

## 2. tétel

Nagy Dániel

2019. június 9.

## Kivonat

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

# 1. Bevezetés

## 1.1. Valószínűesszámitás 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  $\Omega$  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  $\Omega$ -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  $\Omega$  egy felosztása, akkor

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

- Bayes-tétel: Ha  $A_1, A_2, \dots$  az  $\Omega$  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 [1].
- **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  $\rho$  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  $\Theta$  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  $\mu$  és  $\sigma$  meghatározása. Nem-parametrikus modellek azok, amelyeket nem lehet véges számú valós paraméterrel definiálni, pl.  $\mathcal{F} = \{\text{az összes létező PDF}\}$ .

## 2. Bootstrap módszerek

### 2.1. Jackknife módszer

## 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  $\theta$  paramétereknek. Ha a változó eloszlása ismert, akkor ezzel megadható a likelihood függvény:

$$\begin{aligned}\text{diszkrét változóra: } \mathcal{L}(\theta) &= P(X = x|\theta) \\ \text{folytonos változóra: } \mathcal{L}(\theta) &= \rho(x|\theta)\end{aligned}$$

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

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

A maximum likelihood módszer lényege, hogy megkeressük azt a  $\theta$  paramétert, ami a likelihood függvény maximumizálja:

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

**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

## 5. Post-hoc analízis

## 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  $\Theta$  paraméter-teret felosztjuk  $\Theta_0$  és  $\Theta_1$ -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

$$\begin{aligned} X \in R &\Rightarrow H_0\text{-t elvetjük és elfogadjuk } H_1\text{-et} \\ X \notin R &\Rightarrow H_0\text{-t elfogadjuk} \end{aligned}$$

- 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

- **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:

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

## 7.1. z-teszt

## 7.2. t-test

## 7.3. Konfidenciaintervallumok

# 8. Függetlenségvizsgálat, $\chi^2$ -próba

## Hivatkozások

- [1] Bessel's correction [https://en.wikipedia.org/wiki/Bessel%27s\\_correction](https://en.wikipedia.org/wiki/Bessel%27s_correction).