

Függetlenségvizsgálat χ^2 -próbával (kidolgozott feladatok)

Szalay Krisztina

1. feladat

Egy nagyvállalatnál azt vizsgálták, van-e összefüggés a dolgozók neme és a beosztása között. Az eredményeket az alábbi táblázat tartalmazza:

	vezető	beosztott
nő	24	158
férfi	57	321

Elfogadható-e 95%-os szignifikancia szinten az az állítás, hogy a dolgozók neme és beosztása között nincs összefüggés?

Megoldás: Feltételezésünk az, hogy a nem és a beosztás független egymástól, ezért függetlenségvizsgálatot hajtunk végre. Az adatokat a nem szerint $r = 2$, a beosztás szerint $s = 2$ csoportba osztották, a mintaelemek száma összesen $n = 540$.

A fenti, megfigyelt gyakoriságokat tartalmazó táblázatot (2×2 -es kontingencia-tábla) egészítsük ki a peremgyakoriságokkal. Ez azt jelenti, hogy meghatározzuk a nemek szerinti gyakoriságokat (a mintában összesen hány nő és hány férfi szerepel), illetve a beosztás szerinti gyakoriságokat (a mintában összesen hány vezető és hány beosztott szerepel). Ehhez összeadjuk az egyes sorokban, illetve oszlopokban található értékeket. Az alábbi táblázatot kapjuk:

	vezető	beosztott	összes
nő	24	158	182
férfi	57	321	378
összes	81	479	560

Most nézzük meg, hogy függetlenség esetén milyen eloszlást várunk!

Az elvárt gyakoriságok kiszámításához minden esetben az alábbiak szerint kell eljárunk: az adott megfigyelt gyakorisághoz tartozó sor- és oszlopösszeget összeszorozzuk, majd elosztjuk a minta elemszámával. Például a női vezetők száma (megfigyelt gyakorisága) a megfigyelés szerint 24. A női vezetők számának elvárt gyakoriságát úgy számolhatjuk ki, hogy a 24-hez tartozó sorösszeget (182) szorozzuk az oszlopösszeggel (81), majd a szorzatot elosztjuk a minta elemszámával (560).

Így a női vezetők számának elvárt gyakorisága: $\frac{182 \cdot 81}{560} = 26,3250$.

Hasonlóan számítjuk ki a többi értéket is:

A férfi vezetők számának elvárt gyakorisága: $\frac{378 \cdot 81}{560} = 54,6750$.

A női beosztottak számának elvárt gyakorisága: $\frac{182 \cdot 479}{560} = 155,6750$.

A férfi beosztottak számának elvárt gyakorisága: $\frac{378 \cdot 479}{560} = 323,3250$.

A fenti értékekkel elkészíthetjük a várt gyakoriságoknak megfelelő táblázatot (ügyelve arra, hogy minden értéket a megfelelő helyre írjunk):

	vezető	beosztott	összes
nő	26.3250	155.6750	182
férfi	54.675	323.3250	378
összes	81	479	560

Ezután csak annyi dolgunk van, hogy a megfigyelt és várt gyakoriságok különbségének négyzetét osztjuk a várt gyakorisággal minden cella esetén, majd összeadjuk a kapott értékeket:

$$\frac{(24 - 26,3250)^2}{26,3250} + \frac{(158 - 155,6750)^2}{155,6750} + \frac{(57 - 54,6750)^2}{54,6750} + \frac{(321 - 323,3250)^2}{323,3250} = 0,3557$$

A hipotézisvizsgálat lépései ezután az alábbiak:

1. *A nullhipotézis kimondása.*

A feltételezésünk az, hogy ennél a vállalatnál a nem és a beosztás függetlenek egymástól.

2. *A próbafüggvény kiválasztása és értékének kiszámítása.*

A próba függetlenségvizsgálat χ^2 -próbával. A próbastatisztika mintából számított értéke a korábban kiszámolt 0,3557.

3. *A kritikus tartomány kijelölése.*

A próba egyoldali χ^2 -próba, az elfogadási tartomány $(0, \chi_t^2)$. A kritikus értéket a χ^2 -eloszlás táblázatából olvashatjuk ki. Ügyeljünk arra, hogy függetlenségvizsgálatnál $(r - 1) \cdot (s - 1)$ a szabadsági fok, ami jelen példa esetében $(2 - 1) \cdot (2 - 1) = 1$, 0,95 valószínűség mellett. Ez a jegyzet χ^2 -eloszlás táblázatában az első sor, első oszlopban található

$$\chi_t^2 = 3,84$$

érték.

4. *A döntés.*

$0,3557 < 3,84$, a próbastatisztika értéke az elfogadási tartományba esik, a nullhipotézist elfogadjuk 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján elfogadható, hogy a vállalatnál független a dolgozók beosztása a nemüktől.

2. feladat

Egy kutatás során azt vizsgálták, hogy összefügg-e a néző életkora és a legkedveltebb film típusa. A felmérés eredményeit az alábbi táblázat tartalmazza:

	ismeretterjesztő film	kalandfilm	vígjáték	romantikus film
0-18 éves	12	52	34	25
19-49 éves	28	35	43	24
49 év felett	58	14	43	32

Elfogadható-e 95%-os szignifikancia szinten az az állítás, hogy az életkor és a legkedveltebb film típusa között nincs összefüggés?

Megoldás: Az adatokat az életkor szerint $r = 3$, a film típusa szerint $s = 4$ csoportba osztották, a mintaelemek száma összesen $n = 400$.

A megfigyelt gyakoriságokat tartalmazó táblázatban (3×4 -es kontingencia-tábla) az egyes sorokban, illetve az egyes oszlopokban található gyakoriságokat összeadva megkapjuk a peremgyakoriságokat. Foglaltuk ezt össze az alábbi táblázatban:

	ismeretterjesztő film	kalandfilm	vígjáték	romantikus film	összes
0-18 éves	12	52	34	25	123
19-49 éves	28	35	43	24	130
49 év felett	58	14	43	32	147
összes	98	101	120	81	400

A következő lépésben számítsuk ki, hogy függetlenség esetén milyen értékeket várunk az egyes cellákba!
Az előző feladatban láttuk szerint az adott megfigyelt gyakorisághoz tartozó sor- és oszlopösszeget összeszorozzuk, majd elosztjuk a minta elemszámával.

Vegyük észre, hogy az algoritmus a következő számítást adja: a kiválasztott cella és az utolsó sor és utolsó oszlopban található, összes mintaszámnak megfelelő cella, tekinthető egy téglalap két szemközi csúcsának. Pl. tekintsük a 19-49 évesek közül a kalandfilmeket kedvelőket. A cellában a gyakoriság 35, az minta elemszáma 400. Egészítsük ki a táblázatot a téglalap másik két csúcsában található cellák elemszámával (101, 130):

	ismeretterjesztő film	kalandfilm	vígjáték	romantikus film	összes
0-18 éves					
19-49 éves		35			130
49 év felett					
összes		101			400

Ennek segítségével a 19-49 évesek között a kalandfilmeket kedvelők cellájába a várt gyakoriság úgy számolható, hogy a téglalapon a kiválasztott cellával szomszédos csúcsokon található cellák értékeinek szorzatát osztjuk a szemközi csúcs cellájának értékével:

$$\frac{130 \cdot 101}{400} = 32,8250$$

Hasonlóan a többi cellára is kiszámítva a várt gyakoriságot, az alábbi táblázatot kapjuk:

	ismeretterjesztő film	kalandfilm	vígjáték	romantikus film	összes
0-18 éves	30.1350	31.0575	36.9000	24.9075	123
19-49 éves	31.8500	32.8250	39.0000	26,3250	130
49 év felett	36.0150	37.1175	44.1000	29.7675	147
összes	98	101	120	81	400

Kiszámítjuk a próbastatisztika értékét (azaz a megfigyelt és várt gyakoriságok különbségének négyzetét osztjuk a várt gyakorisággal minden cella esetén, majd összeadjuk a kapott értékeket):

$$\begin{aligned} & \frac{(12 - 30,1350)^2}{30,1350} + \frac{(52 - 31,0575)^2}{31,0575} + \frac{(34 - 36,9000)^2}{36,9000} + \frac{(25 - 24,9075)^2}{24,9075} + \\ & \frac{(28 - 31,8500)^2}{31,8500} + \frac{(35 - 32,8250)^2}{32,8250} + \frac{(43 - 39,0000)^2}{39,0000} + \frac{(24 - 26,3250)^2}{26,3250} + \\ & \frac{(58 - 36,0150)^2}{36,0150} + \frac{(14 - 37,1175)^2}{37,1175} + \frac{(43 - 44,1000)^2}{44,1000} + \frac{(32 - 29,7675)^2}{29,7675} = 54,3686 \end{aligned}$$

A hipotézisvizsgálat :

1. A nullhipotézis kimondása.

A feltételezésünk az, hogy az életkor és preferált film típusa függetlenek egymástól.

2. A próbafüggvény kiválasztása és értékének kiszámítása.

A próba függetlenségvizsgálat χ^2 -próbával. A próbastatisztika mintából számított értéke 54,3686.

3. A kritikus tartomány kijelölése.

A próba egyoldali χ^2 -próba, az elfogadási tartomány $(0, \chi_t^2)$. A kritikus értéket olvassuk ki a χ^2 -eloszlás táblázatából. A szabadsági fok $(r - 1) \cdot (s - 1) = (3 - 1) \cdot (4 - 1) = 6$, 0,95 valószínűség mellett. Ez a jegyzet χ^2 -eloszlás táblázatában a hatodik sor, első oszlopban található

$$\chi_t^2 = 12,59$$

érték.

4. A döntés.

$54,3686 > 12,59$, a próbastatisztika értéke az elutasítási tartományba esik, a nullhipotézist elutasítjuk, 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján van összefüggés az életkor és a legkedveltebb film típusa között.