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

7. előadás

Bayes becslések

- Példa: egy szabályosnak tűnő érmét 10-szer feldobtunk és egy fej sem jött ki, azaz $\hat{p} = 0$. De azért nem gondoljuk, hogy $p = 0$.
- Tegyük fel, hogy van valami előzetes információnk az ismeretlen paraméterről: az 1/2-hez közeli értékeket valószínűbbnek gondoljuk, mint a 0-hoz közelieket.
- Matematikailag: egy eloszlást tételezünk fel a paramétertéren ($(0;1)$ intervallum)
- Elnevezés: apriori eloszlás (pl. béta eloszlás a vszg. esetén), jelölése: π.
- A minta prediktív (feltétel nélküli) eloszlása:

$$f_\pi(x) = \int_\Theta f_\vartheta(x)\pi(\vartheta)d\vartheta$$

- Ez diszkrét esetben $P(\mathbf{X} = \mathbf{x})$, a π indexet gyakran ki sem írjuk

Aposzteriori eloszlás

- A minta megfigyelése után más lesz az eloszlás a paramétertéren (aposzteriori eloszlás):

$$\pi^*(\vartheta) = \frac{P(\mathbf{X} = \mathbf{x}|\vartheta)\pi(\vartheta)}{P(\mathbf{X} = \mathbf{x})}$$

- Példa: ha m érmedobásnál ℓ fej jön ki, a fej valószínűségének apriori eloszlása Béta(k, n), akkor a fej valószínűségének aposzteriori eloszlása Béta$(k + \ell, n + m - \ell)$
- Ha a mintaelemszám nő, akkor egyre inkább a minta határozza meg az aposzteriori eloszlást
- Ha az apriori eloszlás szórása kicsi (prekoncepció), akkor ezt nehéz kompenzálni – csak nagy minta esetén lehetséges

Bayes-becslések: veszteségfüggvények

- Ha már tudjuk a paraméter eloszlását (π^*), akkor már csak a számítási módot kell meghatározzuk a becsléshez
- Veszteségfüggvény: $R(\hat{\vartheta}, \vartheta)$: az a veszteség, ami akkor ér bennünket, ha a ϑ paramétert $\hat{\vartheta}$-al becslünk
- Tulajdonságok:
  - nemnegatív,
  - tipikusan $\hat{\vartheta} - \vartheta$ függvénye,
  - R is valószínűségi változó.
- Példák:

$$R(\hat{\vartheta}, \vartheta) = (\hat{\vartheta} - \vartheta)^2,$$

$$R(\hat{\vartheta}, \vartheta) = |\hat{\vartheta} - \vartheta|$$

- Optimális döntés speciális veszteségfüggvényekre: a veszteségfüggvény várható értékét minimalizáljuk
  - A négyzetes veszteségfüggvényre $E(\pi^*)$
  - Az abszolút eltérésre $medin(\pi^*)$

Hipotézisvizsgálat

- **Hipotézis**: egy állítás, aminek igazságát vizsgálni szeretnénk. Egy hipotézist vagy elfogadunk, vagy elutasítunk/elvetünk.
- A paraméterteret diszjunkt részekre bontjuk: $\Theta = \Theta_0 \cup^* \Theta_1$
- A hipotézisvizsgálati alapfeladat (absztraktul, a gyakorlatban konkretizálni szoktuk)

  $H_0 : \vartheta \in \Theta_0 \quad \rightarrow$ nullhipotézis

  $\underline{H_1 : \vartheta \in \Theta_1} \quad \rightarrow$ ellenhipotézis vagy alternatív hipotézis
- A nullhipotézis esetén az elfogadás helyett helyesebb azt mondani, hogy nem tudjuk elvetni. Az okokról később.
- A H_0 hipotézisnek azon állítást szokás választani,
  - ami sok éves tapasztalatnak felel meg
  - amit "remélünk", hogy teljesül
  - aminek az elutasítása, gyakran negatív következményekkel jár (büntetés, bírság, riasztás stb.)

Hipotézisvizsgálati alapfogalmak

Hogyan döntsünk? Vajon H_0 igaz, vagy H_1? $\rightarrow$ jó lenne valamilyen matematikai eljárás

- Statisztikai próba vagy röviden **próba**: az a módszer/eljárás, amely során a minta segítségével döntést hozunk a hipotézis(ek)ről.
- **Paraméteres próba**: Olyan próba, amely során a feladatban lévő ismeretlen eloszlás jellege ismert, és a nullhipotézis az eloszlás valamely paraméterére (vagy annak egy minket érdeklő függvényére) vonatkozik.

- Mintatér felbontása két diszjunkt részre: $\mathcal{X} = \mathcal{X}_e \cup^* \mathcal{X}_k$
- $\mathcal{X}_k$: **kritikus tartomány** – azon $\mathbf{x}$ megfigyelések halmaza, amikre *elutasítjuk* a nullhipotézist
- $\mathcal{X}_e$: **elfogadási tartomány** – azon $\mathbf{x}$ megfigyelések halmaza, amikre *elfogadjuk* a nullhipotézist

Hipotézisvizsgálati alapfogalmak II

- Döntési mátrix hipotézisvizsgálat esetén:

Döntés "Valóság"	H_0-t elfogadjuk ($\mathcal{X}_e$)	H_0-t elutasítjuk ($\mathcal{X}_k$)
H_0 teljesül (Θ_0)	helyes döntés	elsőfajú hiba
H_0 nem teljesül (Θ_1)	másodfajú hiba	helyes döntés

- **Elsőfajú hiba** (type I error): a nullhipotézist elvetettük, de nem szabadott volna, mert a H_0-beli állítás igaz
  Valószínűsége: $\alpha(\vartheta) := P_\vartheta(\mathcal{X}_k)$, ahol $\vartheta \in \Theta_0$
  További szokásos jelölések:
  $\alpha(\vartheta) = P_{\vartheta\in\Theta_0}(\mathcal{X}_k) = P_{H_0}(\mathcal{X}_k) = P_0(\mathcal{X}_k)$
- **Másodfajú hiba** (type II error): a nullhipotézist elfogadtuk, de nem szabadott volna, mert a H_0-beli állítás hamis
  Valószínűsége: $\beta(\vartheta) := P_\vartheta(\mathcal{X}_e)$, ahol $\vartheta \in \Theta_1$
  További szokásos jelölések:
  $\beta(\vartheta) = P_{\vartheta\in\Theta_1}(\mathcal{X}_e) = P_{H_1}(\mathcal{X}_e) = P_1(\mathcal{X}_e)$
- **Erőfüggvény**: $\psi(\vartheta) := P_\vartheta(\mathcal{X}_k)$, ahol $\vartheta \in \Theta_1$

Hipotézisvizsgálati alapfogalmak III

- **Terjedelem**: $\alpha := \sup_{\vartheta\in\Theta_0} \alpha(\vartheta)$
  Hosszabban: a próba pontos terjedelmének is hívják
- A hipotézisvizsgálati feladat elején rögzíteni szokás a terjedelmet, tipikusan 5%-on (esetleg más szám 1% és 10% között). Ezáltal döntésünket
  - 5%-os elsőfajú hiba valószínűsége mellett, vagy másképp:
  - 95%-os megbízhatósággal

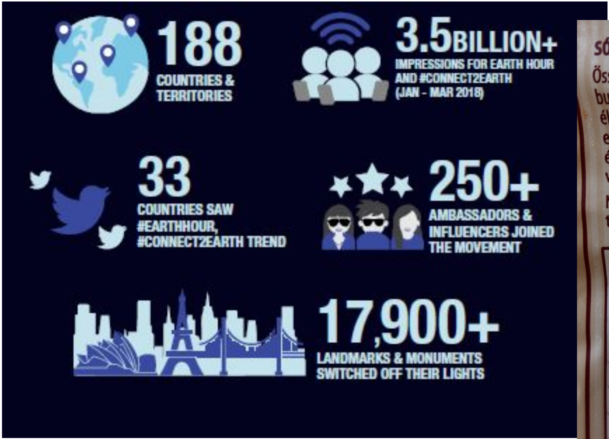
  fogjuk meghozni.

Hipotézisvizsgálati alapfogalmak IV

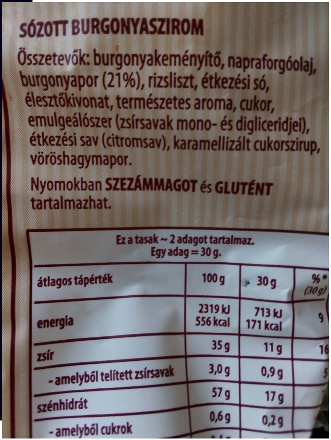
- H_0 **egyszerű**, ha $|\Theta_0| = 1$ (egyelemű)
- H_0 **összetett**, ha $|\Theta_0| > 1$ (legalább kételemű)
  A továbbiakban feltesszük, hogy $\Theta \subset \mathbb{R}$ (valós paraméter esete)
- **Kétoldali próba**: $H_0 : \vartheta = \vartheta_0 \quad H_1 : \vartheta \neq \vartheta_0$
- **Egyoldali próba**: $H_0 : \vartheta = \vartheta_0 \quad H_1 : \vartheta > \vartheta_0$ (vagy $H_1 : \vartheta < \vartheta_0$)
- **Próbastatisztika**: Olyan alkalmas statisztika, amely segítségével a kritikus tartományt meghatározzuk.
  - Ez jellemzően úgy szokott menni, hogy valós értékű $T : \mathcal{X} \to \mathbb{R}$ próbastatisztikát választunk, majd az alábbi alakú kritikus tartományok közül keressük valamelyiket:
    $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : \ T(\mathbf{x}) > c\}$ (egyoldali próbánál)
    $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : \ T(\mathbf{x}) < c\}$ (egyoldali próbánál)
    $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : \ |T(\mathbf{x})| > c\}$ (kétoldali próbánál)
  - c neve: **kritikus érték**, ami jellemzően függ a próba terjedelmétől, ezért c_α-val jelöljük. Ez általában arra utal, hogy c_α a $T(\mathbf{X})$ valószínűségi változó α-kvantilise.
  - A próba meghatározása: előre rögzített α terjedelemhez azt a c_α értéket keressük, amire a próba terjedelme éppen α:
    $\sup P_\vartheta(T(\mathbf{X}) > c_\alpha) = \alpha$

Öt perc szünet: a Föld órája

Egy "érdekes" ábra



Es egy felirat egy 40g-os tasakról



Feladatok

E33.) Minőségellenőrként az a feladatunk, hogy a gyártósorokat szükség esetén leállítsuk, amennyiben túl sok selejtes termék kerül le róluk. A megengedett selejtarány legfeljebb 5%. Összesen 25 terméket vizsgálunk meg. Ha a selejtesek száma legalább k, akkor leállítjuk a gyártósort. Amennyiben kevesebb selejtesünk van k-nál, akkor a termelés mehet tovább.

a.) Írjuk fel a hipotéziseket és a próbát!

b.) Határozzuk meg az elsőfajú hibát!

c.) Mely k érték esetén lesz az elsőfajú hiba valószínűsége legközelebb 5%-hoz?

A továbbiakban tegyük fel, hogy a vizsgált gyártósoron a termékek meghibásodásának valószínűsége 0,1.

d.) Határozzuk meg a másodfajú hibát!

e.) Ábrázoljuk az erőfüggvényt különböző k értékek esetén!

f.) Mely k érték esetén lesz az elsőfajú és a másodfajú hiba valószínűségének összege minimális?

Feladatok

E**34.)** Az (A) gép által termelt termékek egy bizonyos jellemzője $N(11, 1^2)$, míg a (B) gépen termelt termékeké $N(13, 4^2)$ eloszlású.
Legyenek
H_0 : a mintánk az (A) gépen készült
H_1 : a mintánk a (B) gépen készült

Ha egy 16 elemű minta átlaga legfeljebb 12, akkor elfogadjuk H_0-t, különben elvetjük.

- a.) Mekkora az elsőfajú és a másodfajú hiba valószínűsége?
- b.) Milyen c értéket adjunk meg a 12 helyett ahhoz, hogy $0,05$ legyen a próba terjedelme? Ekkor mennyi a másodfajú hiba valószínűsége?
- c.) Milyen c értéket adjunk meg a 12 helyett ahhoz, hogy az elsőfajú és a másodfajú hiba valószínűségének összegét minimalizáljuk?

A hipotézisvizsgálat menete I.

1. A terjedelem (α) lefixálása, ami jellemzően 1% és 10% közötti, tipikusan 5%
   Megbízhatóság$= 1 - \alpha$, általában %-osan írjuk
2. Nullhipotézis (H_0) felírása – sokévi, megszokott, elvárt értékeknek megfelelő paramétertartomány
3. Alternatív hipotézis (H_1) felírása – a feladat alapján bennünket érdeklő kérdésnek megfelelő paramétertartomány
4. A probléma megoldására alkalmas próba vagy próbák kiválasztása – feltételek ellenőrzése
5. Próbastatisztika kiszámítása
6. Kritikus érték kiszámítása, kritikus tartomány ($\mathcal{X}_k$) megállapítása
7. Döntés:
   - $\mathbf{x} \in \mathcal{X}_k \rightsquigarrow$ **erős döntés**, H_1-et elfogadjuk, H_0-t elvetjük/elutasítjuk
   - $\mathbf{x} \in \mathcal{X}_e \rightsquigarrow$ **gyenge döntés**, H_0-t nem tudjuk elutasítani

A hipotézisvizsgálat menete II.

1.) A terjedelem (α) lefixálása
2.) Nullhipotézis (H_0) felírása
3.) Alternatív hipotézis (H_1) felírása
4.) A probléma megoldására alkalmas próba vagy próbák kiválasztása
5.) Számítógéppel dolgozva, az előző dián lévő 6.)-7.) helyett dönthetünk az ún. *p*-érték alapján is:

$$\boxed{p\text{-érték} < \alpha \quad \Leftrightarrow \quad \mathbf{x} \in \mathcal{X}_k \quad \Leftrightarrow \quad H_1\text{-et elfogadjuk}}$$

***p*-érték**: az a terjedelem, amire a kritikus érték megegyezik a próbastatisztikával (másképpen: a legkisebb olyan α, amire az adott minta esetén elvetjük H_0-t)

Ha például a *p*-érték$= 0.06$, akkor 5%-os elsőfajú hiba valószínűség mellett nem tudjuk elvetni H_0-t, de 10%-os elsőfajú hiba valószínűség esetén már elvetjük H_0-t.

Ha a *p*-érték$= 0.16$, akkor a hagyományos, értelmes – 90% fölötti – megbízhatósági szinteken nem tudjuk elvetni H_0-t.

Nevezetes paraméteres próbák – áttekintés

Próbák a normális eloszlás várható értékére vonatkozóan:

- Egymintás *u*-próba, egymintás *t*-próba
- Kétmintás próbák:

	a két minta független		a két minta párosított
σ_1 és σ_2 ismert	kétmintás *u*-próba		egymintás *u*-próba a különbségekre
σ_1 és σ_2 ismeretlen	előzetes *F*-próba		egymintás *t*-próba a különbségekre
	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	
	kétmintás t-próba	Welch-próba	

Próbák normális eloszlás szórásnégyzetére vonatkozóan:

- Egymintás próba: χ^2-próba
- Kétmintás próba: *F*-próba

Összefüggő (párosított) minták: X_i és Y_i ugyanahhoz, az i-edik személyhez, tárgyhoz, objektumhoz tartozó véletlen mennyiség, $i = 1, 2, \ldots$

Nevezetes paraméteres próbák I

Egymintás *u*-próba

$X_1, \ldots, X_n \sim N(m, \sigma^2)$, ahol σ ismert, m ismeretlen paraméter

Kétoldali: $H_0 : m = m_0$
$H_1 : m \neq m_0$

Próbastatisztika: $T(\mathbf{X}) = u := \sqrt{n}\frac{\overline{X} - m_0}{\sigma} \overset{H_0 \text{ esetén}}{\sim} N(0,1)$
Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : |u| > u_{1-\alpha/2}\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.
Egyoldaliak: $H_1 : m > m_0$ $H_1 : m < m_0$
Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : u > u_{1-\alpha}\}$
$\mathcal{X}_k = \{\mathbf{x} : u < -u_{1-\alpha}\}$

Áttekintés

Nevezetes paraméteres próbák II

Egymintás *t*-próba

$X_1, \dots, X_n \sim N(m, \sigma^2)$, ahol m és σ ismeretlen paraméterek

Kétoldali: $H_0 : m = m_0$
$\underline{H_1 : m \neq m_0}$

Próbastatisztika: $T(\mathbf{X}) = t := \sqrt{n}\frac{\overline{X}-m_0}{S_n^*} \overset{H_0 \text{ esetén}}{\sim} t_{n-1}$

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : |t| > t_{n-1,1-\alpha/2}\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.

Egyoldaliak: $H_1 : m > m_0$ $H_1 : m < m_0$

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : t > t_{n-1,1-\alpha}\}$ $\mathcal{X}_k = \{\mathbf{x} : t < -t_{n-1,1-\alpha}\}$

[Áttekintés]

Feladatok

E**35.)** A fogyasztóvédelmi hatóság többszöri lakossági bejelentést kapott, hogy a Portokall nevű, fél literes kiszerelésű narancsitalokban a flakonra írt 500 ml-nél jóval kevesebb üdítő van. Ez alapján vizsgálatot kezdtek, a fogyasztóvédelem munkatársa vásárolt a boltban 10 darabot, majd megnézte a benne lévő ital térfogatát (ml):
483, 502, 498, 496, 502, 483, 494, 491,505, 486.
Tegyük fel, hogy egy fél literes üdítős üvegbe töltött narancslé mennyisége normális eloszlást követ.
Állíthatjuk-e 95%-os megbízhatóság esetén, hogy a Portokall gyártója át akarja verni a vevőket?

E**36.)** Használjuk az első előadáson kitöltött kérdőív eredményeit, és próbáljunk meg az alapján a TTK hallgatóira vonatkozólag következtetéseket levonni!

a.) Állíthatjuk-e, hogy a TTK-n a fiúk várható értékben legalább 10 cm-rel magasabbak a lányoknál;

b.) Elfogadhatjuk-e, hogy a lányok magasságának szórása k cm, ahol $k = 1, 2, \ldots, 15$?