Bevezetés
A paraméteres eljárások alkalmazásához, a célváltozóra nézve szigorú feltételek szükségesek (folytonosság, normalitás, szóráshomogenitás), ekkor a hipotéziseket egy-egy paraméterre (pl. átlag, szórás) fogalmaztuk meg. Ha a feltételek nem teljesülnek, illetve a változók már eleve nominális vagy ordinális szintűek, nem használhatjuk a paraméteres eljárásokat mert nagymértékben torzítanak. Így jöttek létre az ún. nemparaméteres eljárások, amiből sok fajta alakult ki, de nem szükségesek a paraméteres próbáknál előírt megszorítások.

A χ^2- eloszlás

A χ^2-eloszlást a próbastatisztikákban legtöbbször kategorikus adatok elemzésére használjuk, illetve akkor, ha az ordinális, vagy ennél finomabb skálákon nem használjuk fel a változó nagyságrendjére vonatkozó információt.

Ha n darab standard normális eloszlású változót négyzetesen összegzünk, akkor kapjuk a χ^2-eloszlást:

$$\text{Ha: } \eta_1, \eta_2, \eta_3, ..., \eta_n \in N(0,1)$$

Akkor kapjuk a Chi-eloszlást:

$$\chi_n = \eta_1 + \eta_2 + \eta_3 + ... + \eta_n$$

Ha négyzetesen összegzünk, akkor a Chi-négyzet eloszlást kapjuk:

$$\chi_n^2 = \eta_1^2 + \eta_2^2 + \eta_3^2 + ... + \eta_n^2$$

Vagyis az n szabadsági fokú χ^2-eloszlás nem más mint n darab független standard normál eloszlás négyzetösszege.

A χ^2- statisztika

Nullhipotézise általában az, hogy két vagy több nominális változó eloszlása azonos.

$H_0 : F \equiv H$
$H_1 : F \neq H$

Ha a nominális változónak K-darab különböző értéke fordulhat elő, akkor a Chi-négyzet statisztika általános alakja a következő:

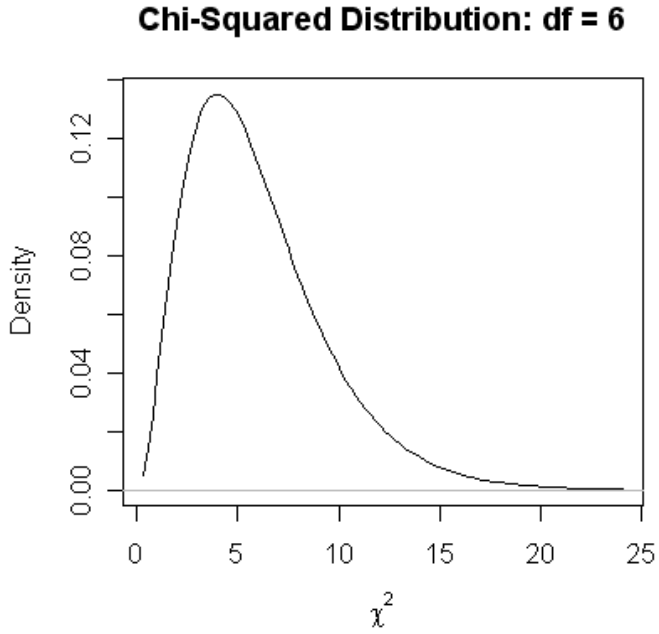
tap = tapasztalt, mért gyakoriság
várt = illeszkedés esetén elvárt, elméleti gyakoriság

$$\chi^2 = \sum \frac{(tap - várt)^2}{várt}$$

n_i: az i-edik cellában tapasztalt gyakoriság
N: elemszám
p_i: az i-edik cellában elvárt valószínűség

$$\sum_{i=1}^{K} \frac{(n_i - Np_i)^2}{Np_i} \in \chi^2_{(K-1),\alpha}$$

A próbastatisztikát természetesen α szignifikancia-szinthez tartozó kritikus érték mellett értelmezzük (táblázati érték). Ha a kiszámított próbastatisztika-érték ennél nagyobb elvetjük a nullhipotézist. Számítógépes alkalmazásoknál általában nem a táblázati F_{krit}-értéket kapjuk (mivel a számítógép nem tudja, hogy mi milyen szigorú szignifikancia szint mellett döntünk majd később), hanem a p-szignifikancia szintet határozza meg. Ha a p-érték 0,05-nél kisebb, akkor elvetjük a H_0-t, egyébként megtartjuk.



Illeszkedésvizsgálat χ^2-próbával

Illeszkedésvizsgálatnál az egyik változó egy elméleti eloszlás, a másik pedig a mért gyakorisági adatok.

H_0: a tapasztalati és a hipotetikus eloszlás megegyezik
H_1: a tapasztalati és a hipotetikus eloszlás nem egyezik meg

Azaz:
$H_0 : F \equiv H$
$H_1 : F \neq H$

Egy telefonos lelkisegély szolgálatnál egy egyhetes időintervallum során következő módon alakul a napi telefonhívások száma: H:29, K:35, Sze:31, Cs:39, P:47, Szo:62, V:51
A gyakorlat szerint a lelki segítők száma a hét első négy napján 1-1, az utolsó három napon 2-2 fő.

Kérdés: A gyakorlat összhangban van-e azzal az elvárással, hogy a kollegák munkaterhelését egyenletesen osszuk el?

	Napok:	n_i	p_i	$N*p_i$	$n_i - N*p_i$	$\frac{(n_i - Np_i)^2}{Np_i}$
1	Hétfő	29	0,1	29,4	-0,4	0,005
2	Kedd	35	0,1	29,4	5,6	1,066
3	Szerda	31	0,1	29,4	1,6	0,087
4	Csütörtök	39	0,1	29,4	9,6	3,134
5	Péntek	47	0,2	58,8	-11,8	2,368
6	Szombat	62	0,2	58,8	3,2	0,174
7	Vasárnap	51	0,2	58,8	-7,8	1,034
	Σ	294	1,0	294		**7,868**

Kézi számolás, és chi-négyzet eloszlási táblázat használata esetén; a df=6, és α=0,05 szignifikancia-szinthez tartozó kritikus érték: 12,592 , így a kiszámolt próbastatisztika értéke (7,87) még belefér az elfogadási tartományba. Vagyis helyes az a gyakorlat miszerint duplázni kell az utolsó három napon a szolgálatot teljesítők létszámát.

Az illeszkedésvizsgálat futtatása R-ben (lelkisegély szolgálat):

```
gyak <- c(29,35,31,39,47,62,51)
prob <- c(1,1,1,1,2,2,2)
chisq.test(gyak,p=prob/10)

Vagy általánosabban:
chisq.test(gyak,p=prob,rescale.p=TRUE)
```

Eredmény:

```
        Chi-squared test for given probabilities

data:  gyak
X-squared = 7.8707, df = 6, p-value = 0.2477
```

A kézi számolással szinkronban, (α=0,05 mellett) itt sem utasítjuk el a H_0-t.

Feladat:
1. Egy pénzérme szabályosságát vizsgáljuk: feldobjuk 100-szor és 60 esetben FEJ lett az eredmény. Szabályos-e a pénzérme.
2. Dobókocka szabályosságát vizsgálva az alábbi dobások születtek: **1-es:**15 , **2-es:**22 , **3-as:**17 , **4-es:**28 , **5-ös:**30 , **6-os:**19 . Szabályos-e a dobókockánk?

Két változó kapcsolata

Két változó kapcsolatával eddig csak folytonos esetben találkoztunk. Itt tanultuk a korrelációt és a regressziót mint a lineáris kapcsolat erősségének mérőszámát. Most nominális- és ordinális változók kapcsolatával folytatjuk, amihez be kell vezetni a kontingenciatábla fogalmát.

Mi a kontingencia tábla, és mire jó?

Megfigyelési egységekről több különböző kategorikus változó adatait összegyűjtve ábrázoljuk a változók különböző értékeinek *együttes* előfordulási gyakoriságait.

Az együttes gyakoriságok táblázatos elrendezése a kontingenciatábla. Az elemzés céljaitól függően több formája lehet, két szempontos esetben a táblázat sorai az egyik, oszlopai a másik változó kategóriáit jelentik, a cellákba pedig a megfigyelt, együttes gyakoriságok kerülnek. Előfordul, hogy folytonos változókra is szerkesztünk kontingenciatáblát, ekkor a változók értékeit intervallumokra bontjuk és ezen intervallumok előfordulási gyakoriságait írjuk a megfelelő cellákba (pl. khi-négyzet-próba, illeszkedésvizsgálatnál normalitásvizsgálat esetén).

A kontingenciatábla elemzése lehetőséget ad a változók közötti függőségi viszonyok feltárására is. Kétszempontos kontingenciatáblán általában a khi-négyzet-próba szolgál a változók függetlenségének vizsgálatára. Ha emellett döntünk, akkor a cellagyakoriságok becsülhetők a marginális gyakoriságok szorzatával, osztva a megfigyelések teljes N számával. Ha a függetlenség nullhipotézisét elutasítjuk, asszociációs v. függőségi mérőszámokkal (association measures) jellemezzük a változók közötti kapcsolat erősségét. Ilyen maga a khi-négyzet-statisztika értéke is. Ha ezt N-nel elosztjuk, a phi-négyzet négyzetes kontingenciát (contingency coefficient) kapjuk. Ez - a sorok és az oszlopok számától függően, alkalmas normalizáló tényezővel - 0 és 1 közé tehető. Az így normalizált kontingencia négyzetgyöke a Cramér-féle V, ezt néha a kapcsolat irányát mutató előjellel is ellátják. Az említett mérőszámok szimmetrikusak, a változók sorrendjét, vagyis a kontingenciatábla sorait és oszlopait felcserélve értékük nem változik. Aszimmetrikus függőségi mérőszám pl. a Goodman-Kruskal-féle lambda, amely azt méri, hogy a sorváltozó mennyire határozza meg az oszlopváltozó értékét. KxK típusú táblázatban a változók egybevágóságát vizsgálja a Cohen-κ mutató.

Egyszerűsége és gyakori alkalmazása miatt külön is említendő a két dichotóm (kétértékű) változóból keletkező 2 x 2-es (négymezős) kontingenciatábla. Kevés megfigyelés esetén a khi-négyzet-próba helyett a Fisher-féle egzakt próbát (Fisher's exact test) érdemes választani, mivel az utóbbi sokkal pontosabb.

A kontingenciatáblákon a hipotézistesztelés legtöbbször visszavezethető a halmazelméletből is ismert függetlenségi formulára, vagyis a:

$$P(A \cap B) = P(A)*P(B)$$

összefüggésre,

Ha az A és B eseményhalmazok egymástól függetlenek, akkor a metszethalmaz (együttes előfordulás) várható valószínűsége egyenlő az elemi halmazok valószínűségeinek szorzatával. Kontingenciatáblán pedig, így módosul: a cellánkénti várható valószínűségek egyenlők az adott cellához tartozó marginális valószínűségek szorzataival, ami csak akkor teljesül, ha a sorokban és az oszlopokban levő változók függetlenek egymástól. Ez utóbbi következménye, hogy egy-egy változóra vonatkozó cellagyakoriságok arányai is megmaradnak a függetlenség, azaz H_0 esetén.

Tegyük fel, hogy G-sorból és K-oszlopból áll a kontingenciatáblánk.

Változók:	B_1	B_2	B_3	.....	B_K	Sor-marginálisok:
A_1	O_{11}	O_{12}	O_{13}	.....	O_{1K}	O_{1+}
A_2	O_{21}	O_{22}	O_{23}	.....	O_{2K}	O_{2+}
A_3	O_{31}	O_{32}	O_{33}	.....	O_{3K}	O_{3+}
.....	.....	.....	.....	.....	.....	.....
A_G	O_{G1}	O_{G2}	O_{G3}	.....	O_{GK}	O_{G+}
Oszlop-marginálisok:	O_{+1}	O_{+2}	O_{+3}	.....	O_{+K}	N

Mivel minden i-edik sorban K-darab cellát összegzünk, a sormarginálisok általános alakja a következő:

$$O_{i+} = \sum_{j=1}^{K} O_{ij}$$

Mivel minden j-edik oszlopban G-darab cellát összegzünk, az oszlopmarginálisok általános alakja a következő:

$$O_{+j} = \sum_{i=1}^{G} O_{ij}$$

A teljes elemszám pedig az összes cella elemszámainak összegeként állítható elő:

$$N = \sum_{i=1}^{G} \sum_{j=1}^{K} O_{ij}$$

<u>Függetlenség vizsgálat</u>

Ha a két változó kategorikus - akár nominális, akár ordinális - a függetlenség vizsgálat Khi-négyzet próbára vezet. Ugyanazt az elvet alkalmazzuk, mint az illeszkedés vizsgálatnál, csak kicsit máshogy.

A G-sorból, és K-oszlopból álló kontingenciatáblán a Chi-négyzet statisztika a következőképp alakul:

$$\sum_{i=1}^{G} \sum_{j=1}^{K} \frac{(n_{ij} - Np_{ij})^2}{Np_{ij}} \in \chi^2_{(G-1)(K-1),\alpha}$$

A kézi számolás során célszerű a tapasztalt és várt gyakoriságokra alapozni, mert kevesebb számolást igényel:
$O_{ij}=n_{ij}$= az i-edik sor j-edik cellájában tapasztalt, megfigyelt, mért (Observed) gyakoriság
E_{ij}= az i-edik sor j-edik cellájában függetlenség esetén várt (Expected) gyakoriság

Alapösszefüggések a kontingenciatáblán:

$$E_{ij} = N \cdot p_{ij} \qquad p_{ij} = p_{i+} \cdot p_{+j} \qquad p_{i+} = \frac{O_{i+}}{N} \qquad p_{+j} = \frac{O_{+j}}{N}$$

Ebből következik, hogy: $E_{ij} = N \cdot p_{ij} = N \cdot p_{i+} \cdot p_{+j} = N \cdot \frac{O_{i+}}{N} \cdot \frac{O_{+j}}{N}$

Némi egyszerűsítés után, csak a marginálisokkal kifejezve: $E_{ij} = \frac{O_{i+} \cdot O_{+j}}{N}$

Így a próbastatisztika jóval egyszerűbb alakot ölt:

$$\sum_{i=1}^{G} \sum_{j=1}^{K} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \in \chi^2_{(G-1)(K-1),\alpha}$$

Hipotézisek:
H_0: az oszlpokban levő gyakoriságok függetlenek a soroktól
H_1: az oszlpokban levő gyakoriságok nem függetlenek a soroktól

Ugyanez jelölésekkel felírva:
$H_0 : \forall i, j : O_{ij} = E_{ij}$
$H_1 : \exists i, j : O_{ij} \neq E_{ij}$

Homogenitás vizsgálat

Formailag ugyanúgy történik, mint a függetlenség vizsgálat, csak más az értelmezése. Mindkét esetben azt kérdezzük, az egyik változó eloszlása eltérő-e a másik változó különbözô értékeinél. Vagyis az a kérdés, hogy a sorváltozó és az oszlopváltozó szerinti gyakoriságok függetlenek-e egymástól?

Példa a homogenitásvizsgálatra:
Egy kutatás során az elsőéves egyetemi hallgatók lakáskörülményeit vizsgálták:

		Lakáskörülmények:				
		Kollégium:	Albérlet:	Család:	Egyéb:	Σ
Neme:	Fiú:	**114**	**157**	**97**	**27**	395
	Lány:	**158**	**255**	**146**	**66**	625
	Σ	272	412	243	93	N=1020

Az elemzés futtatása R-ben:

```
Table <- matrix(c(114,157,97,27,158,255,146,66), 2, 4, byrow=TRUE)
rownames(Table) <- c('Fiu', 'Lany')
colnames(Table) <- c('Koli', 'Alberlet', 'Csalad', 'Egyeb')
Table
Test <- chisq.test(Table, correct=FALSE)
Test
```

Az eredmény:

```
      Koli Alberlet Csalad Egyeb
Fiu    114      157     97    27
Lany   158      255    146    66

     Pearson's Chi-squared test

data:  Table
X-squared = 5.0583, df = 3, p-value = 0.1676
```

A p=0,1676-os szignifikancia szint azt jelzi, hogy a két nem képviselőinek lakóhely szerinti eloszlása homogénnek tekinthető.

Példa a függetlenségvizsgálatra:

Feladat: A gyerek későbbi társadalmi státusza összefügghet-e az apa végzettségével?
A-változó: Apa végzettség: 1= alsó, 2=közép, 3=felső
B-változó: Gyerek státusz: 1= alsó, 2=közép, 3=felső

Adatok:

		Gyerek (B):			
		B_1	B_2	B_3	Σ
Apa (A):	A_1	**30**	**50**	**30**	110
	A_2	**60**	**25**	**20**	105
	A_3	**55**	**45**	**90**	190
	Σ	145	120	140	N=405

Az elemzés futtatása R-ben:

```
Table <- matrix(c(30,50,30,60,25,20,55,45,90), 3, 3, byrow=TRUE)
rownames(Table) <- c('A1', 'A2', 'A3')
colnames(Table) <- c('B1', 'B2', 'B3')
Table
Test <- chisq.test(Table, correct=FALSE)
Test
```

A futtatás eredménye:

```
   B1 B2 B3
A1 30 50 30
A2 60 25 20
A3 55 45 90

     Pearson's Chi-squared test

data:  Table
X-squared = 48.8659, df = 4, p-value = 6.227e-10
```

Az eredmény azt mutatja, hogy a gyerek későbbi társadalmi státusza és az apa végzettsége összefügg: p=0,000, azonban a változók közötti kapcsolat irányáról nem kapunk információt.

Ha a függetlenség vizsgálat során azt kapjuk, hogy a két változó független egymástól, akkor a kérdést le is zárhatjuk. Ha azonban nem függetlenek, akkor a kapcsolat mibenlétét, erősségét

kezdhetjük vizsgálni. Erre szolgálnak a különböző **asszociációs mérőszámok,** melyeket az előbbi Apa-Gyerek vizsgálat χ^2-eredményét felhasználva fogunk bevezetni.

A χ^2 -statisztikából származó asszociációs mérőszámok nominális skálán

A χ^2 statisztika a két diszkrét változó függetlenségét teszteli, H_0-esetén függetlenségről (illetve homogenitásról) beszélünk, ilyenkor a próbastatisztika értéke nulla, vagy nullához közeli. A két változó függése esetén a χ^2 statisztika pozitív értéket vesz fel és minél nagyobb ez az érték, annál nagyobb a függés mértéke is. Mivel a statisztika maximális értéke függ az elemszámtól és a szabadsági foktól is, a felhasználó számára értelmezhetőbb, származtatott mérőszámok kerültek kidolgozásra. A transzformációk célja az eredeti χ^2 statisztika értékét "beszorítani" a [0 , 1] tartományba, hogy ezáltal egy korrelációra emlékeztető mérőszámot kapjunk.

A Φ (Phi) együttható

$$\Phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{48.86}{405}} = 0.347$$

A Φ együttható tulajdonságai:
- H_0-esetén nulla az értéke, ez a függetlenség jele
- 2x2-es kontingncia tábla esetén, az együttható maximális értéke 1
- az együttható értéke túllépheti az 1-gyet, ha a táblázat sorainak, vagy oszlopainak száma kettőnél nagyobb.

Kontingencia (Pearson-féle C) együttható

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{48.86}{48.86 + 405}} = 0.328$$

A C együttható tulajdonságai:
- H_0-esetén nulla az értéke, ez a függetlenség jele
- az együttható mindig 0 és 1 között marad, de maximális értéke az 1-gyet sohasem éri el

Cramer féle V együttható

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{48.86}{405 \cdot 2}} = 0.245$$

ahol ***k*** az oszlopok vagy sorok száma közül a kisebbik.

A Cramer-féle V együttható tulajdonságai:
- H_0-esetén nulla az értéke, ez a függetlenség jele
- A V együttható mindig 0 és 1 között marad, maximális értéke elérheti az 1-gyet bármely kontingenciatábla esetén. Ha két oszlopunk vagy sorunk van, akkor értéke azonos a **Φ** együtthatóval, mivel a tört nevezőjében ekkor csak az N-értéke szerepel.

Az asszociációs mérőszámok kiszámítása R-ben a **vcd-csomagból**:

```
Table <- matrix(c(30,50,30,60,25,20,55,45,90), 3, 3, byrow=TRUE)
rownames(Table) <- c('A1', 'A2', 'A3')
colnames(Table) <- c('B1', 'B2', 'B3')
Table

Test <- assocstats(Table)
Test
```

A futtatás eredménye:

```
                    X^2 df   P(> X^2)
Likelihood Ratio 46.744  4 1.7243e-09
Pearson          48.866  4 6.2273e-10

Phi-Coefficient   : 0.347
Contingency Coeff.: 0.328
Cramer's V        : 0.246
```

Fisher-féle egzakt-próba (Fisher's exact test of significance)

Két dichotóm változó közötti kapcsolat erősségét méri.
A függetlenséget teszteli és közvetlenül számítja ki a szignifikancia szintet.

H_0: A sorok és oszlopok függetlensége (homogenitás)
H_1: A függetlenség / homogenitás sérül

Nem érzékeny:
- az eloszlásra, és
- a mintanagyságra sem.

Általában 2*2-es kontingenciatáblán, és kis elemszámnál használjuk, mivel eléggé számolásigényes.

A χ^2-próbát helyettesíti, ha:
- valamelyik cella gyakorisága n<5, illetve
- ha a mintanagyság N<20

A Fisher-próba működési elve:
Közvetlenül számolja a mért gyakoriságok alapján az aránytalanság mértékét, a tapasztaltnál extrémebb értékek bekövetkezésének valószínűségét H_0 igaz volta esetén. A számítás alapja a hipergeometrikus eloszlás. A számítás során, rögzített marginálisok, és függetlenséget feltételező H_0 esetén, a tapasztaltnál szélsőségesebb elemek elméleti valószínűségeit összegezzük, a hipergeometrikus eloszlás minden további tagjára.

Vizsgálat: Igaz-e hogy a lányok depressziósabbak mint a fiúk?

A mérés során az alábbi eredményeket kaptuk:

	Depressziós	Nem depressziós
Lány	7	2
Fiú	5	6

A Fisher-próba kiszámításának menete:
Megkeressük a legkisebb cellagyakoriságot n_{min} (itt: $n_{min} = 2$).
A legkisebb cellagyakoriságot, és a hozzá tartozó átlót lépésenként 1-gyel csökkentve, a másik átlót pedig 1-gyel növelve egyre „erősebb" kereszttáblákat állítunk elő, amíg: $n_{min}=0$.
(Ha eredetileg $n_{min} = 2$, akkor 3 lépésből áll a számítás.)

A mért gyakoriságokat tartalmazó táblából indulunk ki, majd:
- az n_{min} –hez tartozó diagonális elemeit mindig 1-gyel csökkentjük egészen 0-ig, eközben
- a másik átló elemeit 1-gyel növeljük
- kiszámítjuk minden lépésnél a P_i-t
- addig ismételjük a lépéseket amíg n_{min} –hez tartozó cella 0 lesz
- kiszámítjuk a $P=P_0+P_1+P_2+\ldots P_k$ értéket, vagyis az egyes lépésekből származó valószínűségek összegét.

	Y_1	Y_2	**Sormarginálisok:**
X_1	**a=7**	**b=2**	$r_1=a+b$
X_2	**c=5**	**d=6**	$r_2=c+d$
Oszlopmarginálisok:	$s_1=a+c$	$s_2=b+d$	N=a+b+c+d

Az i-edik lépésben a P_i-valószínűség a következőképp alakul:

$$P_i = \frac{r_1!r_2!s_1!s_2!}{N!a!b!c!d!}$$

Vagyis minden lépésnél úgy számítjuk ki a P_i -t ,hogy a marginálisok faktoriálisainak szorzatát elosztjuk a teljes elemszám, és a cellánkénti elemszámok faktoriálisainak szorzatával.
A számítás k+1 lépésből áll:$P=\Sigma P_i = P=P_0+P_1+P_2+...P_k$

Lássuk a fenti adatokkal a számítás menetét:

P_0 Alaphelyzet:

a=7	**b=2**	$r_1=9$
c=5	**d=6**	$r_2=11$
$s_1=12$	$s_2=8$	N=20

$$P_0 = \frac{9! 11! 12! 8!}{20! 7! 2! 5! 6!} = 0{,}132$$

P_1 Első lépés:

a=8	**b=1**	r_1=9
c=4	**d=7**	r_2=11
s_1=12	s_2=8	N=20

$$P_1 = \frac{9! \cdot 11! \cdot 12! \cdot 8!}{20! \cdot 8! \cdot 1! \cdot 4! \cdot 7!} = 0{,}024$$

P_2 Második lépés:

a=9	**b=0**	r_1=9
c=3	**d=8**	r_2=11
s_1=12	s_2=8	N=20

$$P_2 = \frac{9! \cdot 11! \cdot 12! \cdot 8!}{20! \cdot 9! \cdot 0! \cdot 3! \cdot 8!} = 0{,}001$$

Így: $P = \Sigma P_i = P_0 + P_1 + P_2 = 0{,}132 + 0{,}024 + 0{,}001 = 0{,}157$

Azaz: **p = 0,157**

Ez azt jelenti, hogy a nullhipotézist megtartjuk, vagyis a minta alapján nem mondhatjuk azt, hogy a lányok depressziósabbak lennének a fiúkhoz képest.

A Fisher-próba és χ^2-próba futtatása R-ben:

```
Tabla <- matrix(c(7,2,5,6), 2, 2, byrow=TRUE)
rownames(Tabla) <- c('a', 'b')
colnames(Tabla) <- c('x', 'y')
Tabla
#fisher.test(Tabla, a="less")    # egyoldali/alsó szignifikancia szint
#fisher.test(Tabla, a="two")     # kétoldali szignifikancia szint
fisher.test(Tabla, a="greater")# egyoldali/felső szignifikancia szint
chisq.test(Tabla, correct=FALSE)
remove(Tabla)
```

Elvi lehetőség az R-ben, hogy ki lehet számoltatni az alsó egyoldali-, és a kétoldali szignifikancia szintet is, de a gyakorlatban ennek nincs jelentősége.

R-Commanderrel:

Statistics / Contingency tables / Enter and analyze two-way table

Eredmény:

```
     Fisher's Exact Test for Count Data

data:  Tabla
p-value = 0.1569
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.5762681       Inf
sample estimates:
odds ratio
  3.895711


        Pearson's Chi-squared test
```

```
data:  Tabla
X-squared = 2.1549, df = 1, p-value = 0.1421
```

Amint látható a χ^2- statisztika esetén "szignifikánsabb" lett az eredmény, mert a kis, és kiegyensúlyozatlan elemszám miatt torzulás jelentkezik (másodfajú hiba). A torzulás mértéke az elemszámok csökkenésével egyre nagyobb, ilyen esetben valóban csak a Fisher-próba az ami jól használható.

Kappa (Cohen-féle κ) együttható

Nominális változók egybehangzóságára alkalmazható asszociációs mérőszám

Két nominális változó (A és B) egyezését vizsgálja. Ha ugyanazt az eseményrendszert kétfajta kódolással (A-kódolás és B-kódolás) képezzük le, megvizsgálható, hogy a két kódolás különbözik-e, vagy lényegében ugyanaz. A módszert legtöbbször tesztek validitásvizsgálatára, illetve kódolók (ítészek, bírálók) ítéleteinek egybehangzóságának vizsgálatára használjuk.

H_0: a két kategorizáció (kódolás) egymástól független
H_1: a két kategorizáció egybehangzik, a függetlenségtől pozitív irányban tér el

$$H_0 : A \neq B$$
$$H_1 : A \approx B$$

Gyakorlati probléma:
- Van egy drága, hagyományos teszt (A), és egy új olcsó eljárás (B). A két módszer ugyanazt a jelenséget kívánja mérni. El kell dönteni, hogy kiváltható-e az új módszerrel a régi?

Megoldása:
A mérés során ugyanazt a jelenséget (eseménysort) mindkét teszttel megmérjük, majd megvizsgáljuk, hogy a kétféle teszt által adott kétféle kódolás ("diagnózis") mennyire egyezik meg. Egyezés esetén a kétféle kódból előállított kontingenciatáblán, csak a főátlóban lesznek gyakorisági adatok

Feltétel, hogy a kétféle mérésből származó adatok (A és B) ugyanazt a kategória-rendszert adják outputként (pl. Skizofrén, Neurotikus, Egészséges).

		B-teszt		
		S	**N**	**E**
A-teszt	**S**	45	5	6
	N	10	70	3
	E	7	5	56

Láthatjuk, hogy a kétféle mérés nagyjából ugyanazt adja. Nem tökéletes az egybehangzóság, de a főátló igen erős.

A próbastatisztika kizárólag a **kontingenciatábla főátlójában** levő tapasztalt- és a függetlenség esetén várható gyakoriságokra alapoz.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

ahol: $P_o = \sum_{i=1}^{n} p_{ii}$ és $Pe = \sum_{i=1}^{n} p_{i+} \cdot p_{+i}$

vagy gyakoriságokkal:

$$\text{Ha}: \quad E_{ii} = \frac{O_{i+} \cdot O_{+i}}{N}$$

$$\text{Akkor:} \quad \kappa = \frac{\sum_{i=1}^{n} O_{ii} - \sum_{i=1}^{n} E_{ii}}{N - \sum_{i=1}^{n} E_{ii}}$$

A mutató standard hibája pedig (amit kézzel nem érdemes számolni):

$$ASE(\kappa) = \frac{1}{N(N^2 - \sum O_{i+} \cdot O_{+i})^2}\left[N^2 \sum O_{i+} \cdot O_{+i} + (\sum O_{i+} \cdot O_{+i})^2 - N \sum O_{i+} \cdot O_{+i}(O_{i+} + O_{+i})\right]$$

A kappa együttható lényegében azt méri, hogy a függetlenség állapotához képest, mennyire erősödik fel a kereszttáblában a főátló, azaz mennyire vág egybe a két kódoló kategorizációja. A számítógépes alkalmazásoknál egy Z-transzformált próbastatisztikát alkalmaznak a szignifikanciaszint megállapítására (amely H_0 esetén aszimptotikusan standard normál eloszlású):

$$Z = \frac{\kappa}{ASE(\kappa)}$$

A kappa-mutató értelmezése:

0-0,4-ig	gyenge
0,4-0,6	közepes
0,6-0,8	jó
0,8-1	kiváló

A Cohen-kappa kiszámítása kézzel a fenti adatokkal:

Változók:	B_1	B_2	B_3	*Sormarginális:*
A_1	45	5	6	56
A_2	10	70	3	83
A_3	7	5	56	68
Oszlopmarginális:	62	80	65	N=207

Tapasztalt gyakoriság a főátlóban:

$\sum O_{ii} = 45 + 70 + 56 = 171$

Függetlenség esetén várt elméleti gyakoriság a főátlóban:

$$\sum E_{ii} = \frac{56 \cdot 62}{207} + \frac{83 \cdot 80}{207} + \frac{68 \cdot 65}{207} = 16{,}77 + 32{,}07 + 21{,}35 = 70{,}19$$

Így a kappa értéke:

$$\kappa = \frac{171 - 70{,}19}{207 - 70{,}19} = \frac{100{,}81}{136{,}81} = 0{,}7368$$ ami egyébként a „jó” egybehangzóságot jelenti

A Cohen-kappa mutató az R-ben a *vcd-csomagból* érhető el:

```
library(vcd)
tabla<-matrix(c(45,5,6,10,70,3,7,5,56),3,3,byrow=TRUE)
is.matrix(tabla)
ckappa<-Kappa(tabla)
ckappa
```

Az eredmény pedig kissé hiányos:

```
                value        ASE
Unweighted 0.7368365 0.03986458
Weighted   0.7195975 0.09455902
```

Így (a súlyozatlan κ-ra) a kétoldali szignifikancia szint kiszámítása: a „hagyományos” módszerrel történik:

```
cohensig=2*(1-pnorm(0.7368/0.039))
cohensig

Vagy egyszerűbben, a számok begépelése nélkül:
cohensig=2*(1-pnorm(ckappa$Unweighted[1]/ckappa$Unweighted[2]))
cohensig
```

Ennek értéke: p=0.000
Az eredmény azt mutatja, hogy a két teszt javarészt ugyanazt a jelenséget méri, jól egyezik a kétféle eredmény. Ami azt jelenti, hogy az új és olcsóbb (B) eljárással elég jól helyettesíthető a régi (A) módszer.

Lambda (Goodman-Kruskal-féle λ)

Nominális változók predikciós jellegű kapcsolatának vizsgálatára alkalmazható asszociációs mérőszám

A PRE-elv (**P**roportional **R**eduction in predictive **E**rror)

Két változó kapcsolatának vizsgálatára alkalmazott, egyik legrégebbi alapelv az Y-változóban (célváltozó) tapasztalható előrejelzési hiba egy másik X-változó (prediktor) általi csökkentése. A statisztikai próbák zöme erre az alapelvre vezethető vissza. Lényege, hogy a két változóról akkor gondoljuk, hogy összefüggnek (pl. oksági kapcsolat van közöttük), ha a prediktorváltozó értékeinek ismerete lényegesen (szignifikánsan) csökkenti a célváltozó becslésének hibáját. Az eddig ismert paraméteres próbák (pl. Lin.Reg. , ANOVA) összhangban vannak ezzel az elvvel, a Goodnam-Kruskal-féle λ pedig tökéletesen bele is illik a PRE-elv koncepciójába. Ha az X-változóval kapcsolatos a-paraméter szignifikánsan

csökkenti az Y becslési hibáját, ez általában azt jelenti, hogy a két változó összefügg, valamilyen értelemben, pl. az egyik változó (prediktor) befolyásol egy másik változót (célváltozó).

Lineáris regresszió példáján:

$H_0 : Y_i = m + \varepsilon_i$

$H_1 : Y_i = b + a \cdot X_i + \varepsilon_i$

Nullhipotézis esetén, a legjobb becslés a célváltozó (Y) átlaga. Ezzel szemben, akkor fogadjuk el a H_1-et ha az „*a*” meredekségparaméter bevezetése (és az Y_i-hez tartozó X_i értékeinek ismerete) szignifikánsan csökkenti a becslési hibát.

Egyszempontos ANOVA példáján:

$H_0 : Y_i = m + \varepsilon_i$

$H_1 : Y_i = m + a_i + \varepsilon_i$

Nullhipotézis esetén, a legjobb becslés a célváltozó (Y) átlaga. Akkor fogadjuk el a H_1-et ha az „*a*” csoport-paraméter bevezetése (X_i értékeinek ismerete) szignifikánsan csökkenti a becslési hibát.

Ha az alapelvet megértettük, akkor könnyen generalizálhatjuk egyéb, nominális változókra is:

Nominális változók esetén a változó (**B**) legvalószínűbb értékének legjobb előrejelzése, a B-változó módusza, vagyis a leggyakoribb értéke. Ha ez a B-változó és egy másik, nominális A-változó függvénye vagy következménye, akkor az A-változó értéke szerinti B-móduszokból megbízhatóbban lehet következtetni a B-értékekre, azaz csökken a B-re vonatkozó előrejelzési hiba valószínűsége.

Kérdés: Ha ismert a populáció, egy nominális változó szerinti kategorizációja (A), akkor lehet-e következtetbi ugyanezen populáció másik nominális változójára (B).

Másképp: ha ismerem a populáció egyik kategorizációját, akkor ennek ismerete csökkenti-e egy másik kategorizáció becslésének véletlen hibáját?

A B-változó előrejelzési hibája, ha a B-változó móduszával becslünk:

$$P_{[hibaB]}$$

A B-változó előrejelzési hibája, ha ismerjük az A-változó értékei szerinti B-móduszokat:

$$P_{[hibaB|A]}$$

Abszolút hibacsökkenés:

$$P_{[hibaB]} - P_{[hibaB|A]}$$

Arányos hibacsökkenés:

$$PRE_{B|A} = \frac{P_{[hibaB]} - P_{[hibaB|A]}}{P_{[hibaB]}}$$

P_{+m}: a legnagyobb oszlopmarginális valószínűsége (B-módusz)
P_{im}: az i-edik sor legnagyobb elemének valószínűsége (soronkénti B-móduszok)

$$\lambda_{B|A} = \frac{(1 - P_{+m}) - (1 - \sum P_{im})}{1 - P_{+m}} = \frac{\sum P_{im} - P_{+m}}{1 - P_{+m}}$$

Ugyanez a gyakoriságokkal kifejezve:
O_{+m}: a legnagyobb oszlopmarginális
O_{im}: az i-edik sor legnagyobb eleme

$$\lambda_{B|A} = \frac{\sum O_{im} - O_{+m}}{N - O_{+m}}$$

Azt fejezi ki, hogy milyen arányban csökken a B-változó előrejelzési hibája, ha ismerem ugyanezen sokaság A-változóbeli értékét is. A mutató közvetlenül méri az arányos hibacsökkenés mértékét.
Szemléletesebben: a sorváltozó (A) mennyire határozza meg az oszlopváltozó (B) értékét?

A számítógépes alkalmazásoknál Z-transzformált próbastatisztikát alkalmaznak a szignifikanciaszint megállapítására:

$$Z = \frac{\lambda_{B|A}}{ASE(\lambda_{B|A})}$$

A gyakorlatban, a lambda értéke már néhány tized esetén is erős függést jelez

Korábbi példánk kapcsán már megállapítottuk, hogy a gyerek és az apa társadalmi státusza összefüggött (legalábbis a χ^2 – statisztika ezt mutatta), arról viszont nem kaptunk információt, hogy ez a kapcsolat milyen irányú.

Feladat:
Az apa végzettsége befolyásolja-e a gyerek társadalmi státuszát, vagy fordítva?
A-változó: Apa végzettség: 1= alsó, 2=közép, 3=felső
B-változó: Gyerek státusz: 1= alsó, 2=közép, 3=felső

Adatok:

		Gyerek (B):			
		B_1	B_2	B_3	Σ
Apa (A):	A_1	**30**	**50**	**30**	110
	A_2	**60**	**25**	**20**	105
	A_3	**55**	**45**	**90**	190
	Σ	145	120	140	N=405

A gyerekre nézve:

$$\lambda_{B|A} = \frac{(50 + 60 + 90) - 145}{405 - 145} = \frac{200 - 145}{405 - 145} = \frac{55}{260} = 0{,}211$$

Az apára nézve:

$$\lambda_{A|B} = \frac{(50 + 60 + 90) - 190}{405 - 190} = \frac{200 - 190}{405 - 190} = \frac{10}{215} = 0{,}046$$

A kapott eredmények nem mondanak ellent a józanésznek sem, mivel az apa státusza inkább meghatározhatja a gyerek társadalmi helyzetét, mint fordítva.

Megjegyzés:
A Goodman-Kruskal-féle λ-mutató az R-programcsomagban még nincs implementálva.

Asszociációs mérőszámok ordinális változók esetén

Monotonitási együtthatók:
- Goodman-Kruskal féle (gamma)
- Kendall féle τ τ_b τ_c (tau és tau b, c)
- Somers féle D
- Kendall-féle
- Spearman-féle rangkorreláció

A fagylat-fogyasztási preferenciákat vizsgáljuk a csoki és a vaníliafagyalt esetén.
Kérdés: Mennyire szereti Ön a ……fagylaltot?
1. utálom 2. megeszem 3. szeretem

Változók: / Személyek:	**X (csoki)**	**Y (vanília)**
A	1	2
B	2	3
C	2	2
D	3	2
E	1	2

A személyek X és Y változójának elemei között, ha minden elemet összehasonlítunk, összesen:

$$\frac{N(N-1)}{2}$$ darab elempárt lehet képezni

Ez 5 személy esetén 10 darab párt/összehasonlítást fog jelenteni.

Monoton kapcsolat szempontjából a személyek között megkülönböztetünk konkordáns (P) és diszkordáns (Q) , valamint kapcsolt (T) párokat is.

Definíciók:
P: Konkordáns (egyirányú) az olyan pár, amelynél az egyik személy mindkét változójához tartozó skálán magasabban rangsorol, mint a másik személy. Vagyis akkor monoton, ha $X_2 > X_1$ esetén $Y_2 > Y_1$ is mindig fennáll.
AB, BE

Esetünkben: P=2

Q: Diszkordáns (fordított) az olyan pár, amelynél a két személy mindkét változójában ellentétesen rangsorolt. Tehát: $X_2>X_1$ esetén $Y_2<Y_1$ is mindig igaz.
BD
Esetünkben: Q=1

Tx: (csak az X-változóban kapcsolt (azonos) és Y változójában eltérő pár)
BC
Esetünkben: $T_x=1$

Ty: (csak azY-változóban kapcsolt és X változójában eltérő pár)
AC, AD, CD, CE, DE
Esetünkben: $T_y=5$

A monotonitásra vonatkozó mérőszámok nagy hasonlóságot mutatnak, amennyiben a P-Q és P+Q arányát vizsgálják különböző feltételek mellet. Közös bennük az a törekvés, hogy a mutató értékét a [-1 , 1] tartományba szorítsák be.

Goodman-Kruskal féle Γ:

A gamma megmutatja, hogy mennyivel nagyobb a konkordáns párok valószínűsége a diszkordáns párok valószínűségénél.

$$\Gamma = \frac{P - Q}{P + Q}$$

A gamma értéke az előbbi példában:

$$\Gamma = \frac{P - Q}{P + Q} = \frac{2 - 1}{2 + 1} = \frac{1}{3} \approx 0{,}333$$

Kihagyja azokat az eseteket, ahol kapcsolt pár (egyenlőség) van, ezért csak a monoton változópárokkal foglalkozik. Értéke -1 és 1 között mozoghat, függetlenség esetén *nulla* az értéke. A Γ=0 érték azonban csak a 2x2-es táblázat esetén jelent biztosan függetlenséget.

Somers féle D:
Ez aszimmetrikus mérőszám. A $D_{(X|Y)}$ azt kérdezi, Y_1 és Y_2 különbözősége esetén X_1 és X_2 viszonya jelent-e monotonitást. Ebben az értelemben X-t tekinthetjük függő változónak.

$D_{(X|Y)}$: $$D_{(X|Y)} = \frac{P - Q}{P + Q + Tx} = \frac{2 - 1}{2 + 1 + 1} = \frac{1}{4} = 0{,}25$$

Ha Y a függő változó, akkor:

$D_{(Y|X)}$: $$D_{(Y|X)} = \frac{P - Q}{P + Q + Ty} = \frac{2 - 1}{2 + 1 + 5} = \frac{1}{8} = 0{,}125$$

A szimmetrikus változat a két aszimmetrikus D középértéke a képletben látható módon.

Szimmetrikus D: $$D_{(sym)} = \frac{P - Q}{P + Q + \frac{Tx + Ty}{2}} = \frac{2 - 1}{2 + 1 + \frac{1 + 5}{2}} = \frac{1}{6} \approx 0{,}166$$

Kendall féle τ (Tau)
Értéke azt fejezi ki, hogy mennyivel nagyobb a a konkordáns párok valószínűsége a diszkordáns párokéhoz képest, ha az összes lehetséges párt figyelembe vesszük.

$$\tau = \frac{2(P - Q)}{N(N - 1)} = \frac{2(2 - 1)}{5 \cdot (5 - 1)} = \frac{2}{20} = 0{,}1$$

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

Kendall féle τ_b (Tau b)

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + Tx) \cdot (P + Q + Ty)}} = \frac{2 - 1}{\sqrt{(2 + 1 + 1)(2 + 1 + 5)}} = \frac{1}{\sqrt{4 \cdot 8}} \approx 0{,}177$$

Ami egyébként a két aszimmetrikus Somers-féle D mértani közepével egyenlő, azaz:

$$\tau_b = \sqrt{D_{(X|Y)} \cdot D_{(Y|X)}} = \sqrt{0{,}25 \cdot 0{,}125} \approx 0{,}177$$

Értéke csak akkor érheti el a +1-et vagy -1-et, ha a táblázat sorainak és oszlopainak száma egyenlő.

Kendall féle τ_c (Tau c):
Ennek értéke már bármilyen táblázat esetén elérheti a +1-et vagy -1-et.

$$\tau_c \frac{2m(P - Q)}{N^2(m - 1)} = \frac{2 \cdot 2(2 - 1)}{5^2 \cdot 1} = \frac{4}{25} = 0{,}16$$

Az „m” jelentése: a két változó értékkészlete (kereszttáblán: a sorok ill. oszlopok száma) közül a kisebbik (itt: m=2).

Spearman-féle rangkorreláció

Ha két folytonos változó eloszlása különbözik, illetve sérül a normalitási követelmény, akkor a két folytonos változó lineáris kapcsolatára vonatkozó Pearson-féle (paraméteres) lineáris

korrelációs együttható torzított eredményt adhat. Ugyanis a Pearson-féle r-együttható csak intervallum skálán levő normális eloszlású változókra használható.

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

Számunkra sokszor csak a két változó együttváltozása (monotonitása) a fontos: ha az egyik nagyobb akkor a másik nagyobb kisebb vagy változatlan? Ekkor már nem a változók konkrét értékei fontosak, csak az egymáshoz viszonyított helyzetük. Ebből az alapelvből kiindulva születtek meg a rangsoroláson alapuló eljárások. A **rangsorolásos eljárások** lényege (ld. később is), hogy a számítás nem a változók konkrét értékeivel történik, hanem a rendezett mintában elfoglalt sorszámmal (X_i → helyett: Rang(X_i) = rangszám). A rangsorolásos eljárások általában nem érzékenyek a normalitási feltétel sérülésére, és a minták eloszlásának különbözőségére sem. Csak azt igénylik, hogy a változók legalább ordinális típusúak legyenek, ugyanis ez a rendezhetőség a rangszám-konverzó szükséges és elégséges feltétele.

A Spearman-féle rangkorreláció alapelve:
- mindkét mintát rendezzük
- a rendezett minták elemeihez rangszámokat rendelünk
- a rangszámokra számoljuk ki a hagyományos Pearson-féle (paraméteres) korrelációt

Mindkét minta n-elemű
X : $x_1 ... x_n$ melynek r-különböző értéke lehet
Y : $y_1 ... y_n$ melynek s-különböző értéke lehet

Mindkét minta elemeit rangsoroljuk:
X-rangsor: 1,2...r
Y-rangsor: 1,2...s

Az eredeti értékeket a rendezett mintabeli rangszámokkal helyettesítjük (rangszám-konverzió):
x_i → R_i
y_i → S_i

Ezt követően a konvertált rangszámokra alkalmazzuk a Pearson-képletet:

$$r_s = \frac{\sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2 \sum_{i=1}^{n} (S_i - \bar{S})^2}}$$

Az asszociációs eljárások közül az R-ben egyelőre csak a Pearson-félr r, Spearman-féle rho, és a Kendall-féle tau_b érhető el. Az eljárások a **cor.test()** függvénnyel futtathatóak.

A szükséges adatok (fagyipreferencia) bevitele a futtatáshoz:

```
fagyi<-data.frame(X=c(1,2,2,3,1),Y=c(2,3,2,2,2))
attach(fagyi)
```

A "hagyományos" Pearson-féle paraméteres korreláció futtatása R-ben:

```
cor.test(X,Y,method="pearson")
```

Eredmény:

```
        Pearson's product-moment correlation

data:  X and Y
t = 0.2335, df = 3, p-value = 0.8304
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8486951  0.9087566
sample estimates:
      cor
0.1336306
```

A Spearman-féle nemparaméteres korreláció (rho) futtatása R-ben:

```
cor.test(X,Y,method="spearman")
```

Eredmény:

```
Warning in cor.test.default(X, Y, method = "spearman") :
         p-values may be incorrect due to ties

        Spearman's rank correlation rho

data:  X and Y
S = 16, p-value = 0.7833
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.186339
```

A Kendall-féle *tau b* asszociációs együttható kiszámítása R-ben:

```
cor.test(X,Y,method="kendall")
```

Az eredmény, pedig:

```
Warning in cor.test.default(X, Y, method = "kendall") :
         Cannot compute exact p-value with ties

        Kendall's rank correlation tau

data:  X and Y
z = 0.433, p-value = 0.665
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.1767767
```

A Hmisc-csomagban található asszociációs eljárások:

Goodman–Kruskal gamma (Hmisc-csomagból):

```
GKgamma<-rcorr.cens(X, Y, outx=TRUE)
GKgamma
```

```
sig=2*(1-pnorm(GKgamma[2]/GKgamma[3]))
sig
```

A futtatás eredménye:

```
       C Index            Dxy           S.D.              n        missing
     0.6666667      0.3333333      0.5443311      5.0000000      0.0000000
    uncensored Relevant Pairs     Concordant      Uncertain
     5.0000000      6.0000000      4.0000000      0.0000000
Sig
      Dxy
0.5402914
```

Somers' $D_{(x|y)}$ asszociációs együttható (Hmisc-csomagból):

```
rcorr.cens(X, Y, outx=FALSE)
```

Vagy egyszerűbben:

```
DXY<-rcorr.cens(X, Y)
DXY
sigdxy=2*(1-pnorm(DXY[2]/DXY[3]))
sigdxy
```

A futtatás eredménye:

```
       C Index            Dxy           S.D.              n        missing
     0.6250000      0.2500000      0.4145781      5.0000000      0.0000000
    uncensored Relevant Pairs     Concordant      Uncertain
     5.0000000      8.0000000      5.0000000      0.0000000
sigdxy
      Dxy
0.5464936
```

A futtatás a változók cseréjével:

```
DYX<-rcorr.cens(Y, X)
DYX
sigdyx=2*(1-pnorm(DYX[2]/DYX[3]))
sigdyx
```

A futtatás eredménye:

```
       C Index            Dxy           S.D.              n        missing
     0.5625000      0.1250000      0.2359323      5.0000000      0.0000000
    uncensored Relevant Pairs     Concordant      Uncertain
     5.0000000     16.0000000      9.0000000      0.0000000
sigdyx
      Dxy
0.5962416
```

Az eredményekből az tűnik ki, hogy a csoki és vanília fagyi közötti preferencia enyhén összefügg, mintha a vanília szeretete inkább befolyásolná a csoki fagyi iránti preferencia mértékét (Somers-féle: $D_{X|Y} > D_{Y|X}$), ám ez a kapcsolat nem szignifikáns egyik irány esetén sem.