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

8. előadás

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

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**33.)** 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**34.)** 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 $k = 1, 2, \ldots, 15$?

Nevezetes paraméteres próbák III

Kétmintás *u*-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 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: $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : |u| > u_{1-\alpha/2}\}$

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

Egyoldaliak:	$H_1 : m_1 > m_2$	$H_1 : m_1 < m_2$
Kritikus tartomány:	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : u > u_{1-\alpha}\}$	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : u < -u_{1-\alpha}\}$

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

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

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

Egyoldaliak:	$H_1 : m_1 > m_2$	$H_1 : m_1 < m_2$
Krit. tart.	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t > t_{n+m-2,1-\alpha}\}$	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t < -t_{n+m-2,1-\alpha}\}$

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

Öt perc szünet

- 2011-ben az USA-ban 11000 új állás jött létre (azaz az év végén ennyivel többen dolgoztak, mint az év elején)
- Wisconsin állam kormányzója büszkén jelentette be, hogy az ő állama egyedül az ország állásainak 50%-át hota létre, mert náluk 5500-zal dolgoztak többen.
- Mi a véleményünk?
- A szomszédos Minnesota viszont 13000 új állást hozott létre (>100%)
- Azaz nem szabad százalékot használni ott, ahol negatív értékek is vannak!
- Forrás: Ellenberg: How not to be wrong?

Öt perc szünet

- 2011-ben az USA-ban 11000 új állás jött létre (azaz az év végén ennyivel többen dolgoztak, mint az év elején)
- Wisconsin állam kormányzója büszkén jelentette be, hogy az ő állama egyedül az ország állásainak 50%-át hota létre, mert náluk 5500-zal dolgoztak többen.
- Mi a véleményünk?
- A szomszédos Minnesota viszont 13000 új állást hozott létre (>100%)
- Azaz nem szabad százalékot használni ott, ahol negatív értékek is vannak!
- Forrás: Ellenberg: How not to be wrong?

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

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

Kétoldali: $H_0 : \sigma = \sigma_0$
$H_1 : \sigma \neq \sigma_0$

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

Nevezetes paraméteres próbák VII

F-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, σ_1, σ_2 ismeretlen paraméterek

Kétoldali: $H_0 : \sigma_1 = \sigma_2$
$H_1 : \sigma_1 \neq \sigma_2$

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

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

Feladatok

E**35.)** Használjuk az első előadáson kitöltött kérdőív eredményeit, és próbáljunk meg az alapján a TTK hallgatóira vonatkozólag következtetéseket levonni! Állíthatjuk-e, hogy a TTK-n

- a.) a fiúk legalább 10 cm-rel magasabbak a lányoknál;
- b.) a lányok magasságának szórása k cm, ahol $k = 1, 2, \ldots, 15$?

E**36.)** Bálint gazda este összefut a helyi kocsmában a szomszéd gazdálkodóval, Máté gazdával, aki elmeséli, a tehenei tejének tejzsírszázaléka jelentősen megnőtt, mióta szilázzsal is eteti őket minden nap. Ezen felbuzdulva, Bálint gazda úgy dönt, hogy 6 kedvenc tehenén kipróbálja ezt a "diétát" – egy hónapon keresztül szilázzsal is etette őket, majd megnézte a tejük tejzsírszázalékát:

Mit ettek	Julcsa	Bogár	Riska	Csendes	Bimbó	Mula
Csak füvet	3,84	3,79	3,78	4,00	3,83	3,84
Szilázst is	3,90	4,05	3,8	4,01	3,81	3,9

Vizsgáljuk meg alkalmas statisztikai próbával, hogy a szilázs növeli-e a tej tejzsírszázalékát!

Eredmény, magyarázatok

Az előző feladat végeredménye (számítógéppel megoldva):
p-érték$= 0,0867$ (egymintás t-próba a különbségekre)
Helyes, a szövegkörnyezetbe ágyazott szöveges értelmezések:

- 95%-os megbízhatóság mellett nem tudjuk bizonyítani, hogy a szilázzsal való etetés növeli a tehenek tejének tejzsírszázalékát. ($\rightsquigarrow$ H_1-et elvetjük)
- 5%-os elsőfajú hiba valószínűsége mellett nem tudjuk elutasítani, hogy a szilázzsal való etetés nem változtat a tehenek tejének tejzsírszázalékán. ($\rightsquigarrow$ H_0-t nem tudjuk elvetni)
- 90%-os megbízhatóság esetén azt mondhatjuk, hogy a szilázzsal való etetés növeli a tehenek tejének tejzsírszázalékát.

Nem teljesen korrekt értelmezés:

- 5%-os elsőfajú hiba valószínűsége mellett azt mondhatjuk, hogy a szilázzsal való etetés nem változtat a tehenek tejének tejzsírszázalékán. ($\rightsquigarrow$ H_0-t elfogadjuk)

Eredmény, magyarázatok/2, még egy feladat

- $8,67\%$ az esélye, hogy ilyen (vagy még nagyobb) eltérést kapunk akkor, ha a szilázzsal való etetés valójában nem változtat a tehenek tejének tejzsírszázalékán. ($\rightsquigarrow$ H_0-t elfogadjuk)

Helytelen, hibás szöveges értelmezések:

- 95%-os valószínűséggel elvethetjük, hogy a szilázzsal való etetés növeli a tehenek tejének tejzsírszázalékát.
- 5%-os valószínűséggel nem tudjuk elutasítani, hogy a szilázzsal való etetés nem változtat a tehenek tejének tejzsírszázalékán.
- $8,67\%$ az esélye, hogy a szilázzsal való etetés nem növeli a tehenek tejének tejzsírszázalékát.

Mi történik, ha az *u*-próbák, illetve *t*-próbák feltételei közül nem teljesül az, hogy a minta normális eloszlású, mi mégis hagyományos módon végrehajtjuk a próbát?

E**37.)** Vizsgáljuk meg szimulációval a kétoldali, egymintás *t*-próba terjedelmét, amennyiben a minta $n = 10, 20, 50, 100, 200$ elemű, független normális, illetve exponenciális eloszlású!