Khi-négyzet próba függetlenségvizsgálatra

Szűcs Mónika, Griechisch Erika, Rárosi Ferenc
SZTE ÁOK-TTIK Orvosi Fizikai és Orvosi Informatikai Intézet

Utoljára frissítve: 2020. június 4.

1. Megfigyelt és várt gyakorisági táblázat

Egy olyan táblázat, mely a megfigyelt gyakoriságokat tartalmazza, a két változó alapján csoportosítva. Az egyik változó kimenetelei (A_1, A_2, $\ldots A_r$) kerülnek a sorokba, a másik változóé (B_1, B_2, $\ldots B_c$) pedig az oszlopokba. O_{ij} annak az abszolút gyakorisága, amikor A_i és B_j egyszerre teljesül.

	B_1	B_2	$\ldots$	B_c	Sorösszeg
A_1	O_{11}	O_{12}	$\ldots$	O_{1c}	O_{1+}
A_2	O_{21}	O_{22}	$\ldots$	O_{2c}	O_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\ldots$	O_{rc}	O_{r+}
Oszlopösszeg	O_{+1}	O_{+2}	$\ldots$	O_{+c}	n

- $O_{i+} = \sum_{j=1}^{c} O_{ij}$ az A_i esemény gyakorisága (sorösszegek), ahol $i = 1, 2, \ldots, r$
- $O_{+j} = \sum_{i=1}^{r} O_{ij}$ a B_j esemény gyakorisága (oszlopösszegek), ahol $j = 1, 2, \ldots, c$
- n a megfigyelések száma

A relatív gyakorisági eloszlását a sorösszegeknek illetve az oszlopösszegeknek marginális eloszlásoknak hívjuk. **Tegyük fel, hogy a két változó független**, ekkor az eloszlások minden oszlop esetén azonosak. Minden oszlopra a marginális eloszlást feltételezve, kapjuk a **várt gyakoriságokat**:

$$E_{ij} = \frac{O_{i+} \cdot O_{+j}}{n} = \frac{\text{sorösszeg} \times \text{oszlopösszeg}}{\text{megfigyelések száma}}$$

	B_1	B_2	$\ldots$	B_c	Sorösszeg
A_1	E_{11}	E_{12}	$\ldots$	E_{1c}	O_{1+}
A_2	E_{21}	E_{22}	$\ldots$	E_{2c}	O_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	E_{r1}	E_{r2}	$\ldots$	E_{rc}	O_{r+}
Oszlopösszeg	O_{+1}	O_{+2}	$\ldots$	O_{+c}	n

SZÉCHENYI 2020
MAGYARORSZÁG KORMÁNYA
Európai Unió
Európai Szociális Alap
BEFEKTETÉS A JÖVŐBE

Példa megfigyelt és várt gyakorisági táblázatokra
Van e kapcsolat az influenzás megbetegedések száma és a vakcina típusa között?
Megfigyelt gyakorisági táblázat

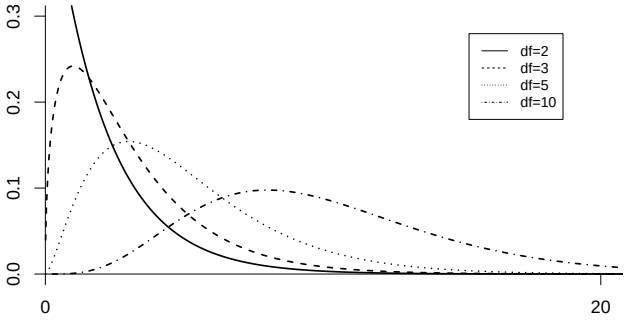
	Influenzás	Nem lett influenzás	Total
Csak szezonális	43 (15.36%)	237 (84.64%)	280 (100%)
Csak H1N1	52 (20.80%)	198 (79.20%)	250 (100%)
Kombinált	25 (9.26%)	245 (90.74%)	270 (100 %)
Total	120	680	800

Várt gyakorisági táblázat

	Influenzás	Nem lett influenzás	Total
Csak szezonális	$\frac{280\times120}{800} = \mathbf{42}$	$\frac{280\times680}{800} = \mathbf{238}$	280
Csak H1N1	$\frac{250\times120}{800} = \mathbf{37.5}$	$\frac{250\times680}{800} = \mathbf{212.5}$	250
Kombinált	$\frac{270\times120}{800} = \mathbf{40.5}$	$\frac{270\times680}{800} = \mathbf{229.5}$	270
Total	120	680	800

2. Khi-négyzet eloszlás

Ha X_1, X_2 … X_m független, standard normális eloszlású véletlen változók, akkor $X_1^2+X_2^2+\ldots X_m^2 = \sum_{i=1}^{m} X_i^2$ khi-négyzet (χ^2) eloszlást követ m szabadsági fokkal.



3. Khi-négyzet próba függetlenségvizsgálatra

Célja annak a vizsgálata, hogy populációban van-e két diszkrét változó közötti kapcsolat.

Feltétele, hogy a **várt gyakoriságok** legfeljebb **20%**-a kisebb 5-nél.
Gyakran ennél szigorúbb, de könnyebben ellenőrizhető feltételt használunk: a várt gyakoriságok mindegyike legalább 5.

Hipotézisek:

- H_0: a két változó független. $P(A_i \cdot B_j) = P(A_i) \cdot P(B_j)$
- H_1: a két változó között van összefüggés.

Számolás és döntés:

I. Próbastatisztika

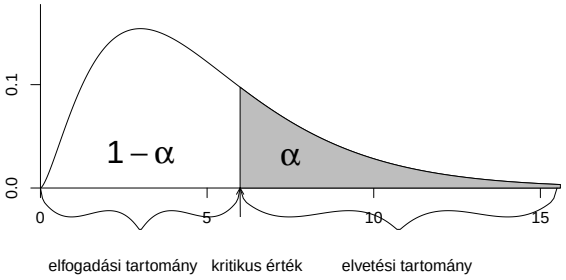
$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Ha a két változó független, a próbastatisztika χ^2 eloszlást követ **(r – 1)(c – 1) szabadsági fokkal**

Döntési szabály:

- Ha $\chi^2 < \chi^2_\alpha$, a nullhipotézist elfogadjuk
- Ha $\chi^2 > \chi^2_\alpha$, a nullhipotézist elvetjük

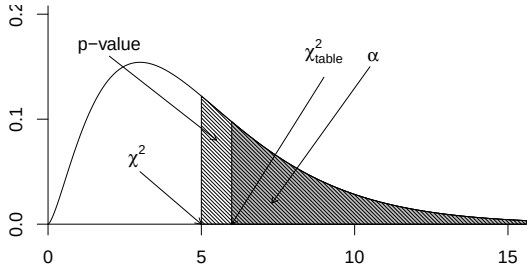
ahol χ^2_α a khi-négyzet eloszlás $(r-1)(c-1)$ szabadságfokhoz és α szignifikancia szinthez tartozó kritikus értéke.



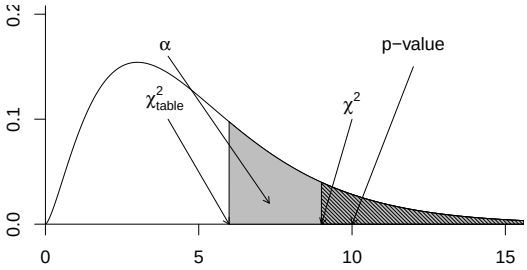
II. p-érték

Döntési szabály:

Ha $p > \alpha$, akkor H_0-t elfogadjuk



Ha $p < \alpha$, akkor H_0-t elvetjük



Számolás R-rel:

```
> chisq.test(nem, orul, correct = FALSE)

  Pearson's Chi-squared test

data:  nem and orul
X-squared = 5.9852, df = 1, p-value = 0.01443


gyakorisag_tabla = table(nem, orul)
> chisq.test(gyakorisag_tabla, correct = FALSE)

  Pearson's Chi-squared test

data:  gyakorisag_tabla
X-squared = 5.9852, df = 1, p-value = 0.01443
```

3.1. Khi-négyzet próba függetlenségvizsgálatra Yates-korrekcióval

Abban az esetben, a szabadsági fok 1 (2x2-es táblázat), a khi-négyzet próba próbastatisztikája pontosabban számolható, ha korrekciót alkalmazunk. Az egyik leggyakrabban alkalmazott korrekció, a Yates-féle folytonossági korrekció. **Ez a korrekció csak két bináris[1] változó közötti kapcsolat elemzése esetén használható.**

[1] kétértékű, pl. igen/nem vagy férfi/nő

	B_1	B_2	Sorösszeg
A_1	$O_{11} = a$	$O_{12} = b$	$O_{1+} = a + b$
A_2	$O_{21} = c$	$O_{22} = d$	$O_{2+} = c + d$
Oszlopösszeg	$O_{+1} = a + c$	$O_{+2} = b + d$	$n = a + b + c + d$

Próbastatisztika Yates-féle korrekcióval:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(\left|O_{ij} - E_{ij}\right| - \frac{1}{2}\right)^2}{E_{ij}} = \frac{n\left(|ad - bc| - \frac{1}{2}n\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Számolás R-rel:

```
> chisq.test(nem, orul)

  Pearson's Chi-squared test with Yates' continuity correction

data:  nem and orul
X-squared = 4.6565, df = 1, p-value = 0.03094
```

4. Fisher-féle egzakt próba

Amennyiben a χ^2 próba feltétele nem teljesül, leggyakrabban a Fisher-féle egzakt próbát használjuk két diszkrét változó közötti összefüggésének vizsgálatára. Gyakorlatban csak kis mintaelemszám esetén használjuk, de **bármekkora minta esetén pontos értéket ad**.

```
> fisher.test(nem,orul)

  Fisher's Exact Test for Count Data

data:  nem and orul
p-value = 0.01991
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.189144 58.646564
sample estimates:
odds ratio
  5.932357
```

Felhasznált irodalom

- Reiczigel Jenő, Harnos Andrea, Solymosi Norbert: Biostatisztika nem statisztikusoknak, Pars Kft. (2014)
- Reiczigel Jenő: Válogatott fejezetek a biostatisztikából, SZIE ÁOTK (2005) http://www2.univet.hu/users/jreiczig/valfej/val-fej-jegyzet-2005-02-05.pdf (2019.05.21.)

Jelen tananyag a Szegedi Tudományegyetemen készült
az Európai Unió támogatásával.
Projekt azonosító: EFOP-3.4.3-16-2016-00014