Matematikai statisztika

Informatika alapszak, "A" 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

Szóródási mutatók számítása

Terjedelem: $R = x_n^* - x_1^*$ (x_n^* a legnagyobb, x_1^* a legkisebb mintaelem; R=range)

Interkvartilis terjedelem: $IQR = Q_3 - Q_1$

Tapasztalati szórás: az átlagtól való átlagos négyzetes eltérés négyzetgyöke

- Számítása közvetlenül az adatokból: $s_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$
- Számítása osztályközös gyakorisági sorból: $s_n = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}{n}}$

Korrigált tapasztalati szórás: az átlagtól való korrigált átlagos négyzetes eltérés négyzetgyöke

- Számítása közvetlenül az adatokból: $s_n^* = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$
- Számítása osztályközös gyakorisági sorból: $s_n^* = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}{n-1}}$
- ezt "szeretjük" a legjobban, minden szoftver, programcsomag szórás számításánál ezt veszi alapértelmezettnek

Szóródási mutatók számítása/2

Relatív szórás vagy **szórási együttható**: az átlagtól való átlagos eltérés százalékban; lehet a korrigált és a korrigálatlan tapasztalati szórásnégyzetből is számítani:

$$V = \frac{s_n^*}{\overline{x}} \text{ vagy } V = \frac{s_n}{\overline{x}}$$

Kevésbé gyakran használt, szóródást mérő mutatók:

- átlagos abszolút eltérés: $\frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$
- Gini-együttható: $G = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$.

Alakmutatók számítása

- Alakmutatók: a szórást ezeknél is választhatjuk a tapasztalati vagy a korrigált tapasztalati szórásnak egyaránt.

Tapasztalati ferdeség

- Számítása közvetlenül az adatokból: $\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{(s_n)^3}$
- Számítása osztályközös gyakoriságí sorból: $\frac{\frac{1}{n}\sum_{i=1}^{n}f_i(x_i-\overline{x})^3}{(s_n)^3}$

Tapasztalati csúcsosság

- Számítása közvetlenül az adatokból: $\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{(s_n)^4}-3$
- Számítása osztályközös gyakoriságí sorból: $\frac{\frac{1}{n}\sum_{i=1}^{n}f_i(x_i-\overline{x})^4}{(s_n)^4}-3$

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: $F \rightsquigarrow F_\vartheta$
- Példák:
  - Poisson eloszlású minta, ekkor $\vartheta \rightsquigarrow \lambda \in \Theta = (0; \infty)$
  - normális eloszlású minta, ekkor $\vartheta \rightsquigarrow (\mu, \sigma) \in \Theta = (-\infty; \infty) \times (0; \infty) \subset \mathbb{R}^2$
  - 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?) $\rightsquigarrow$ *illeszkedésvizsgálat*
- Ha tudom, hogy Poisson-eloszlású, akkor hogyan becsüljem meg a paramétert? $\rightsquigarrow$ *pontbecslés*
- Milyen intervallumban lesz "nagy" valószínűséggel a becsült paraméter? $\rightsquigarrow$ *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 $g(\hat{\vartheta})$.