Matematika Statisztika

8-9. előadás

Arató Miklós

2022.10.24.

Hipotézisvizsgálati alapfogalmak

- **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)

  $H_0 : \vartheta \in \Theta_0$ $\rightsquigarrow$ nullhipotézis

  $\underline{H_1 : \vartheta \in \Theta_1}$ $\rightsquigarrow$ 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
  - amit elutasítva, gyakran negatív következményekkel jár (büntetés, bírság, jobb modell keresésének kényszere stb.)

Hipotézisvizsgálati alapfogalmak

Hogyan döntsünk? Vajon H_0 igaz, vagy H_1? $\rightsquigarrow$ 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 **x** megfigyelések halmaza, amikre *elutasítjuk* a nullhipotézist
- $\mathcal{X}_e$: **elfogadási tartomány** – azon **x** megfigyelések halmaza, amikre *elfogadjuk* a nullhipotézist

Ez oké, de mi alapján rakjunk egy **x** megfigyelt mintát $\mathcal{X}_k$-ba vagy $\mathcal{X}_e$-be? $\rightsquigarrow$ 2 fóliával később

Hipotézisvizsgálati alapfogalmak

- 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 $\vartheta \in \Theta_1$

- **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

- Legyen $H_0 : \vartheta \in \Theta$
  - H_0 **egyszerű**, ha $|\Theta_0| = 1$ (egyelemű)
  - H_0 **összetett**, ha $|\Theta_0| > 1$ (legalább kételemű)
- **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:
    $\sup_{\vartheta \in \Theta_0} P_\vartheta(T(\mathbf{X}) > c_\alpha) = \alpha.$

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 minta 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_1-et elutasítjuk, 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 fólián lévő 5.)-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

Ha például *p*-érték$=0.06$, akkor 5%-os elsőfajú hiba valószínűsége mellett nem tudjuk elvetni H_0-t, de 10%-os elsőfajú hiba valószínűsége esetén már elvetjük H_0-t.
Ha például *p*-érték$=0.16$, akkor a hagyományos, értelmes – 90% és 99% közötti – 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:

<table>
<tr><td></td><td colspan="2">a két minta
független</td><td>a két minta
nem független</td></tr>
<tr><td>σ_1 és σ_2
ismert</td><td colspan="2">kétmintás u-próba</td><td>egymintás u-próba
a különbségekre</td></tr>
<tr><td rowspan="3">σ_1 és σ_2
ismeretlen</td><td colspan="2">előzetes F-próba</td><td rowspan="3">egymintás t-próba
a különbségekre</td></tr>
<tr><td>$\sigma_1 = \sigma_2$</td><td>$\sigma_1 \neq \sigma_2$</td></tr>
<tr><td>kétmintás t-próba</td><td>Welch-próba</td></tr>
</table>

Próbák normális eloszlás szórásnégyzetére vonatkozóan:

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

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

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: $\mathcal{X}_k = \{\mathbf{x} : |u| > u_{\alpha/2}\}$

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:	$\mathcal{X}_k = \{\mathbf{x} : u > u_\alpha\}$	$\mathcal{X}_k = \{\mathbf{x} : u < -u_\alpha\}$

Áttekintés

Egymintás t-próba

$X_1, \ldots, X_n \sim N(m, \sigma^2)$, 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: $\mathcal{X}_k = \{\mathbf{x} : |t| > t_{n-1,\alpha/2}\}$

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

Áttekintés

Kétmintás *u*-próba

$X_1, \ldots, X_n \sim N(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_m \sim N(m_2, \sigma_2^2)$ független minták
ahol m_1, m_2 ismeretlen paraméterek, σ_1, σ_2 ismert

Kétoldali: $H_0 : m_1 = m_2$
$\underline{H_1 : m_1 \neq m_2}$

Próbastatisztika: $T(\mathbf{X}, \mathbf{Y}) = u := \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} N(0, 1)$

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : |u| > u_{\alpha/2}\}$

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

Egyoldaliak:	$H_1 : m_1 > m_2$	$H_1 : m_1 < m_2$
Kritikus tartomány:	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : u > u_\alpha\}$	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : u < -u_\alpha\}$

Áttekintés

Kétmintás *t*-próba

$X_1, \ldots, X_n \sim N(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_m \sim N(m_2, \sigma_2^2)$ független minták
ahol m_1, m_2, $\sigma_1 = \sigma_2$ ismeretlen paraméterek

Kétoldali: $H_0 : m_1 = m_2$
$\underline{H_1 : m_1 \neq m_2}$

Próbastatisztika: $T(\mathbf{X}, \mathbf{Y}) = t := \sqrt{\frac{nm}{n+m}} \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)(S_1^*)^2 + (m-1)(S_2^*)^2}{n+m-2}}} \overset{H_0 \text{ esetén}}{\sim} t_{n+m-2}$

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : |t| > t_{n+m-2,\alpha/2}\}$

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

Egyoldaliak:	$H_1 : m_1 > m_2$	$H_1 : m_1 < m_2$
Krit. tart.:	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t > t_{n+m-2,\alpha}\}$	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t < -t_{n+m-2,\alpha}\}$

Áttekintés

Welch-próba

$X_1, \ldots, X_n \sim N(m_1, \sigma_1^2)$ és $Y_1, \ldots, Y_m \sim N(m_2, \sigma_2^2)$ független minták ahol m_1, m_2, $\sigma_1 \neq \sigma_2$ ismeretlen paraméterek

Kétoldali: $H_0 : m_1 = m_2$
$\underline{H_1 : m_1 \neq m_2}$

Próbastatisztika: $T(\mathbf{X}, \mathbf{Y}) = t' := \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{(S_1^*)^2}{n} + \frac{(S_2^*)^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} t_f$, ahol

$\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$, $\quad c = \dfrac{\frac{(s_1^*)^2}{n}}{\frac{(s_1^*)^2}{n} + \frac{(s_2^*)^2}{m}}$, ha $s_1^* > s_2^*$ (így csináljuk)

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : |t| > t_{f,\alpha/2}\}$

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

Egyoldaliak:	$H_1 : m_1 > m_2$	$H_1 : m_1 < m_2$
Krit. tartomány:	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t > t_{f,\alpha}\}$	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t < -t_{f,\alpha}\}$

Áttekintés

χ^2**-próba** (normális eloszlás szórására)

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

Kétoldali: $H_0 : \sigma = \sigma_0$
$\underline{H_1 : \sigma \neq \sigma_0}$

Próbastatisztika: $T(\mathbf{X}) = h := \frac{(n-1)(S_n^*)^2}{\sigma_0^2} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{n-1}$

Kritikus tartomány: $\mathcal{X}_k = \left\{ \mathbf{x} : h < \chi^2_{n-1,\alpha/2} \text{ vagy } h > \chi^2_{n-1,1-\alpha/2} \right\}$

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

Egyoldaliak:	$H_1 : \sigma > \sigma_0$	$H_1 : \sigma < \sigma_0$
Kritikus tartomány:	$\mathcal{X}_k = \{\mathbf{x} : h > \chi^2_{n-1,1-\alpha}\}$	$\mathcal{X}_k = \{\mathbf{x} : h < \chi^2_{n-1,\alpha}\}$

Áttekintés

F-próba

$X_1, \dots, X_n \sim N(m_1, \sigma_1^2)$
$Y_1, \dots, Y_m \sim N(m_2, \sigma_2^2)$ független minták
ahol m_1, m_2, σ_1, σ_2 ismeretlen paraméterek

Kétoldali: $H_0 : \sigma_1 = \sigma_2$
$\underline{H_1 : \sigma_1 \neq \sigma_2}$

Próbastatisztika: $T(\mathbf{X}, \mathbf{Y}) = F = \frac{(S_1^*)^2}{(S_2^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{n-1,m-1}$

Kritikus tartomány:
$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : F < F_{n-1,m-1,\alpha/2}$ vagy $F > F_{n-1,m-1,1-\alpha/2}\}$

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

Egyoldaliak:	$H_1 : \sigma_1 > \sigma_2$	$H_1 : \sigma_1 < \sigma_2$
Krit. tart.:	$\{(\mathbf{x}, \mathbf{y}) : F > F_{n-1,m-1,1-\alpha}\}$	$\{(\mathbf{x}, \mathbf{y}) : F < F_{n-1,m-1,\alpha}\}$

Áttekintés

A próbafüggvény

Próbafüggvény: $\varphi \colon \mathcal{X} \to [0;1] \rightsquigarrow$ ennyi valószínűséggel vetem el a minta alapján a nullhipotézist

- $\varphi(\mathbf{x}) := I(\mathbf{x} \in \mathcal{X}_k) = \begin{cases} 1 & \text{ha } \mathbf{x} \in \mathcal{X}_k \\ 0 & \text{ha } \mathbf{x} \in \mathcal{X}_e \end{cases}$
- a próbafüggvény is egy statisztika
- a próbafüggvény egyértelműen meghatározza a próbát, ezért gyakran a próbát magával a φ függvénnyel azonosítják
- tipikusan $\quad \varphi(\mathbf{x}) = \begin{cases} 1 & \text{ha } T(\mathbf{x}) \geq c_\alpha \\ 0 & \text{ha } T(\mathbf{x}) < c_\alpha \end{cases}$ alakú,
  ahol T egy alkalmas statisztika,
  c_α pedig a kritikus érték, amit úgy határozunk meg, hogy
  $P_{\vartheta \in \Theta_0}(\mathcal{X}_k) = E_{\vartheta \in \Theta_0}(\varphi(\mathbf{X})) = P_{\vartheta \in \Theta_0}(T(\mathbf{X}) \geq c_\alpha) = \alpha \quad$ teljesüljön
- diszkrét eloszlású minták esetén rendszerint nem lehet úgy meghatározni c_α-t, hogy a terjedelem pontosan α legyen, ezért a próbafüggvény fogalmának általánosítására, úgynevezett **véletlenítés**re (randomizálásra) van szükség ilyen esetek miatt.

A próbafüggvény

Véletlenített próbafüggvény: $\varphi\colon \mathcal{X} \to [0;1]$

- $$\varphi(\mathbf{x}) := I(\mathbf{x} \in \mathcal{X}_k) = \begin{cases} 1 & \text{ha } \mathbf{x} \in \mathcal{X}_k \\ p & \text{ha } \mathbf{x} \in \mathcal{X}_r \\ 0 & \text{ha } \mathbf{x} \in \mathcal{X}_e \end{cases}, \text{ ahol}$$

  $\mathcal{X}_r$ neve: véletlenítési vagy "randomizálási" tartomány; $p \in [0;1]$
- Tipikusan $$\varphi(\mathbf{x}) = \begin{cases} 1 & \text{ha } T(\mathbf{x}) > c_\alpha \\ p_\alpha & \text{ha } T(\mathbf{x}) = c_\alpha \\ 0 & \text{ha } T(\mathbf{x}) < c_\alpha \end{cases} \text{ alakú,}$$

  ahol T egy alkalmas statisztika,
  c_α a kritikus érték és $p_\alpha \in [0;1]$, amiket úgy határozunk meg, hogy
  $E_{\vartheta\in\Theta_0}(\varphi(\mathbf{X})) = P_{\vartheta\in\Theta_0}(T(\mathbf{X}) > c_\alpha) + p_\alpha \cdot P_{\vartheta\in\Theta_0}(T(\mathbf{X}) = c_\alpha) = \alpha$
  teljesüljön
  Az ilyen próbafüggvénnyel végrehajtott próbát **véletlenített próbá**nak hívjuk.

Próbák tulajdonságai

- **Torzítatlan próba (legfeljebb α terjedelemmel)**:
  $P_\vartheta(\mathbf{X} \in \mathcal{X}_k) \leq \alpha \quad \forall \vartheta \in \Theta_0$-ra és
  $P_\vartheta(\mathbf{X} \in \mathcal{X}_k) \geq \alpha \quad \forall \vartheta \in \Theta_1$-re
  Megfontolás a definíció mögött: ha nem teljesül H_0, akkor a minta alapján az elvetés valószínűsége legalább annyi legyen, mintha igaz lenne H_0.
- **Konzisztens próba (α terjedelemmel)**: olyan próba, aminek α a terjedelme és a mintaméret növelésével az erőfüggvény 1-hez konvergál. Formálisan felírva:
  $P_{\vartheta \in \Theta_0}(\mathbf{X} \in \mathcal{X}_k) = \alpha$ és
  $\psi_n(\vartheta) \xrightarrow{n \to \infty} 1 \quad \forall \vartheta \in \Theta_1$-re, ahol
  ψ_n az n elemű mintához tartozó erőfüggvény
  Megj.: ha az erőfüggvény 1-hez konvergál, akkor ebből következik, hogy a másodfajú hiba valószínűsége 0-hoz tart.

Próbák tulajdonságai

- Legyenek φ_1 és φ_2 torzítatlan próbák. A φ_1 **próba erősebb** φ_2 **próbánál**, ha φ_1 próba erőfüggvénye $\forall\vartheta \in \Theta_1$ esetén nagyobb vagy egyenlő, mint φ_2 próba erőfüggvénye
  Megj.: nem biztos, hogy két próba közül az egyik erősebb a másiknál
- **Egyenletesen legerősebb próba**: az adott hipotézisvizsgálati feladat esetén minden más torzítatlan próbánál erősebb
  Megj.: nem biztos, hogy létezik egyenletesen legerősebb próba az adott feladatra
- Mikor létezik egyenletesen legerősebb próba? Ha létezik, akkor hogyan találjuk meg?

Legerősebb próba keresése

Ha mind H_0, mind H_1 egyszerű, akkor adott α terjedelemhez lehet legerősebb próbát találni, ezt pedig úgy hívják, hogy **valószínűség-hányados próba.**
A hipotézisek folytonos esetre (diszkrétre a sűrűségfüggvény helyett a konkrét eloszlást kell írni):

$H_0: \; f = f_0$
$\underline{H_1: \; f = f_1}$

A próba kritikus tartománya: $\mathcal{X}_k = \left\{ \mathbf{x} : \overbrace{\dfrac{f_1(\mathbf{x})}{f_0(\mathbf{x})}}^{T(\mathbf{x})} > c_\alpha \right\}$

Tehát azokat az **x**-eket, amikre a $T(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$ statisztika nagy, bepakoljuk a kritikus tartományba egészen addig, míg az adott α terjedelmet el nem érjük. Diszkrét esetben ehhez általában véletlenítésre van szükség, azaz bizonyos **x**-ek esetén nem 1 vagy 0, hanem egy, e két szám közé eső (jelöljük p_α-val) valószínűséggel vetjük el a nullhipotézist.

A valószínűség-hányados próba elméleti hátterét a *Neyman-Pearson* (alap)*lemma* biztosítja.

Motiváció – χ^2-próbával vizsgálható problémák

- Szabályos-e egy érme/kocka?
- Normális eloszlást követ-e a magyar nők testmagassága?
- Lehet-e exponenciális eloszlású az az idő, amit a Blaha Lujza téri megállóban a következő villamosra várakozással töltünk?
- Állíthatjuk-e, hogy a nők és a férfiak vérnyomása ugyanolyan eloszlást követ?
- Független-e a diákok matematika és irodalom érdemjegye? Aki jó matekból, általában jó magyarból is?
- Független-e egymástól az emberek szemszíne és hajszíne? Igaz-e, hogy a szőke hajúak főleg kék szeműek?

A χ^2-próba

Legyen $A_1, \ldots, A_r$ teljes eseményrendszer.
Végezzünk n darab független megfigyelést, jelölje az i-edik esemény bekövetkezési gyakoriságát N_i ($i = 1, \ldots, r$). A megfigyelések egyes eredményei segítségével definiálható az X_i valószínűségi változó, ami vegyen fel olyan értéket, amelyik számú esemény a teljes eseményrendszerből bekövetkezett. Ezáltal formálisan

$N_i = \sum_{j=1}^{n} I(X_j = i)$ és $\sum_{i=1}^{r} N_i = n$

H_0: $P(A_i) = p_i$, $i = 1, \ldots, r$ $\quad \rightsquigarrow$ tfh. $p_i > 0\ \forall i,\ p_1 + \ldots + p_r = 1$
H_1: a nullhipotézis tagadása

Próbastatisztika: $\quad T_n(\mathbf{X}) := \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \xrightarrow[n\to\infty]{H_0 \text{ esetén}} \chi^2_{r-1}$ eloszlásban
Kritikus tartomány: $\mathcal{X}_k = \left\{\mathbf{x} : T_n(\mathbf{X}) > \chi^2_{r-1;1-\alpha}\right\}$

A χ^2-próba

Alkalmazásai:

- illeszkedésvizsgálat: egy minta adott eloszlást követ-e
- homogenitásvizsgálat: két minta eloszlása megegyezik-e
- függetlenségvizsgálat: két szempont, ismérv, tulajdonság független-e egymástól

Megjegyzések:

- a χ^2-próba **aszimptotikus** próba, ami azt jelenti, hogy "nagy" mintaelemszámra lehet végrehajtani. "Kicsi" minták esetén a kritikus érték nem használható, azt szimulálni kell a konkrét minta alapján.
- Mikor elég "nagy" már egy minta – hüvelykujjszabály: ha legalább 100 elemű. Egyébként eloszlásfüggő, legalább mekkora n-re van szükség, hogy kritikus értéknek a χ^2-eloszlás kvantiliseit lehessen használni.
- Végrehajtásának további feltétele, hogy minden osztályban "elegendő" mennyiségű gyakoriság legyen.
- A próbastatisztikában lévő összeg tagjai $\frac{(O-E)^2}{E}$ alakúak, ahol E: elméleti gyakoriságok, O: tapasztalati gyakoriságok

Illeszkedésvizsgálat

H_0: a minta egy adott eloszlásból származik
H_1: a minta nem ilyen eloszlású

Végrehajtása:

- grafikusan módszerek ("szemmel" jónak tűnik-e az illeszkedés):
  - Q-Q plot
  - P-P plot
  - hisztogram/magfüggvényes sűrűségfüggvény-becslés, valamint az illesztett sűrűségfüggvény egy ábrában
- statisztikai próbák:
  - diszkrét eloszlás esetén χ^2-próba
  - folytonos eloszlás esetén több statisztikai próba közül lehet választani
    - diszkretizálás (mesterséges osztályok létrehozása) révén χ^2-próba
    - Kolmogorov-Szmirnov próba
    - Cramér-von Mises próba
    - Anderson-Darling próba
    - Shapiro-Wilk próba: kizárólag normalitásvizsgálatra, amire ez a legjobb

Illeszkedésvizsgálat grafikusan

Q-Q plot (kvantilis-kvantilis ábra)

- Az illesztett eloszlás kvantiliseit vetjük össze a tapasztalati kvantilisekkel, azaz a következő pontokat ábrázoljuk:
  $\left(F^{-1}\left(\frac{k}{n+1}\right), x_k^*\right) \qquad k = 1, \ldots, n$
  ahol
  - F: az illesztett eloszlás eloszlásfüggvénye
  - x_k^* a k. rendezett mintaelem
- Be szokták húzni a 45 fokos egyenest és minél jobban rásimulnak a pontok az egyenesre, annál jobbnak tekinthető az illeszkedés.
- Felnagyítja az eloszlás szélein az eltéréseket, ezért szinte mindig előnyben részesítik a P-P plot-tal szemben.

qqplot1.pdf

qqplot2.pdf

Illeszkedésvizsgálat grafikusan

P-P plot (percentilis-percentilis ábra)

- Az illesztett eloszlás egyes valószínűségeit vetjük össze a tapasztalati valószínűségekkel, azaz a következő pontokat ábrázoljuk:
  $\left(\frac{k}{n+1}, F\left(x_k^*\right)\right) \qquad k = 1, \ldots, n \qquad$ ahol
  - F: az illesztett eloszlás eloszlásfüggvénye
  - x_k^* a k. rendezett mintaelem
- Be szokták húzni a 45 fokos egyenest és minél jobban rásimulnak a pontok az egyenesre, annál jobbnak tekinthető az illeszkedés.
- Felnagyítja az eloszlás közepén az eltéréseket

A Q-Q plot és P-P plot nem helyettesíti a formális tesztelést, inkább kiegészíti azt!

Illeszkedésvizsgálat χ^2-próbával

Osztályok	1	2	...	r	Összesen
Valószínűségek	p_1	p_2	...	p_r	1
Gyakoriságok	N_1	N_2	...	N_r	n

H_0 : a valószínűségek: $\mathbf{p}=(p_1,\ldots,p_r)$
H_1: nem ezek a valószínűségek

Próbastatisztika: $T_n(\mathbf{X}) = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \xrightarrow{H_0 \text{ esetén}} \chi^2_{r-1}$ elo.-ban, ha $n \to \infty$

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Becsléses illeszkedésvizsgálat: csak annyit "sejtünk", hogy a minta valamilyen eloszlású, viszont a paramétereiről nincs sejtésünk. Ilyenkor amennyiben ML-módszerrel becsüljük meg az s darab ismeretlen paramétert, akkor a próbastatisztika: $T_n(\mathbf{X}) \xrightarrow{H_0 \text{ esetén}} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

A χ^2-próba végrehajtásának feltételei (hüvelykujjszabály): $N_i \geq 4$ és $np_i \geq 4$ minden i-re. Ha ezek nem teljesülnek, akkor osztályokat kell összevonni.

Illeszkedésvizsgálat Kolmogorov-Szmirnov próbával

$H_0 : F_{X_1}(x) = F(x) \ \ \forall x \in \mathbb{R}$ ahol F egy adott eloszlás elofv.-e
H_1: a nullhipotézis tagadása

Próbastatisztika: $D_n(\mathbf{X}) = \sup\limits_{x \in \mathbb{R}} |F_n(x) - F(x)|$

A próbastatisztika $\sqrt{n}$-szeresének eloszlása H_0 esetén az ún. Kolmogorov-eloszláshoz tart ($n \to \infty$). Jelöljük K_α-val a Kolmogorov-eloszlás α-kvantilisét.

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : \sqrt{n} D_n(\mathbf{x}) > K_{1-\alpha}\}$

Megjegyzések:

- D_n kiszámításához elég csak a mintapontokban tekinteni az eltérést.
- Nem lehet használni a határeloszlást, ha paramétereket kell becsülnünk! Ilyen esetben a kritikus értéket szimulációval kaphatjuk meg.
- A Kolmogorov-eloszlás eloszlásfüggvénye: $1 - 2\sum\limits_{k=1}^{\infty}(-1)^{k-1} e^{-2k^2x^2}$

Homogenitásvizsgálat

Adott két független minta, mindkettő egy közös szempont szerint r osztály egyikébe sorolva.

	Osztályok	1	2	...	r	Összesen
1. minta	Valószínűségek	p_1	p_2	...	p_r	1
	Gyakoriságok	N_1	N_2	...	N_r	n
2. minta	Valószínűségek	q_1	q_2	...	q_r	1
	Gyakoriságok	M_1	M_2	...	M_r	m

H_0: a két minta azonos eloszlású, azaz $(p_1, \ldots, p_r) = (q_1, \ldots, q_r)$
H_1: a nullhipotézis tagadása

Próbastatisztika: $T_{n,m}(\mathbf{X},\mathbf{Y}) = nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \xrightarrow[n\to\infty]{H_0 \text{ esetén}} \chi^2_{r-1}$ eloszlásban

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{X},\mathbf{Y}) : T_{n,m}(\mathbf{X},\mathbf{Y}) > \chi^2_{r-1,1-\alpha}\}$

Függetlenségvizsgálat

Feladat: van egy minta, két ismérv szerint csoportosítva. Azt kell eldönteni, hogy a két szempont független-e egymástól.

$p_{i,j} = P$(egy megfigyelés az (i,j) osztályba kerül)
$N_{i,j}$ =ennyi megfigyelés kerül az (i,j) osztályba

		2. szempont					
		1	...	j	...	s	Összesen
	1	N_{11}	...	N_{1j}	...	N_{1s}	$N_{1\bullet}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
1. szempont	i	N_{i1}	...	N_{ij}	...	N_{is}	$N_{i\bullet}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	r	N_{r1}	...	N_{rj}	...	N_{rs}	$N_{r\bullet}$
Összesen		$N_{\bullet 1}$	...	$N_{\bullet j}$	...	$N_{\bullet s}$	n

ahol $N_{i\bullet} = \sum_{j=1}^{s} N_{ij}$ és $N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$

Függetlenségvizsgálat

Itt formálisan a mintánk két dimenziós: a megfigyelések az $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ párok, ahol az X-ek r, az Y-ok pedig s különböző értéket vehetnek fel nemnulla valószínűséggel: $p_{i,j} = P(X_1 = x_i, Y_1 = y_j)$, ahol $i = 1, \ldots, r$ és $j = 1, \ldots, s$.

Továbbá $N_{i,j} = \sum_{k=1}^{r} \sum_{l=1}^{s} I(X_k = x_i, Y_l = y_j)$.

H_0 : az ismérvek függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j} \quad \forall i, j$-re
H_1: az ismérvek nem függetlenek

Próbast.: $T_n(\mathbf{X}, \mathbf{Y}) = n \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{N_{i,j}^2}{N_{i\bullet} N_{\bullet j}} - 1 \right) \xrightarrow[n\to\infty]{H_0 \text{ esetén}} \chi^2_{(r-1)(s-1)}$ elo.-ban

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{X}, \mathbf{Y}) : T_n(\mathbf{X}, \mathbf{Y}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$

Ha $r = s = 2$, akkor a próbastatisztika $T_n = n \cdot \frac{(N_{11} N_{22} - N_{12} N_{21})^2}{N_{1\bullet} N_{2\bullet} N_{\bullet 1} N_{\bullet 2}}$-re egyszerűsödik, az aszimptotikus eloszlás pedig 1 szabadságfokú χ^2.