

Statisztika III – nemparaméteres próbák

Sztochasztika

Horváth Illés

2024/12/03

Az u -próbák és t -próbák a várható értéket tesztelik vagy egy adott érték, vagy egy másik minta várható értéke ellen.

A nemparaméteres próbák ezzel szemben valamilyen absztrakt tulajdonság teljesülését tesztelik egy numerikus érték helyett.

Három nemparaméteres próbát vizsgálunk meg:

- *illeszkedésvizsgálat*: azt teszteli, hogy egy minta egy adott háttéreloszlásból származik-e;
- *homogenitásvizsgálat*: azt teszteli, hogy két minta eloszlása azonos-e;
- *függetlenségvizsgálat*: azt teszteli, hogy egy mintán megfigyelt két tulajdonság független-e.

A három próba együttesen *Pearson-féle χ^2 -próbák* néven is ismertek (mivel mindegyik a χ^2 eloszlást használja).

Legyen adott egy n elemű minta, ahol minden elem r kategória valamelyikébe esik. $1 - \varepsilon$ szignifikanciaszinten akarjuk tesztelni a következőt:

- H_0 : a minta eloszlása egy adott $p_1, \dots, p_r$ elméleti eloszlást követ a kategóriákon, vagy
- H_1 : a minta eloszlása eltér $p_1, \dots, p_r$ -től.

Jelölje a mintaelemek számát az egyes kategóriákban ν_i , $i = 1, \dots, r$. A statisztika

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

A χ^2_{ε} percentilis az $r - 1$ szabadsági fokú χ^2 -eloszlás $(1 - \varepsilon)$ kvantilise.

Ha $\chi^2 < \chi^2_{\varepsilon}$, elfogadjuk H_0 -t. Egyébként H_0 -t elvetjük.

Egy véletlen bit generátor elméletileg 50% eséllyel ad 0-t vagy 1-et. 1000 elemű mintát veszünk, melynek eredménye 471 db 0 és 529 db 1-es. Teszteljük 95%-os szignifikanciaszinten azt, hogy a háttéreloszlás 50%-50%-e.

Illeszkedésvizsgálatot végzünk. $r = 2$ kategória van: 0 és 1.

- H_0 : a háttéreloszlás $p_1 = 0.5$, $p_2 = 0.5$;
- H_1 : a háttéreloszlás ettől eltérő.

A minta

$$\nu_1 = 471, \quad \nu_2 = 529,$$

a minta mérete $n = 1000$.

A statisztika

$$\begin{aligned}\chi^2 &= \sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i} \\ &= \frac{(471 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} + \frac{(529 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} = 3.364.\end{aligned}$$

A percentilis az $r - 1 = 1$ szabadsági fokú χ^2 eloszlás 95% kvantilise:

$$\chi_{\varepsilon}^2 = 3.84$$

a χ^2 eloszlás táblázata alapján.

$$\chi^2 = 3.364 < 3.84 = \chi_{\varepsilon}^2$$

teljesül, tehát elfogadjuk H_0 -t 95%-os szignifikanciaszinten.

De ha például a minta 451 db 0 és 549 db 1-es, akkor a statisztika

$$\begin{aligned}\chi^2 &= \sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i} \\ &= \frac{(451 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} + \frac{(549 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} = 10.404,\end{aligned}$$

$$\chi^2 = 10.404 > 3.84 = \chi_{\varepsilon}^2,$$

tehát erre a mintára elvetjük H_0 -t 95%-os szignifikanciaszinten, azaz a konklúzió az, hogy a véletlen bit generátor nem 50%-ban ad 0-t és 1-est.

Illeszkedésvizsgálat alkalmazható akkor is, ha a háttéreloszlás folytonos. Ilyenkor a folytonos tartományt fel kell darabolni véges sok intervallumra a próba előtt.

Ökölszabályként minden intervallumnak tartalmaznia kell legalább 5 minta elemet, különben a próba nem lesz elég megbízható. Ezen túl az intervallumok pontos megválasztásában van némi szabadságunk.

Amint az intervallumokat lerögzítettük, azok lesznek a kategóriák. A mintaelemeket is ennek megfelelően csoportosítjuk, a p_i elméleti valószínűségek pedig az egyes intervallumok valószínűségei az elméleti háttéreloszlás szerint.

Adott két minta, n és m elemű. Minden egyes mintaelem r kategória valamelyikébe esik. Tesztelni akarjuk $1 - \varepsilon$ szignifikanciaszinten, hogy

- H_0 : a két minta eloszlása azonos, vagy
- H_1 : a két minta eloszlása eltérő.

Legyen az egyes kategóriákba eső mintaelemek száma ν_i , $i = 1, \dots, r$ az első mintára és μ_i , $i = 1, \dots, r$ a második mintára. A statisztika

$$\chi^2 = \sum_{i=1}^r nm \frac{(\nu_i/n - \mu_i/m)^2}{\nu_i + \mu_i}.$$

A χ^2_{ε} percentilis az $r - 1$ szabadsági fokú χ^2 -eloszlás $1 - \varepsilon$ kvantilise.

Ha $\chi^2 < \chi^2_{\varepsilon}$, elfogadjuk H_0 -t. Egyébként elvetjük H_0 -t.

Adott egy n méretű minta, ahol minden elemnek van két tulajdonsága. Az első tulajdonság r kategória valamelyikébe eshet, a második tulajdonság s kategória valamelyikébe. Tesztelni akarjuk $1 - \varepsilon$ szignifikanciaszinten, hogy

- H_0 : a két tulajdonság független, vagy
- H_1 : a két tulajdonság nem független.

Jelölje ν_{ij} az olyan mintaelemek számát, melyekre az első tulajdonság az i -edik kategóriába és a második tulajdonság a j -edik kategóriába esik. Legyen továbbá

$$\nu_{i,.} = \sum_{j=1}^s \nu_{ij} \quad \text{és} \quad \nu_{.,j} = \sum_{i=1}^r \nu_{ij}.$$

A statisztika

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s n \frac{\left(\nu_{i,j} - \frac{\nu_{i,\cdot} \nu_{\cdot,j}}{n}\right)^2}{\nu_{i,\cdot} \nu_{\cdot,j}}.$$

A percentilis az $(r-1)(s-1)$ szabadsági fokú χ^2 -eloszlás $1 - \varepsilon$ kvantilise.

Ha $\chi^2 < \chi_{\varepsilon}^2$, elfogadjuk H_0 -t. Egyébként elvetjük H_0 -t.

9. feladat

Egy tóban háromféle hal él: amúr, makréla és ponty. Ottó bácsi, az öreg horgász azt súgja nekünk, hogy a tóban kétszer annyi a ponty, mint a makréla vagy az amúr. Kifogtunk 60 halat; döntsük el ez alapján 95%-os szinten, hallgathatunk-e Ottó bácsira.

amúr	makréla	ponty
11	14	35

Megoldás. Illeszkedésvizsgálatot végzünk. A kategóriák száma $r = 3$, és az elméleti háttéreloszlás Ottó bácsi szerint

$$p_1 = 0.25, \quad p_2 = 0.25, \quad p_3 = 0.5.$$

- H_0 : a minta ebből a háttéreloszlásból származik;
- H_1 : a minta nem ebből a háttéreloszlásból származik.

9. feladat

A minta mérete $n = 60$, és maga a minta

$$\nu_1 = 11, \quad \nu_2 = 35, \quad \nu_3 = 14.$$

A statisztika

$$\begin{aligned} \chi^2 &= \sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i} = \frac{(11 - 60 \cdot 0.25)^2}{60 \cdot 0.25} + \\ &\quad \frac{(35 - 60 \cdot 0.5)^2}{60 \cdot 0.5} + \frac{(14 - 60 \cdot 0.25)^2}{60 \cdot 0.25} = 1.967. \end{aligned}$$

A percentilis az $r - 1 = 2$ szabadsági fokú χ^2 -eloszlás 95%-os kvantilise:

$$\chi_{\varepsilon}^2 = 5.99.$$

Az összehasonlítás

$$\chi^2 = 1.967 < \chi_{\varepsilon}^2 = 5.99$$

teljesül, tehát elfogadjuk H_0 -t 95%-os szignifikanciaszinten, és úgy döntünk, hallgathatunk Ottó bácsira.

13. feladat

Azt szeretnénk megtudni, hogy motorbalesetek esetén a bukósisak színe és a baleseti sérülések súlyossága között van-e összefüggés. Az utóbbi néhány év adatai alapján a következő táblázatot kaptuk:

	fekete	fehér	narancssárga
nincs sérülés	501	367	31
könnyű sérülés	173	107	7
súlyos sérülés	30	15	1

95%-os konfidenciaszinten döntsünk arról a hipotézisről, hogy a csoport (kontroll vagy balesetes) független a bukósisak színétől.

Függetlenségvizsgálatot végzünk.

- H_0 : a sisak színe és a sérülés súlyossága független;
- H_1 : a sisak színe és a sérülés súlyossága nem független.

$r = 3$ szín kategória és $s = 3$ sérülés kategória van. ν_{ij} a táblázat elemei, és szükségünk van a következő értékekre is:

$$\begin{aligned}\nu_{1,.} &= 501 + 367 + 31 = 899, & \nu_{.,1} &= 501 + 173 + 30 = 704, \\ \nu_{2,.} &= 173 + 107 + 7 = 287, & \nu_{.,2} &= 367 + 107 + 15 = 489, \\ \nu_{3,.} &= 30 + 15 + 1 = 46, & \nu_{.,3} &= 31 + 7 + 1 = 39.\end{aligned}$$

A teljes minta mérete $n = 1232$.

13. feladat

A statisztika

$$\begin{aligned}\chi^2 &= \sum_{i=1}^r \sum_{j=1}^s n \frac{(\nu_{i,j} - \frac{\nu_{i,\cdot} \nu_{\cdot,j}}{n})^2}{\nu_{i,\cdot} \nu_{\cdot,j}} = \\ &1232 \cdot \left(\frac{(501 - \frac{899 \cdot 704}{1232})^2}{899 \cdot 704} + \dots + \frac{(1 - \frac{46 \cdot 39}{1232})^2}{46 \cdot 39} \right) = \\ &= 3.875.\end{aligned}$$

A percentilis az $(r-1)(s-1) = (3-1)(3-1) = 4$ szabadsági fokú χ^2 -eloszlás 95%-os percentilise:

$$\chi_{\varepsilon}^2 = 9.49.$$

Az összehasonlítás

$$\chi^2 = 3.875 < \chi_{\varepsilon}^2 = 9.49$$

teljesül, így elfogadjuk H_0 -t 95%-os szinten, azaz a minta alapján a sisak színe és a sérülés súlyossága függetlennek tekinthető.