RANGSOROLÁSON ALAPULÓ NEM-PARAMÉTERES PRÓBÁK

Sorrendbe állítjuk a vizsgált értékeket (a mintaelemeket) és az aktuális érték helyett a ***rangszámokat*** használjuk a próbastatisztikák értékeinek kiszámítására. Egyes próbáknál a két vagy több mintából származó értékeket összevonjuk és az ***egész*** mintát egy közös sorba rendezzük, majd hozzárendeljük a mintaelemekhez a rangszámokat (pl. Mann-Whitney és Kruskal-Wallis próbák); más próbáknál a két mintát külön-külön rangsoroljuk és mindkettőhöz külön-külön rendelünk rangokat (pl. rangkorrelációs módszereknél).

Általában a nem-paraméteres próbákat akkor használjuk, amikor a paraméteres próbák feltételei nem teljesülnek:

- Ha nem igaz, hogy az alapsokaság normális eloszlású. A nem-paraméteres-próbákat lehet ***eloszlás-független*****nek** is nevezni, mert nincs kikötés erre vonatkozóan.
- Ha az értékek eleve csak ordinális skálán mértek (pl. szubjektív rangsorolás: melyik állat barátságosabb).

A rangsorolásos próbáknál mindenképpen áttérünk ordinális skálára, vagyis információt vesztünk, ha az eredeti adatok arány- vagy intervallum-skálán voltak megadva. Ebből következik ezeknek a módszereknek a gyengesége: csak a nagy eltéréseket tudják kimutatni, a kisebb eltérések esetén a próba eredménye a H_0 megtartása lesz, vagyis nő az II. tipusú hiba valószínűsége. (Mindazonáltal könnyebbség is, hogy csak ordinális skálán kell tudnunk mérni a változókat, pl. érkezési sorrendnél nem szükséges a pontos érkezési idő, elég a sorrend megadása a statisztika elvégzéséhez.)

Kapcsolt rangok: ha többször szerepel ugyanaz az érték, akkor ugyanazt a rangot kell nekik adni, mindegyiknek azt az átlagos rangot, ami a sorszámaik átlaga lenne:

Pl. az 5. és 6. érték ua.: $r_5 = r_6 = \frac{5+6}{2} = 5{,}5$, a 7. mintaérték a 7-es rangszámot kapja.

Pl. az első 3 ua.: $r_1 = r_2 = r_3 = \frac{1+2+3}{3} = 2$, a negyedik a 4-es rangot kapja.

Kapcsolt rangok esetén korrekciós tényezőként szerepel: $E_i = e_i^3 - e_i$, ahol e_i azt adja meg, hogy hány érték egyezik meg az *i.* csoportban.

Mann-Whitney teszt

A kétmintás t-próba illetve d-próba helyett alkalmazható.
Nullhipotézisünk az, hogy a két minta ugyanabból az alapsokaságból származik.
A szignifikancia próbák alapelve, hogy ha a két minta azonos alapsokaságból származik, akkor a rangszámok véletlenszerűen oszlanak meg a minták közt. Ekkor a véletlen csak nagyon ritkán produkál pl. olyan szélsőséges megoszlást, hogy az egyik minta minden eleme kisebb a másik mintáénál:
pl.

1. minta:	**15**	**16**	**17**	**18**				
2. minta:					*20*	*21*	*22*	*23*
rangszámok:	**1**	**2**	**3**	**4**	*5*	*6*	*7*	*8*

Nyilvánvalóan látszik, hogy itt el kell vetni a rangsorszámok véletlen megoszlásának feltételezését, a megoszlást a „kezelés”-nek tudjuk be.

Általánosságban a következő próbastatisztikát kell kiszámítani:

Legyen n_1 a kisebbik minta elemszáma, n_2 a nagyobbiké (tehát a két minta lehet eltérő méretű, mint a kétmintás t-próbánál is, csak alkalmasan kell elnevezni, tehát a kisebbik legyen az 1. minta).
A két mintát összevonva az értékeket helyettesítjük a rangszámokkal.

Az egyik minta minden elemére (célszerű a kisebbre) kiszámoljuk, hogy a másik mintában hány nála kisebb érték van. Ezeket az értékeket összeadjuk, ez lesz a C érték. (Ha az egyik minta elemének rangja ugyanannyi, mint a másik minta egy eleméé, akkor ½-el számolunk.)
A próbastatisztika értéke, U_s a C illetve az n_1n_2-C értékek közül a nagyobb.

$$U_s = \max\{C, n_1n_2 - C\}$$

Az $U_{krit(n_1,n_2)}$ értékeket a Mann-Whitney U-táblázatból nézzük ki: ha kétoldalú próbát végzünk, akkor az $\alpha/2$ sorból, mivel a táblázat egyoldalú. A táblázat szélein a mintaelemszámok szerepelnek, nem pedig a szabadságfokok!

A következő két példa 2 szélsőséges esetet mutat. Minkét esetben α=0,05.
1. példa: a két minta nagyon egyezik:

1. minta:		**13**		**16**		**24**		**29**		**34**
2. minta:	*9*		*15*		*20*		*27*		*33*		*37*
rangsz.:	*1*	**2**	*3*	**4**	*5*	**6**	*7*	**8**	*9*	**10**	*11*

n_1=5, n_2=6, R_1=30, C =1+2+3+4+5=15 n_1n_2 -C=30-15=15
U_s=15 $U_{krit,5,6,\ 0,025}$=

$U_s < U_{krit}$, így elfogadjuk H_0-t.

2. példa: a két minta nagyon különbözik:

1. minta:						**27**	**29**	**33**	**34**	**37**
2. minta:	*9*	*13*	*15*	*16*	*20*	*24*				
rangsz.:	*1*	*2*	*3*	*4*	*5*	*6*	**7**	**8**	**9**	**10**	**11**

n_1=5, n_2=6, R_1=45, C =30 n_1n_2 -C=30-30=0
U_s=30 $U_{krit,5,6,\ 0,025}$=

$U_s > U_{krit}$, elvetjük H_0-t.

A Mann-Whitney teszt U táblázata csak akkor használható, ha a nagyobb elemszám nem nagyobb 20-nál ($n_2 \le 20$). Amennyiben ennél nagyobb minta-elemszámú mintákat akarunk összehasonlítani, úgy egy közelítést kell alkalmaznunk.
Az U_s eloszlása nagy n-re közelíti a normális eloszlást, $\mu = n_1n_2/2$ várható értékkel, és $\sigma = \sqrt{n_1n_2(n_1+n_2+1)/12}$ szórással.

Ekkor kiszámoljuk a következőt: $Z = \frac{U_s - \mu}{\sigma} = \frac{U_s - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$

Ezt a Z értéket kell összehasonlítani a standard normális eloszlás megfelelő szignifikancia szintű értékével.

Megjegyzés: ha a minták **között** kapcsolt rangok vannak (egyező sorszámok, ld. a példában), akkor $n_2 \leq 20$ esetén használható az U_{krit} táblázat, ám $n_2>20$ esetén a fenti Z értéket korrigálni kell.

Egyező sorszámok esetén a közelítő formula alakja egy kicsit más. Minden egyezést tartalmazó csoportra legyen e_i az egyezések száma. Számoljuk ki a $(e_i^3-e_i)$ értéket és adjuk össze őket. Ez esetben a közelítő normális eloszlás szórása:

$$\sigma = \sqrt{\frac{n_1 n_2}{(n_1+n_2)^2 - (n_1+n_2)} \cdot \frac{(n_1+n_2)^3 - (n_1+n_2) - E}{12}} \quad \text{ahol} \quad E = \sum_i E_i = \sum_i \left(e_i^3 - e_i\right)$$

A Kruskal-Wallis próba

Ez a próba az egyszempontos varianciaanalízis nem-paraméteres megfelelője. 3 vagy több minta összehasonlítása esetén alkalmazzuk.
H_0: a minták azonos alapsokaságból származnak.
H_1: legalább egy minta különböző alapsokaságból származik.

Feltételek: a mintaelemek egymástól független, random kiválasztása. NEM feltétel a normális eloszlású alapsokaság.

A próba elve hasonló a Mann-Whitney próbáéhoz: Az összesített mintában kiosztjuk a rangszámokat. Ha a minták ugyanabból az alapsokaságból származnak, akkor a rangszámok eloszlása véletlenszerű lesz az egyes minták között.

h: a minták száma
n_j: a *j*-ik minta elemszáma
R_j: a *j*-ik minta rangszámösszege
$N = \sum_{j=1}^{h} n_j$ az adatok száma= összelemszám

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{h} \frac{R_j^2}{n_j} \right) - 3(N+1) \qquad \text{ha nincsenek egyező (kapcsolt) rangok!}$$

Ez a próbastatisztika *h*-1 szabadságfokú χ^2-eloszlást követ, 5-nél nagyobb mintaelemszámokra jó közelítéssel. (A képlet átalakítható olyan formára, amely hasonlít a χ^2-próbánál megszokotthoz: egy-egy mintára vonatkozó várható és tapasztalt rangösszeg különbségének négyzetét súlyozva összegzi.)

Ha $H > \chi^2_{krit(h-1,\alpha)}$, akkor H_0-t elvetjük, mert a rangösszegek túlságosan eltérnek a várttól, nem valószínű, hogy a rangok pusztán a véletlen miatt oszlanának el ennyire egyenlőtlenül.
Ha $H < \chi^2_{krit(h-1,\alpha)}$, akkor H_0-t megtartjuk, a rangösszegek eltérése a várttól nem túl nagy, betudható a sztochasztikus ingadozásnak.

Ha szignifikáns eltérést találtunk, a variancia anlízishez hasonlóan itt is tovább vizsgálhatjuk különböző összehasonlításokkal, hogy mely minták különbsége okozta ezt.

Korlátok (a Kruskal-Wallis-próbánál):
1) Ha kapcsolt rangok is vannak, akkor H értékét korrigálni kell:
legyen *k* a kapcsolt rangú csoportok száma:

$$H_{korr} = \frac{H}{1 - \frac{\sum_{i=1}^{k} E_i}{N^3 - N}}$$

2) Általában legyen a mintaelemszám 5-nél több minden csoportban. Általában ezt könnyű elérni, ha mégsem teljesül ez a feltétel, akkor a Kruskal-Wallis kritikus-érték táblázatot kell használni a χ^2 táblázat helyett, amely 3 minta összehasonlítására alkalmas. Ezt nem részletezzük.

Két változó kapcsolatának mérése: rangkorrelációs módszerek

Az eddig tanult Pearson-féle korreláció-számítás (r) nem használható,

- ha az adatok nem normális eloszlásúak, illetve
- ha csak ordinális skálán mértek.
  - Pl. pszichológiai kísérlet: állítsuk sorba az objektumokat, hogy mennyire hasonlítanak X-re. Két egyén által adott rangsorolás egyezését rangkorrelációval mérhetjük.
  - Pl. 6 növényen a csírázási sorrend és virágzási sorrend kapcsolata.

n mintaelemen két-két változó: *x* és *y* értékeit mérjük.

Alapelv: külön-külön rangsoroljuk az *x* és *y* értékeket. Figyelni kell az összetartozó adatpárokat, mert az összehasonlítás az alapján történik, hogy az összetartozó rangértékek mennyire különböznek.

Felírjuk az egyik változónak (*x*) megfelelő rangokat növekvő sorrendben, majd mindegyik alá írjuk a másik változóból (*y*) ugyanarra a mintaelemre kapott rangot (ezek már nem lesznek monoton növekvő sorrendben, kivéve a maximális pozitív korreláció esetét).

Pl. *x*: csírázási sorrend: 1 2 3 4 5 6 (r_{ix} értékek)
y: virágzási sorrend: 1 2 3 4 5 6 (r_{iy} értékek)

ez a példa a maximális pozitív korrelációt mutatja, elvárásunk, hogy a rangkorreláció értéke 1 legyen

Pl. x: 1 2 3 4 5 6
y: 6 5 4 3 2 1

ez pedig a maximális negatív korreláció szélsőséges esete lenne, a rangkorrelációnak -1-nek kell lennie.

Ha a két rangsor között nincs kapcsolat, az egyik a másikhoz képest véletlenszerű, illetve nem-szignifikánsan tér el a véletlentől, akkor korrelálatlanok.

A rangkorrelációt ennek megfelelően méri a Spearman-féle $\rho(= r_s)$ és a Kendall-féle τ.

1) Spearman-féle rangkorreláció

$$\rho = r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$ ahol $d_i = r_{ix} - r_{iy}$ a rangok különbsége az i-ik mintaelemre

ha nincsenek kapcsolt rangok, akkor igaz, hogy $-1 \le r_s \le 1$

(kapcsolt rangok esetén bonyolultabb)

Hipotézisvizsgálat:

H_0: r_s értéke nem tér el szignifikánsan a 0-tól, vagyis a két változó valójában független, csupán a véletlen miatt nem 0 a korreláció.
H_1: r_s értéke szignifikánsan eltér a 0-tól, a két változó nem független.

Ha n elég nagy ($n \ge 10$), akkor r_s eloszlása megegyezik r-ével (a lineáris korrelációéval):

$$\hat{t} = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$ df=n-2

2) Kendall-féle τ

Mint fent, x változó szerint növekvő sorrendbe rendezzük a rangokat és alá írjuk a megfelelő y rangot.
Kiszámítjuk a C_i értékeket: az egyik sor i-ik eleme utáni nagyobb rangszámúak száma a második sorban.

$$\tau = \frac{4\sum_{i=1}^{n} C_i - n(n-1)}{n(n-1)}$$

Ha $n > 40$, akkor τ eloszlása közelít a normálishoz, ekkor:

$$t_s = \frac{\tau}{\sqrt{\frac{2(2n-5)}{9n(n-1)}}}$$

Példa: méhanya mérete és utód hossza (parthenogenezis), $n = 15$

r_x:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
r_y:	2	1	13	5	3	4	12	9	10	6	7	8	14	11	15
C_i:	13	13	2	9	10	9	2	4	3	5	4	3	1	1	0

$\sum_i C_i = 79$

τ=(4.79-15.14)/15.14=0,504

t_s=0,504/$\sqrt{2(35)/(9 \cdot 15 \cdot 14)}$ =2,62

t_∞-nél nézzük meg (ez közelítés, persze): $t_{\infty\ 0,05}$=1.96 < t_s, tehát a próba szignifikáns eredményt adott. Pontosabb vizsgálathoz egy speciális táblázat kell.

A két rangkorrelációs koefficiens összehasonlítása:

r_s: nagyobb súlyt ad a „távoleső" rangoknak (d^2), ezért ott célszerű használni, ahol a közeli rangeltérések kevésbé megbízhatóak.

τ: egyenlően súlyozza a rangbeli eltéréseket.