusil zjistit hodnoty $\alpha^{(R)}, \beta_1^{(R)}, \beta_2^{(R)}$ a $\alpha^{(S)}, \beta_1^{(S)}, \beta_2^{(S)}$ pro případ tří skupin, kde $\omega_1 = 1:2$, $\omega_2 = 1:3$, $\omega_3 = 4:1$ (opět můžeme pracovat s „minimálními“ daty obsahujícími 3, 4, respektive 5 pozorování z námi uvažovaných skupin, v nichž pozorované šance odpovídají výše uvedeným poměrům).

Řešení: $\alpha^{(R)} = -0{,}693$, $\beta_1^{(R)} = -0{,}405$, $\beta_2^{(R)} = 2{,}079$ (všimněte si, že první dva parametry jsou stejné jako v příkladě srovnání dvou kategorií v předchozí kapitole), $\alpha^{(S)} = -0{,}135$, $\beta_1^{(S)} = -0{,}558$, $\beta_2^{(S)} = -0{,}963$ (všimněte si, že v této parametrizaci nedochází ke shodě parametrů jako u parametrizace obvyklé v R, protože se mění průměrná šance, s níž jsou šance v jednotlivých skupinách srovnávány).

3.1. Matice popisující parametrizace

Obě parametrizace můžeme přehledně popsat pomocí matic o I řádcích (řádky reprezentují skupiny) a $I-1$ sloupcích (sloupce reprezentují parametry značené β):

$$\boldsymbol{C}^{(R)} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \boldsymbol{C}^{(S)} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -1 & -1 & \cdots & -1 \end{pmatrix}$$

Při označení $\ln(\boldsymbol{\omega}) = (\ln\omega_1, \ldots, \ln\omega_I)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{I-1})'$ a použití horních indexů (R) a (S) k rozlišení parametrizace, můžeme psát

$$\begin{aligned} \ln(\boldsymbol{\omega}) &= \alpha^{(R)}\mathbf{1} + \boldsymbol{C}^{(R)}\boldsymbol{\beta}^{(R)}, \\ \ln(\boldsymbol{\omega}) &= \alpha^{(S)}\mathbf{1} + \boldsymbol{C}^{(S)}\boldsymbol{\beta}^{(S)}. \end{aligned}$$

V R můžeme získat matice $\boldsymbol{C}^{(R)}$ a $\boldsymbol{C}^{(S)}$ příkazy `contr.treatment(I)`, respektive `contr.sum(I)`, kde za `I` dosadíme příslušný počet kategorií.

Poznamenejme, že parametrizací může být víc než jsou zde uváděné dvě. Pro zájemce o tuto problematiku lze doporučit 6. kapitolu monografie [4], obsahující výklad o tzv. *kontrastech*.

3.2. Testování rozdílu mezi skupinami

Ukázali jsme, že při použití různých parametrizací mají parametry modelu rozdílné interpretace. Proto také nulovost jednotlivých parametrů má rozdílnou interpretaci. Hypotéza $\mathrm{H}_0\colon \beta_i^{(R)} = 0$ říká, že šance v $(i+1)$-ní skupině je stejná jako v první skupině. To je (pro případ více než dvou skupin) jiné tvrzení, než $\mathrm{H}_0\colon \beta_i^{(S)} = 0$, které říká, že šance v i-té skupině je stejná jako průměrná šance. Je dobré si uvědomit, že jsou to právě testy nulovosti jednotlivých parametrů, jejichž p-hodnoty nám programy automaticky vrací, a neočekávat (vzhledem k výše uvedené rozdílnosti) shodu p-hodnot.

Někoho by snad mohlo napadnout, že významnost faktoru určujícího jednotlivé skupiny (např. léčby) je tedy dána parametrizací modelu (výběrem vhodné parametrizace můžeme získávat výsledky podle potřeby). Není tomu tak. Ptáme-li se na významost faktoru určujícího skupiny, zajímá nás totiž složená hypotéza $\mathrm{H}_0\colon \beta_i^{(R)} = 0$ pro všechny $i = 1, \ldots, I-1$, respektive

$\mathrm{H}_0\colon \beta_i^{(S)} = 0$ pro všechny $i = 1, \ldots, I-1$. Při použití obou parametrizací dojdeme ke stejné p-hodnotě a tudíž ke stejnému závěru.

4. Závěr

Tímto článkem jsme se pokusili upozornit na nejednoznačnost parametrizace v situaci, kdy chceme porovnávat pravděpodobnosti výskytu určitého jevu ve dvou či více skupinách. Uživatel statistického programu si musí být dobře vědom, která parametrizace je při analýze použita, protože parametry příslušné kvalitativním proměnným mají v různých parametrizacích různou interpretaci. S tím je spojen také problém interpretace výsledku testování hypotézy o nulovosti parametru. Zatímco v jedné z námi uvažovaných parametrizací znamená nulovost parametru shodu pravděpodobnosti výskytu daného jevu v příslušné skupině a v referenční skupině, ve druhé parametrizaci jde o shodu s průměrem všech skupin. Různé programy mají různá standardní nastavení, většinou však umožňují použít i jinou než standardní parametrizaci.

Literatura

[1] Anděl, J. (2005): *Základy matematické statistiky.* Matfyzpress, Praha, 2005.

[2] R Core Team (2019): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`

[3] TIBCO Software Inc. (2017): *STATISTICA Data Analysis Software System, Version 13, 2017.* URL: `http://www.statsoft.com/`

[4] Zvára, K. (2008): *Regrese.* Matfyzpress, Praha, 2008.

SETKÁNÍ ZÁSTUPCŮ VÝBORU ČSTS S DELEGACÍ REGIONÁLNÍHO STATISTICKÉHO ÚŘADU ČÍNSKÉ PROVINCIE ČE-ŤIANG

MEETING OF THE REPRESENTATIVES OF THE CZECH STATISTICAL SOCIETY AND PROVINCIAL BUREAU OF STATISTICS OF ZHÈJIĀNG IN CHINA

Ondřej Vozár

E-mail: ondrej.vozar@czso.cz

Ve čtvrtek 13. června 2019 od 14:00 do 16:30 proběhlo jednání delegace Regionálního statistického úřadu (Provincial Bureau of Statistics) významné čínské provincie Če-ťiang (čínsky: 浙江; pinyin: Zhèjiāng) se zástupci výboru České statistické společnosti.

Vedoucím čínské delegace byl pan Wang Jie, předseda úřadu. Výbor ČStS reprezentovali předseda společnosti Ondřej Vencálek, prof. Jaromír Antoch, prof. Gejza Dohnal a Ondřej Vozár. Jednání bylo zaměřeno na výměnu zkušeností z oblasti výuky statistiky, státní statistiky a informace o možnostech případné další spolupráce.

Po představení zástupců obou stran proběhla diskuse, jež se v první části zaměřila na porovnání systémů vysokoškolského vzdělávání v oblasti statistiky v Číně a v ČR. Systémy jsou obdobné, studují se odděleně matematické obory, ekonomická a sociální statistika, bioinformatika a průmyslová statistika. V obou zemích je uplatnění absolventů obdobné, zaměstnání ve státních i statistických úřadech je zpravidla finančně méně atraktivní než v soukromém sektoru.

Druhá část diskuse se více zaměřila na problematiku oficiální statistiky. Čínská strana se zajímala především o organizaci státní statistické služby v ČR. Dále byla diskutována problematika publikování HDP. Z ČR byl představen systém publikování údajů v předem pevně stanovených termínech a princip neposkytování dat před těmito termíny. Další téma bylo sdílení dat mezi pracovišti státní služby, poskytování mikrodat a nestandardních výstupů výzkumníkům a studentům a publikování metodických informací. Z diskuse vyplynulo, že v Číně je, na rozdíl od ČR, povinné sdílení údajů s ministerstvy. Na druhé straně se zde plně nectí zásada využívat data od respondentů pouze pro statistické účely, takže na rozdíl od ČR se získaná data mohou používat například i pro kontrolní účely. Nicméně čínská strana si je tohoto problému vědoma.

Zajímavé nám přišlo, že na rozdíl od ČR je v Číně povinné ke všem výstupům státní statistiky připojit detailní metodologii, podle které byly publikované výsledky spočteny, respektive mít potřebnou metodologii vystavenu na webu. Důvod, že je tak možno lépe a snáze veřejně kontrolovat práci statistiků, považujeme za více než přínosný. V ČR i v Evropě toto pravidlo platí pouze pro některé klíčové ukazatele. To neznamená, že dané metodologie neexistují, jsou však uživatelům statistiky zpravidla skryty.

Závěrem byly čínskými kolegy prezentovány stručné údaje o provincii Če-ťiang. Provinční statistický úřad mimo jiné projednal s OSN vytvoření centra zabývajícího se výzkumem problematiky tak zvaných velkých dat (big data) ve statistice. Za české kolegy poté prof. Antoch prezentoval stručnou historii Univerzity Karlovy a organizaci studia statistiky na MFF UK.

Čínská delegace byla informována o chystané konferenci CESS 2020, jež se uskuteční v září 2020 v Praze a kterou je ČStS spoluorganizátorem, viz `https://cess2020.nipax.cz/`. Ta představuje vhodnou platformu k výměně poznatků a navazování spolupráce především v oblasti státní a ekonomické statistiky. Čeští zástupci též vyjádřili zájem o spolupráci se státní univerzitou v Če-ťiangu s vznikajícím centrem pro analýzu velkých dat.

POZVÁNKA NA KONFERENCI CESS 2020
INVITATION TO THE CESS 2020 CONFERENCE

Redakce

The Conference of European Statistics Stakeholders 2020 will be held in Prague, the Czech Republic on 29–30 September 2020. It is co-organised by Eurostat, the European Central Bank, the European Statistical Advisory Committee, the United Nations Economic Commission for Europe, the Federation of European National Statistical Societies, the Czech Statistical Office, the Czech Statistical Society, the Charles University and the Czech Technical University in Prague.

Launched for the first time in 2014, with the idea of creating a platform for discussing statistics methodology, results, challenges and best practices between the producers of European statistics, researchers, as users and other stakeholders, the CESS 2020 continues to offer the opportunity for networking and interacting among stakeholders of European Statistics from science, academia, statistical producers and societal groups.

Its aim is to share recent developments and innovative solutions for reflecting and explaining complex economic and social phenomena with European statistics and to discuss challenges associated with their production and use. The Conference will be organized according to three thematic blocks – Statistics, Science and Society.

You are invited to participate in this conference by submitting an abstract/or paper highlighting the relevant thematic blocks by 14 June 2020.

We would like to draw your attention to the publication opportunities for CESS 2020 authors: Selected papers for CESS 2020 will be included within a Conference publication released as an e-publication by Eurostat. Abstracts will be subject to a peer review by members of the Scientific Programme Committee prior to the conference. Authors of selected abstracts will be invited to extend their papers for possible publication in the ECB Statistics Paper Series (see SPS guidance on ECB website, `https://www.ecb.europa.eu/pub/research/statistics-papers/html/index.en.html`). Conference website: `https://cess2020.nipax.cz/`

Local Organizing Committee: `cess2020@npx.cz`

Obsah

Vědecké a odborné články

Zprávy a informace

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in the journal are published in English, Czech and Slovak languages.

Předseda společnosti: Mgr. Ondřej Vencálek, Ph.D., Katedra matematické analýzy a aplikací matematiky, Přírodovědecká fakulta Univerzity Palackého, 17. listopadu 12, 771 46 Olomouc, e-mail: ondrej.vencalek@upol.cz.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.