Nemparaméteres és paraméteres illeszkedésvizsgálat

– egy MATLAB® alapú megközelítés –

Baja Zsolt, Vas Orsolya

Matematika és Informatika Intézet, Babeș–Bolyai Tudományegyetem, Kolozsvár, Románia

(bajazsolt98@gmail.com, vas.orsolya@yahoo.com)

12. labor / 2022. december 19–23.

- A statisztikai sokaságokat általában valamely Y kvalitatív (minőségi), vagy kvantitatív (mennyiségi) jellemző alapján tanulmányozzuk.
- Legyen k az Y jellemző osztályainak száma és jelölje E_i azt az eseményt, miszerint a sokaságnak egy találomra kiválasztott egyede éppen az i-edik osztályhoz $(i = 1, 2, \ldots, k)$ tartozik!
- Nyilván, a $p_i = P(E_i)$ jelöléseket $(i = 1, 2, \ldots, k)$ használva, a $\sum_{i=1}^{k} p_i \equiv 1$ azonosságot kapjuk.
- A továbbiakban azt feltételezzük, hogy az ismeretlen elméleti $\{p_i\}_{i=1}^{k}$ valószínűségek egy bizonyos – általunk helyesnek tartott – elméleti eloszlásnak megfelelően viselkednek.
- Feltételezésünk helyességének eldöntése érdekében a

$$H_0 : p_i = p_i^{(0)},\ i = 1, 2, \ldots, k \tag{1}$$

nullhipotézist és a

$$H_1 : \exists i_0 \in \{1, 2, \ldots, k\} : p_{i_0} \neq p_{i_0}^{(0)} \tag{2}$$

alternatív hipotéziseket fogalmazhatjuk meg, ahol $\left\{p_i^{(0)}\right\}_{i=1}^{k}$ az osztályok általunk feltételezett elméleti bekövetkezési valószínűségét jelöli.

- Tekintsük a vizsgált sokaságnak egy n-elemű független $\{Y_j\}_{j=1}^{n}$ mintavételét!
- Ekkor meghatározhatjuk az osztályok $\{n_i\}_{i=1}^{k}$ abszolút gyakoriságát, vagyis megszámolhatjuk, hogy az adott mintavétel esetén hányszor következtek be az $\{E_i\}_{i=1}^{k}$ események. Az így kapott abszolút gyakoriságokhoz az $\{N_i\}_{i=1}^{k}$ valószínűségi változókat társíthatjuk.
- Tehát a vizsgált Y kvalitatív, vagy kvantitatív tulajdonságból kiindulva, egy k-állapotú, multinomiális eloszlású $N(N_1, N_2, \ldots, N_k)$ valószínűségi vektort kaptunk, amely eloszlását a

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \cdot {p_1}^{n_1} {p_2}^{n_2} \cdot \ldots \cdot {p_k}^{n_k} \qquad (3)$$

valószínűség adja, ahol $n = \sum_{i=1}^{k} n_i$, $0 \leq n_i \leq n$, $\sum_{i=1}^{k} p_i \equiv 1$.

- Összegezve a fentieket, beláthatjuk, hogy az (1)-es nullhipotézis valójában egy multinomiális eloszlású valószínűségi vektor paraméterezésére ad tippet.
- Nagy elemszámú mintavételek esetén a (3)-as képlet kiértékelése túlságosan költségesnek bizonyul. Így hatékonyabb módszerre van szükségünk a feltevésünk ellenőrzésére. Ezzel kapcsolatos a következő tétel.

1. Tétel (Az illeszkedést ellenőrző χ^2-próbák valószínűségi változója)

Az előző oldalakon bevezetett jelöléseket használva a

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \tag{4}$$

valószínűségi változó $\chi^2\,(k-1,1)$-eloszlást követ, amikor $n \to \infty$.

Bizonyítás

Az állítás igazolását megtalálhatjátok a nyomtatott jegyzet 199–201., illetve az elektronikus változat 206–208. oldalain!

- Amint hamarosan látni fogjuk, a (4)-es $\chi^2\,(k-1,1)$-eloszlású valószínűségi változót már kényelmesen használhatjuk mind diszkrét, mind folytonos eloszlások *illeszkedésvizsgálatára*.
- A továbbiakban ismertetendő hipotézisellenőrző próbák esetén különbséget teszünk aszerint, hogy a nullhipotézisben feltételezett elméleti eloszlás minden paraméterét, vagy azoknak csak egy részét, vagy esetleg egyiket sem ismerjük. Amennyiben minden paramétert ismertnek feltételezünk, akkor az illeszkedésvizsgálatot *nemparaméteresnek*, máskülönben *paraméteresnek* nevezzük.
- Paraméteres illeszkedésvizsgálat esetén a rendelkezésünkre álló n-elemű, független $\left\{Y_j\right\}_{j=1}^{n}$ mintavételből adhatunk *pontbecslést* az ismeretlen paraméterekre úgy, hogy *a mintavételi elemekből alkotott folytonos, vagy diszkrét valószínűségi vektor együttes sűrűség-, vagy relatív gyakoriság függvényét maximalizáljuk az ismeretlen paraméterek szerint*. Az így megbecsült paraméterek száma tovább csökkenti a (4)-es valószínűségi változó szabadságfokát, pontosabban fogalmazva, ha az optimalizációs folyamat során összesen $0 \leq \ell < k-1$ paramétert határoztunk meg, akkor a (4)-es χ^2-eloszlású valószínűségi változó szabadságfoka $k-1-\ell$ lesz!
- Ezt a pontbecslési eljárást a *maximum-likelihood módszerének* nevezzük, és azért ad optimális közelítést az ismeretlen paraméterekre, mert, ha a maximalizáció során kapott paramétereket helyettesítjük az együttes sűrűség-, vagy relatív gyakoriság függvénybe, akkor azt érjük el, hogy a rendelkezésünkre álló független $\left\{Y_j\right\}_{j=1}^{n}$ mintavétel bekövetkezése lesz a legvalószínűbb az összes lehetséges n-elemű mintavételek közül.

- A továbbiakban tekintsünk néhány megodott példát a maximum-likelihood pontbecslési módszer alkalmazására!

1. Megoldott feladat (Exponenciális eloszlás)

A maximum-likelihood módszerével adjunk pontbecslést az $\mathcal{E}xp(\lambda)$-eloszlás ismeretlen $\lambda > 0$ paraméterére!

Megoldás

- Tekintsük az adott eloszlásnak egy n-elemű, független $\{Y_j\}_{j=1}^n$ mintavételét! Felhasználva az $\{Y_j\}_{j=1}^n$ valószínűségi változók függetlenségét, valamint az $\mathcal{E}xp(\lambda)$-eloszlás sűrűségfüggvényét, az $Y(Y_1, Y_2, \ldots, Y_n)$ valószínűségi vektor együttes sűrűségfüggvényére az

$$f_Y(\lambda; y_1, y_2, \ldots, y_n) = \prod_{j=1}^{n} \left(\lambda e^{-\lambda y_j}\right) = \lambda^n e^{-\lambda \sum_{j=1}^{n} y_j}, y_1 > 0, y_2 > 0, \ldots, y_n > 0$$

  kifejezést kapjuk, amelyet a továbbiakban a λ paraméter szerint maximalizálunk.

- Az egyszerűség kedvéért, a továbbiakban elhagyjuk az $y_1, y_2, \ldots, y_n$ változók felsorolását az együttes sűrűségfüggvény argumentumából, mert az optimalizáció során amúgy is konstansként kezeljük ezeket az értékeket.

Megoldás – folytatás

- Vegyük észre, hogy az $f_Y(\lambda)$ függvény ugyanazon szélsőértékhelyekre veszi fel az optimumait, mint a jóval egyszerűbben kezelhető $\ln(f_Y(\lambda))$ függvény.
- Ezért maximumpont keresése végett, elégséges a

$$\begin{cases} \dfrac{\partial}{\partial\lambda}\ln(f_Y(\lambda)) & = & 0, \\ \dfrac{\partial^2}{\partial\lambda^2}\ln(f_Y(\lambda)) & < & 0 \end{cases}$$

  rendszert megoldanunk.
- A rendszer első feltételét a

$$\frac{\partial}{\partial\lambda}\left(n\ln(\lambda)-\lambda\sum_{j=1}^{n}y_j\right)=\frac{n}{\lambda}-\sum_{j=1}^{n}y_j=0$$

  alakra hozhatjuk, amelyből a

$$\overline{\lambda}=\frac{n}{\sum_{j=1}^{n}y_j}>0$$

  szélsőértékhelyet kapjuk.

Megoldás – folytatás

- Mivel
$$\frac{\partial^2}{\partial\lambda^2}\ln\left(f_Y\left(\lambda\right)\right)\Big|_{\lambda=\overline{\lambda}} = -\frac{n}{\lambda^2}\Big|_{\lambda=\overline{\lambda}} = -\frac{n}{\overline{\lambda}^2} < 0,$$
következik, hogy a $\lambda = \overline{\lambda}$ szélsőértékhely valóban maximumpontot határoz meg.
- Tehát az azonos $\mathcal{E}xp\left(\lambda\right)$-eloszlású és független $\left\{Y_j\right\}_{j=1}^n$ minták esetén az eloszlás ismeretlen elméleti λ paraméterének közelítésére a
$$\Lambda = \Lambda\left(Y_1, Y_2, \ldots, Y_n\right) = \frac{n}{\sum_{j=1}^n Y_j} = \frac{1}{\frac{1}{n}\sum\limits_{j=1}^n Y_j} = \frac{1}{\overline{Y}}$$
becslőképletet (másképpen statisztikát) használhatjuk.

2. Megoldott feladat (Binomiális eloszlás)

A maximum-likelihood módszerével adjunk pontbecslést a $\mathcal{B}ino\,(m, p)$-eloszlás ismeretlen $p \in (0,1)$ paraméterére, ahol $m \geq 1$ ismert természetes szám!

Megoldás

- Tekintsük az adott eloszlásnak egy n-elemű, független $\{Y_j\}_{j=1}^{n}$ mintavételét!
- Az $\{Y_j\}_{j=1}^{n}$ valószínűségi változók függetlensége és a $\mathcal{B}ino\,(m, p)$-eloszlás

$$f\,(p; m, y) = \binom{m}{y} p^y\,(1-p)^{m-y}\,,\ y \in \{0, 1, \ldots, m\}$$

relatív gyakoriság függvénye alapján az $Y\,(Y_1, Y_2, \ldots, Y_n)$ valószínűségi vektor együttes sűrűségfüggvényére az

$$f_Y\,(p; m, y_1, y_2, \ldots, y_n) = f_Y\,(p) = \prod_{j=1}^{n} \binom{m}{y_j} \cdot p^{\sum_{j=1}^{n} y_j}\,(1-p)^{nm-\sum_{j=1}^{n} y_j}$$

kifejezést kapjuk.

Megoldás – folytatás

- Az 1. feladatbeli meggondolásokhoz hasonlóan, az ismeretlen $p \in (0,1)$ paramétert a

$$\begin{cases} \dfrac{\partial}{\partial p} \ln\left(f_Y\left(p\right)\right) & = & 0, \\ \dfrac{\partial^2}{\partial p^2} \ln\left(f_Y\left(p\right)\right) & < & 0 \end{cases}$$

  feltételek alapján határozzuk meg.

- A rendszer első feltételéből a

$$\frac{\partial}{\partial p} \ln\left(f_Y\left(p\right)\right) = \frac{\partial}{\partial p} \left(\sum_{j=1}^{n} \ln\left(\binom{m}{y_j} \right) + \ln\left(p\right) \sum_{j=1}^{n} y_j + \left(nm - \sum_{j=1}^{n} y_j \right) \ln\left(1-p\right) \right)$$

$$= \frac{1}{p} \sum_{j=1}^{n} y_j - \frac{1}{1-p} \left(nm - \sum_{j=1}^{n} y_j \right) = 0$$

  egyenletet kapjuk, amelyet a

Megoldás – folytatás

$$\frac{p}{1-p} = \frac{\sum_{j=1}^{n} y_j}{nm - \sum_{j=1}^{n} y_j}$$

alakra hozhatunk, amiből az aránypárok egyszerű tulajdonságát felhasználva, a

$$\overline{p} = \frac{1}{mn} \sum_{j=1}^{n} y_j$$

szélsőértékhelyet kapjuk.

- Az olvasóra bízzuk a

$$\frac{\partial^2}{\partial p^2} \ln\left(f_Y\left(p\right)\right)\Big|_{p=\overline{p}} < 0$$

egyenlőtlenség kimutatását.

Megoldás – folytatás

- Végül kijelenthetjük, hogy a $p = \overline{p}$ szélsőértékhely valóban maximumpontja az együttes sűrűségfüggvénynek. Tehát az n-elemű mintavétel alapján az ismeretlen p paraméterre az

$$S = S(Y_1, Y_2, \ldots, Y_n) = \frac{1}{mn} \sum_{j=1}^{n} Y_j = \frac{1}{m} \cdot \frac{1}{n} \sum_{j=1}^{n} Y_j = \frac{1}{m} \cdot \overline{Y}$$

  becslőstatisztikát ajánlhatjuk.

- Diszkrét valószínűségi változók esetén nem mindig lehetséges a relatív gyakoriság függvényt parciálisan deriválni az ismeretlen paraméter szerint. Ilyenkor a sűrűségfüggvény maximumpontját biztosító szélsőértékhelyet egyéb módszerekkel kell meghatároznunk, például egymást követő, relatív gyakoriság függvényértékek összehasonlításával.
- Ugyanakkor folytonos valószínűségi változók és vektorok esetén előfordulhat, hogy több ismeretlen paramétert is meg kell határoznunk az (együttes) sűrűségfüggvény maximumpontjának beazonosítása végett. Erre ad példát a következő vázlatosan megoldott feladat. A részletes számítások kivitelezését az olvasóra bízzuk.

3. Megoldott feladat (Kétdimenziós normális eloszlás)

Alkalmazzuk a maximum-likelihood módszerét az

$$(X,Y) \sim \mathcal{N}_2\left(\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right),$$

$$\mu_1, \mu_2 \in \mathbb{R},\ \sigma_1, \sigma_2 > 0,\ \rho \in [-1,1]$$

eloszlású valószínűségi vektorra, ahol mind az öt paraméter ismeretlen!

Megoldás

- A 2. labor leírásában már láttuk, hogy az (X,Y) binormális valószínűségi vektor együttes sűrűségfüggvényét az

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2\left(1-\rho^2\right)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)},$$

$$(x,y) \in \mathbb{R}^2$$

felület adja.

Megoldás – folytatás

- Ezért a független $\left\{\left(X_j = x_j, Y_j = y_j\right) \sim \mathcal{N}_2\left(\mu, \Sigma\right)\right\}_{j=1}^{n}$ mintavétel esetén, ilyen előállítású együttes sűrűségfüggvényeket kellene összeszoroznunk, majd az így kapott bonyolult kifejezést logaritmálnunk, olyan összeget kapva, amelyben a binormális sűrűségfüggvények természetes alapú logaritmusai jelennek meg, végül pedig az optimalizációs feladat szempontjából az így kapott összegeket parciálisan is kellene deriválnunk az ismeretlen paraméterek szerint.
- Sokkal egyszerűbb, ha előbb az f sűrűségfüggvény természetes alapú logaritmusát parciálisan deriváljuk mind az öt ismeretlen paraméter szerint, majd az így kapott kifejezésbe behelyettesítjük a mintavételi pontokat, végül pedig összegezzük az adódó kifejezéseket. Egyszerűen az elvégzendő műveletek sorrendjének felcseréléséről van szó, amit megtehetünk, hiszen az $\left\{\left(X_j = x_j, Y_j = y_j\right)\right\}_{j=1}^{n}$ mintavételbeli pontok függetlenek egymástól. (Vigyázat, nem a komponensek (koordináták) függetlenségéről van szó!)

Megoldás – folytatás

- A $\ln(f)$ függvény parciális deriváltjaira a

$$\frac{\partial}{\partial \mu_1} \ln\left(f\left(x,y\right)\right) = \frac{1}{\sigma_1\left(1-\rho^2\right)}\left(\frac{x-\mu_1}{\sigma_1} - \rho\frac{y-\mu_2}{\sigma_2}\right),$$

$$\frac{\partial}{\partial \mu_2} \ln\left(f\left(x,y\right)\right) = \frac{1}{\sigma_2\left(1-\rho^2\right)}\left(\frac{y-\mu_2}{\sigma_2} - \rho\frac{x-\mu_1}{\sigma_1}\right),$$

$$\frac{\partial}{\partial \sigma_1} \ln\left(f\left(x,y\right)\right) = -\frac{1}{\sigma_1} + \frac{1}{\sigma_1^2\left(1-\rho^2\right)}\left(\frac{\left(x-\mu_1\right)^2}{\sigma_1} - \rho\frac{\left(x-\mu_1\right)\left(y-\mu_2\right)}{\sigma_2}\right),$$

$$\frac{\partial}{\partial \sigma_2} \ln\left(f\left(x,y\right)\right) = -\frac{1}{\sigma_2} + \frac{1}{\sigma_2^2\left(1-\rho^2\right)}\left(\frac{\left(y-\mu_2\right)^2}{\sigma_2} - \rho\frac{\left(x-\mu_1\right)\left(y-\mu_2\right)}{\sigma_1}\right),$$

$$\frac{\partial}{\partial \rho} \ln\left(f\left(x,y\right)\right) = \frac{\rho}{1-\rho^2} - \frac{\rho}{\left(1-\rho^2\right)^2}\left(\frac{\left(x-\mu_1\right)^2}{\sigma_1^2} - 2\rho\frac{\left(x-\mu_1\right)\left(y-\mu_2\right)}{\sigma_1\sigma_2} + \frac{\left(y-\mu_2\right)^2}{\sigma_2^2}\right) + \frac{1}{1-\rho^2}\frac{\left(x-\mu_1\right)\left(y-\mu_2\right)}{\sigma_1\sigma_2}$$

kifejezéseket kapjuk, és ekkor a maximum-likelihood módszere alapján a

Megoldás – folytatás

$$\begin{cases} \sum_{j=1}^{n} \frac{\partial}{\partial \mu_1} \ln\left(f\left(x_j, y_j\right)\right) & = & 0, \\ \sum_{j=1}^{n} \frac{\partial}{\partial \mu_2} \ln\left(f\left(x_j, y_j\right)\right) & = & 0, \\ \sum_{j=1}^{n} \frac{\partial}{\partial \sigma_1} \ln\left(f\left(x_j, y_j\right)\right) & = & 0, \\ \sum_{j=1}^{n} \frac{\partial}{\partial \sigma_2} \ln\left(f\left(x_j, y_j\right)\right) & = & 0, \\ \sum_{j=1}^{n} \frac{\partial}{\partial \rho} \ln\left(f\left(x_j, y_j\right)\right) & = & 0 \end{cases}$$

egyenletrendszert kell megoldanunk a szélsőértékhelyek végett.

Megoldás – folytatás

- Helyes átalakítások után az

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j,\ \overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j,$$

$$\overline{s}_x = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2},\ \overline{s}_y = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \overline{y}\right)^2},$$

$$\overline{r}_{xy} = \frac{1}{n\overline{s}_1\overline{s}_2}\sum_{j=1}^{n}\left(x_j - \overline{x}\right)\left(y_j - \overline{y}\right)$$

koordinátájú $(\overline{x}, \overline{y}, \overline{s}_x, \overline{s}_y, \overline{r}_{xy})$ szélsőértékhelyet kapjuk.

Megoldás – folytatás

- Kimutatható, hogy ez a szélsőértékhely valóban maximumpont, és ekkor már nyugodtan használhatjuk az

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j,\ \overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j,$$

$$\overline{S}_X = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2},\ \overline{S}_Y = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2},$$

$$\overline{R}_{XY} = \frac{1}{n\overline{S}_1\overline{S}_2}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)\left(Y_j - \overline{Y}\right)$$

  statisztikákat is az ismeretlen paraméterek mintavételi becslésére.
- Vegyük észre, hogy az eredmény nem is annyira meglepő, hiszen a becsült várható érték, a becsült szórás, illetve a becsült korrelációs együttható kifejezését kaptuk vissza!

- A maximum-likelihood módszerének ismertetése után térjünk vissza a nemparaméteres és paraméteres illeszkedésvizsgálatra!

Nemparaméteres χ^2-próba az ismeretlen elméleti eloszlásra

- Tekintsük az első oldalakon bevezetett jelöléseket! Emlékezzünk, hogy a

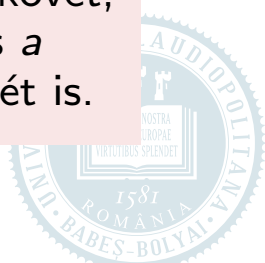
$$H_0 : p_i = p_i^{(0)},\ i = 1, 2, \ldots, k$$

nullhipotézist helyességét kell eldöntenünk a

$$H_1 : \exists i_0 \in \{1, 2, \ldots, k\} : p_{i_0} \neq p_{i_0}^{(0)}$$

alternatív hipotézissel szemben, ahol $\left\{p_i^{(0)}\right\}_{i=1}^{k}$ a statisztikai sokaságon tanulmányozott kvalitatív, vagy kvantitatív Y jellemző k darab osztályának – általunk feltételezett – elméleti, bekövetkezési valószínűségét jelöli.

- Ebben a pontban a $\left\{p_i^{(0)}\right\}_{i=1}^{k}$ valószínűségekről azt feltételezzük, hogy az azokat meghatározó összes paramétert ismerjük. Például nemcsak azt állítjuk, hogy a vizsgált sokaság az Y jellemző szempontjából valamilyen gamma-eloszlást követ, hanem egyúttal rögzítjük a 2. labor anyagában értelmezett $\mathcal{G}(a, b)$-eloszlás a alakparaméterének, valamint b skálázási tényezőjének pontosnak vélt értékét is.



folytatás

- Az Y jellemző szempontjából tanulmányozott sokaságnak egy n-elemű, független $\{Y_j\}_{j=1}^n$ mintavétele alapján már meghatároztuk az osztályok $\{n_i\}_{i=1}^k$ abszolút gyakoriságát ($\sum_{i=1}^k n_i = n$, $0 \leq n_i \leq n$), mi több ezekhez az $\{N_i\}_{i=1}^k$ valószínűségi változókat társítottuk. Az 1. tétel kimondja, hogy ilyen feltételek mellett a

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

valószínűségi változó $\chi^2\,(k-1,1)$-eloszlást követ.

- Akárcsak az eddigi próbák során, a nullhipotézis elfogadását, vagy elutasítását most is adott $\alpha \in (0,1)$ szignifikanciaszinttől függő megbízhatósági intervallum szerkesztésére vezetjük vissza.

folytatás

- Helyettesítsük be a χ^2 változó alakjába egyrészt a tapasztalati eloszlást meghatározó $\{n_i\}_{i=1}^k$ abszolút gyakoriságokat, másrészt az igaznak feltételezett nullhipotézisbeli $\left\{p_i^{(0)}\right\}_{i=1}^k$ elméleti valószínűségeket is! Az így kapott

$$\chi_0^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$$

  mennyiséget a nemparaméteres, illeszkedést ellenőrző χ^2-próba értékének nevezzük.

- Mivel a χ^2 valószínűségi változó csak nemnegatív értékeket vehet fel, a rá vonatkozó konfidenciaintervallum alsó végpontját 0-ra, felső végpontját pedig az $1-\alpha$ valószínűséghez tartozó

$$\chi^2_{k-1,1-\alpha} = F^{-1}_{\chi^2(k-1,1)}(1-\alpha)$$

  kvantilisre állíthatjuk.

- Amennyiben $\chi_0^2 \notin \left(0, \chi^2_{k-1,1-\alpha}\right)$, akkor elvetjük, máskülönben elfogadjuk a $H_0 : p_i = p_i^{(0)}$, $i = 1, 2, \ldots, k$ nullhipotézist.

- Tekintsünk egy példát!

Nemparaméteres χ^2-próba az exponenciális eloszlás ellenőrzésére

Adott $\alpha \in (0,1)$ szignifikanciaszint mellett alkalmazzuk a nemparaméteres χ^2-próbát az *ismert* $\lambda > 0$ paraméterű $\mathcal{E}xp(\lambda)$-eloszlás ellenőrzésére!

Megoldás

- Az $Y \sim \mathcal{E}xp(\lambda)$ valószínűségi változó csak pozitív értékeket vehet fel.
- Éppen ezért az Y „jellemző" osztályait a

$$0 = x_0 < x_1 < x_2 < \ldots < x_k = +\infty$$

osztópontrendszer által szült $[x_{i-1}, x_i)$ részintervallumokkal reprezentálhatjuk, ahol $i = 1, 2, \ldots, k$.

- A feltételezett

$$F_{\mathcal{E}xp(\lambda)}(x) = 1 - e^{-\lambda x}, x \geq 0$$

eloszlásfüggvény segítségével a null- és alternatív hipotéziseket a

Megoldás – folytatás

$$
\begin{aligned}
H_0 : p_i &= p_i^{(0)} \\
&= F_{\mathcal{E}xp(\lambda)}(x_i) - F_{\mathcal{E}xp(\lambda)}(x_{i-1}) \\
&= \begin{cases} 1 - e^{-\lambda x_1}, & i = 1, \\ e^{-\lambda x_{i-1}} - e^{-\lambda x_i}, & 2 \leq i \leq k-1, \\ e^{-\lambda x_{k-1}}, & i = k, \end{cases}
\end{aligned}
$$

illetve a

$$H_1 : \exists i_0 \in \{1, 2, \ldots, k\} : p_{i_0} \neq p_{i_0}^{(0)}$$

alakban fogalmazhatjuk meg.

- Álljon most a rendelkezésünkre az Y valószínűségi változónak egy n-elemű független $\{Y_j = y_j\}_{j=1}^n$ mintavétele! Számoljuk meg, hogy az $\{[x_{i-1}, x_i)\}_{i=1}^k$ osztályok (részintervallumok) hány elemet foglalnak magukba a mintevételi $\{y_j\}_{j=1}^n$ értékek közül, vagyis határozzuk meg az osztályok $\{n_i\}_{i=1}^k$ abszolút gyakoriságát!

Megoldás – folytatás

- Ekkor már minden adat ismert a nemparaméteres, illeszkedést ellenőrző χ^2-próba

$$\chi_0^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$$

értékének kiszámításához.

- Ha

$$\chi_0^2 < \chi_{k-1,1-\alpha}^2 = F_{\chi^2(k-1,1)}^{-1}(1-\alpha),$$

akkor elfogadhatjuk a nullhipotézist, miszerint az adott $\{y_j\}_{j=1}^n$ mintavételi értékek valóban $\mathcal{E}xp(\lambda)$-eloszlást követnek, máskülönben elutasítjuk azt az α kockázati szint mellett.

Paraméteres χ^2-próba az ismeretlen elméleti eloszlásra

- Ebben a pontban azt feltételezzük, hogy az ismeretlen eloszlású Y valószínűségi változó F_Y eloszlásfüggvénye az $\{a_r\}_{r=1}^{\ell} \subset \mathbb{R}$ ismeretlen alakparaméterektől is függ, azaz az F_Y eloszlásfüggvényt az

$$F_Y = F_Y\left(x; a_1, a_2, \ldots, a_\ell\right), x \in \mathbb{R}$$

alakban keressük.

- Ilyen esetben az $\{a_r\}_{r=1}^{\ell}$ alakparaméterekre előbb pontbecslést adunk a maximum-likelihood módszerével a rendelkezésünkre álló független mintavétel alapján. Jelölje $\{\overline{a}_r\}_{r=1}^{\ell}$ a becsült paramétereket, továbbá tekintsük az Y változónak az

$$x_0 < x_1 < \ldots < x_k$$

osztópontrendszer által meghatározott $\{[x_{i-1}, x_i)\}_{i=1}^{k}$ osztályait.

- Ekkor a null- és alternatív hipotéziseket a

$$H_0 : p_i = p_i^{(0)} = F_0\left(x_i; \overline{a}_1, \overline{a}_2, \ldots, \overline{a}_\ell\right) - F_0\left(x_{i-1}; \overline{a}_1, \overline{a}_2, \ldots, \overline{a}_\ell\right), \; i = 1, 2, \ldots, k$$

illetve a

$$H_1 : \exists i_0 \in \{1, 2, \ldots, k\} : p_{i_0} \neq p_{i_0}^{(0)}$$

alakban fogalmazhatjuk meg, ahol

folytatás

$$p_i = F_Y\left(x_i; a_1, a_2, \ldots, a_\ell\right) - F_Y\left(x_{i-1}; a_1, a_2, \ldots, a_\ell\right),\ i = 1, 2, \ldots, k$$

az ismeretlen eloszlásfüggvényből származó valószínűségeket jelöli.

- Tekintsük az Y valószínűségi változónak az n-elemű, független $\{Y_j\}_{j=1}^n$ mintavételét és határozzuk meg az $\{[x_{i-1}, x_i)\}_{i=1}^k$ osztályok $\{n_i\}_{i=1}^k$ abszolút gyakoriságát ($\sum_{i=1}^k n_i = n$, $0 \leq n_i \leq n$)!
- Ha az $\{n_i\}_{i=1}^k$ abszolút gyakoriságokhoz az $\{N_i\}_{i=1}^k$ valószínűségi változókat társítjuk, akkor a

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

valószínűségi változó $\chi^2\left(k - \ell - 1, 1\right)$-eloszlást követ, azaz minden becsült paraméter eggyel csökkenti a (4)-es $\chi^2\left(k - 1, 1\right)$-eloszlású valószínűségi változó szabadságfokát.

folytatás

- Tehát, a nemparaméteres, illeszkedést ellenőrző χ^2-próbával szemben egyrészt az ismeretlen paraméterek maximum-likelihood módszerrel történő becslését, másrészt a becsült paraméterek számával megegyező mértékben csökkent szabadságfokot emelhetjük ki eltérésként. Ezt a két különbséget leszámítva, a paraméteres, illeszkedést ellenőrző χ^2-próba – adott $\alpha \in (0,1)$ szignifikanciaszint esetén – teljesen hasonló lépéseket hajt végre, mint a nemparaméteres változata: ha teljesül a

$$\chi_0^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}} < \chi^2_{k-\ell-1,1-\alpha} = F^{-1}_{\chi^2(k-\ell-1,1)}(1-\alpha)$$

egyenlőtlenség, akkor a nullhipotézist, egyébként az alternatív hipotézist fogadja el.

- Annak érdekében, hogy az olvasó számára egyszerű összehasonlítási alapot biztosítsunk, tárgyaljuk újra a feltételezett $\mathcal{E}xp(\lambda)$-eloszlás ellenőrzését úgy, hogy ezúttal a $\lambda > 0$ paramétert ismeretlennek tekintjük!

Paraméteres χ^2-próba az exponenciális eloszlás ellenőrzésére

Adott $\alpha \in (0,1)$ szignifikanciaszint mellett alkalmazzuk a paraméteres χ^2-próbát az *ismeretlen* $\lambda > 0$ paraméterű $\mathcal{E}xp(\lambda)$-eloszlás ellenőrzésére!

Megoldás

- Tudjuk, hogy az $Y \sim \mathcal{E}xp(\lambda)$ valószínűségi változó csak pozitív értékeket vehet fel.
- Ezért az Y változó osztályait a

$$0 = x_0 < x_1 < x_2 < \ldots < x_k = +\infty$$

osztópontrendszer által meghatározott $\{[x_{i-1}, x_i)\}_{i=1}^{k}$ részintervallumokkal azonosítjuk.

Megoldás – folytatás

- Mivel a feltételezett
$$F_{\mathcal{E}xp(\lambda)}(x) = 1 - e^{-\lambda x}, x \geq 0$$
eloszlásfüggvény λ paraméterét nem ismerjük, előbb értékére maximum-likelihood pontbecslést adunk.
- Tekintsük ezért az Y valószínűségi változónak egy n-elemű független $\{Y_j = y_j\}_{j=1}^n$ mintavételét! Ekkor az 1. feladat alapján az ismeretlen λ paraméter becslésére a tapasztalati várható érték reciprokát használhatjuk, vagyis:
$$\overline{\lambda} = \frac{n}{\sum_{j=1}^{n} y_j} = \frac{1}{\overline{y}}.$$
- Ezért a null- és alternatív hipotéziseket a

$$H_0 : p_i = p_i^{(0)} = F_{\mathcal{E}xp(\overline{\lambda})}(x_i) - F_{\mathcal{E}xp(\overline{\lambda})}(x_{i-1}) = \begin{cases} 1 - e^{-\overline{\lambda} x_1}, & i = 1, \\ e^{-\overline{\lambda} x_{i-1}} - e^{-\overline{\lambda} x_i}, & 2 \leq i \leq k-1, \\ e^{-\overline{\lambda} x_{k-1}}, & i = k, \end{cases}$$

illetve a
$$H_1 : \exists i_0 \in \{1, 2, \ldots, k\} : p_{i_0} \neq p_{i_0}^{(0)}$$
alakban fogalmazhatjuk meg.

Megoldás – folytatás

- Ha az $\{[x_{i-1}, x_i)\}_{i=1}^k$ osztályok (részintervallumok) $\{n_i\}_{i=1}^k$ abszolút gyakoriságát is meghatározzuk, akkor már minden adat ismert a paraméteres, illeszkedést ellenőrző χ^2-próba

$$\chi_0^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$$

értékének kiszámításához.

- Amennyiben a

$$\chi_0^2 < \chi_{k-2,1-\alpha}^2 = F_{\chi^2(k-2,1)}^{-1}(1-\alpha)$$

egyenlőtlenség teljesül, akkor elfogadhatjuk a nullhipotézist, miszerint az adott $\{y_j\}_{j=1}^n$ mintavételi értékek $\mathcal{E}xp\left(\overline{\lambda}\right)$-eloszlást követnek, egyébként elutasítjuk azt az α szignifikanciaszint mellett.

Végső megjegyzések

- A nemparaméteres, illetve a paraméteres illeszkedésvizsgáló χ^2-próbák azt feltételezik, hogy a statisztikai sokaságon tanulmányozott kvalitatív, vagy kvantitatív Y valószínűségi változó $k \geq 1$ darab osztályát ismerjük. A gyakorlati alkalmazások esetében viszont általában csak egy n-elemű független $\{Y_j = y_j\}_{j=1}^{n}$ mintavétel áll a rendelkezésünkre.
- Így felmerül a kérdés, hogy hogyan választhatjuk meg az osztályokat és mi legyen azok optimális száma.
- Amennyiben a nullhipotézisben folytonos eloszlást feltételezünk és egyenlő hosszúságú osztályokat szeretnénk előállítani, akkor az osztályok számát a

$$k = [1 + \log_2 n]$$

Sturges-féle közelítő képlet adja, amivel az osztályokat meghatározó osztópontókat az

$$x_i = y_{\min} + i \cdot \frac{y_{\max} - y_{\min}}{k},\ i = 0, 1, \ldots, k$$

módon vehetjük fel, ahol $y_{\min} = \min\{y_j : j = 1, 2, \ldots, n\}$ és $y_{\max} = \max\{y_j : j = 1, 2, \ldots, n\}$.

Végső megjegyzések – folytatás

- Ha a nullhipotézisben diszkrét eloszlásra gyanakszunk, akkor az osztályokat az adott mintából összegyűjtött és növekvő sorrendbe rendezett különböző egész számoknak feleltethetjük meg, vagy dolgozhatunk ugyancsak a fenti Sturges-féle képlettel úgy, hogy a kapott részintervallumok egésszé kerekített középpontját tekintjük a feltételezett diszkrét eloszlású valószínűségi változó osztályainak.
- Ügyeljünk arra, hogy diszkrét eloszlást feltételezve az osztályok egész értékű pontokká „zsugorodnak össze", következésképpen a nullhipotézis által igaznak tartott valószínűségeket a feltételezett diszkrét eloszlás eloszlásfüggvényének „pillanatnyi változásával", vagyis a relatív gyakoriság függvénnyel határozhatjuk meg.

1. feladat

Ellenőrizzétek az $\alpha = 0.03$ szignifikanciaszint mellett, hogy $(n = 6, \sigma = 1)$-paraméterű Pearson-féle χ^2-eloszlást követnek-e az alábbi számok!

5.078	1.279	2.268	7.382	3.428	11.735	7.065	8.923
10.074	10.230	4.751	1.959	5.659	9.660	4.387	4.236
3.739	5.316	12.310	8.819	13.176	1.580	6.317	7.266
2.993	4.776	7.463	6.324	2.784	4.675	5.321	5.569
6.987	2.821	2.146	5.640	5.336	10.536	2.854	12.928
4.454	5.461	1.318	8.051	4.3 00	4.479	1.681	6.725
4.381	11.152	1.561	4.392	8.841	8.163	11.056	10.824
4.171	5.404	6.311	5.063	3.380	15.393	1.328	10.343
8.528	8.922	4.085	4.501	5.208	2.098	12.286	7.205
14.082	4.522	5.595	9.007	3.354	7.463	2.623	1.477
1.823	10.886	1.884	3.883	5.961	3.378	2.686	5.320
5.635	3.093	1.520	2.345	2.563	6.801	9.627	3.539
3.370	3.282	1.258	6.437	3.098	4.057	6.988	6.285
5.758	9.549	5.893	1.724	11.737	4.856	4.959	1.219
3.564	3.616	1.404	7.234	2.894	4.616	5.170	6.971

2. feladat

Milyen eloszlásúak az alábbi számok és milyen alakparaméterek jellemzik az eloszlást?

0.65	4.03	3.04	5.47	0.90	0.75	1.81	1.21	0.25
0.97	3.68	6.54	1.55	1.61	1.37	5.01	0.01	2.91
2.26	1.04	1.08	3.50	5.27	9.02	0.38	2.88	0.22
1.63	0.81	1.83	5.43	6.05	0.36	0.76	3.70	3.83
0.41	13.18	3.36	1.53	0.14	1.56	1.70	2.13	3.40
1.68	5.62	0.08	7.08	0.22	1.08	2.69	3.38	0.06
2.27	5.13	0.19	5.35	0.25	0.19	0.38	3.31	0.03
3.40	3.72	1.86	2.70	0.39	7.26	0.80	8.37	3.97
0.10	0.25	1.30	0.05	0.09	6.58	9.96	8.36	1.09
0.52	3.10	2.10	6.43	5.15	0.92	1.57	0.78	0.34
1.64	0.17	2.15	0.37	2.93	3.89	6.89	0.74	2.86
3.81	0.52	3.27	3.53	0.77	0.18	0.92	5.01	1.18
2.96	3.39	1.07	2.36	0.35	2.21	0.82	12.72	0.24
9.71	0.69	4.81	1.03	4.65	3.95	4.70	1.72	2.90
4.91	1.37	1.65	3.95	0.48	4.78	2.43	3.29	1.27
1.90	1.36	0.10	2.59	5.69				

3. feladat

Feljegyezve egy hat oldalú dobókockával való gurítások eredményét az alábbi táblázatot kapták:

Dobott érték	1	2	3	4	5	6
Gyakoriság	284	259	241	210	238	268

Az $\alpha = 0.02$ szignifikanciaszint mellett döntsétek el, hogy szabályos hexaéder alakú volt-e a használt dobókocka!

4. feladat

Egy kertészet palántanevelő részlegén végzett vizsgálat során feljegyezve az egy négyzetméteren kikelt csípőspaprika-palánták számát az alábbi összesített táblázatot kapták:

Palánták száma	0	1	2	3	4	5	6	7	8	9	10	≥ 11
Gyakoriság	24	86	170	227	203	134	83	43	19	6	4	1

Az $\alpha = 0.01$ szignifikanciaszint mellett döntsétek el, hogy a csípőspaprika-palánták száma Poisson-eloszlást követ-e!

– legyen ünnep minden pillanatotok –