Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs rendszerek Tanszék

Biostatisztika 2.

Dr. Dinya Elek –Dr. Solymosi Róbert: Biometria a klinikumban
Dr. Dinya Elek: Biostatisztika c. művei alapján

Dr. Hullám Gábor

Nemparaméteres eljárások

Ha egy paraméteres statisztikai eljáráshoz kapcsolódó feltételeket nem tudjuk biztosítani, akkor annak megfelelő *nemparaméteres* eljárást válasszuk.

- A nemparaméteres vagy eloszlásmentes (distribution free) tesztek
  - nem igénylik a változók normalitását,
  - nem igénylik a varianciák homogenitását,
  - felteszik, hogy az összehasonlítandó minták eloszlása formája közel azonos.

ezek gyengébb kritériumok mint a normalitás kritériuma

- A nemparaméteres eljárások egyaránt érvényesek nominális, ordinális és intervallum skáláról származó adatokra
- A nemparaméteres tesztek ereje gyengébb mint a neki megfelelő paraméteres teszté, ami a háttérfeltételek hiányából adódik.

Nemparaméteres eljárások

- A minták összehasonlításakor a minták eloszlásának azonosságát teszteljük
- Nem tesztelhetjük a populáció átlagainak azonosságát, mivel az eljárás eloszlásmentes
- Ha feltesszük, hogy a populációk eloszlása szimmetrikus, akkor viszont a teszt az átlagok tesztelésére vezethető vissza
  - Feltétel: szimmetrikus eloszlásnál a medián és az átlag azonos
- A tesztek az adatok növekvő sorrendbe rendezett sorszámait (rangjait) használják
  - Ezért a név: rendstatisztika

Nemparaméteres eljárások

- A minták összehasonlításakor a minták eloszlásának azonosságát teszteljük
- Nem tesztelhetjük a populáció átlagainak azonosságát, mivel az eljárás eloszlásmentes
- Ha feltesszük, hogy a populációk eloszlása szimmetrikus, akkor viszont a teszt az átlagok tesztelésére vezethető vissza
  - Feltétel: szimmetrikus eloszlásnál a medián és az átlag azonos

Nemparaméteres eljárások

- A tesztek az adatok növekvő sorrendbe rendezett sorszámait (rangjait) használják
  - Ezért a név: rendstatisztika

Eredeti adatok	5, 2, 10, 15, 7, 4, 11, 3, 9, 12
Rendezett adatok	2, 3, 4, 5, 7, 9, 10, 11, 12, 15
Rangszámok (r_i)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10

- Azonos értékek kapcsolt rangot kapnak

Eredeti adatok	1, 2, 4, 4, 4, 7, 8, 8
Kiosztott rangok	1, 2, 3, 4, 5, 6, 7, 8
Valós rang (r_i)	1, 2, 4, 4, 4, 6, 7.5, 7.5

$$\frac{3+4+5}{3} = \frac{12}{3} = 4$$

$$\frac{7+8}{2} = \frac{15}{2} = 7.5$$

Rangszámok

- nem mindig egész számok
- nagyon sok azonos érték rontja az alkalmazott próba érzékenységét
- különösen nem számszerű (nominális) változók esetén előnyös a használatuk

R^2

A rangokra vonatkozóan az alábbi műveletek érvényesek:

- a) rangszámok összege

$$R = \sum_{i=1}^{N} r_i = \frac{N(N+1)}{2}$$

- b) rangszámok négyzetösszege

$$R^2 = \sum_{i=1}^{N} r_i^2 = \frac{N(N+1)(2N+1)}{6}$$

- c) rangszámok átlaga és varianciája

$$\bar{x}_R = \frac{N+1}{2}$$

Előjel teszt (sign teszt)

- Páros minták összehasonlításának nemparaméteres eljárása (~ egymintás *t*–teszt)
- Két összetartozó minta különbségének előjelét vesszük (+ , –) és azt elemezzük az alábbi statisztikával:

$$z = \frac{|D| - 1}{\sqrt{N}}$$

- Ahol z standard normális eloszlás alapján meghatározható
- Vagy binomiális eloszlás alapján (ami N↑ esetén tart normálishoz)

$$z = \frac{x - \mu}{\sigma} = \frac{x - Np}{\sqrt{Npq}}$$

- x: + esetek száma
- μ: Np
- Folytonossági korrekció: x+0.5 ha x<Np; x-0.5 x>Np

Wilcoxon signed-ranked teszt

- Párosított mintákra
- Hipotézis:
  - H0: az érték párok közötti különbség egy 0 körüli szimmetrikus eloszlást követ
  - H1: az érték párok közötti különbség nem követ szimmetrikus eloszlást
- Eljárás lépései:
- Párok közötti különbségek és a hozzá tartozó előjel meghatározása:
- $|x_{2,i}-x_{1,i}|$ és $\mathrm{sgn}(x_{2,i}-x_{1,i})$
- A 0 különbségűek elhagyását követően rangsor kialakítása az abszolút különbség alapján (Nr)
- W statisztika számítása: $W = \sum_{i=1}^{Nr}[sgn(x_{2,i} - x_{1,i}) \cdot R_i]$, ahol Ri az i-edik pár rangja

Wilcoxon signed-ranked teszt

- H0 esetén W eloszlására igaz, hogy
- $\mu=0;\ \sigma^2 = \frac{N_r \cdot (N_r+1) \cdot (2N_r+1)}{6}$
- A kritikus érték táblázat alapján meghatározható (W_{kr})
- Ha $|W| > W_{kr}$ akkor H0 elvethető

Mann–Whitney U – teszt

- Alkalmazás: ha a kétmintás *t*-teszt feltételei (a normalitás vagy a varianciák homogenitása) nem teljesülnek
- Lépései:
- 1. A két minta elemeinek összevonása, majd növekvő sorrendbe állítása. Végül minden értékhez a megfelelő rangszám hozzárendelése.
- 2. A csoportok rangszámainak összegének (R1, R2) meghatározása
- Ha a N1≠ N2 akkor legyen N1<N2.
- A kisebb mintához tartozó U-statisztika számítása:

$$U = N_1 \cdot N_2 + \frac{N_1(N_1+1)}{2} - R_1 \qquad \bar{x}_U = \frac{N_1 \cdot N_2}{2} \qquad s_U^2 = \frac{N_1 \cdot N_2 (N_1 + N_2 + 1)}{12}$$

Mann–Whitney U – teszt

- Ha N1 és N2 ≥ 8, akkor az U közelítőleg normális eloszlású, tehát z-score alapján dönthetünk H0: R1=R2 hipotézis tekintetében

$$U = N_1 \cdot N_2 + \frac{N_2(N_2+1)}{2} - R_2$$

- Az U1 és U2 statisztikára teljesül:

$$U_1 + U_2 = N_1 \cdot N_2$$

$$R_1 + R_2 = \frac{(N_1+N_2)(N_1+N_2+1)}{2}$$

Kolmogorov–Szmirnov teszt

- A tesztet a két minta eloszlásának tesztelésére használjuk.
- H0: a két eloszlás azonos a két minta kumulatív eloszlása összehasonlítása alapján:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

- ahol $F_{1,n}$ az első és $F_{2,n'}$ a második minta tapasztalati eloszlása.
- H0 hipotézist α szinten elvetjük, ha

$$\sqrt{\frac{nn'}{n+n'}} D_{n,n'} > K_\alpha.$$

Kruskal–Wallis féle H próba

- k független minta összehasonlítása
- egyszempontos varianciaanalízis nemparaméteres változata
- Mann–Whitney vagy a Wilcoxon rank sum teszt általánosítása
- H0: a k független minta ugyanabból a populációból való
- Lépései:
  - 1) a megfigyelt értékek összevonása, egy mintává egyesítése (N = N1 + N2 + N3 + ...+ Nk)
  - 2) az értékek növekvő sorba állítása
  - 3) rangszámok (ri) és a rangösszegek (Ri) meghatározása
  - 4) H–statisztika számítása a H0 hipotézis tesztelésére, amely k–1 szabadságfokú χ^2 eloszlást követ.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \left(\frac{R_i^2}{N_i} \right) - 3(N+1)$$

Kruskal–Wallis féle H próba

- Ha a rangok között kapcsolt rangok is előfordulnak, akkor korrekcióra van szükség:

$$H_k = \frac{\frac{12}{N(N+1)}\sum_{i=1}^{k}\left(\frac{R_i^2}{N_i}\right) - 3(N+1)}{1 - \frac{\sum_{j}^{N} T_j}{N^3 - N}}$$

  - ahol Tj = t3 – t és t a kapcsolt rangok száma
- A korrekcióval H értéke is nő.
- Szignifikáns eltérés esetén szükségünk lehet a minták páronkénti összehasonlítására, vagyis egy post–hoc teszt eredményére (pl.: Mann-Whitney U tesztet).

Friedman-teszt

- kétszempontos varianciaanalízis nemparaméteres változata
- k számú összetartozó minta vizsgálata

- Sorok és oszlopok sorrendje véletlenszerűen választott
  - Pl.: sorok – betegek, oszlopok – kezelések

- Cél: az oszlopok közötti eltérés vizsgálata
  - 1) rangsorolás soronként, rang meghatározása (ri)
  - 2) oszlopok rangösszegének meghatározása (Ri)
  - 3) statisztika számítása

$$\chi_r^2 = \frac{12}{Nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3N(k+1)$$

Friedman-teszt

$$\chi_r^2 = \frac{12}{Nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3N(k+1)$$

- N a sorok száma
- k az oszlopok száma
- R_i^2 az oszlopok rangösszegének négyzete
- Az eloszlás χ^2 eloszlást követ k–1 szabadságfokkal, ha a sorok és oszlopok száma nem túlságosan kicsi.

A statisztikával a sorok közötti eltérés is ellenőrizhető

- Az így meghatározott χ^2 eloszlás szabadságfoka N–1

$$x_{rk}^2 = \frac{12}{Nk(k+1)} \sum_{j=1}^{N} R_i^2 - 3k(N+1)$$

Kontingencia táblák vizsgálata

- A nominális és ordinális skáláról származó diszkrét változók analízise a korábbiaktól eltérő típusú vizsgálati módszereket igényel
- Ezeket kontingencia (gyakorisági) táblák formájában vizsgáljuk
- Általános formában a kontingencia táblázat mérete *rxk*, és szabadsági foka a $df=(r-1)\cdot(k-1)$.
- A kontingencia táblák méretét a sorok és oszlopok száma határozza meg.

	Készítmény		
Betegség	Kapott	Nem kapott	Total
Van	g_1	g_2	$g_1 + g_2$
Nincs	g_3	g_4	$g_3 + g_4$
Total	$g_1 + g_3$	$g_2 + g_4$	$N = g_1 + g_2 + g_3 + g_4$

Kontingencia táblák vizsgálata

Alkalmazás

- függetlenség vizsgálat,
- homogenitás vizsgálat,
- eloszlás vizsgálat,
- két független binomiális arány vizsgálata irányul.

A khí-négyzet teszt használatának feltételei

- a) a megfigyelt értékek táblázatában bármely cella értéke lehet *0, sőt sorok és oszlopok teljesen 0 értékűek is lehetnek,*
- b) a várható értékek között nem lehet *0 érték,*
- c) a várható értékek táblázatában az olyan cellák száma, ahol az érték *1-5 közötti, nem lehet több, mint az össz cellaszám 25%-a,*
- d) a teszt ereje *N ≥ 30 mintaszámnál a legerősebb, alatta ne használjuk,*

Kontingencia táblák vizsgálata

- Az elemzés során a megfigyelt és a várható gyakoriságok eltérését vizsgáljuk.
- A nullhipotézis szerint nincs eltérés ezen értékek közt,
- Ezt a gyakoriságokból készített Pearson–féle χ^2 statisztikával ellenőrizhető:

$$\chi^2 = \sum_{j=1}^{n} \sum_{i=1}^{k} \frac{\left(g_{ij} - e_{ij}\right)^2}{e_{ij}}$$

- , ahol g_{ij}: az i-edik sor és j-edik oszlopban lévő cella megfigyelési értéke
- e_{ij}: az i-edik sor és j-edik oszlopban lévő cella várható értéke
- Egy cella várható értékét úgy kapjuk meg, hogy a hozzátartozó sor– és oszlopösszeg szorzatát elosztjuk az N mintaszámmal:

$$e_{ij} = \frac{g_{i.} \cdot g_{.j}}{N}$$

Kontingencia táblák – folytonossági korrekció

- Kis minták esetén a χ^2–statisztika eredménye pontosítható, ha folytonossági korrekciót alkalmazunk (Yates – féle korrekció)

$$\chi^2_{korr} = \sum_{j=1}^{n} \sum_{i=1}^{k} \frac{\left(\left|g_{ij} - e_{ij}\right| - 0.5\right)^2}{e_{ij}}$$

Asszociációs mérőszámok

- Ha a két változó nem független egymástól, akkor a közöttük lévő kapcsolat szignifikáns lesz
- Ekkor a két változó közötti kapcsolat erősségének megállapítására ún. szimmetrikus asszociációs mérőszámok használhatók:
- *Kontingencia együttható:*

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

  - Értéke [0, 1] intervallumban van: 0 a függetlenséget, 1 a "tökéletes kapcsolatot" jelenti.

Asszociációs mérőszámok

- *Phi-együttható*

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

- *Csuprov- együttható*

$$T = \sqrt{\frac{\chi^2}{N \cdot \sqrt{(k-1)(n-1)}}}$$

  - Értéke [0, 1] intervallumban helyezkedik el

- *Cramer-együttható:*

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

  - Értéke: $0 \leq V \leq 1$.

Fisher-egzakt és McNemar teszt

- Kis minták esetén használatos a χ^2–statisztika helyett: Fisher-egzakt teszt

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$$

- Nem független minták vizsgálata: McNemar-teszt

		B orvos véleménye +	B orvos véleménye –	
A orvos véleménye	+	a	b	$r_1 = a + b$
	–	c	d	$r_2 = c + d$
		$c_1 = a + c$	$c_2 = b + d$	

$$\chi 2 = \frac{(|b-c|-1)^2}{b+c}$$

ANOVA (**AN**alysis **O**f **VA**riance)

- **Több minta egyidejű összehasonlítására** (vagy egy mintán belüli több csoport összehasonlítására)

Lényege:

- A mintákból számolt összvarianciát két részre osztjuk: **csoporton belüli** (within) és **csoportok közötti** varianciára (between)
- A két részvarianciát hasonlítjuk össze **F–próbá**val
- Attól függően, hogy melyik hatás (csoporton belüli vagy csoportok közötti) a domináns, döntünk a vizsgálat felől.
- Ha a csoportok közötti eltérés jelentős, akkor a csoportok közötti variabilitás lesz a domináns rész a két variancia között és ilyenkor az F–próba szignifikáns eredményt ad.
- A varianciaanalízis után végrehajtott post–hoc tesztek adják meg, hogy mely jellemzők okozzák az eltéréseket

ANOVA - 2

Attól függően, hogy hány szempont (független faktor) szerint csoportosítjuk a vizsgált változót:

- egyszempontos vagy
- többszempontos ANOVA elrendezésekről beszélhetünk.

A t-tesztek az ANOVA eljárás speciális esetéinek tekinthetők:

- a) Párosított t-teszt: ismételt méréses ANOVA két időpontra vonatkoztatva.
- b) Kétmintás t-teszt: egyszempontos ANOVA két csoportra vonatkoztatva.

ANOVA

Példa: négy fajta készítmény (referens és három új készítmény) terápiás hatását vizsgáljuk, akkor az előbbiek értelmében azt vizsgáljuk, hogy az összvariabilitásból milyen jelentőséggel bír az egyes csoportokon belüli egyedi variabilitás, s mennyit jelent a csoportok közötti variabilitás (a tulajdonképpeni gyógyszerhatás).

Ha a kezelések (gyógyszerhatások) közötti eltérés jelentős, akkor a csoportok közötti variabilitás lesz a domináns rész a két variancia között és ilyenkor az F–próba szignifikáns eredményt ad.

Azt, hogy melyik kezelések okozzák az eltéréseket a varianciaanalízis után végrehajtott post–hoc tesztek adják meg.

Egyszempontos ANOVA

- Feltételek:
  - a) legalább 3 diszjunkt csoport
  - b) a vizsgált változó legyen normális (vagy közel normális)
  - c) a csoportok között a variancia legyen homogén (Bartlett-próba, Levene-teszt)
  - d) az esetszám lehetőleg legyen azonos csoportonként (balanced), ekkor lesz a legnagyobb a teszt ereje
- Ha a feltételek nem teljesülnek, akkor a nemparaméteres Kruskal-Wallis tesztet kell alkalmazni
- Hipotézis:
  1)
  - Ho: $\mu_1 \hat{} = \mu_2 \hat{} = \mu_3 \hat{} = \ldots = \mu_n \hat{}$ (átlagok nem térnek el szignifikánsan egymástól, $p \geq 0.05$)
  - H1: $\mu_1 \hat{} \neq \mu_2 \hat{} \neq \mu_3 \hat{} \neq \ldots \neq \mu_n \hat{}$ (átlagok szignifikánsan eltérnek $p<0.05$)

Egyszempontos ANOVA

- 2) Csoportok közötti variancia homogén (F-próba)
  - Ho: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_n^2$
  - H1: $\sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \neq \ldots \neq \sigma_n^2$

- Lineáris egyenlet
  - $y_{i,j} = \mu + \alpha_i + e_{i,j}$
    - y_{ij}: a függő változó értéke
    - μ : a kísérlet főátlaga, fix hatás
    - α_i : fix hatás
    - e_{ij}: hiba, vagy eltérés

Egyszempontos ANOVA

- A mintára vonatkozó teljes variabilitás: az egyes mintaelemeknek a nagy átlagtól való eltérésének négyzetösszege (Total Sum of Squares)

$$SS(\text{teljes}) = \sum_{j=1}^{k} \sum_{i=1}^{N_j} \left(x_{ij} - \bar{x} \right)^2$$

  - *k* index a csoportszámot,
  - N_j a csoport elemszámot jelöli,
  - *xij* a j–edik csoport i–edik eleme

Egyszempontos ANOVA

- a teljes négyzetösszeg (SS^2_T) két részre bontható: egy csoporton belüli (Within–group: SS^2_W) és egy csoportok közötti (Between–group: SS^2_B) négyzetes összegre:

$$SS^2_T = SS^2_W + SS^2_B$$

Source of variaton	SS (Sum of Suares)	df (Degrees of Freedom	MS (Mean Squares)	F	p
Between	SS^2_B	g-1	s^2_B	$F = \frac{s^2_B}{s^2_W}$	
Within	SS^2_W	N-g	s^2_W		
Total	$SS^2_B + SS^2_W$	N-1			

Egyszempontos ANOVA

- a) a j–edik csoport mintaelemeinek összege

$$\sum_{i=1}^{N_j} x_{ij} = S_j$$

- b) a teljes minta összege

$$\sum_{j=1}^{k}\sum_{i=1}^{N_j} x_{ij} = S$$

- c) teljes mintára vonatkozó négyzetes összeg

$$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{N_j} x_{ij}^2 - \frac{S^2}{N}$$

- d) csoportokon belüli négyzetes összeg

$$SS_W = \sum_{j=1}^{k}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}_j\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{N_j} x_{ij}^2 - \sum_{j=1}^{k}\left(\frac{S_j^2}{N_j}\right)$$

- e) csoportok közötti négyzetes összeg

$$SS_B = \sum_{j=1}^{k} N_j \cdot \left(\bar{x}_j - \bar{x}\right)^2 = \sum_{j=1}^{k}\left(\frac{S_j^2}{N_j}\right) - \frac{S^2}{N}$$

Egyszempontos ANOVA

- Arra nem kaptunk választ, vajon milyen csoportok átlagértékei között van eltérés
- Ehhez páronkénti összehasonlítások szükségesek, aminek a száma $k(k-1)/2 \rightarrow m$
- A többszörös összehasonlítás (többszörös hipotézis-tesztelés) azzal a veszéllyel jár, hogy 'megnövekszik' az elsőfajú hiba elkövetési valószínűsége
  - m tesztnél legalább 1 hiba: $1-(1-\alpha)^m$
- Ennek kiküszöbölésére szignifikancia szint korrekciós módszerek szükségesek
  - Bonferroni
  - Benjamini & Hochberg

Kétszempontos ANOVA (ismétlés nélkül)

- A vizsgált paramétert két szempont (faktor) hatásaként értékeljük
- Azt vizsgáljuk, hogy az egyes faktoroknak van–e hatása az értékek alakulására
- A faktorok diszkrét értékkészletűek (vagy nominálisak)
- Táblázatos elrendezésben:
  - az egyik faktor lehetséges értékei határozzák meg a sorok számát: r
  - a másik faktor lehetséges értékei határozzák meg az oszlopok számát: c

Kétszempontos ANOVA (ismétlés nélkül)

	1	2	3	…	c	
1	x_{11}	x_{12}	x_{13}	…	x_{1c}	$\bar{x}_{1\bullet}$
2	x_{21}	x_{22}	x_{23}	…	x_{2c}	$\bar{x}_{2\bullet}$
3	x_{31}	x_{32}	x_{33}	…	x_{3c}	$\bar{x}_{3\bullet}$
…	…	…	…	…	…	…
r	x_{r1}	x_{r2}	x_{r3}	…	x_{rc}	$\bar{x}_{r\bullet}$
	$\bar{x}_{\bullet 1}$	$\bar{x}_{\bullet 2}$	$\bar{x}_{\bullet 3}$	…	$\bar{x}_{\bullet c}$	$\bar{x}_{\bullet\bullet}$

- x_{ij} : i-dik sorban és j-dik oszlopban álló értéket jelöli
- $\bar{x}_{i\bullet}$: i-dik sor átlaga
- $\bar{x}_{\bullet j}$: j-dik oszlop átlaga

Kétszempontos ANOVA (ismétlés nélkül)

- A teljes négyezetösszeg:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{..})^2$$

- Ez három komponensre osztható:
  - sorok szerinti négyzetösszeg
  - oszlop szerinti négyzetösszeg
  - interakciós vagy residuális négyzetösszeg
- A teljes átlagtól való eltérésből kiindulva:

$$(x_{ij}-\bar{x}_{..})=(x_{i.}-\bar{x}_{..})+(x_{.j}-\bar{x}_{..})+(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{..})^2=c\cdot\sum_{i=1}^{r}(\bar{x}_{i.}-\bar{x}_{..})^2+r\cdot\sum_{j=1}^{c}(\bar{x}_{.j}-\bar{x}_{..})^2+\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})^2$$

Kétszempontos ANOVA (ismétlés nélkül)

Legyen:

- ***Si.*** az i–edik sorösszeg,
- ***S.j*** a j–edik oszlopösszeg,
- ***Sij*** az i–edik sorban és j–edik oszlopban álló érték
- ***S*** az N megfigyelés összege

$$\text{Sor szerinti négyzetösszeg} = \frac{1}{c}\cdot\sum_{i=1}^{r} S_{i.}^2 - \frac{S^2}{N}$$

$$\text{Oszlop szerinti négyzetösszeg} = \frac{1}{r}\cdot\sum_{j=1}^{c} S_{.j}^2 - \frac{S^2}{N}$$

$$\text{Interakció} = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 - \frac{1}{c}\cdot\sum_{i=1}^{r} S_{i.}^2 - \frac{1}{r}\cdot\sum_{j=1}^{c} S_{.j}^2 + \frac{S^2}{N}$$

$$\text{Teljes négyzetösszeg} = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 - \frac{S^2}{N}$$

Kétszempontos ANOVA (ismétlés nélkül)

A kétszempontos variancianalízis matematikai modellje:

$$x_{ij} = \mu + \alpha_i + \beta_j + I_{ij} + \varepsilon_{ij} \,, \qquad \text{ahol}$$

- μ : a teljes minta átlagértéke
- α_i : i–edik sorhatás, $\quad \Sigma_i \, \alpha_i = 0$
- β_j : j–edik oszlophatás, $\quad \Sigma_j \, \beta_j = 0$
- I_{ij} : az i–edik sor és j–edik oszlop interakciója (a két faktor közötti interakció, amit 0–nak tételezünk fel)
- ε_{ij} : hibatag (normális eloszlású valószínűségi változó 0 átlaggal az σ^2 varianciával).
- x_{ij}: is normális eloszlású valószínűségi változó μ átlaggal és σ^2 varianciával

Kétszempontos ANOVA (ismétlés nélkül)

Forrás	Négyzetösszeg (SS)	df	Variancia (MS)	F
Sorok	$c \cdot \sum_{i=1}^{r} (\bar{x}_{i.} - \bar{x}_{..})^2 = S_r$	r–1	$s_r^2 = \frac{S_r}{r-1}$	$\frac{s_r^2}{s_{rc}^2}$
Oszlopok	$r \cdot \sum_{j=1}^{c} (\bar{x}_{.j} - \bar{x}_{..})^2 = Sc$	c–1	$s_c^2 = \frac{S_c}{c-1}$	$\frac{s_c^2}{s_{rc}^2}$
Interakció	$\sum_{i=1}^{r}\sum_{j=1}^{c} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2 = S_{rc}$	(r–1)·(c–1)	$s_{rc}^2 = \frac{S_{rc}}{(r-1)(c-1)}$	
Total	$\sum_{i=1}^{r}\sum_{j=1}^{c} (x_{ij} - \bar{x}_{..})^2 = S_t$	r·c–1		

Az analízis során két nullhipotézist tesztelünk:

- H0(1): minden sorátlag egyenlő $\bar{x}_{1.} = \bar{x}_{2.} = \bar{x}_{3.} = \ldots = \bar{x}_{r.}$
- H0(2): minden oszlopátlag egyenlő $\bar{x}_{.1} = \bar{x}_{.2} = \bar{x}_{.3} = \ldots = \bar{x}_{.c}$

Kétszempontos ANOVA ismétléssel

A kétszempontos ismétléses variancianalízis modellje:

$$x_{ijk} = \mu + \alpha_i + \beta_j + I_{ij} + \varepsilon_{ijk}\ , \qquad \text{ahol}$$

- μ : a teljes minta átlagértéke
- α_i : i–edik sorhatás, $\quad \Sigma_i\ \alpha_i = 0$
- β_j : j–edik oszlophatás, $\quad \Sigma_j\ \beta_j = 0$
- I_{ij} : az i–edik sor és j–edik oszlop interakciója (a két faktor közötti interakció, amit 0–nak tételezünk fel)
- x_{ijk}: **a k–adik megfigyelési érték az i–edik sor és j–edik oszlophatásra vonatkozóan**
- ε_{ijk} : x_{ijk} -hoz tartozó hibatag (normális eloszlású valószínűségi változó 0 átlaggal és σ^2 varianciával).

Kétszempontos ANOVA ismétléssel

Az analízis során három nullhipotézist tesztelünk:

- H0(1): minden sorátlag egyenlő $\overline{x}_{1.} = \overline{x}_{2.} = \overline{x}_{3.} = \ldots = \overline{x}_{r.}$
- H0(2): minden oszlopátlag egyenlő $\overline{x}_{.1} = \overline{x}_{.2} = \overline{x}_{.3} = \ldots = \overline{x}_{.c}$
- **H0(3): nincs kereszthatás a faktorok között**: $I_{ij} = 0$

Az ellenőrzést mindig az interakció szignifikanciájával kezdjük !

- Ha nem szignifikáns, akkor a H0(1) és H0(2) hipotéziseket ellenőrizhetjük az adott α szignifikancia érték mellett
- Ha közel szignifikáns, akkor használjuk a korrekciós tényezőt, amivel S_e^2 helyettesíthető

$$s_k^2 = \frac{S_{rc} + S_e}{(r-1)(c-1) + rc(n-1)}$$

Kétszempontos ANOVA ismétléssel

- Ha szignifikáns, akkor az ANOVA mellett további vizsgálatok szükségek, többek közt a faktorok közötti függőségek vizsgálatára
  - Ekkor:

$$H_0^{(1)} \text{ esetén: } F = \frac{s_r^2}{s_{rc}^2} \qquad \text{és} \qquad H_0^{(2)} \text{ esetén: } F = \frac{s_c^2}{s_{rc}^2}$$

Forrás	Négyzetösszeg	df	Variancia (MS)	F
Sorok	$nc\sum_{i=1}^{r}(\bar{x}_{i..} - \bar{x}_{...})^2 = S_r$	r–1	$s_r^2 = \frac{S_r}{r-1}$	$\frac{s_r^2}{s_e^2}$
Oszlopok	$nr\sum_{j=1}^{c}(\bar{x}_{.j.} - \bar{x}_{...})^2 = S_c$	c–1	$s_c^2 = \frac{S_c}{c-1}$	$\frac{s_c^2}{s_e^2}$
Interakció	$n\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2 = S_{rc}$	(r–1)(c–1)	$s_{rc}^2 = \frac{S_{rc}}{(r-1)(c-1)}$	$\frac{s_{rc}^2}{s_e^2}$
Residuál	$\sum_{i=1}^{i}\sum_{j=1}^{j}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij.})^2 = S_e$	rc(n–1)	$s_e^2 = \frac{S_e}{rc(n-1)}$	
Total	$\sum_{i=1}^{i}\sum_{j=1}^{j}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{...})^2 = S_t$	nrc–1		

Kovarianciaanalízis - ANCOVA

- ANOVA + a vizsgált változóra egy másik folytonos változó (pl. életkor) hatást gyakorol
- Ezek az ún. kovariáns változók
- A modellben ezt / ezeket a változó/ka/t figyelembe kell venni

Kétszempontos ANOVA – randomizált blokkok

- A randomizált blokkok analízise kétszempontos ANOVA módszerrel végezhető

Cél: a kezelés hatásának vizsgálata, ami csak akkor ad megbízható eredményt, ha az interakció kicsi

- A blokk elrendezéssel csökkenthető a hiba nagysága, ezáltal a vonatkozó F–próba érzékenyebbé és megbízhatóbbá válik.

Lépései:

- Egyik faktor mentén történő csoportosítás, pl.: méret
- A csoportokon belül, a másik faktor szerinti besorolás random pl.: kezelés besorolás
- Kétszempontos ANOVA alkalmazása

Kétszempontos ANOVA – randomizált blokkok

- Pl.: Gyakori probléma, az életkor befolyásoló hatásának a figyelembe vétele.
- A kort zavaró (confounded) változónak nevezzük, mert nem lehet igazából tudni, hogy egy vizsgálatban egy adott hatásért a ténylegesen vizsgált faktor vagy a kor a felelős.
- Ilyen esetekben a randomizált blokk segít a blokkok közötti eltérés kiszűrésében.

	Kezelés		
	1.	2.	3.
I. blokk	C	B	A
II. blokk	B	A	C

Kétszempontos ANOVA – latin négyzet

- Véletlen blokk, amely két confounding (két hibaforrás) változó hatását akarja kiegyenlíteni
- Az elrendezés tulajdonsága, hogy minden kezelés csak egyszer fordul elő a sorokban és oszlopokban.
- Az ilyen elrendezést kiegyensúlyozott (balanced) elrendezésnek is nevezzük.
- Három vagy magasabb szempontú ANOVA eljárásokra is alkalmazhatjuk (replikációval vagy anélkül)

	I.	II.	III.	IV.
1.	A	B	C	D
2.	B	D	A	C
3.	C	A	D	B
4.	D	C	B	A

MANOVA

- Többváltozós , azaz Mulitvariate ANOVA vizsgálat
- Függő változók korreláltak
- MANOVA-val megválaszolható kérdések:
  - A független változók értékeinek a változása befolyásolják-e szignifikánsan a függő változókat?
  - Milyen kapcsolat áll fenn a függő változók között?
  - Milyen kapcsolat áll fenn a független változók között?
- Alapja két variancia mátrix
  - Modell (által magyarázott) variancia: Σ_{modell}
  - Reziduális variancia (hiba variancia): Σ_{res} és ennek inverze: Σ_{res}^{-1}
  - Legyen A = $\Sigma_{modell} \times \Sigma_{res}^{-1}$
  - H0: $\Sigma_{modell} = \Sigma_{res}$, vagyis A ~ I

MANOVA

Használt statisztikák:

- Wilks lambda
- Pillai-M. S. Bartlett (trace)
- Lawley-Hotelling trace
- Roy's greatest root

Köszönöm a figyelmet!

- (gabor.hullam-*at-mit*.bme.hu)

Budapest University of Technology and Economics
Department of Measurement and Information Systems