Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs rendszerek Tanszék

Biostatisztika 3.

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

Dr. Hullám Gábor

Spearman-féle rangkorreláció

- Lineáris korrelációs együttható speciális esetének tekinthető.
- A kapcsolat szorosságának mérésére a két változó rangszámainak különbségét használjuk fel:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{N} d_i^2}{N^3 - N}$$

  - di = xi – yi az x és y rangjainak különbsége
  - N = a mintaszám
- Az együttható értékei a $-1 \le r_s \le 1$ intervallumba esnek
- Minél közelebb vannak ezek az értékek a –1–hez vagy +1–hez, annál szorosabb a kapcsolat a két változó között.
- rs ≤ 0 estén a két ismérv szerinti rangsor fordított sorrendben van

Spearman-féle rangkorreláció

- Kapcsolt rangok esetén:

$$r_s = \frac{\frac{1}{6}\left(N^3 - N\right) - \left(T_x + T_y\right) - \sum_i d_i^2}{\sqrt{\left[\frac{1}{6}\left(N^3 - N\right) - 2T_x\right]\left[\frac{1}{6}\left(N^3 - N\right) - 2T_y\right]}} \qquad T = \sum_{j=1}^{i} \frac{1}{12}\left(t_j^3 - t_j\right)$$

  - di = xi – yi az x és y rangjainak különbsége
  - N = a mintaszám
  - T – korrekciós tényező, t – a kapcsolt rangok száma
  - j = 1, 2, 3, ..., i az azonos rangszámú csoportok száma
- H0: a korrelációs koefficiens 0, az alábbi t– statisztikával ellenőrizhető :

$$t = r_s \sqrt{\frac{N-2}{1-r_s^2}}$$

Spearman-féle rangkorreláció

$$t = r_s \sqrt{\frac{N-2}{1-r_s^2}}$$

- N–2 szabadságfokú t–eloszlást követ.

- Ha az így kiszámított $t > t_{krit}$ a táblázatbeli kritikus értéknél, akkor az r_s értéke a két változó kapcsolatának a jellemzésére használható
- Ellenkező esetben nincs valós kapcsolat a két változó között

Kendall-féle rangkorreláció

- Két változó kapcsolatát mérő τ együttható a Spearman–féle korrelációs együttható alternatívája.
- A számításhoz az egyes változók rang adatainak természetes sorrendjét vizsgáljuk
- Legyen $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ megfigyelése halmaza X és Y random változó esetében, úgy hogy mindegyik érték egyedi
- Minden (x i , y i) és (x j , y j) megfigyelés pár ,ahol i ≠ j,
- *konkordáns* ha a rangsor mindkét változónál megegyezik, azaz [x i > x j és y i > y j] vagy [x i < x j és y i < y j]
- *diszkordáns*, ha [x i > x j és y i < y j] vagy [x i < x j és y i > y j]
- Ha x i = x j vagy y i = y j akkor a pár se sem *konkordáns, se nem diszkordáns*

Kendall-féle rangkorreláció

- A τ éréke a [–1, +1] intervallumban helyezkedik el:
  - +1 érték jelenti, hogy a rangpárok sorrendje természetes
  - –1 a fordított sorrendet jelenti

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

- Kapcsolt rangok esetén:

$$\tau = \frac{S}{\sqrt{\left[\frac{1}{2}N(N-1) - T_x\right]\left[\frac{1}{2}N(N-1) - T_y\right]}}$$

$$T_x = \frac{1}{2}\sum_i t_i(t_i - 1) \quad \text{és} \quad T_x = \frac{1}{2}\sum_j t_j(t_j - 1)$$

  - Ha egy párban kapcsolt rang szerepel, akkor értékük 0
  - Tx és Ty az X és Y változók kapcsolt rangjainak a számát jelenti
  - S: n_c-n_d , ahol n_c a konkordáns párok száma;
    n_d a diszkordáns párok száma:

Kendall-féle rangkorreláció

- H0: a változók között nincs valós kapcsolat
- Számítandó statisztika:

$$z = \frac{|S| - 1}{\sqrt{\frac{N(N-1)(2N+5)}{18}}}$$

- H0 elvetése standard normális eloszlás adott α szint melletti kritikus értéke szerint

Korreláció és a regresszió kapcsolata

Célkitűzés

Két vagy több változó közötti kapcsolat vizsgálata

Korreláció

- Változók közötti kapcsolat erősségének számszerű kifejezése

Regresszió

- Egy vagy több változó (független változók) milyen hatással van egy kitüntetett változóra (függő változó)
- A változók közötti sztochasztikus kapcsolatban lévő törvényszerűségeket, tendenciát fejezi ki függvények formájában

- erős korreláció → regressziós összefüggés megfelelő jellemzést ad
- gyenge korreláció → regressziós összefüggés korlátozott jellemzést ad

Korreláció

- Két vagy több változó között a kapcsolat erősségének a megállapítása
- Fajtái a változók eloszlásától függően:
  - a) lineáris korreláció: a változók normális eloszlásúak,
    pl.: *Pearson-féle r együttható*
  - b) nemlineáris korreláció: a változók nem normális eloszlásúak,
    pl.: *Spearman-féle* ρ

- A korrelációs együttható értéke *[-1, +1] tartományban van:*
  - *-1 a maximális negatív,*
  - *+1 a maximális pozitív korrelációs kapcsolatot,*
  - *0 közeli érték a korrelálatlanságot (de nem függetlenséget) jelenti a változók között.*

- Általánosan az alábbi hipotéziseket vizsgáljuk:
- H0: **nincs korrelációs kapcsolat** az X és Y változók között (vagy H0: $r = 0$)
- H1: **van korrelációs kapcsolat** az X és X változók között (vagy H1: $r \neq 0$)

Kovariancia

- Két egymástól különböző valószínűségi változó együttes eloszlására jellemző érték, amely megadja a két változó együttmozgását
- A várható értékektől vett eltérések szorzatának várható értékét fejezi ki: $Cov(X,Y) = E[(X-E(X))\,(Y-E(Y))] = E(XY) - E(X)\,E(Y)$
- Kovariancia számítása:

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y)).$$

$cov(X,X) = Var(X)$

- Nem egyenlő (x_i, y_i) előfordulási gyakoriság esetén

$$\operatorname{cov}(X,Y) = \sum_{i=1}^{n} p_i (x_i - E(X))(y_i - E(Y)).$$

- Mintából számítva: $\frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{n-1}$

Kovariancia

- A függvény értékkészlete: $(-\infty , \infty)$.
- Ha a kovariancia pozitív, akkor a két változó átlagosan ugyanabban az irányban tér el a saját átlagától, X növekedésével átlagosan Y is nő,
- Ha a kovariancia negatív az X növekedésével Y csökken

Lineáris korreláció

Pearson–féle korrelációs együttható: r

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

ahol $\bar{x}$ az x_i értékek, $\bar{y}$ pedig az y_i értékek átlaga

Számításának feltételei:

- a) Az X és Y változók legyenek normális eloszlásúak
- b) Az összes kovariancia legyen lineáris
- c) Az X és Y értékeket egymástól függetlenül mérjük

Megjegyzés:

- A kiugró (outliers) értékek erőteljesen befolyásolják *r* értékét.

Lineáris korreláció

- a) A nevezőben a két változó szórásának szorzata szolgál standardizáló tényezőként: így *r értéke standardizált lesz (és összevethető)*
- b) Akkor értelmes, ha X és Y kapcsolata az adott tartományon belül jó közelítéssel lineáris.
- c)Ha más természetű a kapcsolat, a korrelációs formula akkor is csak a lineáris komponensét méri.
- d) Ha r = 0, (illetve ha r nem különbözik szignifikánsan a 0-tól) akkor korrelálatlanságról beszélünk
  (Nem függetlenségről !)
- e) A korrelációs értékeket $r \geq 0.7$ felett mondjuk erős kapcsolatnak

Korrelációs együttható szignifikanciája

- X és Y változók összes populációbeli N számú mintáját, akkor az így kapott sokaságot kétváltozós sokaságnak nevezzük, amelyről feltételezzük a kétváltozós normális eloszlást.
- E kétdimenziós normális eloszlás korrelációját az elméleti korrelációs együttható méri, amit ρ – val jelölünk. Értékkészlete a [–1, 1] intervallum
- A mintából meghatározott r ennek az elméleti korrelációs együtthatónak a becslése
- Az r eloszlása nem szimmetrikus, a ρ – t a –1, 0, +1 értékek kivételével csak jól közelíti. A végpontok miatt ferde eloszlás, ami $\rho = 0$ estén válik szimmetrikussá.

Korrelációs együttható szignifikanciája

- Az r szignifikancia értékének ellenőrzése N – 2 szabadságfokú t–statisztikával

$$t = r \cdot \sqrt{\frac{N-2}{1-r^2}}$$

- Szignifikáns eltérés esetén a H0: $\rho = 0$ hipotézist elvetjük és az r értékét valós kapcsolatnak minősítjük
- Döntés a t értéke alapján:
  - Ha $t<t_{krit}$, akkor H0-t elfogadjuk, vagyis az r érték nem különbözik szignifikánsan a 0-tól.
  - Ha $t>t_{krit}$, akkor H0-t elvetjük az adott szignifikanciaszinten. Ez esetben r olyan mértékben különbözik 0-tól, amit az adott mintaelemszám mellett a mintavételi hiba már ritkán okoz.

Korrelációs együttható szignifikanciája

- A $\rho \neq 0$ vagy $\rho = \rho_1$ hipotézisek tesztelésénél az r eloszlása aszimmetrikus, de az ún. Fisher–féle Z transzformációval normális eloszlást kapunk

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad \mu_z = \frac{1}{2}\ln\left(\frac{1+\rho_1}{1-\rho_2}\right) \qquad \sigma_z = \frac{1}{\sqrt{N-3}}$$

$$z_A = z - \frac{1.96}{\sqrt{N-3}}$$

$$z_F = z + \frac{1.96}{\sqrt{N-3}}$$

Korrelációs együttható szignifikanciája

- Inverztranszformációval visszakapjuk a korrelációs együttható konfidencia intervallumának alsó és a felső korlátját

$$r_F = \frac{e^{2 \cdot Z_F} - 1}{e^{2 \cdot Z_F} + 1} \qquad r_A = \frac{e^{2 \cdot Z_A} - 1}{e^{2 \cdot Z_A} + 1}$$

- A két együttható eltérésének szignifikanciája tesztelhető:

$$z = \frac{z_1 - z_2 - \mu_{z_1} - \mu_{z_2}}{\sigma_{z_1 - z_2}}$$

- μ_{z1}, μ_{z2} : az r1 és r2 együtthatók z eloszlásbeli átlagai
- $\sigma_{z1\text{-}z2}$: az r1 és r2 együtthatók z eloszlásbeli szórásainak különbsége

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

Többszörös korreláció

- Kettőnél több változó kapcsolatrendszerét vizsgáljuk
- Legyen három változónk X1, X2 és X3, a közöttük lévő páronkénti korrelációk r12, r13 és r23
- Ahhoz, hogy ezek tisztán két változó közötti kapcsolat erősségét mutassák ki kell szűrni a többi változó hatását
- Ezt a cél szolgálja a parciális korreláció, ahol a többi változó hatása konstansként kezelt
- Az r12.3 parciális korrelációs együtthatója:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

Többszörös korreláció

- 12.3 index–ben a pont utáni szám jelenti azt a változót, amelynek hatását kiszűrjük.
- Az r12.3 a reziduálok közötti korrelációt jelenti, az X3 hatásának kiszűrése után.
- A parciális korrelációs együttható szignifikanciáját, következő statisztikával ellenőrizhetjük a H0: r12.3 = 0 hipotézis mellett:

$$t = \frac{r_{12.3}}{\sqrt{\frac{1 - r_{12.3}^2}{N - 3}}}$$

- amely df = N – 3 szabadságfokú t–eloszlást követ.

Lineáris regresszió

- Két változó közötti kapcsolat jellegét becsli
- A lineáris regressziós függvény alakja:

$$\hat{y} = a + b \cdot x$$

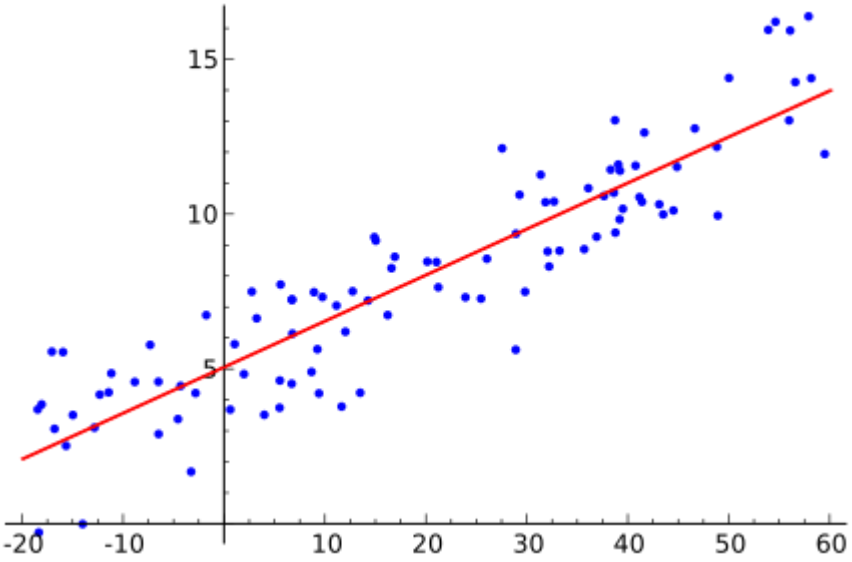
ahol

$\hat{y}$: a függő változó

x: a független változó

a: az y tengely metszete

b: az egyenlet meredeksége (az α szög tangense).



https://commons.wikimedia.org/wiki/File:Linear_regression.svg

Lineáris regresszió

A regressziószámítás feltételei:

- Y változó eloszlása legyen normális
- X változóra hibamentes adatfelvétel
- a minta legyen reprezentatív

Célkitűzés:

- Az egyenes paramétereinek meghatározásakor keressük azokat az a és b értékeket, amelyek mellett a mérési pontokra a regressziós egyenes a legjobban illeszkedik.
- A feladatot a legkisebb négyzetek módszerével végezzük el.

Lineáris regresszió

- Határozzuk meg az egyenlet (a, b) paramétereit, hogy a rezidum értékek eltérésének négyzetösszege minimális legyen:

$$y_D = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \longrightarrow \text{Min.}$$

- Helyettesítsük be az egyenletbe a regressziós függvény általános alakját :

$$y_D = \sum_{i=1}^{N} (y_i - a + bx_i)^2 \longrightarrow \text{Min.}$$

- A feltételnek eleget tevő a és b értékét szélsőérték számítással kapjuk meg:

$$b = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \qquad a = \bar{y} - b\bar{x}$$

Lineáris regresszió

Jelentés

- *A b paraméter jelentése: az X független változó egységnyi változása milyen nagyságú változást okoz az Y függővváltozóban.*
- *Az a értéke a tengelymetszet magasságát adja.*

Eredmény értelmezése

- A regressziós összefüggés szignifikanciáját az ANOVA táblázat alapján vizsgáljuk.
- H0: nincs kapcsolat X és Y változók között
- H1: van kapcsolat X és Y változók között
- Ha az eredmény szignifikáns az adott α érték mellett, akkor fogadhatjuk csak el a valósnak a változók közötti kapcsolatot.

Lineáris regresszió

- Az ANOVA táblázat felépítése

Forrás	SS	df	MS	F	p
Regresszió	$\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = SS_R$	1	$\frac{SS_R}{1} = s_R$	$\frac{s_R}{s_H}$	
Reziduális (hiba)	$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = SS_H$	$N - 2$	$\frac{SS_H}{N-2} = s_H$		
Total	$\sum_{i=1}^{N}(y_i - \bar{y})^2$	$N - 1$			

- Az F_{krit} kritikus értéket az F-táblázatból df = (1, N – 2) szabadságfoknál keressük

Lineáris regresszió

- Ez egy egyoldalú próba ($s_R^2 \geq s_H^2$),
- Ha $F>F_{krit}$ adott α mellett, akkor elvetjük H0-t és a b eltérése a 0-tól szignifikáns,
- Ekkor a lineáris egyenlet predikcióra használható : adott x érték mellett jósolható az y várt értéke
- Az egyenlet használata csak azon a tartományon belül valid, ahol a regressziót végeztük
- Az egyenes alakjától függően lehet: *pozitív irányú regresszió (x és y értéke együtt nő) vagy negatív irányú regresszió (x értéke nő az y értéke csökken).*

Többváltozós lineáris regresszió

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Az alábbi hipotéziseket vizsgáljuk:

- ***H0: nincs kapcsolat az X_i és Y változók között vagy H0: $\beta_i = 0$***
- ***H1: van kapcsolat az X_i és Y változók között vagy H1: $\beta_i \neq 0$***

Az eljárás arra is választ ad, hogy az Xi változók közül melyek az Y szempontjából fontos (releváns) változók, melyek ténylegesen befolyásolják az értékét

A módszer használatának feltétele:

- a) az Xi változók és Y között a kapcsolat lineáris
- b) Xi változók legyenek függetlenek (kollinearitás vizsgálat)

- A független változók között nemcsak folytonos, hanem nominális változók is megengedettek

Többváltozós lineáris regresszió

Multikollinearitás

- Ha az *Xi* változók kapcsolat áll fenn, akkor azokat a változókat ki kell hagyni a további elemzésből.
- Multikollinearitás vizsgálatára a változók korrelációs mátrixának determinánsa is felhasználható: |R| = 0 estén a változók között a kapcsolat maximális, |R| = 1–nél a változók függetlenek.

R^2 és az ún. módosított R^2 (adjusted R^2) számítása

- Jelentése: az *Xi* változók az *Y* varianciájának hány %–át magyarázzák.
- A módosított R^2 érték kisebb, és megbízhatóbb mértéke a regresszió jóságának, mivel ez az érték már mintafüggetlen.

Nemlineáris regresszió

- Olyan esetekben, amikor a függő és független változók között a kapcsolat nem lineáris, az y becslésére a nemlineáris regressziós eljárást alkalmazzuk.
- Segítséget jelent, ha a kapcsolat jellegéről van előzetes információnk pl. polinommal írható le a kapcsolat, ismerjük a polinom fokszámát stb.

A becslő függvénynél törekedni kell arra, hogy

- a) minél kevesebb paramétert tartalmazzon,
- b) jól illeszkedjen a modell
- c) a residuálisok kicsik legyenek.

- ***H0: nincs kapcsolat az X és Y változók között.***
- ***H1: van kapcsolat az X és Y változók között.***