Leíró és matematikai statisztika

Matematika alapszak, matematikai elemző szakirány

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

3. előadás

Szóródási mutatók számítása

Terjedelem: $R = x_n^* - x_1^*$ (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

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:

- Az átlagtól vett átlagos abszolút eltérés: $\frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$
- A mediántól vett átlagos abszolút eltérés: $\frac{\sum_{i=1}^{n} |x_i - Median(x)|}{n}$
- Átlagos abszolút eltérés: $D = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$.

Tapasztalati eloszlásfüggvény, alakmutatók számítása

- Tapasztalati eloszlásfüggvény: a tapasztalati eloszlás (minden mintaelem valószínűsége 1/n) eloszlásfüggvénye
- Alakmutatók:

Tapasztalati ferdeség

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

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 gyakorisági sorból: $\frac{\frac{1}{n}\sum_{i=1}^{k} f_i(x_i-\overline{x})^4}{(s_n)^4} - 3$

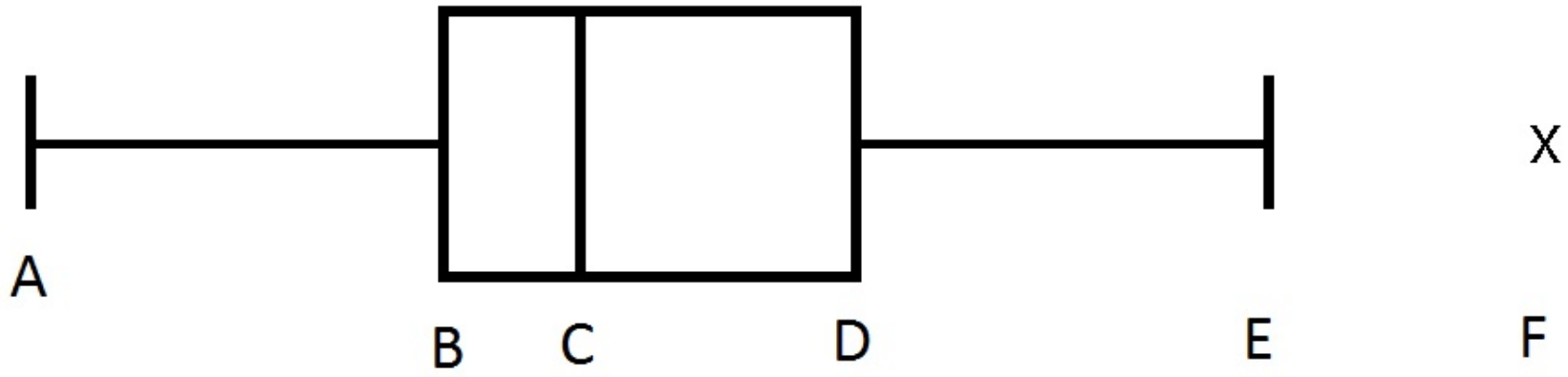
Fontos leíró statisztikai ábrák I

Hisztogram – Ha a mennyiségi ismérv folytonos vagy sok ismérvérték van, akkor alkalmas módon osztályokat képezünk, majd minden egyes adatot pontosan egy osztályhoz rendeljük. A hisztogram az osztályok gyakoriságait ábrázolja.

- az osztályok száma: k
- ha azonos hosszúságú (h) osztályközöket akarunk létrehozni, akkor $h = \frac{x_n^* - x_1^*}{k}$
- az f_i gyakoriságokat ábrázoljuk a függőleges tengelyen
- **ha az osztályközök különböző hosszúságúak, akkor a gyakoriságokat egy közös hosszra kell arányosítani**
- sűrűséghisztogramnál a $g_i = \frac{f_i}{nh_i}$ relatív gyakoriság/intervallumhossz értéket ábrázoljuk a függőleges tengelyen (területarányos, összterület=1)

Fontos leíró statisztikai ábrák II

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



A betűk a következő értékeket jelentik:

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

ahol $IQR = Q_3 - Q_1$ az interkvartilis terjedelem

Mintavétel a gyakorlatban

Az adatokkal szemben támasztott követelmények:

- pontosság – ne legyenek hibásak és a szükséges pontosságban álljanak rendelkezésre
- gyorsaság – hamar be lehessen őket szerezni
- gazdaságosság – az adatgyűjtés legyen "olcsó"

Az adatgyűjtés fajtái:

- teljes körű – például a népszámlálás
- részleges – a gyakorlatban ez a jellemző

A részleges adatgyűjtés fajtái:

- reprezentatív (mintavételes): a teljes sokaság jellemzőit megfelelően tükröző részsokaságból, ún. mintasokaságból szerezzük be az adatokat
- monográfia: egy vagy néhány kiemelt egyed részletes vizsgálata
- egyéb – például önkéntes kitöltésen alapuló internetes teszt

Az adatelemzés elemei (leíró statisztikák alk.)

1.) Adathibák keresése, irreális adatok, értékek törlése. Ha lehet, akkor a hibák korrigálása.
2.) Ha sok a különböző adat, akkor alkalmas osztályközös gyakorisági sor készítése
3.) Középértékek kiszámítása:
   - átlag (számtani vagy mértani – amelyiknek értelme van)
   - helyzeti középértékek: módusz (az osztályközös gyakorisági sorból) és medián
4.) Szóródási mutatók kiszámítása:
   - szórás és relatív szórás
   - terjedelem és interkvartilis terjedelem
5.) Alakmutatók kiszámítása:
   - ferdeség
   - csúcsosság
6.) Ábrák készítése:
   - hisztogram/sűrűséghisztogram
   - boxplot ábra
   - Lorenz-görbe (értékösszeg sor esetén)
7.) Visszacsatolás $\rightsquigarrow$ a felfedezett adathibák javítása

Feladatok

E11.) Azonos felhasználási körülmények között megmérték 15 azonos típusú mobiltelefon akkumulátorának lemerülési idejét teljes feltöltöttségről: (óra)

18	16	15	20	12	16	-15	23
14	11	17	15	200	19	18	20

a.) Nézzük át nagy vonalakban az adatokat, reálisak-e! Próbáljuk meg kijavítani az esetleges adathibákat!

b.) Ábrázoljuk a tapasztalati eloszlásfüggvényt! Számítsuk ki és értelmezzük a 16 helyen!

c.) Készítsünk alkalmas sávszélességű hisztogramot!

d.) Elemezzük a lemerülési időt az alapstatisztikák: az átlag, a korrigált tapasztalati szórás, szórási együttható és boxplot ábra (kvartilisek) segítségével! Számítsuk ki a tapasztalati ferdeséget és csúcsosságot! Értelmezzük is az eredményeket!

Megoldás (értelmezések)

a.) Adatjavítás: -15 és 200, a helyes értékek vélhetően 15 és 20

b.) Az akkumulátorok 3/8-ad része 16 óránál hamarabb merült le.

c.) Ld. korábban

d.) Az akkumulátorok átlagosan 16,8 óra alatt merültek le. Az egyes akkumulátorok lemerülési ideje az átlagos lemerülési időtől átlagosan 3,19 órával, azaz 18,96%-kal tért el.
Az akkumulátorok egynegyede legfeljebb 15 óra alatt lemerült, míg háromnegyede legalább 15 órán keresztül ébren volt. Az akkumulátorok egyik fele legfeljebb 16,5 óra alatt lemerült, míg másik fele legalább 16 és fél órán keresztül tudta árammal ellátni a telefont. Az akkumulátorok 75%-a legfeljebb 19,75 óra alatt lemerült.
Az akkumulátorok lemerülési idejének eloszlása nagyjából szimmetrikus, csúcsossága a normális eloszláséhoz viszonyítva laposabb.

Feladatok

E12.) Egy megyében a kistermelő gazdaságok termőterület szerinti megoszlása:

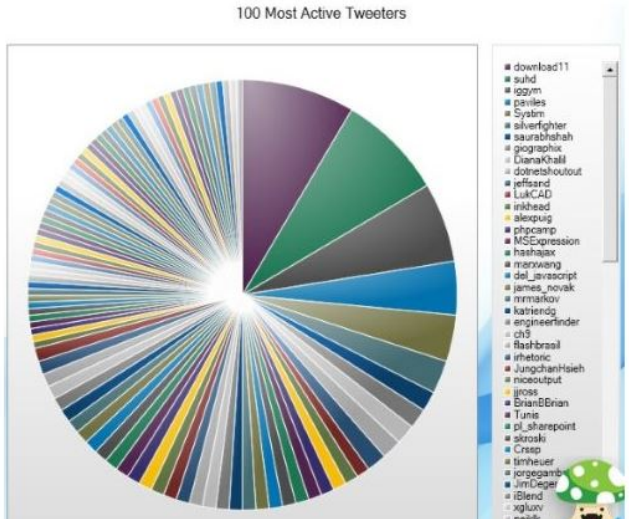
Termőterület (hektár)	Gazdaságok száma
– 4	200
4 – 10	90
10 – 20	80
20 – 30	60
30 – 60	20
Összesen	450

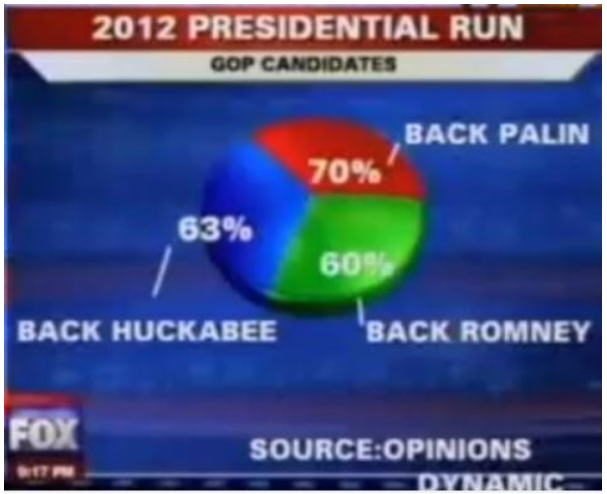
a.) Készítsünk hisztogramot! Milyen az eloszlás ferdesége?

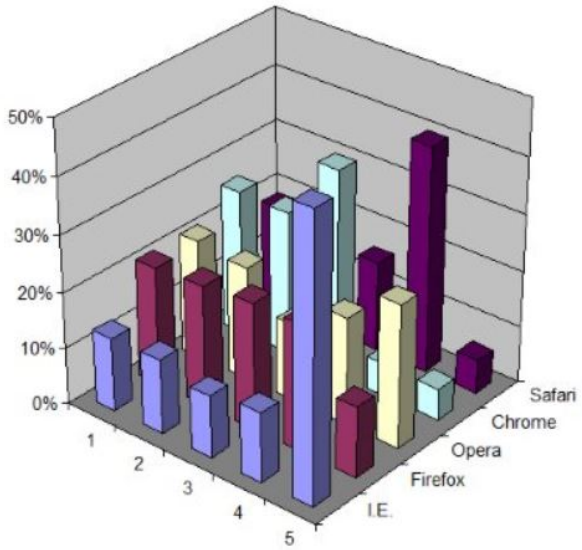
b.) Jellemezzük (szövegesen is) a kistermelők termőterület szerinti eloszlását alapstatisztikák (mintaátlag, korrigált tapasztalati szórás, tapasztalati módusz és kvartilisek) segítségével!

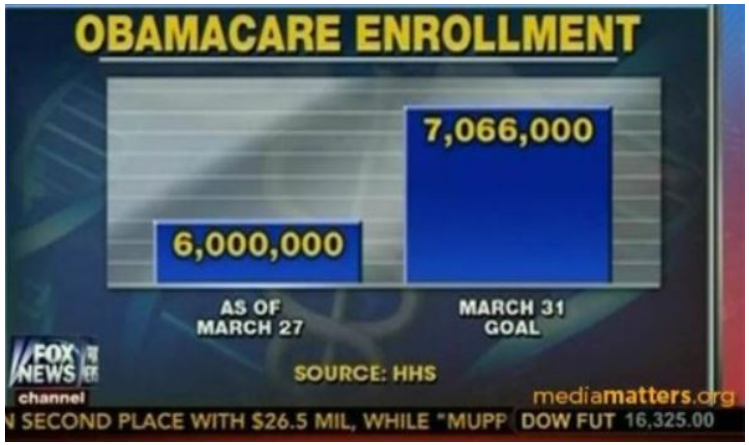
c.) Mennyire koncentrálódik a termőterület? Készítsünk Lorenz-görbét!

Rövid szünet: néhány katasztrofális ábra









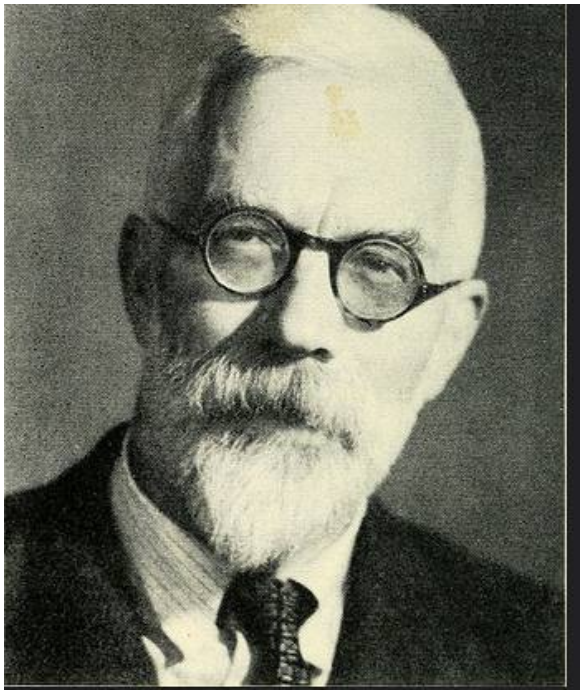
Matematikai statisztika: bevezető

- Statisztikai mező: $(\Omega, \mathcal{A}, P_\vartheta) : \vartheta \in \Theta$
- Paramétertér: Θ. Ez lehet egydimenziós, de akár végtelen dimenziós is
- Minta: $\mathbf{X} = (X_1, \ldots, X_n)$ 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

Karl Pearson (1857 – 1936)

- angol matematikus, statisztikus
- a matematikai statisztika atyja
- hisztogram
- Pearson-féle korreláció és kapcsolata a lineáris regresszióval
- momentum módszer
- hipotézisvizsgálat elméletének lefektetése, *p*-érték
- χ^2-próba
- főkomponens analízis (principal component analysis, PCA)
- "Statistics is the grammar of science."

Ronald Fisher (1890 – 1962)



- angol statisztikus és biológus
- F-eloszlás, Student-féle t-eloszlás
- elégséges statisztika
- Fisher-információ
- a statisztika bayes-i megközelítése

- diszkriminancia analízis
- extrémérték-elmélet (extreme value theory)
- újramintavételezés – Fisher-féle permutációs teszt

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

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

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

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? $g(\vartheta)$	Ha mivel becsüljük? $T_n(\mathbf{X})$	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