Matematikai statisztika

Informatika alapszak, "B" spec.

Zempléni András

Valószínűségelméleti és Statisztika Tanszék
Matematikai Intézet
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Honlap: `zempleni.elte.hu`

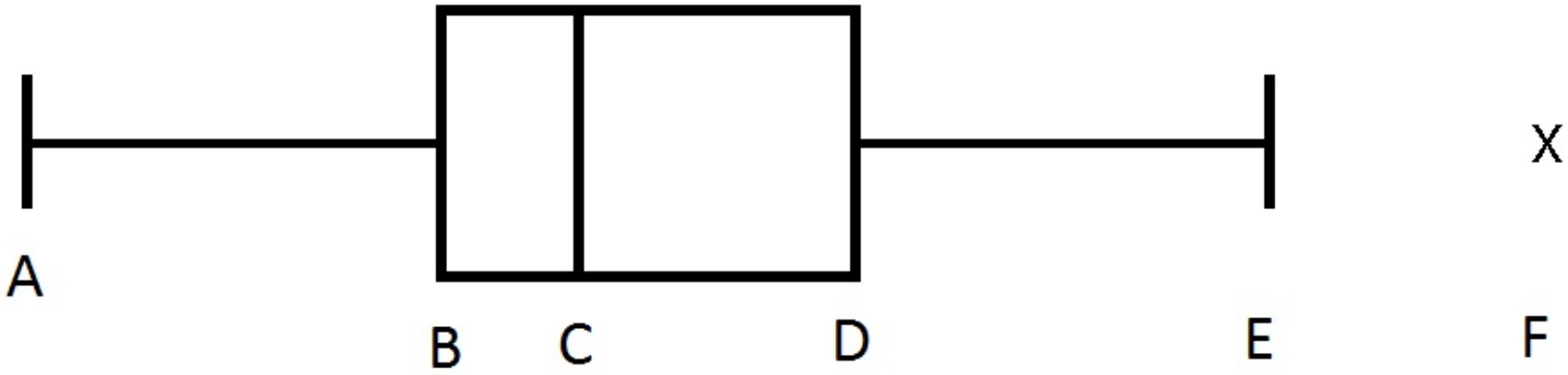
E-mail: `andras.zempleni@ttk.elte.hu`

Szoba: D 3-310

2. előadás

Fontos leíró statisztikai ábrák II

Boxplot ábra (Box&Whiskers diagram) – ez fekvő, de lehet álló is



A betűk a következő értékeket jelentik:

- $A = \max\{x_1^*, Q_1 - 1,5 \cdot IQR\}$
- $B = Q_1$ (első kvartilis)
- $C = Me$ (medián)
- $D = Q_3$ (harmadik kvartilis)
- $E = \min\{x_n^*, Q_3 + 1,5 \cdot IQR\}$
- F: kiugró érték (outlier) $\rightarrow$ azokat az adatpontokat tüntetjük fel, amik A-n vagy E-n kívülre esnek

ahol $IQR = Q_3 - Q_1$ az interkvartilis terjedelem

Tapasztalati eloszlás

- **Tapasztalati eloszlás**: minden megfigyeléshez azonos, $\frac{1}{n}$ súlyt rendelünk $\Rightarrow$ ez egy diszkrét eloszlás
- A mintaátlag éppen ennek a várható értéke
- A tapasztalati eloszlás eloszlásfüggvényét hívjuk **tapasztalati eloszlásfüggvény**nek, ami egy tiszta ugrófüggvény, értéke minden mintaelem helyén $\frac{1}{n}$ nagyságot ugrik felfelé.
  A tapasztalati eloszlásfüggvény az x helyen:

$$\frac{I(x_1 < x) + I(x_2 < x) + \ldots + I(x_n < x)}{n} = \frac{\sum_{i=1}^{n} I(x_i < x)}{n}$$

Azt mutatja meg, hogy a mintaelemek hányad része kisebb x-nél.

Matematikai statisztika: bevezető

- Statisztikai mező: $(\Omega, \mathcal{A}, P_\vartheta) : \vartheta \in \Theta$
- Paramétertér: Θ. Ez lehet egydimenziós, de akár végtelen dimenziós is
- Minta: $X = (X_1, \ldots, X_n)$ független, azonos F eloszlású valószínűségi változók
- Mintatér $\mathcal{X} : \mathbb{R}^n$ azon része, ahova a mintaelemek eshetnek
- A mintaelemek eloszlása ismeretlen, de paraméterezhető: $F \to F_\vartheta$
- Példák:
  - Poisson eloszlású minta, ekkor $\vartheta \to \lambda \in \Theta = (0; \infty)$
  - normális eloszlású minta, ekkor $\vartheta \to (\mu, \sigma) \in \Theta = (-\infty; \infty) \times (0; \infty) \subset \mathbb{R}^2$
  - F-ről nem tudunk semmit, ekkor Θ végtelen dimenziós. De ekkor is lehet egydimenziós paramétereket értelmezni, például várható érték, szórás

Motiváció – becsléselmélet

Az Asus kicseréli táblagépeit, amennyiben a vevők 8-nál több pixelhibát jelentenek be vásárlástól számítva 3 napon belül. A Samsung már egyetlen, 3 napon belül bejelentett pixelhiba esetén is új készüléket biztosít. A Sony-nál legalább 2 pixelhiba esetén jár új táblagép.
Hogyan tudnánk megbecsülni, hogy a gyártónak éves szinten milyen mértékű vesztesége származik ezekből a cserékből?

- Kulcskérdés: mi az esélye, hogy egy, a gyártósorról véletlenszerűen leemelt készüléket pixelhiba miatt ki kell cserélni?
- Ha X a pixelhibák száma, akkor a kérdéses valószínűség például a Sony-nál: $P(X \geq 2)$
- Milyen eloszlású lehet X (Poisson?) $\rightarrow$ *illeszkedésvizsgálat*
- Ha tudom, hogy Poisson-eloszlású, akkor hogyan becsüljem meg a paramétert? $\rightarrow$ *pontbecslés*
- Milyen intervallumban lesz "nagy" valószínűséggel a becsült paraméter? $\rightarrow$ *intervallumbecslés*
- Ezután készíthető a kérdéses valószínűségre intervallumbecslés, abból pedig egy intervallumbecslés a várható veszteségre.

Forrás: `http://www.origo.hu/techbazis/20130329-ra-lehet-fazni-a-pixelhibas-okosmobillal.html`

Becslések és alapdefiníciók

- Legyen $\mathsf{X} = (X_1, \ldots, X_n)$ i.i.d. minta egy ϑ valós paraméterű eloszláscsaládból. $T : \mathcal{X} \to \mathbb{R}$ becslés ϑ-ra.
- Tulajdonságai:
  - Torzítatlanság: $E_\vartheta T(\mathsf{X}) = \vartheta$ minden $\vartheta \in \Theta$ paraméterre
  - Aszimptotikus torzítatlanság: $E_\vartheta T_n(\mathsf{X}) \to \vartheta$ (ha $n \to \infty$) minden $\vartheta \in \Theta$ paraméterre
  - Konzisztencia: $T_n(\mathsf{X}) \to \vartheta$ sztochasztikusan (ha $n \to \infty$) minden $\vartheta \in \Theta$ paraméterre (ez a gyenge, erős, ha 1 vszgű a konv.)
- Megj.: A konzisztenciához elégséges, hogy T_n aszimptotikusan torzítatlan legyen és $D^2(T_n) \to 0$

Definíció. [Likelihood függvény] $L(\vartheta; \mathsf{x}) = f_\vartheta(\mathsf{x}) = \prod_{i=1}^{n} f_\vartheta(x_i)$, ha az eloszlás folytonos és $L(\vartheta; \mathsf{x}) = P_\vartheta(\mathsf{X} = \mathsf{x}) = \prod_{i=1}^{n} P_\vartheta(X_i = x_i)$, ha az eloszlás diszkrét

Definíció. [Log-likelihood függvény] $\ell(\vartheta; \mathsf{x}) = \ln(L(\vartheta; \mathsf{x}))$

Fontos becslések tulajdonságai

Tétel. Legyen $X_1, \ldots, X_n$ i.i.d. minta egy ϑ paraméterű eloszláscsaládból, $h : \mathbb{R} \to \mathbb{R}$ (mérhető) függvény. Tegyük fel, hogy a táblázatban szereplő összes várható érték/szórás létezik minden ϑ esetén.

Mit becsülünk? $g(\vartheta)$	Mivel becsüljük? $T_n(X)$	Torzítatlan?	Aszimptotikusan torzítatlan?	Gyengén/erősen konzisztens?
$E_\vartheta X_1$	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	igen	igen	igen
$D^2_\vartheta X_1$	$S_n^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$	nem	igen	igen
$D^2_\vartheta X_1$	$(S_n^*)^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$	igen	igen	igen
$F_\vartheta(x)$	$F_n(x) = \frac{\sum_{i=1}^{n} I(X_i < x)}{n}$	igen	igen	igen
$E_\vartheta h(X_1)$	$\frac{\sum_{i=1}^{n} h(X_i)}{n}$	igen	igen	igen

Becslési módszerek I

- **Maximum likelihood módszer** (ML-módszer): Azt a paraméterértéket keressük, ahol a likelihood függvény a legnagyobb értéket veszi fel: $\max_{\vartheta} L(\vartheta; \mathrm{x})$.
- Ez nyilván megegyezik azzal a paraméterértékkel, ahol a log-likelihood függvény veszi fel a legnagyobb értéket, azaz: $\max_{\vartheta} \ell(\vartheta; \mathrm{x})$.
- Amennyiben a függvény deriválható ϑ szerint, akkor a maximumot kereshetjük a deriváltak segítségével
- Célszerűbb a likelihood függvény helyett a log-likelihood függvény maximumhelyét keresni.
- Ha ϑ 1 dimenziós, akkor $\partial_\vartheta \ell(\vartheta, \mathrm{x}) = 0$, míg ha $\vartheta = (\vartheta_1, ..., \vartheta_p)$ p dimenziós, akkor $\partial_{\vartheta_i} \ell(\vartheta, \mathrm{x}) = 0$ megoldásából kapjuk a becslést.

Tétel. [ML-becslés invariáns tulajdonsága] Ha ϑ ML-becslése $\hat{\vartheta}$, akkor tetszőleges g függvény esetén $g(\vartheta)$ ML-becslése $g(\hat{\vartheta})$.

Becslési módszerek II

Momentum módszer:

- A mintából számítható tapasztalati momentumokat ($m_i := \frac{1}{n}\sum_j x_j^i$) egyenlővé tesszük az elméleti momentumokkal ($M_i(\vartheta) := E_\vartheta X^i$),
- mégpedig annyit, amennyiből a paramétereket meg tudjuk határozni. p darab ismeretlen paraméter esetén p ismeretlenes egyenletrendszert oldunk meg ϑ-ra: $M_1(\vartheta) = m_1, \ldots, M_p(\vartheta) = m_p$ (megjegyzés: $m_1 = \overline{x}$)
- Ha valamelyik egyenlet nem ad információt a keresett paraméterre, akkor magasabb hatványokat nézünk (míg megoldható nem lesz az egyenletrendszer)
  Példák: Határozzuk meg a momentum és a maximum likelihood becslést az alábbi eloszlások paraméterére:
  - Poisson
  - exponenciális
- Legyen X_1 Bin$(2; p)$ eloszlású (egyelemű) minta, ahol $p \in (0; 1)$ ismeretlen paraméter. Adjunk X_1 segítségével torzítatlan becslést $g(p) = \frac{1}{p}$-re!