Vasja Vehovar: ANALIZA KONTINGENČNIH TABEL NA OSNOVI RAZLIK STRUKTURNIH DELEŽEV

ANALYZING CONTINGENCY TABLES WITH PERCENTAGE DIFFERENCE: Percentages (or proportions) are often considered as low form of statistics when analyzing contingency tables. But percentages and the differences between them have some properties which are not so obvious and which enable some interesting techniques for further analysis. In this article the approach based on percentage difference is exposed, illustrated and evaluated. Special attention is paid to standardization procedure with reference to work of Rosenberg and Davis.

I.UVOD

Metode za analizo nominalnih spremenljivk postajajo matematično vse bolj zahtevne. Pristop, ki izhaja iz enostavnih strukturnih deležev, je zato še vedno aktualen. Giblje se v okviru, ki je neposredno vezan na podatke in je zato razumljiv tudi laičnemu bralcu. Kljub priljudnosti pa skrivajo strukturni deleži in njihove razlike v sebi precej več informacij, kot je videti na prvi pogled.

S pregledovanjem strukturnih deležev analizo kontingenčnih tabel praviloma začenjamo in k njim se vračamo tudi pri interpretaciji. Poglobljena analiza strukturnih deležev je zato precej naravna pot za nadaljnji študij. Pristop ni nov, opis strukturnih deležev in njihovih razlik najdemo v večini učbenikov statistike (npr. Blejec, 1976), v pričujočem prikazu je le sistematiziran in dopolnjen z lastnostmi, ki jih ima razlika strukturnih deležev kot poseben primer mere asociacije (Reynolds, 1977). Nekoliko bolj specifično je le preverjanje vzročnih modelov (Rosenberg 1962, Davis 1976).

V nadaljevanju bodo torej prikazane elementarne lastnosti razlik strukturnih deležev in osnovne ideje postopkov, ki na njih temeljijo. Namen prispevka je predvsem v sistematični predstavitvi metode in deloma tudi v njenem ovrednotenju.

Analiza razlik strukturnih deležev ne potrebuje zahtevnejšega matematično-statističnega instrumentarija, čeprav se zamudnim - vendar povsem enostavnim - preračunavanjem ni mogoče izogniti. V pomoč je več programov kot npr. CHIPENDALE (Davis, 1985), ki je služil tudi izračunom v pričujočem prikazu.

II. DELEŽI IN RAZLIKE DELEŽEV[1]

Pri analizi kontingenčnih tabel zaradi nazornosti do uporabnikov pogosto uporabljamo strukturne odstotke. V angleščini se je celo ustalil izraz "percentage table". Gre za tabele relativnih frekvenc, ki jih računamo po vrsticah dvorazsežnih tabel. Če imamo vzorčne podatke, so to ocene pogojnih verjetnosti odvisne spremenljivke pri pogoju, da neodvisna spremenljivka zavzema določene vrednosti. V nadaljevanju ne bomo najbolj dosledni pri ločevanju strukturnih deležev in strukturnih odstotkov[2]; imenovali jih bomo deleži. Enako nedosledni bomo tudi pri verjetnostih, ki jih bomo občasno interpretirali z odstotki.

1. Primer. Oglejmo si najprej enostaven primer, ki bo služil (izključno) za ilustracijo tudi v nadaljevanju. Gre za anketne podatke o branju[3], kjer so anketiranke na vprašanje, ali so v zadnjem času prebrale kakšno knjigo, odgovarjale z DA in NE. Poglejmo povezavo s starostjo:

TABELA 1A: Podatki o starosti in branju

	BRANJE		
STAROST	ne	da	skupaj
pod 45 let	506	477	983
45 let in več	550	317	867
skupaj	1056	794	1850

Izračunajmo deleže po vrsticah. S tem smo:

- prevedli podatke na interval od 0 do 100,
- odstranili vpliv velikosti posameznih starostnih skupin,
- pripravili podatke za vzročno analizo.

TABELA 1B: Strukturni deleži za branje po starostnih skupinah

	BRANJE		
STAROST	ne	da	skupaj
pod 45 let	51.5%	48.5%	100%
45 let in več	63.4%	36.6%	100%
skupaj	57.1%	42.9%	100%

1 Osnovne opredelitve pojmov analize nominalnih spremenljivk temeljijo na Momirovic, 1988. Pri slovenski terminologiji so upoštevana nekatera priporočila Komisije za statistično terminologijo Slovenskega statističnega društva.
2 Razlikujejo se za faktor 100.
3 Ker v konkretnih raziskavah (npr. Slovensko javno mnenje) ni bilo primera, ki bi povsem ustrezal za zgled v pričujočem članku, so bili posebej za ta prikaz izbrani in obdelani podatki iz neobjavljene študije J.Hajda, ki se nanašajo na vzorec žensk v Baltimoru.

Iz tabele 1B je razvidno, da mlajše ženske berejo več; bere jih 48.5%, med starejšimi jih bere le 36.6%. Razlika deležev je 12.0% (48.52%-36.56%). Interpretacija je enostavna: med mlajšimi je za 12.0% več žensk, ki berejo, kot med starejšimi.

2. Razlika deležev v tabelah 2x2. Razlika deležev je mera[4] asociacije za tabele 2x2. Če označimo spremenljivki z X, Y, njuni kategoriji z 0,1 in posamezne frekvence v tabeli z

	Y 0	1
X		
1	a	b
0	c	d ,

potem definiramo razliko deležev D_{yx} kot[5]

$$D_{yx} = b/(a+b) - d/(c+d) = (bc-ad)/(bc+ad+ac+bd).$$

a) Z izborom D_{yx} je vzročni red določen - predpostavlja vpliv X na Y (kategorije neodvisne spremenljivke X imamo vedno v stolpcu). Seveda je vzročni red prepuščen subjektivni presoji, kar je za razlike deležev pomembno, saj so asimetrične:

$$D_{xy} = b/(b+d)-a/(a+c) = (bc-ad)/(ab+ad+cb+cd) \neq D_{yx}.$$

Primer: D_{xy} dobimo z odštevanjem deležev po stolpcih:

$$D_{xy} = (477/794 - 506/1056)*100 = 12.2\% \neq D_{yx} = 12.0\%.$$

b) Razlike deležev so podobne Yulovemu Q, ki je ena starejših (Yule, 1912) mer asociacije; očitno je, da so zaradi dodatnih členov v imenovalcu razlike deležev manjše, kvečjemu enake Q, vendar enakega predznaka:

$$Q = (bc-ad)/(ad+bc).$$

c) Za razlike deležev velja, da so neobčutljive za množenje posameznih skupnih[6] frekvenc s konstanto, kar omogoča hipotetično variiranje velikosti vzorcev.

---

4 Mera asociacije izraža stopnjo povezanosti dveh nominalnih spremenljivk.
5 Če zamenjamo vrstni red kategorij pri eni spremenljivki, se razliki deležev spremeni predznak.
6 Skupne frekvence izražajo število enot v posameznih kategorijah določene nominalne spremenljivke.

d) Goodman in Kruskal (1954) sta pokazala, da ima Q pomen tudi v smislu redukcije napake napovedi. Podobno velja za D_{xy} (Goodman, 1972). Ker so bc, ad, bd in ac pari[7], ki se v kategorijah spremenljivke X razlikujejo, bc pari, pri katerih so kategorije obeh spremenljivk urejene enako (konsistentni pari) in ad pari, kjer so kategorije obeh spremenljivk urejene različno (nekonsistentni pari)[8], se izkaže (Davis,1978:145), da je razlika Dyx odstotek, za katerega informacija o urejenosti para na spremenljivki X izboljša slučajnostno napoved o urejenosti para na spremenljivki Y, če predpostavljamo enako urejenost kategorij pri X in Y. Paru $(0,Y_1),(1,Y_2)$ torej prirejamo (0,0),(1,1), paru $(1,Y_1),(0,Y_2)$ pa (1,1),(0,0).

Primer: Denimo, da slučajnostno izbiramo mlajšo in starejšo žensko in za tako dobljene pare napovedujemo odgovor glede na branje. Če bi vedno za mlajšo trdili, da bere, za starejšo pa obratno, bi imeli za 12% boljši uspeh, kot v primeru, ko bi vsaki pripisali enako (50%) verjetnost za branje.

e) V primeru vzorčnih podatkov je treba populacijske vrednosti razlike deležev oceniti. Pri enostavnem slučajnem vzorčenju in dovolj velikih vzorcih so vzorčni deleži p (velikost vzorca N) nepristranske cenilke populacijskih deležev p' (velikost populacije n)[9], pri čemer se cenilka porazdeljuje normalno. Ker je varianca binomsko porazdeljene spremenljivke X (z verjetnostjo p') enaka p'*(1-p'), je varianca cenilke p enaka

$$\text{var } p = \text{var}X*(n-N)/(n-1) \doteq p(1-p)*(n-N)/(n-1).$$

D_{yx}' lahko zapišemo tudi kot razliko pogojnih verjetnosti

$$D_{yx}' = P_{y=1/x=1}' - P_{y=1/x=0}',$$

zato je tudi vzorčna razlika deležev D_{yx} nepristranska cenilka populacijske D_{yx}'; porazdeljuje se normalno in v primeru neodvisnih pogojnih verjetnosti je njena varianca enaka vsoti obeh pogojnih varianc. Za razlike deležev torej na enostaven način oblikujemo intervale zaupanja in preizkuse značilnosti. Podobno velja tudi za njihove izpeljanke: pogojne in parcialne razlike in razlike višjega reda[10].

Primer: Izračunajmo za naš primer (D_{yx}=0.12):

- $\text{var}(D_{yx}) \doteq 0.515*0.485/983 + 0.634*0.366/867 = 5.2*10^{-4}$,
- standardna napaka ocene $SE(D_{yx}) = 0.023$,
- 95% interval zaupanja za D_{yx}' pa je (0.075, 0.165).

---

7 Iz TABELE 1 je razvidno, da se med pari oblike (Y,X) na spremenljivki X razlikujejo naslednji pari: (1,1)-(0,0), (1,0)-(0,1), (1,1)-(1,0), (0,1)-(0,0).

8 Konsistenten je par (1,1)-(0,0), nekonsistenten pa par (1,0)-(0,1).

9 Populacijskim vrednostim bomo v nadaljevanju pripisali znak "'".

10 Za zanesljivejše ocene je priporočljivo imeti v celicah vsaj 20 elementov (Reynolds, 1977:20).

f) Imejmo regresijo dihotomnih spremenljivk Y in X (neodvisna spremenljivka). Označimo s $p_{y=1/x}'$ pogojno verjetnost za Y=1 pri različnih vrednostih X. Sledi:

$$\triangle p_{y=1/x}' / \triangle X = (p_{y=1/x=1}' - p_{y=1/x=0}')/(1-0) = D_{yx}'.$$

D_{yx} je torej ocena regresijskega koeficienta D_{yx}' v enačbi:

$$p_{y=1/x}' = p_{y=1/x=0}' + D_{yx}' * X.$$

Primer: V našem primeru je ocena za $p_{y=1/x} = 0.366 + 0.120*X$.
Za vrednostih X= 0,1 (starost 45 in več, pod 45) pa dobimo oceni za $p_{y=1/x}$ iz TABELE 1B: 0.366, 0.485.

g) Opozoriti velja na trivialno lastnost zgornje enačbe, ki le ni povsem očitna: če vstavimo za X njegove strukturne deleže, dobimo na levi strukturne deleže spremenljivke Y.

Primer: $p_{y=1/x=(867/1850)} = 0.366+0.120*(867/1850) = 794/1850$.

h) Razlike deležev so tesno povezane tudi z ostalimi merami asociacije v tabeli 2x2, npr.:

$$D_{yx} = D_{xy} * varY/varX, \quad r^2 = D_{yx} * D_{xy}, \quad \chi^2 = D_{yx} * D_{xy}.$$

3. Razlika deležev kot mera asociacije v tabelah 2x2. Preglejmo za D_{yx} kriterije za mere asociacije (Momirović,1988: 65):

- spremenljivki sta neodvisni natanko takrat, kadar je D=0;
- kadar je D=1, sta spremenljivki popolnoma odvisni;
- D meri stopnjo povezanosti s podobno učinkovitostjo kot Q;
- obstaja enostavna porazdelitvena funkcija cenilke D;
- nadaljnji postopki in izpeljevanja so razmeroma enostavni.

Toda: Bolj ko sta spremenljivki porazdeljeni asimetrično, bolj se D_{yx} in D_{xy} razlikujeta - subjektivnost pri opredeljevanju vzročnosti ima vse večjo težo. Poleg tega je zgornja meja 1 dosežena le v redkih primerih, ko sta dve celici enaki. Brez dodatnega računanja tako ne vemo, kakšna je zgornja meja za določeno tabelo. Primerjanje tabel je zato bistveno oteženo.

Primer. Tudi če bi v našem primeru - ob danih skupnih frekvencah - vse ženske pod 45 let brale (odgovor DA), bi imeli D_{yx}=0.808 (80.8%) in ne 1.00 (100%).

4. Spremenljivke z več kategorijami. Razlike deležev je mogoče (v smislu opredelitve v II.2) računati le za tabele 2x2. Pri večjem številu kategorij smo prisiljeni spremenljivke dihotomizirati. Temu se lahko izognemo, če razlike deležev računamo za vsako kategorijo posebej:

--- Pri večjem številu kategorij odvisne spremenljivke (npr. večje število kategorij spremenljivke BRANJE) računamo razlike deležev za vse kategorije in D_{yx} interpretiramo kot vpliv spremenljivke X na določeno kategorijo spremenljivke Y.

--- Kadar ima več kategorij neodvisna spremenljivka (npr. več starostnih kategorij), računamo razlike deležev glede na izbrano bazno kategorijo neodvisne spremenljivke. S tem je problem dihotomizacije sicer rešen, namesto tega pa imamo večje število razlik deležev, ki analizi manjšajo preglednost.

5. Posplošitev razlike deležev - Somersov D. Videli smo, da je na množici urejenih parov mogoče razlike deležev interpretirati v smislu redukcije napake napovedi. Ker je kategorije nominalnih spremenljivk mogoče urejati le pri dihotomnih spremenljivkah, pomenijo posplošitev pri večjem številu kategorij mere asociacije za ordinalne spremenljivke. Za Q je to Goodman-Kruskalova gama (Goodman, 1972), za razliko deležev pa Somersov Ds_{yx}' (Somers, 1968):

$$Ds_{yx}' = (p_k' - p_n')/(p_k' + p_n' + p_{uY}').$$

Pri tem sta p_k' oziroma p_n' verjetnosti, da iz populacije izberemo konsistenten oziroma nekonsistenten par, p_{uY}' pa verjetnost, da se par ujema samo na spremenljivki Y. Ds_{yx}' je torej razlika pogojnih verjetnosti med pari, ki se ujemajo v vrstnem redu vrednosti obeh spremenljivk in pari, ki se ne ujemajo, pri pogoju, da se vrednosti za X v okviru vsakega para razlikujejo. V primeru 2x2 imamo $Ds_{yx}'=D_{yx}'$:

$$D_{yx}'=(p_{y=1,x=1}'*p_{y=0,x=0}'-p_{y=1,x=0}'*p_{y=0,x=1}')/(p_{x=1}'*p_{x=0}'),$$

kar je po kratkem računu enako $D_{yx}'= p_{y/x=1}'- p_{y/x=0}$.

V najbolj enostavnem primeru tabele 2x2 ocenimo D_{yx}' z

$D_{yx}= b/(a+b) - d/(c+d))$ in $varD_{yx} = ab/(a+b)^2 + cd/(c+d)^2$.

III. TEHNIKE ANALIZE

Razlike deležev omogočajo oblikovanje učinkovitih tehnik za analizo nominalnih spremenljivk.

1. Pogojne razlike. Ko analiziramo vpliv tretje (kontrolne, testne) spremenljivke Z na zvezo X in Y, dobimo pri vsaki njeni kategoriji pogojno tabelo za X in Y, za katero računamo pogojne mere asociacije, v našem primeru pogojno razliko oziroma pogojni D, kar zvezo med X in Y dodatno osvetli.

Primer. Iz TABELE 1A in 1B je razvidna določena povezanost med starostjo in branjem. Uvedimo spremenljivko IZOBRAZBA:

TABELA 2: Branje po starostno-izobrazbenih skupinah

IZOBRAZBA	STAROST	BRANJE ne	da	skupaj	%da	Di
višja&visoka	pod 45	46	163	209	78.0%	
	45 in več	36	104	140	74.3%	D_1=3.7%(4.7%)
srednja	pod 45	327	290	617	47.0%	
	45 in več	179	159	338	47.0%	D_2=0.0%(3.9%)
ostalo (manj	pod 45	133	24	157	15.3%	
kot srednja)	45 in več	335	54	389	13.9%	D_3=1.4%(3.3%)
skupaj		1056	794	1850	42.9%	

Pogojne razlike D_i so majhne in neznačilno različne od nič (v oklepajih so standardne napake ocene - SE), kar kaže, da v posameznih izobrazbenih skupinah ni povezanosti med starostjo in branjem. Zveza v TABELI 1B je torej nastala zaradi različne izobrazbene strukture žensk v različnih starostnih skupinah.

Pri treh spremenljivkah obstaja več različnih povezav. Lasarsfield (1950:135-167) npr. jih je tipiziral v interpretacijo, primer brez učinka, prikrivanje, specifikacijo ter primere nejasnih povezav, kar vse lahko analiziramo s pogojnimi razlikami. Seveda pri ne-eksperimentalnem raziskovanju ne vemo, če smo kontrolirali res vse spremenljivke in zato tudi ne vemo dokončno, če sta spremenljivki v resnici povezani.

2. Parcialne razlike. Kadar so podatki obsežnejši, je pregledovanje pogojnih tabel zamudno in sintetiziranje pogojnih D težavno. V pomoč so parcialne mere asociacije (Reynolds, 1977: 56). V ta namen izračunamo mere asociacije za pogojne tabele, iz njih pa parcialno mero - v našem primeru parcialno razliko oziroma parcialni D.

Ugodna je tehtana aritmetična sredina pogojnih D_i, z utežmi, ki so sorazmerne zanesljivosti ocen za D_i - dobimo homogeno parcialno razliko D_h (Davis, 1976). Ker je zanesljivost ocene D_i obratno sorazmerna z varianco D_i, imamo:

$$D_h=\sum_i (D_i*(1/varD_i)/\sum_i (1/varD_i)).$$

Primer: Pogojne D_i in njihove variance (TABELA 2) vstavimo v gornjo enačbo in dobimo D_h= 1.3% (SE=2.1%), kar potrjuje, da je čisti učinek starosti na branje - potem ko odstranimo (kontroliramo) vpliv izobrazbe - zanemarljiv.

Če kontroliramo več spremenljivk, se število elementov v tehtani vsoti poveča in je enako produktu njihovih dimenzij.

3. Razlike višjega reda. Kadar se povezanost med spremenljivkama s kategorijami tretje spremenljivke močno spreminja, govorimo o interakciji. V takih primerih parcialni D učinke pogojnih razlik nevtralizira. Interakcijo odkrijemo z odštevanjem pogojnih razlik. Dobimo razlike drugega reda - DD, ki so nepristranska cenilka populacijskih DD'.

Primer. Razlika drugega reda med kategorijama "višja in visoka izobrazba" in "ostalo" je enaka (TABELA 2):

DD = 3.7% - 1.4% = 2.3% ,

kar je različno od 0 (SE = 5.75%). Interakcije med starostnima skupinama torej ni; razlike v branju, ki so nastale zaradi različne starosti, se z izobrazbo ne spreminjajo.

Če imamo štiri spremenljivke računamo razlike tretjega reda - DDD. Izkušnje kažejo, da so značilne interakcije štirih (in več) spremenljivk redke, čeprav jim je mogoče na intuitiven način slediti s pomočjo razlik višjih redov - razlike nižjega reda odštevamo naprej. Pri tem moramo seveda vsakič upoštevati dihotomizacijo neodvisne spremenljivke.

4. Vzročni modeli. Pri preverjanju modelov koristimo dejstvo, da nastopajo razlike deležev v regresiji dihotomnih spremenljivk (II.2f) in čiste vzročne učinke ocenimo s parcialnimi razlikami. Ker potrebujemo za to pogojne tabele, smo omejeni na modele brez povratnih zank: povezave morajo biti vzročne (za relacijo vzročnosti so spremenljivke delno urejene). Pri samem računanju čistega učinka X na Y kontroliramo vse učinke, ki vplivajo na povezanost X in Y: predhodne (spremenljivke, ki vplivajo na X), navidezne (spremenljivke, ki vplivajo na X in Y), posredne (spremenljivke, preko katerih vpliva X na Y). Naknadno lahko vstavimo v model še začetne strukturne deleže in popravimo izhodiščno vrednost. Oblikovanje tovrstnih modelov je preprosto, vendar zahteva jasno razumevanje vzročne analize (Davis, 1986; Asher, 1976).

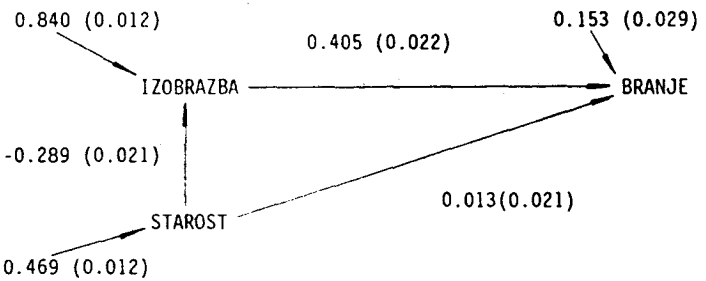
Primer. Teoretično je model za naš primer naslednji (SLIKA 1): starost vpliva na branje neposredno in posredno, prek izobrazbe. Dvojni vpliv starosti na izobrazbo in na branje pa ustvarja pri zvezi izobrazba-branje navidezen učinek. Razliko deležev zato obakrat parcializiramo, to je, kontroliramo posredno oziroma navidezno povezavo. Povezave starost-izobrazba zaradi vzročne urejenosti ni treba parcializirati. Zaradi poenostavitve - pa tudi vsebinske utemeljenosti - sta v SLIKI 1 (ne pa tudi v podatkih TABELE 2) kategoriji "višja in visoka izobrazba" ter "srednja izobrazba" združeni. V splošnem spremenljivk ni treba dihotomizirati, saj lahko izberemo bazno kategorijo; preostale kategorije s tem nastopijo v modelu.

Označimo: STAROST - delež žensk s starostjo "nad 45 let",
IZOBRAZBA - delež žensk s "srednjo, višjo in visoko" izobrazbo,
BRANJE - delež žensk, ki bere - odgovor "DA".

Deleži nad puščicami pomenijo parcialne D_h (D_h = 0.013 za starost in izobrazbo smo izračunali že v III.2), deleže nad puščicami, ki prihajajo od zunaj, pa razberemo iz TABELE 2:

- 46.9% je delež žensk s starostjo 45 in več,
- 84.0% je delež žensk z najmanj srednjo izobrazbo med ženskami s starostjo pod 45,
- 15.3% je delež žensk, ki bere, med ženskami, ki imajo manj kot srednjo izobrazbo in starost pod 45.

SLIKA 1: Vzročni model za starost, izobrazbo in branje



V oklepajih so standardne napake ocene (SE). Razberemo dve močni povezavi starost-izobrazba (-0.289) in izobrazba-branje (0.405). Slednje npr. pomeni: povečanje deleža izobraženih žensk (najmanj srednja šola) za enoto, poveča delež žensk, ki berejo za 0.405 enote.

Grafični predstavitvi je enakovreden sistem linearnih enačb (uporabljamo lastnost iz točke II.2g).

IZOBRAZBA = 0.840 - 0.289 * STAROST
BRANJE = 0.153 + 0.405 * IZOBRAZBA - 0.005 * STAROST + E.

E pomeni konstanto, s katero lahko popravimo začetno vrednost (0.153), če želimo, da bi naš model povsem ustrezal vsem začetnim strukturnim deležem. V našem primeru je E = 0.009.

5. Standardizacija. Še nazornejši je postopek standardizacije. V sociološkem raziskovanju se je pojavil dolgo zatem, ko se je v demografiji že povsem uveljavil. Kljub temu je prisoten že več desetletij (Hagood, 1941), posebej odkar ga je poenostavil Rosenberg (1962). Z uporabo računalnikov je bilo mogoče pristop tudi razširiti (Davis, 1984).

Ideja standardizacije je v odstranitvi predhodnih, navideznih in posrednih vplivov. Sam postopek se reducira na izenačevanje deležev v ustreznih tabelah; razlike deležev so zato enake nič in povezanost izgine. Pomembna je izbira deležev, s katerimi standardiziramo, ponavadi vzamemo deleže celotne populacije.

Primer. Za naš primer določimo najprej vzročni red:

STAROST ---> IZOBRAZBA ---> BRANJE

Radi bi dobili čiste učinke starosti in izobrazbe na branje, zato odstranimo povezavo STAROST -> IZOBRAZBA. Oglejmo si najprej osnovno starostno-izobrazbeno strukturo iz TABELE 2:

TABELA 3A: Podatki o starosti in izobrazbi

	IZOBRAZBA			
STAROST	višja&visoka	srednja	ostalo	skupaj
pod 45	209	617	157	983
nad 45	140	338	389	867
skupaj	349	955	546	1850

TABELA 3B: Strukturni deleži za izobrazbo po starostnih skupinah

	IZOBRAZBA			
STAROST	višja&visoka	srednja	ostalo	skupaj
pod 45	21.3%	62.8%	16.0%	100%
nad 45	16.1%	39.0%	44.9%	100%
skupaj	18.9%	51.6%	29.5%	100%

Razvidno je, da so mlajše ženske bolj izobražene, pogojni D_i so značilno različni od 0. Standardizacija pa priredi ženskam v vseh starostnih skupinah enako izobrazbeno strukturo, to je strukturo iz celotnega vzorca (zadnja vrstica TABELE 3B):

TABELA 3C: Standardizirani strukturni deleži za izobrazbo po starostnih skupinah

	IZOBRAZBA				
STAROST	višja&visoka	srednja	ostalo	skupaj	.
pod 45	18.9%	51.6%	29.5%	100%	(983)
nad 45	18.9%	51.6%	29.5%	100%	(867)
skupaj	18.9%	51.6%	29.5%	100%	(1850)

V TABELI 3A so se posamezna polja seveda spremenila. Z množenjem deležev in skupnih frekvenc iz TABELE 3C (npr. 18.9*867/100=163.6) dobimo nove standardizirane frekvence v TABELI 3D, to je frekvence, kakršne bi bile v primeru neodvisnosti med izobrazbo in starostjo.

TABELA 3D: Standardizirani podatki za starost in izobrazbo

	IZOBRAZBA			
STAROST	visja&visoka	srednja	ostalo	skupaj
pod 45	185.4	507.4	289.1	983.0
nad 45	163.6	447.6	255.9	867.0
skupaj	349.0	955.0	546.0	1850.0

Seveda so se spremenile tudi frekvence za spremenljivko BRANJE. Čeprav so odstotki žensk (TABELA 2), ki berejo, v posameznih starostno-izobrazbenih kategorijah enaki, so spremenjene frekvence v TABELI 3D predrugačile TABELI 1A in 1B Namesto npr. 317 v TABELI 1A dobimo iz TABELE 2 in TABELE 3D (163.8*74.3+447.4*47.0+255.8*13.9)/100 = 367.6 v TABELI 3E.

TABELA 3E:
Standardizirani podatki za starost in branje

STAROST	BRANJE		
	ne	da	skupaj
pod 45	555.5	427.5	983
45 in več	499.4	367.6	867
skupaj	1054.0	795.0	1850

TABELA 3F:
Strukturni deleži za branje iz standardiziranih podatkov

	BRANJE		
	ne	da	skupaj
	56.5	43.5	100%
	57.6	42.4	100%
	57.0	43.0	100%

Zveza starost-branje iz standardiziranih podatkov (TABELA 3F) je D_s = 1.1% (43.5%-42.4%) in je neznačilno različna od nič (SE=2.3%). Zveza starost-branje (TABELA 1B) je bila navidezna. Interpretacija je enostavna: Če bi imele starejše in mlajše ženske enako izobrazbo, bi razlike v branju izginile.

Rosenberg (1962) je pokazal, da je razlika diferenc standardiziranih podatkov enaka enostavni tehtani aritmetični sredini pogojnih razlik D_i, če za utezi izberemo število enot N_i v odgovarjajoči pogojni tabeli:

$$D_s = \sum_i (D_i * N_i / \sum_i N_i),$$

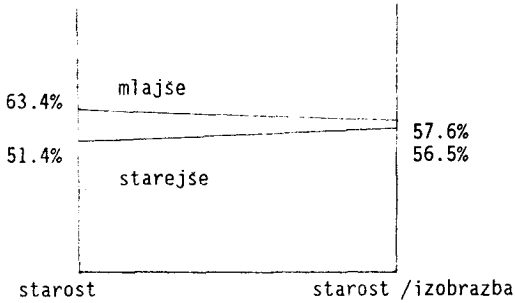
Primer. Izračunajmo D_s za standardizirane podatke še po zgornji formuli (podatki so iz TABELE 2):

D_s = (3.7*349+0.0*955+1.4*546)/1850*100%= 1.1%.

Formalno je standardizacija poseben primer parcialnih razlik. Vsebinsko pa ima dodatno dimenzijo zaradi posebej enostavne interpretacije, ki omogoča, da spremenljivk med analizo ni treba dihotomizirati ali izražati v baznih kategorijah.

Rezultate standardizacije najbolj nazorno ponazorimo z grafom.

Primer. SLIKA 2: Odstotek žensk, ki bere



Na levi strani so odstotki branja izračunani iz osnovnih podatkov, na desni pa je prikazano stanje po standardizaciji, ko je odstranjen vpliv izobrazbe.

Učinke v modelu lahko nazorno razstavimo (dekomponiramo) na iste in vzročne, z odštevanjem od skupnih učinkov pa še na posredne in navidezne, vse izražene z razlikami deležev.

Primer: Razstavimo učinek STAROST ---> BRANJE:

	razlike deležev
A. čisti učinek (TABELA 3F)	1.1%
B. posredni učinek, preko IZOBRAZBE (C-A)	10.9%
C. skupni učinek (TABELA 1B)	12.0%

Podobno dekompozicijo lahko napravimo tudi na osnovi homogenih parcialnih razlik, kot lahko tudi na osnovi standardizacije oblikujemo modele v obliki grafov. Rezultati se v glavnem ujemajo. Razlike izvirajo iz različnih uteži in specifičnosti vzročne analize pri postopku standardizacije. Pomanjkljivost obeh postopkov pa je v odsotnosti koeficienta, ki bi izražal stopnjo prileganja modela k podatkom. Posebej očitno je to v primeru močnih interakcij, kjer zaradi narave parcialnih razlik model ne more biti najbolj ustrezen.

Davis (1984) je postopek razširil na več kot tri spremenljivke. Standardizacija se v tem primeru izvaja suksesivno v smislu vzročnega reda. Vzročna urejenost je zato nadvse pomembna. Spremenljivke moramo zato glede na relacijo vzročnosti urediti linearno in ne le delno. Skupne frekvence spremenljivk, ki v vzročnem redu določeni spremenljivki sledijo, se namreč s standardizacijo spreminjajo; standardiziranje pa ni komutativna operacija. Problem se pri treh spremenljivkah še ne pojavi, ker je za standardizacijo na voljo ena sama spremenljivka.

IV. ZAKLJUČEK

1. Razlika deležev je v osnovi mera asociacije za tabele 2x2. Kljub temu omogoča analizo spremenljivk z večjim številom kategorij. Z izračunavanjem pogojnih in parcialnih razlik ter razlik višjega reda pa je mogoče analizirati tudi več spremenljivk, kar lahko izrazimo v obliki vzročnih modelov. Specifičen primer vzročnih modelov je standardizacija, ki omogoča posebej učinkovito in nazorno analizo.

2. Razlike deležev omogočajo intuitiven, enostaven in deskriptiven pristop, ki zahteva od raziskovalca le najelementarnejše statistično znanje. Po drugi strani uporabnik brez težav razume celotno analizo. Komunikativnost in nazorna interpretacija sta največji odliki opisane metode, čeprav razlikam deležev ni moč odrekati uporabnosti tudi za potrebe same analize.

3. Povzemimo še pomanjkljivosti:

- Dihotomizacija neodvisne spremenljivke je predpogoj za izračun razlike deležev, s čimer so povezane določene težave (izguba informacije, nasilni posegi). V primeru izbora bazne kategorije pa sledi nepreglednost analize oziroma modela. V vsakem primeru - razen (delno) pri standardizaciji - moramo analizo reducirati na tabele 2x2.

- Razlika deležev ima kot mera asociacije nejasno zgornjo mejo, s čimer je primerjava različnih tabel močno otežena.

- Postopkom, ki temeljijo na razliki deležev, je vgrajena predpostavka vzročnosti, ki sili k arbitrarnosti pri opredeljevanju vzrokov, modele pa zožuje na delno oziroma linearno urejene spremenljivke glede na relacijo vzročnosti.

- Pri modelih manjka koeficient, ki bi izražal stopnjo prileganja modela k podatkom.

4. Ker ponujajo razlike deležev po eni strani enostavnost, po drugi strani pa imajo nekaj nedvomnih pomanjkljivosti, jim povsem pristaja naslednja misel[11]: Seek simplicity and distrust it!

11 Išči enostavnost in ji ne zaupaj (motto v Davis, 1971).

REFERENCE

1. Asher H B: Causal Modeling, Quantitative Applications in the Social Sciences, Sage University Papers, 1983.

2. Blejec M: Statistične metode za ekonomiste, Ekonomska Fakulteta, Ljubljana, 1976.

3. Davis J A: Elementary Survey Analysis, 195 str., Prentice-hall, Inc, New Jersey, 1971.

4. Davis J A: Analysing Contingency Tables with Linear Flow Graphs: D-System. In D Heisse (ed), Sociological Methodology, Jossey-Bass, 1976a.

5. Davis J A: Extending Rosemberg Techiques for Standardizing Procentage Tables, Social Forces, Vol 62/3, str.679- 708, March 1984.

6. CHIPENDALE, A System for Sociological Table-building, 1985, by James A. Davis, Published by True BASIC Inc.

7. Davis J A: Logic of the Causal Order, Quantitative Applications in the Social Sciences, Sage University Papers, 1986.

8. Davis J A: Social Differencies in Contemplorary America. Harcourt, Brace Jovanovich, 1987.

9. Goodman L A, Kruskal W: A Measures of association for cross-classification, Journal of the American Statistical Association, 1954, 49, 732-764.

10. Goodman L A, Kruskal W: A Measures of association for cross classification, Journal of the American Statistical Association, 1972, 67, 415-421.

11. Hagood M J: Statistics for Sociologists, New York, Reynal and Hitchock, 1941, chp. 27.

12. Lasarsfield P F, Kendall P L: Problems of survey analysis, Free Press, 1950, str. 135-167.

13. Liebetrau A M: Measures of Association, Quantitative Applications in the Social Sciences, Sage University Papers, 1983.

14. Momirović K: Uvod u analizu nominalnih varijabli, Metodološke sveske 2, RI FSPN, 1988.