

Összefüggés vizsgálat

1. Elméleti összefoglaló

Empirikus vizsgálatok során kutatási problémaként gyakran kerülhetünk szembe a „miért”, a „mi ennek az oka”, „milyen összefüggés van a változók között” kérdésekkel. Ezen kérdések kezelésére mind kvalitatív, mind kvantitatív eljárások ismertek. E kérdéskör statisztikai vizsgálatára úgynevezett **összefüggés vizsgálatot** alkalmazhatunk. Ekkor vizsgálhatjuk a változók kapcsolatának, együttmozgásának tulajdonságait, ez esetben **kapcsolatvizsgálatról** beszélünk, illetve a kapcsolat során ok-okozati viszonyt is vizsgálhatunk, ekkor úgynevezett **magyarázó modellek** alkalmazásáról beszélhetünk. Magyarázó modellek esetén az okok **magyarázóváltozóknak**, míg az okozatot **függő változónak** nevezzük. Például ha az alkoholizmus és a hajléktalanság összefüggését vizsgáljuk, akkor a két jelenség között kapcsolatot igen, de ok-okozati viszonyt nem vizsgálhatunk. Ezzel szemben, ha az a kérdés, hogy a tanulmányi átlag és a tanulással töltött idő összefüggését vizsgáljuk, akkor a változók közötti kapcsolat jellemzése mellett ok-okozati viszonyt is vizsgálhatunk.

Mit jelent az összefüggés vizsgálat végrehajtása?

1. **Első lépésben** meg kell adnunk a nullhipotézist. **E próbák nullhipotézise szerint a változók között nincs szignifikáns kapcsolat, míg az alternatív hipotézis szerint szignifikáns kapcsolat van a változók között.**
2. Második lépésben ki **kell választanunk az alkalmazandó eljárást.** Ehhez tisztázni kell a vizsgálatba bevont változók **mérési szintjét.** E vizsgálatoknál jelen kurzus során csak kategorikus (nominális és ordinális), illetve metrikus (skála) változókat különböztetünk meg.
 - Amennyiben **két kategoriális változó** összefüggését vizsgáljuk, **keresztábla elemzést** végzünk. Ez az eljárás nem ok-okozati viszonyt, csupán a változók „együttmozgását”, a kategóriák asszociációját vizsgálja. Például milyen szoros kapcsolat van az alkoholizmus (igen/nem) és a hajléktalanság (igen/nem) között, mennyire asszociálhatunk az egyik jelenségről a másikra. Ordinális mérési szintű változók kapcsolatának a vizsgálatára a rangkorrelációs együttható is alkalmazható.
 - Amennyiben **két, vagy több metrikus változó** összefüggését elemezzük, vizsgálhatjuk akár a **változók közötti kapcsolat tulajdonságát**, ez esetben **korrelációs számításról** beszélünk, illetve a metrikus változók rangsorai közötti kapcsolatot a rangkorreláció számítással jellemezhetjük. Továbbá vizsgálhatunk változók között ok-okozati viszonyt is. Ekkor **regresszió számításról** beszélünk. Például milyen szoros kapcsolat van a termék értékesített mennyisége és ára között, illetve hogyan befolyásolja egy termék értékesítését a termék ára. Ezzel a vizsgálati típussal a későbbi fejezetekben foglalkozunk.

- Amennyiben **metrikus és kategorikus változók** közötti összefüggést vizsgálunk, akkor
 - amennyiben **függőváltozó kategorikus**, a magyarázóváltozók körében pedig metrikus változókat szerepeltetünk, akkor **logisztikus regressziót** célszerű alkalmaznunk. Például befolyásolja-e egy termék ára azt, hogy megvesszük, vagy sem. Logisztikus regressziós modellben magyarázóváltozóként a metrikus változók mellé kategorikus változókat is betehetünk. Ezzel a vizsgálati típussal ezen a képzési szinten nem foglalkozunk.
 - Amennyiben a **függőváltozó metrikus**, a magyarázóváltozók kategorikusak, akkor **varianciaanalízist (ANOVA)** végezhetünk. Például befolyásolja-e egy termék értékesítését a termék minősége. Ezzel a vizsgálati típussal a későbbi fejezetekben foglalkozunk, az alapjaival az I. félévben foglalkoztunk (varianciafelbontás, külső és belső eltérés-négyzetösszeg). Természetesen vannak kevert modellek is, amikor a magyarázóváltozó oldalon keverednek a változó típusok. Ekkor ANCOVA modellről beszélünk.

3. A megfelelő eljárás kiválasztása után kiszámíthatunk egyrészt érdemes megvizsgálni **leíró statisztikai** mutatókkal (például gyakoriságok, relatív gyakoriságok vagy átlagok kiszámításával) vagy diagramokkal (például oszlop- vagy pontdiagramokkal) azt, hogy hogyan jelenik meg, hogyan rajzolódik ki a kapcsolat? Ez gyakorlatilag valamilyen összefüggés, mintázat keresését jelenti. A leíró statisztikai mutatókkal a minta és (amennyiben teljes körű megfigyelésről van szó) az alapsokaság is jellemezhető. Kiszámítható továbbá a kapcsolatmérő mutató (mind a mintára mind teljes körű megfigyelés esetén az alapsokaságra), amely a vizsgált változók közötti kapcsolat erősségét jellemzi. Ez gyakorlatilag egy a kiválasztott eljáráshoz kötődő kapcsolatmérő mutató kiszámítását jelenti. E mutatók közös jellemzője, hogy értékük abszolút értékben 0 és 1 közé esik. Ha a két változó között függvényyszerű kapcsolat van, akkor a mutató értéke egy, míg függetlenség esetén a mutató értéke nulla. A mutató egyre nagyobb értéke erősebb kapcsolatot jelent.

Keresztábra elemzés és varianciaanalízis esetén az alkalmazandó mutató csak a kapcsolaterősségre ad információt, míg korrelációs számítás esetén a mutató a kapcsolat erősségét és irányát is mutatja. Ezt a korrelációs számítás esetén részletesen taglaljuk.

4. Amennyiben következtetést akarunk levonni a (reprezentatív) minta alapján az alapsokaságra (**induktív statisztika** alkalmazása), akkor arra a kérdésre keressük a választ, hogy **szignifikánsnak minősül-e a kapcsolat**, avagy sem. A kiválasztott statisztikai próbafüggvény (teszt) lefuttatása után döntünk a nullhipotézis elfogadásáról, illetve elvetéséről. A nullhipotézis elvetése csupán azt jelenti, hogy a két változó nem tekinthető egymástól függetlennek, azaz ettől még a köztük lévő kapcsolat erőssége – gyakorlati szempontból – még jelentéktelen is lehet. Tehát ha szignifikáns kapcsolat van két változó között, akkor még a kapcsolaterősségét is jellemeznünk kell.

Meg kell jegyezni, hogy a kapcsolatvizsgálat pusztán matematikailag vizsgálja a változók együttmozgását, ami nem feltétlen jelent valós kapcsolatot. Mindig az

elemző felelőssége, hogy mely változók között keresünk összefüggéseket. Például, matematikailag szoros kapcsolat mutatható ki a csecsemőhalandóság és a Magyarországra érkező külföldi turisták száma között, de ez mégsem jelent valós kapcsolatot.

Kapcsolatvizsgálat során az **elsőfajú hiba** elkövetése azt jelenti, hogy **ott mutatunk ki kapcsolatot, ahol nincs is**.

A fejezet hátralevő részében kereszt táblák vizsgálatával és a foglalkozunk.

2. Kereszt táblák vizsgálata

Két kategoriális változó kapcsolatának vizsgálata a változók egy kombinációs táblája (kereszt táblája) alapján végezhető el. Ezt az Excelben a kimutatás készítés segítségével végezhetjük el.

Első lépésben a minta vagy (teljes körű megfigyelés esetén) az alapsokaság **leíró statisztikai** vizsgálatára azt érdemes megvizsgálnunk, mit is jelenthet a kapcsolat kategorikus változók esetén? Hogyan rajzolódik ki a kapcsolat? Ez gyakorlatilag valamilyen összefüggés, mintázat keresését jelenti, amelyet (mivel kategorikus változókkal foglalkozunk) gyakoriságokkal, relatív gyakoriságokkal (sor-, oszlop-, összeg százalék) jellemezhetünk. Továbbá kiszámíthatjuk a kapcsolat erősségét mérő mutatót.

A második lépésben Khí-négyzet teszt segítségével (**induktív statisztikai** eljárás alkalmazásával) megvizsgálhatjuk, hogy a vizsgált kapcsolat szignifikánsnak minősül vagy sem. A próba nullhipotézise szerint a két ismerv (változó) közötti kapcsolat nem szignifikáns. Például, ha azt vizsgáljuk, hogy a jegyzet használat független-e a tagozattól, akkor a nullhipotézis azzal ekvivalens, hogy mind a nappalis, mind a távos hallgatók esetében a jegyzethasználat szerinti megoszlás szignifikánsan nem különbözik.

A mind a leíró, mind az induktív statisztikai eljárások esetén a vizsgálat alapja az a valószínűségi számításból tanult összefüggés, mely szerint, ha A és B események egymástól függetlenek, akkor $P(AB) = P(A)P(B)$. E vizsgálatok során a táblázatok egy mintaadatait tartalmazza. Nekünk meg kellene határozni azt, hogy a vizsgált táblázat belsejében milyen adatok szerepelnének függetlenség esetén, majd ezeket össze kellene hasonlítani. A következőkben ezt szemléltetem. A vizsgálat során f_{ij} azt jelenti, hogy a megfigyelések alapján az egyes cellákba hány elem került, f_{ij}^* azt jelenti, hogy az egyes cellákba hány elem kerülne, ha a két ismerv független lenne egymástól.

Nézzünk egy példát egy város véletlenszerűen (FAE) kiválasztott 1156 lakásának komfortságára és a benne élők háztartástípusára vonatkozó adatok alapján.

Komfortfokozat	Háztartástípus		Összesen
	Egyedülálló és gyermektelen házaspárok száma	Gyermekecs házaspárok száma	
Komfort nélküli	20	10	30
Félkomfortos	90	236	326
Komfortos	300	500	800
Összesen	410	746	1156

Mivel a vizsgált két ismérv (komfortfokozat, háztartástípus) minőségi ismérv, ezért a kereszttábla elemzést hajthatunk végre. Először számítsuk ki, hogy milyen értékek szerepelnének az egyes cellákban, ha a két ismérv teljesen független lenne egymástól!

A megadott táblázat egy empirikus vizsgálat eredményét tartalmazza. A táblázat belsejében szereplő f_{ij} értékek azt jelentik, hogy a megfigyelések alapján az egyes cellákba hány elem került. A kérdés azt jelenti, hogy az egyes cellákba milyen f_{ij}^* értékek kerülnének az f_{ij} értékek helyére. Az f_{ij}^* azt mutatja, hogy az egyes cellákba hány elem kerülne, ha a két ismérv független lenne egymástól. Kérdés az, hogy hogyan. Ennek szemléltetésére néhány megjegyzést teszek, amelyben átmenetileg használjuk a valószínűség terminust arra a kifejezésre, hogy valami az esetek hány százalékában teljesül.

1. Például, annak a valószínűsége, hogy egy vizsgált lakás komfort nélküli 30/1156.
2. Annak a valószínűsége, hogy egy vizsgált lakásban egyedülálló, vagy gyermektelen házaspárok laknak 410/1156.
3. Annak a valószínűsége, hogy egy vizsgált lakásban egyedülálló, vagy gyermektelen házaspárok laknak, valamint a vizsgált lakás komfort nélküli 20/1156.
4. Annak a valószínűsége, hogy egy vizsgált lakásban egyedülálló, vagy gyermektelen házaspárok laknak, valamint a vizsgált lakás komfort nélküli és a két ismérv (esemény) egymástól független $\frac{410}{1156} \cdot \frac{30}{1156}$. Ennek a magyarázata az, hogy ha A és B események egymástól függetlenek, akkor $P(AB)=P(A)P(B)$.
5. Ebből következően, ha a két ismérv teljesen független egymástól, akkor azon vizsgált lakások száma, ahol egyedülálló, vagy gyermektelen házaspárok laknak, valamint a vizsgált lakás komfort nélküli $\frac{410}{1156} \cdot \frac{30}{1156} \cdot 1156$. Ennek a magyarázata az, hogy a

valószínűséget leegyszerűsítve úgy foghatjuk fel, hogy az esetek hány százalékában teljesül valami. Tehát a függetlenség esetén fennálló gyakoriságokat gyakorlatilag a sorok, az oszlopok összege és a minta elemszáma alapján határozhatjuk meg.

Ezek után számítsuk ki az összes f_{ij}^* értékeket. Ennek menetét az alábbi táblázat zárójelben szereplő képletei mutatják.

Komfortfokozat	Háztartástípus		Összesen
	Egyedülálló és gyermektelen házaspárok száma $f_{ij} (f_{ij}^*)$	Gyermekek házaspárok száma $f_{ij} (f_{ij}^*)$	
Komfort nélküli	$20 (\frac{30410}{1156} = 10,64)$	$10 (\frac{30746}{1156} = 19,36)$	30
Félkomfortos	$90 (\frac{326410}{1156} = 115,62)$	$236 (\frac{326746}{1156} = 210,38)$	326
Komfortos	$300 (\frac{800410}{1156} = 283,74)$	$500 (\frac{800746}{1156} = 516,26)$	800
Összesen	410	746	1156

Látható, hogy egy-egy cellában az f_{ij} és az f_{ij}^* értékek nem azonosak, így a komfortfokozat és a háztartás típus nem teljesen független, kérdés az, hogy ezek az eltérések szignifikánsak-e, azaz szignifikáns kapcsolat van-e a két változó között.

A vizsgálat lefuttatásához egyetlen képlettel kell összehasonlítani az összes összetartozó f_{ij} és f_{ij}^* értéket. Ezt különféle matematikai okok miatt a

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$$

képlettel tehetjük meg. A dupla szumma jel azt jelenti, hogy a képletben szereplő mennyiséget minden cella esetében külön-külön kiszámítjuk, majd ezeket összeadjuk. A képletben – az angol szavak kezdőbetűiből – az r a sorok számát, míg a c az oszlopok számát jelöli.

Látható, hogy teljesen függetlenség esetén (mindencellában $f_{ij} = f_{ij}^*$), a próbafüggvény értéke nulla.

Ebben a konkrét példában

$$\chi^2 = \frac{(20 - 10,64)^2}{10,64} + \frac{(10 - 19,36)^2}{19,36} + \dots + \frac{(500 - 516,26)^2}{516,26} = 23,0022$$

A kérdés az, hogy ez az érték nagy eltérést jelez-e a nulla értéktől. Ennek megválaszolására kiszámíthatjuk a Cramer-együtthatót (a kapcsolat erősségének vizsgálatára) az alábbi módon:

$$C = \sqrt{\frac{\chi^2}{n \cdot \min\{(r-1), (c-1)\}}}$$

Ahol a nevezőben megjelenő szorzat második tényezője azt jelenti, hogy az (r-1) és (c-1) értékek közül a kisebbet kell behelyettesíteni. A mutató esetén a 0,2-nél alacsonyabb értékek gyenge, a [0,2;0,3) intervallumba eső értékek közepes erősségű, és a 0,3-mal egyenlő, vagy afeletti értékek erős kapcsolatra utalnak.

Az előző konkrét példa szerint

$$C = \sqrt{\frac{\chi^2}{n \cdot \min\{(r-1), (c-1)\}}} = \sqrt{\frac{23,0022}{1156 \cdot \min\{(3-1), (2-1)\}}} = \sqrt{\frac{23,0022}{1156 \cdot 1}} = 0,14$$

Azaz a komfortfokozat és a háztartástípus között gyenge kapcsolat van.

További elemzési lépésként merülhet fel az induktív statisztika alkalmazása. Azaz arra a kérdésre keressük a választ, hogyha mintánk van, akkor milyen következtetést tudunk az alapsokaságra levonni? Azaz mondhatjuk-e, hogy az eddig leírt összefüggés szignifikáns kapcsolatot jelent? Ennek eldöntésére használjuk a hipotézisvizsgálatot. Ennek végrehajtását Excelben és SPSS-ben mutatjuk be.

Megjegyezendő, hogy a hipotézisvizsgálat alkalmazásának a feltétele az, hogy a függetlenség esetén fennálló gyakoriságok ériék el legalább az 5-ös értéket. Ennek nem teljesülése a keresztábra celláinak legfeljebb 20 százalékaig tolerálható. Az előfeltétel ellenőrzését az SPSS program végrehajtja, míg az Excel alkalmazásánál ezt nekünk kell ellenőriznünk.

SZEGEDI TUDOMÁNYEGYETEM
GAZDASÁGTUDOMÁNYI KAR
KÖZGAZDÁSZ KÉPZÉS
TÁVOKTATÁSI TAGOZAT
LECKESOROZAT
COPYRIGHT © SZTE GTK 2017/2018

A LECKE TARTALMA, ILLETVE ALKOTÓ ELEMEI ELŐZETES,
ÍRÁSBELI ENGEDÉLY MELLETT HASZNÁLHATÓK FEL.

JELLEN TANYAG
A SZEGEDI TUDOMÁNYEGYETEMEN KÉSZÜLT
AZ EURÓPAI UNIÓ TÁMOGATÁSÁVAL.
PROJEKT AZONOSÍTÓ: EFOP-3.4.3-16-2016-00014

SZÉCHENYI 2020



MAGYARORSZÁG
KORMÁNYA

Európai Unió
Európai Szociális
Alap



BEFEKTETÉS A JÖVŐBE