R nyelv – statisztikai tesztek

Vadon Viktória

2024/2025/I. félév

Tartalomjegyzék

1. Illeszkedésvizsgálat

ld. qqplot a 9. hétről.

2. Konfidenciaintervallum

- adott egy eloszlás, pl. standard normális – legyen X a val. vált.
- adott egy konfidenciaszint, pl 90%, 0.9
- konfidenciaintervallum: keresek egy $[a, b]$ intervallumot, hogy $\mathbb{P}(a \leq X \leq b) = 0.9$.
- más szavakkal: keresek egy intervallumot, ahol a sűrűségfüggvény (példánkban: a Gauss haranggörbe) integrálja 0.9
- legegyszerűbb kvantilisekkel leírni az intervallum végpontjait
- erre több megoldás is lehetséges.
  - tipikusan a várható érték körül szimmetrikusan jelöljük ki az intervallumot: 0.05-kvantilistől 0.95-kvantilisig – pozitívban és negatívban is a legritkább 5%-ot „dobjuk el”, összesen 10%-ot dobtunk el, 90%-ot tartottunk meg.

  - de 0-kvantilistől 0.9-kvantilisig is 0.9 eséllyel esik a val. vált. – itt felső 10%-ot „dobtuk el”.
  - vagy 0.1-kvantilistől 1-kvantilisig is 0.9 az integrál.

3. Tesztek elmélete és várható érték

- statisztikai teszt bemutatása az (egymintás) T-teszt példáján keresztül
- **T-teszt feltételei:**
  - független, azonos eloszlású minta $X_1, X_2, \ldots, X_n$
  - normális (Gauss) eloszlás – fontos! de gyakorlatban sokszor feltehető, CHT miatt.
  - várható érték ismeretlen μ.
  - szórást a mintából becsüljük, *korrigált* empirikus szórásnégyzettel
- T-teszt célja: várható érték ismeretlen μ egyenlő-e 0-val? (általánosabban: 0 helyett egy adott μ_0 számmal egyenlő-e?)
- konyhanyelven, a statisztikai teszt konyhanyelven egy „fogadás”.
  - 2 alternatíva közül szeretnénk tippelni, és minél nagyobb eséllyel eltalálni a helyes választ
  - ún. kétoldalas T-teszt esetén:
  - H_0, ún. „nullhipotézis”: $\mu = 0$
  - H_1, ún. „ellenhipotézis”: $\mu \neq 0$
  - ha H_0-ra tippelünk, azt mondjuk, elfogadjuk a nullhipotézist; ha H_1-re tippelünk, azt mondjuk, elutasítjuk a nullhipotézist.
  - mikor van igazunk, és mikor tévedünk?

		TIPPÜNK	
		H_0	H_1
IGAZ	H_0	igazunk van	elsőfajú hiba: „hamis negatív”
	H_1	másodfajú hiba: „hamis pozitív”	igazunk van

- **hogyan tudunk „okosan fogadni”?**
  - nyilván szeretnénk, hogy minél nagyobb eséllyel igazunk legyen. de hogyan?
- teszt-statisztika fogalma:
  - aggregáljuk az adatokat, a sok adatból egyetlen mérőszámot számolunk: ún. teszt-statisztika – a statisztika a mintának egy függvénye.
  - T-teszt esetén, ha a várható érték a kérdés: nagy számok törvénye, centrális határeloszlástétel szerint a mintaátlag $\bar{X} = (X_1 + \cdots + X_n)/n$ konvergál a várható értékhez!
  - T-teszt esetén T-statisztika: $T = \bar{X}/(S/\sqrt{n})$
  - itt S a mintából számolt korrigált empirikus szórás.
- a statisztika értéke alapján tippelünk:
  - nyilván a mintaátlag még mindig egy véletlen szám, de közel van a várható értékhez – ha 0 környéki számot látok, jó eséllyel csak véletlen ingadozás; ha

0-tól távolit, akkor viszont jó eséllyel tényleg más a várható érték!
  - a statisztika egy valós értékű függvény, lehetséges értékei szerint előre felszeleteljük a számegyenest:
  - ún. elfogadási tartomány: ha a statisztika ide esik, H_0-t fogok tippelni – T-tesztnél egy 0 körüli intervallum
  - ún. kritikus tartomány: ha a statisztika ide esik, H_1-t fogok tippelni – kiugró, nem túl esélyes értékek halmaza
- mekkora a tévedés esélye?
  - ha H_0 igaz lenne, azaz az ismeretlen várható érték μ tényleg 0 (μ_0) lenne, akkor pontosan ismernénk az adatok eloszlását – következésképpen a statisztikának az eloszlását is, mert az a minta függvénye!
  - T-teszt esetén a T-statisztika Student t-eloszlást követne, df=n-1 szabadsági fokkal (n: az adatok száma – de a szórás S egy becsült paraméter)
  - elsőfajú hiba: H_0 igaz, a várható érték 0, de a véletlen folytán mégis kiugró átlagot kaptunk, ami a kritikus tartományba esik
  - ha már kiválasztottam a kritikus tartományt, az elsőfajú hibának pontosan ki tudjuk számolni az esélyét, mert H_0-t feltételezve tudjuk, hogy T milyen eloszlást követ, tudjuk, milyen eséllyel esik a kritikus tartományba!
  - a másodfajú hiba esélye ellentétesen mozog az elsőfajúval: úgy tudom csökkenteni az elsőfajú hiba, „hamis negatív" esélyét, minél nagyobb tartományban fogadom el H_0-t – akkor viszont annál nagyobb eséllyel kapok „hamis pozitívat", másodfajú hibát!
  - a másodfajú hibát nem tudjuk számszerűsíteni, mert az épp azt jelenti, hogy mi elhittük a várható érték 0, pedig a tényleges várható érték nem 0, de fogalmunk sincs, mennyi!
  - mivel az elsőfajú hibát tudom megszabni, sokszor a „bizonyítandó" eredmény H_1-be kerül – pl. hogy a vizsgált gyógyszer hatása a beteg vérnyomására nem 0.
- az elsőfajú hiba mértékét szabjuk meg
  - előre eldöntjük, mekkora esélyt engedünk meg „hamis negatív"-nak – alkalmazástól függően pl. 10%, 5%, 1%
  - nyilván minél nagyobb a minta, a CHT szerint annál közelebb kerül az átlag a várható értékhez, annál inkább csökken mindkét hibám.
  - kisebb mintánál érdemes nagyobbra hagyni az elsőfajú hiba esélyét – különben a másodfajú hiba lesz aránytalanul nagy.
  - a fix elsőfajú hibához szabjuk meg az elfogadási és kritikus tartományt – pl. 10% hibaesély más szavakkal 90%-os konfidenciaszintet jelent – az elfogadási tartomány 90%-os konfidenciaintervallum (ld. föntebb) lesz a megfelelő t-eloszlásból!
  - fenti példánkban, amikor H_1 opció $\mu \neq 0$, szimmetrikus konfidenciaintervallumot választunk

3.1. T-teszt R-ben

- futtatás: tegyük fel, hogy z egy vector, ami a minta elemeit tartalmazza.
- hívás: t.test(z), eredménye komplex objektum
- néhány adattag:
- t: számolt t-statisztika, lekérés: $stat
- df: (számolt) szabadsági fok, lekérés: $par
- p: p-value, lekérés: $p.value
  - ez alapján döntünk H_0 és H_1 tipp között, de kicsit más a logika, mint korábban
  - p.value azt mutatja, (legalább) mekkora elsőfajú hibát kellene megengednem, hogy az adott t-statisztika már a kritikus tartományba kerüljön
  - azaz, pl. ha 10%-os elsőfajú hibát engeek meg:
  - ha p.value túl nagy (> 0.1), akkor H_0-t fogok tippelni.
  - ha p.value elég kicsi (p.value < 0.1), akkor H_1-t fogok tippelni – ilyenkor mondjuk, hogy a teszt **szignifikáns** – az eltérés (feltételezhetően) nem csak a véletlen műve – a tesztelt gyógyszer statisztikailag „igazoltan" hatásos.
  - (mellesleg ez is kvantilises alapon működik)
- confidence interval: konfidencia-intervallum a tényleges várható értékre – feltéve a normális eloszlást, és a számolt szórást, az átlag körüli intervallum, amibe nagy eséllyel beleesik a tényleges várható érték is.

3.2. T-teszt variánsai, paraméterek R-ben

- súgóoldal l ?t.test
- 0 helyett általános várható érték: pl. mu = 2: H_0 $\mu = 2$
- conf.level = 0.95 – a megadott konfidencia-intervallum szintjének beállítása; 0.95-ös konfidencia-szint 0.05-ös hibavalószínűségnek felel meg. mint a p.value-nál írtuk fentebb, a tesztre magára nincs hatással, a p.value-t kiköpi, és mi döntünk!
- amiről eddig beszéltünk, az szimmetrikus, avagy kétoldali teszt volt. ez az alapértelmezés: `alternative = "two.sided"`
  - létezik egyoldalas teszt is!
  - `alternative="less"`: H_1: $\mu < 0$ – de továbbra is H_0: $\mu = 0$. korábbi példában: a gyógyszer nincs hatással a beteg vérnyomására, VS csökkenti azt.
  - itt H_0-nak és H_1-nek nem kell a teljes számegyenest lefedni!
  - de ha $= 0$ és < 0 a két opció, akkor nyilván nagy pozitív átlagra inkább = 0-t fogunk tippelni.
  - az elfogadási és kritikus tartománynak viszont le kell fednie a teljes számegyenest!
  - a korábbi példánknál maradva, legyen 0.9 a konfidencia-szint – 0.1 az elsőfajú hiba. most csak a negatívban akarunk kritikus tartományt – most 10%-ot egyben dobunk el, az eloszlás legalsó 10%-át! (avagy: az elfogadási tartománynak a 90%-os konfidencia-intervallum a 0.1-kvantilistől az 1-kvantilisig

terjed.)
  - analóg módon `alternative = "greater"` kulcsszóval H_1: $\mu > 0$.
- eddig egymintás tesztről volt szó: azonos eloszlásból származó val. változókról próbáltuk eldönteni, a várható értéke 0-e.
  - létezik kétmintás teszt is:
  - 2 független mintával: t.test(x,y,paired=FALSE) – pl. x vector: $X_1, \ldots, X_n$, férfiak magassága, y vector: $Y_1, \ldots, Y_m$: nők magassága; itt megengedett, hogy különböző hosszú vector-ok legyenek! de feltsszük, hogy mindkettő normális eloszlás. mire jó? ha arra vagyunk kíváncsiak, férfiak és nők magassága azonos eloszlást követ-e – pontosabban, feltéve a normális eloszlást és azonos szórást*, azonos-e a várható érték.
  - ha a szórás egyenlő a két mintában, var.equal = TRUE – de csak ha tudjuk, vagy mert az adatok szakértője megsúgta, vagy mert teszteltük (F-teszt) – ha nem garantált az egyenlő szórás, az R automatikusan kompenzál és másképp számol, hagyjuk alapértelmezett FALSE-on, ha nem biztos.
  - szórás azonos-e? F-próba: var.test(x,y) – H_0: azonos szórás, H_1: különböző – kis p.value-ra H_1-et tippelünk
  - 2 függő mintával: t.test(x,y,paired=TRUE) – vérnyomásos példában, x,y, vector-okban ugyanazon betegek vérnyomása a gyógyszer bevétele előtt, majd utána – itt tulajdonképpen az x-y különbségre számol egy egymintást tesztet.

4. Függetlenség vizsgálata

- tegyük fel, hogy adott (X_i, Y_i) minta, együttes eloszlásból
- X_i-k egy x vector-ban (vagy factor-ban), Y_i-k egy y vector-ban (vagy factor-ban); vagy data.frame oszlopaiként
- pl. ugyanazon betegek vérnyomása és vércukorszintje.
- kérdés, hogy az X, Y változók függők vagy függetlenek-e?
- ehhez: Pearson-féle χ^2 (khí-négyzet)-teszt
- ha diszkrét változók: tekintsünk egy együttes gyakoriság-táblázatot

	Y=1	Y=2	Y=3
X=1	2	5	4
X=2	6	3	3

  jelentése: pl. 5 olyan (X_i, Y_i) pár van az összes 23 adatból, ahol $X_i = 1$, $Y_i = 2$.
- ha folytonos(ak): cut() függvénnyel intervallumokra vágjuk:
  - cut(x,breaks=n): n intervallumra vág, automatikusan számol töréspontokat
  - vagy cut(x,breaks=vekt), vekt az intervallumhatárok vektora – azaz benne van a min és max is! ha valami kilóg belőle, ott NA az érték.
  - eredménye mindig egy factor változó, ami azonos hosszú x-szel.
  - alapértelmezett factor szintek: az intervallumok, kezdőponttal és végponttal – labels = kulcsszóval megadhatók egy vektorban.

- feltételek: „elég nagy” méretű minta ; legalább (átlagosan) 5 elem / cella (az együttes gyakoriság-táblázatban)
- diszkrét változónál is előfordulhat, hogy kicsi a minta, és túl sokféle érték lehetséges – itt is szükség lehet értékek csoportosítására, intervallumokra vágásra!
- χ^2-teszt:
  - kiszámolja X és Y marginális eloszlását / relatív gyakoriságokat
  - ez alapján kiszámolja, hogy X és Y ha függetlenek lennének, mennyi lenne egy-egy cellában az elemek száma = várt érték
  - χ^2 statisztika = a (tényleges-várt)^2/várt cellaértékek összege
  - H_0: függetlenek, H_1: függőek
  - df: szabadsági fok, (sorok száma-1)*(oszlopok száma-1)
  - feltéve a függetlenséget, a statisztika χ^2(df) eloszlást követ (normális változók négyzetösszege) – mivel négyzetösszegből adódik, mindig pozitív!
  - mindig egyoldalas teszt: függetlenség esetén a tényleges és várt értékek közel egyeznek, a statisztika értéke kicsi; ha függőek, a tényleges értékek jobban eltérnek a várttól, a statisztika kiugró pozitív értéket kap.
- teszt R-ben:
  - chisq.test(frequency.table), vagy chisq.test(x,y)
  - X-squared, $stat: a számolt χ^2-statisztika
  - df, $par: szabadsági fok: (sorok száma-1)*(oszlopok száma-1).
  - $p.value: hasonló p-value, mint a t-test-nél – ha p.value kicsi, H_1-t, azaz függőséget tippelünk; ha p.value nagy, H_0-t, függetlenséget tippelünk.