

2. tétel

Nagy Dániel

2020. június 1.

Kivonat

Bootstrap módszerek. A maximum likelihood módszer. Hipotézis tesztelés. Extrém statisztikák. Post hoc analízis. Regresszió. Függetlenségvizsgálat. Egzakt tesztek.

Tartalomjegyzék

1. Bevezetés	
1.1. Valószínűségszámítás alapfogalmak
1.2. Statisztikai következtetés (inference)
2. Bootstrap módszerek [1]	
2.1. Jackknife módszer [2]
3. Maximum likelihood	
4. Extrém statisztikák	
5. Post-hoc analízis	
6. Regresszió	
7. Hipotézistesztelés	
7.1. Egymintás u -próba [3]
7.2. Egymintás t -próba [4]
8. Függetlenségvizsgálat, χ^2-próba	

1. Bevezetés

1.1. Valószínűségszámítás alapfogalmak

- **Eseménytér** (ez egy absztrakt fogalom): $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ pl. kockadobás esetén $\Omega = \{\omega_1 = \text{"1est dobok"}, \omega_2 = \text{"2est dobok"}, \omega_3 = \text{"párosat dobok"} \dots\}$
- **Valószínűségi változó**: $X : \Omega \rightarrow \mathbb{R}$ pl. kockadobás esetén $X(\omega_1) = 1, X(\omega_2) = 2, \dots$
- **Valószínűség**: P egy mérték, amely Ω részhalmazaihoz számot rendel:

- $P : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$
- $P(\Omega) = 1$ és $P(\emptyset) = 0$
- $0 \leq P(A) \leq 1 \forall A \in \Omega$
- Ha A_1, A_2, \dots diszjunkt részhalmazai Ω -nak, akkor

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

- **Hasznos összefüggések:**

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Két esemény független $\iff P(A \cap B) = P(A)P(B)$
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Teljes valószínűség: Ha A_1, A_2, \dots az Ω egy felosztása, akkor

$$P(B) = \sum_k P(B|A_k)P(A_k)$$

- Bayes-tétel: Ha A_1, A_2, \dots az Ω egy felosztása, akkor

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)} = \frac{P(B|A_k)P(A_k)}{\sum_j P(B|A_j)P(A_j)}$$

- **Eloszlásfüggvény** (CDF - cumulative distribution function):

$$F_X(x) = P(X < x) = P(\{\omega \in \Omega | X(\omega) < x\})$$

diszkrét esetben

$$F_X(x) = P(X = x) = P(\{\omega \in \Omega | X(\omega) = x\})$$

Ha az X változó F eloszlást követ, akkor így jelöljük: $X \sim F$.

- **Sűrűségfüggvény** (PDF - Probability density function):

Ha az X változó eloszlásfüggvénye $F_X(x)$, akkor a sűrűségfüggvény definíciója

$$F_X(x) = \int_{-\infty}^x \rho_X(\xi) d\xi \iff P(a \leq X(\omega) \leq b) = \int_a^b \rho_X(x) dx$$

Megjegyzés: sűrűségfüggvénye csak folytonos eloszlású valószínűségi változónak van.

- **Várható érték**

$$\text{folytonos eset } E(X) = \langle X \rangle = \int_{-\infty}^{\infty} x \rho(x) dx$$

$$\text{diszkrét eset } E(X) = \langle X \rangle = \sum_k x_k p_k = \sum_k x_k P(X = x_k)$$

- **Várható értékre vonatkozó azonosságok:**

$$- \text{ Ha } Y = g(X) \Rightarrow E(Y) = E(g(X)) = \int_{-\infty}^{\infty} g(x) \rho(x) dx$$

$$- E\left(\sum_k a_k X_k\right) = \sum_k a_k E(X_k)$$

$$- \text{ Ha } X_1, X_2, \dots \text{ független változók, akkor } E\left(\prod_k X_k\right) = \prod_k E(X_k)$$

- **Variancia** (szórásnégyzet)

Ha $E(X) = \mu$, akkor a szórásnégyzet a változó és a várható értéke közötti különbség négyzetének várható értéke:

$$\sigma^2(X) = V(X) = E((X - \mu)^2) = \langle (X - \mu)^2 \rangle = \langle X^2 \rangle - \mu^2$$

- Ha X_1, X_2, \dots függetlenek, akkor

$$\sigma^2\left(\sum_k (a_k X_k + b_k)\right) = \sum_k a_k^2 \sigma^2(X_k)$$

- **Szórás** (standard deviation) definíciója:

$$\sigma(X) = \sqrt{\sigma^2(X)} = \sqrt{\langle X^2 \rangle - \langle X \rangle^2}$$

- **Minta**

Matematikailag egy statisztikai minta megfelel N darab azonos eloszlású, független (iid) változónak egy adott F eloszlásból.

- **Minta átlaga:** $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ (p_k -t a relatív gyakorisággal közelítjük)

- **Minta varianciája:** $s^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$, standard hibája $SE = \sqrt{s^2}$. A nevezőben az $N-1$ faktor az ún. Bessel-korrekciónak [5].

- **Megjegyzés:** Ha egy teljes populáció esetén $E(X) = \mu$ és $V(X) = \sigma^2$, attól még általában $\bar{X} \neq \mu$ illetve $s^2 \neq \sigma^2$.

- Egy minta esetében \bar{X}, s^2, SE maguk is valószínűségi változók, hiszen minden mintavételezés esetén más-más értéket vehetnek fel. Ezért van értelme arról beszélni, hogy pl. s^2 értéke milyen eloszlást követ. Ha a minta (mérési pontok) iid változók, és $E(X_i) = \mu$, $V(X_i) = \sigma^2$, akkor

$$\begin{aligned} E(\bar{X}) &= \mu \\ V(\bar{X}) &= \sigma^2/N \\ E(s^2) &= \sigma^2 \end{aligned}$$

1.2. Statisztikai következtetés (inference)

- Az alapprobléma: van egy adathalmaz, ami tartalmazza a méréseket. Ezek $X_1, X_2, \dots, X_N \sim F$ független, azonos F eloszlást követő valószínűségi változók.
- A statisztikai következtetés feladata, hogy a minta alapján meghatározzuk az F eloszlásfüggvényt. Ezzel ekvivalens, ha F helyett a ρ sűrűségfüggvényt határozzuk meg.
- Ehhez használhatunk parametrikus és nem-parametrikus modelleket. A parametrikus modell egy olyan \mathcal{F} halmaz, ami a lehetséges PDF-eket tartalmazza:

$$\mathcal{F} = \{\rho(x|\theta) : \theta \in \Theta\},$$

ahol Θ a lehetséges paraméterek halmaza. Pl. ha normális eloszlást feltételezünk, akkor a parametrikus modell

$$\mathcal{F} = \left\{ \rho(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) : \mu \in \mathbb{R}, \sigma > 0 \right\},$$

a feladat pedig μ és σ meghatározása. Nem-parametrikus modellek azok, amelyeket nem lehet véges számú valós paraméterrel definiálni, pl. $\mathcal{F} = \{\text{az összes létező PDF}\}$.

- **bias** Egy becsült $\hat{\theta}$ paraméter esetén a bias (előítélet, torzítás) alatt a becsült paraméter várható értéke és a valódi értéke közötti különbséget értjük [6]:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

2. Bootstrap módszerek [1]

A Bootstrap módszerek arra jók, hogy egy statisztikai minta esetén meghatározzuk a mérés pontosságát. A módszer lényege a következő: ahhoz, hogy az adott mérendő paraméter (pl. átlag) hibáját meg tudjuk mondani, ismerni kellene a populáció adott paraméterét. Mivel ezt nem ismerjük, ezért a paraméter mérésének hibáját úgy becsüljük, hogy a meglévő mintából sokszor újramintavételezünk, majd az újbóli mintavételezések alapján számítjuk ki a hibát. Ezzel az eljárással egy becslést kapunk az adott paraméter eloszlására. Tehát a bootstrap módszer azzal a feltételezéssel él, hogy a (minta \rightarrow populáció) következtetés egyenértékű az (újramintavételezés \rightarrow minta) következtetéssel.

Példa. arra vagyunk kíváncsiak, hogy Magyarországon átlagosan milyen magasak az emberek. Mivel nem tudunk megmérni mindenkit, ezért kiválasztunk 1000 embert és lemérjük a magasságukat. Az átlagra a $\hat{\mu}$ értéket kapjuk. A kérdés az, hogy ez a $\hat{\mu}$ érték mennyiben tér el a valós μ értéktől? A választ úgy keressük, hogy az 1000 adatból véletlenszerűen újra mintavételezünk, így lesz sok $\hat{\mu}_i$ átlagunk és erre már kiszámíthatjuk a $V(\hat{\mu})$ szórását.

2.1. Jackknife módszer [2]

A Jackknife egy olyan mintavételezési módszer, melynek segítségével különböző statisztikai paraméterek határozhatók meg, mint pl. a variancia és a bias. A Jackknife módszer lényege, hogy az adott mintából mindig kihagyunk egy elemet, majd az így keletkező mintára újra kiszámoljuk az adott paramétert. Így kapunk egy eloszlást a becsült paraméterre, amelyből pontosabb eredményt kaphatunk.

Példa. A populáció átlagot akarjuk megbecsülni egy n elemű mintából. Ehhez kiszámoljuk minden $n - 1$ elemű almintából az \bar{x}_i átlagokat:

$$\bar{x}_i = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n x_j$$

Ezután az átlag becslése az előbb kapott értékek átlaga:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i$$

A Jackknife módszer alapján a becslés varianciája:

$$V(\bar{x}) = \frac{n-1}{n} \sum_{i=1}^n (\bar{x}_i - \bar{x})^2$$

3. Maximum likelihood

A maximum likelihood módszer egy olyan becslési eljárás, amelynek segítségével egy parametrikus modell paramétereinek értékét próbáljuk a minta alapján meghatározni. Ehhez felírjuk az ún. likelihood-függvényt, ami azt fejezi ki, hogy a mért adatok esetén mekkora a valószínűsége a θ paramétereknek. Ha a változó eloszlása ismert, akkor ezzel megadható a likelihood függvény:

$$\text{diszkrét változóra: } \mathcal{L}(\theta) = P(X = x|\theta)$$

$$\text{folytonos változóra: } \mathcal{L}(\theta) = \rho(x|\theta)$$

A gyakorlatban sokszor a log-likelihood függvényt vagy az átlagolt log-likelihood függvényt használjuk:

$$\ell(\theta|x) = \ln \mathcal{L}(\theta|x)$$

$$\hat{\ell}(\theta|x) = \frac{1}{N} \ln \mathcal{L}(\theta|x)$$

A maximum likelihood módszer lényege, hogy megkeressük azt a θ paramétert, ami a likelihood függvényt maximalizálja:

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta|x_1, x_2, \dots, x_N)$$

Példa: normális eloszlás paraméterei

$$\rho(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Ezért egy N elemű minta esetén azt feltételezve, hogy a mintát egy normális eloszlást követő populációból vesszük, az N minta sűrűségfüggvénye:

$$\begin{aligned}\rho(x_1, x_2, \dots, x_N | \mu, \sigma^2) &= \prod_{i=1}^N \rho(x_i | \mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^N (x_i - \mu)^2}{2\sigma^2}\right) = \mathcal{L}(\mu, \sigma^2)\end{aligned}$$

A log-likelihood függvény pedig

$$\ell(\mu, \sigma^2) = \ln \mathcal{L}(\mu, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2$$

Ahhoz, hogy megkapjuk a $\hat{\mu}_{MLE}$ és $\hat{\sigma}_{MLE}$ becsült paramétereket, az alábbi két egyenletet kell megoldani:

$$\begin{aligned}\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} &= 0 \\ \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma} &= 0\end{aligned}$$

Ha ezeket megoldjuk, az jön ki, hogy $\hat{\mu}_{MLE} = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ és $\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$.

4. Extrém statisztikák

Az extrém statisztikák (extrémérték-elméletek) olyan események modellezésére alkalmazhatók, amelyek extrém ritkán fordulnak elő: pl. szökőár, tűzsdei összeomlás, 100 éves hőmérsékleti rekord, stb. [7].

Egy egyszerű egyváltozós modell a következő képpen írható le: Legyen I_n annak a valószínűsége, hogy n esemény során bekövetkezik az extrém érték. Ezt úgy modellezzük, hogy veszünk X_1, \dots, X_n iid változót az F eloszlásból. Ekkor $M_n = \max(X_1, \dots, X_n)$ esetén:

$$P(M_n \leq z) = P(X_1 \leq z, \dots, X_n \leq z) = (F(z))^n$$

Az $I_n = I_n(M_n \geq z)$ így binomiális eloszlást követ, amelyre $p(z) = 1 - (F(z))^n$. Az egy extrém esemény bekövetkezéséhez szükséges próbálkozások száma pedig geometriai eloszlást követ ugyanezzel a $p(z)$ paraméterrel.[8]

5. Post-hoc analízis

A post-hoc analízis azt jelenti, hogy csak azután választjuk ki a teszt statisztikát, miután már az adatokat ismerjük. ("post-hoc" = "ez után"). Ez a módszer többféle problémához vezethet, bővebben: [9, 10]

6. Regresszió

Tegyük fel, hogy a megfigyelt adathalmaz $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$. Ekkor az X a független változó (feature variable, predictor, regressor), Y pedig a függő változó (outcome, response variable). Az $r(X)$ **regressziós függvény** az Y várható értéke X függvényében:

$$r(x) = E(Y|X = x)$$

- Ha $r(x)$ megadható véges számú valós paraméterrel, akkor **parametrikus regresszióról** beszélünk.
- Ha Y meghatározása a cél, ismert X esetén, akkor **predikcióról** beszélünk.
- Ha Y diszkrét (pl. kutya vagy macska látható a képen), akkor **klasszifikációról** beszélünk.
- Ha a cél az $r(x)$ görbe meghatározása, akkor **regresszióról** vagy **görbeillesztésről** beszélünk.

7. Hipotézistesztesztelés

A hipotézistesztesztelés lényege, hogy a rendelkezésre álló adatok alapján egy feltevés (hipotézis) igazságtartalmára akarunk kijelentést tenni. Fontos, hogy a feltevés a teljes populációra vonatkozik, tehát egy hipotézis elfogadásának vagy elutasításának mindig van valamennyi bizonytalansága. Példák: 1. Dobunk egy dobókockával 100-szor, majd feltesszük azt a hipotézist, hogy a dobott számok egyenletes eloszlást követnek. 2. Megnézzük 1000 ember jövedelmét majd feltételezzük, hogy a jövedelem olyan lognormális eloszlást követ, amelyre $\mu = 150000\text{HUF}$ és $\sigma = 50000\text{HUF}$.

Formális definíció:

- A Θ paraméter-teret felosztjuk Θ_0 és Θ_1 -re. ($\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$)
- A **nullhipotézis** $H_0 : \theta \in \Theta_0$
- Az **alternatív hipotézis** $H_1 : \theta \in \Theta_1$
- Az összegyűjtött $X \in \mathcal{X}$ adatok alapján akarjuk a hipotézist eldönteni.
- Az elvetési régió (rejection region) egy $R \subset \mathcal{X}$ halmaz, amelyre

$$X \in R \Rightarrow H_0\text{-t elvetjük és elfogadjuk } H_1\text{-et}$$

$$X \notin R \Rightarrow H_0\text{-t elfogadjuk}$$

- Egy tesztfüggvény (test statistic) egy $T : \mathcal{X} \rightarrow \mathbb{R}$ függvény, amelyre az elvetési régió így írható:

$$R = \{x \in \mathcal{X} | T(x) > c\},$$

ahol c a teszt kritikus értéke.

- A hipotézis tesztelés lényege keresni egy olyan T -t és c -t, amivel a legjobb (legkevésbé káros) döntést hozhatjuk.

- **Egyoldalú teszt:** Ha X a rendelkezésre álló adathalmaz, és $\theta \in \Theta$ egy paraméter (pl. átlag, variancia, stb), akkor egy egyoldalú teszt a következőt jelenti:

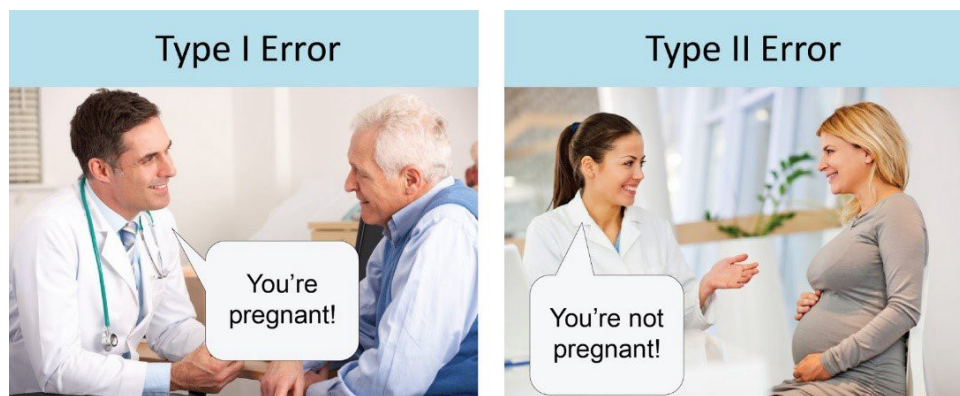
$$H_0 : \theta = \theta_0 \text{ és } H_1 : \theta < \theta_0 \text{ (vagy } > \theta_0)$$

- **Kettős hipotézisteszt:** Ha X a rendelkezésre álló adathalmaz, és $\theta \in \Theta$ egy paraméter (pl. átlag, variancia, stb), akkor egy kettős teszt a következőt jelenti:

$$H_0 : \theta = \theta_0 \text{ és } H_1 : \theta \neq \theta_0$$

- **Hibák**

	H_0 igaz	H_0 hamis
Elfogadjuk H_0 -t	OK	type 2 hiba (false positive)
Elvetjük H_0 -t	type 1 hiba (false negative)	OK



1. ábra. A két hibatípus.

- **Statisztikai erő:** Egy kettős hipotézisteszt statisztikai ereje nem más, mint az a valószínűség, hogy a teszt helyesen veti el a nullhipotézist, amikor az alternatív hipotézis igaz:

$$\text{statisztikai erő} = \beta = P(H_0 \text{ elutasítva} \mid H_1 \text{ igaz})$$

- **szignifikancia-szint** Egy statisztikai teszt α szignifikancia szintű, ha a nullhipotézis elvetésének a valószínűsége α feltéve, hogy a nullhipotézis igaz.

$$\alpha = P(H_0\text{-t elvetjük} \mid H_0 \text{ igaz})$$

- **p-érték:** A p-érték azt mutatja meg, hogy ha igaz a nullhipotézis, akkor mekkora valószínűséggel kapunk olyan eredményt, amely legalább annyira extrém, mint a mért eredmény. Pl. Ha egy minta átlaga \bar{X} , akkor a $p = 5\%$ azt jelenti, hogy ha a populáció átlaga μ , akkor 5% valószínűséggel mérhetek legalább $|\bar{X} - \mu|$ eltérést. A p-értéket így is definiálhatjuk:

$$p = P(\text{ezt az eredményt mérem} \mid H_0 \text{ igaz})$$

7.1. Egymintás u -próba [3]

(angolul z-test [11])

- Egymintás u -próbával a minta alapján a populáció átlagára vonatkozó hipotézist lehet tesztelni.
- A populációt normális eloszlásúnak feltételezzük, melynek ismerjük a σ szórását.
- Ha a σ szórás nem ismert, akkor a u -próba helyett t -próbát kell végezni.
- Azt akarjuk megvizsgálni, hogy a minta alapján a μ érték tekinthető-e a populáció átlagának.
- A teszt elvégzéséhez kiszámoljuk a minta \bar{X} átlagát, majd ebből a próbastatisztikát:

$$u = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- Kihaszználjuk azt a feltételt, hogy az u próbastatisztika standard normális eloszlást követ, és adott előre meghatározott p -értékre kiszámoljuk a $u_{p/2} = \Phi^{-1}(1 - p/2)$ -t.
- Ha $|u| \geq u_{p/2}$, akkor ez azt jelenti, hogy a mért \bar{X} és a feltételezett μ érték között p szignifikancia-szint mellett az eltérés szignifikáns. Tehát elvetjük a nullhipotézist, miszerint a populáció átlaga μ .
- **Példa:** Begyűjtünk 3 tonna almát, amiről tudjuk, hogy az almák tömegének szórása $\sigma = 40g$. Találomra kiválasztunk 100db almát, amelyeket megmérve azt kapjuk, hogy a tömegek átlaga $\bar{X} = 150g$. Kijelenthető-e 95%-os biztonsággal, hogy a begyűjtött almák átlagos tömege $\mu = 140g$? És 99%-os biztonsággal?

$$- H_0 : \mu = 140g$$

$$- H_1 : \mu \neq 140g$$

$$- p_1 = 0.05, p_2 = 0.01$$

$$- u_{p_1/2} = \Phi^{-1}(1 - p_1/2) = \Phi^{-1}(1 - 0.05/2) \approx 1.96$$

$$- u_{p_2/2} = \Phi^{-1}(1 - p_2/2) = \Phi^{-1}(1 - 0.01/2) \approx 2.58$$

$$- u = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{150 - 140}{40/10} = 2.5$$

$$- \text{WTF???$$

7.2. Egymintás t -próba [4]

- A vizsgált valószínűségi változó normális eloszlást követ, de nem ismerjük a σ szórását.
- A kérdés ugyanaz, mint az előbb, hogy a populáció átlaga megegyezik-e statisztikai szempontból a feltételezett μ értékkel.

- t -próba esetén a próbastatisztika a következő lesz:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}},$$

ahol $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ a minta szórása. A próbastatisztika ebben az esetben Student t -eloszlást követ, $f = n - 1$ szabadsági fokkal.

- Az előbbihez hasonlóan az adott p szignifikancia szinthez megkeressük a t_p értéket.
- Ha $|t| \geq t_p$, akkor a nullhipotézist elvetjük, mert a minta átlaga szignifikánsan eltér a μ értéktől.
- Ellenkező esetben, vagyis ha $|t| < t_p$, a nullhipotézist megtartjuk.

8. Függelenségvizsgálat, χ^2 -próba

A függetlenségvizsgálat arra szolgál, hogy eldöntsük, két diszkrét változó között, adott konfidenciaszinten van-e összefüggés, vagy ez csupán a véletlen mintavételezés következménye. A változók lehetnek nominálisak (pl. fizetés) vagy kategoriálisak (pl. nem). Folytonos változók (pl. életkor, fizetés, magasság) esetén ezeket diszkrété tesszük, úgy hogy felosztjuk véges számú intervallumra. Ahhoz, hogy megállapítsuk, a mért gyakoriságok szignifikánsan eltérnek-e a véletlen mértékétől, ún. Pearson-féle χ^2 -próbát kell végezni. Ennek a menete a következő:

- első lépésben kiszámoljuk a teszt statisztikát: $\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$, ahol
 - O_i a mért értékek
 - E_i a várt elméleti értékek
 - χ^2 a teszt-statisztika, ami aszimptotikusan a χ^2 eloszláshoz tart
- Megállapítjuk a szabadsági fokok számát. Függelenségvizsgálat esetén ez $df = (R - 1)(C - 1)$, ahol R az egyik C pedig a másik változó kategoriáinak száma (pl. ha azt vizsgáljuk, hogy van-e összefüggés a nemek és a magasság között, akkor $R = 2$ és $C = 10$, ha a magasságot 10 csoportra osztjuk).
- Meghatározunk egy szignifikancia szintet pl 95%.
- Összehasonlítjuk a kiszámolt χ^2 értéket kritikus χ^2_t értékkel. χ^2_t függ a választott szignifikancia szinttől és a szabadsági fokok számától.
- A két érték összehasonlításával elfogadjuk, vagy elvetjük a nullhipotézist (a nullhipotézis általában az, hogy a két változó között adott szignifikancia szinten van összefüggés.)

Konkrét számolások példákat ebben a jegyzetben lehet találni: [\[12\]](#)

Hivatkozások

- [1] Bootstrapping Wikipedia [https://en.wikipedia.org/wiki/Bootstrapping_\(statistics\)](https://en.wikipedia.org/wiki/Bootstrapping_(statistics)).
- [2] Jackknife resampling https://en.wikipedia.org/wiki/Jackknife_resampling.
- [3] Egymintás u-próba https://hu.wikipedia.org/wiki/Egymintás_u-próba.
- [4] Egymintás t-próba https://hu.wikipedia.org/wiki/Egymintás_t-próba.
- [5] Bessel's correction https://en.wikipedia.org/wiki/Bessel%27s_correction.
- [6] Bias of an estimator https://en.wikipedia.org/wiki/Bias_of_an_estimator.
- [7] Extrémérték-elmélet <https://hu.wikipedia.org/wiki/Extrémérték-elmélet>.
- [8] Extreme value theory https://en.wikipedia.org/wiki/Extreme_value_theory#Univariate_theory.
- [9] Post hoc analysis https://en.wikipedia.org/wiki/Post_hoc_analysis.
- [10] Data dredging https://en.wikipedia.org/wiki/Data_dredging.
- [11] Z-test <https://en.wikipedia.org/wiki/Z-test>.
- [12] Függetlenségvizsgálat χ^2 -próbával http://www.sze.hu/~harmati/valszam/vizsga/khi_fugg_vizsg.pdf.