

Dotazníkový prieskum II. – overovanie hypotéz Questionnaire research II. – hypotheses testing

Alena Kaščáková, Gabriela Nedelová

Abstract: The aim of the paper is to give the overview of the most common and the mostly used techniques in the process of testing the hypotheses from the questionnaire research. After the formulation of the aim of the research and exploratory analysis the hypotheses testing follows with placing importance on the appropriate methods. Testing of representativeness is the first step of the inference process of questionnaire data survey. For the comparison of one, two or more sample means could be used parametric or non-parametric tests, while the second ones have the lower power, but the wider usage.

Abstrakt: Cieľom príspevku je podať prehľad najčastejšie využívaných techník pri overovaní hypotéz z dotazníkového prieskumu. Po stanovení cieľa výskumu a prieskumnej analýze dát nasleduje overenie stanovených hypotéz s výberom vhodných metód. Prvým krokom v zovšeobecňovacom procese výsledkov dotazníkového prieskumu je overenie reprezentatívnosti výberu. Na porovnávanie jednej, dvoch alebo viacerých stredných hodnôt sa využívajú parametrické alebo neparametrické testy, pričom druhé z nich majú síce nižšiu silu, ale širšie možnosti využitia.

Key words: questionnaire, questionnaire research, variable, database, parametric and non-parametric tests

Kľúčové slová: dotazník, dotazníkový prieskum, štatistický znak, databáza, parametrické a neparametrické testy

JEL classification: C18

Úvod

V sociálno-ekonomických výskumoch je potrebné analyzovať množstvo informácií. Nie vždy sú tieto informácie dostupné z zdrojov štátnej štatistiky, prieskumov orgánov štátnej správy či súkromných alebo marketingových firiem. Preto je potrebné zvoliť cestu získania údajov vlastným primárnym prieskumom. Najčastejším spôsobom zberu údajov býva dotazníkový prieskum, ktorý poskytuje informácie kvalitatívneho alebo kvantitatívneho charakteru. Tieto môžu mať formu číselných alebo nečíselných dát, pričom túto formu je potrebné brať do úvahy pri voľbe metód, potrebných na ich vyhodnotenie. Bez ohľadu na voľbu prístupu k zberu dát je možné vo výskume identifikovať štyri základné etapy: prípravná fáza, zber a kontrola údajov, analýza a vyhodnotenie, interpretácia výsledkov a ich publikovanie. Vzhľadom na zameranie príspevku sa budeme v ďalšom bližšie venovať problematike testovania hypotéz o úrovni znakov získaných z dotazníkového prieskumu a metódam, ktoré sa pri testovaní využívajú. Nadvižeme tak na príspevok Dotazníkový prieskum I., publikovaný v predchádzajúcom roku.

1. Overenie reprezentatívnosti

Medzi základné požiadavky na informácie z empirických výskumov patrí okrem objektívnosti, reliability a validity, reprezentatívnosť výberového súboru. Reprezentatívny výberový súbor je taký výberový súbor, v ktorom sú štatistické jednotky zastúpené v rovnakej štruktúre ako v analyzovanom základnom súbore. Reprezentatívnosť sa overuje ako zhoda rozdelenia pravdepodobnosti skúmaných výberových znakov s odpovedajúcimi v základnom súbore. V prípade viacerých znakov sa často uvažuje reprezentatívnosť pre každý znak zvlášť a súbor sa považuje za reprezentatívny, ak je reprezentatívny pre všetky jednotlivé znaky (Luha, 2007). Znaky, podľa ktorých sa overuje reprezentatívnosť súboru sa vyberajú v závislosti od zamerania prieskumu. Znaky sledované v dotazníkových prieskumoch sú

častejšie kvalitatívne ako kvantitatívne. Ak uvažujeme kvalitatívny znak X s kategóriami $x_1, x_2, \dots, x_k$, ktoré nadobúdajú s pravdepodobnosťami $\pi_1, \pi_2, \dots, \pi_k$, kde $\sum_{i=1}^k \pi_i = 1$, označme

absolútne početnosti hodnôt znaku v jednotlivých kategóriách vo výberovom súbore $n_1, n_2, \dots, n_k$ s rozsahom n a príslušné relatívne početnosti ako $p_1, p_2, \dots, p_k$. Pre reprezentatívny výber platí hypotéza H_0 o zhode pravdepodobnostného rozdelenia $(\pi_1, \pi_2, \dots, \pi_k)$ s predpokladaným pravdepodobnostným rozdelením znaku v základnom súbore $(\pi_{10}, \pi_{20}, \dots, \pi_{k0})$. Odhadom rozdelenia $(\pi_1, \pi_2, \dots, \pi_k)$ sú relatívne početnosti $(p_1, p_2, \dots, p_k)$ z výberového súboru. Hypotézu H_0 overíme napríklad χ^2 -testom dobrej zhody (Luha, 2006): $H_0 : \pi_i = \pi_{i0}$ pre všetky $i = 1, 2, \dots, k$ oproti alternatívnej hypotéze H_A : aspoň pre jedno i $\pi_i \neq \pi_{i0}$. Podmienkou použitia testu je dostatočne veľký rozsah výberového súboru a aby aspoň 80% očakávaných početností $n\pi_{i0}$ bolo väčších ako 5 a žiadna nebola menšia ako 1. Splnenie tejto podmienky sa často zabezpečuje zlučovaním tried. Testovacia štatistika

$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\pi_{i0})^2}{n\pi_{i0}}$ má asymptoticky χ^2 -rozdelenie s $k-1$ stupňami voľnosti. Uvedený test

môžeme využiť aj pre kvantitatívne premenné po ich kategorizácii. V prípade kvalitatívnych znakov nadobúdajúcich len dve možné hodnoty kategórií, napríklad znak zisťujúci prítomnosť (prvá kategória) alebo neprítomnosť (druhá kategória) určitej vlastnosti, môžeme reprezentatívnosť overiť testom zhody pravdepodobnosti prvej kategórie s nejakou očakávanou hodnotou: $H_0: \pi = \pi_0$ oproti $H_A: \pi \neq \pi_0$. Za podmienky dostatočného rozsahu

súboru $n\pi_0(1-\pi_0) > 9$ má testovacia štatistika $U = \frac{|p - \pi_0| - \frac{1}{2n}}{s_{\pi_0}}$ ($s_{\pi_0} = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$, $p = \frac{n_1}{n}$)

asymptoticky normálne rozdelenie so strednou hodnotou 0 a smerodajnou odchýlkou 1 (ozn. $N[0,1]$). Ak podmienky predchádzajúceho testu nie sú splnené použijeme exaktný binomický test (Rezanková, 2007, s. 58), v ktorom sa počíta exaktná p-hodnota α' . V programe SPSS sú hypotézy stanovené nasledovne:

a) ak $n_1 < n - n_1$ a $p < \pi_0$, potom $H_0: \pi = \pi_0$ oproti $H_A: \pi < \pi_0$ a p-hodnota

$$\alpha' = \sum_{i=0}^{n_1} \binom{n}{i} \pi^i (1-\pi)^{n-i}, \text{ čo je pravdepodobnosť, že pri binomickom rozdelení}$$

s parametrami n a π nadobudne premenná hodnoty menšie alebo rovné n_1 ,

b) ak $n_1 < n - n_1$ a $p > \pi_0$ sú hypotézy stanovené nasledovne $H_0: \pi = \pi_0$ oproti $H_A: \pi > \pi_0$,

$$\text{potom } \alpha' = \sum_{i=n_1}^n \binom{n}{i} \pi^i (1-\pi)^{n-i},$$

c) pre $H_0: \pi = \pi_0$ oproti $H_A: \pi \neq \pi_0$ je p-hodnota α' dvojnásobkom p-hodnoty z jednostrannej hypotézy po a) alebo b).

2. Neparametrické testy

Použitie testov o hodnotách parametrov známych pravdepodobnostných modelov, tzv. parametrických testov, je viazané na splnenie určitých predpokladov. Používajú sa, keď sa skúmajú znaky, ktorých rozdelenie sa dá aspoň približne popísať niektorým z pravdepodobnostných modelov. Ak neplatí niektorý z predpokladov a použijeme parametrické techniky, znižujeme preukaznosť záverov. Pri práci s rozdelením neznámeho typu sa používajú neparametrické testy a potom pravdepodobnostné závery testov nezávisia od tvaru rozdelenia náhodných veličín. Obvykle stačí splnenie predpokladu o spojitosti

rozdelenia. Nevýhodou neparametrických testov je, že majú menšiu silu testu, ako podobný parametrický test.

Jednovýberové testy

Okrem testovania, či rozdelenie pravdepodobnosti sledovaného základného súboru je v zhode s nejakým teoretickým rozdelením, nás často zaujímajú testy súvisiace s úrovňou hodnoty sledovaného znaku, pričom o rozdelení sledovaného znaku máme minimálne informácie. Predpokladanú úroveň hodnôt ordinálneho alebo číselného znaku môžeme overiť znamienkovým alebo jednovýberovým Wilcoxonovým testom. V oboch testoch v nulovej hypotéze predpokladáme, že medián $\tilde{x}$ je rovný číslu x_0 .

Znamienkový test

Predpokladáme, že máme výber zo spojitého rozdelenia s mediánom x_0 (Anděl, 2005, s. 231 - 233) a overujeme $H_0: \tilde{x} = x_0$ proti alternatívnej hypotéze $H_A: \tilde{x} \neq x_0$. Označme Y počet hodnôt z výberového súboru väčších ako medián, potom Y má pri platnosti nulovej hypotézy binomické rozdelenie $Bi(n; 0,5)$ a nulovú hypotézu zamietame, ak je Y blízko nuly alebo n , teda ak $Y \leq k_1$ alebo $Y \geq k_2$, kde k_1 je najväčšie a k_2 je najmenšie z čísel, pre ktoré platí $P(Y \leq k_1) \leq \frac{\alpha}{2}$ a $P(Y \geq k_2) \leq \frac{\alpha}{2}$. Pre výber s rozsahom nad 20 môžeme použiť

aproximačné kritérium $U = \frac{2Y - n}{\sqrt{n}}$, kde U má asymptoticky normálne rozdelenie $N[0,1]$. Ak rozdelenie sledovaného znaku je výrazne zošikmené, test má malú silu a preto je potrebný väčší rozsah výberu.

Jednovýberový Wilcoxonov test

Test je silnejší ako znamienkový test (Anděl, 2005, s. 233-235). Predpokladajme, že $X_1, X_2, \dots, X_n$ je náhodný výber zo spojitého rozdelenia s hustotou symetrickou okolo mediánu x_0 . Položíme $Y_i = X_i - x_0$, pozorovania x_i rovné x_0 vynechávame. Zoradíme Y_i v absolútnej hodnote vzostupne a označíme R_i poradie Y_i ; $S^+ = \sum_{Y_i \geq 0} R_i$ a $S^- = \sum_{Y_i < 0} R_i$. Ak je menšie z čísel

$S^+; S^-$ menšie alebo rovné tabelovanej hodnote (Anděl, 1978, s. 334) zamietame nulovú hypotézu. Platí, že S^+ má asymptoticky normálne rozdelenie $N[0,1]$ a test sa dá založiť na

štatistike $U = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}$. Jedným z predpokladov tohto testu je symetria hustoty

okolo mediánu, k zamietnutiu hypotézy môže dôjsť aj vtedy, keď platí nulová hypotéza, ale hustota je výrazne nesymetrická.

Dvojvýberové testy

Pre porovnávanie súborov alebo premenných (znakov) z hľadiska úrovne hodnôt sú skupiny hodnôt prislúchajúce premenným často označované ako súbory a rozlišujú sa tzv. závislé a nezávislé súbory. Pri nezávislých výberoch porovnávame určitý znak v podskupinách výberového súboru alebo v porovnateľných výberových súboroch. Pri porovnávaní premenných obsahujúcich rovnaké kategórie alebo údaje zistené v rôznych časových obdobiach hovoríme o závislých výberoch. Pre dve premenné ide o tzv. párové testy. Pri každej z n štatistických jednotiek sa zisťujú hodnoty dvoch spolu súvisiacich veličín. Máme teda k dispozícii pozorovania dvojice náhodných veličín $(Y_1, Z_1), (Y_2, Z_2), \dots, (Y_n, Z_n)$.

Veličiny Y_i a Z_i nie sú vzájomne nezávislé, pretože sú zisťované na tom istom i-tom objekte. Vytvoríme diferencie párových hodnôt $X_1=Y_1-Z_1, X_2=Y_2-Z_2, \dots, X_n=Y_n-Z_n$.

Znamienkový párový test

Pri tomto teste overujeme nulovú hypotézu, že medián diferencií je rovný nule oproti obojstrannej alternatívnej hypotéze. Počet nenulových diferencií (kladných alebo záporných hodnôt x_i), ktorý je menší, označíme ako $n_{\min}$ a počet všetkých nenulových diferencií (počet nenulových x_i) ako n_d . Pre dostatočne veľký rozsah výberového súboru, ak $n_d > 25$, má pri

platnosti nulovej hypotézy testovacia štatistika
$$U = \frac{n_{\min} - \frac{n_d}{2} + 0,5}{\frac{\sqrt{n_d}}{2}}$$
 asymptoticky normálne

rozdelenie $N[0,1]$.

Wilcoxonov párový test

Je rozšírením znamienkového testu. Pri výpočte diferencií sa každej z nich priradí poradie na základe jej absolútnej hodnoty. Všetkým diferenciám s rovnakou absolútnou hodnotou sa priradí priemerné poradie, potom súčty poradií zvlášť pre kladné (S^+) a pre záporné diferencie (S^-) a menší zo získaných súčtov označíme $S_{\min}$. Nech n_d je celkový počet nenulových

diferencií, potom testovacie kritérium
$$U = \frac{S_{\min} - \frac{1}{4}n_d(n_d + 1)}{\sqrt{\frac{1}{24}n_d(n_d + 1)(2n_d + 1)}}$$
 má pri platnosti hypotézy

o nulovosti mediánu diferencií asymptoticky normálne rozdelenie $N[0,1]$.

Wilcoxonov – Mannov – Whitneyov test

Pri porovnávaní znaku v podskupinách výberového súboru alebo v porovnateľných výberových súboroch sa využíva Wilcoxonov – Mannov – Whitneyov test (Anděl, 2007, s. 100 – 104). Overujeme hypotézu $H_0: F(\mathbf{x}) = G(\mathbf{x})$ oproti $H_A: F(\mathbf{x}) \neq G(\mathbf{x})$. Všetkých $m + n$ hodnôt usporiadame podľa veľkosti, každej hodnote priradíme jej poradie, T_1 bude súčet poradií $X_1, X_2, \dots, X_m$ v prvom súbore a T_2 súčet poradií $Y_1, Y_2, \dots, Y_n$ v druhom súbore. Ak m a n sú dostatočne veľké, teda $\min(m, n) > 20$ a $\max(m, n) > 30$, potom testovacia štatistika

$$U = \frac{mn + \frac{1}{2}m(m+1) - T_1 - \frac{1}{2}mn}{\sqrt{\frac{1}{12}mn(m+n+1)}}$$
 má pri platnosti nulovej hypotézy približne normálne

rozdelenie $N[0,1]$.

Viacvýberové testy

Aj pri porovnávaní úrovne znaku vo viac ako dvoch súboroch je potrebné rozlišovať, či výbery jednotiek sú závislými alebo nezávislými výbermi zo základných súborov. Pri porovnávaní úrovne hodnôt viacerých premenných (tzv. závislé výbery) je používaný Friedmanov test.

Friedmanov test

Testom overíme zhodu mediánov pre k ($k > 2$) premenných s rovnakými kategóriami (napr. odpovede respondentov na otázky s rovnakou bodovou stupnicou hodnotenia), teda pre k závislých výberov s rovnakým rozsahom n . Je neparametrickou obmenou analýzy rozptylu

pre poradové dáta (Reiterová, E. 2008, s. 26 – 27). Nulová hypotéza predpokladá, že všetky výbery sú z jedného základného súboru, distribučná funkcia je rovnaká. Kódom odpovedí získaným od respondenta priradíme poradie k hodnôt zistených pre k premenných. Predpokladáme, že ak nie je rozdiel medzi výbermi a platí nulová hypotéza, budú poradia pre každú jednotku a aj poradia pre každú premennú rozdelené náhodne. Označme S_j súčet poradových čísel pre j -tu premennú, a t_i je počet premenných, na ktoré odpovedal i -ty

respondent rovnako, štatistika $F = \frac{\frac{12}{nk \cdot (k+1)} \sum_{j=1}^k S_j^2 - 3n(k+1)}{1 - \frac{\sum_{i=1}^n (t_i^3 - t_i)}{nk \cdot (k^2 + 1)}}$ má pri platnosti nulovej

hypotézy približne χ^2 rozdelenie s $k-1$ stupňami voľnosti.

Kruskalov – Wallisov test

Zovšeobecnením dvojvýberového Wilcoxonovho testu je Kruskalov – Wallisov test a používa sa najmä vtedy, ak ide o výbery z rozdelení veľmi sa odlišujúcimi od normálneho rozdelenia. (Anděl, 2007, s. 123 – 128). Predpokladáme, že máme K výberov so spojitými distribučnými funkciami, ktoré sú navzájom nezávislé. Spolu tvoria združený výber s rozsahom n . Sledovaná premenná môže mať veľký počet kategórií. Budeme testovať hypotézu $H_0: F_1(x) = F_2(x) = \dots = F_K(x)$ pre všetky x oproti alternatívnej hypotéze, že H_0 neplatí. Ak nezamietame nulovú hypotézu, potom nezmietame vlastne hypotézu, že vo všetkých súboroch je rovnaký medián premennej. Združený náhodný výber sa usporiada do neklesajúcej postupnosti a určí sa poradie každej veličiny. Pre každý výber s rozsahom n_i , potom vypočítame priemerné poradie $\bar{R}_i$, kde $i=1, 2, \dots, K$. Ako testová štatistika sa použije

$$KW^* = \frac{\frac{12}{n \cdot (n+1)} \sum_{i=1}^K n_i \bar{R}_i^2 - 3(n+1)}{1 - \frac{\sum_{i=1}^{n_s} (t_i^3 - t_i)}{n^3 - n}}, \text{ kde } t_i \text{ sú počty zhodných poradí (pre určité poradie)}$$

v rôznych výberoch a n_s je počet variantov poradí, ktoré sa vo výberoch vyskytujú. Štatistika má pri platnosti nulovej hypotézy približne χ^2 rozdelenie s $K-1$ stupňami voľnosti.

3. Záver

Nezanedbateľnou súčasťou každého výskumu je prípravná fáza prieskumu, ktorá umožní pripraviť relevantné údaje pre ich ďalšie spracovanie a interpretovanie a dosiahnutie tak cieľa výskumu. Nasledujúca fáza spracovania údajov po ich očistení, exploratórnej a deskriptívnej analýze pokračuje overovaním reprezentatívnosti výberu a testovaním štatistických hypotéz. Pri výbere metód je potrebné identifikovať či je známe rozdelenie testovaného znaku a použiť tak parametrické alebo neparametrické testy. Druhá skupina testov má síce menšiu silu, ale nemá stanovené požiadavky na typ rozdelenia testovaného znaku. Vyžaduje citlivé rozlíšenie, či výbery znaku sú závislé alebo nezávislé. Procedúry neparametrických testov bývajú súčasťou štatistických softvérov. Je na vedomostnej a odbornej vyspelosti a erudovanosti užívateľa a jeho praktických skúsenostiach, aby na overenie stanoveného predpokladu vybral vhodný test a správne interpretoval jeho výsledok.

Tento príspevok vznikol s podporou grantov VEGA 1/0935/13 a VEGA 1/0647/14.

4. Literatúra

- [1]ANDĚL, J. 1978. Matematická statistika. Praha: SNTL, 1978. 346 s.
- [2]ANDĚL, J. 2005. Základy matematické statistiky. Praha: Matfyzpress 2005. 358 s. ISBN 80-86732-40-1.
- [3]ANDĚL, J. 2007. Statistické metody. Praha: Matfyzpress 2007.299 s. ISBN 80-7378-003-8.
- [4]LUHA, J. 2003. Matematickoštatistické aspekty spracovania dotazníkových výskumov . In: Štatistické metódy vo vedecko-výskumnej práci ' 2003. Bratislava: Slovenská štatistická a demografická spoločnosť, s. 3 – 32. ISBN 80-88946-32-8.
- [5]LUHA, J. 2006. Overovanie reprezentatívnosti výberového súboru. In: FORUM STATISTICUM SLOVACUM, 2006, č. 5. Bratislava: Slovenská štatistická a demografická spoločnosť, s. 102 – 107. ISSN 1336-7420.
- [6]LUHA, J. 2007. Dva pohľady na reprezentatívnosť. In: FORUM STATISTICUM SLOVACUM, 2007, č. 4. Bratislava: Slovenská štatistická a demografická spoločnosť, s. 94 – 97. ISSN 1336-7420.
- [7]LUHA, J. 2008. Prvotná štatistická analýza kvalitatívnych dát. In: FORUM STATISTICUM SLOVACUM, 2008, č. 2. Bratislava: Slovenská štatistická a demografická spoločnosť, s. 2 – 21. ISSN 1336-7420.
- [8]PEČÁKOVÁ, I. 2008. Statistika v terénnych průzkumech. Praha: Professional Publishing, 2008. 231 s. ISBN 978-80-86946-74-0.
- [9]REITEROVÁ, E. 2008. Základy psychmetrie. Olomouc: Univerzita Palackého v Olomouci, 2008. 117 s. ISBN 978-80-24420-65-3
- [10] ŘEZANKOVÁ, H. 2007. Analýza dat z dotazníkových šetření. Praha: Professional Publishing, 2007. 212 s. ISBN 978-80-86946-49-8

Adresa autorov:

Alena Kaščáková, Ing. PhD.
KKMaIS, EF UMB
Tajovského 10
975 90 Banská Bystrica
alena.kascakova@umb.sk

Gabriela Nedelová, RNDr. PhD.
KKMaIS, EF UMB
Tajovského 10
975 90 Banská Bystrica
gabriela.nedelova@umb.sk