Statisztika III – nemparaméteres próbák

Sztochasztika

Horváth Illés

2023/11/30

Nemparaméteres próbák

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

Nemparaméteres próbák

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.

Nemparaméteres próbák

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.

Nemparaméteres próbák

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).

Illeszkedésvizsgálat

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 $p_1, \ldots, p_r$ elméleti eloszlást követ a kategóriákon, vagy
- H_1: a minta eloszlása eltér $p_1, \ldots, p_r$-től.

Illeszkedésvizsgálat

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 $p_1, \ldots, p_r$ elméleti eloszlást követ a kategóriákon, vagy
- H_1: a minta eloszlása eltér $p_1, \ldots, p_r$-től.

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

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

Illeszkedésvizsgálat

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 $p_1, \ldots, p_r$ elméleti eloszlást követ a kategóriákon, vagy
- H_1: a minta eloszlása eltér $p_1, \ldots, p_r$-től.

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

Illeszkedésvizsgálat

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 $p_1, \ldots, p_r$ elméleti eloszlást követ a kategóriákon, vagy
- H_1: a minta eloszlása eltér $p_1, \ldots, p_r$-től.

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

Példa

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.

Példa

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.

Példa

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

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

a minta mérete $n = 1000$.

Példa

A statisztika

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

Példa

A statisztika

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

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

$$\chi^2_\varepsilon = 3.84$$

a χ^2 eloszlás táblázata alapján.

Példa

A statisztika

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

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

$$\chi^2_\varepsilon = 3.84$$

a χ^2 eloszlás táblázata alapján.

$$\chi^2 = 3.364 < 3.84 = \chi^2_\varepsilon$$

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

Példa

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

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

Példa

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

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

és

$$\chi^2 = 10.404 > 3.84 = \chi^2_\varepsilon,$$

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 folytonos háttéreloszlásra

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

Illeszkedésvizsgálat folytonos háttéreloszlásra

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.

Illeszkedésvizsgálat folytonos háttéreloszlásra

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.

Homogenitásvizsgálat

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.

Homogenitásvizsgálat

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

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

Homogenitásvizsgálat

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

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

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

Homogenitásvizsgálat

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

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

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.

Függetlenségvizsgálat

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.

Függetlenségvizsgálat

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 $\nu_{i,j}$ az olyan mintaelemek számát, melyekre az első tulajdonság az i-edik kategóriába és a második tulajdonság a j-edik kategóriába esik. Legyen továbbá

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

Függetlenségvizsgálat

A statisztika

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

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.

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 $\chi^2 < \chi^2_\varepsilon$, 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

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

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

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

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

- 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

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

9. feladat

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

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

A statisztika

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

9. feladat

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

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

A statisztika

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

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

$$\chi^2_\varepsilon = 5.99.$$

9. feladat

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

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

A statisztika

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

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

$$\chi^2_\varepsilon = 5.99.$$

Az összehasonlítás

$$\chi^2 = 1.967 < \chi^2_\varepsilon = 5.99$$

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.

13. feladat

Függetlenségvizsgálatot végzünk.

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

13. feladat

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. $\nu_{i,j}$ a táblázat elemei, és szükségünk van a következő értékekre is:

$$\begin{aligned} \nu_{1,.} &= 501 + 367 + 31 = 905, & \nu_{.,1} &= 501 + 173 + 30 = 704, \\ \nu_{2,.} &= 173 + 107 + 7 = 287, & \nu_{.,2} &= 367 + 107 + 15 = 489, \\ \nu_{3,.} &= 30 + 15 + 1 = 46, & \nu_{.,3} &= 31 + 7 + 1 = 39. \end{aligned}$$

A teljes minta mérete $n = 1232$.

13. feladat

A statisztika

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

$$1232 \cdot \left(\frac{\left(501 - \frac{905\cdot 704}{1232}\right)^2}{905 \cdot 704} + \cdots + \frac{\left(1 - \frac{46\cdot 39}{1232}\right)^2}{46 \cdot 39} \right) =$$

$$= 3.875.$$

13. feladat

A statisztika

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

$$1232 \cdot \left(\frac{\left(501 - \frac{905\cdot 704}{1232}\right)^2}{905 \cdot 704} + \cdots + \frac{\left(1 - \frac{46\cdot 39}{1232}\right)^2}{46 \cdot 39} \right) =$$

$$= 3.875.$$

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

$$\chi^2_\varepsilon = 9.49.$$

13. feladat

A statisztika

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

$$1232 \cdot \left(\frac{\left(501 - \frac{905\cdot 704}{1232}\right)^2}{905 \cdot 704} + \cdots + \frac{\left(1 - \frac{46\cdot 39}{1232}\right)^2}{46 \cdot 39} \right) =$$

$$= 3.875.$$

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

$$\chi^2_\varepsilon = 9.49.$$

Az összehasonlítás

$$\chi^2 = 3.875 < \chi^2_\varepsilon = 9.49$$

teljesül, így elfogadjuk H_0-t 95%-os szinten, és arra következtetünk a minta alapján, hogy a sisak színe és a sérülés súlyossága függetlenek.