4. fejezet

KERESZTTÁBLA-ELEMZÉS

A fejezet céljai

A fejezet célja, hogy az olvasó a fejezet elolvasása után:

- ismerje az egyik legegyszerűbb kétváltozós elemzés, azaz a kereszttábla-elemzés lényegét,
- azonosítani tudja a kereszttábla-elemzéshez szükséges változótípusokat,
- ismerje a kereszttábla elemzés és a Khi-négyzet próba elvégzésének feltételeit,
- tudja előállítani és elemezni a kereszttáblát és a releváns mutatókat,
- képes legyen következtetéseket levonni az elvégzett elemzések alapján.

Az egyváltozós elemzések után a következő fejezetekben a két- és többváltozós elemzési technikákat tekintjük át, amelyek közül elsőként a kereszttábla-elemzéssel foglalkozunk. A kereszttábla-elemzés talán a legegyszerűbb és ezáltal talán a legelterjedtebb módszerek egyike a gyakorlati életben.

4.1. Az elemzés gyakorlati alkalmazhatósága

A kereszttábla-elemzés széles körben elterjedt elemzési módszer, amely két vagy több változó közötti összefüggést vizsgál, illetve ezek kombinált gyakorisági eloszlását mutatja. Az elemzés egyszerűsége és a kapott információk könnyű értelmezhetősége óriási előnyt jelent mind a kutatók, mind a felhasználók számára, ezért az egyik leggyakrabban alkalmazott módszer. Az elemzés elvégzésével arra keresünk választ, hogy két nominális vagy ordinális változó kapcsolatban áll-e egymással. A 4.1. táblázat a módszer gyakorlati alkalmazására kínál példákat.

	Menedzseri probléma	Kutatási probléma	Kutatási kérdés	Hipotézis	Változók	Következtetés
1.	A gyártó mobiltelefon-értékesítése csökkenő tendenciát mutat.	Mobiltelefon-használók (újra) szegmentálása	A mobil telefon használatának gyakorisága (gyakran használók, közepesen gyakran használók, ritkán használók és nem használók) kapcsolatban van-e a nemmel?	A gyakran használók férfiak.	Független változó[1] – nem (férfi, nő) Függő változó: használati gyakoriság (magas, közepes, alacsony)	Megállapítható, melyik nem melyik használati kategóriába tartozik.
2.	Új regionális tv- csatorna indítása	Az egyes régiók tv-nézési szokásai	Van-e kialakult csatornapreferenciája a városi és falusi lakosságnak?	A falusiak kedvenc csatornája az MTV1.	Független változó – lakóhely (város, falu), illetve régió (Kelet-Magyarország stb.) Függő változó: tv- csatornák	Megállapítható, hogy vannak-e regionális eltérések a csatornaválasztásban, illetve preferenciákban.
3.	Hogyan változtassuk meg termékfejlesztési stratégiánkat?	A (potenciális) fogyasztóink által preferált fő termékjellemzők	Van-e összefüggés a design fontossága és a korcsoportok között a termékvásárlásnál?	A design, mint termékjellemző egyáltalán nem játszik szerepet a vásárlásnál az 50 feletti vásárlók körében.	Független változó – korcsoport (18–25, 51–65, stb.) Függő változó: design fontossága (igen, nem)	Megállapítható, melyek azok a termékjellemzők, amelyek az egyes korcsoportoknak a leginkább számítanak.

4.1. táblázat. A menedzseri probléma kialakításától a hipotézis megfogalmazásáig terjedő folyamat bemutatása

A 4.1. táblázat példáinál kereszttábla-elemzéssel állapítható meg, hogy a két változó között van-e kapcsolat. A következőkben előbb a kapcsolat kimutatási módját vizsgáljuk, majd a levonható következtetéseket tekintjük át.

4.2. Általános elméleti áttekintés

A 4.2. táblázat a struktúravizsgáló, -igazoló módszereket tartalmazza, amelyek között láthatjuk a kereszttábla-elemzés helyét. A kereszttábla egy olyan statisztikai technika, amely két vagy több változót ír le egyidejűleg egy olyan táblával, amely megmutatja két vagy több – korlátozott számú kategorizált vagy értéket felvevő változó együttes eloszlását (Malhotra, 2001). A kereszttábla-elemzés segítségével

[1] Az első fejezetben tárgyalt függőségi technikák (lásd 1. fejezet 4. ábra) esetében megkülönböztetünk független és függő változókat, amelyek a kutatási hipotézisünk szerint összefüggnek. Ez azt jelenti, hogy a független változó változást idéz elő a függő változóban, míg a függő változó változik a független változó hatására.

két nominális, ordinális, illetve kategorizált metrikus változó összefüggését elemezzük. Másképp megfogalmazva: a kereszttábla-elemzés nem más, mint két gyakorisági elemzés együttes vizsgálata két nem metrikus változó esetében.

Függő változó	Független változó: Nem metrikus	Független változó: Metrikus
Nem metrikus	Kereszttábla-elemzés	Diszkriminancia-elemzés
Metrikus	Variancia-elemzés	Korreláció, regresszióelemzés

4.2. táblázat. A struktúravizsgáló módszerek egy részének összefoglalása

Tételezzük fel, hogy azt szeretnénk megvizsgálni, hogy egy adott vállalat két márkáját Magyarország mely régiójában vásárolják a leginkább. A kereszttábla-elemzés lényege a fentiekben bemutatottak alapján, hogy két – az elemzés szempontjából releváns – változó kapcsolatát próbáljuk feltárni, amelyre a kutatás elméleti megközelítése, illetve akár egy véletlen megérzés alapján jutottunk. Esetünkben a nullhipotézis (részletesebben lásd az kutatási folyamat fejezetnél) azt fejezi ki, hogy a vállalat márkái és az adott régióban való értékesítése között nincs összefüggés. Ha a vizsgálat során a nullhipotézist elvetjük, az azt jelenti, hogy van összefüggés a két változó között, amennyiben azonban elfogadjuk, akkor nincsen. A továbbiakban bemutatjuk a kereszttábla-elemzés folyamatát az elemzés statisztikai mutatóin keresztül.

4.3. A kereszttábla-elemzés statisztikái

A következőkben rövid említést teszünk az egyes mutatók természetéről és számításáról. A továbbiakban a legfontosabb, azaz a leggyakrabban előforduló statisztikák kerülnek bemutatásra: Khi-négyzet (χ^2), Φ (phi), Kontingencia-együttható (C), Cramer-féle V, lambda, Goodman és Kruskal tau, bizonytalansági együttható (uncertainty coefficient), Kendall-tau b, Kendall-tau c és gamma.

4.3.1. A változók összefüggése, (χ^2)-próba

A kereszttáblával kapcsolatos statisztikák közül talán a leggyakrabban használt a Pearson-féle χ^2- *(Khi-négyzet) statisztika*, amely a két változó összefüggésének statisztikai szignifikanciáját méri. Ezen mutatószám alapján megállapítható,

hogy van-e statisztikai összefüggés a két változó, esetünkben a márka és a régió között. A H_0 nullhipotézis az, hogy nincsen összefüggés.

Az áttekinthetőség kedvéért egy kis elemszámú mintán bemutatjuk a χ^2 értékének számítását. A 4.4. táblázat abszolút értékben mutatja mini adatfelvételünk eredményeit, amely alapján megállapíthatjuk, hogy a 11 válaszadóból összességében 6 nyugat-magyarországi és 5 kelet-magyarországi válaszadó van, és 5 válaszadó az A, 6 pedig a B márkát a vásárolja. A megfigyelt (tényleges) értékek mellett, a Khi-négyzet próba számításához szükség van az elvárt értékekre is. Az elvárt érték a megfigyeléseknek egy olyan eloszlását jelenti, amely esetben nincs összefüggés a két változó között.

Egy kereszttábla számításához minden esetben abszolút számok szükségesek, mint a 4.3. táblázatban is látható. Ha csak relatív (százalékos) értékek állnak rendelkezésünkre, először transzformálnunk kell azokat. Ha az oszlopok és sorok összesen értékei (peremgyakoriságok) adottak, akkor ezek segítségével az egyes cellák elvárt értékei kiszámíthatók. Az egyes cellák elvárt értékeinek számítása a

$$f_e = \frac{n_r\, n_c}{n}$$

képlet alapján történik, ahol n_r sorösszesen, n_c oszlopösszesen, n pedig a teljes mintanagyság. Ha vesszük a Nyugat-Magyarországot és az A márkát, akkor az elvárt értéket a bal felső cellára a következőképp számíthatjuk ki:

$$f_e = \frac{6 * 5}{11},$$

így az összefüggés nélküli állapotban lévő elvárt (elméleti) értékeket a zárójelben találjuk (2,7).

Márka/ Régió	A márka	B márka	Összesen
Nyugat-Magyarország	1 (2,7)	5 (3,3)	6
Kelet-Magyarország	4 (2,3)	1 (2,7)	5
Összesen	5	6	11

4.3. táblázat. A vállalat márkái és az ország régiói közötti összefüggés

A χ^2 értékének számítása a

$$\chi^2 = \sum_{\substack{\text{összes} \\ \text{cella}}} \frac{\left(f_o - f_e\right)^2}{f_e}$$

képlet alapján történik, amely elsődlegesen az elvárt (f_e) és a megfigyelt értékek (f_o) összevetésén alapszik. A Khi-négyzet próba az egyes cellákban lévő megfigyelt eseteknek a számát hasonlítja össze, azzal az – elvárt – esetszámmal, amelyet akkor kapnánk, ha nem lenne kapcsolat a két változó között. Elvégezve a számítást mind a négy cellára, amely például a nyugat-magyarországi válaszadók és az A márkát vásárlók közös cellájánál:

$$\frac{(1-2{,}7)^2}{2{,}7} = 1{,}0704,$$

majd az értékeket összegezve megkapjuk az empirikus, azaz a tapasztalati χ^2 értéket (4,412). Ezt az értéket az elméleti (táblázatokban rendelkezésre álló) értékhez viszonyítva megállapíthatjuk, hogy elvetjük vagy elfogadjuk a nullhipotézist. Az elméleti érték mint küszöbérték minden esetben két tényezőtől függ: az egyik a szabadságfok (df) nagysága (amelynek számítása df = (sor–1) * (oszlop–1)), illetve a szignifikanciaszint, azaz az előbbiekben tárgyalt α, amelyet az ún. χ^2 eloszlástáblákban találhatunk meg.

A χ^2 eloszlás ferde, alakja kizárólag a szabadságfok értékétől függ, és ahogy ennek értéke növekszik, az eloszlás egyre szimmetrikusabbá válik (Malhotra, 2001). Ez főleg arra utal, hogy a nagyobb táblák esetében, ahol akár csak az egyik, akár mindkét változónak több választási lehetősége van, egyre nagyobb szabadságfokot és egyre szimmetrikusabb eloszlást eredményez.

A χ^2 eloszlástáblázat a szabadságfok ((2–1) * (2–1) = 1) és egy adott hibavalószínűség (0,05) függvényében megadja a viszonyítási küszöbértéket (esetünkben ez: 3,841). Az általunk számított tényleges, tapasztalati (4,412) érték nagyobb, mint a küszöbérték, s ez alapján a nullhipotézist elvetjük. Ez azt jelenti, hogy van kapcsolat a két változó között, tehát attól függően, hogy a keleti vagy a nyugati országrészben lakik a válaszadó, más-más valószínűséggel vásárolja a két vizsgált márkát.

A Khi-négyzet statisztika egyik fő jellemzője, hogy érzékeny a mintanagyságra, ugyanis a Khi-négyzet lineárisan függ a minta elemszámától, azaz ugyanolyan eloszlásoknál előfordulhat az a jelenség, hogy két változó alacsony mintaelemszámnál (10 fő) nem mutat szignifikáns eredményt, míg viszonylag magas elemszám (1000 fő) esetén már igen. Ez a márka-régió példánál azt jelenti, hogy kis mintaelemszámnál a megfigyelt és az elvárt érték között nagy különbséget kell találni ahhoz, hogy szignifikáns eredményt kapjunk.

4.3.2. A kapcsolat erőssége

Általában a kapcsolat erőssége akkor érdekes, ha az összefüggésről bebizonyítottuk, hogy statisztikailag szignifikáns. A kapcsolat erősségét eltérő mutatókkal mérjük attól függően, hogy a változók nominális (4.4. táblázat) vagy ordinális skálán (4.6. táblázat) mértek.

4.3.2.1. Nominális skálák

Nominális skáláknál egyrészt a Φ (phi), a kontingencia (C), a Cramer V együtthatók mint szimmetrikus mutatók alkalmazhatók, amelyek kapcsolódnak a Khi-négyzet statisztika értékéhez; másrészt a lambda, Goodman és Kruskal tau és bizonytalansági együttható (uncertainty coefficient) mint aszimmetrikus mutatók alkalmazhatók. Az, hogy egy mutató szimmetrikus, azt jelenti, hogy a független és a függő változók felcserélése nem változtatja meg az eredményt, míg az aszimmetrikus mutatókra ez nem igaz. A kapcsolaterősséget vizsgáló együtthatók értéke általában 0 és 1 között mozog, ahol a nulla (0) a kapcsolat hiányát, míg az egy (1) az erős kapcsolatot jelenti a két változó között.

	Nominális skála	
	Szimmetrikus	*Aszimmetrikus*
2*2-es táblánál:	Φ (phi) együttható: $\phi = \sqrt{\frac{\chi^2}{N}}$	
Bármely táblánál (2*2-esre is):	Kontingencia-együttható (C): $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$ Cramer V: $V = \sqrt{\frac{\chi^2}{N(k-1)}}$	Lambda: $\lambda = \frac{SUM(f_i - F_d)}{N - F_d}$ Goodman és Kruskal tau, Bizonytalansági együttható (Uncertainty coefficient)

4.4. táblázat. A kapcsolat erősségét mérő mutatószámok és ezek képlete nominális skáláknál

A *phi együtthatót* 2*2-es kereszttáblák, azaz 2 sorral és 2 oszloppal rendelkező tábláknál alkalmazzuk. (Ebbe az összesítő sorok és oszlopok nem számítanak bele.) A phi együttható a Khi-négyzetnek a mintanagysággal (N) korrigált (normált) értéke, amely megegyezik a korrelációs együtthatóval 2*2-es táblák esetében. A phi maximális értéke 2*2-es táblák esetében 1; más táblaméretnél nincs felső korlátja, s ezért eltérő (nem 2*2-es) táblaméretnél a mutató értelmezése bonyolult, ezért nem is alkalmazzák.

A phi alkalmazásakor a Yates folytonossági korrekció (Continuity Correction) nem alkalmazható (lásd később). Esetünkben (2*2-es tábla) a phi értéke

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{4{,}412}{11}} = 0{,}633,$$

ami azt jelenti, hogy elég erős kapcsolat van a két változó között.

A kontingencia-együttható (C) bármely, így akár a 2*2-es táblánál is alkalmazható, szintén a mintanagyságot használja a számításánál. A kontingencia-együttható értéke a tábla méretétől függ, maximum 1 lehet, azonban ezt ritkán éri el. A mutató értelmezése nem egyszerű, ezért érdemesebb a Cramer V-t alkalmazni. A mutató értéke:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{4{,}412}{4{,}412 + 11}} = 0{,}535$$

A *Cramer V* bármely kereszttábla esetében alkalmazható, és 2*2-es táblák esetében megegyezik a Φ (phi)-vel. A Cramer V számos kutató szerint a „legmegbízhatóbb" mutató. Az előző kettővel szemben, és ezt érdemes minden esetben megvizsgálni. A Cramer V számításához a mintanagyságra (N) és a kettő közül a kevesebb lehetőséget felkínáló ismert kategóriáinak számára (k) van szükség. (Ha például az egyik változónak 3 kategóriája van (magas, közepes, alacsony), míg a másiknak 4 (különböző korcsoportok), azaz egy 3*4-es táblánk van, akkor k=3. Egy 2*2-es tábla esetében a k=2.)

A Cramer V egy 2*2-es táblára megegyezik a Φ (phi)-vel, ugyanis:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{4{,}412}{11(2-1)}} = \sqrt{\frac{4{,}412}{11}} = 0{,}633$$

(Az SPSS-ben a két mutató egy parancs alatt található meg.)

A *lambda* azt méri, hogy a független változó milyen mértékben képes a függő változót előre jelezni, amit százalékos formában fejez ki. A mutató a hiba csökkenésének mértékét mutatja, azaz egy megfigyelés egy változó szerinti hovatartozását használjuk arra, hogy előre jelezzük egy másik változó szerinti hovatartozását. (Például a nemek szerinti hovatartozás alapján igyekszünk eldönteni valakiről, hogy dohányzik-e.) A lambdával azonos jelentéssel bír a Goodman és Kruskal tau és a bizonytalansági együttható[2] (uncertainty coefficient) mutató. Ezek mindegyike százalékos formában mutatja ki azt, hogy mennyivel csökken a becsléshiba valószínűsége azáltal, hogy a független változót bevontuk az elemzésbe.

A lambda nagyon robusztus mutató, és értéke számos esetben nulla. A lambda számításához a következő elemek szükségesek (lásd 4.4. táblázat): a független változó kategóriáinak legnagyobb értékei, összesítve (SUM f_i) azaz a peremen belül a két legnagyobb érték az 5 és a 4; a függő változó (márka) legnagyobb pe-

[2] A bizonytalansági együttható azon az elven működik, hogy akkor tudunk a legkevesebbet mondani az eloszlásról, azaz akkor a legnagyobb a bizonytalanság, ha homogén az eloszlás.

remeloszlása (F_d), amely esetünkben 6 (B márka), illetve a mintaelemszám (N), amely alapján a lambda értéke:

$$\lambda = \frac{SUM(f_i - F_d)}{N - F_d} = \frac{SUM((5+4)-6)}{11-6} = 0{,}6$$

Ez azt jelenti, hogy a válaszadó lakhelyének régió szerinti ismerete a márkaválasztásra adott előrejelzési hibát 60 százalékkal csökkenti, azaz ez nagyon jó előre jelző (prediktor) változó. A képlet nevezőjéből kiolvasható, hogy a régió kategóriáinak ismeretében 11 próbálkozásból, 6-szor helyesen találnánk el a márkát, míg 5 esetben helytelenül tippelnénk.

4.3.2.2. Ordinális skálák

A *sorrendi (ordinális) skáláknál* az előbbiektől eltérő mutatókat kell alkalmaznunk (4.5. táblázat), hiszen itt a két változó értékeinek sorrendje között keresünk összefüggést. Kiszámításuk során azokat a megfigyeléseknél, amelyeket mindkét változó szerint rangsoroltak, a sorrendi helyezéseket összehasonlítjuk, azt vizsgálva, hogy az egyik változó szerinti helyezés megegyezik-e a másik változó szerinti sorrenddel vagy sem. Ha a két változó szerinti sorrend megegyezik, akkor ezeket konkordáns (előzés), ha ellentétes, akkor diszkordáns (holtverseny) pároknak nevezzük. Ha értékük éppen azonos a két változó szerint, kötött párokról beszélünk. A sorrendi skáláknál a táblák szimmetrikussága dönti el, hogy melyik mutató a legalkalmasabb az adott elemzés során.

	Ordinális skála
Szimmetrikus táblák esetén	Kendall tau-b: $\tau_b = \frac{k-d}{\sqrt{(k+d+v_x)*(k+d+v_y)}}$
Nem szimmetrikus táblák esetén	Kendall tau-c: $\tau_c = \frac{2m*(k-d)}{N^2*(m-1)}$
Bármely táblaméret esetén	Gamma: $\gamma = \frac{k-d}{k+d}$ Somers-féle d

4.5. táblázat. A kapcsolat erősségét mérő mutatószámok ordinális skáláknál

A szimmetrikus tábláknál a Kendall tau-b (τ_b), a nem szimmetrikus tábláknál a Kendall tau-c (τ_c) használatos, míg a gamma (γ) bármilyen táblaméretnél alkalmazható. A *Kendall tau-b* (τ_b) képlete a konkordáns (k), a diszkordáns (d) és a kötött párokat (v_x, v_y) foglalja magában. A mutató értéke –1 és +1 között változhat, ahol +1 azt jelenti, hogy a párok sorrendje hasonló, míg –1-nél a párok sorrendje épp ellentétes. *Kendall tau-c* (τ_c) képletében a k, a konkordáns, d a diszkordáns párokat, m a legkisebb kategóriaszámot és N a mintaelemszámot jelzi. A mutató értéke szintén –1 és +1 közötti értéket vehet fel, és értéke általában hasonló a Kendall tau-b értékéhez.

A *gamma* értéke 0–1 között változhat, ahol 0 azt jelenti, hogy a változók függetlenek egymástól, míg az 1 azt, hogy a változók teljes mértékben függnek egymástól.

A *Somers-féle d* két ordinális változó közötti kapcsolatot méri, értéke –1 és +1 között változhat. A mutató 1-hez közeli értéke abszolút értékben erős kapcsolatot jelent, míg 0 közeli értéke gyenge kapcsolatot vagy a kapcsolat hiányát jelenti.

4.4. A kereszttábla-elemzés lényege és feltételei

A mutatószámok után tekintsük át a *kereszttábla-elemzés feltételeit*. Ezeknek teljesülniük kell ahhoz, hogy az előbb látott mutatók értékeit helyesen tudjuk értelmezni.

- A megfigyeléseknek függetleneknek kell lenniük, azaz egy megfigyelés (megkérdezett) nem szerepelhet egyszerre két vagy több cellában.
- A kereszttábla két vagy több változó közötti összefüggés vizsgálatára alkalmas. Számos esetben kulcskérdés azonban, hogy melyik a független és melyik a függő változó, azaz melyik változó befolyásolja a másikat, ugyanis a két változót ennek megfelelően vizsgáljuk, aminek hatása van a tábla értelmezésére is. A régió és a márkaválasztás példában a kérdés logikai úton egyszerűen eldönthető volt. (A márkapreferenciától aligha függ a lakóhely.) Amikor azonban a viszony nem egyértelmű (például egy vásárló márkahűsége és bolthűsége közötti kapcsolat), valójában csak azt tudjuk megállapítani, hogy a változók összefüggenek-e, milyen erős a kapcsolat közöttük, s melyik változó van nagyobb hatással a másikra.
- A kereszttábla-elemzés nagy előnye, hogy alapjában véve mindenféle skálán mért változó vizsgálatára alkalmas, azonban a leginkább névleges (nominális), sorrendi (ordinális), illetve kategorizált metrikus skálánál használható.
- A skála típusa minden elemzésnél fontos szerepet játszik, ebben az esetben főleg amiatt, mert az elemzés egyik alapfeltétele a cellákban szereplő megfigyelések alulról korlátos száma. A gyakorlatban ez azt jelenti, hogy ha a skálakategóriák száma magas, akkor valószínűleg számos olyan cella

szerepel majd a kereszttáblánkban, amely kis számú megfigyelést tartalmaz. Ezért annak érdekében, hogy metrikus változók is elemezhetők legyenek, kategorizálnunk kell ezeket. Azonban mivel a Khi-négyzet próbánál a cellában található tényleges és ezáltal elvárt értékek száma a kritikus tényezők közé tartozik, a válasz(kategóriák) száma kialakításának megfelelő figyelmet kell szentelni. A Khi-négyzet próba korlátai a következők:

- Az elvárt értéknek minden cellában legalább 1-nek kell lennie (ez az érték a peremgyakoriságoktól függ).
- A cellák maximum húsz százalékában lehet a várható érték kisebb mint 5. Ennek a kritériumnak létezik egy szigorúbb változata is, amely szerint a kereszttábla nem tekinthető megbízhatónak, amíg minden egyes cella elvárt értéke el nem éri az 5-öt. A fenti márka-régió példában ez egyik cellára sem igaz.

Ha az előbbiek nem teljesülnek, két megoldás kínálkozik. A leggyakrabban alkalmazott megoldás a megfigyelések újrakódolása, azaz új kategóriák létrehozása úgy, hogy az egyes cellák megfeleljenek az elvárt kritériumoknak. A másik megoldás a további adatgyűjtés, amely nyomán a peremgyakoriságok és az elvárt értékek növekedhetnek.

- A kereszttábla egyes celláinak számát tekintve, néhány szerző megszabja, hogy a kereszttáblának több mint 5 cellából kell állnia, ugyanis 2*2-es kereszttábla és alacsony esetszám mellett megbízhatósági problémák léphetnek fel (Brosius és Brosius, 1995). Amennyiben 2*2-es kereszttáblánál az egyik cella elvárt értéke 5 alatt van, akkor az SPSS automatikusan elvégzi a Fisher egzakt tesztet, amelyre a későbbiek során részletesen kitérünk.
- Az egyes változóknál a kategóriák száma nem csak az elemzés feltételeire hat vissza (cellagyakoriság), hanem a táblázat áttekinthetőségére is. Ennek eredményeként érdemes átgondolni, melyik változó a függő és melyik a független, ugyanis általános szabályként elfogadható, hogy a független változó szerint számítjuk a százalékokat a függő változóra (Malhotra, 2002). Ez esetünkben azt jelenti, hogy a régió mint független változó szerint vizsgáljuk a márkaválasztás mint függő változó megoszlását.

4.5. Szemléltető példa SPSS-ben

A korábban használ régió-márka példa adatbázisa két változót tartalmaz: a márka és a régió változókat. A márka változó egy adott vállalat két márkáját tartalmazza, A és B márkát (értékcímke), míg a régión belül Kelet- és Nyugat-Magyarországot különböztetünk meg (lásd 4.1. ábra).

adat [DataSet1] - SPSS Data E

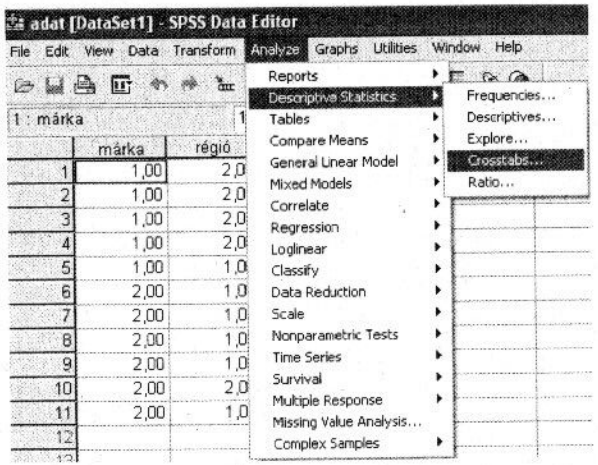
File Edit View Data Transform Ar

1 : márka 1

	márka	régió
1	1,00	2,00
2	1,00	2,00
3	1,00	2,00
4	1,00	2,00
5	1,00	1,00
6	2,00	1,00
7	2,00	1,00
8	2,00	1,00
9	2,00	1,00
10	2,00	2,00
11	2,00	1,00

4.1. ábra. Adatbeviteli nézet

A kereszttábla-elemzés az ANALYZE/DESCRIPTIVE STATISTICS/CROSSTABS útvonalon érhető el 4.2. ábra).



4.2. ábra. Kereszttábla parancssor

Három fontos lépést kell végrehajtani a kereszttábla-elemzésnél:

Az *első*, hogy a változókat át kell vinni az oszlop (COLUMN(S)) és a sor (ROW(S)) változók ablakába (4.3. ábra, bal oldali párbeszédablak). Annak eldöntésére, hogy melyik változó legyen a sor-, melyik az oszlopváltozó, nincsenek meghatározott szabályok. Aki először találkozik a kereszttáblával mindenesetre érdemes mindkét verziót kipróbálni, azaz ugyanazt a változót egyszer sor-, egyszer pedig

oszlopváltozóként szerepeltetni és megvizsgálni a különbséget. A különbség az egyes cellák és az azokban lévő számok helyzetében lesz megfigyelhető. Ugyanakkor a statisztikában és a társadalomtudományokban alapvető szabályként elfogadott az, hogy a sorváltozó a független (X), míg az oszlop a függő változó (Y).

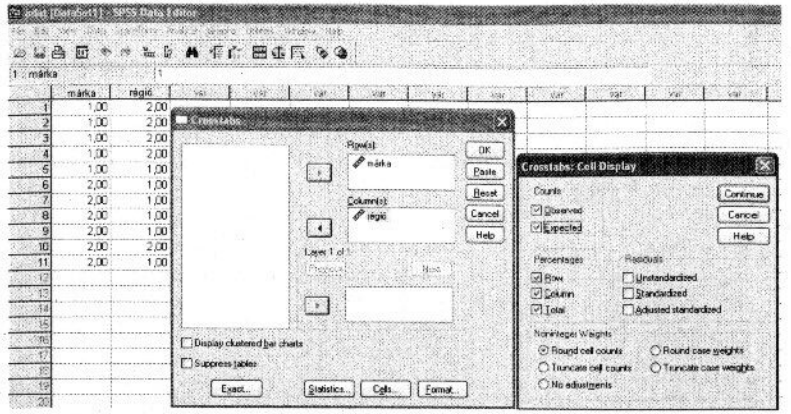
A *második* lépés a cellatartalom meghatározása (lásd 4.3. ábra, jobb oldali párbeszédablak). A CELLS párbeszédpanelen belül a COUNTS – OBSERVED (megfigyelt értékek) menüpont alapbeállításként mindig működik, amely megmutatja a cellák elemszámát abszolút értékben. Ez alatt a COUNTS – EXPECTED (várható értékek) menüpont a cella várható értékét jelenti, független eloszlás esetén. A CELLS parancson belül meg tudjuk jelölni azokat az értékeket, amelyekre szükségünk lehet az elemzés során. A három legfontosabb tényező ebben az esetben a sor (ROW), az oszlop (COLUMN), és a teljes (TOTAL) megoszlás megjelölése, s ezáltal megkapjuk százalékos (relatív) gyakoriságként ezek megoszlását az adott mintán belül, annak függvényében, hogy melyik változónk volt a sor-, illetve melyik volt az oszlopváltozó. Esetünkben a régió a sorváltozónk, azaz a sormegoszlás erre a változóra fogja megadni a válaszlehetőségek (Kelet- és Nyugat-Magyarország) közötti százalékos sormegoszlási adatokat, míg a márkánál (oszlopváltozó) az oszlopmegoszlást, illetve a TOTAL parancs segítségével az teljes mintanagyságra vonatkozó százalékos megoszlást.

A reziduumok (RESIDUALS) közül a *nem standardizált* (UNSTANDARDIZED) reziduum a tényleges és a várható érték különbségét mutatja, míg a *standardizált* (STANDARDIZED) reziduum a nem standardizált reziduumoknak a becsült standard hibájával való korrekciója, ez utóbbi az úgynevezett Pearson-reziduum. Az ADJ. STANDARDIZED (korrigált standardizált reziduum) az előző standardizált mutatónak egy olyan változata, amely szintén a megfigyelt és az elvárt érték különbségét mutatja, a korrekció azonban a standard hibával történik. Ez azt jelenti, hogy a különbség a megfigyelt és az elvárt értékek között szórásegységekben fejeződik ki. Ez azért fontos, mert ez a mutató tartalmát tekintve megmutatja a kereszttáblán belüli szignifikáns relációkat. Ha a mutató +2, illetve e feletti értéket vesz fel, akkor azok szignifikánsan összefüggnek, míg –2, illetve ez alatti értékek bizonyos, hogy nem függnek össze egymással, míg a többi érték esetében nem tudunk semmi bizonyosat az adott relációról. Ennek a szabálynak az alkalmazása azonban óvatosságot és a Khi-négyzet statisztika egyidejű figyelembevételét igényli.

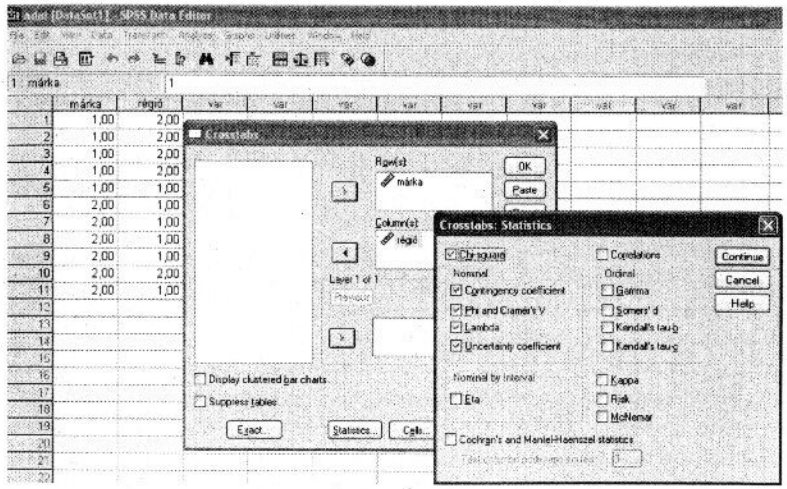
A 4.4. ábrán a STATISTICS menüpont felépítését láthatjuk, ahol a Khi-négyzeten kívül a különböző skáláknál alkalmazható statisztikák is fel vannak tüntetve. Ezek többségét már ismerjük, a még be nem mutatott lehetőségek pedig a következők.

CORRELATIONS (Korrelációs együttható): A korrelációs együtthatóval a Korreláció és regresszió fejezetben részletesen foglalkozunk.

A Kappa két értékelés, illetve értékelő közötti egyetértést mér. Értéke 0 és 1 között változhat. Ha a kappa értéke 0, a két értékelés között nincs összefüggés, azaz nincs egyetértés, ha értéke 1, akkor a kettő között teljes egyetértés van.



4.3. ábra. Cellatartalom meghatározása



4.4. ábra. A Statisztikák menüpont

Általánosan elfogadott, hogy Kappa 0,4 alatti értéke nagyon gyenge egyetértés, 0,4–0,75 között elfogadható egyetértés, míg 0,75 fölött kiváló az egyetértés szintje. Csak szimmetrikus táblákra alkalmazott.

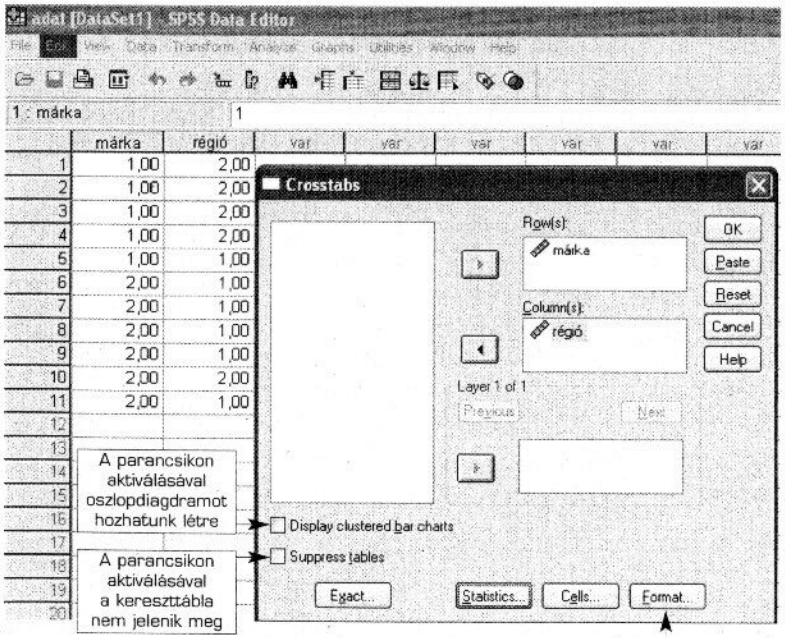
RISK (kockázati hányados). A kockázati hányados 2*2-es tábláknál a két változó kapcsolatát vizsgálja. Az egyik változó egy tényező, például zsíros ételek fogyasztása, a másik egy esemény, szívinfarktus bekövetkezte. A mutató értékének alsó határa 0, felső határa nincs, mindamellett a program nemcsak egy értéket, hanem egy konfidenciaintervallumot is ad, amelyet figyelembe kell venni a kapott értékek értelmezésénél. Amennyiben a mutató értéke 1, illetve azt a konfiden-

ciaintervallum magába foglalja, akkor ez azt jelenti, hogy nincs összefüggés a tényező és az esemény között. Amennyiben a mutató értéke 0, illetve nagyobb, mint 1, akkor valószínűsíthető, hogy összefüggés van a két változó között.

McNemar-teszt. A teszt egy megkérdezett csoporton végzett két mérés közötti változást vizsgál. A McNemar-teszt általában 2*2-es táblákra alkalmazott, ugyanis a mérés dichotóm skálán történik. (Dichotóm skálának nevezzük az olyan skálákat, amelyeknek két kimenetele van, általában 0 és 1). A teszt azt mutatja, hogy hány válaszadó választotta ugyanazt az opciót az első és a második mérés alkalmával egyaránt.

Az Eta mutató egy intervallum (függő) és egy nominál vagy ordinális skálán mért változó (független) közötti kapcsolatot mér. A program két eta-értéket számít, az egyik a sorváltozót tekinti intervallumskálának, a másik az oszlopváltozót. Ezzel a mutatóval a varianciaelemzésnél bővebben foglalkozunk.

Cochran and Mantel-Haenszel statistics (Cochran és Mantel-Haenszel-féle statisztika): a Cochran és Mantel-Haenszel statisztika két dichotóm változó közötti összefüggést vizsgál, egy vagy több kontrollváltozó hatását feltételezve. A statisztika alkalmazásának előnye, hogy egyszerre veszi figyelembe az összes kontrollváltozó hatását.



4.5. ábra. Segédfunkciók bemutatása

Példánkban ezúttal a Statistics menüpontban (4.4. ábra) négy, nominális skálákra vonatkozó mutatószámot jelöltünk meg, amelyek fontosak lehetnek az elsődleges elemzés során: a Khi-négyzet próbát, a kapcsolat erősségét illetően pedig a phi és Cramer V, a kontingencia és a lambda, valamint a bizonytalansági együtthatót. (A kapott eredményeket a 4.6. fejezetben mutatjuk be.)

Végül tekintsünk át néhány segédfunkciót a kereszttáblával kapcsolatban a 4.5. ábrán, amelyek közül hármat kívánunk kiemelni. A „display clustered bar charts" opció segítségével oszlopdiagramot hozhatunk létre a kereszttábla értékei alapján, a „suppress tables" opció bejelölése esetén pedig maga a kereszttábla nem jelenik meg. Végül a format parancson belül kérhetjük a kereszttáblán belüli értékek növekvő vagy csökkenő sorrendben való megjelenését.

4.6. Elemzés és értelmezés

Az output nézetben minden elemzés elején találhatunk egy összegzést (Case Processing Summary) a minta megoszlásáról annak tekintetében, hogy hány valós (Valid), hiányzó (missing) érték – eset, válaszadó, megfigyelés – van és mekkora a teljes mintanagyság (Total). Leegyszerűsített példánkban összesen 11 valós esetünk van, azaz minden válaszadó összes válaszát ismerjük, tehát nincs hiányzó érték, amely alapján mind a 11-et, azaz 100 százalékot értékelni tudjuk (lásd 4.6. táblázat).

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
márka * régió	11	100,0%	0	,0%	11	100,0%

4.6. táblázat. A valós és a hiányzó eseteket összesítő táblázat

A kereszttábla címe mutatja a két változó rövidített nevét, ahol a régió a sor és a márka az oszlopváltozó. A cellák tartalma a következő:

- a cellamegfigyelések száma abszolút értékben (Count),
- a cellák várható értéke (Expected Count),
- a sormegoszlás (% within Régió),
- az oszlopmegoszlás (% within Márka),
- a teljes mintanagyság szerinti megoszlás (% of Total) és
- a standardizált, korrigált reziduum (Adjusted Residual),

amelyekről a 4.4. ábra kapcsán már említést tettünk.

A megfigyelt és a várható értékekre a továbbiakban nem térünk ki, ugyanis abszolút értékek alapján nagyon ritkán elemzünk egy kereszttáblát, sokkal inkább a kereszttábla szempontjából releváns relatív gyakoriságokra koncent-

rálunk. A 4.7. táblázatban a sorváltozó szerinti megoszlás (% within RÉGIÓ) 16,7 százalék, illetve 83,3 százalék a nyugat-magyarországi válaszadóknál, ami azt jelenti, hogy a 16,7 százaléka a nyugat-magyarországi lakosoknak az A márkát vásárolja, míg 83,3 százalékuk a B-t. Ezt az analógiát követve elemezhetjük a kelet-magyarországi lakosokat is márkavásárlás szempontjából, ahol megállapíthatjuk, hogy az ottani lakosok 80 százaléka az A márkát választja. A táblát elemezhetjük még az oszlopváltozó, azaz a márka szempontjából (% within MÁRKA), ahol az A márkát választók 20,0 százaléka nyugat-magyarországi, 80,0 százaléka kelet-magyarországi lakos. Mind a sor-, mind az oszlopmegoszlások összege 100 százalékot eredményez az összesen (Total) oszlopokban. A teljes mintára vonatkozó megoszlások (% of Total) alapján megállapítható, hogy a nyugat-magyarországi lakosok, akik B márkát használnak, azaz 5 fő, a minta 45,5 százalékát teszik ki.

régió * márka Crosstabulation

			márka		
			A márka	B márka	Total
régió	Nyugat Mo	Count	1	5	6
		% within régió	16,7%	83,3%	100,0%
		% within márka	20,0%	83,3%	54,5%
		% of Total	9,1%	45,5%	54,5%
		Adjusted Residual	-2,1	2,1	
	Kelet Mo	Count	4	1	5
		% within régió	80,0%	20,0%	100,0%
		% within márka	80,0%	16,7%	45,5%
		% of Total	36,4%	9,1%	45,5%
		Adjusted Residual	2,1	-2,1	
Total		Count	5	6	11
		% within régió	45,5%	54,5%	100,0%
		% within márka	100,0%	100,0%	100,0%
		% of Total	45,5%	54,5%	100,0%

4.7. táblázat. A régió és a márka kapcsolatát bemutató kereszttábla

A kereszttábla bemutatása után meg kell vizsgálnunk, hogy vajon összefüggnek-e a változók, és ha igen, miként és milyen mértékben. A két változó közötti összefüggésre a Pearson-féle Khi-négyzet (CHI-SQUARE TESTS) táblázat adja meg a választ (lásd 4.9. táblázat). A mutató megfigyelt értéke 4,412, amely még 0,036-os (ASYMP. SIG. (2-sided)) kétoldali szignifikanciaszinten[3] vizsgálva is meghaladja az elméleti (küszöb) értéket, azaz a szignifikanciaszint kisebb, mint az általunk

[3] Az aszimptotikus szignifikancia (Asymp. Sig.) kifejezés arra utal, hogy az esetek egy részében, amikor összefüggéseket vizsgálunk, a tesztstatisztika eloszlása nem ismert, azonban amennyiben az esetek száma növekszik, az eloszlás hasonlítani kezd egy ismert eloszlásra. A továbbiakban ezt az ismert eloszlást alkalmazzuk a tesztstatisztika szignifikanciaszintjének számítására.

választott 0,05-ös (5 százalékos) szignifikanciaszint. Ez azt jelenti, hogy elvetjük a nullhipotézist, amely szerint nincs összefüggés a két változó között. Ennek megfelelően feltételezzük, hogy a régió és a márka között szignifikáns összefüggés van. Egyelőre azonban nem tudunk semmit arról, hogy ez a kapcsolat mit jelent.

Ennek megállapítására a 4.7. táblázatban a sorváltozó, azaz például a nyugat-magyarországi kategóriában az oszlop, azaz a márka megoszlását kell megvizsgálni, illetve ugyanezt kell tenni a kelet-magyarországi kategóriánál is. Természetesen figyeljünk arra, hogy a független változó szerint elemezzük a táblázatot. Ez alapján azt állapíthatjuk meg, hogy a nyugat-magyarországi kategóriában az 54,5 százalékhoz (TOTAL) hasonlítjuk a két márkára kapott százalékokat (A: 20,0; B: 83,3), ahol a B márkát fogyasztók aránya jóval magasabb az átlagosnál. Ez azt jelenti, hogy ez a reláció szorosan összefügg, azaz a nyugat-magyarországi lakosok a B márkát választják. Ugyanezt elvégezve a kelet-magyarországi kategóriára azt kapjuk, hogy ők az A márkát preferálják. A 4.7. táblázatban szereplő standardizált reziduumok szintén alátámasztják az eddigi megállapításainkat, ugyanis két cella esetében találtunk 2 feletti értéket, ahol valószínűsíthetjük, hogy a tényleges és az elvárt értékek között jelentős különbség van, azaz az adott relációk összefüggnek egymással.

A 4.8. táblázat alatt található *a* megjegyzés a Yates folytonossági korrekcióra utal (CONTINUITY CORRECTION), amely a Khi-négyzet korrekciója 2*2-es táblák esetén. (Ez automatikusan megjelenik 2*2-es tábláknál.) A Yates folytonossági korrekció a Khi-négyzet próba „konzervatívabb" változata, amely abban nyilvánul meg, hogy a szignifikanciaszint értéke magasabb (0,136), mint a Pearson Khi-négyzeté. Eszerint a mutató szerint a tábla két változója független egymástól, ugyanis az elutasítást indokló szignifikanciaszint (0,136) magasabb, mint az általunk választott 5 százalékos szignifikanciaszint. A valószínűségi arány (LIKELIHOOD RATIO) szintén hasonló a Khi-négyzethez, és nagy minták esetében értéke azonos a Khi-négyzetével. Ez alapján viszont a kapcsolat szignifikáns, ugyanis a 0,029 kisebb, mint a 0,05-ös érték.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,412[b]	1	,036		
Continuity Correction[a]	2,228	1	,136		
Likelihood Ratio	4,747	1	,029		
Fisher's Exact Test				,080	,067
Linear-by-Linear Association	4,011	1	,045		
N of Valid Cases	11				

a. Computed only for a 2x2 table

b. 4 cells (100,0%) have expected count less than 5. The minimum expected count is 2,27.

4.8. táblázat. Khi-négyzet teszt (Chi-Square Tests)

Egy másik mutató, a kis mintáknál használható Fisher EXACT teszt szerint azonban sem a kétoldali, sem egyoldali próba nem mutatott szignifikáns összefüggést. (2*2-es tábláknál az SPSS automatikusan számítja a Fisher-tesztet.) A lineáris kapcsolat mutató (LINEAR-BY-LINEAR ASSOCIATION), amely hasonló elven működik, mint a Khi-négyzet statisztika, ebben az esetben nem alkalmazható, mivel az intervallum- és arányskálák közötti összefüggést vizsgál, mi viszont nominális skálákat elemeztünk.

Ezek alapján megállapíthatnánk azt, hogy nem jutottunk előbbre, ugyanis néhány statisztika szerint van, néhány szerint viszont nincs szignifikáns összefüggés a két változó között. Ezt az ellentmondást két módon lehet feloldani: a gyakorlatban egyrészt sokszor előfordul, hogy ha az általánosan alkalmazott 5 százalékos szinten nincs összefüggés, akkor az elsőfajú hiba valószínűségét 10 százalékra növelik, amely mellett már esetünkben is elfogadható lenne a két változó közötti szignifikáns összefüggés. Másrészt viszont a Khi-négyzet és az ehhez hasonló statisztikák kritériumai sérültek (lásd a 4.9. táblázat alatti, alacsony kategóriánkénti elemszámra utaló b) megjegyzést) azzal, hogy bár teljesül a cellákra a minimum 1 várható érték, azonban a cellák több mint 20 százalékánál a várható érték nem haladja meg az 5-öt. Ez azt jelenti, hogy a tábla nem megbízható (reliability), azaz egy feltétel sérült, így a Khi-négyzet alkalmazása nem helyénvaló, illetve használata a változók transzformálását igényli (lásd 1. fejezet). Ennek alapján a FISHER'S EXACT teszt értékei leginkább mérvadók, azaz az eddigiek alapján a két változó között 5 százalékos szignifikanciaszint mellett nincs szignifikáns kapcsolat.

Vizsgáljuk meg a 4.5. táblázatban említett többi mutatót (lásd 4.9. táblázat) is! A lambda, Goodman és Kruskal-féle tau, illetve a bizonytalansági együttható aszimmetrikus mutatók, s a becslés hibavalószínűségének csökkenését jelzik. Ez például azt jelenti, hogy a lambda szerint 60 százalékkal (0,60)[4], a Goodman és Kruskal tau szerint 40,1 százalékkal (0,401), míg a bizonytalansági együttható szerint 31,3 százalékkal (0,313) javítja a régió ismerete a márkaválasztásra adott becslésünket. Ez azt jelenti, hogy a régió kitűnő előre jelző változónak tűnik az adott márkaválasztás tekintetében, azonban látni kell, hogy 5 százalékos szignifikanciaszinten, csak a GOODMAN ÉS KRUSKAL TAU (approx. Sig. = 0,045) és a bizonytalansági együttható (Approx. Sig. = 0,029) szignifikáns, a LAMBDA viszont nem (Approx. Sig. = 0,118). Az értékek közötti jelentős különbségek valamelyest csökkenek a Lambda kizárásával (40,1% vs. 30,1%), azonban ezeknek a mutatóknak a kezelése még így is komoly óvatosságot igényel.

Ugyanakkor a 4.9. táblázatban látható, hogy az egyes mutatókhoz több érték tartozik. A program ugyanis kiszámítja a mutató értékét annak a függvényében, hogy a függő változó a régió vagy a márka, illetve azon esetekben, ahol nem ismert, melyik a függő, illetve számol egy úgynevezett szimmetrikus (symmetric) mutatót, amely a másik két érték átlaga. Esetünkben az egyes mutatók értékei egyenlők, amely azt jelenti, hogy mindkét változó azonos mértékben hatással van a másikra, azaz akár a márka is lehetne független változó és a régió a függő.

[4] Az érték megegyezik a 4.3.2. alfejezetben általunk számítottal.

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,600	,277	1,563	,118
		régió Dependent	,600	,283	1,467	,142
		márka Dependent	,600	,283	1,467	,142
	Goodman and Kruskal tau	régió Dependent	,401	,297		,045[c]
		márka Dependent	,401	,297		,045[c]
	Uncertainty Coefficient	Symmetric	,313	,253	1,238	,029[c]
		régió Dependent	,313	,253	1,238	,029[d]
		márka Dependent	,313	,253	1,238	,029[d]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

d. Likelihood ratio chi-square probability.

4.9. táblázat. Lambda, Goodman és Kruskal tau, bizonytalansági együttható

A 4.10. táblázatban található szimmetrikus mutatók (Phi, Cramer V, kontingencia-együttható) mind szignifikánsak, és megegyeznek az általunk számított értékekkel (lásd 4.3.2. alfejezet), kivéve a phi előjelét. A Phi azért negatív, mert az SPSS egy az általunk használttól eltérő képletet alkalmaz, amely bizonyos mértékben magában foglalja a kapcsolat irányát is, ezért értéke –1-től +1-ig terjedhet. Ez nem változtat az értékén, csak az előjelén, amely figyelmen kívül hagyható. Így a 4.10. táblázat értékei alapján megállapítható, hogy a két változó között közepesnél erősebb szignifikáns kapcsolat van.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-,633	,036
	Cramer's V	,633	,036
	Contingency Coefficient	,535	,036
N of Valid Cases		11	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

4.10. táblázat. Phi, Cramer V és kontingencia-együttható

Mindezek alapján tekintsük át még egyszer a kereszttábla-elemzés folyamatát!

Gyakorlati útravaló

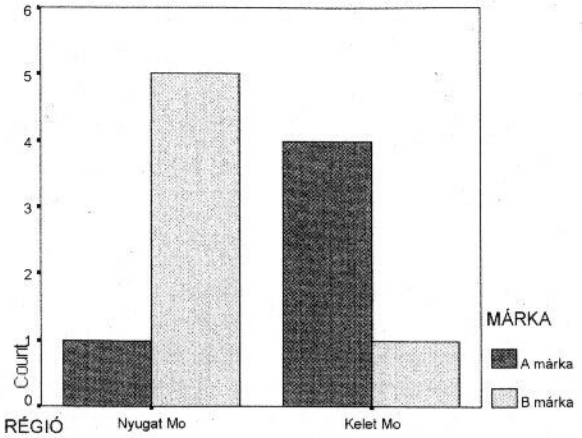
A kereszttábla-elemzés során a következő lépések elvégzése ajánlott:

- Megvizsgáljuk a tesztstatisztika (χ^2-próba) segítségével, hogy van-e szignifikáns összefüggés a változók között. Amennyiben nincs, akkor is elemezhetjük a kereszttáblát, de a megállapításainkhoz hozzá kell tennünk, hogy az eredmény nem bizonyult szignifikánsnak (azaz igen nagy a valószínűsége annak, hogy egy helyes nullhipotézist elutasítunk).
- Amennyiben van összefüggés, megvizsgáljuk – a skálák típusától függően – a kapcsolat erősségét.
- Ezzel egyidejűleg megvizsgáljuk, hogy a cellagyakorisággal kapcsolatos feltételek teljesülnek-e, amennyiben nem, a táblánk nem bizonyult megbízhatónak.
- Amennyiben van összefüggés és a feltételeink teljesülnek, levonjuk a következtetést a kereszttáblából a független változó szerinti a függő változóra.

Több mint 3 változót is lehet kereszttáblával elemezni, de ennek értelmezése igen bonyolult, mivel a cellák száma megsokszorozódik, az egy cellában lévő válaszadók vagy esetek minimális számának fenntartása problematikussá válik. Ha több mint két változó közötti kapcsolat vizsgálatáról van szó, akkor más struktúraelemző vagy -feltáró módszer, például regresszió-, variancia- vagy diszkriminanciaelemzés alkalmazása javasolt a változók mérési szintjének függvényében.

4.7. Az eredmények bemutatása

Az eredmények prezentálását tekintve a kereszttábla-elemzésnek a megadott változók arányait kell tartalmaznia, amelyet számos módon tudunk ábrázolni. Ennek egyik módszere az oszlopdiagram, ahol a független változó egyes kategóriái szerint feltüntetjük a függő változó kategóriáit. A 4.6. ábra bemutatja a kereszttábla-elemzés prezentálásának egyik módszerét, amelyet a kereszttábla-elemzés párbeszédpanelen belül a DISPLAY CLUSTERED BAR CHARTS menüponttal aktiváltunk (4.5. ábra). A 4.6. ábrán látható, hogy a nyugat-magyarországi lakosok a B márkát, míg a kelet-magyarországi lakosok az A márkát vásárolják.



4.6. ábra. Kereszttábla összefüggésének megjelenítése

A fentebb elemzett kereszttábla eredményeinek írásbeli bemutatása esetén azt állapíthatjuk meg, hogy a régió és a márkaválasztás szignifikánsan összefügg egymással, ahol a nyugat-magyarországi lakosok a B márkát, míg a kelet-magyarországi lakosok az A márkát vásárolják. Ez azonban nem megbízható eredmény, mert a megfigyelések száma alacsony. A kijelentés után zárójelben fel kell tüntetni a tesztstatisztika, illetve a szabadságfok (df) értékét, és a szignifikanciaszintet: ($\chi^2 = 4{,}412$; df = 1, $p = 0{,}036$).

Esettanulmány

Egy kereszttábla-elemzéssel arra kerestük a választ, hogy a gyógyszertárak elhelyezkedése (k22) és a tulajdonostárs megválasztása (k27) között milyen összefüggés van. Az összefoglaló táblázatból (4.11. táblázat) kiderül, hogy a mintában összesen 172 gyógyszertár van, amelyből 92,4 százalék, azaz 159 válaszolt mindkét kérdésre, 13 (7,6 százalék) pedig legalább az egyik kérdést kihagyta. Ennek eredményeként 159 esetet elemezhetünk.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Település típusa * Tulajdonostársai a családból kerülnek ki?	159	92,4%	13	7,6%	172	100,0%

4.11. táblázat. Összesítő táblázat

A kereszttáblában (4.12. táblázat) látható a két kérdés, ahol a település típusa a sor, míg a tulajdontárs (A tulajdonostárs a családból került-e ki?) az oszlopváltozó, ahol feltételezzük, hogy a település típusa meghatározza a tulajdonostárs megválasztását. A táblázat celláiban látható az abszolút érték, a sor, az oszlop és a teljes megoszlás, valamint a korrigált standardizált reziduum. A táblázatban a standardizált reziduumokra pillantva megállapítható, hogy a táblázat egyes relációi összefüggnek, azonban mielőtt alaposabban elemeznénk a kereszttáblát, vizsgáljuk meg, hogy az eredmény szignifikáns-e.

Település típusa * Tulajdonostársai a családból kerülnek ki? Crosstabulation

			Tulajdonostársai a családból kerülnek ki?		
			igen	nem	Total
Település típusa	Budapest	Count	9	20	29
		% within Település típusa	31,0%	69,0%	100,0%
		% within Tulajdonostársai a családból kerülnek ki?	9,8%	29,9%	18,2%
		% of Total	5,7%	12,6%	18,2%
		Adjusted Residual	-3,2	3,2	
	50 000-nél nagyobb lélekszámú város	Count	20	14	34
		% within Település típusa	58,8%	41,2%	100,0%
		% within Tulajdonostársai a családból kerülnek ki?	21,7%	20,9%	21,4%
		% of Total	12,6%	8,8%	21,4%
		Adjusted Residual	,1	-,1	
	egyéb	Count	63	33	96
		% within Település típusa	65,6%	34,4%	100,0%
		% within Tulajdonostársai a családból kerülnek ki?	68,5%	49,3%	60,4%
		% of Total	39,6%	20,8%	60,4%
		Adjusted Residual	2,4	-2,4	
Total		Count	92	67	159
		% within Település típusa	57,9%	42,1%	100,0%
		% within Tulajdonostársai a családból kerülnek ki?	100,0%	100,0%	100,0%
		% of Total	57,9%	42,1%	100,0%

4.12. táblázat. Kereszttábla

A Pearson-féle Khi-négyzet próba (4.13. táblázat) szerint a két változó szignifikáns (χ^2 =10,946; df=2, p=0,04), és a táblázat alatt található állítás szerint az elvárt értékek alapján a táblázat megbízható. A kapcsolat erősségét illetően a táblázat mérete alapján a Cramer V és a kontingencia-együttható alkalmazható, amelyek hozzávetőlegesen ugyanolyan, a közepesnél gyengébb szignifikáns eredményt mutatnak (0,262 és 0,254). (Lásd 4.14. táblázat.)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,946[a]	2	,004
Likelihood Ratio	10,930	2	,004
Linear-by-Linear Association	9,773	1	,002
N of Valid Cases	159		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 12,22.

4.13. táblázat. Khi-négyzet próba

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,262	,004
	Cramer's V	,262	,004
	Contingency Coefficient	,254	,004
N of Valid Cases		159	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

4.14. táblázat. Phi, Cramer-féle V és kontingencia-együttható

Ha a tulajdonostárs megválasztása a függő változó, akkor a Lambda értéke 0,164 (16,4%), a tau értéke 0,069 (6,9%) és a bizonytalansági együttható értéke 0,050 (5%), amely értékek szignifikánsak, azonban igen eltérők (4.15. táblázat). Mindazonáltal megállapítható, hogy a kapcsolat szignifikáns, és a független változó (régió) előre jelző képessége igen alacsony, azaz elég valószínű, hogy más változók is szerepet játszanak a tulajdonostárs megválasztásában.

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,085	,039	2,070	,038
		Település típusa Dependent	,000	,000	[c]	[c]
		Tulajdonostársai a családból kerülnek ki? Dependent	,164	,073	2,070	,038
	Goodman and Kruskal tau	Település típusa Dependent	,034	,021		,005[d]
		Tulajdonostársai a családból kerülnek ki? Dependent	,069	,039		,004[d]
	Uncertainty Coefficient	Symmetric	,042	,025	1,683	,004[e]
		Település típusa Dependent	,036	,021	1,683	,004[e]
		Tulajdonostársai a családból kerülnek ki? Dependent	,050	,030	1,683	,004[e]

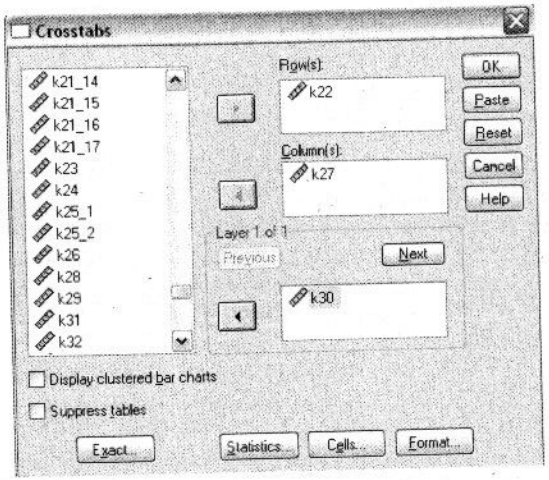
a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Cannot be computed because the asymptotic standard error equals zero.
d. Based on chi-square approximation
e. Likelihood ratio chi-square probability.

4.15. táblázat. A lambda, a Goodman and Kruskal tau és a bizonytalansági együttható

Lévén, hogy a kereszttábla szignifikáns, és habár statisztikailag a kapcsolat nem túl erős a két változó között, érdemes elemezni a táblát, és levonhatjuk a legfontosabb megállapításokat.

A táblázatból rögtön kitűnik, hogy a megfigyelések jelentős része – mintegy 60 százaléka – az 50 ezer főnél kisebb településekről származik, és nagyjából egyenletesen oszlik el a másik két kategória között (Budapest, 50 ezer főnél nagyobb település). Ha megvizsgáljuk a táblán belüli arányokat, és az ezeket alátámasztó standardizált reziduumokat, megállapítható, hogy a két reláció összefügg, mégpedig a budapesti gyógyszertáraknál a tulajdonostársak nem a családból kerülnek ki, míg az 50 ezer főnél nagyobb településeken igen. Ez alapján úgy tűnik, hogy a kisebb településeken családi vállalkozásokként működnek a patikák, Budapesten viszont nem. Emögött az állhat, hogy valószínűleg a családi vállalkozásoknak jobban kedvez a vidéki háttér.

Ha a településtípust egy másik változóval elemezzük, mégpedig hogy a patika új vagy már korábban létezett (jogelőd, k30), akkor azt tapasztaljuk, hogy a kapcsolat szignifikáns ($\chi^2 = 12{,}456$; df = 2, p = 0,002) és hogy Budapesten új patikákat hoztak létre, míg vidéken a már működőket üzemeltették tovább. Ugyanakkor ha a tulajdonostárs (k27) és a jogelőd (k30) változókat elemzem együtt, akkor ezek között semmilyen szignifikáns összefüggést nem találunk ($\chi^2 = 0{,}055$; df = 1, p = 0,815). Ezek alapján a következő relációkat találtuk szignifikánsnak: 1. Budapest és új patika, 2. Budapest és nem családi patika, 3. 50 ezer főnél nagyobb település és korlábban már működő patika, illetve 4. 50 ezer főnél nagyobb település és a családi patika. A három változó együttes elemzése sem hozott szignifikáns eredményt (jogelőd nélküli gyógyszertár esetén $\chi^2 = 2{,}464$; df = 2, p = 0,292; korábban létezett gyógyszertár esetén $\chi^2 = 10{,}041$; df = 2, p = 0,007). A három változó együttes elemzése úgy érhető el esetünkben, hogy a jogelőd változót bevisszük a legalsó „layer" ablakba, míg a másik két változót változatlanul hagyjuk a sor- és oszlopváltozó ablakokban (lásd 4.7. ábra).



4.7. ábra. Harmadik változó bevonása a kereszttábla-elemzésbe

Fontos tanulság azonban, hogy egy meglévő kapcsolatot érdemes más, az elméleti vagy a logikai modellünkbe beleillő változókkal, azaz egy harmadik változóval tesztelni. Ugyanis ilyen változók bevonásával mélyebb betekintést nyerhetünk az adatstruktúrába, és vagy megerősíthetjük a már feltárt összefüggést, vagy leleplezhetünk egy látszatkapcsolatot.

Természetesen felmerül a kérdés, hogy a középső csoportra (50 ezer fő feletti település, de nem Budapest) miért nem sikerült érdemleges megállapítást tennünk. Emögött az állhat, hogy a csoport túl nagy és általános, és további, részletesebb kategóriákra lenne szükség. Ennek általában – és ebben az esetben is – a kérdőívben előre megadott kategóriák, illetve az alacsony esetszám szab határt. Az eredmények pontosítására szóba jöhetne még a súlyozás lehetősége is, amennyiben az alapsokasági peremeloszlások rendelkezésünkre állnak.