Valószínűségszámítás

2021. december 8.

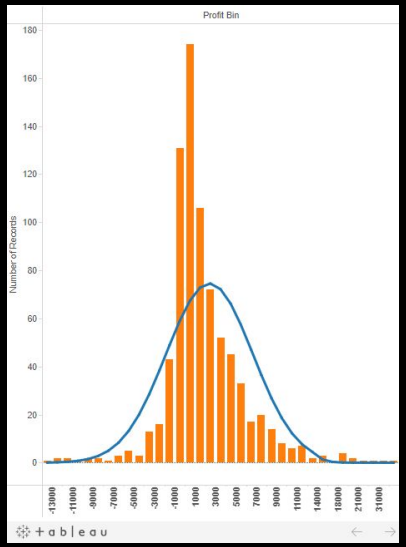
Mészáros Szabolcs

Tárgyhonlap:
cs.bme.hu/valszam

A prezentáció anyagát és az abból készült videofelvételt a tárgy hallgatói jogosultak használni, kizárólag saját célra. A felvétel másolása, videómegosztókra való feltöltése részben vagy egészben tilos, illetve csak a tantárgyfelelős előzetes engedélyével történhet.

Irodalomjegyzék

- Bolla M., Krámli A., Nagy-György J. - Többváltozós statisztikai módszerek
- J. L. Devore, K. N. Berk - Modern Mathematical Statistics with Application
- G. James, D. Witten, T. Hastie, R. Tibshirani - Intro to Statistical Learning
- R. W. Keener - Theoretical Statistics



Statisztika, problémafelvetés

Főhősünk, a 19. századi Kincskereső Kiss Ödön hallott az aranylázról, és fontolgatja, hogy ő is útnak indul. Néhány ismerőséről már tudja, milyen mennyiségű aranyat gyűjtöttek össze egy-egy út alatt. Hogyan tudná ezek alapján

a) megbecsülni az összeszedhető vagyon várható értékét, illetve
b) eldönteni, megéri-e költeni erre az utazásra?

becsléselmélet

hipotézisvizsgálat

Várható érték becslése

Adott: $X_1, X_2, \ldots, X_n$ független, azonos eloszlású v. v.

Ötlet 1: $\dfrac{\min_i(X_i) + \max_i(X_i)}{2}$ Egyenletes elo.-ra jó, másra nem.

Ötlet 2: $\overline{X}_n = \dfrac{1}{n}(X_1 + \cdots + X_n)$

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}(X)$$

$$\lim_{n\to\infty} \overline{X}_n = \mathbb{E}(X)$$

Ötlet 3: sorba rendezzük a mintát, és

nézzük a (tapasztalati) mediánt (≃ középsőt)

Szórásnégyzet becslése

Tapasztalati szórásnégyzet: $$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Szimulálva látható, pl: $X_1, X_2, \ldots, X_{10} \sim B(5;0{,}4)$

$s_X^2 \approx 1{,}08$ pedig $n \cdot p \cdot (1-p) = 5 \cdot 0{,}4 \cdot 0{,}6 = 1{,}2$

Állítás: $$\mathbb{E}(s_X^2) = \frac{n-1}{n}\mathbb{D}^2(X_1)$$

Korrigált tap. szórás: $$\frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\big((X_i - \overline{X}_n)^2\big)$$

Korrigált szórásnégyzet

Bizonyítás: $\mathbb{E}(s_X^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big((X_i - \overline{X}_n)^2\big)$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big(X_i^2 + \overline{X}_n^2 - 2X_i\overline{X}_n\big)$$

$$= \mathbb{E}(X_1^2) + \mathbb{E}(\overline{X}_n^2) - 2\cdot\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}X_i\overline{X}_n\Big)$$

$$= \mathbb{E}(X_1^2) - \mathbb{E}(\overline{X}_n^2)$$

$$\mathbb{E}(\overline{X}_n^2) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(X_i^2)$$

$$+\frac{2}{n^2}\sum_{i<j}\mathbb{E}(X_iX_j)$$

$$= \frac{n}{n^2}\mathbb{E}(X_1^2) + \frac{n(n-1)}{n^2}\mathbb{E}(X_1)^2$$

Paraméter becslése, általánosan

Definíció: $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$

statisztikai mező, ha minden $\theta \in \Theta$ esetén $(\Omega, \mathcal{F}, P_\theta)$ val. mező.

Minta: $X_1, \ldots, X_n$ független, azonos eloszlású val. változók.

Statisztika: $T(X_1, \ldots, X_n)$ val. változó, ahol $T : \mathbb{R}^n \to \mathbb{R}$

Torzítatlan a $\psi(\theta)$ paraméterre nézve, ha $\mathbb{E}(T(X_1, \ldots, X_n)) = \psi(\theta)$

Függ θ-tól

Példa: $\Theta = \{(a, b) \mid 0 < a < b < 1\}$

$P_{(a,b)} : X_1 \sim U(a, b) \quad \psi(a, b) = \frac{a+b}{2} \qquad T$: lásd első dia

Torzítatlanság, megjegyzések

- A szórásnak nem torzítatlan becslése a korrigált szórásnégyzet gyöke.
- Adott paraméterhez nem feltétlenül létezik torzítatlan becslése
  pl. eloszlás mediánjának becslése
- Van amikor létezik torzítatlan becslés, csak rossz:

$$\Theta = (0, \infty) \qquad P_\theta : X_1 \sim \mathrm{Pois}(\theta) \qquad \psi(\theta) = \mathbb{P}_\theta(X_1 = 0)^2$$

$$T(X_1) \text{ torzítatlan} \Rightarrow T(X_1) = (-1)^{X_1}$$

- “Bias-Variance trade-off”:

$$\mathbb{E}\big((T(X) - \psi(\theta))^2\big) = \mathbb{D}^2\big(T(X)\big) + \big(\mathbb{E}(T(X)) - \psi(\theta)\big)^2$$

A becslés átlagos négyzetes hibája — becslés szórásnégyzete — torzítás négyzete

Becslés, példa

Legyen $X_1, \ldots, X_n \sim \mathbf{1}(p)$
ahol p a becsülendő paraméter.

$$T_1(X) = \overline{X}_n \qquad T_2(X) = \frac{\sum_{i=1}^{n} X_i + 2}{n+4}$$

$$\mathbb{E}(T_1(X)) = p \qquad \mathbb{E}(T_2(X)) = \frac{np+2}{n+4} = p + \frac{2-4p}{n+4}$$

$$\mathbb{D}^2(T_1(X)) = \frac{p(1-p)}{n} \qquad \mathbb{D}^2(T_2(X)) = \frac{np(1-p)}{(n+4)^2}$$

Becslés, példa

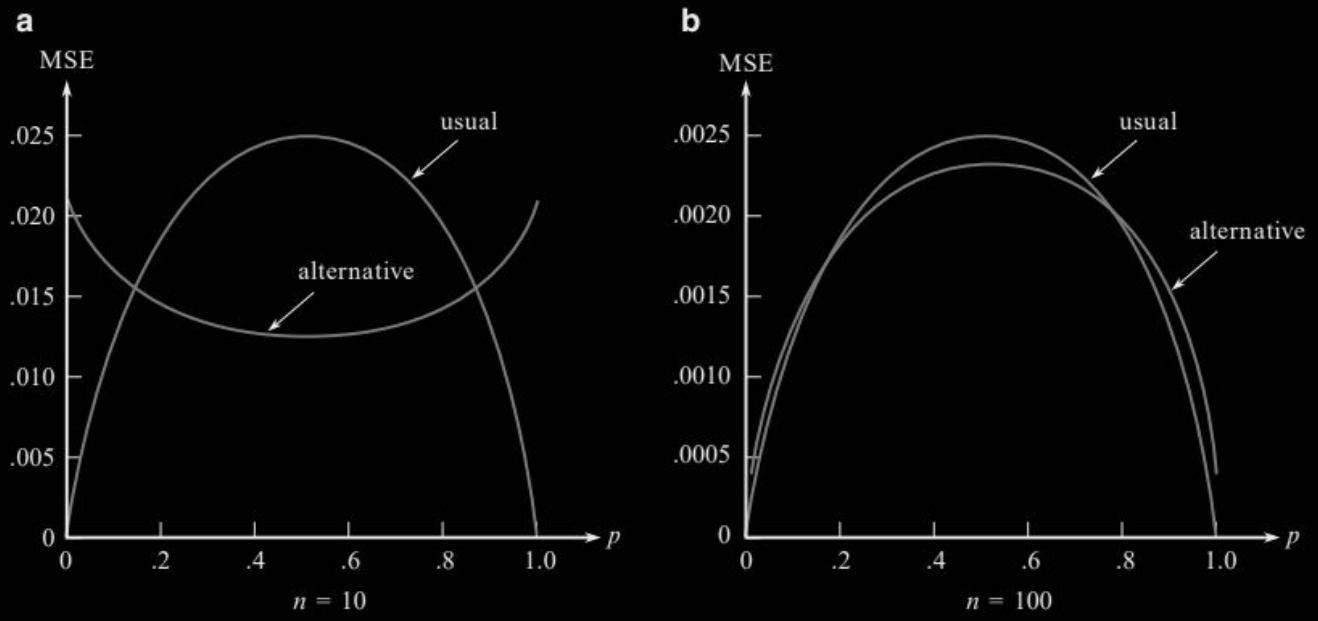


Figure 7.1 Graphs of MSE for the usual and alternative estimators of p ■

Maximum likelihood becslés

Intuitív kérdés: melyik θ paraméter a legvalószínűbb, ha ezt a mintát látom.

Probléma: θ nem val. változó, nincs “legvalószínűbb” értéke.

$$\mathbb{P}(\theta = t \mid X = x) \to \max_t$$

Válasz 1: Tegyük fel, hogy mégis val. változó => Bayes-becslések

Válasz 2: új kérdés, melyik θ esetén a legvalószínűbb, hogy ezt a mintát látom?

$$\mathbb{P}_\theta(X = x) \to \max_\theta$$

=> Maximum-likelihood becslés

Maximum-likelihood becslés

Folytonos esetben: $f_{\underline{X},\theta}(\underline{x}) \to \max_{\theta}$ $\qquad f_{\underline{X},\theta}(\underline{x}) = \prod_{i=1}^{n} f_{X_i,\theta}(x_i)$

Ekvivalensen, maximalizáljuk: $\ln f_{\underline{X},\theta}(\underline{x}) = \sum_{i=1}^{n} \ln f_{X_i,\theta}(x_i)$

Példa: $X_1, \ldots, X_n \sim \mathrm{Exp}(\theta)$

$$= \sum_{i=1}^{n} \ln(\theta e^{-\theta x_i}) = n \ln(\theta) - \theta \sum_{i=1}^{n} x_i \overset{\partial_\theta}{\Longrightarrow} \frac{n}{\theta} = \sum_{i=1}^{n} x_i$$

A paraméter M-L becslése az átlag reciproka.

Hipotézisvizsgálat, példa

Tesztelni szeretnénk, hogy egy adott szolgáltatás esetében egy változtatás növeli-e a szolgáltatás addigi 25%-os sikerarányát.

$$X_1, \dots, X_{20} \sim \mathbf{1}(p) \qquad T(X) : \text{sikerek száma}$$

$$T(X) > c \Rightarrow \text{javít} \qquad c = ?$$

fals pozitív, elsőfajú hiba (Type I)

$$\mathbb{P}_{(\text{nem javít})}(T(X) > c) = 1 - \mathbb{P}\big(B(20; 0{,}25) \leq c\big)$$

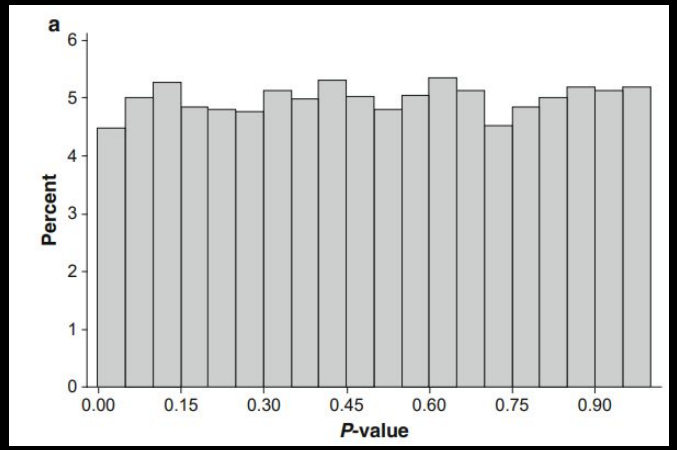
$$\mathbb{P}_{(\text{javít})}(T(X) \leq c) = \mathbb{P}\big(B(20; p) \leq c\big)$$

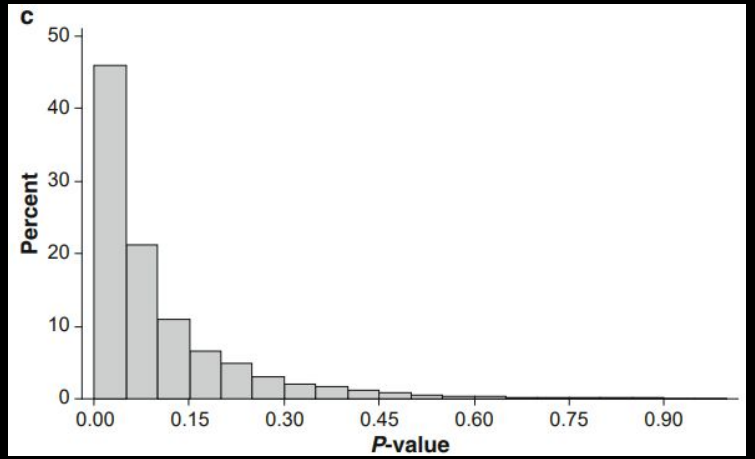
fals negatív, másodfajú hiba (Type II)

Hipotézisvizsgálat, fogalmak

- Tesztstatisztika: $T(X)$
- Null hipotézis, H_0 : status quo
- Ellenhipotézis, H_a : találtunk valamit
- Kritikus tartomány: amilyen $T(X)$ értékek esetén elvetjük H_0 -t.
- Elsőfajú hiba: elvetjük H_0 -t, pedig igaz.
- Másodfajú hiba: nem vetjük el H_0 -t pedig hamis.
- Terjedelem: elsőfajú hiba valószínűsége, α
- Erőfüggvény: 1 - (másodfajú hiba valószínűsége), $b(\theta) = 1 - \beta$
  a paraméter függvényében
- p-érték: tegyük fel, hogy H_0 igaz, és a tesztstatisztika értéke t .
  Ekkor a p-érték $= \mathbb{P}_{H_0}(T(X) > t)$
  (“mennyire mond ellent a nullhipotézisnek”)

p-érték, vizualizálva





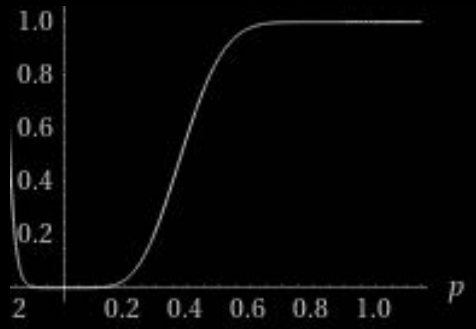
Hipotézisvizsgálat, példa

Terjedelemet rögzítjük: $\alpha = 0.1$

Ebből a kritikus érték meghatározható: $c = 7$

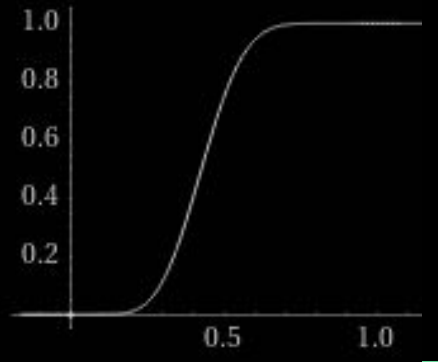
=> erő függvény:

$$b(0{,}3) = \mathbb{P}\big(B(20; 0{,}3) > 7\big) \approx 0{,}228$$



Ha ehelyett $c = 8$-at használnánk, akkor $\alpha = 0.41$

de az erőfüggvény gyengébb lesz.



u-próba (z-test)

Minta: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

ahol σ^2 ismert, μ ismeretlen paraméter

$H_0: \mu = \mu_0$ H_a : attól függ, legyen most $\mu > \mu_0$

Próbastatisztika:

$$T(X) = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma}$$

Állítás: H_0 esetén

$$T(X) \sim N(0,1)$$

Tehát $\alpha = \Phi(c)$

amiből c meghatározható.

t-próba

Kérdés: mit csináljunk, ha σ nem ismert?

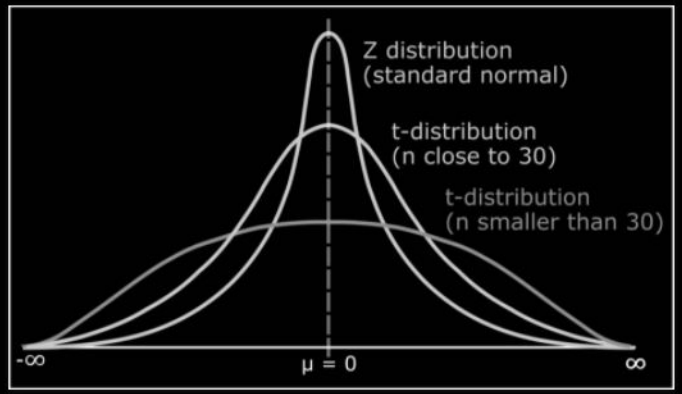
Próbastatisztika:

$$T(X) = \sqrt{n}\frac{\overline{X}_n - \mu_0}{s_X^*}$$

$$s_X^* = \sqrt{\frac{n}{n-1}s_X^2}$$

Állítás: H_0 esetén

$$T(X) \sim t_{n-1}$$



Köszönöm a figyelmet!