Két ismérv szerinti elemzés

Egy sokaságot egyszerre több ismérv szerint is vizsgálhatunk.

A következő ismérveket különböztetjük meg:

- minőségi ismérv (pl. férfi vagy nő)
- területi ismérv (pl. városban vagy faluban lakik)
- mennyiségi ismérv (pl. milyen magasak)

Megnézem a kapcsolódó epizódot

Ha mindkét ismérv minőségi (vagy területi), akkor asszociációs kapcsolatról beszélünk.

Ilyen például egy cég alkalmazottjainak megoszlása neme és beosztása szerint.

Minőségi				
		Nő	Férfi	Össz.
	Vezető	7	18	25
Minőségi	Közép-vezető	11	23	34
	Beosztott	756	185	941
	Össz.	774	226	1000

Az így létrejövő táblát kombinációs táblának nevezzük. Átlagot, szórást és egyéb mutatókat egyik ismérv szerint sem tudunk számolni.

Megnézem a kapcsolódó epizódot

Ha az egyik ismérv minőségi (vagy területi), a másik mennyiségi, akkor vegyes kapcsolatról beszélünk.

Ilyen például egy város szállodáinak megoszlása az éjszakák ára és a szállodák besorolása alapján.

Minőségi					
	Árak (EUR/fő/éj)	Szálloda típusa			Össz.
		**	***	****	
	0-50	37	8	1	46
Mennyiségi	51-100	15	40	3	58
	101-150	10	33	12	55
	151-200	4	22	15	41
	Össz.	66	103	31	200

Átlagot, szórást és egyéb mutatókat csak az egyik ismérv, az árak szerint tudunk számolni.

Megnézem a kapcsolódó epizódot

Ha mindkét ismérv mennyiségi, akkor korrelációs kapcsolatról beszélünk.

Ilyen például Európa 4 országának egy főre jutó GDP-je és az ezer főre jutó gépjárművek számának megoszlása.

	Mennyiségi	Mennyiségi
Ország	GDP/fő (USD)	Gépjárművek (db/1000 fő)
Ausztria	50 380	550
Belgium	46 237	503
Hollandia	52 646	481
Svájc	82 484	539

A táblázatban mindkét ismérv szerint tudunk átlagot, szórást és egyéb mutatókat számolni.

Megnézem a kapcsolódó epizódot

Ha mindkét ismerv sorrendi, akkor rangkorrelációs kapcsolatról beszélünk.

Ilyen példáuk ha két társadalmi csoportot kérdezünk meg, hogy 1-től 10-ig rangsorolják az alábbi országokat, az alapján, hogy mennyire szívesen nyaralnának ott.

Ország	Egyik csoport	Másik csoport
Ausztria	10	2
Belgium	9	6
Csehország	4	7
Franciaország	3	3
Görögország	1	8
Hollandia	7	5
Lengyelország	2	9
Magyarország	6	10
Németország	5	4
Svájc	8	1

Megnézem a kapcsolódó epizódot

Két ismérv akkor független, ha minden feltételes megoszlás egyforma és megegyezik a feltétel nélküli megoszlással.

Megnézem a kapcsolódó epizódot

Két ismérv között függvényszerű kapcsolat van, ha nem minden feltételes megoszlás egyforma, de minden feltételes eloszlás szórása nulla.

Függvényszerű kapcsolatnál az egyik ismérv ismeretében a másik egyértelműen kitalálható.

A két ismérv kapcsolata akkor függvényszerű, ha nem minden feltételes megoszlás egyforma, de a feltételes megoszlások szórása nulla.

Megnézem a kapcsolódó epizódot

Ha a két ismérv közötti kapcsolat nem független és nem is függvényszerű, akkor sztochasztikus kapcsolatról beszélünk. Kicsit összefüggnek ugyan az adatok, de olyan nagyon azért nem.

A két ismérv kapcsolata akkor sztochasztikus, ha nem minden feltételes megoszlás egyforma de a feltételes megoszlások szórása nem mind nulla.

Megnézem a kapcsolódó epizódot

A Cramer-féle asszociációs együttható arra való, hogy amikor mindkét ismérv minőségi, rávilágítson a két ismérv közötti kapcsolat szorosságára.

$$C = \sqrt{\frac{\chi^2}{N \cdot min\{(r-1);(c-1)\}}}$$

Itt N= az összes elem, r= a táblázat sorainak száma és $c =$ a táblázat oszlopainak száma, továbbá

$$\chi^2 = \sum \frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*}$$

A Cramer-mutató függvényszerű kapcsolat esetén 1, független esetén pedig 0.

Megnézem a kapcsolódó epizódot

A kombinációs tábla általános sémája:

	C_1	C_2	$\cdots$	C_j		Össz.
R_1	f_{11}	f_{12}	$\cdots$	f_{1j}		$f_{1\bullet}$
R_2	f_{21}	f_{22}	$\cdots$	f_{2j}		$f_{2\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
R_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}		$f_{i\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Össz.	$f_{\bullet 1}$	$f_{\bullet 2}$	$\cdots$	$f_{\bullet j}$		N

Az első oszlop elemei, amint látjuk f_{11} aztán f_{21} és így tovább az általános tag f_{1i}, ami közös bennük az az, hogy a második indexe mindegyiknek 1-es.

Az oszlop alján összegezzük őket, az összeg $f_{\bullet 1}$ ami azt jelenti, hogy ez azoknak az elemeknek az összege, ahol a második index 1, az első index pedig tökmindegy, hogy mi, ezt hivatott jelezni a ● jel.

Aztán a második oszlopban pontosan ugyanez a helyzet, az oszlopban lévő elemek f_{12} alatta f_{22} és így tovább, összegük pedig $f_{\bullet 2}$.

Ugyanez megy a sorokra is, az első sor elemei f_{11} aztán f_{12} és így tovább, itt az elemek első indexe egyezik meg, mindegyiknek 1-es, összegüket pedig úgy jelöljük, hogy $f_{1\bullet}$.

Megnézem a kapcsolódó epizódot

A Csuprov-féle mutató segítségével két ismérv közötti kapcsolatot vizsgálhatjuk.

A Csuprov-féle mutató:

$$\Gamma = \sqrt{\frac{\chi^2}{N \cdot \sqrt{r-1} \cdot \sqrt{c-1}}}$$

Megnézem a kapcsolódó epizódot

Ha azt vizsgáljuk, hogy az egyes értékek mennyire térnek el a részátlagoktól, akkor belső szórást számolunk.

Jele: σ_B

A rész-szórásokból úgy lesz belső szórás, hogy súlyozzuk őket a rész-sokaságok szórásával.

A belső szórást kiszámolhatjuk a rész-szórások nélkül is.

Megnézem a kapcsolódó epizódot

Ha a részátlagoknak nézzük a főátlagtól való eltérését, az a külső szórás.

Jele: σ_K

Megnézem a kapcsolódó epizódot

Ha az egyes értékeknek nézzük a főátlagtól való eltérést, az a teljes szórás.

A teljes szórás az egész sokaság szórását jelenti, vagyis ha nem bontjuk fel a sokaságot rész-sokaságokra.

Jele :σ

A háromféle szórásra mindig teljesül, hogy

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

Megnézem a kapcsolódó epizódot

A belső eltérés-négyzetösszeg a belső szórás gyök alatti részének számlálója.

Jele: SSB

Megnézem a kapcsolódó epizódot

A külső eltérés-négyzetösszeg a külső szórás gyök alatti részének számlálója.

Jele: SSK

Megnézem a kapcsolódó epizódot

A teljes eltérés-négyzetösszeg a teljes szórás gyök alatti részének számlálója.

Jele: SST

$$SST = SSB + SSK$$

Megnézem a kapcsolódó epizódot

A PRE egy rövidítés, Proportional Reduction Errors, ami relatív hibacsökkenésnek fordítható.

A módszer lényege, hogy a PRE érték kiszámolásával megállapítható, az egyik ismérv ismerete hány százalékkal csökkenti a másik ismérv nagyságával kapcsolatos bizonytalanságot.

Ha $PRE = 0$, az azt jelenti, hogy ez a bizonytalanság egyáltalán nem csökken. Ebben az esetben a két ismérv egymástól független.

Ha $PRE = 1$, akkor a bizonytalanság 100%-kal csökken. Ilyenkor a két ismérv között fügvényszerű kapcsolat van.

Ha pedig PRE értéke valahol nulla és egy között van, akkor a kapcsolat nem független és nem is függvényszerű, tehát sztochasztikus.

$$PRE = H^2 = \frac{\sigma^2 - \sigma_B^2}{\sigma^2} = \frac{\sigma_K^2}{\sigma^2}$$

vagy

$$PRE = H^2 = \frac{SST - SSB}{SST} = \frac{SSK}{SST}$$

Megnézem a kapcsolódó epizódot

Ha két ismérv között korrelációs kapcsolat van, akkor a két ismérv közötti kapcsolat szorosságát a lineáris korrelációs együttható írja le:

$$r = \frac{\sum dX \cdot dY}{\sqrt{\sum d^2X \cdot \sum d^2Y}}$$

A lineáris korrelációs együttható azt méri, hogy X és Y között milyen szoros lineáris kapcsolat van. Értéke mindig $-1 \geq r \geq 1$.

Ha $r = \pm 1$, akkor X és Y között függvényszerű lineáris kapcsolat van, ha $r = 0$, akkor nincs lineáris kapcsolat. De attól, hogy nincs lineáris kapcsolat, másfajta kapcsolat még lehet, tehát $r = 0$ esetén X és Y nem biztos, hogy független.

Megnézem a kapcsolódó epizódot

A determinációs együttható a lineáris korrelációs együttható négyzete, azaz r^2.

A determinációs együttható pontosan úgy értelmezhető, mint a PRE mutató a vegyes kapcsolatnál.

Megnézem a kapcsolódó epizódot

Ha pl. egy verseny eredményét ketten is megtippelik, és el kell döntenünk melyikük találta el jobban a valós eredményt...

Erre való a rangkorrelációs együttható:

$$\rho = 1 - \frac{6 \cdot \sum (R_X - R_Y)^2}{N(N^2-1)}$$

Hogyha valaki éppen eltalálja a helyes sorrendet, akkor a rangkorreláció értéke 1.

Ha pedig éppen a fordított sorrendet találja el, akkor -1.

És minél inkább eltalálja valaki a valós sorrendet, a rangkorreláció annál nagyobb.

Megnézem a kapcsolódó epizódot