Nemparaméter próbák

Nemparaméteres próba alatt olyan próbát értünk, amelyben a hipotéziseink nem az eloszlás valamely paraméterére vonatkoznak, hanem például arra, hogy két valószínűségi változó azonos eloszlású vagy független-e, adataink származhatnak-e egy adott eloszlásból, stb. Ezek a próbák egy olyan statisztikán alapulnak, amelyek eloszlása a nullhipotézis fennállása esetén független az eloszlásfüggvénytől.

A nemparaméteres próbák száma napjainkban egyre növekszik. Mi itt csak a leggyakrabban használt, ún. χ^2-próbát tárgyaljuk részletesen.

Először a χ^2-eloszlás fogalmát ismételjük át:

Legyenek $X_1, X_2, \ldots, X_r$ független standard normális eloszlású v.v-k. Ekkor a

$$Y = X_1^2 + X_2^2 + \cdots + X_r^2$$

v.v-t r-szabadságfokú χ^2-eloszlású v.v-nal mondjuk, a jelölése: $\chi^2(r)$

A $\chi^2(r)$ eloszlású v.v. eloszlásfüggvényének értekei táblázatból számolhatók.

A következő tétel magyarázza el, hogy miért olyan fontos ez az eloszlás:

Tétel Legyen X egy olyan v.v, melynek lehetséges kimenetelei:

$$x_1, x_2, \ldots, x_r.$$

Legyen

$$P(X = x_i) = p_i.$$

Végezzünk n kisérletet az X v.v-ra, jelölje v_i az $X = x_i$ esemény relatív gyakoriságát. Ekkor ha n értéke nagy, akkor

$$\sum_{i=1}^{r} \frac{(v_i - np_i)^2}{np_i}$$

véletlen mennyiség eloszlásfüggvénye jól közelíthető egy $(r-1)$-szabadságfokú χ^2-eloszlású v.v eloszlásfüggvényével.

Illeszkedésvizsgálat χ^2-próbával

Tiszta illeszkedésvizsgálat

Diszkrét eset

Legyen X egy olyan v.v, melynek lehetséges kimenetelei:

$$x_1, x_2, \ldots, x_r.$$

A hipotézisünk a következő:

$$H_0 : P(X = x_i) = p_i, \quad i = 1, 2, \ldots, r$$

H_1 pedig ennek tagadása.

Végezzünk n kisérletet az X v.v-ra, jelölje v_i az $X = x_i$ esemény relatív gyakoriságát. Ekkor a χ^2-tel jelölt próbastatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{(v_i - np_i)^2}{np_i}$$

ϵ elsőfajú hiba mellett akkor fogadjuk el a H_0-t, ha

$$\chi^2 < \chi^2_\epsilon(r-1)$$

és akkor utasítjuk el, ha

$$\chi^2 \geq \chi^2_\epsilon(r-1)$$

Példák: Vizsgáljuk meg, hogy szabályosnak tekinthető-e az a dobókocka, melyet $n = 1200$-szor feldobva az egyes oldalak gyakoriságai

$$v_1 = 184, \quad v_2 = 212, \quad v_3 = 190,$$

$$v_4 = 208, \quad v_5 = 212, \quad v_6 = 194.$$

A döntésünket $\epsilon = 0,05$ elsőfajú hiba mellett hozzuk meg.

Megoldás: A nullhipotézis:

$$H_0 : P(X = i) = \frac{1}{6} \quad i = 1, 2, 3, 4, 5, 6$$

azaz $p_i = \frac{1}{6}$, ezért $np_i = 200$. Így a próbastatisztika:

$$\chi^2 = \sum_{i=1}^{6} \frac{(v_i - 200)^2}{200} =$$

$$\frac{16^2 + 12^2 + 10^2 + 8^2 + 12^2 + 6^2}{200} = 3,72.$$

$r = 6$, ezért

$$\chi^2_{0,05}(r-1) = \chi^2_{0,05}(5) = 11,1,$$

ezért

$$\chi^2 < \chi^2_{0,05}(5)$$

így elfogadjuk a H_0-t.

Abszolút folytonos eset

Legyen X egy folytonos eloszlású v.v. A

$$H_0 : P(X < x) = F(x), \quad x \in \mathbb{R}$$

nullhipotézist teszteljük adott $F(x)$ esetén. Az ellenhipotézis tehát az, hogy nem $F(x)$ az eloszlásfüggvény.

Végezzünk n független kisérletet az X v.v-ra: az eredmény $X_1, X_2, \ldots, X_n$. Osszuk fel a számegyenest

$$-\infty = x_0 < x_1 < \cdots < x_{r-1} < x_r = \infty$$

osztópontokkal r rszre. A felosztást úgy végezzük, hogy minden részbe essen néhány (legalább 3) kisérleti eredmény. Jelölje v_i azon kisérletek számát, amikor a kisérlet eredménye az $[x_{i-1}, x_i[$ intervallumba esik, azaz

$$v_i = |\{k : X_k \in [x_{i-1}, x_i[\}|$$

Legyen

$$p_i = P(X \in [x_{i-1}, x_i[) = F(x_i) - F(x_{i-1})$$

A próbastatisztika változatlanul:

$$\sum_{i=1}^{r} \frac{(v_i - np_i)^2}{np_i}$$

és akkor fogadjuk el a H_0-t, ha

$$\chi^2 < \chi^2_\epsilon(r-1)$$

és akkor utasítjuk el, ha

$$\chi^2 \geq \chi^2_\epsilon(r-1).$$

Példa: Tekintsük az alábbi megfigyeléssort (az adatokat az egyszerűség kedvéért nagyság szerint rendezve adjuk meg):

$$0,03, \quad 0,05, \quad 0,08, \quad 0,08, \quad 0,11, \quad 0,12, \quad 0,14$$

$$0,19, \quad 0,23, \quad 0,24, \quad 0,31, \quad 0,43, \quad 0,45, \quad 0,56$$

$$0,61, \quad 0,67, \quad 0,69, \quad 0,78, \quad 0,85, \quad 0,88$$

$\epsilon = 0,05$ elsőfajú hiba mellett döntsön arról a hipotézisról, hogy a minta a $[0,1]$ intervallumon egyenletes eloszlásból származik-e (RND-e)!

Megoldás: Osszuk fel a $[0,1]$ intervallumot 4 egyenlő részre

$[0, \frac{1}{4}[$	$[\frac{1}{4}, \frac{2}{4}[$	$[\frac{2}{4}, \frac{3}{4}[$	$[\frac{3}{4}, 1]$
$v_1 = 10$	$v_2 = 3$	$v_3 = 4$	$v_4 = 3$

Ekkor a próbastatisztika:

$$\chi^2 = \frac{(10-5)^2 + (3-5)^2 + (4-5)^2 + (3-5)^2}{5} = 6,8$$

másrészt $r = 4$, így

$$\chi^2_\epsilon(r-1) = \chi^2_{0,05}(3) = 7,81$$

ezért

$$\chi^2 < \chi^2_\epsilon(r-1),$$

így elfogadjuk H_0-t, a minta az RND-ből származik.

Amennyiben a diszkrét esetben a valószínűségek, folytonos esetben az eloszlásfüggvény becsültek (úgy értve, hogy az ismert jellegű eloszlás paramétereit előzetesen ugyanabból a mintából becsültük), akkor a próbastatisztika számlása a korábban megismert módon történik, csak a kritikus értéket adó χ^2 eloszlás szabadsági foka $r - 1 - b$ lesz, ahol b a becsült paraméterek száma. A paraméterek becslését maximum-likelihood módszerrel kell elvégezni.

Példa: 1000 négygyermekes családban megvizsgálták a fiúk számát és a következőt kapták:

Gyermekek száma	0	1	2	3	4
Családok száma	56	223	401	251	69

Döntsünk $\epsilon = 0,05$ elsőfajú hibavalószínűség mellett arról, hogy a mintánk binomális eloszlásból származik-e!

Először a binomiális eloszlás paraméterét becsüljül meg:

$$\widehat{p} = \frac{\text{fiúk száma}}{\text{gyerekek száma}} =$$

$$\frac{56 * 0 + 223 * 1 + 401 * 2 + 251 * 3 + 69 * 4}{4 * 1000} = 0,5135$$

Legyen X egy n=4, $\widehat{p} = 0,5135$ paraméterű binomiális eloszlású v.v. Ekkor

$$p_0 = P(X = 0) = \binom{4}{0}(1 - 0,5135)^4 0,5135^0 = 0,056$$

$$p_1 = P(X = 1) = \binom{4}{1}(1 - 0,5135)^3 0,5135^1 = 0,2365$$

$$p_2 = P(X = 2) = \binom{4}{2}(1 - 0,5135)^2 0,5135^2 = 0,3744$$

$$p_3 = P(X = 3) = \binom{4}{3}(1 - 0,5135)^1 0,5135^3 = 0,2635$$

$$p_4 = P(X = 4) = \binom{4}{4}(1 - 0,5135)^0 0,5135^4 = 0,0695$$

Így a próbastatisztika:

$$\chi^2 = \sum_{i=0}^{4} \frac{(v_i - np_i)^2}{np_i} =$$

$$\frac{(56-56)^2}{56} + \frac{(223-236,5)^2}{236,5} + \frac{(401-374,4)^2}{374,4} +$$

$$\frac{(251-263,5)^2}{263,5} + \frac{(69-69,5)^2}{69,5} = 3,257$$

Másrészt $r = 5$ és a becsült paraméterek száma 1, ezért most a

$$\chi^2_\epsilon(r-1-1) = \chi_{0,05}(3) = 7,815,$$

ezért

$$\chi^2 < \chi^2_\epsilon(r-1-1)$$

így elfogadjuk, hogy a gyerekek száma binomiális eloszlást követ.

Tegyük fel, hogy egy gyáregységben az alkalmazottak havi fizetése (ezer Ft-ban) a következőképpen alakul:

$$25, \quad 26, \quad 18, \quad 20, \quad 51, \quad 45, \quad 80, \quad 20, \quad 30, \quad 25,$$

$$250, \quad 46, \quad 38, \quad 142, \quad 40, \quad 21, \quad 25, \quad 26, \quad 52, \quad 90,$$

$$44, \quad 21, \quad 24, \quad 40, \quad 51$$

Döntsön $\epsilon = 0,01$ elsőfajú hibavalószínűség mellett arról, hogy a bérek normális eloszlásból származnak-e.

Először a normális eloszlás paramétereit: várható értékét és szórását becsüljük meg az átlaggal ill. a korrigált tapasztalati szórással:

$$\widehat{m} = \frac{25 + 26 + \cdots + 51}{25} = 50$$

$$(S^*)^2 = \frac{(25-50)^2 + (26-50)^2 + \cdots + (51-50)^2}{25-1} = 2500$$

így

$$S^* = 50$$

A valós számegyenest 5 részre bontjuk és ott megnézzük a gyakoriságot ill az $\widehat{m} = 50$, $\widehat{\sigma} = 50$ paraméterű normális eloszlású v.v. intervallumba esésének valószínűségét:

intervallumok	$]-\infty, 20]$	$]20, 40]$
a vszg. képlete	$\Phi(-0,6)$	$\Phi(-0,2) - \Phi(-0,6)$
a vszg. értéke	0,2744	0,1464
elméleti gyak. (np_i)	6,86	3,66
megfigyelt gyak. (v_i)	3	12

intervallumok	$]40,60]$	$]60,80]$
a vszg. képlete	$\Phi(0,2)-\Phi(-0,2)$	$\Phi(0,6)-\Phi(0,2)$
a vszg. értéke	0,1584	0,1464
elméleti gyak. (np_i)	3,96	3,66
megfigyelt gyak. (v_i)	6	1

intervallumok	$]80,\infty]$
a vszg. képlete	$1-\Phi(0,6)$
a vszg. értéke	0,2744
elméleti gyak. (np_i)	6,86
megfigyelt gyak. (v_i)	3

Igy

$$\chi^2 = \sum_{i=1}^{5} \frac{(v_i - np_i)^2}{np_i} = 26,33$$

másrészt most $r = 5$ és két paramétert becsültünk, ezért

$$\chi^2_\epsilon(r-1-2) = \chi^2_{0,01}(2) = 9,21$$

értékhez viszonyítunk, amire

$$\chi^2 > \chi^2_\epsilon(r-1-2),$$

ami alapján elutasítjuk H_0-t, nem normális eloszlásból származnak az adatok.

Homogenitásvizsgálat χ^2-próbával

Azt akarjuk megvizsgálni, hogy két v.v tekenthető-e azonos eloszlásúnak. Legyen $X_1, X_2, \ldots, X_n$ az X háttérváltozóból származó és $Y_1, Y_2, \ldots, Y_m$ az Y háttérváltozóból származó minta.

Kérdés: azonos eloszlású háttérváltozóból származnak-e.

Például két betegcsoportnál a vérnyomás azonos eloszlású-e; két nagy évfolyam szigorlati eredményei azonosak-e?

A

$$H_0 : P(X < x) = P(Y < x), \text{ minden } x \in \mathbb{R}$$

nullhipotézist teszteljük a tagadása mint alternativa ellen.

Diszkrét eset

Ha X és Y eloszlása diszkrét, akkor H_0 fennállása csak akkor kérdéses, ha azonos az értékkészletük: $x_1, \ldots, x_r$. Legyen

$$v_i = |\{k : X_k = x_i\}$$

$$u_i = |\{k : Y_k = x_i\}$$

A próbastatisztika:

$$\chi^2 = nm \sum_{i=1}^{r} \frac{(\frac{v_i}{n} - \frac{u_i}{m})^2}{u_i + v_i}$$

Adott ϵ elsőfajú hiba esetén ezt kell $\chi^2_\epsilon(r-1)$-gyel összehasonlítani.

Példa: Két dobokockával dobva az alábbi gyakoriságokat figyeltük meg:

	1	2	3	4	5	6
1. kocka	27	24	26	23	18	32
2. kocka	18	12	15	21	14	20

$\epsilon = 0,05$ elsőfajú hiba mellett döntsön arról, hogy a két kocka tekinthető-e azonos eloszlásúnak!

Megoldás: $n = 150$ és $m = 100$, a próbastatisztika

$$\chi^2 = 150 \cdot 100 \sum_{i=1}^{6} \frac{(\frac{u_i}{150} - \frac{v_1}{100})^2}{u_i + v_i} = 2,2$$

és most

$$\chi^2_\epsilon(r-1) = \chi^2_{0,05}(5) = 11,1,$$

tehát

$$\chi^2 < \chi^2_\epsilon(r-1),$$

ezért elfogadjuk, hogy azonos eloszlásúak.

Folytonos eset

Ha X és Y folytonos eloszlású v.v, akkor az illeszkedésvizsgálatnál tanult módon felosztjuk a számegyenest r részre a

$$-\infty = x_0 < x_1 < \cdots < x_{r-1} < x_r = \infty$$

osztópontokkal, és legyen

$$v_i = |\{k : X_k \in [x_{i-1}, x_i[\}|$$

és

$$u_i = |\{k : y_k \in [x_{i-1}, x_i[\}|$$

A próbastatisztika változatlanul

$$\chi^2 = nm \sum_{i=1}^{r} \frac{(\frac{v_i}{n} - \frac{u_i}{m})^2}{u_i + v_i}$$

Függetlenségvizsgálat χ^2-próbával

Azt szeretnénk megtudni, hogy a kétdimenziós (X, Y) valószínűségi változó komponensei tekinthetők-e függetlennek az $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$ minta alapján.

Például független-e a testmagasság a testsúlytól; a szemszín a hajszíntől; középiskolások matematikai osztályzatai függetlenül alakulnak-e a történelem osztályzattól?

Mind a disztkrét, mind a folytonos esetben találni fogunk olyan $A_1, \ldots, A_r$ illetve $B_1, \ldots, B_s$ teljes eseményrendszereket (melyek X illetve Y értékkészletére vonatkoznak), hogy a nullhiputézisünk az alábbi alakban fogalmazható meg:

$$H_0 : P(A_i B_j) = P(A_i)P(B_j), \qquad \text{minden } i = 1, \ldots, r, j = 1, \ldots, s$$

az alternativa, hogy ez nem teljesül legalább egy i, j párra.

Becsléses függetlenségvizsgálat: Ha X és Y elméleti eloszlása nem ismert

Diszkrét eset

Ha X és Y diszkrét eloszlásúak, akkor jelölje $x_1, \ldots, x_r$ illetve $y_1, \ldots, y_s$ az értékkészletüket és legyen

$$v_{i.} = |\{k : X_k = x_i\}|, \quad i = 1, \ldots, r$$

$$v_{.j} = |\{k : Y_k = y_j\}|, \quad j = 1, \ldots, s$$

$$v_{ij} = |\{k : X_k = x_i, Y_k = y_j\}|, \quad i = 1, \ldots, r, j = 1, \ldots s.$$

Az ezekkel a mennyiségekkel képzett próbastatisztika:

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(v_{ij} - \frac{v_{i.} v_{.j}}{n})^2}{v_{i.} v_{.j}},$$

amit a $\chi^2_\epsilon((r-1)(s-1))$ értékkel kell összehasonlítani.

Példa: Az alábbi táblázat mutatja, hogy egy évfolyamon hányan kapták az egyes jegyeket statisztikából aszerint csoportosítva a diákokat, hogy sportolnak-e rendszeresen. $\epsilon = 0,05$ elsőfajú hiba mellett döntsünk arról, hogy független-e a statisztika jegy a sportolástól!

jegyek	2	3	4	5
nem sportolnak	50	35	35	20
sportolnak	10	15	25	10

Megoldás: A sportolás kérdésnél két eset van, ezért $r = 2$, míg az osztályzatoknál 4 eset lehet, emiatt $s = 4$. Az ún. koningenciatáblázat a következő:

jegyek	2	3	4	5	
nincs sport	50	35	35	20	$v_{1.} = 140$
sportolnak	10	15	25	10	$v_{2.} = 60$
	$v_{.1} = 60$	$v_{.2} = 50$	$v_{.3} = 60$	$v_{.4} = 30$	

Ekkor

$$\chi^2 = n \sum_{i=1}^{2} \sum_{j=1}^{4} \frac{(v_{ij} - \frac{v_{i.} v_{.j}}{n})^2}{v_{i.} v_{.j}} =$$

$$200 \left(\frac{(50 - \frac{140*60}{200})^2}{140 * 60} + \frac{(35 - \frac{140*50}{200})^2}{140 * 50} + \cdots + \frac{(10 - \frac{60*30}{200})^2}{60 * 30} \right) =$$

$$9,13$$

és

$$\chi^2_\epsilon((r-1)(s-1)) = \chi^2_{0,05}(3) = 7,81,$$

tehát elutasítjuk, hogy a sportolás és a jegyek függetlenek lennének

Folytonos eset

Vesszük a szokásos felosztását a valós számegyenesnek és legyen

$$v_{i.} = |\{k : X_k \in [x_{i-1}, x_i[\}|, \quad i = 1, \ldots, r$$

$$v_{.j} = |\{k : Y_k \in [y_{j-1}, y_j[\}|, \quad j = 1, \ldots, s$$

A többi megegyezik a diszkrét esettel.

Az alábbi táblázat három - a TV-ben különböző intenzitással reklámozott - fogkrém fogyasztására vonatkozó adatokat tartalmaz a TV-nézés idejének függvényében:

fogkrém fajta TV nézés hetente	A 1 óra reklám	B 5 perc reklám	C 0 perc reklám
5 óránál kevesebb	80	60	60
5-15 óra közt	70	70	60
15 óra felett	90	65	45

Döntsön $\epsilon = 0,1$ elsőfajú hiba mellet arról, hogy van-e összefüggés a kedvelt fogkrém márkája és a TV-nézés időtartama közt!

Megoldás: A kontingenciatáblázat:

80	60	60	$v_{1.} = 200$
70	70	60	$v_{2.} = 200$
90	65	45	$v_{3.} = 200$
$v.1 = 240$	$v_{.2} = 195$	$v_{.3} = 165$	

Ezekkel az értékekkel:

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(v_{ij} - \frac{v_{i.} v_{.j}}{n})^2}{v_{i.} v_{.j}} = 6$$

A feladatban $r = 3$ és $s = 3$, ezért

$$\chi^2_{\epsilon}((r-1)(s-1)) = \chi^2_{0,1}(4) = 7,78$$

tehát nem állíthatjuk, hogy van a kedvelt fogkrém az a TV nézés közt összefüggés.