PARAMÉTERES PRÓBÁK

Szórásra, varianciára vonatkozó próbák

Mennyire egységes (egyforma)? Mennyire homogén? Mekkora az ingadozása?

Egy minta esetén: ***X^2-próba varianciára***

Két minta esetén: **F-próba**

Több minta esetén: **Levene-próba** *VAGY* **Bartlett-próba**

Várható értékre vonatkozó próbák

Mekkora (az értéke)? Ugyanakkora-e? Nagyobb-e?

Egy minta esetén: ***egymintás t-próba***

Két független minta esetén: **Welch-próba** *(a kétmintás t-próba egy speciális esete, korrekciót tartalmaz arra*

Két (páronként) összefüggő minta esetén: **páros t-próba**

Több független minta esetén: **Varianciaanalízis (ANOVA)**

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1: vannak olyan i, j indexek, amelyekre $\mu_i \neq \mu_j$ (az i. alapsokaság és a j. alapsokaság várható értéke nem egyezik meg), *azaz nem mindegyik várható érték egyenlő*

Feltételek:
1. Szóráshomogenitás: a szórások egyenlősége minden csoportban.
2. A reziduumok (az értékeknek a csoportátlagoktól való eltérése) normalitása
3. A minták függetlenek. (Helyes mintavétel esetén teljesül.)

Páronkénti összehasonlítás – *post hoc* tesztek (pairwise comparisons)

(Csak) ha szignifikáns különbséget találunk, fontos kérdés, hogy mely csoportok várható értéke különbözik egymástól!

A fent tárgyalt ANOVA modell az úgynevezett ***egytényezős, teljes véletlen elrendezésű ANOVA***.

NEMPARAMÉTERES PRÓBÁK

Nemparaméteres próbáknál nem követeljük meg a normális (vagy egyéb) eloszlást.

Ha a normalitás erősen sérül, akkor a paraméteres próbáink helyett is rendelkezésre állnak nemparaméteres próbák (*Kruskal-Wallis* – az egytényezős, teljes véletlen elrendezésű ANOVA helyett; *Mann-Whitney* – a Welch-próba helyett, *Wilcoxon* – a páros t-próba, illetve az egymintás t- próba helyett).

> *Ezeket most nem tárgyaljuk részletesen, az R Commander-ben elérhetőek, de vigyázzanak!, ezeknek is vannak feltételei, például – ha a normalitást nem is – a szóráshomogenitást nem mindig nélkülözhetjük. Ha ilyen próbákat használnálnak, nézzenek utána a szakirodalomban, vagy kérdezzenek!*

A következőkben néhány olyan nemparaméteres próbáról lesz szó, ami az eloszlással kapcsolatos kérdéseket feszeget.

X^2-próba illeszkedésvizsgálatra

> *Erre láttunk már példákat, ide tartoznak a normalitásvizsgálatok! (SW, KS)*

Kérdés: a minta alapján az alapsokaság eloszlása egy feltételezett (hipotetikus) eloszlással megegyezik-e?

Fajtái:

- *tiszta*: a kérdéses eloszlás paramétereivel együtt adott
- *becsléses:* csak az eloszlás típusa van meghatározva (a paramétereket a mintából becsüljük)

H_0: az alapsokaság eloszlása megegyezik a hipotetikus eloszlással
H_1: az alapsokaság eloszlása nem egyezik meg a hipotetikus eloszlással

Feltétel: a megfigyelt és a várt gyakoriság kategóriánként legalább 5 legyen.

Ha a változó eloszlása folytonos, szükséges lehet kategóriákat / osztályokat / csoportokat létrehozni, mert a nemparaméteres, X^2-próbáknál ilyenekkel dolgozunk. Ugyanakkor sokszor már eleve ilyen formában vesszük fel az adatokat, ilyenkor a meglévő osztályokat használhatjuk. Előfordul az is, hogy olyan jellegű adatokat vizsgálunk, amelyeket nem is tudnánk másként, mint csoportokként jellemezni.

Számítások – *példa*

Szabályosnak tekinthető-e egy dobókocka, ha 120 dobásból 18-szor kaptunk egyest, 17-szer kettest, 25-ször hármast, 17-szer négyest, 15-ször ötöst és 28-szor hatost?

H_0: az eloszlás (diszkrét) egyenletes *(minden értéket azonos valószínűséggel vesz fel)*, a kocka szabályos
H_1: az eloszlás nem (diszkrét) egyenletes tiszta illeszkedésvizsgálat!

Kategóriák / osztályok létrehozása.	**a dobások lehetséges értékei: 1, 2, 3, 4, 5, 6**
A megfigyelt gyakoriságok meghatározása kategóriánként. *A mintából!* ⟶ f_i *Legalább 5! (ellenőrizzük)*	**könnyű:** **18, 17, 25, 17, 15, 28 *(> 5)***
A várt gyakoriságok meghatározása kategóriánként. *A hipotetikus eloszlásból!* ⟶ e_i *Legalább 5! (ellenőrizzük)*	**Egyenletes eloszlás, minden érték valószínűsége egyforma (1/6):** **20, 20, 20, 20, 20, 20 *(> 5)***
Számított érték: $X^2_{számított} = \sum_{i=1}^{k} \frac{(e_i - f_i)^2}{e_i}$, ahol *k* a kategóriák / osztályok száma	$\mathbf{X^2_{számított}} = \sum_{i=1}^{k} \frac{(e_i - f_i)^2}{e_i}$ **= 6,8**
Ez (k–1)* szabadsági fokú X^2-eloszlást követ, innen kaphatjuk a kritikus értéket vagy a p-értéket.	**a szabadsági fok: df = 5** **p = 0,236 > 0,05 = α, tehát H_0-t megtartjuk**

R-ben: **chisq.test(x=c(18,17,25,17,15,28),p=c(20/120,20/120,20/120,20/120,20/120,20/120))**

* Megjegyzés: becsléses illeszkedésvizsgálatnál a szabadsági fok k – 1 – b, ahol *b* a becsült paraméterek száma

2. példa

A Kis-Balatonból vett vízminták segítségével azt szeretnénk megállapítani, hogy a szervesanyag-tartalomra mint valószínűségi változóra feltételezhetjük-e a μ = 17 és σ = 2 paraméterű normális eloszlást? (α = 0,05)

Osztályok	Gyakoriságok
14,51-15,5	8
15,51-16,5	12
16,51-17,5	22
17,51-18,5	41
18,51-19,5	12
19,51-20,5	5

H_0: az eloszlás μ = 17 és σ = 2 paraméterű normális eloszlás
H_1: az eloszlás nem μ = 17 és σ = 2 paraméterű normális eloszlás

tiszta illeszkedésvizsgálat – megadott (μ = 17 és σ = 2) paraméterek

<table>
<tr><td>Kategóriák / osztályok létrehozása.</td><td>léteznek („Osztályok”)</td></tr>
<tr><td>A megfigyelt gyakoriságok meghatározása kategóriánként.
A mintából! ⟶ f_i
Legalább 5! (ellenőrizzük)</td><td>megvannak („Gyakoriságok”)

mindegyik legalább 5 ✓</td></tr>
<tr><td>A várt gyakoriságok meghatározása kategóriánként. A hipotetikus eloszlásból! ⟶ e_i
Legalább 5! (ellenőrizzük)</td><td>
<table>
<tr><th></th><th>Eloszlásfv (<)</th><th>Különbsége</th><th>Várt gyakoriságok</th></tr>
<tr><td>14,51-15,5</td><td>0,1056</td><td>0,1056</td><td>10,56</td></tr>
<tr><td>15,51-16,5</td><td>0,2255</td><td>0,1199</td><td>11,99</td></tr>
<tr><td>16,51-17,5</td><td>0,4013</td><td>0,1758</td><td>17,58</td></tr>
<tr><td>17,51-18,5</td><td>0,5987</td><td>0,1974</td><td>19,74</td></tr>
<tr><td>18,51-19,5</td><td>0,7734</td><td>0,1747</td><td>17,47</td></tr>
<tr><td>19,51 fölött</td><td>1</td><td>0,2266</td><td>22,66</td></tr>
</table>
mindegyik legalább 5 ✓</td></tr>
<tr><td>Számított érték:
$X^2_{számított} = \sum_{i=1}^{k} \frac{(e_i - f_i)^2}{e_i}$,
ahol k a kategóriák / osztályok száma</td><td>$X^2_{számított} = \sum_{i=1}^{k} \frac{(e_i - f_i)^2}{e_i}$ = 40,105</td></tr>
<tr><td>Ez (k–1) szabadsági fokú X^2-eloszlást követ, Innen kaphatjuk a kritikus értéket vagy a p-értéket.</td><td>a szabadsági fok: df = 5 (6 osztály volt!)
p = 1,422 · 10^{-7} < 0,05 = α, tehát H_0-t elvetjük</td></tr>
</table>

R-ben: **chisq.test(x=c(8,12,22,41,12,5),p=c(0.1056,0.1199,0.1758,0.1974,0.1747,0.2266))**

X^2-próba függetlenségvizsgálatra

Kérdés: a minta alapján elfogadhatjuk-e, hogy két változó független egymástól?

H_0: a két változó független egymástól
H_1: a két változó nem független egymástól *(…!)*

Feltétel: a megfigyelt és a várt gyakoriság kategóriánként legalább 5 legyen.

Függetlenek-e egymástól a szemszín és a testtömeg?

Mit jelent ez?

- ***Hogy a vékony, közepes teltségű, illetve telt emberek között egyforma arányban fordulnak-e elő a kék, zöld, barna és fekete szeműek.***
- ***Hogy ha egy vékony emberről meg kell mondanom, hogy milyen valószínűséggel kék szemű, pontosan ugyanazt mondom-e, mintha csak általában (egy ismeretlen tömegű emberről) kell megmondanom.***

H_0: a szemszín és a testtömeg független egymástól
H_1: a szemszín és a testtömeg nem független egymástól

A megfigyelések elrendezése: ***kontingenciatáblázat****ban, azaz kétdimenziós gyakorisági táblázatban*

megfigyelések	kék szemű	zöld szemű	barna szemű	fekete szemű
vékony	7	10	14	5
közepes	16	18	34	9
telt	10	6	15	6

Feltétel: a megfigyelt ✓ és a várt gyakoriság kategóriánként legalább 5 legyen.

Számítások

A megfigyelt gyakoriságok oszlop- és sorösszegeinek kiszámítása

7	10	14	5	**36**
16	18	34	9	**77**
10	6	15	6	**37**
33	**34**	**63**	**20**	**150**

☺

Ez alapján a várt gyakoriságok meghatározása az $e_{ij} = \frac{s_i \cdot o_j}{n}$ képlettel, ahol s_i az i. sor összege, o_i a j. oszlop összege, n pedig a táblázat teljes összege

7,9	**8,2**	**15,1**	**4,8**	36,0
16,9	**17,5**	**32,3**	**10,3**	**77,0**
8,1	**8,4**	**15,5**	**4,9**	37,0
33,0	**34,0**	63,0	20,0	150,0

77 * 34 / 150 = 17,45

Mj.: a feltétel – a várt gyakoriság kategóriánként legalább 5 legyen – enyhén sérül (4,8 illetve 4,9)!

Számított érték:

$X^2_{számított} = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(e_{ij} - f_{ij})^2}{e_{ij}}$,

ahol *r* a sorok, *c* az oszlopok száma

0,11	**0,41**	**0,08**	**0,01**	
0,05	**0,02**	**0,09**	**0,16**	
0,43	**0,68**	**0,02**	**0,23**	

$(14–15,1)^2 / 15,1 = 0,0801$

$X^2_{számított} = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(e_{ij} - f_{ij})^2}{e_{ij}}$ = **2,28**

Ez (r–1)(c–1) szabadsági fokú X^2-eloszlást követ, Innen kaphatjuk a kritikus értéket vagy a p-értéket.
(Mit jelöl az r és a c?... ☺)

A szabadsági fok: df = (3 – 1)(4 – 1) = 2 · 3 = 6
p = 0,89 > α, H_0-t megtartjuk, a szemszín és a testtömeg függetlenek egymástól.

Alternatív lehetőség függetlenségvizsgálatra: a ***Fisher-féle egzakt próba***

H_0: a két változó független egymástól

H_1: a két változó nem független egymástól *(…!)*

- kis minta esetén is használható (amikor a függetlenségvizsgálatra vonatkozó X^2-próba a feltételek sérülése miatt már nagyon bizonytalan eredményt adna)
- 2x2-es esetben lehetséges egyoldali ellenhipotézissel is alkalmazni
- csak számítógéppel

X^2-próba homogenitásvizsgálatra

Kérdés: elfogadhatjuk-e, hogy a vizsgált minták azonos eloszlású alapsokaságokból származnak?

H_0: a minták azonos eloszlású alapsokaságokból származnak

H_1: a minták nem azonos eloszlású alapsokaságokból származnak

Technikailag: a függetlenségvizsgálattal megegyezik!

a minták azonos eloszlású alapsokaságokból származnak	↔	a kérdéses (vizsgált eloszlású) változó és a minta eredete függetlenek egymástól

Mindhárom fenti próbánál a feltételek között szerepelt, hogy a megfigyelt, illetve a várt gyakoriságok ne legyenek túl kicsiek (5-nél kisebbek). Előfordul azonban, hogy mégis ilyen adataink vannak. Ez lehet az alacsony mintaelemszám miatt (próbáljunk – időnk, pénzünk, energiánk korlátai között maradva – nagyobb mintákkal dolgozni!), illetve az osztályok kialakítása is történhet szerencsétlenül.

Ilyenkor gyakran összevonunk osztályokat – ezzel persze információt is veszítünk –, és így biztosítjuk a számítások megbízható voltát.

Megjegyzés: Ahogy azt látjuk, a X^2-próbák esetében gyakoriságokkal dolgozunk. Ezt úgy tudjuk megtenni, ha a változó(i)nk értékeiből kategóriákat alakítunk ki – ha már eleve nem ilyen jellegű adataink voltak. A kategóriák kialakításánál – amennyiben ezt nekünk kell megtennünk –, kisebb részben figyelembe vehetjük az adatok eloszlását (leíró statisztika, hisztogram segítségével), azonban mindig szerencsésebb – egyértelműbb és megbízhatóbb is –, ha a csoportok, kategóriák kialakításánál ***szakmai szempontokat*** veszünk figyelembe, ***szakirodalmi források***ra támaszkodunk!

Hol tartunk?

A statisztika feladatai
– adatgyűjtés
az adatok bemutatása, szemléltetése (statisztikák, ábrák) ✓

Megismertünk néhány statisztikai mutatót (várható érték, variancia, szórás, medián, kvartilisek), és az adatok szemléltetésére használt ábrák közül párat (hisztogram, boxplot, …).

– becslések készítése az adatok, tehát a minta alapján (pontbecslés, intervallumbecslés), ✓

Pontbecslések, konfidenciaintervallum várható értékre (és varianciára).

– hipotézisvizsgálat
tipikus esetei:
→ megcáfoljuk, hogy két (vagy több) mennyiség átlaga (…) egyenlő ✓
→ megcáfoljuk, hogy két mennyiség független ✓

Megismertük a hipotézisvizsgálat alapjait, felépítését, a döntési algoritmusokat, és a számtalan hipotézisvizsgálat közül néhány, gyakran előforduló módszerről részletesebben is beszéltünk. Még néhányat megemlítettünk, sok százról nem esett szó.

– asszociáció, korreláció, regresszió (változók közötti kapcsolat kimutatására, jellemzésére)

Na, ezekről még egyáltalán nem esett szó!