MATEKING.HU

Képletgyűjtemény

STATISZTIKA tantárgy

Kiadás dátuma: 2026. 02. 07.

Tartalomjegyzék

Alapfogalmak

Az ismérvek olyan vizsgálati szempontok, amelyek alapján a sokaság részekre osztható.

Vannak olyanok, amik csak két részre osztják a sokaságot, például azokra, akik megbuktak statisztikából és azokra akik nem. Olyan is van, ami mondjuk öt részre osztja a sokaságot, sőt olyan is lehet, hogy végtelen sok részre osztja.

Négy fő ismérvfajta különböztethető meg:

- területi
- időbeli
- mennyiségi
- minőségi

Megnézem a kapcsolódó epizódot

Az ismérveken belül négyféle mérési szintet tudunk megkülönböztetni.

A nominális (névleges) mérési skála a sokaság elemeit valamilyen tulajdonság szerint csoportokba sorolja, de a csoportok között nincsen semmilyen sorrendiség.

Ilyenek pl.:

- Az áldozatok halálának oka (MINŐSÉGI)
- A terroristák származási országa (TERÜLETI)
- A pilóták állampolgársága (MINŐSÉGI)

Megnézem a kapcsolódó epizódot

Az ismérveken belül négyféle mérési szintet tudunk megkülönböztetni.

Az ordinális (sorrendi) mérési skála a sokaság elemeit valamilyen tulajdonság szerint csoportokba sorolja, és a csoportok között van sorrendiség.

Ilyenek pl.:

- A statisztika vizsga osztályzata (MINŐSÉGI)
- A hotelek csillagos értékelése (MINŐSÉGI)

Megnézem a kapcsolódó epizódot

Az intervallum-skála a sokaság elemeit valamilyen mérés szerint rendezi sorba.

Az intervallum-skála jellegzetes tulajdonsága, hogy a "mennyivel több" kérdésre választ tud adni, de a "hányszor annyi" kérdésre nem.

Ilyenek pl.:

- Hőmérséklet (MENNYISÉG)
- Születési dátum (IDŐBELI)

Megnézem a kapcsolódó epizódot

Az arány-skála a sokaság elemeit szintén valamilyen mérés szerint rendezi sorba, de abban különbözik az intervallum-skálától, hogy ennek van valódi nullpontja.

Innen ered az arány-skála elnevezés is, van értelme az egymáshoz viszonyított arányoknak.

Ilyenek pl.:

- Életkor (MENNYISÉGI)
- Testmagasság (MENNYISÉGI)

Megnézem a kapcsolódó epizódot

Hogyha egy sokaság elemei egymástól jól elkülöníthető egységek, akkor a sokaság diszkrét.

Egy évfolyam hallgatói például diszkrét sokaság, egy borászat által termelt éves bormennyiség viszont már nem.

Megnézem a kapcsolódó epizódot

Hogyha egy sokaság nem diszkrét, akkor az folytonos.

Egy évfolyam hallgatói például diszkrét sokaság, egy borászat által termelt éves bormennyiség viszont már folytonos sokaság.

Megnézem a kapcsolódó epizódot

Az időpontra vonatkozó sokaságokat álló sokaságnak nevezzük.

Például egy város lakosainak száma január elsején egy álló sokaság.

Megnézem a kapcsolódó epizódot

Az időtartamra vonatkozó sokaságokat mozgó sokaságnak nevezzük.

Például egy városban január elsején született lakosok száma mozgó sokaság.

Megnézem a kapcsolódó epizódot

A viszonyszámok jele V és kiszámolásának módja meglehetősen semmitmondó:

$$V = \frac{A}{B}$$

A képletben A és B bármi lehet, de a képlet mégis roppant fontos.

Megnézem a kapcsolódó epizódot

Ha több viszonyszámunk van, fölmerülhet az igény ezek átlagolására.

Az egyik ilyen lehetőség a számtani átlag.

$$\overline{V} = \frac{V_1 \cdot B_1 + V_2 \cdot B_2}{B_1 + B_2} = \frac{A_1 + A_2}{B_1 + B_2}$$

Megnézem a kapcsolódó epizódot

Ha több viszonyszámunk van, fölmerülhet az igény ezek átlagolására.

Az egyik ilyen lehetőség a harmonikus átlag.

$$\overline{V} = \frac{A_1+A_2}{\frac{A_1}{V_1}+\frac{A_2}{V_2}} = \frac{A_1+A_2}{B_1+B_2}$$

Megnézem a kapcsolódó epizódot

A dinamikus viszonyszámok idősorok adataiból számított hányadosok.

Megnézem a kapcsolódó epizódot

A megoszlási viszonyszám egy sokaság valamely részének az egészhez viszonított arányát írja le.

Megnézem a kapcsolódó epizódot

Az intenzitási viszonyszám két, egymással valamilyen kapcsolatban álló sokaság mennyiségeinek hányadosa.

Megnézem a kapcsolódó epizódot

Azokat az adatsorokat nevezzük idősornak, melyek egy vagy több ismérv időben történő megoszlását írják le.

Megnézem a kapcsolódó epizódot

A bázisviszonyszámok mindig a bázishoz viszonyítanak.

$$b_m = \frac{y_m}{y_1}$$

A bázisviszonyszámok és a láncviszonyszámok közötti kapcsolat a következő:

$$l_2 \cdot l_3 \cdot \dots \cdot l_m = \frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \dots \cdot \frac{y_m}{y_{m-1}} = b_m$$

Egy másik nagyon fontos összefüggés, hogy

$$l_m = \frac{y_m}{y_{m-1}} = \frac{b_m}{b_{m-1}}$$

Megnézem a kapcsolódó epizódot

A láncviszonyszámok mindig az előző évhez viszonyítanak.

$$l_m = \frac{y_m}{y_{m-1}}$$

A bázisviszonyszámok és a láncviszonyszámok közötti kapcsolat a következő:

$$l_2 \cdot l_3 \cdot \dots \cdot l_m = \frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \dots \cdot \frac{y_m}{y_{m-1}} = b_m$$

Egy másik nagyon fontos összefüggés, hogy

$$l_m = \frac{y_m}{y_{m-1}} = \frac{b_m}{b_{m-1}}$$

Megnézem a kapcsolódó epizódot

A változás átlagos mértéke:

$$\overline{d} = \frac{\sum d_1}{n-1} = \frac{y_n - y_1}{n-1}$$

Tehát összeadogatjuk a változásokat, aztán elosztjuk...

Az évek száma n, de nem n-el osztunk, azért nem, mert a változások számával kell osztanunk, ebből pedig egyel kevesebb van.

Megnézem a kapcsolódó epizódot

A változás üteme azt adja meg, hogy hány százalékos volt a változás.

A változás üteme:

$$\overline{l} = \sqrt[n-1]{l_2 \cdot l_3 \cdot l_4 \cdot \dots \cdot l_n} = \sqrt[n-1]{\prod_{t=2}^{n} l_t} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

A gyökkitevőben azért van $n-1$, mert nem az évek száma kell, hanem a változások száma, egyik évről a másikra. Ez pedig $n-1$.

Megnézem a kapcsolódó epizódot

A tartamidősorok egy vizsgált időtartamra vonatkozó megfigyeléseket tartalmaznak.

Például egy év baleseteinek a számát, egy hónapban eladott fogkrémek számát, stb. Ilyenkor az adatok összeadása értelmes eredményt ad.

Megnézem a kapcsolódó epizódot

Az állapotidősorok egy vizsgált időtartam egy adott pillanatára vonatkozó megfigyeléseket tartalmazzák, például az ország lakosságának számát egy adott év adott pillanatában, vagy a raktáron lévő fogkrémkészletet egy adott hónap adott pillanatában, stb. és ilyenkor az adatok összeadásával nem kapunk értelmezhető eredményt.

Megnézem a kapcsolódó epizódot

Egy speciális átlag, például ha négy hónap adataiból számoljuk ki az átlagot, viszont csak három hónapos időtartamra.

Az állapotidősornál mindig kronologikus átlagot számolunk:

$$\bar{y}_k = \frac{\frac{y_1}{2}+y_2+y_3+\cdots+\frac{y_n}{2}}{n-1}$$

Megnézem a kapcsolódó epizódot

A statisztikai táblákat három fő csoportba tudjuk sorolni.

A legegyszerűbb típust egyszerű táblának nevezzük.

Ezek tulajdonképpen egymás mellett szerepeltetett valamilyen adatok.

A táblában nincsen sem "Összesen" sor, sem pedig "Összesen" oszlop.

Például:

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

Megnézem a kapcsolódó epizódot

A statisztikai táblákat három fő csoportba tudjuk sorolni.

Az egyik típus az úgynevezett csoportosító tábla, aminek lényege, hogy az adatokat valamelyikismérv szerint tudjuk összesíteni.

Például:

Ország	Népesség (millió)	Sertések száma (millió)
Ausztria	8,9	2,8
Belgium	11,5	6,2
Hollandia	17,4	11,9
Svájc	8,6	1,4
Összesen:	46,4	22,3

Megnézem a kapcsolódó epizódot

A statisztikai táblákat három fő csoportba tudjuk sorolni.

Az egyik típus az úgynevezett kombinációs tábla vagy más néven kontingencia tábla, amely esetében mindegyik ismérv szerint tudjuk az adatokat összesíteni.

Például:

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

Megnézem a kapcsolódó epizódot

Egy ismérv szerinti elemzés

Az adatsor legyakoribb értéke a módusz. Hogyha például Bob matekjegyei ezek:

2, 3, 1, 4, 1, 2, 2, 3, 5, 2, 3, 2, 3, 2, 4, 3, 2, 4, 2, 4

Akkor egyszerűen meg kell számolni, hogy melyikből van a legtöbb, és az a matekjegy lesz a módusz. Most 2-esből van a legtöbb, így Bob matekjegyeinek a módusza 2. A módusz jele Mo és így most Mo=2.

Léteznek olyan eloszlások is, amelyeknek több módusza van. Hogyha például Bob jegyei:

1, 2, 2, 3, 5, 3, 3, 4, 2

Itt 2-esből és 3-asból ugyanannyi van, mindkettőből 3 darab. Ez egy kétmóduszú eloszlás.

Megnézem a kapcsolódó epizódot

A medián a növekvő sorba rendezett adatsor középső értéke. Ha az adatsorban páros sok elem van, akkor nincs középső elem, ilyenkor a két középső elem átlagát vesszük.

Hogyha például Bob matekjegyei ezek:

2, 3, 1, 4, 1, 2, 2, 3, 5, 2, 3, 2, 3, 2, 4, 3, 2, 4, 2, 4

Akkor egyszerűen növekvő sorba kell rakni..

1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5

És aztán meg kell keresni melyik a középső. Most nincsen középső, mert páros sok elem van, így ilyenkor a két középen lévőt átlagoljuk:

1, 1, 2, 2, 2, 2, 2, 2, 2, **2**, **3**, 3, 3, 3, 3, 4, 4, 4, 4, 5

Ezeknek az átlaga 2,5 vagyis a medián most 2,5. A medián jele Me, így Me=2,5

Megnézem a kapcsolódó epizódot

Az átlagot úgy kapjuk meg, hogy az összes elemet összeadjuk, és aztán elosztjuk az elemek számával.

Jele: $\overline{x}$

Megnézem a kapcsolódó epizódot

Az átlagtól való átlagos eltérés egyik legjobb mérőszáma a szórás. Hátránya, hogy egy kicsit ronda a szórás képlete. A szórást egy szigma nevű görög betűvel jelöljük.

$$\sigma = \sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}}$$

Megnézem a kapcsolódó epizódot

Az adatsor első felének a felezőpontja az alsó kvartilis.

Az alsó kvartilis jele: Q_1

Megnézem a kapcsolódó epizódot

Az adatsor második felének a felezőpontja a felső kvartilis.

A felső kvartilis jele: Q_3

Megnézem a kapcsolódó epizódot

A kvartilisek és a medián azt szemlélteti, hogyan oszlanak el az adatsorban szereplő adatok. Ezek segítségével készíthető el a doboz-ábra, vagy másnéven dobozdiagram. Szokás még sodrófa diagramnak is nevezni, és az angol elnevezést is gyakran használják, ami a box plot.

Egy sobarendezett adatsorban öt darab speciális negyedelőpontot fogunk használni. Az első az adatsor legkisebb értéke, ez a Q_0 . Aztán a következő negyedelő az alsó kvartilis, ami Q_1 utána jön a felezőpont vagyis a medián, ezt Me-vel és Q_2-vel is jelöljük, végül a felső kvartilis, ami a Q_3. Az adatsor legnagyobb értéke pedig Q_4. A legnagyobb és a legkisebb érték különbsége a terjedelem, míg a két kvartilis különbségét félterjedelemnek vagy más néven interkvartilisnek hívjuk. Ezekből épül föl a doboz-ábra vagy másként dobozdiagram.

Előfordulhat, hogy az adatsorban kiugró értékek is szerepelnek. A kiugró érték az, ami az alsó kvartilisnál legalább a félterjedelem másfélszeresénél kisebb, vagy pedig a felső kvartilisnél legalább a félterjedelem másfélszeresénél nagyobb. Huh, ez elég bonyolultan hangzik. De valójában nagyon egyszerű, csak nézd meg kapcsolódó epizódot és kiderül.

Megnézem a kapcsolódó epizódot

A relatív szórás azt mondja meg, hogy a szórás az átlagnak hány százaléka:

$$V = \frac{\sigma}{\overline{X}}$$

Megnézem a kapcsolódó epizódot

$$Mo = mo + \frac{k_1}{k_1+k_2} \cdot h_{mo}$$

A képletben mo a nyers módusz, ami a legnagyobb gyakoriságú osztály alsó határa. A k_1-et úgy kapjuk, ha ennek az osztályköznek a gyakoriságából kivonjuk az előtte lévő osztályköz gyakoriságát. A k_2-t pedig úgy kapjuk, ha ennek az osztályköznek a gyakoriságából az utána lévő osztályköz gyakoriságát vonjuk le. A h_{mo} pedig ennek az osztályköznek a hosszát jelöli.

Megnézem a kapcsolódó epizódot

$$Me = me + \frac{\frac{N}{2} - f'_{me-1}}{f_{me}} \cdot h_{me}$$

Itt f' a kumulált gyakoriság. me a mediánt tartalmazó osztályköz eleje, f_{me} a mediánt tartalmazó osztályköz gyakorisága, f'_{me-1} a medián előtti osztályköz kumulált gyakorisága, h_{me} pedig a mediánt tartalmazó osztályköz hossza.

Megnézem a kapcsolódó epizódot

$$Q_{\frac{k}{m}} = a_i + \frac{\frac{k}{m} \cdot N - f'_{i-1}}{f_i} \cdot h_i$$

Az alsó kvartilis:

$$Q_{\frac{1}{4}} = a_i + \frac{\frac{1}{4} \cdot N - f'_{i-1}}{f_i} \cdot h_i$$

A felső kvartilis:

$$Q_{\frac{3}{4}} = a_i + \frac{\frac{3}{4} \cdot N - f'_{i-1}}{f_i} \cdot h_i$$

Megnézem a kapcsolódó epizódot

A relatív gyakoriság jele g_i, és úgy kell kiszámolni, hogy a gyakoriságot osztjuk az összes elemszámmal:

$$g_i = \frac{f_i}{N}$$

Megnézem a kapcsolódó epizódot

Az értékösszeg jele S_i és úgy kapjuk meg, hogy az osztályközepeket megszorozzuk a gyakorisággal.

$$S_i = X_i \cdot f_i$$

Megnézem a kapcsolódó epizódot

A Herfindahl-index egy eszköz a koncentráció vizsgálatára.

$$HI = \sum Z_i^2$$

A Herfindahl-index 1/N és 1 között vesz fel értékeket és minél közelebb van az 1-hez, annál nagyobb a koncentráció.

Megnézem a kapcsolódó epizódot

A Lorenz-görbe egy eszköz a koncentráció vizsgálatára.

A Lorenz-görbe az úgynevezett koncentrációs területtel szemlélteti a koncentráció mértékét.

Minél nagyobb ez a terület, a koncentráció annál erősebb.

Olyankor pedig, amikor a Lorenz görbe egybeesik a négyzet átlójával, a koncentráció nulla.

Megnézem a kapcsolódó epizódot

Az alakmutatók arról szólnak, hogy az eloszlás mennyire asszimetrikus.

Az egyik legegyszerűbb és leggyakrabban használt alakmutatók, az úgynevezett Pearson-féle mérőszámok:

$$P = 3\frac{\overline{X}-Me}{\sigma} \qquad A = \frac{\overline{X}-Mo}{\sigma}$$

A negatív értékek jobb oldali asszimetriát jelentenek. A pozitív értékek esetén pedig bal oldali asszimetria van.

A Pearson-féle P és A mutatók általában -1 és 1 között tartózkodnak és csak extrém esetekben vesznek föl 1-nél nagyobb vagy -1-nél kisebb értéket.

Megnézem a kapcsolódó epizódot

Az alakmutatók arról szólnak, hogy az eloszlás mennyire asszimetrikus.

Az egyik legegyszerűbb és leggyakrabban használt alakmutatók, az úgynevezett Pearson-féle mérőszámok mellett az F-mutatók:

$$F_{0,25} = \frac{(Q_3-Me)-(Me-Q_1)}{(Q_3-Me)+(Me-Q_1)}$$

$$F_{0,1} = \frac{(D_9-Me)-(Me-D_1)}{(D_9-Me)+(Me-D_1)}$$

ahol D_1 az első, D_9 pedig a kilencedik decilist jelenti.

A negatív értékek jobb oldali asszimetriát jelentenek. A pozitív értékek esetén pedig bal oldali asszimetria van.

Az F mutató csak -1 és 1 között lehet.

Megnézem a kapcsolódó epizódot

A csúcsosság azt jelenti, hogy az eloszlás görbéje mennyire csúcsosodik ki.

A csúcsosság mérésére a következő mutató van forgalomban:

$$\alpha_4 = \frac{M_4(\overline{X})}{\sigma^4} - 3$$

Itt $M_4(\overline{X})$ az úgynevezett negyedik momentum, és így számolható ki:

$$M_4(\overline{X}) = \frac{\sum\left(\overline{X}-X_i\right)^4}{N}$$

Megnézem a kapcsolódó epizódot

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				
Minőségi		Nő	Férfi	Össz.
	Vezető	7	18	25
	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 egyikismérv 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					
Mennyiségi	Árak (EUR/fő/éj)	Szálloda típusa			Össz.
		**	***	****	
	0-50	37	8	1	46
	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 egyikismérv 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 kétismérv egymástól független.

Ha $PRE = 1$, akkor a bizonytalanság 100%-kal csökken. Ilyenkor a kétismérv 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

Standardizálás

Az első szint a főátlagok összehasonlítása:

$$K = \overline{V}_0 - \overline{V}_1$$

A második szint a részhatás különbség kiszámolása:

(ilyenkor az összetételhatást vesszük standardnek)

$$K' = \overline{V}'_0 - \overline{V}'_1 = \frac{\sum B_0 \cdot V_0}{\sum B_0} - \frac{\sum B_0 \cdot V_1}{\sum B_0}$$

A harmadik szint az összetételhatás különbség kiszámolása:

$$K'' = \overline{V}''_0 - \overline{V}''_1 = \frac{\sum B_0 \cdot V_1}{\sum B_0} - \frac{\sum B_1 \cdot V_1}{\sum B_1}$$

Megnézem a kapcsolódó epizódot

Területi összehasonlítás:

$$K = \overline{V}_1 - \overline{V}_0$$

$$K' = \overline{V}'_1 - \overline{V}'_0$$

$$K'' = \overline{V}''_1 - \overline{V}''_0$$

$$K = K' + K''$$

Megnézem a kapcsolódó epizódot

Időbeli összehasonlítás:

$$I = \frac{\overline{V}_1}{\overline{V}_0}$$

$$I' = \frac{\overline{V}'_1}{\overline{V}'_0}$$

$$I'' = \frac{\overline{V}''_1}{\overline{V}''_0}$$

$$I = I' \cdot I''$$

Megnézem a kapcsolódó epizódot

Az első szint a főátlagok összehasonlítása:

$$I = \frac{\overline{V}_1}{\overline{V}_0}$$

A második szint a részhatásindex kiszámolása:

(ilyenkor az összetételhatást vesszük standardnek és mindig a tárgyidőszakét)

$$I' = \frac{\overline{V}'_1}{\overline{V}'_0} = \frac{\frac{\sum B_1 \cdot V_1}{\sum B_1}}{\frac{\sum B_1 \cdot V_0}{\sum B_1}}$$

A harmadik szint az összetételhatás-index kiszámolása:

(ilyenkor a részhatást vesszük standardnek és mindig a bázisidőszakit, de ha I és I' már megvan, akkor egyszerűbb az $I = I' \cdot I''$ alapján számolni.)

$$I'' = \frac{\overline{V}''_1}{\overline{V}''_0} = \frac{\frac{\sum B_1 \cdot V_0}{\sum B_1}}{\frac{\sum B_0 \cdot V_0}{\sum B_0}}$$

Megnézem a kapcsolódó epizódot

Indexszámítás

A volumenindex a forgalom változásának mértéke, amit súlyozhatunk a bázisidőszak vagy a tárgyidőszak áraival.

Volumenindex Laspeyres /bázis időszak szerint/:

$$I_q^0 = \frac{\sum p_0 q_1}{\sum p_0 q_0}$$

Volumenindex Paasche /tárgyidőszak szerint/:

$$I_q^1 = \frac{\sum p_1 q_1}{\sum p_1 q_0}$$

Megnézem a kapcsolódó epizódot

Az árindex a szektort érintő árváltozást méri, és súlyozhatjuk a bázisidőszak vagy a tárgyidőszak volumeneivel.

Árindex Laspeyres /bázis időszak szerint/:

$$I_p^0 = \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

Árindex Paasche /tárgyidőszak szerint/:

$$I_p^1 = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

Megnézem a kapcsolódó epizódot

Az indexek átlagformái:

$$I_p^0 = \frac{\sum p_0 q_0 \cdot i_p}{\sum p_0 q_0} = \frac{\sum p_1 q_0}{\sum \frac{p_1 q_0}{i_p}} = \frac{\sum v_0 \cdot i_p}{\sum v_0}$$

$$I_p^1 = \frac{\sum p_0 q_1 \cdot i_p}{\sum p_0 q_1} = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}} = \frac{\sum v_1}{\sum \frac{v_1}{i_p}}$$

$$I_q^0 = \frac{\sum p_0 q_0 \cdot i_p}{\sum p_0 q_0} = \frac{\sum p_0 q_1}{\sum \frac{p_0 q_1}{i_q}} = \frac{\sum v_0 \cdot i_q}{\sum v_0}$$

$$I_q^1 = \frac{\sum p_1 q_0 \cdot i_q}{\sum p_1 q_0} = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_q}} = \frac{\sum v_1}{\sum \frac{v_1}{i_q}}$$

Megnézem a kapcsolódó epizódot

A Fischer-féle árindex és volumenindex:

$$I_p^F = \sqrt{I_p^0 \cdot I_p^1} \qquad I_q^F = \sqrt{I_q^0 \cdot I_q^1}$$

Megnézem a kapcsolódó epizódot

Az értékindex:

$$I_v = \frac{\sum p_1 q_1}{\sum p_0 q_0} = I_p^1 \cdot I_q^0 = I_q^1 \cdot I_p^0 = I_p^F \cdot I_q^F$$

Megnézem a kapcsolódó epizódot

A vásárlóerő mérésére van forgalomban az úgynevezett vásárlóerő-paritás. Angol megfelelője Purchasing Power Parity alapján rövidítése PPP.

A vásárlóerő-paritás az egyedi vásárlóerő-paritásoknak a fogyasztással súlyozott átlaga.

Megnézem a kapcsolódó epizódot

Idősorok

A dekompozíciós modellek lényege, hogy az idősorok négy, egymástól elkülöníthető komponensből tevődnek össze:

- a hosszú távú folyamatokat leíró trendből,
- az ettől szabályos ingadozással eltérő szezonális komponensből,
- a többnyire hosszú távú hullámzást kifejező ciklikus komponensből és
- a véletlen összetevőből.

Megnézem a kapcsolódó epizódot

A lineáris trend egyenlete nagyon egyszerű:

$$\hat{y}_t = \hat{b}_0 + \hat{b}_1 \cdot t$$

A $\hat{b}_0$ és $\hat{b}_1$ paramétereket Excelben vagy bármilyen statisztikai programban néhány kattintással megkapjuk.

Ha kézzel szeretnénk őket kiszámolni, akkor pedig ezekre a normálegyenletekre lesz hozzá szükség:

$$\sum_{t=1}^{n} y_t = n \cdot \hat{b}_0 + \hat{b}_1 \sum_{t=1}^{n} t \qquad \sum_{t=1}^{n} t \cdot y_t = \hat{b}_0 \cdot \sum_{t=1}^{n} t + \hat{b}_1 \sum_{t=1}^{n} t^2$$

Megnézem a kapcsolódó epizódot

A szezonalitást úgy kell elképzelni, hogy az minden nyári szezonban ugyanannyit hozzáad, minden téliben pedig ugyanannyit elvesz a trendvonal által meghatározott értékből.

Pl. ha a négy évszakot vesszük, akkor négy szezonunk van, van egy téli, egy tavaszi, egy nyári és egy őszi, ezért négy szezonalitást kell számolnunk. Más idősorok esetében természetesen ez lehet több is és kevesebb is.

A szezonalitás képlete a következő:

$$s_j = \frac{\sum_{i=1}^{n/p} (y_{ij} - \hat{y}_{ij})}{n/p}$$

A képlet roppant barátságos, de némi magyarázatra szorul. Mindössze arról van szó, hogy minden egyes szezonra átlagoljuk a trendvonal és a tényleges értékek közötti eltéréseket.

Vagyis a képletben p a szezontípusok száma, n pedig az összes szezon száma, y_{ij} jelenti a tényleges értéket, ahol az ij-t úgy kell érteni, hogy az i-edik év j-edik szezonja. $\hat{y}_{ij}$ pedig ennek a trend szerinti megfelelője.

Megnézem a kapcsolódó epizódot

Ha összeadjuk a nyers szezonális eltéréseket, és ezek összege nem nulla, akkor vesszük az átlagukat.

$\overline{s} = \frac{s_1+s_2+s_3+s_4}{4}$

És ezt az átlagot mindegyik nyers szezonális eltérésből levonjuk.

$\tilde{s}_1 = s_1 - \overline{s}$

$\tilde{s}_2 = s_2 - \overline{s}$

$\tilde{s}_3 = s_3 - \overline{s}$

$\tilde{s}_4 = s_4 - \overline{s}$

Így kapjuk meg a korrigált szezonális eltéréseket

Megnézem a kapcsolódó epizódot

Minden olyan függvényt, ami az y tengelyre szimmetrikus, páros függvénynek hívunk. Ezek a függvények azt tudják, hogy bármely x-re amelyre értelmezve vannak:

$$f(-x) = f(x)$$

Azokat a függvényeket, amelyek az origóra szimmetrikusak, páratlan függvénynek nevezzük. A páratlan függvények úgy működnek, hogy bármely x-re amelyre értelmezve vannak:

$$f(-x) = -f(x)$$

Megnézem a kapcsolódó epizódot

Ha az x különböző pozitív egész kitevős hatványait összeadjuk vagy kivonjuk, akkor polinomokat kapunk.

A polinomfüggvény általános alakja:

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots a_1 x + a_0$$

A legmagasabb fokú tag együtthatóját hívjuk főegyütthatónak.

Megnézem a kapcsolódó epizódot

A mozgóátlagolás lényege, hogy az idősor egyes elemeit a körülötte lévő elemek átlagával helyettesítjük, kisimítva ezzel az esetleges erős hullámzásokat.

A mozgóátlagok kiszámolásának képlete páros és páratlan tagú átlagok esetén eltérő.

Ha a tagok száma páratlan:

$\hat{y}_t = \frac{y_{t-k}+\cdots+y_{t-1}+y_t+y_{t+1}+\cdots+y_{t+k}}{2k+1}$

Ha pedig a tagok száma páros:

$\hat{y}_t = \frac{\frac{y_{t-k}}{2}+\cdots+y_{t-1}+y_t+y_{t+1}+\cdots+\frac{y_{t+k}}{2}}{2k}$

Megnézem a kapcsolódó epizódot

Jelentősen csökkenthetjük a normálegyenletek által okozott szenvedéseket, ha az idő múlását jelentő t paramétert úgy adjuk meg, hogy az összege éppen nulla legyen.

Ekkor

$$\sum y_t = n \cdot \hat{b}_0 \qquad \sum t \cdot y_t = \hat{b}_1 \sum t^2$$

Megnézem a kapcsolódó epizódot

Becslések

Olyan esetekben, amikor valamiért nem tudjuk vagy nem akarjuk a teljes sokaságot megvizsgálni, hogy meghatározzuk a fontosabb statisztikai mutatóit, becslést alkalmazunk. A becslés lényege, hogy egy minta alapján próbálunk ezekre a mutatókra következtetni.

Megnézem a kapcsolódó epizódot

A megbízhatósági szintet konfidencia szintnek nevezzük. A konfidencia szint szokásos jelölése $1-\alpha$.

Megnézem a kapcsolódó epizódot

Az $1-\alpha$ megbízhatósági szinthez, vagy másként konfidencia szinthez tartozó konfidencia intervallumok azok az intervallumok, amik a sokasági átlagot $1-\alpha$ valószínűséggel tartalmazzák.

A konfidencia intervallum végpontjai:

$\overline{x} \pm Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ ahol

$\overline{x} =$ a minta átlaga

$n =$ a minta elemszáma

$\sigma =$ a teljes sokaság szórása

$Z_{1-\frac{\alpha}{2}}$ pedig a standard normális eloszlás $1-\frac{\alpha}{2}$-höz tartozó Z értéke.

Megnézem a kapcsolódó epizódot

$\overline{x} \pm Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ ahol

$\overline{x} =$ a minta átlaga

$n =$ a minta elemszáma

$\sigma =$ a teljes sokaság szórása

$Z_{1-\frac{\alpha}{2}}$ pedig a standard normális eloszlás $1-\frac{\alpha}{2}$-höz tartozó Z értéke.

Megnézem a kapcsolódó epizódot

A FAE minta azt jelenti, hogy a mintavétel során bármely mintaelemet azonos eséllyel választunk ki. Ilyen a visszatevéses mintavétel, vagy pedig abban az esetben ha az alapsokaság elemszáma nagyon nagy, akkor a visszatevés nélküli mintavétel is.

Megnézem a kapcsolódó epizódot

$\overline{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$ ahol

$1-\alpha =$ konfidencia szint

$\overline{x} =$ a minta átlaga

$n =$ a minta elemszáma

$s =$ a teljes sokaság szórása, a sokasági szórás nem ismert

$t_{1-\frac{\alpha}{2}}$ pedig a t-eloszlás $1-\frac{\alpha}{2}$-höz tartozó értéke.

Megnézem a kapcsolódó epizódot

$\overline{p} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$ ahol

$1-\alpha =$ konfidencia szint

$\overline{p} =$ a minta alapján kapott valószínűség

$n =$ a minta elemszáma

$Z_{1-\frac{\alpha}{2}}$ pedig a standard normális eloszlás $1-\frac{\alpha}{2}$-höz tartozó Z értéke.

Megnézem a kapcsolódó epizódot

$\frac{(n-1)\cdot s^2}{\chi^2_{1-\frac{\alpha}{2}(v)}} < \sigma^2 < \frac{(n-1)\cdot s^2}{\chi^2_{\frac{\alpha}{2}(v)}}$ ahol

$1-\alpha =$ konfidencia szint

$\overline{p} =$ a minta alapján kapott valószínűség

$n =$ a minta elemszáma

$s =$ a minta szórása, a sokasági szórás nem ismert

$\chi^2(v)$ pedig a khi-négyzet eloszlás megfelelő értéke

Megnézem a kapcsolódó epizódot

Az EV-minta abban különbözik a FAE-mintától, hogy a kiválasztott mintaelemek nem függetlenek egymástól.

Ez olyankor fordulhat elő, ha a teljes sokaság mérete viszonylag kicsi a minta elemszámához képest. EV-minták esetén tehát a minta fontos jellemzőjévé válik, hogy mekkora a teljes sokaság, amelynek elemszámát N jelöli.

Megnézem a kapcsolódó epizódot

$\overline{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \cdot \sqrt{1-\frac{n}{N}}$ ahol

$1-\alpha =$ konfidencia szint

$\overline{x} =$ a minta átlaga

$n =$ a minta elemszáma

$N =$ a teljes sokaság elemszáma

$s =$ a minta szórása

$t_{1-\frac{\alpha}{2}}$ pedig a t-eloszlás $1-\frac{\alpha}{2}$-höz tartozó értéke.

Megnézem a kapcsolódó epizódot

$\overline{p} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \cdot \sqrt{1-\frac{n}{N}}$ ahol

$1-\alpha =$ konfidencia szint

$\overline{p} =$ a minta alapján kapott valószínűség

$n =$ a minta elemszáma

$N =$ a teljes sokaság elemszáma

$Z_{1-\frac{\alpha}{2}}$ pedig a standard normális eloszlás $1-\frac{\alpha}{2}$-höz tartozó Z értéke.

Megnézem a kapcsolódó epizódot

Ha a teljes sokaságot felosztjuk viszonylag homogén rétegekre, és a mintát is ezen a rétegek szerint vizsgáljuk, a variancia csökkenthető.

$\hat{\overline{x}}_R \pm Z_{1-\frac{\alpha}{2}} \cdot s_{\hat{\overline{X}}_R}$

$1-\alpha =$ konfidencia szint

$\overline{x} =$ a minta átlaga

$n =$ a minta elemszáma

$n_j =$ a minta j-edik rétegének elemszáma

$N =$ a teljes sokaság elemszáma

$N_j =$ a teljes sokaság j-edik rétegének elemszáma

$W_j =$ a teljes sokaság j-edik rétegének a teljes sokasághoz viszonyított aránya

$s_j =$ a minta j-edik rétegének szórása

$\hat{\overline{X}}_R = \sum_{j=1}^{M} W_j \overline{x}_j$

$s^2_{\hat{\overline{X}}_R} = \sum_{j=1}^{M} W_j^2 \frac{s_j^2}{n_j}\left(1-\frac{n_j}{N_j}\right)$

Megnézem a kapcsolódó epizódot

A kétmintás becslésekre akkor van szükség, amikor kétsokaság valamilyen paraméterét, leginkább az átlagát szeretnénk összehasonlítani.

A kétmintás becslések lehetnek független mintás becslések vagy páros mintás becslések.

Megnézem a kapcsolódó epizódot

Ha mindkét sokaság közel normális eloszlású, akkor az átlagok különbségének becslésére ez a formula van forgalomban.

$d \pm t_{1-\frac{\alpha}{2}} \cdot s_d$ ahol $d = \overline{x} - \overline{y}$

$s_d = s_c \cdot \sqrt{\frac{1}{n_Y} + \frac{1}{n_X}}$ itt $s_c^2 = \frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2}$

$1-\alpha =$ konfidencia szint

$\overline{x} =$ az egyik minta átlaga

$\overline{Y} =$ a másik minta átlaga

$n_X =$ az egyik minta elemszáma

$n_Y =$ a másik minta elemszáma

A szabadságfok $v = n_X + n_Y - 2$

Megnézem a kapcsolódó epizódot

Egy becslést torzítatlannak nevezünk, ha az egyes mintákból kapott becslések várható értéke megegyezik a becsülni kívánt mennyiséggel.

Ez a tulajdonság azt jelenti, hogy a becslés során kapott értékek a becsülni kívánt érték körül ingadoznak, és ez az ingadozás szimmetrikus. A torzítatlan becsléseket mindig előnyben részesítjük a torzítottakkal szemben.

Megnézem a kapcsolódó epizódot

A kérdés az, hogy ha egy sokasági jellemzőre több becslés jöhet szóba, hogyan válasszunk közülük, vagyis mikor tekintünk egy becslést jónak, kettő közül melyiket tekintjük jobbnak és kijelenthetjük-e valamelyikről, hogy a legjobb?

Két alapvető szempont alapján szoktuk a becsléseket versenyeztetni. Az egyik, a már jól ismert torzítatlanság, vagyis a becslésnek az a tulajdonsága, hogy az összes lehetséges mintán vett becslések átlaga megegyezik a becsülni kívánt sokasági jellemzővel. A másik az úgynevezett minimális variancia kritérium.

A minimális variancia kritérium azt jelenti, hogy ha van két torzítatlan becslésünk, akkor a kettő közül azt tekintjük jobbnak, aminek az összes mintán vett értékeinek varianciája kisebb.

Megnézem a kapcsolódó epizódot

$$MSE(\hat{\theta}) = var(\hat{\theta}) + (E(\theta) - \theta)^2$$

Az első tag a varianciát, a második tag a várható értéktől való eltérést, vagyis a torzítottságot méri. Ha a becslés torzítatlan, $E(\hat{\theta}) = \hat{\theta}$ így ez a második tag nulla. Két becslés közül azt részesítjük előnyben, amelyre MSE kisebb.

Az $E(\hat{\theta}) - \theta$ különbségre, vagyis a torzítás mértékére az angol bias szó alapján a$Bs\hat{\theta}$ jelölés van forgalomban. Használatos tehát az

$$MSE(\hat{\theta}) = var(\hat{\theta}) + Bs^2(\hat{\theta})$$

Képlet is.

Megnézem a kapcsolódó epizódot

Eddigi vizsgálódásaink egyik legfontosabb eredménye a mintaátlagot eloszlásának jellemzése. Ha a teljessokaság átlaga μ és szórása pedig σ, akkor az ebből vett n elemű minták átlagai olyan eloszlással helyezkednek el, aminek átlaga szintén μ, a szórása pedig $\frac{\sigma}{\sqrt{n}}$.

Ezt az utóbbit a minta standard hibájának szokás nevezni. A standard hiba tehát azt mondja meg, hogy a mintaátlagok mekkora szórással ingadoznak a tényleges sokasági átlag körül.

Megnézem a kapcsolódó epizódot

Mintavételi hibának azokat a hibákat nevezzük, amik kimondottan azért fordulnak elő, mert nem tudjuk, vagy nem akarjuk a teljes sokaságot vizsgálni. A mintavételi hiba tehát a sokaság eloszlásán és a mintavételi eljáráson kívül főleg a minta elemszáma határozza meg. Mivel pedig ezeket általában már a mintavételt megelőzően ismerjük, a mintavételi hibának megvan az a kellemes tulajdonsága, hogy legtöbbször előre megállapítható. Vagyis még el sem végeztük a mintavételt, de már tudjuk, hogy mekkora lesz a mintavétel során elkövetett hiba. Ez a kellemes tulajdonság lesz a kiindulópont a becslések és később a hipotizésvizsgálatok elméletének kiépítésében.

Megnézem a kapcsolódó epizódot

Hipotézisvizsgálat

Az elfogadási tartomány az a tartomány, ahová ha a próba értéke kerül, akkor a nullhipotézist elfogadjuk.

Megnézem a kapcsolódó epizódot

A kritikus tartomány az a tartomány, ahová ha a próba értéke kerül, akkor a nullhipotézist elvetjük.

Megnézem a kapcsolódó epizódot

A szignifikanciaszint a hibás döntés valószínűsége.

Megnézem a kapcsolódó epizódot

ELSŐ LÉPÉS: A HIPOTÉZIS MEGFOGALMAZÁSA

Minden hipotézisvizsgálat két egymásnak ellentmondó felvetés felírásával kezdődik. Az egyiket nullhipotézisnek nevezzük és H_0-al jelöljük, a másikat pedig ellenhipotézisnek és jele H_1.

MÁSODIK LÉPÉS: A PRÓBAFÜGGVÉNY KIVÁLASZTÁSA

A próbafüggvények kiválasztása magától a hipotézistől, illetve a mintavétel módjától is függ.

HARMADIK LÉPÉS: SZIGNIFIKANCIASZINT ÉS KRITIKUS TARTOMÁNY

Ha a próbafüggvény értéke az elfogadási tartományba fog esni, akkor ezt a tényt a nullhipotézist igazoló jelnek fogjuk tekinteni. Hogyha pedig a kritikus tartományba, akkor a nullhipotézist elvetjük.

NEGYEDIK LÉPÉS: MINTAVÉTEL ÉS DÖNTÉS

Ha a mintavétellel kapott eredményünk szerint a próbafüggvény az elfogadási tartományba esik, akkor a H_0 nullhipotézist tekintjük igaznak, a H_1 ellenhipotézist pedig elvetjük.

Ha viszont a próbafüggvény a minta alapján a kritikus tartományba esik, akkor a H_0 nullhipotézist vetjük el és a H_1 ellenhipotézist tekintjük igaznak.

Megnézem a kapcsolódó epizódot

A sokaság normális eloszlású, szórása σ, H_0 a sokaság átlagára vonatkozik, a minta elemszáma n.

$$Z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Megnézem a kapcsolódó epizódot

A sokaság normális eloszlású, szórása nem ismert, H_0 a sokaság átlagára vonatkozik, a minta elemszáma n.

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Megnézem a kapcsolódó epizódot

A sokaság tetszőleges eloszlású, szórása nem ismert, H_0 a sokaság átlagára vonatkozik, a minta n elemű, elemszáma nagy.

$$Z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Megnézem a kapcsolódó epizódot

A sokaság tetszőleges eloszlású, H_0 a sokasági arányra vonatkozik, a minta n elemű, elemszáma nagy.

$$Z = \frac{P - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

Megnézem a kapcsolódó epizódot

A sokaság normális eloszlású, H_0 a sokasági szórásra vonatkozik, a minta n elemű.

$$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma_0^2}$$

Megnézem a kapcsolódó epizódot

A sokaság eloszlására irányuló vizsgálat.

H_0: mindegyik osztályköz valószínűsége egy adott eloszlásnak megfelelő érték, vagyis minden i-re az i-edik osztályköz valószínűsége a P_i érték.

Az ellenhipotézis pedig, H_1: van olyan osztályköz, ami nem az adott eloszlásnak megfelelő P_i érték. A próbát $\chi^2_{1-\alpha}(v)$ jobb oldali kritikus értékkel végezzük el, a nullhipotézist az ennél kisebb, az ellenhipotézist az ennél nagyobb értékek igazolják. A minta elemszáma n.

$$\chi^2(v) = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{nP_i}$$

ahol a v szabadságfok: $v = k - b - 1$.

Itt $k =$ az osztályközök száma és $b =$ az adott eloszlás azon paramétereinek száma, amit a mintából becsléssel határozunk meg.

Megnézem a kapcsolódó epizódot

A sokaságon belül két ismérv függetlenségére irányuló vizsgálat. H_0: a két ismérv független, az ellenhipotézis pedig, H_1: a két ismérv közti kapcsolat sztochasztikus vagy függvényszerű.

A próbát $\chi^2_{1-\alpha}(v)$ jobb oldali kritikus értékkel végezzük el, a nullhipotézist az ennél kisebb, az ellenhipotézist az ennél nagyobb értékek igazolják. A minta elemszáma n, a minta alapján készített kontingencia tábla sorainak száma r, oszlopainak száma c.

$$\chi^2(v) = \sum \frac{(n_{ij}-n^*_{ij})^2}{n^*_{ij}}$$

ahol a v szabadságfok $v = (r-1)(c-1)$.

Megnézem a kapcsolódó epizódot

Két sokaságban valamely változó eloszlásának egyezőségére irányuló vizsgálat. H_0: a két sokaságban az eloszlás egyező, az ellenhipotézis pedig, H_1: a két eloszlás nem egyező.

A próbát $\chi^2_{1-\alpha}(v)$ jobb oldali kritikus értékkel végezzük el, a nullhipotézist az ennél kisebb, az ellenhipotézist az ennél nagyobb értékek igazolják. Mintát ezúttal mindkét sokaságból veszünk, az X sokaságból vett minta elemszáma n_X az Y sokaságból vett mintáé n_Y mindkét mintában az osztályközök száma k.

\(\chi^2(v) = n_X \cdot n_Y \cdot \sum_{i=1}^{k} \frac{1}{n_{Xi}+n_{Yi}} \cdot \left(\frac{n_{Xi}}{n_X} - \frac{n_{Yi}}{n_Y} \right)^2 \)

ahol a v szabadságfok $v = k-1$.

Megnézem a kapcsolódó epizódot

Mindkét sokaság normális eloszlású, szórásaik σ_X és σ_Y.

$$Z = \frac{(\overline{y}-\overline{x})-\delta_0}{\sqrt{\frac{\sigma_Y^2}{n_Y}+\frac{\sigma_X^2}{n_X}}}$$

A nullhipotézis: H_0: $\mu_X - \mu_Y = \delta_0$, ahol δ tetszőleges, de előre megadott érték. A minták elemszáma n_X és n_Y.

Megnézem a kapcsolódó epizódot

A két sokaság normális eloszlású és szórásaik egyformák.

$$t(v) = \frac{(\overline{y}-\overline{x})-\delta_0}{s\cdot\sqrt{\frac{1}{n_Y}+\frac{1}{n_X}}}$$

itt $s^2 = \frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2}$

A nullhipotézis H_0: $\mu_X - \mu_Y = \delta_0$, ahol δ tetszőleges, de előre megadott érték.

A minták elemszáma n_X és n_Y, szórása s_X és s_Y, a szabadságfok $v = n_Y + n_X - 2$

Megnézem a kapcsolódó epizódot

A két sokaság eloszlása és szórása nem ismert, mindkettő szórása véges, és mindkét minta elemszáma elég nagy.

$$Z = \frac{(\overline{y}-\overline{x})-\delta_0}{\sqrt{\frac{s_Y^2}{n_Y}+\frac{s_X^2}{n_X}}}$$

A nullhipotézis $H_0: \mu_X - \mu_Y = \delta_0$, ahol δ tetszőleges, de előre megadott érték.

A minták elemszáma n_X és n_Y, szórása s_X és s_Y.

Megnézem a kapcsolódó epizódot

Két sokaság szórásának összehasonlítására irányuló próba, ha mindkét sokaság normális eloszlású. A nullhipotézis $H_0: \sigma_1^2 = \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} \qquad F_{1-p}(v1;v2) = \frac{1}{F_p(v_2;v_1)}$$

Az F-eloszlás két szabadságfoka

$v_1 = n_1 - 1$ és $v_2 = n_2 - 1$

Bal oldali kritikus érték: $\frac{1}{F_{1-\alpha}(v_2;v_1)}$

Jobb oldali kritikus érték: $F_{1-\alpha}(v_1;v_2)$

Kétoldali kritikus érték:

$\frac{1}{F_{1-\frac{\alpha}{2}}(v_2;v_1)}$ és $F_{1-\frac{\alpha}{2}}(v_1;v_2)$

Megnézem a kapcsolódó epizódot

Több sokaság várható értékének összehasonlítására vonatkozó próba, ha mindegyik sokaság normális eloszlású és azonos szórású.

A H_0 nullhipotézis: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_M = \mu$, vagyis az, hogy a várható értékek az összes sokaságra (M db) megegyeznek, míg az ellenhipotézis az, hogy van olyan μ_j amire $\mu_j \neq \mu$.

Megnézem a kapcsolódó epizódot

A Bartlett-próba több sokaság szórásának összehasonlítására vonatkozó próba, ha mindegyik sokaság normális eloszlású.

A H_0 nullhipotézis: $\sigma_1 = \sigma_2 = \sigma_3 = \cdots = \sigma_M = \sigma$, vagyis az, hogy az összes sokaság (M db.) szórása megegyezik, míg az ellenhipotézis az, hogy van olyan σ_j, amire $\sigma_j \neq \sigma$.

$$SSB = \sum_{j=1}^{M}(n_j - 1)s_j^2 \qquad s_b = \frac{SSB}{n-M}$$

A próbafüggvény

$$B^2 = \frac{1}{c}\left(v \cdot \ln s_b^2 - \sum_{j=1}^{M} v_j \ln s_j^2\right)$$

$$c = 1 + \frac{1}{3(M-1)}\left(\sum_{j=1}^{M}\frac{1}{v_j} - \frac{1}{v}\right)$$

Jobb oldali kritikus érték: $\chi^2_{1-\alpha}(M-1)$

Megnézem a kapcsolódó epizódot

Regressziószámítás

A regresszió egyenes egyenlete:

$$y = b_0 + b_1 \cdot x$$

Ahol $b_1 = \frac{\sum dx \cdot dy}{\sum d^2x}$ és $b_0 = \overline{y} - b_1 \cdot \overline{x}$

A regressziós egyenes egyenletében szereplő regressziós paraméterek közül b_1 az egyenes meredeksége. A b_0 érték kevésbé jelentős, ez azt adja meg, hogy a magyarázó változó nulla értékéhez milyen y érték tartozik.

Megnézem a kapcsolódó epizódot

A regressziós egyenes egyenlete $\hat{y} = \hat{b_0} + \hat{b_1} \cdot x$

Ez egy lineáris függvény, ami mindegyik x-hez hozzárendel valamilyen y-t. Ezek általánan eltérnek a valódi y-októl. Ezeket az eltéréseket reziduumoknak nevezzük.

Megnézem a kapcsolódó epizódot

A reziduumokból képzett mutató az úgynevezett SSE, jelentése sum of squares of the errors vagyis eltérés-négyzetösszeg.

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

Ha a regresszió tökéletesen illeszkedik, akkor az $e_i = y_i - \hat{y}_i$ különbségek mindegyike 0, így SSE=0. Ha az illeszkedés nem tökéletes, akkor SSE egy pozitív érték, ami az illeszkedés pontatlanságát méri.

Megnézem a kapcsolódó epizódot

Ha az SSE értékeit elosztjuk a megfigyelt pontok számával és a kapott eredménynek vesszük a gyökét, akkor kapjuk a reziduális szórást:

$$s_e^* = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{\sum e_i^2}{n}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

Megnézem a kapcsolódó epizódot

Az illeszkedés egy mérőszáma a lineáris korrelációs együttható:

$$r = \frac{\sum dx \cdot dy}{\sqrt{\sum d^2x \cdot \sum d^2y}}$$

A lineáris korrelációs együttható azt méri, hogy x és y között milyen szoros lineáris kapcsolat van. Értéke mindig $-1 \leq r \leq 1$.

Megnézem a kapcsolódó epizódot

A magyarázóerőt méri az úgynevezett determinációs együttható, melynek jele R^2. Ez a kétváltozós lineáris modell esetében megegyezik r^2-tel.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Itt SSE az eltérés-négyzetösszeg, míg SSR az úgynevezett regressziós, vagy magyarázó négyzetösszeg, SST pedig a teljes négyzetösszeg, a köztük lévő kapcsolat pedig...

$$SST = \sum d^2y \quad SSR = \sum(\hat{y}_i - \overline{\hat{y}})^2 = b_1^2 \sum d^2x \quad SSE = \sum(y_i - \hat{y}_i)^2 = \sum e_i^2$$

Megnézem a kapcsolódó epizódot

A regresszió egyenes egyenlete:

$$\hat{y} = \hat{b}_0 \cdot x^{\hat{b}_1}$$

Amiből

$$\lg \hat{y} = \lg \hat{b}_0 + \hat{b}_1 \cdot \lg x$$

Ahol $\hat{b}_1 = \frac{\sum d\lg x \cdot d\lg y}{\sum d^2 \lg x}$ és $\lg \hat{b}_0 = \overline{\lg y} - \overline{\lg x} \cdot \hat{b}_1$

Megnézem a kapcsolódó epizódot

A regresszió egyenes egyenlete:

$$\hat{y} = \hat{b}_0 \cdot \hat{b}_1^x$$

Amiből

$$\lg \hat{y} = \lg \hat{b}_0 + x \cdot \hat{b}_1$$

Ahol $\lg \hat{b}_1 = \frac{\sum dx \cdot d\lg y}{\sum d^2 x}$ és $\lg \hat{b}_0 = \overline{\lg y} - \overline{x} \cdot \lg \hat{b}_1$

Megnézem a kapcsolódó epizódot

Az elaszticitás két összefüggő jelenség közti kapcsolat.

Lineáris modellben: $El(\hat{y}, x) = \frac{\hat{b}_1 x}{\hat{y}} = \frac{\hat{b}_1 x}{\hat{b}_0 + \hat{b}_1 x}$

Hatványkitevős modellben: $El(\hat{y}, x) = \hat{b}_1$

Exponenciális modellben: $El(\hat{y}, x) = x \cdot \ln \hat{b}_1$

Megnézem a kapcsolódó epizódot

I. A magyarázó változók nem valószínűségi változók.

II. A magyarázó változók lineárisan független rendszert alkotnak.

III. Az eredményváltozó közel lineáris függvénye a magyarázó változónak.

IV. Az ϵ hibatag feltételes eloszlása normális, várható értéke nulla.

V. Az ϵ hibatag különböző x-ekhez tartozó értékei korrelálatlanok.

Megnézem a kapcsolódó epizódot

A paraméterek becslése:

$$\hat{b}_i \pm t_{1-\frac{\alpha}{2}} \cdot (n-k-1) \cdot s_{\hat{b}_i}$$

A regresszió becslése:

$$\hat{y}_* \pm t_{1-\frac{\alpha}{2}} \cdot (n-k-1) \cdot s_{\hat{y}_*}$$

Megnézem a kapcsolódó epizódot

A többváltozós regressziós modelleket olyankor alkalmazzuk, amikor az eredményváltozó alakulását több magyarázó változó tükrében vizsgáljuk.

A többváltozós lineáris regresszió egyenlete:

$$y = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_k x_k + \epsilon$$

Az y eredményváltozó itt k darab magyarázó változótól és a hibatagtól függ.

A képletben a $\hat{b}_0$ paraméter a tengelymetszet, a többi $\hat{b}_i$ paraméter pedig azt jelenti, hogy az i-edik magyarázó változó egy egységgel történő változása, mennyivel változtatja az $\hat{y}$ értéket, ha a többi magyarázó változót rögzítjük.

Megnézem a kapcsolódó epizódot

A kétváltozós esethez hasonlóan a korreláció itt is a változók közti kapcsolat szorosságát írja le, csakhogy itt egy fokkal rosszabb a helyzet, ugyanis most bármely két változó korrelációját vizsgálhatjuk. Ezt tartalmazza a korrelációmátrix.

$$R=\begin{pmatrix} 1 & r_{y1} & r_{y2} & \dots & r_{yk} \\ r_{1y} & 1 & r_{12} & \dots & r_{1k} \\ r_{2y} & r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{ky} & r_{k1} & r_{k2} & \dots & 1 \end{pmatrix}$$

Itt r_{ij} az x_i és az x_j magyarázó változók közti korrelációt írja le, tehát például r_{12} az x_1 és az x_2 közti korrelációt jelenti.

r_{iy} pedig az x_i magyarázó változó és az y eredményváltozó közti kapcsolatot jelenti.

Mivel $r_{ij}=r_{ji}$ a korreláció-mátrix szimmetrikus. Az áttekinthetőbb felírás kedvéért a felső háromszöget el is szokták hagyni.

Megnézem a kapcsolódó epizódot

A lineáris regresszió egyenlete: $\hat{y}=\hat{b}_0+\hat{b}_1x_1+\hat{b}_2x_2+\dots+\hat{b}_kx_k$

A tesztelés úgy zajlik, hogy nullhipotézisnek tekintjük a $H_0:b_i=0$ feltevést, ellenhipotézisnek pedig azt, hogy $H_1:b_i\neq 0$.

A nullhipotézis azt állítja, hogy a modellben a b_i paraméter szignifikánsan nulla, vagyis az i-edik magyarázó változó felesleges, annak hatása az eredményváltozóra nulla. Az ellenhipotézis ezzel szemben az, hogy $b_i\neq 0$ vagyis az i-edik magyarázó változónak a regresszióban nem nulla hatása van.

Megnézem a kapcsolódó epizódot

Szóródás oka	Négyzetösszeg	Szabadságfok	Átlagos négyzetösszeg	F
Regresszió	SSR	k	$MSR=\frac{SSR}{k}$	$F=\frac{MSR}{MSE}$
Hiba	SSE	$n-k-1$	$MSE=\frac{SSE}{n-k-1}$	
Teljes	SST	$n-1$		

Megnézem a kapcsolódó epizódot

A multikollinearitás röviden összefoglalva azt jelenti, hogy két vagy több magyarázó változó között túl szoros korrelációs kapcsolat van, és ez zavarja a becslést.

A multikollinearitás mérésére az úgynevezett VIF (variance inflator factor) variancia növelő faktor van forgalomban.

$$VIF_j=\frac{1}{1-R_j^2}$$

A képletben szereplő R_j^2 a j-edik magyarázó változó és az összes többi magyarázó változó közti determinációs együttható.

Megnézem a kapcsolódó epizódot

Az autokorreláció a regresszió maradéktagjának a saját későbbi értékeivel való korrelációját jelenti, vagyis egyfajta szabályszerűséget a maradékváltozóban. Ideális esetben a maradéktagnak véletlenszerűnek kell lennie, bármiféle szabályszerűségért a magyarázó változók felelnek a regresszióban.

Az autókorreláció tesztelésére a Durbin-Wattson-tesztet használjuk.

Megnézem a kapcsolódó epizódot

A Durbin-Wattson-teszt lényegében egy hipotizésvizsgálat, aminek részletezésére nem térünk ki, mindössze a használatát nézzük meg.

Maga a próbafüggvény

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2}$$

A szignifikanciaszint α, a próba elvégzése pedig az alábbi módon történik:

d_L és d_U értékeket kikeressük a táblázatból,

n = a megfigyelések száma,

k = a magyarázó változók száma,

végül megnézzük, hogy a próbafüggvény melyik tartományba esik.

Megnézem a kapcsolódó epizódot
