

# Illeszkedésvizsgálat $\chi^2$ -próbával

Szalay Krisztina

## 1. feladat

### (tisztá illeszkedésvizsgálat)

Négy pénzérmét 160-szor feldobunk. A kapott gyakoriságok:

|             |   |    |    |    |    |        |
|-------------|---|----|----|----|----|--------|
| fejek száma | 0 | 1  | 2  | 3  | 4  | Összes |
| gyakoriság  | 5 | 35 | 67 | 41 | 12 | 160    |

Elfogadható-e 95%-os szignifikancia szinten, hogy az érmék szabályosak?

**Megoldás:** Feltételezésünk az, hogy az érmék szabályosak (mindegyiken  $\frac{1}{2}$  valószínűséggel dobunk fejet). A mintaelemek száma  $n = 160$  (160 kísérletet hajtottunk végre), a mintaelemeket a dobott fejek száma szerint osztályoztuk,  $r = 5$  csoportba soroltuk.

A dobott fejek számát tekintjük az  $\xi$  valószínűségi változónak. Ha helyes a feltételezésünk, akkor  $\xi$  binomiális eloszlású valószínűségi változó  $n = 4, p = \frac{1}{2}$  paraméterekkel. A paraméterek elméleti értékek, ebben a feladatban nem a mintából becsüljük a várható értéket, így tisztá illeszkedésvizsgálatot hajtunk végre.

A hipotézisvizsgálat során azt kell eldöntenünk, hogy a megfigyelt valószínűségi változó (a dobott fejek száma), a minta alapján tekinthető-e adott paraméterű binomiális eloszlásúnak. Ennek alapján a próba illeszkedésvizsgálat.

Először meg szeretnénk határozni azt, hogy ha az érmék szabályosak, akkor az egyes osztályokba hány mintaelemet várnánk. Számítsuk ki a dobott fejek számának lehetséges értékeihez tartozó valószínűségeket (azaz írjuk fel  $\xi$  eloszlását)!

$$\begin{aligned}p_0 &= P(\xi = 0) = \left(\frac{1}{2}\right)^4 = \frac{1}{16} \\p_1 &= P(\xi = 1) = \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 \binom{4}{1} = \frac{4}{16} \\p_2 &= P(\xi = 2) = \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 \binom{4}{2} = \frac{6}{16} \\p_3 &= P(\xi = 3) = \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 \binom{4}{3} = \frac{4}{16} \\p_4 &= P(\xi = 4) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}\end{aligned}$$

A kiszámolt valószínűségek alapján meg tudjuk mondani, hogy mennyi az egyes osztályokba eső mintaelemek elvárt gyakorisága, azaz a binomiális eloszlás fennállása esetén átlagosan hány mintaelem esik az egyes osztályokba. Ehhez nincs más dolgunk, mint a fenti valószínűségekkel szorozni a mintaelemek számát ( $np_i$ ) (tehát pl. a 160 kísérletből átlagosan  $p_0 = \frac{1}{16} \cdot 160 = 10$  esetben kapunk olyan dobáseredményt, ahol a négy dobás egyike sem fej). Az utolsó oszlopba a megfelelő helyre a sorösszeg kerül.

Az áttekinthetőbb számoláshoz az alábbi táblázatot hozzuk létre:

|                                |     |       |        |       |     |        |
|--------------------------------|-----|-------|--------|-------|-----|--------|
|                                |     |       |        |       |     | összeg |
| dobott fejek száma             | 0   | 1     | 2      | 3     | 4   |        |
| megfigyelt gyakoriság: $\mu_i$ | 5   | 35    | 67     | 41    | 12  |        |
| elvárt gyakoriság: $np_i$      | 10  | 40    | 60     | 40    | 10  |        |
| $(\mu_i - np_i)^2$             | 25  | 25    | 49     | 1     | 4   |        |
| $(\mu_i - np_i)^2 / (np_i)$    | 2,5 | 0,625 | 0,8167 | 0,025 | 0,4 | 4,3667 |

A hipotézisvizsgálat lépései ezután az alábbiak:

1. *A nullhipotézis kimondása.*

A feltételezésünk most az, hogy az érmék szabályosak, a dobott fejek száma binomiális eloszlású,  $(4, \frac{1}{2})$  paraméterekkel.

2. *A próbafüggvény kiválasztása és értékének kiszámítása.*

A próba illeszkedésvizsgálat,  $\chi^2$ -próbával. A próbastatisztika mintából számított értéke a táblázatunk utolsó sorában szereplő értékek összege, azaz 4,3667.

3. *A kritikus tartomány kijelölése.*

A próba egyoldali  $\chi^2$ -próba. A kritikus értéket a  $\chi^2$ -eloszlás táblázatából olvashatjuk ki,  $r - 1 = 5 - 1 = 4$  szabadsági fok (amely az osztályok száma mínusz 1) és 0,95 valószínűség mellett. Ez a jegyzet  $\chi^2$ -eloszlás táblázatában a negyedik sor, első oszlopban található

$$\chi^2_t = 9,49$$

érték. (A kritikus tartomány ábráján a korábban megszokott jelölésekkel látható, hogy a próbastatisztika értéke az elfogadási tartományba esik.)

4. *A döntés.*

$4,3667 < 9,49$ , a próbastatisztika értéke az elfogadási tartományba esik, a nullhipotézist elfogadjuk 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján elfogadható, hogy az érmék szabályosak.

## 2. feladat

### (becsléses illeszkedésvizsgálat)

Egy focicsapat meccsenként lőtt góljainak számát az alábbi táblázat tartalmazza::

|                  |    |    |    |    |    |   |   |   |
|------------------|----|----|----|----|----|---|---|---|
| lőtt gólok száma | 0  | 1  | 2  | 3  | 4  | 5 | 6 | 7 |
| meccsek száma    | 14 | 18 | 29 | 18 | 10 | 7 | 3 | 1 |

Modellezhető-e a meccsenként lőtt gólok száma olyan Poisson-eloszlású valószínűségi változóval, melynek várható értéke a fenti értékekből számolt meccsenkénti góllátlag?

### Megoldás:

A feladat megoldása során több lépésre különösen oda kell figyelni, ezért érdemes alaposan áttanulmányozni a megoldást!

Elsőként vegyük észre a következőt: a véges mintánkat akarjuk modellezni egy olyan valószínűségi változóval, **amelynek végtelen sok lehetséges értéke van** (tudjuk, hogy egy Poisson-eloszlású valószínűségi változó tetszőleges nemnegatív egész értéket felvehet). Felmerül a kérdés, hogy ezt megtehetjük-e? A focicsapatunk hétnél több gólt egyetlen mérkőzésen sem lőtt. Ha Poisson-eloszlásúnak tekintjük a lőtt gólok számát, akkor viszont az eloszlásból a hétnél több lőtt gól valószínűsége nem nulla lesz. Mit tehetünk ilyen esetben? Megnézzük, hogy a Poisson-eloszlásból mekkora valószínűség-et kapunk a 7-nél több gólra. Ha **ez a valószínűség kellően kicsi (majdnem 0)**, akkor elhanyagolhatónak tekintjük. Ekkor modellezhetünk a végtelen eloszlással. (Emlékezzünk a valószínűség-számítás tanulmányainkból a binomiális eloszlás Poisson-eloszlással való közelítésére!)

A mintaelemek száma  $n = 100$  (összesen 100 mérkőzés eredményét osztályoztuk a táblázatban), a mintaelemeket a lőtt gólok száma szerint osztályoztuk,  $r = 8$  csoportba soroltuk.

A lőtt gólok számát tekintjük az  $\xi$  valószínűségi változónak. Azt feltételezzük, hogy  $\xi$  Poisson-eloszlású valószínűségi változó, melynek  $\lambda$  paraméterét a mintából számolt átlaggal becsüljük, így becsléses illeszkedésvizsgálatot

hajtunk végre. A feladat szerint a  $\xi$  várható értéke az összesen lőtt gólok számának és az összes megfigyelt mérkőzés számának hányadosa:

$$M(\xi) = \frac{0 \cdot 14 + 1 \cdot 18 + 2 \cdot 29 + \dots + 6 \cdot 3 + 1}{14 + 18 + 29 + \dots + 3 + 1} = \frac{230}{100} = 2,3$$

A hipotézisvizsgálat során azt kell eldöntenünk, hogy a megfigyelt valószínűségi változó (meccsenként lőtt gólok száma) a minta alapján tekinthető-e  $\lambda = 2,3$  paraméterű Poisson-eloszlásúnak. Ennek alapján a próba becsléses illeszkedésvizsgálat.

Először nézzük meg, a Poisson-eloszlásból mekkora valószínűségeket kapunk a lőtt gólok számának lehetséges értékeihez.

$$\begin{aligned} p_0 &= P(\xi = 0) = \frac{2,3^0}{0!} \cdot e^{-2,3} = 0,1003 \\ p_1 &= P(\xi = 1) = \frac{2,3^1}{1!} \cdot e^{-2,3} = 0,2306 \\ p_2 &= P(\xi = 2) = \frac{2,3^2}{2!} \cdot e^{-2,3} = 0,2652 \\ p_3 &= P(\xi = 3) = \frac{2,3^3}{3!} \cdot e^{-2,3} = 0,2033 \\ p_4 &= P(\xi = 4) = \frac{2,3^4}{4!} \cdot e^{-2,3} = 0,1169 \\ p_5 &= P(\xi = 5) = \frac{2,3^5}{5!} \cdot e^{-2,3} = 0,0538 \\ p_6 &= P(\xi = 6) = \frac{2,3^6}{6!} \cdot e^{-2,3} = 0,0206 \\ p_7 &= P(\xi = 7) = \frac{2,3^7}{7!} \cdot e^{-2,3} = 0,0068 \\ p_8 &= P(\xi > 7) = 1 - P(\xi \leq 7) = \\ &= 1 - (P(\xi = 0) + P(\xi = 1) + \dots + P(\xi = 7)) = 0,0025 \end{aligned}$$

A kiszámolt valószínűségek alapján meg tudjuk mondani, hogy mennyi az egyes osztályokba eső mintaelemek elvárt gyakorisága, azaz a Poisson-eloszlás fennállása esetén átlagosan hány mérkőzés esik az egyes osztályokba. Ehhez nincs más dolgunk, mint a fenti valószínűségeket szorozni a mintaelemek számával ( $np_i$ ) (tehát pl. 100 mérkőzésből átlagosan  $\mu_0 = 0,1003 \cdot 100 = 10,03$  esetben lesz a lőtt gólok száma 0). Azt látjuk azonban, hogy a 6 és 7 lőtt gólhoz tartozó elvárt gyakoriságok értékei nem érik el az ötöt ( $\mu_6 = 0,0206 \cdot 100 = 2,06$ ,  $\mu_7 = 0,0068 \cdot 100 = 0,68$ ). **Illeszkedésvizsgálatnál ugyanakkor nem lehet az elvárt gyakoriságok celláiban 5-nél kisebb szám,** mert az a próba képletében az összeadandó tagok szélsőségesen nagy értékei miatt potenciálisan torz eredményre vezetne. **Cellaösszevonással** érhetjük el, hogy a cellaértékeink megfelelőek legyenek. Módosítsuk úgy az eredeti táblázatot, hogy az utolsó oszlopokat összevonjuk. Nem elegendő csak az utolsó két oszlopot összevonni, mert a cellaérték az összevonás után is kisebb 5-nél ( $0,206 + 0,68 = 2,74$ ), így az utolsó három oszlopot vonjuk össze (az összevonást már a megfigyelt gyakoriságok kis értéke alapján is megtehettük volna, a 6 és 7 gólhoz tartozó mérkőzések kis száma miatt).

Ugyanakkor  $p_8 = 0,0025$  közel nulla (a legalább 8 gólhoz tartozó elvárt gyakoriság 0,25), így közelíthetünk végtelen eloszlással. Ezt a gyakoriságot is az utolsó (összevont) oszlophoz vonjuk.

Az áttekinthetőbb számoláshoz a kiindulási táblázatot egészítsük az alábbiak szerint:

|                                |         |         |        |        |        |            | Összeg |
|--------------------------------|---------|---------|--------|--------|--------|------------|--------|
| lőtt gólok száma               | 0       | 1       | 2      | 3      | 4      | legalább 5 |        |
| megfigyelt gyakoriság: $\mu_i$ | 14      | 18      | 29     | 18     | 10     | 11         |        |
| elvárt gyakoriság: $np_i$      | 10,03   | 23,06   | 26,52  | 20,33  | 11,69  | 8,37       |        |
| $(\mu_i - np_i)^2$             | 15,7609 | 25,6036 | 6,1504 | 5,4289 | 2,8561 | 6,9169     |        |
| $(\mu_i - np_i)^2 / (np_i)$    | 1,5713  | 1,1103  | 0,2319 | 0,2670 | 0,2443 | 0,8264     | 4,2512 |

A hipotézisvizsgálat lépései ezután:

1. *A nullhipotézis kimondása.*

A feltételezésünk most az, hogy a meccsenként lőtt gólok száma  $\lambda = 2,3$  paraméterű Poisson-eloszlású valószínűségi változó.

2. *A próbafüggvény kiválasztása és értékének kiszámítása.*

A próba illeszkedésvizsgálat,  $\chi^2$ -próbával. A próbastatisztika mintából számított értéke a táblázatunk utolsó sorában szereplő értékek összege, azaz 4,2512.

3. *A kritikus tartomány kijelölése.*

A próba egyoldali  $\chi^2$ -próba. Az eloszlás szabadsági fokára viszont figyelniünk kell! A feltételezett eloszlás paraméterét (várható érték) az adott mintából számolt értékkel (átlaggal) közelítettük. Ekkor a próba szabadsági foka még eggyel csökken, azaz nem osztályok száma (az összevont táblázatban!) mínusz egy, hanem osztályok száma mínusz 2 lesz!!! A kritikus értéket a  $\chi^2$ -eloszlás táblázatából olvashatjuk ki,  $6 - 2$  szabadsági fok és 0,95 valószínűség mellett. Ez a jegyzet  $\chi^2$ -eloszlás táblázatában a negyedik sor, első oszlopban található

$$\chi^2_t = 9,49$$

érték. (A kritikus tartomány ábráját felrajzolva a korábban megszokott jelölésekkel látható, hogy a kritikus érték az elfogadási tartományba esik.)

4. *A döntés.*

$4,2512 < 9,49$ , a próbastatisztika értéke az elfogadási tartományba esik, a nullhipotézist elfogadjuk, 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján modellezhető a meccsenként lőtt gólok száma olyan Poisson-eloszlással, melynek paramétere a mintából számolt átlag.

### 3. feladat (becsléses illeszkedésvizsgálat)

Egy gyártósornál rendszeresen 5 elemű mintát vesznek a termékekből. Egy hét alatt 500 mintát vettek. A mintákban talált selejtek gyakorisága az alábbi volt:

|                |     |     |     |    |   |   |
|----------------|-----|-----|-----|----|---|---|
| selejtek száma | 0   | 1   | 2   | 3  | 4 | 5 |
| gyakoriság     | 170 | 180 | 120 | 20 | 8 | 2 |

Modellezhető-e a mintában levő selejtek száma olyan binomiális eloszlással, melynek várható értéke a fentiekből számolt átlag?

**Megoldás:**

Az egyes mintákban levő selejtek számát tekintjük az  $\xi$  valószínűségi változónak. Azt feltételezzük, hogy  $\xi$  binomiális-eloszlású valószínűségi változó, melynek **várható értékét** a mintából(!) számított **átlaggal becsüljük** (becsléses illeszkedésvizsgálatot hajtunk végre) oly módon, hogy az összes kihúzott selejt számát osztjuk a kísérletek számával:

$$M(\xi) = \frac{0 \cdot 14 + 1 \cdot 180 + 2 \cdot 120 + 3 \cdot 20 + 4 \cdot 8 + 5 \cdot 2}{170 + 180 + 120 + 20 + 8 + 2} = \frac{522}{500} = 1,044$$

A mintaelemek száma  $n = 500$  (500 kísérletet hajtottunk végre), a mintaelemeket a kihúzott selejtek száma szerint osztályoztuk, ezzel  $r = 6$  csoportba soroltuk.

Ha  $\xi$ -t binomiális eloszlással akarjuk modellezni, ismernünk kell a binomiális eloszlás két paraméterét, az  $n$  kísérlet-számot, amely 5, illetve a megfigyelt esemény (selejthúzás)  $p$  valószínűségét. Ez utóbbi nem ismert, de a binomiális eloszlás várható értékéből becsülhetjük: az  $1,044 = M(\xi) = n \cdot p$  egyenletből

$$p = \frac{M(\xi)}{n} = \frac{1,044}{5} = 0,2088$$

Az illeszkedésvizsgálathoz először meghatározzuk, hogy ha  $\xi$  binomiális eloszlású az  $n = 5$ ,  $p = 0,2088$  paraméterekkel, akkor az egyes osztályokba átlagosan hány mintaelemet várunk. Számítsuk ki a kihúzott selejtek számának lehetséges értékeihez tartozó valószínűségeket! (Kiszámoljuk, hogy mekkora annak a valószínűsége, hogy az 5 elemű mintában éppen  $k$  darab selejt van, ahol  $k = 0, 1, 2, 3, 4, 5$ . Célszerű minél több tizedes pontossággal számolni!).

$$\begin{aligned}
p_0 &= P(\xi = 0) = 0,7912^5 = 0,31005 \\
p_1 &= P(\xi = 1) = 0,2088^1 \cdot 0,7912^4 \binom{5}{1} = 0,40912 \\
p_2 &= P(\xi = 2) = 0,2088^2 \cdot 0,7912^3 \binom{5}{2} = 0,21593 \\
p_3 &= P(\xi = 3) = 0,2088^3 \cdot 0,7912^2 \binom{5}{3} = 0,05699 \\
p_4 &= P(\xi = 4) = 0,2088^4 \cdot 0,7912^1 \binom{5}{4} = 0,00752 \\
p_5 &= P(\xi = 5) = 0,2088^5 = 0,000397
\end{aligned}$$

A fenti valószínűségekkel szorozzuk a mintaelemek számát, így megkapjuk az  $i$ -edik osztályhoz tartozó elvárt gyakoriságot ( $np_i$ ), (így pl. az 500 kísérletből átlagosan  $p_1 = 500 \cdot 0,40912 = 204,56$  esetben kapunk olyan eredményt, ahol az 5 kihúzott termékből pontosan egy selejtes). Az elvárt gyakoriságokat beírjuk a táblázatba:

|                       |        |        |        |       |      |     |
|-----------------------|--------|--------|--------|-------|------|-----|
| selejtelek száma      | 0      | 1      | 2      | 3     | 4    | 5   |
| megfigyelt gyakoriság | 170    | 180    | 120    | 20    | 8    | 2   |
| elvárt gyakoriság     | 155.03 | 204.56 | 107.97 | 28.49 | 3.75 | 0.2 |

A kapott elvárt gyakoriságok között azonban két 5-nél kisebb érték is szerepel, ami illeszkedésvizsgálatnál nem megengedett! Tehát cellákat kell összevonnunk. Mivel az utolsó két osztályban a várt gyakoriságok összege ( $3,75 + 0,2 = 3,95$ ) kisebb 5-nél, ezért az utolsó három osztályt vonjuk össze. Ezzel az osztályok száma 4-re csökkent. Ezután a táblázatot az alábbiak szerint kitöltjük:

|                              |          |          |          |             |        |
|------------------------------|----------|----------|----------|-------------|--------|
| selejtelek száma             | 0        | 1        | 2        | 3 v. 4 v. 5 | összes |
| megfigyelt gyakoriság: $\mu$ | 170      | 180      | 120      | 30          |        |
| elvárt gyakoriság: $np_i$    | 155,03   | 204,56   | 107,97   | 32,44       |        |
| $(\mu_i - np_i)^2$           | 224,1009 | 603,1936 | 144,7209 | 5,9536      |        |
| $(\mu_i - np_i)^2 / (np_i)$  | 1,4455   | 2,9487   | 1,3404   | 0,1835      | 5,9181 |

A hipotézisvizsgálat lépései ezután az alábbiak:

1. A nullhipotézis kimondása.

A feltételezésünk most az, hogy  $\xi$  binomiális eloszlású az  $n = 5$  és  $p = 0,2088$  paraméterekkel.

2. A próbafüggvény kiválasztása és értékének kiszámítása.

A próba illeszkedésvizsgálat,  $\chi^2$ -próba. A próbastatisztika mintából számított értéke a táblázatunk utolsó sorában szereplő értékek összege, azaz 5,9181.

3. A kritikus tartomány kijelölése.

A próba egyoldali  $\chi^2$ -próba. A kritikus értéket a  $\chi^2$ -eloszlás táblázatból olvashatjuk ki. A szabadsági fok a várható érték mintából való becslése miatt osztályok száma-2, azaz  $4 - 2 = 2$ , a szignifikancia szint nem volt megadva, ezért a valószínűséget a szokásos 0,95-nek tekintjük. Ez a jegyzet  $\chi^2$ -eloszlás táblázatában a második sor, első oszlopban található

$$\chi_t^2 = 5,99$$

érték. (A kritikus tartomány ábráját felrajzolva látható, hogy a kritikus érték az elfogadási tartományba esik.)

4. A döntés.

$5,9181 < 5,99$ , a próbastatisztika értéke az elfogadási tartományba esik, a nullhipotézist elfogadjuk 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján elfogadható, hogy dobott fejek száma binomiális eloszlású,  $n = 5$ ,  $p = 0,2088$  paraméterekkel.

Ugyanakkor vegyük észre, hogy a számított és a kritikus érték nagyon közel van egymáshoz!