A SOKASÁG TÖBB ISMÉRV SZERINTI VIZSGÁLATA

NÉHÁNY FOGALOMRÓL

- Egy sokaság egységei különféle tulajdonságaik felsorolásával jellemezhetőek. (Pl.: nem, életkor, iskolai végzettség, jövedelem)
- **Részsokaság**: ha a vizsgált sokaság egységeinek valamely nem közös tulajdonságát rögzítjük, akkor mindig egy részsokasághoz jutunk
- **Feltétel nélküli megoszlás**: a fősokaság egységeinek valamely ismérv szerinti megoszlása. (Pl.: az N megfigyelt személyt havi nettó jövedelmük szempontjából vizsgáljuk.)
- **Feltételes megoszlás**: a fősokaságból kijelölt egyes részsokaságok egységeinek ugyanezen ismérv szerinti megoszlása. (Pl.: amikor az elemzést csak az egyetemi diplomával rendelkezőkre terjesztjük ki, tehát azzal a feltétellel vizsgáljuk a jövedelem szerinti megoszlást, hogy a megfigyelt személy egyetemi diplomával rendelkezik)

A SZÓRÓDÁS FOGALMÁRÓL ISMÉT

- Szóródás: az ismérvértékek egymás közötti különbözősége, változékonysága.
  - Általános értelemben: egy ismérv akkor nevezhető szóródónak valamely sokaságon belül, ha annak az adott sokaság egységeinél előforduló változatai nem mind egyformák.
  - Mérőszámai pl.: $RANGE, IRQ, \sigma, s$

A feltétel nélküli megoszlások mindig szóródóak. Ha ugyanis a sokaság minden egységénél azonos lenne az adott változó értéke, akkor nem lenne értelme az adott ismérv szerinti elemzésnek.

A feltételes megoszlások nem szükségképpen szóródóak. Valamilyen alkalmas osztályozással el lehet érni, hogy az egy-egy osztályba – részsokaságba – a vizsgált ismérv szempontjából már csak kevésbé vagy egyáltalán nem szóródó elemek kerüljenek.

A RÉSZSOKASÁGOKON BELÜLI FELTÉTELES MEGOSZLÁSOK ÉS EZEK VISZONYULÁSA A FELTÉTEL NÉLKÜLI MEGOSZLÁSHOZ

Függetlenség

Minden feltételes megoszlás egyforma, s így megegyezik a feltétel nélküli megoszlással.

Példa:

	%
nem vallásos	51,3
vallásos	48,7
összesen	100,0

Csoportképző ismérv: életkor

A fenti eset áll fenn:

	50 évesnél fiatalabb	50 éves vagy idősebb	
nem vallásos	51,3	51,3	51,3
vallásos	48,7	48,7	48,7
összesen	100,0	100,0	100,0

A részsokaságok képzésére használt csoportképző ismérv és a részsokaságon belüli elemzésre használt ismérv függetlenek egymástól.

Determinisztikus kapcsolat

Nem minden feltételes eloszlás egyforma és a feltételes megoszlásokon belül nincsen szóródás.

Példa:

	50 évesnél fiatalabb	50 éves vagy idősebb
nem vallásos	100,0	0,0
vallásos	0,0	100,0
összesen	100,0	100,0

A részsokaságok képzésére használt csoportképző ismérv és a részsokaságon belüli elemzésre használt ismérv között függvényszerű (determinisztikus) kapcsolat van.

Sztochasztikus kapcsolat
Nem minden feltételes eloszlás egyforma és a feltételes megoszlásokon belül van szóródás.
Példa:

	50 évesnél fiatalabb	50 éves vagy idősebb	
nem vallásos	64,7	33,8	51,3
vallásos	35,3	66,2	47,7
összesen	100,0	100,0	100,0

A részsokaságok képzésére használt csoportképző ismérv és a részsokaságon belüli elemzésre használt ismérv között sztochasztikus kapcsolat van.

Az ismérvek közötti kapcsolat fogalma és fajtái összefoglalva

Két ismérv (X és Y) között háromféle természetű kapcsolat lehet:

1. A két ismérv független egymástól: az X ismérv szerinti hovatartozás ismerete nem ad semmiféle többletinformációt az Y szerinti hovatartozásról az Y szerinti feltétel nélküli megoszláshoz képest.

2. A két ismérv között sztochasztikus kapcsolat van: a megfigyelt sokaság egységeinek X ismérv szerinti hovatartozását ismerve levonható ugyan bizonyos következetés az egységek Y szerinti hovatartozásáról, de az a következtetés nem teljesen egyértelmű.

3. A két ismérv függvényszerű kapcsolatban áll egymással: a vizsgált egységek X szerinti hovatartozásának ismeretében teljes egyértelműséggel megmondható azok Y szerinti hovatartozása is.

A megválaszolandó alapkérdések

1. Van-e kapcsolat a vizsgált ismérvek között? (minta vs. teljes populáció)
2. Milyen erős ez a kapcsolat?
   Hol helyezkedik el a két lehetséges szélsőség, a kapcsolat teljes hiánya (függetlenség) és a lehető legerősebb (függvényszerű) kapcsolat között?

KAPCSOLAT-FAJTÁK AZ EGYIDEJŰLEG VIZSGÁLT KÉT ISMÉRV JELLEGE/MÉRÉSI SZINTJE SZERINT:

1. Asszociáció(s kapcsolat): mindkét ismérv nominális- vagy ordinális mérési szintű.

2. Vegyes kapcsolat: az egyik vizsgált ismérv legalább intervallum mérési szintű, a másik maximum ordinális mérési szintű.

3. Korreláció(s kapcsolat): mindkét vizsgálat ismérv legalább intervallum mérési szintű.

4. Rangkorrelációs(s kapcsolat): mindkét változó sorrendi skálán mérhető.

A HIPOTÉZISVIZSGÁLATRÓL

- A populációval kapcsolatban **hipotéziseket fogalmazhatunk meg**. Ez vonatkozhat a populáció eloszlására, illetve egy paraméterére.
- **Hipotézisvizsgálat**: annak meghatározása, hogy ezek a feltevések, tényleg teljesülnek-e a valóságban.
- Feltevésünket **nullhipotézis** formájában fogalmazzuk meg. A nullhipotézissel szembeni hipotézist **alternatív hipotézis**nek nevezzük.
- **Kérdésünk**: Mennyire valószínű az amit megfigyeltünk, ha a null-hipotézis igaz?
- **H_o=igaz és a megfigyelések valószínűsége kicsi ⇒ elvetjük a nullhipotézist** (Ez nem jelenti azt, hogy a hipotézis nem igaz, csak annyit, hogy nincs okunk elfogadni.)
- **H_o=igaz és a megfigyelések valószínűsége nagy ⇒ elfogadjuk a nullhipotézist** (Ez nem jelenti azt, hogy a hipotézis igaz, csak annyit, hogy nincs okunk elvetni.)

SZIGNIFIKANCIASZINT

Mennyire valószínű, hogy a mintában tapasztalt összefüggést pusztán a mintavételi hiba okozza.

Tehát egy összefüggés 0,05-ös szignifikanciája annyit jelent, hogy legfeljebb 5 százalék valószínűséggel kaphatnánk pusztán a mintavételi hibából eredően olyan erős összefüggést, amilyet megfigyeltünk, feltéve, hogy a nullhipotézis igaz.

A χ^2-PRÓBA

H_o: a két változó független

$$\chi^2 = \sum_{\substack{\min den \\ cella}} \frac{(f_o - f_e)^2}{f_e}$$, ahol f_o a ténylegesen megfigyelt gyakoriság *(observed)*

f_e a függetlenség esetén várt gyakoriság *(expected)*

χ^2-STATISZTIKA KISZÁMÍTÁSA

Megfigyelt cellagyakoriságok

NEM a kérdezett neme by TEMPL volt-e templomban

	volt	nem volt	összesen
férfi	20	50	70
nő	20	10	30
összesen	40	60	100

Várható cellagyakoriságok (független tábla)

NEM a kérdezett neme by TEMPL volt-e templomban

	volt	nem volt	összesen
férfi	28	42	70
nő	12	18	30
összesen	40	60	100

A várható cellagyakoriságok kiszámítása

$$28 = \frac{70 \cdot 40}{100}$$

$$42 = \frac{70 \cdot 60}{100}$$

$$12 = \frac{30 \cdot 40}{100}$$

$$18 = \frac{30 \cdot 60}{100}$$

$$\chi^2 = \frac{(20-28)^2}{28} + \frac{(50-42)^2}{42} + \frac{(20-12)^2}{12} + \frac{(10-18)^2}{18} = 12{,}70$$

$$df = (C-1)(R-1)$$

A χ^2-eloszlás táblázatában a következő értékek szerepelnek az 1 szabadságfok mellett

df	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.0001	.0006	.0039	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827

A példánkban számolt 12,7-es khí-négyzet értéket a mintáknak kevesebb, mint 1 ezrelékében várhatnánk. Azaz, ekkora khí-négyzet értéknek 0,001-nél kisebb a valószínűsége, ha véletlen mintavételt alkalmazunk, s ha a változók függetlenek az alapsokaságban. Ha azonban ennyire valószínűtlen, hogy a megfigyelt eltérés pusztán a mintavételi hibából eredjen, akkor hajlunk a nullhipotézis elvetésére, és inkább úgy döntünk, hogy összefüggés van a változók között.

PHI ÉS CRAMER'S V

χ^2 alapú asszociációs mérőszám

$$\varphi = \sqrt{\frac{\chi_p^2}{N}} \qquad \text{Cramer's V} = \sqrt{\frac{\chi_p^2}{N(q-1)}} \text{ , ahol } q = \min(R, C)$$

Tehát, ha a sorok, vagy az oszlopok száma 2, akkor a két mérőszám értéke megegyezik.

A PRE-MODELL (PROPORTIONATE REDUCTION OF ERROR)

mennyire segíti az egyik változó ismerete a másik értékének az előrejelzését

λ (LAMBDA ASSZOCIÁCIÓS MÉRŐSZÁM)

mekkora aránylagos hibacsökkenést okoz az egyik változó értékeinek megtippelésében a másik változó értékének ismerete (értéke 0 és 1 közé eshet)

$$\lambda = \frac{\text{hibacsökkenés a kétváltozós eloszlás ismeretében}}{\text{hibák száma az egyváltozós eloszlás ismeretében}}$$

	férfi	nő	összesen
alkalmazásban áll	900	200	1100
munkanélküli	100	800	900
összesen	1000	1000	2000

ha meg kellene tippelnünk az emberekről, hogy alkalmazásban állnak-e vagy munkanélküliek, és csak ennek a változónak az eloszlását ismernénk, akkor mindig „alkalmazásban állt" mondanánk, és így **2000-ből 900-szor tévednénk**.
ha most úgy kellene tippelnünk, hogy előtte megmondanák az illető nemét, akkor a férfiaknál az „alkalmazásban állra", míg a nőknél a „munkanélkülire" tippelnénk, és így **2000-ből 300-szor tévednénk**

ekkor $\lambda = \frac{600}{900} = 0{,}6\dot{7}$

A lambda mérőszám nem szimmetrikus

ha meg kellene tippelnünk az emberekről, hogy milyen neműek, és csak ennek a változónak az eloszlását ismernénk, akkor mindegy lenne, hogy mit mondunk mindig, de mindenképpen **2000-ből 1000-szer tévednénk**.
ha most úgy kellene tippelnünk, hogy előtte megmondanák, hogy az illető alkalmazásban áll-e, akkor az alkalmazásban állóknál mindig „férfi"-re, míg az alkalmazásban nem állóknál mindig a „nő"-re tippelnénk, és így **2000-ből 300-szor tévednénk**

ekkor $\lambda = \frac{700}{1000} = 0{,}7$

γ (GAMMA ASSZOCIÁCIÓS MÉRŐSZÁM)

$$\gamma = \frac{\text{egyező} - \text{ellentétes}}{\text{egyező} + \text{ellentétes}}$$

(értéke –1 és 1 közé eshet, így értéke nem csak a változók közötti összefüggés erősségéről, hanem irányáról is informál)

	alsó osztály	középosztály	felsőosztály
alacsony előítéletesség	200	400	700
közepes előítéletesség	500	900	400
magas előítéletesség	800	300	100

egyező nagyságviszonyú párok száma = mindegyik cellában az elemek számát megszorozzuk az alatta és egyben tőle jobbra fekvő cellában lévő elemek számának összegével, majd ezeket összeadjuk

200 (900 + 400 + 300 + 100) + 400 (400 + 100) + 500 (300 + 100) + 900 · 100 = 830 000

ellentétes nagyságviszonyú párok száma = mindegyik cellában az elemek számát megszorozzuk az alatta és egyben tőle balra fekvő cellában lévő elemek számának összegével, majd ezeket összeadjuk

700 (900 + 500 + 300 + 800) + 400 (500 + 800) + 400 (300 + 800) + 900 · 800 = 3 430 000

ekkor $\gamma = \frac{830000 - 3430000}{830000 + 3430000} = -0{,}61$

Kereszttábla készítése az SPSS-ben

Statistics → Summarize → Crosstabs ... →
Suppress tables: a kereszttáblát nem, csak a statisztikákat közli

A 'Crosstabs' ablakon belül lehetőségünk van arra, hogy beállítsuk, milyen adatokat akarunk a cellákban megjeleníteni: **Cells ...**

- **Counts ...**
  - **Observed**: a megfigyelt gyakoriságok
  - **Expected**: a várt[1] gyakoriságok
- **Percentiges ...**
  - **Row** (sor): sorszázalék
  - **Column** (oszlop): oszlopszázalék
  - **Total** (teljes): totálszázalék. Az adott cellába eső esetek aránya az összes megfigyelthez képest.

A 'Crosstabs' ablakon belül lehetőségünk van, hogy különböző statisztikákat kérjünk. Ezek elsősorban a változóink közötti összefüggés meglétének, illetve bizonyos statisztikáknál az összefüggés nagyságának mérésére szolgáló mérőszámok. **Statistics ...**

- **Chi-square**
- **Correlation**: Pearson's R: két, legalább intervallum szintű, változó lineáris összefüggésének mérésére alkalmas mérőszám. Értéke a [-1; 1] zárt intervallumba esik. A negatív értékek negatív (az egyik változó értékének emelkedésével a másik értéke csökken), a pozitívak pozitív összefüggést jelentenek (az egyik változó értékének emelkedésével a másik értéke is nő), ahol a –1 és 1 a teljes lineáris meghatározottságot a 0 pedig azt jelenti, hogy a két változó között nincs lineáris összefüggés vagy más szavakkal a két változó korrelálatlan. Mivel azonban kereszttábla-elemzéssel legfeljebb ordinális mérési szintű változók összefüggését vizsgáljuk, ezt a statisztikát itt sohasem használjuk.
- **Nominal Data** (nominális adatok)
  - **Phi and Cramer's V**: Phi és Cramer's V asszociációs mérőszámok kérése
  - **Lambda**: lambda asszociációs mérőszám kérése
- **Ordinal Data** (ordinális adatok)
  - **Gamma**: gamma asszociációs mérőszám kérése

[1] A várt gyakoriság az adott cellába eső megfigyelések száma a sor- és az oszlopváltozó függetlensége esetén.

MINTAELEMZÉS

Kereszttábla-elemzés nominális független változóval

TEMPLOM milyen gyakran jár templomba by FELEKEZ milyen vallású

FELEKEZ Page 1 of 1

TEMPLOM	Count / Row Pct / Col Pct	római katolikus	református	görög katolikus	evangéliskus	Row Total
		1,00	2,00	3,00	4,00	
,00 soha		930	384	27	44	1385
		67,1	27,7	1,9	3,2	33,2
		32,1	40,1	14,9	32,6	
1,00 ritkán		1562	482	116	68	2228
		70,1	21,6	5,2	3,1	53,5
		54,0	50,3	64,1	50,4	
2,00 gyakran		402	92	38	23	555
		72,4	16,6	6,8	4,1	13,3
		13,9	9,6	21,0	17,0	
Column Total		2894	958	181	135	4168
		69,4	23,0	4,3	3,2	100,0

Chi-Square	Value	DF	Significance
Pearson	58,90508	6	,00000
Likelihood Ratio	62,39757	6	,00000
Linear-by-Linear Association	,11881	1	,73033

Minimum Expected Frequency - 17,976

Statistic	Value	ASE1	Val/ASE0	Approximate Significance
Phi	,11888			,00000
Cramer's V	,08406			,00000
Lambda :				
symmetric	,00000	,00000		
with TEMPLOM dependent	,00000	,00000		
with FELEKEZ dependent	,00000	,00000		
Goodman & Kruskal Tau :				
with TEMPLOM dependent	,00687	,00172		,00000
with FELEKEZ dependent	,00440	,00153		,00000

Number of Missing Observations: 329

A kereszttábla segítségével a felekezet és a templomba járás gyakoriságának összefüggését vizsgáljuk. Fontos, hogy mivel független változónk nominális, az elemzéskor nem beszélhetünk tendenciákról, így ugorva át bizonyos százalékokat.

1. **Hipotézis**
   Hipotézisünk az, hogy a római katolikusok és a görög katolikusok gyakrabban, míg a reformátusok és az evangélikusok ritkábban járnak templomba.
2. **Figyelmeztetés**
   Nincsen figyelmeztetés, tehát nincs olyan cella, ahol a függetlenség esetén várható elemszám 5-nél kisebb.
3. **Van-e összefüggés a változók között?**
   Mivel a khí-négyzet próba szignifikanciája 0,00000, tehát kisebb, mint 0,05, ezért elvetjük a khí-négyzet próba nullhipotézisét, vagyis van összefüggés a változók között.
4. **Milyen erős az összefüggés a változók között?**
   A Cramer's V értéke 0,08, tehát a változók között gyenge erősségű összefüggés van.
5. **A függő és független változó**
   A független változó a felekezet, a függő változó a templomba járás gyakorisága. Tehát az egy sorban lévő oszlopszázalékokat hasonlítjuk majd össze.
6. **Elemzés**
   A templomba soha nem járók aránya a görög katolikusok között a legalacsonyabb (14,9 százalék), a római katolikusok és az evangélikusok között közel egyenlő (32,1, illetve 32,6 százalék) és a reformátusok között a legmagasabb (40,1 százalék).
   A templomba ritkán járók aránya nem különbözik olyan nagyon egymástól a különböző felekezetűek esetében. Azt mondhatjuk, hogy a reformátusok és az evangélikusok között egyenlő arányban szerepelnek (50,3, illetve 50,4 százalék), a római katolikusok között valamennyivel magasabb (54,0 százalék) és a görög katolikusok között a legmagasabb (64,1 százalék).
   A gyakran templomba járók aránya a görög katolikusok között a legmagasabb (21 százalék), az evangélikusok között valamivel alacsonyabb (17 százalék), a római katolikusok között ennél is alacsonyabb (13,9 százalék) és a reformátusok között pedig a legalacsonyabb (9,6 százalék).
7. **Végkövetkeztetés**
   A tábla elemzése alapján tehát azt mondhatjuk, hogy a leggyakrabban a görög katolikusok járnak templomba, közöttük mind a ritkán, mind a gyakran templomba járók aránya a legmagasabb. Utánuk az evangélikusok következnek, akik között a templomba soha nem járók aránya ugyan megegyezik a katolikusok közöttivel, de a gyakran templomba járók 3,1 százalékponttal nagyobb arányban szerepelnek közöttük. Harmadik helyen a római katolikusok állnak. Míg a reformátusok a többiektől lemaradva állnak az utolsó helyen: közöttük a legalacsonyabb a gyakran templomba járók és legmagasabb a soha templomba járók aránya.
   Hipotézisünk tehát csak részben igazolódott be.

Kereszttábla-elemzés ordinális független változóval

TEMPLOM milyen gyakran jár templomba by KOR4 4 kategóriás életkor

	KOR4				Page 1 of 1
Count / Row Pct / Col Pct	18-34 év es	35-49 év es	50-64 év es	65+ éves	Row Total
	1,00	2,00	3,00	4,00	
TEMPLOM					
,00 soha	512 / 33,1 / 41,3	489 / 31,6 / 38,6	309 / 20,0 / 29,0	236 / 15,3 / 28,2	1546 / 35,1
1,00 ritkán	651 / 28,4 / 52,5	695 / 30,3 / 54,8	564 / 24,6 / 53,0	381 / 16,6 / 45,6	2291 / 52,0
2,00 gyakran	78 / 13,6 / 6,3	84 / 14,7 / 6,6	191 / 33,4 / 18,0	219 / 38,3 / 26,2	572 / 13,0
Column Total	1241 / 28,1	1268 / 28,8	1064 / 24,1	836 / 19,0	4409 / 100,0

Chi-Square	Value	DF	Significance
Pearson	264,14757	6	,00000
Likelihood Ratio	255,71545	6	,00000
Linear-by-Linear Association	168,88022	1	,00000

Minimum Expected Frequency - 108,458

Statistic	Value	ASE1	Val/ASE0	Approximate Significance
Phi	,24477			,00000
Cramer's V	,17308			,00000
Gamma	,24008	,01948	12,01473	,00000

Number of Missing Observations: 88

A kereszttábla segítségével a templomba járás gyakoriságát mérő változó és az életkor összefüggését vizsgáljuk. Független változónk ordinális, tehát az elemzéskor beszélhetünk tendenciákról.

1. **Hipotézis**
   Hipotézisünk az, hogy az idősebbek gyakrabban járnak templomba.
2. **Figyelmeztetés**
   Nincsen figyelmeztetés, tehát nincs olyan cella, ahol a függetlenség esetén várható elemszám 5-nél kisebb.
3. **Van-e összefüggés a változók között?**
   Mivel a khí-négyzet próba szignifikanciája 0,00000, tehát kisebb, mint 0,05, ezért elvetjük a khí-négyzet próba nullhipotézisét, vagyis van összefüggés a változók között.
4. **Milyen erős az összefüggés a változók között?**
   A Cramer's V értéke 0,17, tehát a változók között gyenge (majdnem közepes) erősségű összefüggés van.
5. **A függő és független változó**
   A független változó az életkor, a függő változó a templomba járás gyakorisága. Tehát az egy sorban lévő oszlopszázalékokat hasonlítjuk majd össze.
6. **Elemzés**
   Míg a 18–34 évesek 41,3 százaléka nem jár soha templomba, addig a 65 évesnél idősebbek között ez az arány 28,2 százalék és a kettő között folyamatos csökkenést látunk (38,6, illetve 29 százalék). Tehát a fiatalabbak között magasabb a templomba soha nem járók aránya.
   A 18–34 évesek között 52,5 százalék a templomba ritkán járók aránya, a 35–49 évesek között 54,8 százalék, az 50–64 évesek között 53 százalék, azonban a 65 évesnél idősebbek között csak 45,6 százalék. Tehát míg az első három korcsoport esetében a templomba ritkán járók aránya nem különbözik lényegesen, addig a legidősebb korosztályban ennél lényegesen alacsonyabb.
   Míg a 18–34 évesek 6,3 százaléka jár gyakran templomba, addig a 65 évesek között ez az arány 26,2 százalék és a közötte folyamatos emelkedést látunk (6,6, illetve 18,0 százalék). Tehát az idősebbek között magasabb a templomba gyakran járók aránya.
7. **Végkövetkeztetés**
   A tábla elemzése alapján tehát azt mondhatjuk, hogy valóban minél idősebb valaki annál valószínűbb, hogy gyakrabban jár templomba. Hipotézisünk tehát beigazolódott.