



EÖTVÖS LORÁND TUDOMÁNYEGYETEM

SZAKDOLGOZAT

A t -próba

Kisházi Réka

Matematika BSc

Matematikai elemző szakirány

Témavezető:

Dr. Csiszár Villő

2023. június

NYILATKOZAT

Név: Kisházi Réka

ELTE Természettudományi Kar, szak: Matematika BSc

NEPTUN azonosító: CZS6DZ

Szakedolgozat címe: A t-próba

A **szakedolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2023. 06. 05.

Kisházi Réka

a hallgató aláírása

Köszönetnyilvánítás

Szeretném megköszönni Csiszár Villőnek a rendszeres konzultációkat, a támogatást és az odafigyelést, amelyek nélkül ez a szakdolgozat nem jöhetett volna létre.

Tartalomjegyzék

1. Bevezető	4
2. A t-próba eredete, tulajdonságai, példa	5
2.1. A t-próba eredete	5
2.2. A t-próba	7
2.3. Példa	8
2.4. A Welch-próba	10
3. A minta normalitása	12
3.1. Ferdeség és csúcsosság fogalma	12
3.2. Mi befolyásolja a t-próba robusztusságát?	14
3.3. Lambda-eloszlás	15
4. Az egymintás t-próba robusztus változatai	18
4.1. Gayen- és Johnson-próba	18
4.2. Saját szimuláció	21
4.3. Szimulációk és regresszió	24
4.4. A trimmelt t-próba	26
4.5. Példa	27
5. Többváltozós t-próba	30
5.1. Valószínűséghányados próba	30
5.2. A t-próbastatisztika levezetése valószínűséghányados próbával	31
5.3. Többváltozós t-próba levezetése valószínűséghányados próbával	34
6. Összegzés	36

1. fejezet

Bevezető

Számos statisztikai próbának az az előfeltétele, hogy a minta, amire a próbát végezzük, normális eloszlású legyen. Ám a gyakorlatban ez gyakran nem biztosított, ezért felmerülhet a kérdés, hogy az egyes próbák hogyan viselkednek normálistól eltérő minta esetén. Dolgozatomban a t-próbát járom körül ilyen szempontból, illetve a t-próba néhány robusztus változatát.

Ismertetem a Gayen- illetve Johnson-féle t-próbát és az ezekre végzett szimulációs vizsgálatokat, illetve egy kisebb, saját szimulációt is. Később a trimmelt és a többváltozós t-próbát is bemutatom.

2. fejezet

A t-próba eredete, tulajdonságai, példa

2.1. A t-próba eredete

A következő részfejezet megírásához a [2] és [4] cikkeket vettem alapul.

A 20. század elején leginkább a normális eloszlást használták statisztikai problémák modellezésében. Ez lehetővé tette jól működő statisztikai próbák alkalmazását nagy elemszámú minták esetén. Kisebb elemszámú minták esetén azonban ezek a próbák nem működtek jól. William Sealy Gosset statisztikus és vegyész ennek a problémának a megoldására dolgozta ki a t-eloszlást használó t-próbát. Ezzel két minta átlagának viszonyáról tudott pontosabb képet alkotni. Munkája nagy hatással volt azokra az alkalmazásokra, amelyeknél praktikus vagy etikai okok miatt nem állnak rendelkezésre nagy elemszámú minták.

Miután Gosset kémiai és matematikából diplomát szerzett, a Guinness sörgyárban kezdett el dolgozni. Itt többek között az alapanyagok költséghatékony tesztelésének kidolgozásával bízták meg. Munkája során észrevette, hogy a kis elemszámú minták átlaga nem normális eloszlást követ, így az eddig használt módszerek nem bizonyultak megfelelőnek. Megfigyelte, hogy



2.1. ábra. William Sealy Gosset (1876– 1937), 1908 körül

minél kisebb a minta elemszáma, annál keskenyebb és magasabb a minta alapján kirajzolt hisztogram. A 20. század elején ezt a megfigyelését egy belső, céges jelentésben közzétette, amire felfigyelt a vezetőség, és javasolta, hogy egy hivatásos matematikus segítségével dolgozza ki alaposabban az elméletét. Ennek hatására Gosset felkereste Karl Pearsont, akinek segítségével kidolgozta az elméletét, és 1908-ben megjelent a „The probable error of a mean” [3] című cikke. Ezt a cikket Student álnéven tette közzé, miokán a vezetőség nem kívánta publikálni a cég adatait. Innen származik a Student-féle t-próba elnevezés is.

Később Pearson, Fisher és Gosset leveleztek többek között a t-próbáról, amely később heves vitává is fajult. A vitában felmerült annak a kérdése is, hogy mi történik a normális eloszlásra kidolgozott tesztekkel, ha nincs biztosítva a tesztelendő minta normalitása. Fisher felvetette, hogy ebben az esetben a minta más mutatóit is érdemes lehet figyelembe venni, az átlagon és a tapasztalati szóráson kívül. Többek között erről is szó lesz a szakdolgozatban.

További érdekesség, hogy a vita egyik pontján Gosset felvetette, hogy egy nullhipotézist akkor lehet elutasítani, ha van egy erős ellenhipotézis. Ebből

alakult ki a hipotézisvizsgálat Neyman és Pearson által kidolgozott máig használt formája.

2.2. A t-próba

A t-próba tehát a statisztikában használt egyik legelterjedtebb próba. Azt kapjuk meg a segítségével, hogy egy feltételezetten normális eloszlású, ismeretlen szórású minta várható értéke szignifikánsan különbözik-e egy előre megadott m értéktől. Ennek eldöntéséhez a hipotézisvizsgálat folyamata vezet.

A hipotézisvizsgálatban két, egymásnak ellentmondó állítást fogalmazunk meg, a nullhipotézist (H_0), illetve az ellenhipotézist (H_1). A nullhipotézis az előzetes tudásunkat, az ellenhipotézis a bebizonyítandó hipotézist fogalmazza meg. A módszer során a nullhipotézist tudjuk elvetni, vagy elfogadni.

Természetesen szeretnénk minél kisebb valószínűséggel rosszul dönteni a vizsgálat során. Hogyha a nullhipotézis igaz, de elutasítjuk, azt elsőfajú hibának hívjuk, ha viszont a nullhipotézis hamis, de elfogadjuk, azt másodfajú hibának hívjuk. A statisztikában az elsőfajú hiba súlyosabbnak számít, ezért annak mértékét a vizsgálat elején szeretnénk maximalizálni egy α érték megadásával, ami általában 5, esetleg 10 százalék szokott lenni. A másodfajú hiba valószínűsége β , az elemszám növelésével csökkenthető. Fontos, hogy ha csökken α , nő β , és viszont.

Legyen $X_1, X_2, \dots, X_n \sim N(m, \sigma^2)$, ahol m és σ ismeretlen. A hipotézisvizsgálat az α megadásával és a hipotézisek megfogalmazásával kezdődik, ami után a próbastatisztika kiszámolása következik, ami a t-próba esetén:

$$t = \frac{\bar{X} - m_0}{s_n^*} \sqrt{n}, \quad (2.1)$$

ahol

- \bar{X} a minta átlaga,

- m_0 a hipotetikus várható értéke és
- s_n^* a minta korrigált tapasztalati szórása.

Ezután a próbastatisztika értékét össze kell vetnünk a megfelelő c_α kritikus értékkel, amit a t-próba esetén a t-eloszlás táblázatából nyerhetünk ki, hiszen a t-próbastatisztika $f = n - 1$ szabadsági fokú t-eloszlást követ. Ehhez szükség van α -ra, illetve a fent említett szabadsági fokra.

A felállított hipotézisek alapján három esetet különböztetünk meg:

- ha $H_0 : m \geq m_0$, $H_1 : m < m_0$ alakúak a hipotéziseink, akkor azt alsó egyoldalú hipotézisnek hívjuk, és ebben az esetben, ha $t < -c_\alpha$, akkor a nullhipotézist elvetjük,
- ha $H_0 : m \leq m_0$, $H_1 : m > m_0$ alakúak a hipotéziseink, akkor azt felső egyoldalú hipotézisnek hívjuk, és ebben az esetben, ha $t > c_\alpha$, akkor a nullhipotézist elvetjük,
- ha $H_0 : m = m_0$, $H_1 : m \neq m_0$ alakúak a hipotéziseink, akkor azt kétoldalú hipotézisnek hívjuk, és ebben az esetben, ha $|t| > c_{\frac{\alpha}{2}}$, akkor a nullhipotézist elvetjük.

A t-eloszlás sűrűségfüggvénye:

$$t_f(x) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{\sqrt{f\pi}\Gamma\left(\frac{f}{2}\right)} \left(1 + \frac{x^2}{f}\right)^{-(f+1)/2},$$

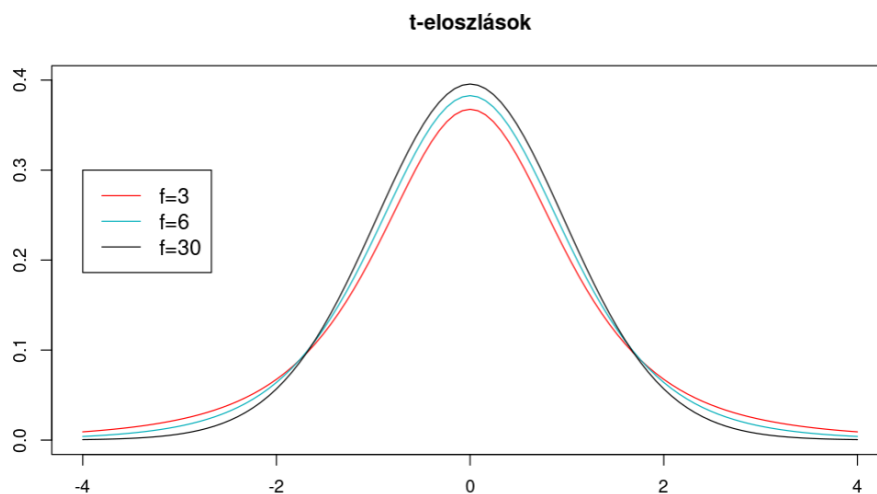
ahol

- f a szabadsági fok,
- Γ pedig a gamma-függvény.

Néhány különböző szabadsági fokú t-eloszlás sűrűségfüggvény grafikonja a 2.2 ábrán látható.

2.3. Példa

Nézzünk egy példát! Egy gyógyszerceg azt állítja, hogy az általuk forgalmazott fogyasztó kapszuláktól egy hónap alatt legalább 10 kg-ot lehet fogyni.



2.2. ábra. t-eloszlások 3, 6 és 30 szabadsági fokokkal

A tablettát 12 túlsúlyos tesztelő egy hónapig az előírásoknak megfelelően szedte, majd a következő súlyoktól szabadult meg (egész számra kerekítve):

$$-3, 3, 5, 7, 7, 8, 8, 11, 12, 12, 12, 16$$

ahol az értékeket kg-ban értjük, és a negatív értékek azt jelentik, hogy az adott illető hízott. Igaza van-e a gyógyszercégnek $\alpha = 0,05$ -ös szignifikancia szinten az ismert adatok alapján, hogy a tablettájuk valóban legalább 10 kg-os fogyást eredményez?

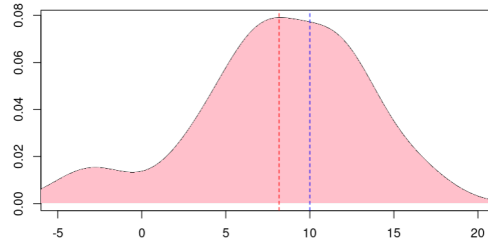
A feladat alapján a következő hipotéziseket tudjuk megfogalmazni:

$$H_0 : m \geq 10,$$

$$H_1 : m < 10.$$

A próbastatisztika a (2.1) képlet alapján:

$$t = \frac{8,17 - 10}{5,02} \sqrt{12} \approx -1,26$$



2.3. ábra. Parzen-Rosenblatt-becsléssel a mintára illesztett sűrűségfüggvény. Piros vonal jelzi a mintából számolt tényleges átlagot, a kék vonal pedig a hipotetikus várható értéket.

az $f = 11$ és $\alpha = 0,05$ értékekhez tartozó kritikus érték $c_\alpha = 1,796$. Mivel $-1,26 \geq -1,796$, így nem tudjuk elvetni a nullhipotézist, miszerint a tabletta legalább 10 kg-os fogyást eredményez. Vizuális szemléltetés a 2.3 ábrán.

Általában nem kézzel, hanem számítógéppel végezzük ezeket a próbákat, ilyenkor a p-érték lesz segítségünkre, ami annak a valószínűsége, hogy ha a nullhipotézis igaz, akkor megkaphatjuk az adott mintát. Ennél a feladatnál 0,1162 adódik, amiből az látszik, hogy jónak tűnik a próbastatisztika alapján hozott döntés.

Az imént részletezett feladat egy egymintás, egyoldali esetet vizsgált. Kétmintás esetben is működik a t-próba, ilyenkor a minták várható értékek az egyenlőségükről állítunk fel hipotéziseket. A kétoldaliság pedig a hipotézisekben használt relációkra utal, kétoldali esetekben $=$, illetve \neq jeleket használunk.

2.4. A Welch-próba

Kétmintás esetben tehát két mintasorozatunk van, amiknek a várható értékeiknek az egymáshoz való viszonyát szeretnénk megvizsgálni úgy, hogy közben feltesszük, hogy a szórásaikat nem ismerjük, de azok egyenlők. Legyen $X_1, X_2, \dots, X_{n_1} \sim N(m_1, \sigma^2)$, illetve $Y_1, Y_2, \dots, Y_{n_2} \sim N(m_2, \sigma^2)$, ahol m_1, m_2 és σ ismeretlen, illetve a két mintasorozat független.

A kétmintás t-próba próbastatisztikája:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1 - 1) s_{n_1}^{*2} + (n_2 - 1) s_{n_2}^{*2}}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2}},$$

ami $m_1 = m_2$ esetén $f = n_1 + n_2 - 2$ szabadsági fokú t-eloszlást követ.

Abban az esetben, hogyha szintén normális eloszlásaink vannak, és a várható értéket szeretnénk tesztelni, viszont a szórások egyenlőségét nem tudjuk garantálni, Welch-próbát alkalmazhatunk.

A Welch-próba próbastatisztikája:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_{n_1}^{*2}}{n_1} + \frac{s_{n_2}^{*2}}{n_2}}},$$

ami tart a

$$f \approx \frac{\left(\frac{s_{n_1}^{*2}}{n_1} + \frac{s_{n_2}^{*2}}{n_2}\right)^2}{\frac{s_{n_1}^{*4}}{n_1^2(n_1-1)} + \frac{s_{n_2}^{*4}}{n_2^2(n_2-1)}}$$

szabadsági fokú t-eloszláshoz. Jobb közelítéshez érdemes feltenni, hogy $n_1 > 5$ és $n_2 > 5$.

3. fejezet

A minta normalitása

A t-próba használatának egyetlen feltétele a vizsgált X változó normalitása. A való életben azonban gyakran nem tudjuk ezt biztosítani, ezért felmerülhet a kérdés, hogy miképp befolyásolja a próba érvényességét, ha nincs biztosítva a normalitás. Azt mondjuk, hogy egy statisztikai eljárás robusztus, hogyha kevésbé érzékeny azokra a feltételekre, amiktől függ. Mennyire robusztus a t-próba?

A fejezet megírásához Vargha [8] cikkét vettem alapul.

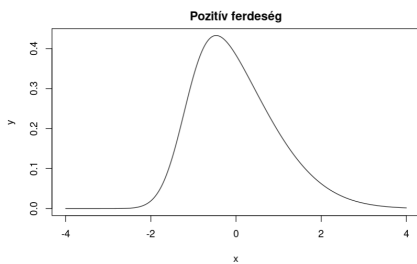
3.1. Ferdeség és csúcsosság fogalma

A ferdeség az eloszlás egy jellemzője, ami azt fejezi ki, hogy egy eloszlás mennyire aszimmetrikus. A következő képlettel lehet megadni:

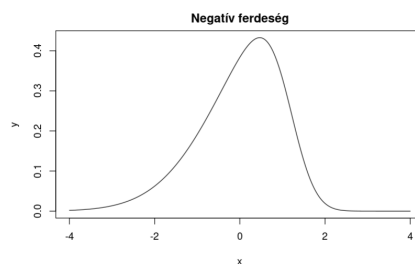
$$\alpha_3 = \frac{E(X - m)^3}{\sigma^3},$$

ahol σ a szórást jelöli, az m pedig a várhatóértéket.

- ha szimmetrikus az eloszlás, akkor $\alpha_3 = 0$,
- ha $\alpha_3 < 0$, akkor balra "nyúlik el" a sűrűségfüggvény,
- ha $\alpha_3 > 0$, akkor jobbra "nyúlik el" a sűrűségfüggvény.

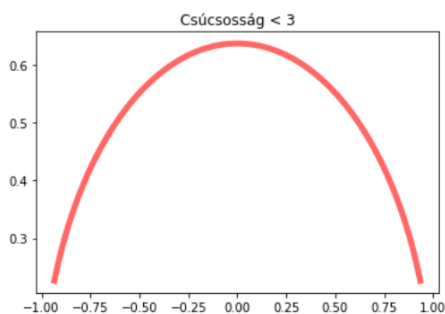


(a) $\alpha_3 = 0.7$ ferdeséggel

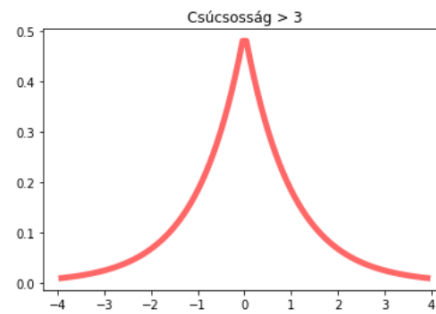


(b) $\alpha_3 = -0.7$ ferdeséggel

3.1. ábra. Különböző α_3 ferdeségre példa



(a) $\alpha_4 = 2$ csúcsossággal



(b) $\alpha_4 = 6$ csúcsossággal

3.2. ábra. Különböző α_4 csúcsosságra példa

A csúcsosság a ferdeséghez hasonlóan az eloszlás formájának egyik jellemzője. Azt fejezi ki, hogy egy eloszlás mennyire vastagszélű, illetve hegyes. Következő a képlete:

$$\alpha_4 = \frac{E(X - m)^4}{\sigma^4},$$

ahol σ a szórást jelöli, az m pedig a várhatóértéket.

- a normális eloszlások csúcsosságára $\alpha_4 = 3$,
- ha $\alpha_4 < 3$, akkor kevésbé csúcsos és vékonyszélű az eloszlás,
- ha $\alpha_4 > 3$, akkor csúcsosabb és vastagszélűbb.

3.2. Mi befolyásolja a t-próba robusztusságát?

Korábbi (1935-1996) tanulmányok alapján több dolgot is kimutattak az egymintás t-próbával kapcsolatban:

- A próba terjedelmére nagyobb hatással van az eloszlás ferdesége, mint annak a csúcsossága.
- A próba egyoldali ellenhipotézis esetén érzékenyebb, mint a kétoldali esetben.
- Kétoldali ($H_1 : m \neq m_0$) és alsó egyoldali ($H_1 : m < m_0$) ellenhipotézis esetén minél nagyobb az α_3 ferdeség, annál nagyobb az elsőfajú hiba valószínűsége.
- Érdekes módon felső egyoldali ($H_1 : m > m_0$) ellenhipotézis esetén pont az ellenkezőjét figyelték meg, tehát minél ferdebb az eloszlás, annál jobban csökken az elsőfajú hiba valószínűsége (ezzel csökken a próba ereje is, vagyis annak a valószínűsége, hogy helyesen elutasítjuk a nullhipotézist).

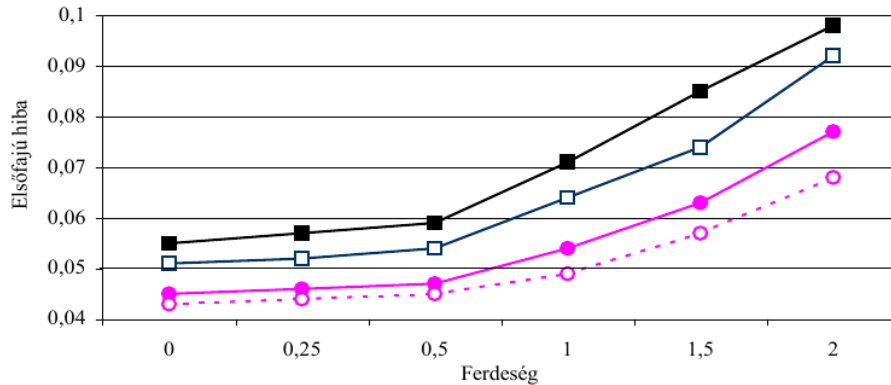
Ezek alapján láthatjuk, hogy ezek a megfigyelések leginkább a ferdeséget vették figyelembe, holott több korábbi munkában is látszik, hogy a ferdeség mellett a csúcsosság is fontos jellemző a téma vizsgálatához.

Később igazolták, hogy az α_4 csúcsosság is szerepet játszik a kis és közepes nagyságú mintájú egymintás t-próba robusztusságában.

A csúcsosságot is figyelembe véve, az figyelhető meg, hogy ahogy nő α_3 , úgy tolódik feljebb a lehetséges α_4 értékek tartománya. Ezenfelül, ahogyan azt a [8] cikkben lévő ábra is mutatja, minél nagyobb α_3 és minél kisebb α_4 , annál nagyobb az elsőfajú hiba valószínűsége (alsó egyoldali ellenhipotézis esetén).

Az, hogy mi számít alacsony, magas, stb. csúcsosságnak a 3.3 ábránál, azt Ramberg [6] munkája alapján definiálta Vargha a [8] munkájában.

Ramberg adott α_3 és α_4 értékek alapján adta meg a többek között általa általánosított lambda-eloszlás λ_1 , λ_2 , λ_3 és λ_4 paramétereit. Ezt használta fel Vargha.



3.3. ábra. Az egymintás t-próba elsőfajú hibája 5 százalékos szignifikancia-szint és $n = 10$ esetén különböző ferdeségi és csúcsossági szinteken [8, 1. ábra]. Az ábrán a telítetlen fekete négyzet jelzi a normális csúcsosságot, a telített fekete négyzet az alacsony csúcsosságot, a telített rózsaszín kör a magas csúcsosságot, és a telítetlen rózsaszín kör az igen magas csúcsosságot.

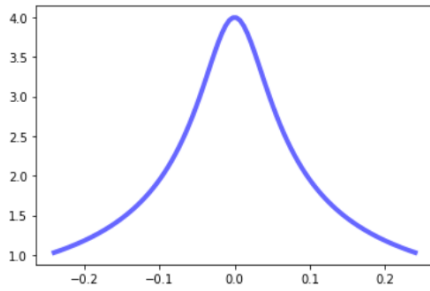
3.3. Lambda-eloszlás

A lambda-eloszlás különlegessége, hogy nem igazán használják modellezésre, hanem annál inkább más eloszlások közelítésére. Ennek oka többek között az, hogy pár kivételtől eltekintve nincs zárt alakja se a sűrűségfüggvénynek, se az eloszlásfüggvénynek. Kvantilisfüggvénnyel (az eloszlásfüggvény inverzével) szokás megadni, ami a következőképp néz ki:

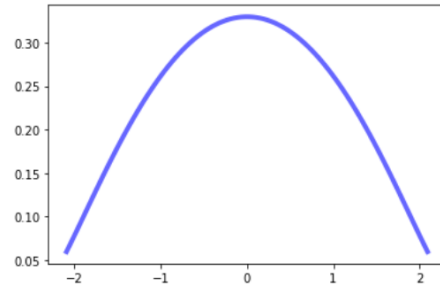
$$Q(p; \lambda) = \frac{p^\lambda - (1-p)^\lambda}{\lambda}, (0 \leq p \leq 1)$$

Néhány tulajdonsága:

- $\lambda \rightarrow 0$ esetén pontosan a logisztikus eloszlást kapjuk,
- $\lambda = 1$ esetén pontosan a $[-1, 1]$ intervallumon egyenletes eloszlást kapjuk,
- $\lambda = -1$ esetén a Cauchy eloszlást közelíti,
- $\lambda = 0,14$ esetén a normális eloszlást közelíti.



(a) $\lambda = 4$



(b) $\lambda = 0,4$

3.4. ábra. Egyparaméterű lambda-eloszlások sűrűségfüggvényei különböző paraméterek mellett

Ahogy fentebb volt említve, többek között Ramberg a lambda-eloszlás egyik általánosítását a következőképp adta meg négy új paraméter bevezetésével:

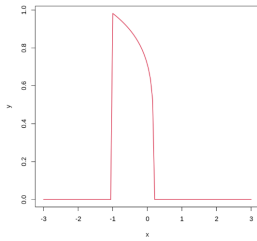
$$Q(p; \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, (0 \leq p \leq 1)$$

ahol

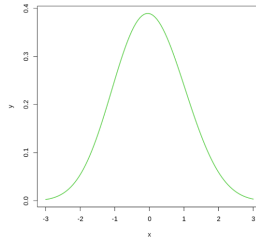
- λ_1 az eltolás paramétere,
- λ_2 a skála paramétere,
- λ_3, λ_4 az alakparaméterek.

A paraméterek által felvehető értékek a következőképpen alakulnak:

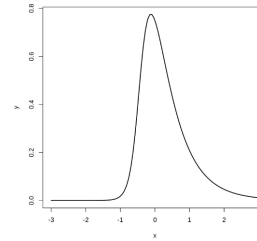
	λ_1	λ_2	λ_3	λ_4
1	$\lambda_1 \in \mathbb{R}$	< 0	< -1	> 1
2	$\lambda_1 \in \mathbb{R}$	< 0	> 1	< -1
3	$\lambda_1 \in \mathbb{R}$	> 0	> 0	> 0
4	$\lambda_1 \in \mathbb{R}$	< 0	< 0	< 0
5	$\lambda_1 \in \mathbb{R}$	< 0	$\lambda_3 \in (-1, 0)$	> 1
6	$\lambda_1 \in \mathbb{R}$	< 0	> 1	$\lambda_4 \in (-1, 0)$



$$\begin{aligned} \text{(a)} \quad \lambda_1 &= -1,009, \\ \lambda_2 &= 0,8397, \\ \lambda_3 &= 0, \\ \lambda_4 &= 0,8541 \end{aligned}$$



$$\begin{aligned} \text{(b)} \quad \lambda_1 &= -0,083, \\ \lambda_2 &= 0,2429, \\ \lambda_3 &= 0,1625, \\ \lambda_4 &= 0,1903 \end{aligned}$$



$$\begin{aligned} \text{(c)} \quad \lambda_1 &= -0,187, \\ \lambda_2 &= 0,1118, \\ \lambda_3 &= -0,0146, \\ \lambda_4 &= -0,0647 \end{aligned}$$

3.5. ábra. Ramberg-féle lambda-eloszlások sűrűségfüggvényei különböző paraméterek mellett

Az 5-ös esetben az alábbi egyenlőtlenségnek igaznak kell lennie:

$$\frac{(1 - \lambda_3)^{1-\lambda_3} (\lambda_4 - 1)^{\lambda_4-1}}{(\lambda_4 - \lambda_3)^{\lambda_4-\lambda_3}} < -\frac{\lambda_3}{\lambda_4}.$$

A 6-os egyenletnél pedig a következő egyenlőtlenségnek kell igaznak lennie:

$$\frac{(1 - \lambda_4)^{1-\lambda_4} (\lambda_3 - 1)^{\lambda_3-1}}{(\lambda_3 - \lambda_4)^{\lambda_3-\lambda_4}} < -\frac{\lambda_4}{\lambda_3}.$$

Megjegyzés: Ramberg óta mások is foglalkoztak a lambda-eloszlás általánosításával, több olyan eredmény is született, amiknél nem ennyire szigorúak a paraméterekre vonatkozó megkötések.

4. fejezet

Az egymintás t-próba robosztus változatai

4.1. Gayen- és Johnson-próba

1949-ben A. K. Gayen publikált egy olyan eljárást, ami figyelembe veszi a minta ferdeségét és csúcsosságát is. Ehhez a tesztelni kívánt eloszlás sűrűségfüggvényének a hatodrendű Edgeworth sorfejtését használja a következő módon:

$$f(x) = \phi(x) - \frac{\alpha_3}{3!}\phi^{(3)}(x) + \frac{\alpha_4}{4!}\phi^{(4)}(x) + \frac{10\alpha_3^2}{6!}\phi^{(6)}(x),$$

ahol

- α_3 az eloszlás ferdesége,
- α_4 az eloszlás csúcsossága,
- $\phi^{(i)}(x)$ pedig a standard normális eloszlás sűrűségfüggvényének a i . deriváltja.

Ezután hosszas számolás után megkapja a t-statisztika sűrűségfüggvényét:

$$p(t) = p_0(t) + \alpha_3 p_{\alpha_3}(t) - \alpha_4 p_{\alpha_4}(t) + \alpha_3^2 p_{\alpha_3}(t),$$

ahol $p_0(t)$ a t-eloszlás sűrűségfüggvénye, a további tagok pedig a α_3 és α_4 miatti módosulások.

Ezt a kifejezést használja ahhoz, hogy pontosabban kiszámolja a terjedelmet. Tehát a próba során használt táblázat értékeit pontosítja, a próbastatisztika képlete változatlan marad. Eljárásának egyik hátránya lehet, hogy feltételezi az elméleti momentumok ismeretét. Illetve, képletének felírásakor véges számú tagját meghagyva a sornak, előfordulhat, hogy a kapott valószínűség nem 0 és 1 közötti.

Később, 1979-ben Norman J. Johnson publikálta a t-próba egy másik módosítását, ami figyelembe veszi a minta ferdeségét:

$$t_1 = t + a_3 \sqrt{n} \left(\frac{1}{6n} + \frac{(\bar{x} - m_0)^2}{3s^2} \right), \quad (4.1)$$

ahol

- \bar{x} a minta átlaga,
- m_0 a hipotetikus várható értéke,
- s a minta szórása,
- a_3 pedig az α_3 elméleti ferdeségi együttható mintabeli becslése.

Johnson az általa adott képlethez többek között Cornish-Fisher sorfejtést használt, ami lehetővé teszi egy tetszőleges X valószínűségi változó felírását a momentumaival és egy $\zeta \sim N(0, 1)$ valószínűségi változó segítségével. Az \bar{X} másodrendű CF(\bar{X}) Cornish-Fisher-féle közelítésére a következő adódik:

$$\text{CF}(\bar{X}) = m_0 + \frac{\sigma}{\sqrt{n}} \zeta + \frac{\mu_3}{6n\sigma^2} (\zeta^2 - 1) + O\left(n^{-\frac{3}{2}}\right)$$

ahol

- $\zeta \sim N(0, 1)$,
- m_0 X várható értéke,

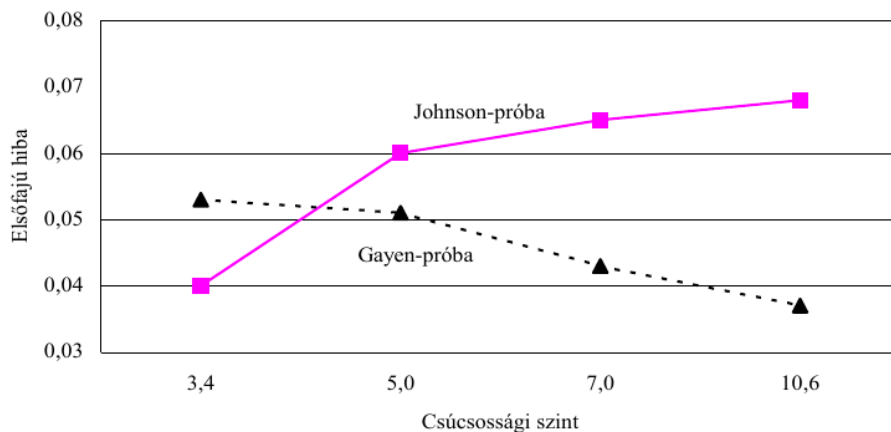
- σ a szórás,
- $\mu_3 = \sigma^3 \alpha_3 (= E(X - \mu)^3)$.

Ennél a próbánál a szokásos t-eloszlás táblázatát használjuk, a próbasta-tisztika kiszámolási módja pedig a fent leírt módon van változtatva. Em-pirikusan tesztelték, hogy a Johnson-féle t-próba legalább 13 elemű minták esetén működik jól.

Ha nem szimmetrikus a minta, akkor ez a próba jobbnak bizonyult két-oldali és alsó egyoldali ellenhipotézis esetén, mint a korábban tárgyalt egy-mintás t-próba.

Johnson úgy vélte, hogy az ő eljárása egymintás kétoldali ellenhipotézis esetén jobb, mint Gayené. Ezt az állítást más kutatók is elfogadták, ezért sokáig nem volt a kutatások témája Gayen módszere. Később azonban úgy látták, mégsem jelentéktelen tényező a minta csúcsossága.

Szimulációs kísérleteket végeztek lambda-eloszlásokon Johnson- és Gayen-próbával. Azt figyelték meg, hogy a Johnson-próba esetén, ahogy nő α_4 , úgy nő az elsőfajú hiba valószínűsége. Gayen-próba esetén viszont, ahogy nő α_4 , úgy csökken az elsőfajú hiba valószínűsége.



4.1. ábra. A Johnson- és a Gayen-próba elsőfajú hibájának függése az α_4 csúcsossági szinttől $\alpha_3 = 1$ ferdeségi szinten, $\alpha = 0,05$ és $n = 10$ esetén [8, 2. ábra].

4.2. Saját szimuláció

Az R programnyelv segítségével saját szimulációkat is végeztem a hagyományos t-próba és a Johnson-féle t-próba tesztelésére. Lambda eloszlásokat használtam, 5 féle csúcossági szinten, és minden csúcossági szinten belül 5 féle ferdeségi szintet vizsgáltam. Az elemszámot 5-re, 7-re, 9-re, 13-ra, 17-re és 21-re néztem meg.

Egy adott n elemszám, csúcosság és ferdeség esetén 1000 alkalommal generáltam n elemű mintát a megfelelő csúcosságú és ferdeségű lambda-eloszlásból. A generált mintákra 5%-os terjedelemmel hagyományos t-próbát és Johnson-féle t-próbát alkalmaztam.

A táblázatokban a kötőjellel (-) elválasztott értékek közül az első jelenti a t-próba esetén azt, hogy hány százalékban utasítottuk el a nullhipotézist, a második ugyanezt jelenti, csak a Johnson-féle t-próba esetén.

A nullhipotézis minden esetben: $H_0 : m = 0$, ami mindegyik paraméterezésnél igaznak bizonyul. Tehát minél közelebb van a táblázatban egy érték a 0,05-höz, annál jobb a teszt.

A 4.1 táblázatból látszik, hogy felső egyoldali ellenhipotézis esetén minden alkalommal a Johnson-féle t-próba legalább olyan jól teljesít, mint a hagyományos t-próba, de általában jóval jobb is.

A 4.2 táblázatból látszik, hogy alsó egyoldali esetben is jobbnak bizonyul a Johnson-féle t-próba (kivétel az $n=5$ és $n=7$ esetekben néhány ferdeségi és csúcossági szint).

Míg felső egyoldali esetben nagyobbak az értékek a 0,05-nél, tehát túl nagy a terjedelem, addig alsó egyoldali esetben kisebbek, tehát úgy tűnik, hogy az elvártnál kisebb a terjedelem.

α_4	α_3	n=5	n=7	n=9	n=13	n=17	n=21
5	1,4	0,125 - 0,071	0,14 - 0,079	0,117 - 0,063	0,097 - 0,052	0,093 - 0,054	0,109 - 0,071
5	1,3	0,129 - 0,098	0,121 - 0,082	0,104 - 0,067	0,101 - 0,073	0,092 - 0,054	0,083 - 0,054
5	1,2	0,095 - 0,081	0,104 - 0,084	0,102 - 0,076	0,093 - 0,073	0,092 - 0,065	0,083 - 0,053
5	1,1	0,112 - 0,087	0,082 - 0,067	0,091 - 0,079	0,102 - 0,076	0,068 - 0,056	0,09 - 0,069
5	1	0,092 - 0,087	0,091 - 0,078	0,083 - 0,067	0,08 - 0,068	0,073 - 0,055	0,069 - 0,064
4,6	1,2	0,109 - 0,087	0,115 - 0,078	0,105 - 0,074	0,097 - 0,07	0,094 - 0,066	0,081 - 0,05
4,6	1,1	0,113 - 0,094	0,106 - 0,077	0,089 - 0,072	0,095 - 0,072	0,088 - 0,07	0,101 - 0,076
4,6	1	0,1 - 0,079	0,111 - 0,087	0,077 - 0,06	0,083 - 0,067	0,073 - 0,062	0,084 - 0,071
4,6	0,9	0,08 - 0,076	0,088 - 0,068	0,091 - 0,08	0,085 - 0,066	0,09 - 0,078	0,067 - 0,055
4,6	0,8	0,089 - 0,072	0,078 - 0,072	0,06 - 0,053	0,082 - 0,077	0,066 - 0,057	0,06 - 0,053
4,2	1,2	0,119 - 0,07	0,113 - 0,064	0,114 - 0,07	0,075 - 0,047	0,087 - 0,044	0,079 - 0,044
4,2	1,1	0,123 - 0,096	0,12 - 0,084	0,091 - 0,064	0,099 - 0,063	0,087 - 0,058	0,092 - 0,056
4,2	1	0,107 - 0,085	0,103 - 0,076	0,09 - 0,07	0,075 - 0,055	0,078 - 0,058	0,077 - 0,06
4,2	0,9	0,087 - 0,07	0,09 - 0,079	0,105 - 0,08	0,075 - 0,059	0,08 - 0,066	0,072 - 0,054
4,2	0,8	0,094 - 0,077	0,09 - 0,071	0,079 - 0,065	0,071 - 0,058	0,063 - 0,051	0,068 - 0,054
3,8	1,1	0,113 - 0,08	0,112 - 0,057	0,108 - 0,066	0,112 - 0,076	0,084 - 0,048	0,095 - 0,061
3,8	1	0,109 - 0,074	0,09 - 0,07	0,088 - 0,062	0,085 - 0,053	0,092 - 0,057	0,065 - 0,039
3,8	0,9	0,085 - 0,067	0,085 - 0,071	0,092 - 0,067	0,058 - 0,045	0,07 - 0,048	0,088 - 0,061
3,8	0,8	0,097 - 0,073	0,091 - 0,069	0,076 - 0,06	0,085 - 0,071	0,084 - 0,067	0,073 - 0,058
3,8	0,7	0,087 - 0,07	0,087 - 0,073	0,068 - 0,05	0,062 - 0,054	0,059 - 0,053	0,076 - 0,068
3,2	0,9	0,104 - 0,063	0,094 - 0,06	0,106 - 0,058	0,084 - 0,048	0,078 - 0,047	0,082 - 0,042
3,2	0,8	0,111 - 0,082	0,086 - 0,056	0,072 - 0,054	0,076 - 0,048	0,069 - 0,054	0,074 - 0,049
3,2	0,7	0,075 - 0,061	0,102 - 0,081	0,078 - 0,054	0,073 - 0,047	0,084 - 0,067	0,074 - 0,057
3,2	0,6	0,072 - 0,072	0,081 - 0,065	0,071 - 0,058	0,065 - 0,049	0,062 - 0,05	0,069 - 0,057
3,2	0,5	0,079 - 0,066	0,066 - 0,057	0,069 - 0,061	0,066 - 0,059	0,074 - 0,059	0,087 - 0,07

4.1. táblázat. Felső egyoldali ellenhipotézis esetén 5%-os szignifikanciaszinten a hagyományos t-próba és a Johnson-féle t-próba elutasításának az aránya H_0 mellett

α_4	α_3	n=5	n=7	n=9	n=13	n=17	n=21
5	1,4	0,02 - 0,02	0,017 - 0,022	0,014 - 0,027	0,019 - 0,032	0,02 - 0,038	0,03 - 0,041
5	1,3	0,018 - 0,021	0,01 - 0,019	0,015 - 0,019	0,022 - 0,034	0,026 - 0,044	0,026 - 0,039
5	1,2	0,016 - 0,017	0,029 - 0,04	0,026 - 0,038	0,028 - 0,039	0,022 - 0,028	0,029 - 0,043
5	1,1	0,02 - 0,026	0,027 - 0,038	0,023 - 0,031	0,028 - 0,04	0,027 - 0,04	0,032 - 0,043
5	1	0,015 - 0,02	0,018 - 0,02	0,029 - 0,037	0,024 - 0,031	0,029 - 0,042	0,034 - 0,048
4,6	1,2	0,019 - 0,02	0,02 - 0,024	0,022 - 0,03	0,015 - 0,025	0,025 - 0,044	0,027 - 0,039
4,6	1,1	0,024 - 0,02	0,021 - 0,025	0,022 - 0,026	0,021 - 0,034	0,029 - 0,037	0,034 - 0,046
4,6	1	0,025 - 0,028	0,022 - 0,038	0,02 - 0,03	0,024 - 0,039	0,022 - 0,032	0,023 - 0,037
4,6	0,9	0,031 - 0,03	0,025 - 0,03	0,024 - 0,033	0,034 - 0,042	0,028 - 0,039	0,025 - 0,04
4,6	0,8	0,028 - 0,025	0,026 - 0,026	0,023 - 0,035	0,034 - 0,039	0,039 - 0,052	0,028 - 0,034
4,2	1,2	0,014 - 0,014	0,013 - 0,015	0,018 - 0,031	0,023 - 0,035	0,024 - 0,038	0,023 - 0,043
4,2	1,1	0,024 - 0,022	0,015 - 0,023	0,017 - 0,027	0,036 - 0,044	0,022 - 0,026	0,038 - 0,057
4,2	1	0,02 - 0,026	0,02 - 0,02	0,023 - 0,03	0,029 - 0,038	0,032 - 0,044	0,037 - 0,045
4,2	0,9	0,025 - 0,028	0,019 - 0,031	0,025 - 0,033	0,025 - 0,039	0,042 - 0,053	0,041 - 0,053
4,2	0,8	0,03 - 0,031	0,02 - 0,032	0,029 - 0,043	0,023 - 0,04	0,032 - 0,053	0,039 - 0,044
3,8	1,1	0,016 - 0,02	0,014 - 0,018	0,024 - 0,029	0,032 - 0,042	0,03 - 0,042	0,031 - 0,04
3,8	1	0,022 - 0,016	0,02 - 0,028	0,03 - 0,033	0,023 - 0,028	0,037 - 0,047	0,023 - 0,031
3,8	0,9	0,03 - 0,031	0,026 - 0,027	0,023 - 0,031	0,029 - 0,034	0,031 - 0,042	0,031 - 0,04
3,8	0,8	0,027 - 0,03	0,021 - 0,033	0,029 - 0,033	0,027 - 0,04	0,029 - 0,04	0,037 - 0,045
3,8	0,7	0,029 - 0,027	0,018 - 0,024	0,029 - 0,032	0,038 - 0,05	0,041 - 0,056	0,031 - 0,041
3,2	0,9	0,036 - 0,034	0,027 - 0,025	0,032 - 0,03	0,028 - 0,035	0,029 - 0,035	0,025 - 0,037
3,2	0,8	0,024 - 0,02	0,04 - 0,04	0,026 - 0,027	0,023 - 0,035	0,029 - 0,036	0,039 - 0,046
3,2	0,7	0,021 - 0,026	0,031 - 0,031	0,029 - 0,037	0,031 - 0,033	0,035 - 0,039	0,041 - 0,051
3,2	0,6	0,03 - 0,036	0,023 - 0,031	0,029 - 0,035	0,037 - 0,044	0,044 - 0,06	0,037 - 0,041
3,2	0,5	0,031 - 0,033	0,031 - 0,039	0,028 - 0,04	0,044 - 0,05	0,034 - 0,042	0,033 - 0,039

4.2. táblázat. Alsó egyoldali ellenhipotézis esetén 5%-os szignifikanciaszinten a hagyományos t-próba és a Johnson-féle t-próba elutasításának az aránya H_0 mellett

4.3. Szimulációk és regresszió

Kíváncsiak lehetünk a Johnson- és Gayen-féle próbák elsőfajú hibájának és erejének az előrejelzésére, előre rögzített szignifikanciaszinten a mintaelem-szám és az eloszlásnak becsült ferdeségi és csúcsossági szintje alapján.

Ehhez először szimulációs elemzéseket végzett Vargha a fent tárgyalt lambda-eloszlásokon, illetve kevert normális eloszlásokon.

A lambda-eloszlásnak az a tulajdonsága, hogy a könnyen számolható kvantilisfüggvényével van megadva, hasznos amikor lambda-eloszlású véletlen változókat szeretnénk generálni. Ugyanis, ha Q tetszőleges kvantilisfüggvény, F a hozzá tartozó eloszlásfüggvény, Y pedig $[0, 1]$ -en egyenletes eloszlású valószínűségi változó, akkor az $X = Q(Y)$ valószínűségi változó kvantilisfüggvénye pont Q . Hiszen

$$\begin{aligned} F_X(t) &= P(X < t) = P(Q(Y) < t) = \\ &= P(F^{-1}(Y) < t) = P(Y < F(t)) = F(t), \end{aligned}$$

tehát X eloszlásfüggvénye éppen F , azaz kvantilisfüggvénye éppen Q . Ezzel a módszerrel könnyedén tudunk az egyenletes eloszlás és a kvantilisfüggvény segítségével adott eloszlású mintát generálni.

A kevert normális eloszlások szimulációs alkalmazása azért lehet jó ötlet, mert gyakori eset, hogy egy populáció több különböző jelentős alpopuláció-ra bontható. Ilyen alpopulációk lehetnek például: nemek szerint, jövedelem szerint, lakhely szerint, végzettség szerint, stb. bonthatóak. Még hogyha az alpopulációk normális eloszlást is követnek, az eredeti populáció általában még ebben az esetben sem lesz normális. Az ilyen több normális eloszlásból összetevődő eloszlást kevert normális eloszlásnak nevezzük. Ha egy ilyen kevert Z eloszlást egy $N(0, 1)$ eloszlású X és egy $N(\mu, \sigma)$ eloszlású Y változóból hozzuk létre úgy, hogy az X eloszlásának aránya p , Y -é pedig $q = 1 - p$, akkor a Z eloszlás paraméterei a következőképpen határozhatók meg:

Z várható értéke:

$$E(Z) = (1 - p)\mu = q\mu,$$

Z szórásnégyzete:

$$D^2(Z) = p + q\sigma^2 + pq\mu^2,$$

Z ferdeségi együtthatója:

$$\alpha_3(Z) = \frac{pq [3\mu (\sigma^2 - 1) - (1 - 2p)\mu^2]}{D^2(Z)^{3/2}},$$

Z csúcsossági együtthatója:

$$\alpha_4(Z) = \frac{3p + 3q\sigma^4 + 6p^2q\mu^2\sigma^2 + 6pq^2\mu^2 + pq(1 - 3pq)\mu^4}{D^2(Z)^2}.$$

Vargha különböző, meghatározott p, μ, σ értékekkel szimulált kevert normális eloszlásokat, amiket be akart sorolni kategóriákba a csúcsossági szintjüktől függően. Ehhez alapul vette a Ramberg táblájából számolt alacsony csúcsossági szinteket, és polinomiális regresszióval becslőfüggvényt készített:

$$\alpha_4(\min) = 1,8 + 0,74 |\alpha_3| + 0,52 (\alpha_3)^2 + 0,41 |\alpha_3|^3.$$

Ami tehát adott α_3 -hoz megadja a legkisebb α_4 értéket. Ennek segítségével már kifejezhető a relatív csúcsosság:

$$\alpha_4(\text{dev}) = \alpha_4 - \alpha_4(\min).$$

A relatív csúcsosság alapján már be lehet sorolni az eloszlásokat csúcsossági kategóriákba. Így a lambda-eloszlásokhoz hasonlóan, szimulációk segítségével kiszámította az elsőfajú hiba valószínűségét és a próbák erejét.

Vargha a két eloszlástípusra külön végzett lineáris regressziós elemzéseket. A magyarázott változók a hagyományos egymintás kétoldali t-próba, a Gayen- illetve a Johnson-féle t-próba ereje és elsőfajú hiábája különbö-

ző szignifikanciaszinten. A magyarázó változók minden esetben: n , α_3 , α_4 , $\alpha_4(dev)$. A modell illeszkedését az R korrelációs együttható és a standard hiba vonatkozásában vizsgálta.

A 4.3, 4.4 táblázatok Vargha táblázatainak egy részhalmaza, korlátozva az elsőfajú hiba becslésére 5%-os szignifikanciaszinten. A magyarázott változó oszlopában az első karakter a próba fajtáját jelenti, a második pedig a szignifikanciaszintet.

Ha például az egymintás t-próba elsőfajú hibájára vagyunk kíváncsiak 5 százalékos szignifikanciaszinten (t5), $n = 21$ esetén egy olyan eloszlás esetében, amelyre $\alpha_3 = 1,4$, $\alpha_4 = 5$, akkor a következő számítási lépéseket kell elvégezni:

$$\begin{aligned} \alpha_4(dev) &= \alpha_4 - \alpha_4(\min) = \\ 6,6 - (1,8 + 0,74 \cdot 1,4 + 0,52 \cdot 1,4^2 + 0,41 \cdot 1,4^3) &= 1,6198; \\ \hat{\alpha} &= 0,053 + 21 \cdot (-0,0002) + 1,4 \cdot 0,0024 + 5 \cdot 0,0029 + 1,6198 \cdot (-0,0045) = \\ &= 0,0594 \end{aligned}$$

Magyarázott változó	A	n	α_3	α_4	$\alpha_4(dev)$	R
t5	0,053	-0,00020	0,0024	0,0029	-0,0045	0,91
J5	0,046	-0,00010	0,0030	-0,0002	0,0036	0,88
G5	0,040	0,00008	0,0016	0,0023	-0,0036	0,88

4.3. táblázat. Lambda eloszlások esetén

Magyarázott változó	A	n	α_3	α_4	$\alpha_4(dev)$	R
t5	0,050	-0,00033	-0,0008	0,0063	-0,0094	0,82
J5	0,028	-0,00036	0,0072	0,0102	-0,0030	0,69
G5	0,034	-0,00003	0,0033	0,0063	-0,0084	0,70

4.4. táblázat. Kevert normális eloszlások esetén

4.4. A trimmelt t-próba

A következő alfejetben leírtak az [1] tankönyv megfelelő fejezetén alapulnak. A t-próba egy másik fajtája a trimmelt t-próba. Ennek a használatához elő-

ször bevezetek két fogalmat. A trimmelt átlag a várható értéknek egy olyan becslése, ami egy adott γ , általában 0,1 és 0,2 közötti trimmelési arány mellett figyelmen kívül hagyja a minta két szélét. Az így kiszámolt trimmelt átlag sokszor jobban írja le egy minta jellemzőjét, mint a hagyományos átlag. A winsorizált szórás gondolata hasonló a trimmelt átlagéhoz. Ugyanúgy adott egy γ trimmelési arány, amivel kijelöli a minta legnagyobb és legkisebb értékeit. Ezután a megmaradt mintának vesszük a minimumát és a maximumát, majd ezekkel helyettesítjük a trimmeléskor kijelölt értékeket. A trimmelt egymintás t-próbával a

$$H_0 : m = m_0$$

alakú nullhipotéziseket tesztelhetjük. Ennek a próbának a végrehajtásához a következő próbastatisztikát kell kiszámítani:

$$T_t = \frac{(1 - 2\gamma)(\bar{x}_t - m_0)}{s_w/\sqrt{n}},$$

ahol

- \bar{x}_t az n-elemű minta trimmelt átlaga,
- s_w a mintabeli winsorizált szórás,
- γ pedig az előre rögzített trimmelési arány.

Amennyiben H_0 igaz, úgy T_t próbastatisztika $f = n - 2g - 1$ szabadsági fokú t-eloszlást követ, ahol $2g$ a trimmelés miatt figyelmen kívül hagyott adatok száma, ami egyenlő $\gamma \cdot n$ egész részének a kétszeresével. A próba elvégzésének a lépései a hagyományos t-próbához képest csak a fentiekben tér el.

4.5. Példa

Vegyük a következő mintát és egy megfogalmazható nullhipotézist és ellenhipotézist:

$$-7, 5, 5, 6, 7, 8, 8, 9, 11, 13, 13, 30$$

$$H_0 : m = 5$$

$$H_1 : m \neq 5$$

A trimmelési arány pedig legyen $\gamma = 0,15$.

$$2g = 2 \cdot [\gamma \cdot n] = 2 \cdot [0,15 \cdot 12] = 2 \cdot [1,8] = 2 \cdot 1 = 2$$

Tehát a mintából összesen 2 adatot hagyunk el, a legkisebbet (-7) és a legnagyobbat (30). A trimmelt mintaátlag a maradék 10 adat átlaga, vagyis

$$\bar{x}_t = \frac{5 + 5 + 6 + 7 + 8 + 8 + 9 + 11 + 13 + 13}{10} = \frac{85}{10} = 8,5.$$

s_w winsorizált szóráshoz az új minta:

$$5, 5, 5, 6, 7, 8, 8, 9, 11, 13, 13, 13,$$

ennek a mintának a szórása:

$$s_w = 3,2$$

Behelyettesítve a T_t -statisztika képletébe:

$$T_t = \frac{0,7 \cdot (8,5 - 5)}{3,2/\sqrt{12}} = 2,649$$

A T_t -statisztika szabadságfoka:

$$f = n - 2g - 1 = 9$$

A t-eloszlás táblázatából a megfelelő szabadsági fokú és szignifikanciaszinthez tartozó kétoldali kritikus érték:

$$c_{\frac{\alpha}{2}} = 2,262$$

mivel $|2,649| > 2,262$, ezért a nullhipotézist elvetjük.

Hogyha ugyanerre a mintára és hipotézisekre a hagyományos t-próbát alkalmazzuk, akkor

$$t = 1,64 \text{ és } c_{\frac{\alpha}{2}} = 2,20$$

adódik. Itt $|1,64| \leq 2,20$, ezért a nullhipotézist elfogadjuk.

Az eltérő eredményre az eredeti minta nagy szórása (8,42) a magyarázat, ami a nevezőben csökkenti a próbastatisztika értékét, így ez a szignifikancia ellenében hat. A trimmelt esetben a szórást jelentősen sikerült csökkenteni.

Az [5] cikkben a trimmelés és a Johnson-féle t-próba együttes alkalmazásával jobb elsőfajú hiba értékeket értek el, mint a hagyományos t-próba esetén.

5. fejezet

Többváltozós t-próba

A többváltozós t-próba a hagyományos t-próba többváltozós változata, ami hasznos lehet a gyakorlatban bonyolultabb problémák megoldásában. Ebben az esetben is a várható értéket teszteljük, ismeretlen szórásmátrix esetén, tehát:

Legyenek $X_1, \dots, X_n \in N_d(m, \Sigma)$ független valószínűségi változók és tekintsük a következő hipotéziseket:

$$H_0 : m = m_0,$$

$$H_1 : m \neq m_0.$$

A többváltozós t-próba próbastatisztikája levezethető a valószínűséghányados próbával a később leírt módon.

5.1. Valószínűséghányados próba

A valószínűséghányados próba egy statisztikai próba, ami az egyes hipotézisek által felírható likelihood-függvények hányadosának az értéket hasonlítja össze egy meghatározott k értékkel, ami alapján dönteni lehet a nullhipotézisről.

$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$ alakú hipotézisek esetén:

$$\lambda = \frac{f(x_1, \dots, x_n, \theta_0)}{\sup_{\theta \in \Theta} f(x_1, \dots, x_n, \theta)}$$

$-2 \ln \lambda$ határeloszlása χ_d^2 , ahol d az eloszlás szabadsági foka, ami egyenlő a teljes paraméterter dimenziójának és a nullhipotézis paraméterterének a dimenziójának a különbségével. Amikor $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$ alakú, tehát egyszerű hipotéziseket fogalmazzunk meg, akkor nincs paraméter, amit becsülnünk kéne, ezért λ a következőképpen egyszerűsödik:

$$\lambda = \frac{f(x_1, \dots, x_n, \theta_0)}{f(x_1, \dots, x_n, \theta_1)}$$

- ha $\lambda < k$, akkor elutasítjuk a nullhipotézist,
- ha $\lambda > k$, akkor elfogadjuk a nullhipotézist,
- ha $\lambda = k$, akkor egy q valószínűséggel utasítjuk el a nullhipotézist.

A Neyman-Pearson lemma kimondja, hogy ez a próba a legerősebb a nála nem nagyobb terjedelműek között.

5.2. A t-próbastatisztika levezetése valószínűség-hányados próbával

Mielőtt a többváltozós t-próbát levezetnénk a valószínűség-hányados próbával, nézzük meg először a [7] cikk alapján az egyváltozós esetet:

Legyen $X_1, X_2, \dots, X_n \sim N(m, \sigma^2)$, ahol m és σ ismeretlen. A következő hipotéziseket fogalmazzuk meg:

$$H_0 : m = m_0$$

$$H_1 : m \neq m_0$$

Ekkor a paraméterek maximum-likelihood becslését használva:

$$\hat{m} = \bar{x} \text{ és } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\begin{aligned} f(x_1, \dots, x_n, \hat{m}, \hat{\sigma}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\hat{\sigma}}} \exp\left[-\frac{(x_i - \hat{m})^2}{2\hat{\sigma}^2}\right] = \left[\frac{1}{2\pi\hat{\sigma}^2}\right]^{\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^n (x_i - \hat{m})^2}{2\hat{\sigma}^2}\right] \\ &= \left[\frac{1}{2\pi \left(\frac{1}{n}\right) \sum_{i=1}^n (x_i - \bar{x})^2}\right]^{\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2 \left(\frac{1}{n}\right) \sum_{i=1}^n (x_i - \bar{x})^2}\right] \end{aligned}$$

Nézzük meg a becsült paraméterekkel a likelihood-függvényt:

$$f(x_1, \dots, x_n, \hat{m}, \hat{\sigma}) = \left[\frac{ne^{-1}}{2\pi \sum_{i=1}^n (x_i - \bar{x})^2}\right]^{n/2}$$

A nullhipotézis paraméterterében hasonlóan:

$$f(x_1, \dots, x_n, m_0, \hat{\sigma}) = \left[\frac{ne^{-1}}{2\pi \sum_{i=1}^n (x_i - m_0)^2}\right]^{n/2}$$

Vegyük a valószínűséghányadost:

$$\begin{aligned} \lambda &= \frac{f(x_1, \dots, x_n, m_0, \hat{\sigma})}{f(x_1, \dots, x_n, \hat{m}, \hat{\sigma})} = \left[\frac{ne^{-1}}{2\pi \sum_{i=1}^n (x_i - m_0)^2}\right]^{n/2} / \left[\frac{ne^{-1}}{2\pi \sum_{i=1}^n (x_i - \bar{x})^2}\right]^{n/2} \\ &= \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - m_0)^2}\right]^{n/2} \end{aligned}$$

Alakítsuk át a nevezőt:

$$\sum_{i=1}^n (x_i - \bar{x} + \bar{x} - m_0)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - m_0) \sum_{i=1}^n (x_i - \bar{x}) + \sum_{i=1}^n (\bar{x} - m_0)^2.$$

Mivel $\sum_{i=1}^n (x_i - \bar{x}) = 0$, ezért a következő módon egyszerűsödik a képlet:

$$\sum_{i=1}^n (x_i - m_0)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - m_0)^2$$

Ezt felhasználva λ képletébe:

$$\lambda = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - m_0)^2} \right]^{n/2} = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - m_0)^2} \right]^{n/2}.$$

A számlálót és a nevezőt is osztva $\sum_{i=1}^n (x_i - \bar{x})^2$ -nel adódik, hogy:

$$\lambda = \left[\frac{1}{1 + \frac{n(\bar{x} - m_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]^{n/2} \leq k,$$

ahol k az a választott szám, aminél ha kisebb λ , akkor elutasítjuk a nullhipotézist.

Mindkét oldalt a $\frac{2}{n}$ -edikre emelve, reciprokot véve, kivonva 1-et, szorozva $(n-1)$ -gyel, a bal oldalon a számlálót és a nevezőt is n -nel osztva, illetve gyököt vonva mindkét oldalból adódik, hogy:

$$\frac{|\bar{x} - m_0|}{s_n^*/\sqrt{n}} \geq k^*,$$

ahol $k^* = \sqrt{(n-1)(k^{-2/n} - 1)}$.

Mivel a bal oldalon pont egy t -eloszlású valószínűségi változó abszolútértékét kapjuk, ezért k^* -ot $t_{\frac{\alpha}{2}, n-1}$ -nek választva éppen α -terjedelmű próbát kapunk.

5.3. Többváltozós t-próba levezetése valószínűséghányados próbával

Az egyváltozós esethez hasonlóan a többváltozós t-próbát is levezethetjük a valószínűséghányados próba segítségével. Az alfejezet a [9] tankönyv megfelelő fejezete alapján készült.

Nézzük először a többváltozós normális eloszlás sűrűségfüggvényét:

Ha $X \in N_d(m, \Sigma)$, tehát X d -dimenziós normális eloszlású m várható érték vektorral és $\Sigma > 0$ kovarianciamátrixszal, akkor sűrűségfüggvénye:

$$f_X(x) = (2\pi)^{-\frac{d}{2}} |\det(\Sigma)|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x - m)^T \Sigma^{-1} (x - m) \right\}.$$

Legyenek $X_1, \dots, X_n \in N_d(m, \Sigma)$ független valószínűségi változók és tekintsük a következő hipotéziseket:

$$H_0 : m = m_0,$$

$$H_1 : m \neq m_0.$$

Tehát erre szeretnénk használni a valószínűséghányados próbát. Nézzük először a likelihood-függvényt:

$$\begin{aligned} f(x_1, \dots, x_n, m, \Sigma) &= \\ &= (2\pi)^{-\frac{nd}{2}} (\det \Sigma)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \operatorname{tr} \left[\Sigma^{-1} \left(\sum_{i=1}^n (x_i - m) (x_i - m)^T \right) \right] \right\}. \end{aligned}$$

A nullhipotézis paraméterterében a likelihood-függvény maximuma éppen:

$$(2\pi)^{-\frac{nd}{2}} \det \left(\frac{1}{n} \sum_{i=1}^n (x_i - m_0) (x_i - m_0)^T \right)^{-\frac{n}{2}} \exp \left\{ -\frac{nd}{2} \right\}.$$

A teljes paraméterterben pedig a likelihood-függvény maximumára

$$(2\pi)^{-\frac{nd}{2}} \left(\det \frac{S}{n} \right)^{-\frac{n}{2}} \exp \left\{ -\frac{nd}{2} \right\}$$

adódik, ahol

$$S = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$$

a szórásmatrix n -szeresének empirikus becslése. Ezért a likelihood-hányados-próba a

$$\left(\frac{\det S}{\det \sum_{i=1}^n (x_i - m_0)(x_i - m_0)^T} \right)^{\frac{n}{2}} = \left[1 + n (\bar{x} - m_0)^T S^{-1} (\bar{x} - m_0) \right]^{-\frac{n}{2}}$$

kifejezés értékén alapul. Mivel ez $n (\bar{x} - m_0)^T S^{-1} (\bar{x} - m_0)$ monoton függvénye, ezért a likelihood-hányados-próba alapján a $H_0 : m = m_0$ hipotézist elvetjük, ha $T^2 = n (\bar{x} - m_0)^T S^{-1} (\bar{x} - m_0) \geq k$, egy előre meghatározott k -ra. Ezt a k -t az alapján számolhatjuk ki, hogy $\frac{n-d}{d} T^2$ eloszlása H_0 mellett éppen $d, n-d$ paraméterű F -eloszlás, azaz $F_{(d, n-d)}$.

6. fejezet

Összegzés

Dolgozatomban először ismertettem a t-próbát, illetve a Welch-próbát. Majd a ferdeség és a csúcsosság fogalmát is mint a t-próba robosztusságát potenciálisan befolyásoló minta jellemzőit. Összefoglaltam a próba robosztusságával kapcsolatos eredményeket, és az ehhez szükséges számítógépes szimulációk okán írtam a lambda-eloszlásokról is. Saját szimulációmban a t-próba és a Johnson-féle t-próba teljesítményét vetettem össze az elsőfajú hiba szempontjából, amiben a Johnson-féle t-próba teljesített jobban, összhangban a korábbi kutatásokkal. Később ismertettem egy regressziós eredményt, ami a hagyományos t-próba, Johnson-féle t-próba, illetve a Gayen-féle t-próba elsőfajú hibáját és erejét hivatott előre jelezni különböző szignifikanciaszinteken. Utána a trimmelt t-próbáról is írtam, illetve a többváltozós t-próbáról is.

Irodalomjegyzék

- [1] Vargha András. Matematikai statisztika. *Pólya kiadó*, 2000.
- [2] Angus Brown. The strange origins of the Student’s t-test. *Physiology News Magazine*, 16:19, 2008.
- [3] WS Gosset. The probable error of a mean. *Biometrika*, 6:1–25, 1908.
- [4] Erich L Lehmann. “Student” and small-sample theory. *Selected works of EL Lehmann*, pages 1001–1008, 2012.
- [5] Wei-Ming Luh and Jiin-Huarng Guo. Transformation works for non-normality? On one-sample transformation trimmed t methods. *British Journal of Mathematical and Statistical Psychology*, 54(2):227–236, 2001.
- [6] John S Ramberg, Edward J Dudewicz, Pandu R Tadikamalla, and Edward F Mykytka. A probability distribution and its uses in fitting data. *Technometrics*, 21(2):201–214, 1979.
- [7] Laura Simon. Introduction to mathematical statistics, 2023.
- [8] András Vargha. Robusztussági vizsgálatok az egymintás t-próbával. *Statisztikai Szemle*, 81(10):872–890, 2003.
- [9] Móri Tamás és Székely Gábor. Többváltozós statisztikai analízis. *Műszaki Tankönyvkiadó. Budapest*, 1986.