2. tétel

Nagy Dániel

2019. június 8.

Kivonat

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

1. Bevezetés

1.1. Valószínűségszámítás alapfogalmak

- **Eseménytér** (ez egy abstrakt fogalom): $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ pl. kockadobás esetén $\Omega = \{\omega_1 = \text{"1est dobok"}, \omega_2 = \text{"2est dobok"}, \omega_3 = \text{"párosat dobok"}...\}$
- **Valószínűségi változó**: $X : \Omega \to \mathbb{R}$ pl. kockadobás esetén $X(\omega_1) = 1, X(\omega_2) = 2, ...$
- **Valószínűség**: P egy mérték, amely Ω részhalmazaihoz számot rendel:
  - $P : \mathcal{P}(\Omega) \to \mathbb{R}$
  - $P(\Omega) = 1$ és $P(\varnothing) = 0$
  - $0 \leq P(A) \leq 1 \ \forall A \in \Omega$
  - Ha $A_1, A_2, ...$ 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 $\Longleftrightarrow P(A \cap B) = P(A)P(B)$
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - Teljes valószínűség: Ha $A_1, A_2, ...$ az Ω egy felosztása, akkor

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

  - Bayes-tétel: Ha $A_1, A_2, ...$ az Ω egy felosztása, akkor

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

- **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

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

  Megjegyzés: sűrűségfüggvénye csak folytonos eloszlású valószínűségi változónak van.

- **Várható érték**

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

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

- **Várható értékre vonatkozó azonosságok**:
  - Ha $Y = g(X) \Rightarrow E(Y) = E(g(X)) = \int\limits_{-\infty}^{\infty} g(x)\rho(x)\mathrm{d}x$
  - $E\left(\sum\limits_k a_k X_k\right) = \sum\limits_k a_k E(X_k)$
  - Ha $X_1, X_2, ...$ független változók, akkor $E\left(\prod\limits_k X_k\right) = \prod\limits_k E(X_k)$

- **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 $X_1, X_2, ...$ 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 error 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**: $\overline{X} = \dfrac{1}{N}\sum\limits_{i=1}^{N} X_i$

- **Minta varianciája**: $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$, standard hibája $SE = \sqrt{s^2}$.
- Egy minta esetében $\overline{X}, s^2, SE$ 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

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

1.2. Statisztikai következtetés (inference)

- Az alapprobléma: van egy adathalmaz, ami tartalmazza a méréseket. Ezek $X_1, X_2, ..., X_N \sim F$ 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

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

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

2. Bootstrap módszerek

2.1. Jackknife módszer

3. Maximum likelihood

4. Extrém statisztikák

5. Post-hoc analízis

6. Regresszió

7. Hipotézis tesztelés

8. z-teszt, t-test

8.1. Konfidenciaintervallumok

8.2. Függetlenségvizsgálat, χ^2-próba

Hivatkozások