Statisztika 1 előadás

Baran Sándor

A tananyag elkészítését az EFOP-3.4.3-16-2016-00021 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

Irodalom

- Hunyadi László., Vita László: *Statisztika I.* Akadémiai Kiadó, Budapest, 2018. Online verzió (2019): `https://mersz.hu/hunyadi-vita-statisztika-i`
- Hunyadi László, Vita László: *Statisztika II.* Akadémiai Kiadó, Budapest, 2018. Online verzió (2019): `https://mersz.hu/hunyadi-vita-statisztika-ii`
- Keresztély Tibor, Sugár András, Szarvas Beatrix: *Statisztika közgazdászoknak. Példatár és feladatgyűjtemény.* Nemzeti Tankönyvkiadó, Budapest, 2005.
- Fazekas István: *Valószínűségszámítás.* Egyetemi Kiadó, Debrecen, 2009.
- Denkinger Géza: *Valószínűségszámítás gyakorlatok.* Nemzeti Tankönyvkiadó, Budapest, 2008.

Tartalom

1. Alapfogalmak
2. Sokaság egy ismérv szerinti leírása
3. Sokaság több ismérv szerinti leírása
4. Összehasonlítás standardizálással és indexszámítással
5. Mintavétel
6. Pontbecslések és tulajdonságaik
7. Nagy számok törvényei
8. Intervallumbecslések
9. Hipotézisvizsgálat

Mi a statisztika?

A statisztika olyan *gyakorlati tevékenység*, illetve *tudományos módszertan*, amely arra szolgál, hogy a valóság tényeinek nagy tömegét tömören, számszerűen jellemezze.

- **Gyakorlati tevékenység**: alapadatokat gyűjt, feldolgoz, elemez, majd közzéteszi ezek eredményét.
- **Tudományos módszertan**: az elemzéshez szükséges megfontolások, eljárások megadása.

A statisztika mindig a tények valamilyen nagy – esetleg végtelen nagy – tömegéről igyekszik tömör, számszerű képet adni.

Példa

Az alkalmazásban állók létszáma a nemzetgazdaságban 2022. május: 3232.4 ezer fő.
Pénzügyi, biztosítási tevékenység: 63.8 ezer fő.

A teljes munkaidőben alkalmazásban állók havi bruttó átlakeresete a nemzetgazdaságban 2022. május: 495 863 Ft.
Pénzügyi, biztosítási tevékenység: 848 688 Ft. *Forrás: KSH*

A valóság jellemezni kívánt tényei bizonyos *egységekhez* köthetőek.

Példa

- Az alkalmazásban állók.
- Adott időszakban Magyarországra érkező külföldiek.

A vizsgálat tárgyának egyedeiről szerzett, megfelelő módon rögzített különféle információkat alapadatoknak, más néven elemi adatoknak nevezzük. Az alapadatok nem feltétlenül számszerűek. A vizsgált egységek bizonyos körét összességében jellemző számszerű információkat általánosságban adatoknak, bizonyos speciális esetekben pedig mutatószámoknak hívjuk. A *mutatószám* elnevezés többnyire a szabványosított tartalmú, egy-egy jelenség jellemzésére visszatérően használt számszerű információk megjelölésére szolgál.

A továbbiakban adatok és mutatószámok helyett csak adatokat fogunk emlegetni.

Példák

Alapadatok

- A Magyarországra érkező külföldiek nemzetisége.
- A Magyarországra érkező külföldiek életkora.
- A Magyarországra érkező külföldiek tartózkodási ideje.

Adatok

- Egy adott időszak alatt Magyarországra érkező külföldiek száma (ezer fő)
  2018: 57 667; 2019: 61 397; 2020: 31 641; 2021: 36 688.
- Egy adott időszak alatt Magyarországra érkező külföldiek összes pénzköltése (millió Ft).
  2018: 2 066 780; 2019: 2 310 110; 2020: 1 054 342; 2021: 1 345 559.

Mutatók

- Egy idelátogató külföldi napi átlagos pénzköltése (ezer Ft).
  2018: 15.9; 2019: 16.7; 2020: 14.9; 2021: 15.7.
- Az vendégéjszakák átlagos száma (éjszaka). 2018: 2.3; 2019: 2.3; 2020: 2.2; 2021: 2.3.

Fontosak a mértékegységek!

Sokaságok

A vizsgálat tárgyát képező egységek összességét, halmazát statisztikai sokaságnak, röviden sokaságnak (populációnak) nevezzük. A sokaság egységei különféle tulajdonságaik megadásával jellemezhetők. E tulajdonságok egy része a sokaság minden egységére nézve közös, más része azonban nem.

Egy sokaság megadható:

- egységeinek felsorolásával;
- eloszlásával.

Sokaságok típusai, I.:

- *Diszkrét*, például:
  - a magyar népesség 2022. január 1-én (9 689 010 fő);
  - 2021-ben Magyarországra érkező külföldiek száma (36 688 ezer fő).
- *Folytonos*, önkényesen elkülöníthető egységek, például:
  - 2021 teljes búzatermése (5 316 074 tonna);
  - a belföldön közúton szállított áruk mennyisége 2021-ben (184 218 tonna).
- *Fiktív*, valamilyen eloszlással megadott, például:
  - 2022 lehetséges búzatermés eredményei.

Sokaságok

Sokaságok típusai, II.:

- *Álló*, azaz valamely időpontra vonatkozik (stock), például:
  - a magyar népesség 2022. január 1-én;
  - az IK beiratkozott hallgatói 2022. szeptember 5-én.
- *Mozgó*, azaz valamely időtartamra értendő (flow), például:
  - 2021 teljes búzatermése;
  - a belföldön közúton szállított áruk mennyisége 2021-ben;
  - az IK hallgatói által a 2021/22 tanév 2. félévének szorgalmi időszakában elfogyasztott sör mennyisége.

Sokaságok típusai, III.:

- *Véges*, például:
  - a magyar népesség 2022. január 1-én.
- *Végtelen*, például:
  - 2022 lehetséges búzatermés eredményei.

Sokaságok

Aggergált sokaság: különféle dolgokból elfogyasztott/felhasznált termékek vagy szolgáltatások összértéke. Például:

- Magyarország teljes exportja 2021-ben (42 781.5 milliárd forint);
- az IK hallgatói által a 2021/22 tanév 2. félévének szorgalmi időszakában elfogyasztott alkoholtartalmú italok összértéke.

Az aggregált sokaság nagysága (*aggregátum*):

$$A = \sum_{i=1}^{n} q_i p_i = \sum_{i=1}^{n} \nu_i$$

q_i: az i-edik fajta egységeinek mennyisége valamilyen alkalmas mértékegységben;

p_i: az i-edik fajta egység egységára;

ν_i: az i-edik fajta egységek összértéke;

n: az egységek száma.

Ismérvek

Az ismérvek olyan vizsgálati szempontok, melyek alapján a sokaság részekre bontható. A sokaság egységeinek valamely adott szempont szerint lehetséges tulajdonságait ismérvváltozatoknak nevezzük.

Ha számszerűek az ismérvváltozatok, akkor ezeket ismérvértékeknek, magát az ismévet pedig változónak nevezzük.

Az ismérvek fajtái

- *területi,* például lakhely, születési hely;
- *időbeli,* például születési idő, munkába állás időpontja;
- *minőségi,* például nem, foglalkozás;
- *mennyiségi,* például életkor, testmagasság, testtömeg, tanulmányi átlag.

Mérési skálák

Ismérvváltozatok átkódolhatóak számokká. Csak olyan műveletek megengedettek, amik az eredeti változatokkal is.

Mérési szintek

- *Nominális*: csak az vizsgálható, két érték egyenlő-e, például név, lakhely, foglalkozás. Nincs mértékegysége.
- *Ordinális*: csak az értékek sorendje számít, távolsága nem, például vizsgajegyek, végzettség. Nincs mértékegysége.
- *Különbségi*: az értékek különbsége is információt hordoz, de az arányuk nem, például hőmérséklet. A skála kezdőpontja önkényes (Celsius, Kelvin, Fahrenheit fok), van mértékegysége.
- *Arány*: kezdőpont egyértelműen adott, az arány is értelmezhető, például havi jövedelem, testmagasság. Van mértékegysége.

Példa

Sokaság	Egy konkrét egység	Ismérv	Ismérvváltozat	Ismérvfajta	Mérési skála
A Mordorba irányuló idegenforgalom a harmadkor 3018. évében	Zsákos Frodó	állampolgárság	megyelakó	minőségi	nominális
		tartózkodási idő (nap)	7	mennyiségi	arány
		életkor (év)	50	mennyiségi	arány
		útitársak száma	2	mennyiségi	arány
		igénybe vett szállás	szabad ég alatt	minőségi	nominális
A Galaktikus Birodalom népessége YU[1] 4-ben	Luke Skywalker	faj	ember	minőségi	nominális
		nem	férfi	minőségi (alternatív)	nominális
		születési hely	Polis Massa bolygó	területi	nominális
		születési idő	YE[1] 19	időbeli	intervallum
		anyabolygó	Tatuin	területi	nominális
		életkor (év)	24	mennyiségi	arány
		foglalkozás	Jedi lovag	minőségi	nominális

[1] YE, illetve YU: a yavini csata (a Halálcsillag megsemmisítése) előtt, illetve után.

Hibák

A statisztikai adatok, mutatószámok célszerű megadási módja:

$$A \pm a$$

A: közelítő érték.

a: abszolút hibakorlát.

$$A - a \leq \text{valódi érték} \leq A + a$$

Relatív hibakorlát: $\alpha = a/A$.

Példa. Magyarország népessége 2022. január 1-én: 9 689 ezer fő.

$A = 9689,\ a = 0.5,\ \alpha = 0.5/9689 \approx 0.00005 = 0.005\%$.

Két közelítő érték összegének vagy különbségének abszolút hibakorlátja a megfelelő *abszolút* hibakorlátok *összege.* Két közelítő érték szorzatának vagy hányadosának *relatív* hibakorlátja nagyjából a megfelelő *relatív* hibakorlátok összege.

Statisztikai alapműveletek

I. A sokaság jellemzése valamely erre alkalmas adattal vagy mutatószámmal.

A sokasághoz hozzárendelünk egy annak egészét jellemző adatot, például a nagyságát, átlagát, várható értékét.

II. Összehasonlítás.

Egy adott jelenség időbeli alakulásáról, területi eltéréseiről, vagy egymáshoz valamilyen módon kapcsolódó jelenségek viszonyáról ad számszerű információt. Fontos, hogy az adatok *összehasonlíthatóak* legyenek.

Az adatokból képezhetünk *különbségeket* vagy *hányadosokat*.

Az alkalmazásban állók havi bruttó átlagkeresete									
Év	2013	2014	2015	2016	2017	2018	2019	2020	2021
Kereset (Ft)	230714	237695	247924	263171	297017	329943	367833	403616	438814
Előző év=100%	103.4	103.0	104.3	106.1	112.9	111.3	111.4	109.7	108.7

Több sokaság adatainak összehasonlítása

A sokaságok viszonya egymáshoz	**A sokaságok adatainak**		**A hányados mértékegysége**
	felsorolására	**hányadosára**	
	használt elnevezés		
Időben és/vagy térben különböző sokaságok	összehasonlító sor (idősor/területi sor)	összehasonlító viszonyszám (dinamikus viszonyszám/ területi összehasonlító viszonyszám)	–, illetve %
Időben és/vagy térben különböző aggregált sokaságok	összehasonlító sor (idősor/területi sor)	index(szám) (területi/időbeli)	–, illetve %
Időben és/vagy térben azonos, de különböző fajta egységekből álló sokaságok	–	intenzitási viszonyszám	a két adat mértékegységének hányadosa

Hunyadi, Vita (2018, 1.4 táblázat)

Példa

Sor-szám	Megnevezés	Mérték-egység	2005	2006	Dinamikus viszonyszám (2005=100)
1.	Alkalmazottak évi átlagos száma	fő	307	236	76.9
2.	Ebből: fizikai foglalkozású	fő	261	208	79.7
3.	Feldolgozott cukorrépa	1000 t	650	475	73.1
4.	Cukortermelés	1000 t	85	70	82.4
5.	Fizikai foglalkozásúak által teljesített munkaórák száma	1000 h	520	360	69.2

Hunyadi, Vita (2018, 1.5 táblázat)

Intenzitási viszonyszámok

- Termelékenység 2005: $\frac{650 \text{ ezer t}}{520 \text{ ezer h}} = 1.25 \text{ t/h}$.
- Egy fizikai dolgozóra eső munkaórák száma 2005, illetve 2006:
  $\frac{520 \text{ ezer h}}{261 \text{ fő}} = 1992.3 \text{ h/fő}$, illetve $\frac{360 \text{ ezer h}}{208 \text{ fő}} = 1730.8 \text{ h/fő}$

Dinamikus viszonyszám

- Egy fizikai dolgozó munkaóráinak változása: $\frac{1730.8 \text{ h/fő}}{1992.3 \text{ h/fő}} = 0.8687 = 86.87\%$

Statisztikai alapműveletek

III. Osztályozás.

Valamely adott sokaság egy vagy több ismérv szerinti tagolása, osztályozása. A csoportok valamilyen szempontból homogénebbek, mint az egész sokaság.

Osztályok: az osztályozás során kapott csoportok.

Csoportképző ismérv(ek): az osztályok elhatárolására szolgáló ismérv(ek).

Példa. Az évfolyam osztályozása a *Mikroökonómia* jegyei alapján.

Elvárások egy osztályozással szemben:

- legyen teljes;
- legyen átfedésmentes;
- eredményezzen homogén osztályokat.

Nómenklatúra: szabványosított osztályozási rendszer.

Foglalkozások Egységes Osztályozási Rendszere (FEOR'08). 10 főcsoport, 42 csoport, 136 alcsoport, 632 foglalkozás.

Csoportosító sor

Osztály	Egységek száma
C_1	f_1
C_2	f_2
$\vdots$	$\vdots$
C_i	f_i
$\vdots$	$\vdots$
C_k	f_k
Összesen	N

C_i: az i-edik osztály azonosítója $(i = 1, 2, \ldots, k)$;

f_i: az i-edik osztály gyakorisága;

k: az osztályok száma, általában k a legkisebb egész, melyre $2^k \geq N$;

N: a sokaság nagysága. $N = \sum_{i=1}^{k} f_i$.

Osztály másik elnevezése: részsokaság. Akkor használjuk, ha az osztályokat külön is tovább akarjuk vizsgálni.

Kombinációs (kontingencia) tábla

Az X ismérv szerinti osztályok	Az Y ismérv szerinti osztályok						$\sum j$
	C_1^Y	C_2^Y	$\dots$	C_j^Y	$\dots$	C_c^Y	
C_1^x	f_{11}	f_{12}	$\dots$	f_{1j}	$\dots$	f_{1c}	$f_{1\cdot}$
C_2^x	f_{21}	f_{22}	$\dots$	f_{2j}	$\dots$	f_{2c}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
C_i^x	f_{i1}	f_{i2}	$\dots$	f_{ij}	$\dots$	f_{ic}	$f_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
C_r^x	f_{r1}	f_{r2}	$\dots$	f_{rj}	$\dots$	f_{rc}	$f_{r\cdot}$
$\sum i$	$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot j}$	$\dots$	$f_{\cdot c}$	N

C_i^X: az X szerinti i-edik osztály azonosítója ($i = 1, 2, \dots, r$); r: az osztályok száma;

C_j^Y: az Y szerinti j-edik osztály azonosítója ($j = 1, 2, \dots, c$); c: az osztályok száma;

f_{ij}: azon elemek száma, melyek mind C_i^X, mind pedig C_j^Y elemei;

$$\sum_{j=1}^{c} f_{ij} = f_{i\cdot}, \quad \sum_{i=1}^{r} f_{ij} = f_{\cdot j}, \qquad \sum_{j=1}^{c} f_{\cdot j} = \sum_{i=1}^{r} f_{i\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} = N.$$

Viszonyszámok

A viszonyszám két adat hányadosa:

$$V = A/B.$$

V: viszonyszám;

A: a viszonyítás tárgya;

B: a viszonyítás alapja.

$$A = B \cdot V, \qquad B = A/V.$$

Típusai

- Dinamikus: idősorok adataiból számított hányadosok.
- Intenzitási: két egymással kapcsolatban lévő, de nem feltétlenül azonos fajta egységekből álló sokaság nagyságából képzett hányadosok.
- Megoszlási: valamely sokaságrésznek az egészhez viszonyított nagyságát mutatja.

Példa

A magyar lakásállomány megoszlása adott év január 1-én

Szobák száma	1990	2010	2022
1	645 064	525 228	458 437
2	1 680 918	1 739 538	1 696 221
3 és több	1 527 306	2 065 915	2 364 613
Összesen	3 853 288	4 330 681	4 519 271

Forrás: KSH

Szobák száma	Százalékos megoszlás			2010. évi állomány (1990=100)	2022. évi állomány (1990=100)	2022. évi állomány (2010=100)
	1990	2010	2022			
1	16.74	12.13	10.15	81.42	71.07	87.28
2	43.62	40.17	37.53	103.49	100.91	97.51
3 és több	39.64	47.70	52.32	135.27	154.82	114.46
Összesen	100.00	100.00	100.00	112.39	117.28	104.35

Dinamikus viszonyszámok kettőnél több adat esetén

$Y_1, Y_2, \ldots, Y_t, \ldots, Y_n$: az idősor adatai.

Bázisviszonyszám: $b_t = Y_t/Y_b, \qquad t = 1, 2, \ldots, n.$

Láncviszonyszám: $\ell_t = Y_t/Y_{t-1}, \qquad t = 2, 3, \ldots, n.$

- Egymást követő bázisviszonyszámok hányadosa:

$$b_t/b_{t-1} = \big(Y_t/Y_b\big) : \big(Y_{t-1}/Y_b\big) = Y_t/Y_{t-1} = \ell_t.$$

- Új bázisra (például Y_b-ről Y_c-re) való áttérés:

$$b_t/b_c = \big(Y_t/Y_b\big) : \big(Y_c/Y_b\big) = Y_t/Y_c.$$

- Láncviszonyszámok szorzata:

$$\ell_1 \cdot \ell_2 \cdot \ldots \cdot \ell_k = \frac{Y_{b+1}}{Y_b} \cdot \frac{Y_{b+2}}{Y_{b+1}} \cdot \ldots \cdot \frac{Y_{b+k}}{Y_{b+k-1}} = \frac{Y_{b+k}}{Y_b} = b_{b+k}.$$

- Két ugyanazon időegységre vonatkozó bázisviszonyszámsor hányadosa:

$$\big(A_t/A_b\big) : \big(B_t/B_b\big) = \big(A_t/B_t\big) : \big(A_b/B_b\big) = V_t/V_b.$$

Példa

A házi gyerekorvosi betegellátás adatai 2005–2009

Év	Házi gyerekorvosok száma (fő)	Bejelentkezett lakosok száma (ezer fő)	Betegforgalom (ezer eset)
2005	1571	1475.5	9634.4
2006	1557	1474.2	9856.7
2007	1554	1461.3	9676.1
2008	1559	1463.4	9780.6
2009	1548	1452.3	10284.2

Forrás: KSH

A házi gyerekorvosi betegellátás időbeli változása

Év	Orvosok száma	Bejelentkezettek száma	Betegforgalom	Orvosok száma	Bejelentkezettek száma	Betegforgalom
	2005=100			Előző év=100		
2005	100.0	100.0	100.0	–	–	–
2006	99.1	99.9	102.3	99.1	99.9	102.3
2007	98.9	99.0	100.4	99.8	99.1	98.2
2008	99.2	99.2	101.5	100.3	100.1	101.1
2009	98.5	98.4	106.7	99.3	99.2	105.1

Példa

A házi gyerekorvosi betegellátást jellemző intenzitási viszonyszámok

Év	Egy házi gyerekorvosra jutó bejelentkezett lakos (fő)	Egy házi gyerekorvosra jutó betegforgalom (eset)	Százezer bejelentkezett lakosra jutó házi gyerekorvos	Egy lakosra jutó betegforgalom (eset)
2005	939	6132	106.5	6.53
2006	947	6331	105.6	6.69
2007	940	6227	106.3	6.62
2008	939	6274	106.5	6.68
2009	938	6644	106.6	7.08

Az előző táblázatból számolt dinamikus viszonyszámok

Év	Egy házi gyerekorvosra jutó bejelentkezett lakos (fő)		Egy házi gyerekorvosra jutó betegforgalom (eset)		Százezer bejelentkezett lakosra jutó házi gyerekorvos		Egy lakosra jutó betegforgalom (eset)	
	2005 =100	előző év =100	2005 =100	előző év =100	2005 =100	előző év =100	2005 =100	előző év =100
2005	100.0	–	100.0	–	100.0	–	100.0	–
2006	100.9	100.9	103.2	103.2	99.2	99.2	102.5	102.5
2007	100.1	99.3	101.5	98.4	99.8	100.7	101.4	99.0
2008	100.0	99.9	102.3	100.8	100.0	100.2	102.3	100.9
2009	99.9	99.9	108.3	105.9	100.1	100.1	108.4	106.0

Grafikus ábrázolás

Idősorok ábrázolása: vonaldiagram

Mennyiségi ismérvek kapcsolata: pontdiagram

Szerkezeti megoszlás ábrázolása: osztott kör-, oszlop- vagy szalagdiagram.

Mennyiségi ismérv eloszlásának ábrázolása: hisztogram.

Mennyiségi sorok

Y: mennyiségi ismérv;

N: a sokaság elemszáma;

$Y_1, Y_2, \ldots, Y_N$: az Y ismérv változatai, amik különbségi, vagy arány skálán mért számértékek.

Diszkrét: csak megszámlálható számosságú értéket vehet fel. Valamilyen számlálás eredménye, például háztartás nagysága, családban lévő gépjárművek száma. Megadható a pontos értéke.

Folytonos: kontinuum számosságú értéket vehet fel. Valamilyen mérés eredménye, például a háztartás összjövedelme, a családban lévő gépjárművek összértéke. Értéke csak bizonyos pontosságra kerekítve adható meg.

Ha a diszkrét ismérv nagyon sok értéket vehet fel, kezelhetjük folytonosként, például nagyvárosok népessége.

A rangsor a megfigyelési egységekhez tartozó Y_i ismérvértékeknek az Y_i monoton nemcsökkenő sorrendjében történő felsorolása. A rangsor i-edik tagját Y_i^*-gal jelöljük.

Gyakoriségi sor

Az Y szerint képzett osztály		Osztályközép	Abszolút	Relatív
alsó határa	felső határa		gyakoriság	
Y_{10}	Y_{11}	Y_1	f_1	g_1
Y_{20}	Y_{21}	Y_2	f_2	g_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{i0}	Y_{i1}	Y_i	f_i	g_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{k0}	Y_{k1}	Y_k	f_k	g_k
Összesen		–	N	1

Y_{i0} és Y_{i1}: az Y ismérv szerint képzett C_i osztály határai. Egybe is eshetnek.

Osztályközös gyakoriségi sor: Y_{i0} és Y_{i1} nem esik egybe.

f_i: a C_i osztály gyakorisága.

$g_i = f_i/N$: a C_i osztály relatív gyakorisága.

$Y_i = \big(Y_{i0} + Y_{i1}\big)/2$: osztályközép.

Gyakoriségi sor

a) Y diszkrét és kevés értéket vesz fel.

Példa. 405 személygépkocsi hengerszám szerinti megoszlása.

A hengerek száma (darab)	A személygépkocsik	
	száma	százalékos megoszlása
Y_i	f_i	g_i
3	4	1.0
4	207	51.1
5	3	0.8
6	84	20.7
8	107	26.4
Összesen	405	100

A rangsor egyértelműen felírható.

Gyakoriségi sor

b) Y folytonos vagy diszkrét és sok értéket vesz fel.

Példa. Magyarország városainak népességszám szerinti megoszlása, 2006. január 1. (Hunyadi, Vita, 2018, 2.3. táblázat).

A népesség száma (fő)	Osztályköz hosszúság	A városok			
		száma	számának megoszlása	népességének száma (fő)	népességének megoszlása
1001 – 5000	4000	56	19.4	199629	4.0
5001 – 10000	5000	95	33.0	685534	13.6
10001 – 20000	10000	76	26.4	1078313	21.3
20001 – 40000	20000	39	13.5	1088993	21.5
40001 – 70000	30000	11	3.8	622350	12.3
70001 – 110000	40000	5	1.7	436468	8.6
110001 – 160000	50000	3	1.0	400349	7.9
160001 – 210000	50000	3	1.0	541758	10.7
Összesen	–	288	99.8	5053394	99.9

Osztályközhossz: $h_i = Y_{i1} - Y_{i0}$

Ha Y_{10} és/vagy Y_{k1} nem ismert, akkor értelmesen megbecsüljük.

Értékösszegsor

Az értékösszegsor az Y ismérv alapján kialakított osztályokhoz az egyes osztályokba tartozó egységeknél fellépő ismérvértékek S_i-vel jelölt összegét rendeli hozzá.

$$S_i = \sum_{Y_{i0} \le Y \le Y_{i1}} Y, \qquad i = 1, 2, \ldots, k.$$

S_i: tényleges értékösszeg.

Ha $Y_{i0} = Y_i = Y_{i1}$, akkor $S_i = f_i \cdot Y_i$.

Osztályközös gyakorisági sor – becsült értékösszeg.

$$\widetilde{S}_i = f_i \cdot Y_i, \qquad i = 1, 2, \ldots, k.$$

Relatív értékösszeg:

$$Z_i = \frac{S_i}{\sum_{i=1}^{k} S_i} \quad \text{vagy} \quad \widetilde{Z}_i = \frac{\widetilde{S}_i}{\sum_{i=1}^{k} \widetilde{S}_i}.$$

Példa

Tényleges és becsült értékösszegek.

A népesség száma	f_i	Y_i	$\widetilde{S}_i = f_i Y_i$	$\widetilde{Z}_i$	S_i	Z_i
1001 – 5000	56	3000	168000	3.23	199629	3.95
5001 – 10000	95	7500	712500	13.69	685534	13.56
10001 – 20000	76	15000	1140000	21.90	1078313	21.34
20001 – 40000	39	30000	1170000	22.48	1088993	21.55
40001 – 70000	11	55000	605000	11.62	622350	12.32
70001 – 110000	5	90000	450000	8.64	436468	8.64
110001 – 160000	3	135000	405000	7.78	400349	7.92
160001 – 210000	3	185000	555000	10.66	541758	10.72
Összesen	288	–	5205500	100.00	5053394	100.00

Hunyadi, Vita (2018, 2.6. táblázat)

Kumulálás

Gyakoriságok, értékösszegek: $f_i' = \sum_{j=1}^{i} f_j$, $S_i' = \sum_{j=1}^{i} S_j$.

Kvantilisek

Az $Y_{i/k}$ i-edik k-adrendű kvantilis az a szám, amelynél az összes előforduló ismérvérték legfeljebb i/k-ad része kisebb és legfeljebb $(1 - i/k)$-ad része nagyobb, $k \geq 2,\ i = 1, 2, \ldots, k-1$. Az i/k helyett tetszőleges $0 < p < 1$ szerepelhet.

k	Elnevezés	Jelölés	Lehetséges kvantilisek
2	medián	Me	Me
4	kvartilis	Q_i	Q_1, Q_2, Q_3
5	kvintilis	K_i	K_1, K_2, K_3, K_4
10	decilis	D_i	$D_1, D_2, \ldots, D_9$
100	percentilis	P_i	$P_1, P_2, \ldots, P_{99}$

$Y_1^*, Y_2^*, \ldots, Y_N^*$: rangsor

$$s_p = p(N+1).$$

Ha $s_p \in \mathbb{Z}$: $Y_p = Y_{s_p}^*$.

Ha $s_p \notin \mathbb{Z}$: $Y_p = Y_{[s_p]}^* + \{s_p\}\big(Y_{[s_p]+1}^* - Y_{[s_p]}^*\big)$.

Példa

Néhány alsó középkategóriás személygépkocsi vegyes fogyasztása.

	Kia cee'd 1.4 CVVT	Citroën C4 1.4 Vti	Ford Focus 1.6 Ti-VCT	Honda Civic 1.4i
Teljesítmény (LE)	100	95	105	100
Fogyasztás (l/100km)	6.0	6.1	5.9	5.4
	Mazda 3 1.6 MZR	Opel Astra 1.4 Ecotec	Renault Mégane 1.6	Volkwagen Golf 1.2 TSI
Teljesítmény (LE)	105	100	100	105
Fogyasztás (l/100km)	6.5	5.5	6.7	5.7

Forrás: Az Autó, 2012/9.

Rangsor: 5.4, 5.5, 5.7, 5.9, 6.0, 6.1, 6.5, 6.7

Alsó kvartilis: $p = 1/4$, $s_p = 9/4$, $[s_p] = 2$, $\{s_p\} = 0.25$.

$$Q_1 = 5.5 + 0.25(5.7 - 5.5) = 5.55.$$

Medián: $p = 1/2$, $s_p = 9/2$, $[s_p] = 4$, $\{s_p\} = 0.5$.

$$Q_2 = Me = 5.9 + 0.5(6.0 - 5.9) = (5.9 + 6.0)/2 = 5.95.$$

Felső kvartilis: $p = 3/4$, $s_p = 27/4$, $[s_p] = 6$, $\{s_p\} = 0.75$.

$$Q_3 = 6.1 + 0.75(6.5 - 6.1) = 6.4.$$

Kvantilisek

Osztályközös gyakorisági sor: a kvantilis közelítése adható meg.

$$\widetilde{Y}_p = Y_{q0} + \left(pN - f'_{q-1}\right)\frac{h_q}{f_q}.$$

q: annak a *legelső* osztálynak a sorszáma, melyre $f'_q \geq pN$ (a kvantilist tartalmazó osztály).
Y_{q0}, h_q, f_q: a kvantilist tartalmazó osztály alsó határa, szélessége, illetve gyakorisága.
f'_{q-1}: a kvantilist tartalmazó osztály előtti osztállyal záródó kumulált gyakoriság.

Relatív gyakoriságokkal való megadás:

$$\widetilde{Y}_p = Y_{q0} + \left(p - g'_{q-1}\right)\frac{h_q}{g_q},$$

$$\widetilde{Y}_p = Y_{q0} + \left(100p - 100g'_{q-1}\right)\frac{h_q}{100g_q}.$$

Példa

A népesség száma	h_i	f_i	$100g_i$	f_i'	$100g_i'$
1001 – 5000	4000	56	19.44	56	19.44
5001 – 10000	5000	95	32.99	151	52.43
10001 – 20000	10000	76	26.39	227	78.82
20001 – 40000	20000	39	13.54	266	92.36
40001 – 70000	30000	11	3.82	277	96.18
70001 – 110000	40000	5	1.74	282	97.92
110001 – 160000	50000	3	1.04	285	98.96
160001 – 210000	50000	3	1.04	288	100.00
Összesen	–	288	100.00	–	–

Alsó kvartilis: $p = 1/4$, $pN = 72$, $q = 2$, $Y_{q0} = 5000$, $h_q = 5000$, $f_q = 95$, $f'_{q-1} = 56$.

$$\widetilde{Q}_1 = 5000 + (72 - 56) \cdot 5000/95 = 5842.$$

Medián: $p = 1/2$, $q = 2$, $Y_{q0} = 5000$, $h_q = 5000$, $g_q = 0.3299$, $g'_{q-1} = 0.1944$.

$$\widetilde{Q}_2 = Me = 5000 + (0.5 - 0.1944) \cdot 5000/0.3299 = 9632.$$

Felső kvartilis: $p = 3/4$, $100p = 75.00$, $q = 3$, $Y_{q0} = 10000$, $h_q = 10000$, $100g_q = 26.39$, $100g'_{q-1} = 52.43$.

$$\widetilde{Q}_3 = 10000 + (75.00 - 52.43) \cdot 10000/26.39 = 18552.$$

Gyakoriségi eloszlások grafikus ábrázolása

Leveles ág ábra (stem-and-leaf)

Függőleges vonal. Tőle balra az ismérvértékek legelső helyiértékű számjegyei (*ágak*). A vonal jobb oldalán az ismérvértékek további azonosításához szükséges számjegyek szóközzel vagy vesszővel elválasztva (*levelek*).

Doboz ábra (box plot, box-and-whiskers plot)

Vízszintes vagy függőleges tengelyen ábrázolja a kvartiliseket, ezek alkotják a *dobozt*, valamint a legnagyobb és a legkisebb ismérvértéket. Q_1: a doboz alja; Q_3: a doboz teteje; Me: a doboz osztóvonala.

Hisztogram

Osztályközös gyakoriségi sorban az osztályközök fölé oszlopokat emelünk, melyek területe arányos az adott osztály gyakoriségával (gyakoriség hisztogram), vagy relatív gyakoriségával (sűrűség hisztogram).

Helyzetmutatók (középértékek). Medián

A medián az az ismérvérték, amelyiknél az összes előforduló ismérvérték legfeljebb fele kisebb és legfeljebb fele nagyobb. $p = 1/2$-hez tartozó kvantilis.

$$\sum_{i=1}^{N} |Y_i - A| \quad \text{minimumhelye} \quad A = Me.$$

Rangsorból számolva:

$$\begin{aligned} N = 2k+1: &\quad Me = Y^*_{k+1}, \\ N = 2k: &\quad Me = \left(Y^*_k + Y^*_{k+1}\right)/2. \end{aligned}$$

Osztályközös gyakorisági sorból számolva:

$$\widetilde{Me} = Y_{me,0} + \left(N/2 - f'_{me-1}\right)\frac{h_{me}}{f_{me}}.$$

me: a legelső olyan osztályköz sorszáma, ahol $f'_{me} \geq N/2$.

Helyzetmutatók (középértékek). Módusz

Diszkrét ismérv esetén a módusz a leggyakrabban előforduló ismérvérték, *folytonos ismérv* esetén pedig a sűrűségfüggvény maximumhelye.

Osztályközös gyakorisági sorból számolható:

$$\widetilde{Mo} = Y_{mo,0} + \frac{d_a}{d_a + d_f} \cdot h_{mo}.$$

mo: a móduszt tartalmazó osztályköz sorszáma;

$$d_a = f_{mo} - f_{mo-1}, \qquad d_f = f_{mo} - f_{mo+1}.$$

Egyenlő osztályközök: d_a és d_f a tényleges gyakoriságok különbségei.

Nem egyenlő osztályközök: d_a és d_f az egységesített (korrigált) gyakoriságok különbségei.

Példa

Magyar városok gyakoriságainak egységesítése.

A népesség száma (fő)	Osztályköz hosszúság	Eredeti	Egységnyi	5000 fő
		hosszúságú osztályközök gyakorisága		
1001 – 5000	4000	56	0.014	70
5001 – 10000	5000	95	0.019	95
10001 – 20000	10000	76	0.0076	38
20001 – 40000	20000	39	0.00195	9.75
40001 – 70000	30000	11	0.000367	1.84
70001 – 110000	40000	5	0.000125	0.63
110001 – 160000	50000	3	0.00006	0.30
160001 – 210000	50000	3	0.00006	0.30
Összesen	–	288		

Hunyadi, Vita (2018, 2.11. táblázat).

$mo = 2$, $f_{mo} = 95$, $f_{mo-1} = 70$, $f_{mo+1} = 38$, $Y_{mo,0} = 5000$, $h_{mo} = 5000$,
$d_a = 95 - 70 = 25$, $d_f = 95 - 38 = 57$.

$$\widetilde{Mo} = 5000 + \frac{25}{25 + 57} \cdot 5000 = 6524.$$

Helyzetmutatók (középértékek). Számtani átlag

$Y_1, Y_2, \ldots, Y_N$: ismérvértékek.

Számtani átlag (súlyozatlan eset):

$$\overline{Y} = \frac{Y_1 + Y_2 + \ldots + Y_N}{N} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{\sum Y}{N}.$$

$$\sum_{i=1}^{N} (Y_i - A)^2 \quad \text{minimumhelye} \quad A = \overline{Y}.$$

S értékösszegből:

$$\overline{Y} = S/N.$$

Gyakorisági sorból (súlyozott eset):

$$\overline{Y} = \frac{\sum_{i=1}^{k} f_i Y_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i Y_i}{N} = \frac{\sum fY}{N} = \frac{\sum \frac{f}{N} Y}{\sum \frac{f}{N}} = \frac{\sum gY}{\sum g} = \sum gY.$$

Y_i: az i-edik osztály egyedi értéke vagy osztályközepe; f_i: az i-edik osztály gyakorisága.

Szóródási mutatók. Terjedelemmutatók

Terjedelem: $R = Y_N^* - Y_1^* = Y_{\max} - Y_{\min}$.

Osztályközös gyakorisági sor: problémás.

Interkvartilis távolság: $R_{0.5} = Q_3 - Q_1$.

Példa. Személygépkocsik vegyes átlagfogyasztása.

Rangsor: 5.4, 5.5, 5.7, 5.9, 6.0, 6.1, 6.5, 6.7

$$R = 6.7 - 5.4 = 1.3 \text{ l/100km}.$$

Kvartilisek: $Q_1 = 5.55$, $Q_3 = 6.4$.

$$R_{0.5} = 6.4 - 5.55 = 0.85 \text{ l/100km}.$$

Szóródási mutatók. Szórás, variancia, relatív szórás

Szórás:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \qquad \text{súlyozatlan eset,}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (Y_i - \overline{Y})^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{\sum_{i=1}^{k} f_i d_i^2}{\sum_{i=1}^{k} f_i}} \qquad \text{súlyozott eset.}$$

$d_i = Y_i - \overline{Y}$: az átlagtól való eltérés.

Variancia (szórásnégyzet): σ^2.

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N} Y_i^2 - N\overline{Y}^2, \quad \text{azaz} \quad \sigma^2 = \overline{Y^2} - \overline{Y}^2.$$

Relatív szórás:

$$V = \sigma / \overline{Y}.$$

Példa

Személygépkocsik vegyes átlagfogyasztása.

Ismérvértékek: 6.0, 6.1, 5.9, 5.4, 6.5, 5.5, 6.7, 5.7.

$$\overline{Y} = \frac{6.0 + 6.1 + \ldots + 5.7}{8} = 5.975 \qquad \text{l/100 km.}$$

$$\sigma = \sqrt{\frac{(6.0 - 5.975)^2 + (6.1 - 5.975)^2 + \ldots + (5.7 - 5.975)^2}{8}} = 0.4265 \qquad \text{l/100 km.}$$

$V = 0.4265/5.975 = 0.0714.$

Példa

Magyar városok népessége

A népesség száma	f_i	Y_i	$f_i Y_i$	$d_i = Y_i - \overline{Y}$	$f_i d_i^2$
1001 – 5000	56	3000	168000	-15075	12726315000
5001 – 10000	95	7500	712500	-10575	10623909375
10001 – 20000	76	15000	1140000	-3075	718627500
20001 – 40000	39	30000	1170000	11925	5546019375
40001 – 70000	11	55000	605000	36925	14998011875
70001 – 110000	5	90000	450000	71925	25866028125
110001 – 160000	3	135000	405000	116925	41014366875
160001 – 210000	3	185000	555000	166925	83591866875
Összesen	288	–	5205500	–	195085145000

$$\overline{Y} = 5205500/288 = 18074.65 \approx 18075.$$

$$\sigma = \sqrt{\frac{195085145000}{288}} = 2.6026.51 \approx 26027.$$

$$V = 26026.51/18074.65 = 1.4399.$$

Koncentráció

A sokasághoz tartozó teljes értékösszeg jelentős részének vagy egészének kevés egységre való összpontosulását koncentrációnak nevezzük.

Kicsi sokaság: abszolút koncentráció.

Nagy sokaság: relatív koncentráció.

Lorenz görbe

Egyedi adatok: $Y_1^*, Y_2^*, \ldots, Y_N^*$ rangsor, S értékösszeg.

$$(0,0) \quad \text{és} \quad \left(\frac{k}{N}, \frac{\sum_{i=1}^{k} Y_i^*}{S} \right), \ k = 1, 2, \ldots, N, \quad \text{pontok.}$$

Osztályközök: g_i', Z_i' osztályközös kumulált relatív gyakoriságok és relatív értékösszegek.

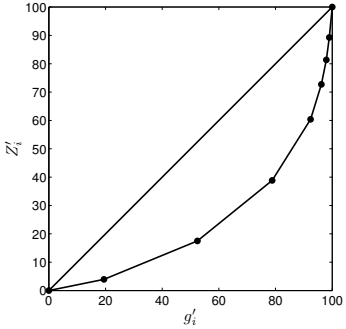
$$(0,0) \quad \text{és} \quad (g_i', Z_i'), \ i = 1, 2, \ldots, k, \quad \text{pontok.}$$

Koncentrációs együttható: a görbe és a négyzet átlója által bezárt terület (koncentrációs terület) aránya a négyzet feléhez. Jele: L.

Példa

Magyar városok népességeinek relatív gyakoriságai és értékösszegei.

A népesség száma	f_i	g_i	g_i'	S_i	Z_i	Z_i'
1001 – 5000	56	19.44	**19.44**	199629	3.95	**3.95**
5001 – 10000	95	32.99	**52.43**	685534	13.56	**17.51**
10001 – 20000	76	26.39	**78.82**	1078313	21.34	**38.85**
20001 – 40000	39	13.54	**92.36**	1088993	21.55	**60.40**
40001 – 70000	11	3.82	**96.18**	622350	12.32	**72.72**
70001 – 110000	5	1.74	**97.92**	436468	8.64	**81.36**
110001 – 160000	3	1.04	**98.96**	400349	7.92	**89.28**
160001 – 210000	3	1.04	**100.00**	541758	10.72	**100.00**
Összesen	288	100.00	–	5053394	100.00	–



Koncentrációs együttható: $L = 0.523$.

Közepes koncentráció.

Koncentráció

Herfindahl index:

$$HI = \sum_{i=1}^{N} Z_i^2.$$

$HI = 1/N$: nincs koncentráció. $HI = 1$: teljes koncentráció.

Példa. Autógyárak piaci részesedése (%) az Európai Unióban 2020-ban és 2021-ben:

Gyártó	VW csop.	Stellanis	Renault csop.	Hyundai csop.	BMW csop.
2021	25.1	21.9	10.6	8.5	6.8
2020	25.8	21.8	11.5	7.0	6.5
Gyártó	Toyota csop.	Daimler	Ford	Volvo	Egyéb
2021	6.3	5.6	4.1	2.3	8.8
2010	5.7	6.3	4.9	2.2	8.3

Forrás: Európai Autógyártók Szövetsége (ACEA)

2021: $HI = 0.251^2 + 0.219^2 + \ldots + 0.088^2 = 0.1511$.

2020: $HI = 0.258^2 + 0.218^2 + \ldots + 0.083^2 = 0.1534$.

Momentumok

A körüli *r*-edik momentum:

$$M_r(A) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - A)^r = \frac{1}{N}\sum_{i=1}^{N} d_i^r(A) \qquad \text{súlyozatlan eset,}$$

$$M_r(A) = \frac{\sum_{i=1}^{k} f_i(Y_i - A)^r}{\sum_{i=1}^{k} f_i} = \frac{1}{N}\sum_{i=1}^{k} f_i d_i^r(A) \qquad \text{súlyozott eset.}$$

$d_i(A) = Y_i - A$: az A értéktől való eltérés.

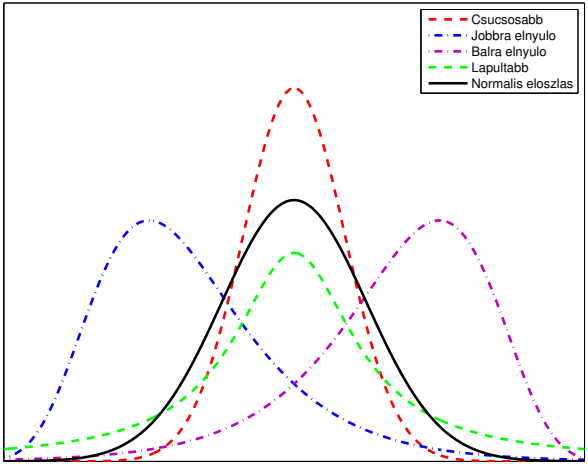
$A = 0$: *r*-edik momentum;

$A = \overline{Y}$: *r*-edik centrális momentum.

Speciális esetek:

- $M_1(0) = \overline{Y}$, $M_1(\overline{Y}) = 0$.
- $M_2(0) = \overline{Y^2}$, $M_2(\overline{Y}) = \sigma^2$.

Alakmutatók



Aszimmetria: jobbra vagy balra elnyúló.

Csúcsosság: hegyesebb vagy lapultabb, mint az ugyanolyan paraméterű normális eloszlás.

Aszimmetriamutatók

Ferdeség (skewness):

$$\alpha_3 = \frac{M_3(\overline{Y})}{\sigma^3}.$$

Pearson-féle mutató:

$$P = \frac{3(\overline{Y} - Me)}{\sigma}.$$

Decilisek és a medián eltérésén alapuló mutató:

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

Mutató	Jobbra elnyúló	Szimmetrikus	Balra elnyúló
Ferdeség	$\alpha_3 > 0$	$\alpha_3 \approx 0$	$\alpha_3 < 0$
Pearson	$P > 0$	$P \approx 0$	$P < 0$
Decilisek	$F_{0,1} > 0$	$F_{0,1} \approx 0$	$F_{0,1} < 0$

Példa

Személygépkocsik vegyes átlagfogyasztása.

Ismérvértékek: 6.0, 6.1, 5.9, 5.4, 6.5, 5.5, 6.7, 5.7.

$\overline{Y} = 5.975,\ \sigma = 0.4265,\ Me = 5.95.$

$$M_3(\overline{Y}) = \frac{(6.0 - 5.975)^3 + (6.1 - 5.975)^3 + \ldots + (5.7 - 5.975)^3}{8} = 0.0262.$$

$$\alpha_3 = \frac{0.0262}{0.4265^3} = 0.3372,$$

$$P = \frac{3(5.975 - 5.95)}{0.4265} = 0.1759.$$

Jobbra elnyúló.

Csúcsossági mutató

Lapultság (kurtosis):

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

Normális eloszlás esetén az elméleti értéke 0. Azonos paraméterű normálishoz hasonlítjuk.

Normálisnál csúcsosabb	Megegyező	Normálisnál lapultabb
$\alpha_4 > 0$	$\alpha_4 \approx 0$	$\alpha_4 < 0$

Példa. Személygépkocsik vegyes átlagfogyasztása.

Ismérvértékek: 6.0, 6.1, 5.9, 5.4, 6.5, 5.5, 6.7, 5.7.

$$M_4(\overline{Y}) = \frac{(6.0 - 5.975)^4 + (6.1 - 5.975)^4 + \ldots + (5.7 - 5.975)^4}{8} = 0.0648.$$

$$\alpha_4 = \frac{0.0648}{0.4265^4} - 3 = -1.0408.$$

Lapultabb, mint az 5.975 várható értékű, 0.4265 szórású normális.

Heterogén sokaságok

Az elemzés Y ismérve szempontjából lényegesen eltérő jellegzetességeket mutató részekre bontható sokaságokat az adott ismérv szempontjából heterogén sokaságoknak nevezzük.

A *fősokaságot* M darab *részsokaságra* bontjuk valamilyen csoportképző ismérv alapján.

Részviszonyszámok:

$$V_j = A_j / B_j, \qquad j = 1, 2, \ldots, M.$$

Összetett viszonyszám:

$$\overline{V} = \frac{\sum_{j=1}^{M} A_j}{\sum_{j=1}^{M} B_j} = \frac{\sum A}{\sum B} = \frac{\sum_{j=1}^{M} B_j V_j}{\sum_{j=1}^{M} B_j} = \frac{\sum_{j=1}^{M} A_j}{\sum_{j=1}^{M} \frac{A_j}{V_j}}.$$

A_j és B_j helyett a belőlük képzett *megoszlási viszonyszámok* is használhatóak.

Példa

A magyar lakásállomány megoszlása adott év január 1.-én

Szobák száma	Lakások száma			Százalékos megoszlás			2016. évi állomány (2010=100)
	1990	2010	2022	1990	2010	2022	
1	645 064	525 228	458437	16.74	12.13	10.15	87.28
2	1 680 918	1 739 538	1 696 221	43.62	40.17	37.53	97.51
3 és több	1 527 306	2 065 915	2 364 613	39.64	47.70	52.32	114.46
Összesen	3 853 288	4 330 681	4 519 271	100.00	100.00	100.00	104.35

Utolsó oszlop első 3 sor: *részviszonyszámok*.

Utolsó oszlop utolsó sor: *összetett viszonyszám*.

$$\frac{525228 \cdot 87.28 + 1739538 \cdot 97.51 + 2065915 \cdot 114.46}{4330681} = \frac{451928881.1}{4330681} = 104.3552\,\%;$$

$$\frac{12.13 \cdot 87.28 + 40.17 \cdot 97.51 + 47.70 \cdot 114.46}{100} = \frac{10435.4251}{100} = 104.3543\,\%.$$

$$\frac{4519271}{\frac{458437}{87.28} + \frac{1696221}{97.51} + \frac{2364613}{114.46}} = 104.3550\,\%; \qquad \frac{100}{\frac{10.15}{87.28} + \frac{37.53}{97.51} + \frac{52.32}{114.46}} = 104.3538\,\%.$$

Rész- és főátlagok

Y_{ij}: a j-edik részsokaság ($j = 1, 2, \ldots, M$) i-edik értéke ($i = 1, 2, \ldots, N_j$).

$N = \sum_{j=1}^{M} N_j$: a fősokaság nagysága.

A j-edik részátlag:

$$\overline{Y_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ij} = \frac{S_j}{N_j}, \quad j = 1, 2, \ldots, M.$$

$S_j = \sum_{i=1}^{N_j} Y_{ij}$: a j-edik részsokaság értékösszege.

A főátlag:

$$\overline{Y} = \frac{1}{N} \sum_{j=1}^{M} \sum_{i=1}^{N_j} Y_{ij} = \frac{1}{N} \sum_{j=1}^{M} S_j = \frac{\sum_{j=1}^{M} N_j \overline{Y_j}}{\sum_{j=1}^{M} N_j} = \frac{\sum_{j=1}^{M} S_j}{\sum_{j=1}^{M} \frac{S_j}{\overline{Y_j}}},$$

mivel

$$S_j = N_j \overline{Y_j}, \quad \text{azaz} \quad N_j = \frac{S_j}{\overline{Y_j}}.$$

Teljes szórás és variancia

Átlagtól való eltérés:

$$d_{ij} = Y_{ij} - \overline{Y} = B_{ij} + K_j, \quad i = 1, 2, \ldots, N_j,\ j = 1, 2, \ldots, M.$$

Belső eltérés:

$$B_{ij} = Y_{ij} - \overline{Y_j}, \quad i = 1, 2, \ldots, N_j,\ j = 1, 2, \ldots, M.$$

Külső eltérés:

$$K_j = \overline{Y_j} - \overline{Y}, \quad j = 1, 2, \ldots, M.$$

Teljes szórás:

$$\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j}(Y_{ij} - \overline{Y})^2} = \sqrt{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j} d_{ij}^2}.$$

Teljes variancia:

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j}(Y_{ij} - \overline{Y})^2 = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j} d_{ij}^2.$$

Belső szórás és variancia

Részszórások vagy csoporton belüli szórások:

$$\sigma_j = \sqrt{\frac{1}{N_j}\sum_{i=1}^{N_j}(Y_{ij} - \overline{Y_j})^2} = \sqrt{\frac{1}{N_j}\sum_{i=1}^{N_j} B_{ij}^2}, \quad j = 1, 2, \ldots, M.$$

Belső szórás:

$$\sigma_B = \sqrt{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j}(Y_{ij} - \overline{Y_j})^2} = \sqrt{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j} B_{ij}^2}.$$

A σ_B belső szórás azt mutatja, hogy a fősokaság egyes egységeihez tartozó Y_{ij} ismérvértékek átlagosan mennyivel térnek el a *saját részátlaguktól*. A belső szórás négyzete a belső variancia.

Részvarianciák és belső variancia kapcsolata:

$$\sigma_B^2 = \frac{1}{N}\sum_{j=1}^{M} N_j \sigma_j^2.$$

Külső szórás és variancia

Külső szórás:

$$\sigma_K = \sqrt{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j}(\overline{Y_j} - \overline{Y})^2} = \sqrt{\frac{1}{N}\sum_{j=1}^{M} N_j K_j^2}.$$

A σ_K külső szórás azt mutatja, hogy a részátlagok átlagosan mennyire térnek el a főátlagtól.

Kapcsolat a varianciák között:

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

Négyzetösszegek közötti összefüggés:

$$\sum_{j=1}^{M}\sum_{i=1}^{N_j}(Y_{ij} - \overline{Y})^2 = \sum_{j=1}^{M}\sum_{i=1}^{N_j}(Y_{ij} - \overline{Y_j})^2 + \sum_{j=1}^{M} N_j(\overline{Y_j} - \overline{Y})^2.$$

$$SST = SSB + SSK$$

SST a teljes, SSB a belső, SSK a külső négyzetösszeg.

Példa

Középfölde népei évenkénti fogathajtó versenye döntőjének másodpercekben mért eredményei:

j	Népcsoport	Eredmény Y_{ij}				N_j	S_j	$\sum_{i=1}^{N_j}(Y_{ij}-\overline{Y_j})^2$
1	Tündék	54.3	59.7	49.5		3	163.5	52.08
2	Törpök	55.0	45.2			2	100.2	48.02
3	Emberek	52.1	54.5	56.9	50.7	4	214.2	22.35
4	Hobbitok	44.8	47.4	54.2		3	146.4	47.12
Összesen						12	624.3	169.57

$$\overline{Y_1} = 163.5/3 = 54.50, \qquad \overline{Y_2} = 100.2/2 = 50.10,$$
$$\overline{Y_3} = 214.0/4 = 53.55, \qquad \overline{Y_4} = 146.4/3 = 48.80.$$

$$\overline{Y} = \frac{3 \cdot 54.50 + 2 \cdot 50.10 + 4 \cdot 53.55 + 3 \cdot 48.80}{12} = \frac{624.3}{12} = 52.025.$$

Példa

j	Népcsoport	Eredmény Y_{ij}				N_j	S_j	$\sum_{i=1}^{N_j}(Y_{ij}-\overline{Y_j})^2$
1	Tündék	54.3	59.7	49.5		3	163.5	52.08
2	Törpök	55.0	45.2			2	100.2	48.02
3	Emberek	52.1	54.5	56.9	50.7	4	214.2	22.35
4	Hobbitok	44.8	54.2	47.4		3	146.4	47.12
Összesen						12	624.3	169.57

$$\overline{Y_1} = 54.50,\ \overline{Y_2} = 50.10,\ \overline{Y_3} = 53.55,\ \overline{Y_4} = 48.80,\ \overline{Y} = 52.025.$$

$$SST = (54.3 - 52.025)^2 + \ldots + (47.4 - 52.025)^2 = 235.8625,$$
$$SSK = 3 \cdot (54.50 - 52.025)^2 + \ldots + 3 \cdot (48.80 - 52.025)^2 = 66.2925,$$
$$SSB = SST - SSK = 169.57.$$

$$\sigma = \sqrt{235.8625/12} = 4.4334, \quad \sigma_B = \sqrt{169.57/12} = 3.7591, \quad \sigma_K = \sqrt{66.292/12} = 2.3504.$$

$$\sigma_1 = \sqrt{52.08/3} = 4.1665, \qquad \sigma_2 = \sqrt{48.02/2} = 4.9000,$$
$$\sigma_3 = \sqrt{22.35/4} = 2.3638, \qquad \sigma_4 = \sqrt{47.12/3} = 3.9632.$$

Az ismérvek közötti kapcsolat fajtái

Lehetséges kapcsolatok két ismérv között:

- Az ismérvek függetlenek egymástól. Például: hajszín, testmagasság.
- A két ismérv között sztochasztikus kapcsolat van, azaz például a sokaság egységeinek X szerinti hovatartozásából, milyenségéből következtetni lehet az Y szerinti hovatartozásra, milyenségre. Például: hajszín, szemszín.
- A két ismérv között függvényszerű, azaz determinisztikus kapcsolat van. Például: ösztöndíjátlag, tanulmányi ösztöndíj.

Az ismérvek *fajtái* szerinti csoportosítás:

- Asszociáció: mindkét ismérv minőségi vagy területi (nominális skála).
- Vegyes kapcsolat: az egyik ismérv mennyiségi, a másik minőségi vagy területi (különbségi vagy arány és nominális skála).
- Korreláció: mindkét ismérv mennyiségi (különbségi vagy arány skála).
- Rangkorreláció: mindkét ismérvet sorrendi skálán mérjük.

Asszociáció

A kontingenciatábla általános alakja:

Az X ismérv szerinti osztályok	Az Y ismérv szerinti osztályok						$\sum j$
	C_1^Y	C_2^Y	$\dots$	C_j^Y	$\dots$	C_c^Y	
C_1^x	f_{11}	f_{12}	$\dots$	f_{1j}	$\dots$	f_{1c}	$f_{1\cdot}$
C_2^x	f_{21}	f_{22}	$\dots$	f_{2j}	$\dots$	f_{2c}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
C_i^x	f_{i1}	f_{i2}	$\dots$	f_{ij}	$\dots$	f_{ic}	$f_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
C_r^x	f_{r1}	f_{r2}	$\dots$	f_{rj}	$\dots$	f_{rc}	$f_{r\cdot}$
$\sum i$	$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot j}$	$\dots$	$f_{\cdot c}$	N

Utolsó sor: Y szerinti megoszlás.

Utolsó oszlop: X szerinti megoszlás.

Ha a soronkénti megoszlások azonosak, akkor X és Y *függetlenek*.

Ha a soronként legfeljebb egy nem 0 gyakoriság van, akkor X értéke egyértelműen meghatározza Y értékét, azaz *függvényszerű a kapcsolat*.

Tényleges és várt gyakoriságok

A $C_i^X \cdot C_j^Y$ tényleges, illetve relatív gyakorisága: f_{ij}, illetve f_{ij}/N.

$$\mathrm{P}\big(C_i^X \cdot C_j^Y\big) \approx \frac{f_{ij}}{N}, \qquad \mathrm{P}\big(C_i^X\big) \approx \frac{f_{i\cdot}}{N}, \qquad \mathrm{P}\big(C_j^Y\big) \approx \frac{f_{\cdot j}}{N}.$$

Ha C_i^X és C_j^Y függetlenek, akkor

$$\mathrm{P}\big(C_i^X \cdot C_j^Y\big) = \mathrm{P}\big(C_i^X\big) \cdot \mathrm{P}\big(C_j^Y\big) \approx \frac{f_{i\cdot}}{N} \cdot \frac{f_{\cdot j}}{N}.$$

$C_i^X \cdot C_j^Y$ várt gyakorisága (ha függetlenek): $N\mathrm{P}\big(C_i^X \cdot C_j^Y\big) \approx \frac{f_{i\cdot} \cdot f_{\cdot j}}{N}$.

Várt gyakoriságok (a függetlenség feltételezése mellett): $f_{ij}^* = \frac{f_{i\cdot} \cdot f_{\cdot j}}{N}$.

A kapcsolat szorossága

Khi-négyzet (chi-square) mutató:

$$\mathcal{X}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-f_{ij}^*)^2}{f_{ij}^*} = N\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_{i\cdot}\cdot f_{\cdot j}} - 1\right).$$

Lehetséges értékei: $O \le \mathcal{X}^2 \le N \min\big\{(r-1),(c-1)\big\}$.

$\mathcal{X}^2 = 0$: X és Y függetlenek.

$\mathcal{X}^2 = N \min\big\{(r-1),(c-1)\big\}$: X és Y között függvényszerű a kapcsolat.

Cramér-féle asszociációs együttható (SPSS: Cramer's V):

$$C = \sqrt{\frac{\mathcal{X}^2}{N \min\big\{(r-1),(c-1)\big\}}}, \qquad 0 \le C \le 1.$$

$C = 0$: X és Y függetlenek.

$C = 1$: X és Y között függvényszerű a kapcsolat.

Példa

Egy kutatócsoport azt vizsgálta, milyen szoros az összefüggés egy bizonyos betegség lefolyásának súlyossága és a betegek életkora között. A vizsgálati adatok:

		Életkor			Összesen
		40 alatti	40–60	60 fölötti	
Lefolyás	enyhe	41	34	9	84
	közepes	25	25	12	62
	súlyos	6	33	15	54
Összesen		72	92	36	200

$r = c = 3$, $N = 200$. Várt gyakoriságok:

30.24	38.64	15.12	84
22.32	28.52	11.16	62
19.44	24.84	9.72	54
72	92	36	200

$$\mathcal{X}^2 = \frac{(41-30.24)^2}{30.24} + \frac{(34-38.64)^2}{38.64} + \ldots + \frac{(15-9.72)^2}{9.72} = 22.5230.$$

$$C = \sqrt{\frac{22.5230}{200 \cdot 2}} = 0.2373. \qquad \text{Gyenge kapcsolat.}$$

PRE eljárás a kapcsolat szorosságának mérésére

Határozzuk meg annak a többletinformációnak a mennyiségét, amit a sokaság egységeinek az X szerinti hovatartozása nyújt az Y szerinti hovatartozásról.

PRE eljárás:

1. Meghatározzuk, hogy összességében mekkora hibával járna, ha a sokaság egységenek Y szerinti hovatartozását kizárólag azok Y szerinti megoszlása alapján próbálnánk meg megadni. Jelölés: E_1.
2. Meghatározzuk a hibát akkor is, ha ismerjük az egységek X szerinti hovatartozását. Jelölés: E_2.
3. Meghatározzuk a relatív hibacsökkenést:

$$0 \leq PRE = \frac{E_1 - E_2}{E_1} \leq 1.$$

$PRE = 0$: X és Y függetlenek.

$PRE = 1$: X és Y között függvényszerű a kapcsolat.

Vegyes kapcsolat

Y a mennyiségi, X a minőségi vagy területi ismérv.

Y_{ij}: a X szerint képzett j-edik részsokaság i-edik egységéhez tartozó Y ismérvérték $(j = 1, 2, \ldots, M, \;\; i = 1, 2, \ldots, N_j)$.

X ismerete nélkül az Y értékének becslése $\overline{Y}$. A becslési hiba:

$$E_1 = \sum_{j=1}^{M} \sum_{i=1}^{N_j} (Y_{ij} - \overline{Y})^2 = SST.$$

Ha tudjuk, a vizsgált egység az X szerinti j-edik részsokaságba tartozik, akkor az Y értékének becslése $\overline{Y_j}$. A becslési hiba:

$$E_2 = \sum_{j=1}^{M} \sum_{i=1}^{N_j} (Y_{ij} - \overline{Y_j})^2 = SSB.$$

PRE mutató:

$$PRE = \frac{E_1 - E_2}{E_1} = \frac{SST - SSB}{SST} = \frac{SSK}{SST} = 1 - \frac{\sigma_B^2}{\sigma^2} = \frac{\sigma_K^2}{\sigma^2} = H^2.$$

Varianciahányados

Varianciahányados:

$$0 \le H^2 = \frac{SST - SSB}{SST} = \frac{SSK}{SST} \le 1.$$

H^2 az Y ismérv szórásnégyzetének az X ismérv által megmagyarázott hányada.

$$H^2 = 0 \iff SSK = \sum_{j=1}^{M} N_j(\overline{Y_j} - \overline{Y})^2 = 0 \iff \overline{Y} = \overline{Y_j}.$$

Ez teljesül, ha X és Y független.

$$H^2 = 1 \iff SSB = \sum_{j=1}^{M} \sum_{i=1}^{N_j} (Y_{ij} - \overline{Y_j})^2 = 0 \iff Y_{ij} = \overline{Y_j}.$$

Ekkor X és Y között függvényszerű a kapcsolat.

Szóráshányados: $H = \sqrt{H^2}$.

Példa

Középfölde népei évenkénti fogathajtó versenye döntőjének másodpercekben mért eredményei:

j	Népcsoport	Eredmény Y_{ij}				N_j
1	Tündék	54.3	59.7	49.5		3
2	Törpök	55.0	45.2			2
3	Emberek	52.1	54.5	56.9	50.7	4
4	Hobbitok	44.8	47.4	54.2		3
Összesen						12

$$SST = 235.8625, \quad SSB = 169.57, \quad SSK = 66.2925.$$

$$H^2 = \frac{66.292}{235.8625} = 0.2811 \quad (28.11\%), \qquad H = 0.5302.$$

A népcsoporthoz való tartozás az Y szórásnégyzetének 28.11%-át magyarázza. $H = 0.5302$ közepesen erős kapcsolatot jelez.

Empirikus (tapasztalati) regressziófüggvény

X és Y mennyiségi ismérvek (akár fel is cserélhető a szerepük).

X: csoportképző ismérv. X szerinti osztályokat sorrendbe tudjuk állítani X értékei szerint.

Vizsgálható az X és az Y közötti kapcsolat *iránya*. Ha X növekedésével Y értéke is nő, az irány pozitív, ellenkező esetben negatív.

Az X szerint képzett részsokaságokhoz hozzárendelt $\overline{Y_j}$ részátlagok sorozatát az Y változó X változóra vonatkozó (X szerinti) empirikus regressziófüggvényének nevezzük.

Grafikus ábrázolás: az $(X_i, \overline{Y_i})$ pontokat összekötő *vonaldiagram*.

Y-nak X-re vonatkozó determinációs hányadosa (az X szerinti osztályokból számolt varianciahányados):

$$\eta^2_{Y|X} = \frac{\sigma^2_K(Y)}{\sigma^2(Y)}.$$

$\sigma^2_K(Y)$, illetve $\sigma^2(Y)$: Y külső, illetve teljes szórásnégyzete.

Példa

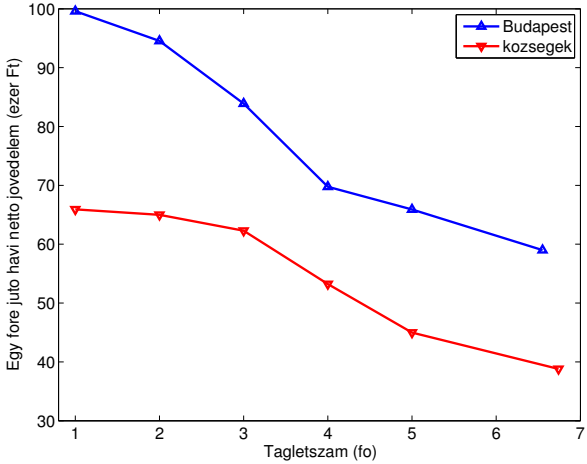
A háztartások taglétszáma és átlagos egy főre jutó havi nettó jövedelme 2004-ben Budapesten és a községekben.

A háztartás tagjainak száma	Az adott taglétszámú háztartásban lévő személyek			
	százalékos megoszlása		egy főre jutó jövedelme (Ft)	
	Budapesten	községekben	Budapesten	községekben
1	13.0	7.9	99586	65921
2	26.6	19.7	94538	64996
3	25.6	20.2	83887	62287
4	21.4	26.7	69762	53235
5	8.5	14.9	65900	44985
6 és több	4.9	10.6	58996	38796
Összesen	100.0	100.0	82974	55619

Forrás: KSH

Legalább 6 fős háztartások átlagos létszáma Budapesten 6.55 fő, a községekben 6.74 fő.

Empirikus regressziófüggvény



Az egy főre jutó havi nettó jövedelem és a háztartások taglétszáma közötti kapcsolat empirikus regressziófüggvényei.

Analitikus regressziófüggvény

X és Y mennyiségi ismérvek. Az (X_i, Y_i) párokat vizsgáljuk.

Kérdés: Felhasználható-e az X változó X_i értéke az Y változó ugyanazon egységéhez tartozó Y_i érték előrejelzésére?

Az X és Y közötti sztochasztikus kapcsolat természetét egy $f(X)$ függvénnyel, analitikus regressziófüggvénnyel akarjuk leírni.

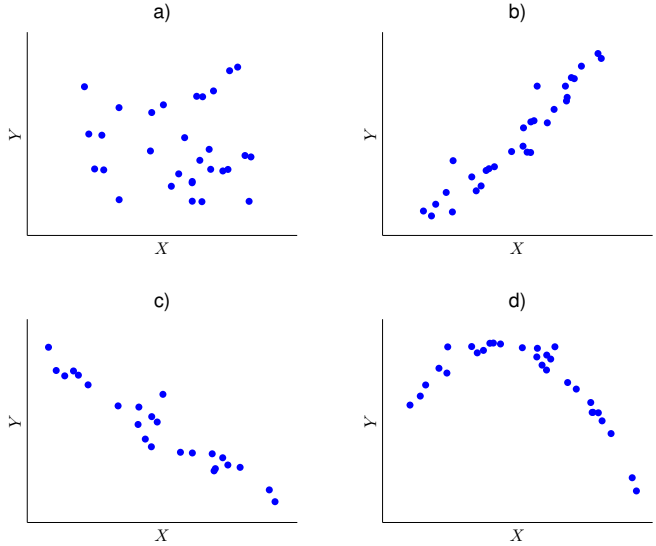
Például:

$$f(X) = \beta_0 + \beta_1 \cdot X, \qquad \text{lineáris regresszió;}$$
$$f(X) = \beta_0 \cdot \beta_1^X, \qquad \text{exponenciális regresszió.}$$

Az Y változó X_i-hez tartozó értékének előrejelzése $f(X_i)$.

Pontdiagram: az (X_i, Y_i) párok ábrázolása a kétdimenziós tér pontjaiként. Utal az $f(X)$ létezésére, illetve alakjára.

Pontdiagram típusok



a) X és Y független.

b) X és Y között *pozitív irányú* (lineáris) kapcsolat.

c) X és Y között *negatív irányú* (lineáris) kapcsolat.

d) X és Y között *nemlineáris* kapcsolat.

Korreláció

(Lineáris) korrelációs együttható:

$$r(X,Y) = \frac{\sum d_{X_i} d_{Y_i}}{\sqrt{\sum d_{X_i}^2 \sum d_{Y_i}^2}} = \frac{\sum X_i Y_i - N\overline{X}\,\overline{Y}}{\sqrt{\left(\sum X_i^2 - N(\overline{X})^2\right)\left(\sum Y_i^2 - N(\overline{Y})^2\right)}}$$

$$= \frac{N\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{\left(N\sum X_i^2 - \left(\sum X_i\right)^2\right)\left(N\sum Y_i^2 - \left(\sum Y_i\right)^2\right)}}.$$

Az $-1 \le r(X,Y) \le 1$ korrelációs együttható abszolút értéke az X és Y közötti lineáris kapcsolat szorosságát méri, előjele pedig a kapcsolat irányát mutatja.

$r(X,Y) = \pm 1$: függvényszerű lineáris kapcsolat;

$r(X,Y) = 0$: nincs lineáris kapcsolat. Korrelálatlanok. Nem feltétlenül függetlenek!

Minél nagyobb $|r(X,Y)|$, annál szorosabb a kapcsolat.

Kovariancia

X és Y kovarianciája:

$$C(X,Y) = \frac{\sum d_{X_i} d_{Y_i}}{N}.$$

Kapcsolata a *korrelációval*:

$$r(X,Y) = \frac{C(X,Y)}{\sigma_X \sigma_Y}.$$

σ_X, σ_Y: X, illetve Y szórása.

Kapcsolata a *varianciával*: $C(X,X) = \sigma_X^2$.

$C(X,Y) > 0$: X és Y között pozitív irányú kapcsolat;

$C(X,Y) < 0$: X és Y között negatív irányú kapcsolat.

$C(X,Y) = 0$: nincs lineáris kapcsolat. Nem feltétlenül függetlenek!

Determinációs együttható, súlyozott alakok

Determinációs együttható: r^2.

PRE mutató. $100r^2$ azt mutatja, hogy az X ismerete hány százalékkal csökkenti az Y nagyságával kapcsolatos bizonytalanságot, ha X és Y között lineáris kapcsolat van.

Súlyozott alakok:

$$C(X,Y) = \frac{\sum f_i \cdot d_{X_i} d_{Y_i}}{N}, \qquad r(X,Y) = \frac{\sum f_i \cdot d_{X_i} d_{Y_i}}{\sqrt{\sum f_i \cdot d_{X_i}^2 \sum f_i \cdot d_{Y_i}^2}}.$$

f_i: az (X_i, Y_i) pár gyakorisága.

$N = \sum f_i$: a sokaság elemszáma.

Példa

Néhány alsó középkategóriás személygépkocsi vegyes fogyasztása és CO_2 kibocsátása.

	Kia cee'd 1.4 CVVT	Citroën C4 1.4 Vti	Ford Focus 1.6 Ti-VCT	Honda Civic 1.4i
Teljesítmény (LE)	100	95	105	100
Fogyasztás (l/100km)	6.0	6.1	5.9	5.4
CO_2 (g/km)	139	140	136	128
	Mazda 3 1.6 MZR	Opel Astra 1.4 Ecotec	Renault Mégane 1.6	Volkwagen Golf 1.2 TSI
Teljesítmény (LE)	105	100	100	105
Fogyasztás (l/100km)	6.5	5.5	6.7	5.7
CO_2 (g/km)	149	129	155	134

Forrás: Az Autó, 2012/9.

X : 6.0, 6.1, 5.9, 5.4, 6.5, 5.5, 6.7, 5.7; $\overline{X} = 5.975$;

Y : 139, 140, 136, 128, 149, 129, 155, 134; $\overline{Y} = 138.75$.

Példa

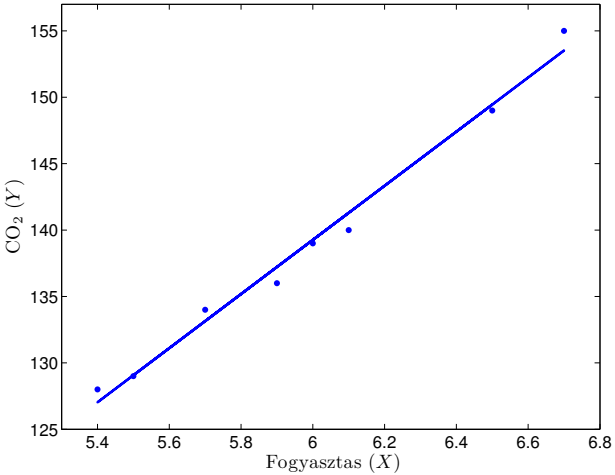
$$d_x: \ 0.025,\ 0.125,\ -0.075,\ -0.575,\ 0.525,\ -0.475,\ 0.725,\ -0.275;$$
$$d_y: \ 0.25,\ 1.25,\ -2.75,\ -10.75,\ 10.25,\ -9.75,\ 16.25,\ -4.75.$$

$$\sum d_x^2 = 0.025^2 + 0.125^2 + \ldots + (-0.275)^2 = 1.455,$$
$$\sum d_y^2 = 0.25^2 + 1.25^2 + \ldots + (-4.75)^2 = 611.5,$$
$$\sum d_x d_y = 0.025 \cdot 0.25 + \ldots + (-0.275) \cdot (-4.75) = 29.65.$$

$$C(X,Y) = \frac{\sum d_{X_i} d_{Y_i}}{N} = \frac{29.65}{8} = 3.7062,$$
$$r(X,Y) = \frac{\sum d_{X_i} d_{Y_i}}{\sqrt{\sum d_{X_i}^2 \sum d_{Y_i}^2}} = \frac{29.65}{\sqrt{1.455 \cdot 611.5}} = 0.9940.$$

Nagy az együttmozgás a vizsgált

Pontdiagram



Korreláció: $r(X, Y) = 0.9940$. Determinációs együttható: $r^2 = 0.9881$.

Az egyenes egyenlete: $f(X) = 16.9914 + 20.3780 \cdot X$.

Rangkorreláció

Mindkét ismérv *sorrendi skálán* mérhető.

R_X és R_Y: az X és Y változó szerinti rangok.

Kapcsolt rangok: az adott ismérv több értéke is megegyezik. A hozzájuk tartozó rangok átlagát kapja meg mindegyik azonos ismérvérték.

Spearman-féle rangkorrelációs együttható:

$$-1 \leq \varrho = 1 - \frac{6 \sum (R_X - R_Y)^2}{N(N^2 - 1)} \leq 1.$$

$\varrho = 1$: tökéletesen egyező rangsorolás.

$\varrho = -1$: tökéletesen ellentétes rangsorolás.

$\varrho = 0$: nincs kapcsolat a rangsorolások között.

Nincsenek kapcsolt rangok – ϱ megegyezik a rangokból számolt r korrelációs együtthatóval.

ϱ^2: a kapcsolat szorosságát mérő *PRE* mutató.

Példa

A hazai informatikai képzőhelyek 2015-ös, a hallgatói, illetve az oktatói kiválóság szerinti rangsorai. (Forrás: `eduline.hu`)

Intézmény	Hallgatók (R_X)	Oktatók (R_Y)	$(R_X - R_Y)^2$
BME-VIK	1	5.5	20.25
ELTE-IK	2	8	36
SZTE-TTIK	3	2	1
PE-MIK	7	3	16
DE-IK	5	5.5	0.25
DF	10	7	9
OE-NIK	4	4	0
PPKE-ITK	6	1	25
GDF	9	9.5	0.25
Kf-GAMFK	8	9.5	2.25
Összesen	55	55	110

$$\varrho = 1 - \frac{6 \cdot 110}{10 \cdot (100-1)} = \frac{1}{3} = 0.3(3), \quad \varrho^2 = \frac{1}{9} = 0.1(1), \quad r = 0.3293.$$

Gyenge kapcsolat a rangsorok között.

Összetett intenzitási viszonyszámok összehasonlítása

Két azonos tartalmú, de különböző összetett viszonyszámot kívánunk összehasonlítani.

$V_{0i} = A_{0i}/B_{0i}$, $V_{1i} = A_{1i}/B_{1i}$: *részviszonyszámok*.

Összetett viszonyszámok:

$$\overline{V}_s = \frac{\sum_j A_{sj}}{\sum_j B_{sj}} = \frac{\sum_j B_{sj} V_{sj}}{\sum_j B_{sj}} = \frac{\sum_j A_{sj}}{\sum_j \frac{A_{sj}}{V_{sj}}}, \qquad s = 0, 1.$$

$\overline{V}_0$ és $\overline{V}_1$ eltérésének okai:

- *eltérőek lehetnek* a két sokaság ugyanazon részeire számított V_{0i} és V_{1i} *részviszonyszámok*, és/vagy
- *eltérő lehet* a két sokaság *szerkezete (összetétele)*.

Jelölések

Részsokaság sorszáma	Első sokaság			Második sokaság			Összehasonlítás	
	számláló	nevező	viszonyszám	számláló	nevező	viszonyszám	különbség	hányados
1	A_{01}	B_{01}	V_{01}	A_{11}	B_{11}	V_{11}	k_1	i_1
2	A_{02}	B_{02}	V_{02}	A_{12}	B_{12}	V_{12}	k_2	i_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	A_{0j}	B_{0j}	V_{0j}	A_{1j}	B_{1j}	V_{1j}	k_j	i_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M	A_{0M}	B_{0M}	V_{0M}	A_{1M}	B_{1M}	V_{1M}	k_M	i_M
Fősokaság	$\sum A_{0j}$	$\sum B_{0j}$	$\overline{V}_0$	$\sum A_{1j}$	$\sum B_{1j}$	$\overline{V}_1$	K	I

Részviszonyszám különbségek: $k_j = V_{1j} - V_{0j}$.

Részviszonyszám hányadosok: $i_j = V_{1j}/V_{0j}$.

Összetett viszonyszám különbségek: $K = \overline{V}_1 - \overline{V}_0$.

Összetett viszonyszám hányadosok: $I = \overline{V}_1/\overline{V}_0$.

Különbségfelbontás

Teljes különbség: $K = \overline{V}_1 - \overline{V}_0$.

$$K = \frac{\sum B_1 V_1}{\sum B_1} - \frac{\sum B_0 V_0}{\sum B_0}.$$

Részhatás különbség(ek):

$$K' = K'_s = \frac{\sum B_s V_1}{\sum B_s} - \frac{\sum B_s V_0}{\sum B_s} = \frac{\sum B_s (V_1 - V_0)}{\sum B_s} = \frac{\sum B_s k}{\sum B_s}, \quad s = 0, 1.$$

A részviszonyszámok közötti eltérések hatását mutatja.

Összetétel hatás különbség(ek):

$$K'' = K''_s = \frac{\sum B_1 V_s}{\sum B_1} - \frac{\sum B_0 V_s}{\sum B_0}, \quad s = 0, 1.$$

A sokaságok eltérő összetételének a hatását mutatja.

Feltétel: $K = K' + K''$.

a) Ha K'-ben $B_s = B_0$, akkor K''-ben $V_s = V_1$ $(K = K'_0 + K''_1)$.

b) Ha K'-ben $B_s = B_1$, akkor K''-ben $V_s = V_0$ $(K = K'_1 + K''_0)$.

Példa

Korcsoport (év)	Népesség száma (millió fő)		Halálozások száma (fő)		Halálozási arányszám (‰)	
	Mexikó	Svédország	Mexikó	Svédország	Mexikó	Svédország
0–14	33.86	1.53	110 471	904	3.3	0.6
15–59	53.01	5.17	140 238	9 674	2.7	1.9
60–69	4.74	1.12	61 826	13 751	13.1	12.3
70–	1.40	0.95	133 913	66 001	95.7	69.5
Összesen	93.01	8.77	446 448	90 330	4.8	10.3

Keresztély, Sugár, Szarvas (2005, B.7 feladat, 87. old.)

Korcsoport (év)	Népesség megoszlása (%)	
	Mexikó	Svédország
0–14	36.4	17.4
15–59	57.0	59.0
60–69	5.1	12.8
70–	1.5	10.8
Összesen	100.0	100.0

A_1: halálozások száma, Svédország;
A_0: halálozások száma, Mexikó;
B_1: népesség száma, Svédország;
B_0: népesség száma, Mexikó;
V_1: halálozási arány, Svédország;
V_0: halálozási arány, Mexikó.

$$K = \overline{V}_1 - \overline{V}_0 = 10.3 - 4.8 = 5.5 \text{ ‰}$$

Példa

Korcsoport (év)	Népesség megoszlása (%)		Halálozások száma (fő)		Halálozási arányszám ‰)	
	Mexikó	Svédország	Mexikó	Svédország	Mexikó	Svédország
0–14	36.4	17.4	110 471	904	3.3	0.6
15–59	57.0	59.0	140 238	9 674	2.7	1.9
60–69	5.1	12.8	61 826	13 751	13.1	12.3
70–	1.5	10.8	133 913	66 001	95.7	69.5
Összesen	100.0	100.0	446 448	90 330	4.8	10.3

$$K_1' = 10.3 - \frac{17.4 \cdot 3.3 + 59 \cdot 2.7 + 12.8 \cdot 13.1 + 10.8 \cdot 95.7}{100} = -3.9 \text{ ‰}$$

$$K_0' = \frac{36.4 \cdot 0.6 + 57 \cdot 1.9 + 5.1 \cdot 12.3 + 1.5 \cdot 69.5}{100} - 4.8 = -1.8 \text{ ‰}$$

$$K_1'' = K - K_0' = 5.5 + 1.8 = 7.3 \text{ ‰} \qquad K_0'' = K - K_1' = 5.5 + 3.9 = 9.4 \text{ ‰}$$

$$K_{10}' = \frac{K_1' + K_0'}{2} = \frac{-3.9 - 1.8}{2} = -2.85 \text{ ‰} \qquad K_{10}'' = \frac{K_1'' + K_0''}{2} = \frac{7.3 + 9.4}{2} = 8.35 \text{ ‰}$$

Hányadosfelbontás

Összhatásindex: $I = \overline{V}_1/\overline{V}_0$.

$$I = \frac{\sum A_1}{\sum B_1} : \frac{\sum A_0}{\sum B_0} = \frac{\sum A_1}{\sum A_0} : \frac{\sum B_1}{\sum B_0} = \frac{\sum B_1 V_1}{\sum B_1} : \frac{\sum B_0 V_0}{\sum B_0}.$$

Részhatásindex(ek):

$$I' = I'_s = \frac{\sum B_s V_1}{\sum B_s} : \frac{\sum B_s V_0}{\sum B_s} = \frac{\sum B_s V_1}{\sum B_s V_0}, \quad s = 0, 1.$$

A részviszonyszámok változásának hatását mutatja.

Összetételhatás index(ek):

$$I'' = I''_s = \frac{\sum B_1 V_s}{\sum B_1} : \frac{\sum B_0 V_s}{\sum B_0}, \quad s = 0, 1.$$

A sokaságok összetétele megváltozásának a hatását mutatja.

Feltétel: $I = I' \cdot I''$.

a) Ha I'-ben $B_s = B_0$, akkor I''-ben $V_s = V_1$ $(I = I'_0 \cdot I''_1)$.

b) Ha I'-ben $B_s = B_1$, akkor I''-ben $V_s = V_0$ $(I = I'_1 \cdot I''_0)$.

Példa

Legmagasabb iskolai végzettség	2007			2010			2007=100
	Létszám		Havi bruttó	Létszám		Havi bruttó	
	fő B_0	megosz-lás (%)	átlagkereset, eFt V_0	fő B_1	megosz-lás (%)	átlagkereset, eFt V_1	$i = V_1/V_0$
8 általános alatt	8866	0.4	126	6236	0.3	127	100.8
8 általános	315625	14.2	109	263124	12.7	123	112.8
Szakiskola	647433	29.2	129	564221	27.3	144	111.6
Középiskola	726217	32.8	172	689006	33.3	186	108.1
Főiskola	329675	14.9	270	348364	16.9	300	111.1
Egyetem	189436	8.5	407	196527	9.5	440	108.1
Összesen	2217252	100.0	184.9	2067478	100.0	209.7	113.4

Forrás: *Munkaügyi adattár.* 2008, 2011.

$I = 113.4\%,$

$I_1' = 209.7 : 190.96 = 1.0981\ (109.81\%),$ $\qquad I_0'' = 113.4 : 109.81 = 1.0327\ (103.27\%),$

$I_0' = 203.18 : 184.9 = 1.0989\ (109.89\%),$ $\qquad I_1'' = 113.4 : 109.89 = 1.0320\ (103.20\%).$

Aggregátumok összehasonlítása

Aggregátum:

$$A = \sum_{i=1}^{n} q_i p_i = \sum_{i=1}^{n} \nu_i$$

q_i: az i-edik fajta egységeinek (termékeinek) mennyisége valamilyen alkalmas mértékegységben;

p_i: az i-edik fajta egység egységára;

ν_i: az i-edik fajta egységek összértéke.

Ha a q_i adatok:

- *termelt mennyiségek*, akkor A a *termelés*;
- *eladott mennyiségek*, akkor A a *forgalom*;
- *fogyasztott mennyiségek*, akkor A a *fogyasztás*.

q_i: valamilyen időszakra értelmezhető.

p_i: valamilyen időpontra értelmezhető.

Továbbiakban: q_i – *termelt mennyiség*; p_i – *egységár*.

Két időszak közötti összehasonlítás

n termék két időszakra vonatkozóan: *bázisidőszak*, *tárgyidőszak*

Termék sorszáma (i)	Termelt mennyiség a bázisidőszakban	Egységár a bázisidőszakban	Termelt mennyiség a tárgyidőszakban	Egységár a tárgyidőszakban
1	q_{01}	p_{01}	q_{11}	p_{11}
2	q_{02}	p_{02}	q_{12}	p_{12}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	q_{0i}	p_{0i}	q_{1i}	p_{1i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	q_{0n}	p_{0n}	q_{1n}	p_{1n}
Rövid jelölés	q_0	p_0	q_1	p_1

Kérdések *egy termékkel*, vagy a *termékek összességével* kapcsolatban:

a) hogyan változott a *termelés értéke*;

b) hogyan változott a *termelés mennyisége* (volumene);

c) hogyan változott az *ár*, illetve az *árszínvonal*.

Egyedi indexek

Egy termék vizsgálata – *dinamikus* viszonyszámok.

$$i_\nu = \frac{q_{1i}p_{1i}}{q_{oi}p_{oi}} = \frac{\nu_{1i}}{\nu_{0i}}, \qquad i_q = \frac{q_{1i}}{q_{oi}}, \qquad i_p = \frac{p_{1i}}{p_{oi}}.$$

Az egy termékre vonatkozóan meghatározott i_ν, i_q és i_p dinamikus viszonyszámokat egyedi indexeknek nevezzük. Az egyedi indexek rendre azt mutatják meg, hogy hogyan (hány százalékkal) változott az adott termékre vonatkozó

- termelési érték,
- termelt mennyiség,
- egységár

a bázisidőszakról a tárgyidőszakra.

Összefüggés az egyedi indexek között: $i_\nu = i_q \cdot i_p$.

Érték- és volumenindex

Értékindex:

$$I_\nu = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum \nu_1}{\sum \nu_0}.$$

Az I_ν értékindex azt mutatja, hogy hogyan (hány százalékkal) változott a teljes termelés értéke a bázisidőszakról a tárgyidőszakra.

Volumenindex:

$$I_q = \frac{\sum q_1 p_s}{\sum q_0 p_s}.$$

p_s: *mindkét időszakra érvényesnek feltételezett* egységár.

Az I_q volumenindex azt mutatja, hogy a termelt mennyiségek összességükben hogyan (hány százalékkal) változtak, vagyis hogyan változott a termelés volumene a bázisidőszakról a tárgyidőszakra.

Árindex

Árindex:

$$I_p = \frac{\sum q_s p_1}{\sum q_s p_0}.$$

q_s: *mindkét időszakra érvényesnek feltételezett* mennyiség.

Az I_p árindex azt mutatja, hogy az egységárak összességükben hogyan (hány százalékkal) változtak, amit az *árszínvonal-változás* mértékének is szokás nevezni.

I_ν, I_q, I_p: indexek aggregát formái.

$$I_\nu = \frac{\sum q_1 p_1}{\sum q_0 p_0}, \qquad I_q = \frac{\sum q_1 p_s}{\sum q_0 p_s}, \qquad I_p = \frac{\sum q_s p_1}{\sum q_s p_0}.$$

Legfontosabb volumen- és árindexformulák

Meg kell választani a p_s egységárakat és a q_s mennyiségeket. Használhatunk

- bázisidőszaki adatokat: I_q-ban $p_s = p_0$, I_p-ben $q_s = q_0$;
- tárgyidőszaki adatokat: I_q-ban $p_s = p_1$, I_p-ben $q_s = q_1$;
- a *bázisidőszaki* és *tárgyidőszaki* indexek mértani átlagát.

Bázisidőszaki súlyozású, avagy Laspeyres-féle indexek:

$$I_q^0 = \frac{\sum q_1 p_0}{\sum q_0 p_0}, \qquad I_p^0 = \frac{\sum q_0 p_1}{\sum q_0 p_0}.$$

Tárgyidőszaki súlyozású, avagy Paasche-féle indexek:

$$I_q^1 = \frac{\sum q_1 p_1}{\sum q_0 p_1}, \qquad I_p^1 = \frac{\sum q_1 p_1}{\sum q_1 p_0}.$$

Mértani átlagolású, ún. Fisher-féle indexek:

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

Példa

Egy fővárosi piacon egy büfében a jellegzetes termékek téli és nyári árai és a fogyasztott mennyiségek.

Megnevezés	December		Július	
	ár (Ft)	eladott mennyiség	ár (Ft)	eladott mennyiség
Nagyfröccs (pohár)	70	1500	80	1800
Sör (korsó)	120	1740	130	2110
Lángos (db)	100	2100	100	2000
Bableves (tál)	400	650	410	660
Hurka (10 dkg)	80	980	85	1060

Keresztély, Sugár, Szarvas (2005, Gy.108 feladat, 103. old.)

$$I_\nu = \frac{80 \cdot 1800 + 130 \cdot 2110 + 100 \cdot 2000 + 410 \cdot 660 + 85 \cdot 1060}{70 \cdot 1500 + 120 \cdot 1740 + 100 \cdot 2100 + 400 \cdot 650 + 80 \cdot 980} = \frac{979000}{862200} = 1.1355 \quad (113.55\,\%).$$

Példa

Megnevezés	December		Július	
	ár (Ft)	eladott mennyiség	ár (Ft)	eladott mennyiség
Nagyfröccs (pohár)	70	1500	80	1800
Sör (korsó)	120	1740	130	2110
Lángos (db)	100	2100	100	2000
Bableves (tál)	400	650	410	660
Hurka (10 dkg)	80	980	85	1060

$$I_q^0 = \frac{70\cdot 1800+120\cdot 2110+100\cdot 2000+400\cdot 660+80\cdot 1060}{70\cdot 1500+120\cdot 1740+100\cdot 2100+400\cdot 650+80\cdot 980} = \frac{928000}{862200} = 1.0763 \quad (107.63\,\%).$$

$$I_q^1 = \frac{80\cdot 1800+130\cdot 2110+100\cdot 2000+410\cdot 660+85\cdot 1060}{80\cdot 1500+130\cdot 1740+100\cdot 2100+410\cdot 650+85\cdot 980} = \frac{979000}{906000} = 1.0806 \quad (108.06\,\%).$$

$$I_q^F = \sqrt{1.0763 \cdot 1.0806} = 1.0784 \quad (107.84\,\%).$$

Példa

Megnevezés	December		Július	
	ár (Ft)	eladott mennyiség	ár (Ft)	eladott mennyiség
Nagyfröccs (pohár)	70	1500	80	1800
Sör (korsó)	120	1740	130	2110
Lángos (db)	100	2100	100	2000
Bableves (tál)	400	650	410	660
Hurka (10 dkg)	80	980	85	1060

$$I_p^0 = \frac{80\cdot 1500+130\cdot 1740+100\cdot 2100+410\cdot 650+85\cdot 980}{70\cdot 1500+120\cdot 1740+100\cdot 2100+400\cdot 650+80\cdot 980} = \frac{906000}{862200} = 1.0508 \quad (105.08\,\%).$$

$$I_p^1 = \frac{80\cdot 1800+130\cdot 2110+100\cdot 2000+410\cdot 660+85\cdot 1060}{70\cdot 1800+120\cdot 2110+100\cdot 2000+400\cdot 660+80\cdot 1060} = \frac{979000}{928000} = 1.0550 \quad (105.50\,\%).$$

$$I_q^F = \sqrt{1.0508 \cdot 1.0550} = 1.0529 \quad (105.29\,\%).$$

Indexek átlagformái

Minden aggregát formában felírható index egyben a megfelelő egyedi indexek súlyozott átlaga, azaz összetett viszonyszám.

Indexformula	A	B	V
I_ν	$q_1 p_1$	$q_0 p_0$	i_ν
I_q^0	$q_1 p_0$	$q_0 p_0$	i_q
I_q^1	$q_1 p_1$	$q_0 p_1$	i_q
I_p^0	$q_0 p_1$	$q_0 p_0$	i_p
I_p^1	$q_1 p_1$	$q_1 p_0$	i_p

Volumenindexek felírása átlagforma alakban:

$$I_q^0 = \frac{\sum q_0 p_0 i_q}{\sum q_0 p_0} = \frac{\sum \nu_0 i_q}{\sum \nu_0} = \frac{\sum q_1 p_0}{\sum \frac{q_1 p_0}{i_q}},$$

$$I_q^1 = \frac{\sum q_0 p_1 i_q}{\sum q_0 p_1} = \frac{\sum q_1 p_1}{\sum \frac{q_1 p_1}{i_q}} = \frac{\sum \nu_1}{\sum \frac{\nu_1}{i_q}}.$$

Összefüggések

Egyedi termékekre vonatkozó összefüggések:

$$q_0 p_0 \cdot i_q = q_1 p_0, \quad q_0 p_0 \cdot i_p = q_0 p_1, \quad q_0 p_0 \cdot i_\nu = q_1 p_1,$$

$$\frac{q_1 p_1}{i_q} = q_0 p_1, \quad \frac{q_1 p_1}{i_p} = q_1 p_0, \quad \frac{q_1 p_1}{i_\nu} = q_0 p_0.$$

Indexformulák összefüggései:

$$I_q^0 \cdot I_p^1 = I_\nu, \quad I_q^1 \cdot I_p^0 = I_\nu, \quad I_q^F \cdot I_p^F = I_\nu.$$

Példa

A karácsonyi fenyőfapiac forgalmi adatairól az alábbiakat tudjuk:

Fenyő fajtája	A forgalom értékének %-os megoszlása 2021-ben	Árváltozás (2020=100%)
Lucfenyő	50	105
Ezüstfenyő	30	107
Nordmann fenyő	20	115

Ismert, hogy a fenyők összes forgalma 2020-ról 2021-re 20%-kal emelkedett. Számítsa ki a forgalom érték-, ár- és volumenindexét. Minden kapott eredményt szövegesen is értékeljen.

A Laspeyres- és a Paache féle indexek eltérése

Volumenindexek:

$$I_q^0 = \frac{\sum q_0 p_0 i_q}{\sum q_0 p_0} = \frac{\sum \nu_0 i_q}{\sum \nu_0}, \qquad I_q^1 = \frac{\sum q_0 p_1 i_q}{\sum q_0 p_1}.$$

Az egyedi volumenindexek súlyozott átlagai. A súlyok:

$$w_L = \frac{q_0 p_0}{\sum q_0 p_0} = \frac{\nu_0}{\sum \nu_0}, \qquad w_P = \frac{q_0 p_1}{\sum q_0 p_1}.$$

Összefüggés a súlyok között:

$$w_P = \frac{q_0 p_0}{\sum q_0 p_0} \cdot \frac{p_1/p_0}{\frac{\sum q_0 p_1}{\sum q_0 p_0}} = w_L \cdot \frac{i_P}{I_p^0}.$$

Bortkiewicz formula

Az I_q^0 és I_q^1 volumenindexek minden olyan esetben eltérő eredményt adnak, amikor

- szóródnak az egyedi volumenindexek és
- szóródnak az egyedi árindexek és
- az egyedi volumen- és egyedi árindexek között sztochasztikus kapcsolat van.

Ha az egyedi volumen- és árindexek közötti sztochasztikus kapcsolat pozitív irányú, akkor $I_q^0 < I_q^1$, ha negatív irányú, akkor $I_q^0 > I_q^1$.

Bortkiewicz formula:

$$\frac{I_q^1}{I_q^0} = 1 + V_{i_q} \cdot V_{i_p} \cdot r_{i_q,i_p}.$$

V_{i_q}, V_{i_p}: az egyedi indexek relatív szórása;

r_{i_q,i_p}: az egyedi ár- és volumenindexek korrelációs együtthatója.

Árollók

Két egymással valamilyen kapcsolatban lévő csoport indexeit hasonlítjuk össze, legtöbbször árindexeket.

Árolló: két árindex hányadosa. A leggyakoribb árollók:

- agrárolló: a mezőgazdasági termékek termelőiár-indexének és a mezőgazdasági ráfordítások árindexének hányadosa.
- cserearányindex (terms of trade): az exportált és importált termékek árindexének hányadosa.

Az *árolló* azt mutatja, hogy a bevételt biztosító termékek bázisidőszakival azonos, illetve egységnyi volumenéért mennyivel nagyobb vagy kisebb volumenű másféle termék kapható cserébe a tárgyidőszakban.

Cserearányindex

A *cserearényindex* azt mutatja, hogy az exportált termékek és szolgáltatások bázisidőszakival azonos, illetve egységnyi volumenéért hányszor akkora volumenű terméket és szolgáltatást lehet importálni a tárgyidőszakban, mint a bázisidőszakban.

$$I_{cs} = \frac{I_p^x}{I_p^m}.$$

I_p^x: az exportált termékek árindexe;

I_p^m: az importált termékek árindexe.

Példa. 2020-ban az export forintban mért árszínvonala 4.7 %-kal, az importé 2.6 %-kal nőtt az előző évihez képest.

$$I_p^x = 104.7, \qquad I_p^m = 102.6, \qquad I_{cs} = 104.7/102.6 = 1.0205.$$

A cserearány 2.05 %-kal javult.

Forrás: KSH, *Helyzetkép a külkereskedelemről, 2020.*

Több időszak közötti összehasonlítás

Az indexsor valamely index – egy érték-, egy ár- vagy egy volumenindex – kettőnél több időszakra vonatkozó sorozata.

Aggregátumok:

$$A_{ij} = \sum q_i p_j, \qquad i, j = 1, 2, \ldots, k.$$

- az összegzés mindig a termékek *ugyanazon körére* terjed ki minden időszakban;
- az i index azt jelzi, *melyik időszak termelési értékéről* van szó, melyik időszak *mennyiségi adatait* vesszük alapul;
- a j index azt jelzi, a termelési értéket *melyik időszak egységárain* számították;
- k: az időszakok száma.

Az A_{ij} aggregátum nem más, mint az i-edik időszak termelési értéke a j-edik időszak egységárain számítva.

Aggregátummátrix

Aggregátummátrix: $A = \left[A_{ij}\right]$.

A főátlója: *folyóáras* aggregátumok.

Az A mátrix főátlójában található aggregátumok hányadosai értékindexeket, valamely adott oszlop aggregátumainak hányadosai voumenindexeket, valamely sor adataiból számított hányadosok pedig árindexeket adnak.

Példa

Magyarországon a háztartásokban az egy főre jutó fogyasztás értékei 1997–2000 között.

Fogyasztás éve	1997	1998	1999	2000
	évi árakon (ezer Ft)			
1997	219	223	238	251
1998	240	244	261	275
1999	261	266	284	299
2000	288	293	313	330

Keresztély, Sugár, Szarvas (2005, Gy.121 feladat, 111. old.)

Indexsorok

1. Ha a főátlóban vagy valamely adott sorban/oszlopban lévő minden aggregátumot egy bázisidőszaknak választott időszak aggregátumával osztunk, akkor bázisindexsorokat kapunk, ha pedig minden aggregátumot a főátlóban vagy valamely adott sorban/ oszlopban közvetlenül előtte található aggregátummal osztunk, akkor láncindexsorokhoz jutunk.

2. Ha a volumen- és árindexsorok esetében az indexsor minden egyes tagját ugyanazon oszlop vagy sor aggregátumaiból számítjuk, akkor állandó súlyú indexsorokhoz jutunk, ha viszont az indexsor minden egyes tagját más-más oszlopban/sorban található két aggregátum hányadosaként számítjuk, akkor változó súlyú indexsorokat kapunk.

3. A változó súlyú láncvolumen- vagy láncárindexek tagjairól mindig egyértelműen eldönthető, azok Laspeyres- vagy Paache-féle indexek-e. Meg kell nézni, az adott ,,láncszem" milyen súlyozású.

4. Egymást követő *láncindexek szorzata* (érték vagy állandó súlyú volumen ill. árindexek esetén) *bázisindexet* ad.

Területi indexek

Területi indexek: indexszámításban az időszakok szerepét *területi* egységek veszik át, a kapcsolódó aggregátumokat hasonlítjuk össze.

Sajátosságai:

- A területi egységeknek nincs sorrendje. Fontos, hogy az azonos relációjú összehasonlítások azonos eredményhez vezessenek.
- Adott ország régiói vagy azonos valutájú országok esetén a területi indexek jelentése ugyanaz, mint az eddigi indexeké. Eltérő valutáju országok: nemzetközi indexek. Az egyes országok aggregátumai más valutában vannak kifejezve.

Vásárlóerő-paritás: nemzetközi árindex.

Jelölése: PPP (Purchasing Power Parity)

A $PPP(A/B)$-vel jelölt A/B relációjú nemzetközi árindex azt mutatja, hogy B ország egy valutaegysége A ország hány valutaegységével egyenértékű, ha azt a vizsgált termékek megvásárlására fordítjuk.

Vásárlóerő-paritás

Egyedi vásárlóerő-paritás: p_A/p_B, azaz a B ország egy valutaegysége hány A valutát ér.

A *nemzetközi árindex* az *egyedi vásárlóerő-paritások* súlyozott átlaga.

Példa. Big-Mac index (Economist, 2022 július).

Egy Big Mac átlagára Magyarországon 1030 HUF, az Egyesült Államokban 5.15 USD.

1 USD vásárlóereje 1030/5.15=200.00 HUF vásárlóerejével egyezik.

HUF/USD árfolyam: 389.05 HUF/1 USD.

2022 júliusában a forint **48.6 %**-kal volt alulértékelve a dollárral szemben.

2020 júliusához képest (40.0 % alulértékelés) **8.6 %** esés.

2010 júliusához képest (5.6 % alulértékelés) **43.0 %** esés.

Az összehasonlítandó országok lakosságszáma, területe, mérete általában jelentősen különböző.
Megoldás: *egy főre vetített* aggregátumokat vizsgálnak.

Bilaterális összehasonlítás

A, B: két eltérő valutájú ország.

$$PPP^A(A/B) = \frac{\sum q_A p_A}{\sum q_A p_B} = \frac{\sum q_A p_B \cdot \frac{p_A}{p_B}}{\sum q_A p_B} = \frac{\sum q_A p_A}{\sum \frac{q_A p_A}{p_A/p_B}},$$

$$PPP^B(A/B) = \frac{\sum q_B p_A}{\sum q_B p_B} = \frac{\sum q_B p_B \cdot \frac{p_A}{p_B}}{\sum q_B p_B} = \frac{\sum q_B p_A}{\sum \frac{q_B p_A}{p_A/p_B}}.$$

A Paache- és a Laspeyres-féle árindexek megfelelői. A Fisher index megfelelője:

$$PPP^F(A/B) = \sqrt{PPP^A(A/B) \cdot PPP^B(A/B)} = 1/PPP^F(B/A).$$

Árszínvonalindex (Price Level Index): $PLI(A/B) = \frac{PPP(A/B)}{ER(A/B)}$.

$ER(A/B)$: valutaárfolyam (Exchange Rate).

Az *árszínvonalindex* megadja, hogy az A ország árszínvonala hányszorosa a B ország árszínvonalának, ha az egységárakat azonos valutában fejezzük ki az $ER(A/B)$ átváltási kulcs segítségével.

Sokaságok megadása

Egyetlen ismérv szerint vizsgált sokaság megadásának módjai.

- *Véges* N elemű sokaság esetén az elemek felsorolásával:

$$Y_1, Y_2, \ldots, Y_N.$$

- *Végtelen* sokaság esetén *diszkrét* ismérvre megadjuk a

$$\mathrm{P}(Y = y_k) = P_k$$

eloszlást vagy az

$$F(y) = \mathrm{P}(Y < y)$$

eloszlásfüggvényt, *folytonos* ismérvre pedig vagy az $F(y)$ eloszlásfüggvényt, vagy a $f(y)$ sűrűségfüggvényt, melyre

$$F(y) = \int_{-\infty}^{y} f(t)\mathrm{d}t.$$

Sokasági várható érték és szórásnégyzet

Elemeivel adott sokaság esetén:

$$\mathrm{E}(Y) = \overline{Y} = \frac{1}{N}\sum_{i=1}^{n} Y_i = \mu, \quad \mathrm{Var}(Y) = \frac{1}{N}\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sigma^2.$$

Eloszlásával adott sokaság esetén diszkrét esetben:

$$\mathrm{E}(Y) = \sum_k y_k \mathrm{P}(Y = y_k) = \mu,$$
$$\mathrm{Var}(Y) = \sum_k \big(y_k - \mathrm{E}(Y)\big)^2 \mathrm{P}(Y = y_k) = \sigma^2;$$

folytonos esetben:

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y f(y)\mathrm{d}y = \mu, \quad \mathrm{Var}(Y) = \int_{-\infty}^{\infty} \big(y - \mathrm{E}(Y)\big)^2 f(y)\mathrm{d}y = \sigma^2.$$

Minta

Minta (n elemű):

$$\mathbf{y} = (y_1, y_2, \ldots, y_n).$$

Elemeivel megadott sokaság esetén egy n elemű minta elemei a sokaság elemei közül kerülnek ki.

Véletlen mintavétel: minden sokasági elem előre megadott valószínűséggel kerül a mintába.

Eloszlásával megadott sokaság esetén a mintaelemek a megadott eloszlású valószínűségi változók.

A mintalemek a *mintavétel előtt* valószínűségi változóknak tekinthetőek. A *mintavétel után* megkapjuk a minta egy realizációját, ami konkrét (szám)értékeket tartalmaz.

A mintából számított tetszőleges *mintajellemző* (pl. átlag, szórás, kvantilisek) szintén valószínűségi változó.

A mintajellemzők eloszlását mintavételi eloszlásnak nevezzük.

Független, azonos eloszlású (FAE) minta

Véges homogén sokaság esetén FAE mintát kapunk, ha minden sokasági elemet azonos valószínűséggel kiválasztva veszünk *visszatevéses* mintát.

Nagyon nagy sokaság esetén a *visszatevés nélküli* mintavétel is közel FAE mintát ad.

Eloszlásával megadott sokaság esetén a FAE minta független, a megadott eloszlással bíró valószínűségi változókból áll.

$y_1, y_2, \ldots, y_n$: FAE minta egy sokaságból, melynek várható értéke és szórása rendre μ és σ.

$\overline{y} = \frac{1}{n}(y_1 + \ldots + y_n)$: mintaátlag.

$$\mathsf{E}(\overline{y}) = \frac{1}{n}\big(\mathsf{E}(y_1) + \ldots + \mathsf{E}(y_n)\big) = \frac{1}{n}\big(\underbrace{\mu + \ldots + \mu}_{n \text{ darab}}\big) = \mu,$$

$$\mathsf{Var}(\overline{y}) = \sigma_{\overline{y}}^2 = \frac{1}{n^2}\big(\mathsf{Var}(y_1) + \ldots + \mathsf{Var}(y_n)\big) = \frac{1}{n^2}\big(\underbrace{\sigma^2 + \ldots + \sigma^2}_{n \text{ darab}}\big) = \frac{\sigma^2}{n}.$$

$\sigma_{\overline{y}} = \sigma/\sqrt{n}$: standard hiba.

Egyszerű véletlen (EV) minta

Egyszerű véletlen mintavételt használunk homogén, véges elemszámú sokaság esetén, amikor a mintát *visszatevés nélkül* választjuk ki, minden lehetséges n elemű minta kiválasztásának azonos valószínűséget biztosítva.

$y_1, y_2, \ldots, y_n$: EV minta egy N elemű sokaságból, melynek várható értéke és szórása rendre μ és σ.

$$\mathrm{E}(\overline{y}_{EV}) = \mu, \qquad \mathrm{Var}(\overline{y}_{EV}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right).$$

Szisztematikus kiválasztás

n elemű EV mintát akarunk venni N elemű sokaságból.

$k = N/n$: lépésköz.

k_0: véletlen kiindulópont ($1 \leq k_0 \leq k$).

Ciklikusan haladva a k_0-ból kiindulva minden k-adik elemet kiválasztunk. Ha a sokaság a vizsgált ismérv szerint véletlenszerűen van rendezve, akkor EV mintához jutunk.

Rétegzett (R) minta

Heterogén sokaság jellemzőit vizsgáljuk, pl. a lakosság jövedelmi viszonyait vagy iskolázottságát a lakhely jellege szerint (Budapest, megyeszékhely, stb).

A rétegzett mintavétel végrehajtása úgy történik, hogy először a sokaságot többé-kevésbé homogén rétegekbe soroljuk be, úgy, hogy a rétegek átfedésmentesen és teljesen lefedjék a sokaságot, majd az egyes rétegeken belül, egymástól függetlenül EV (ritkábban FAE) mintavételt hajtunk végre.

M: rétegek száma;

$N_1, N_2, \ldots, N_M$: az egyes rétegek elemszáma, $\sum_{j=1}^{M} N_j = N$;

$n_1, n_2, \ldots, n_M$: az egyes rétegekből kiválasztott minták elemszáma, $\sum_{j=1}^{M} n_j = n$;

$\mu_1, \mu_2, \ldots, \mu_M$: az egyes rétegek várható értékei;

$\sigma_1, \sigma_2, \ldots, \sigma_M$: az egyes rétegek szórásai.

Mintavételi tervek

a) Egyenletes elosztás. Minden rétegből azonos elemszámú mintát veszünk: $n_i = n/M$.

Egyszerű, az egyes rétegek jellemzői könnyen számolhatóak.

b) Arányos elosztás. Az egyes rétegekből vett minták elemszámai úgy aránylanak egymáshoz, mint maguknak a rétegeknek az elemszámai:

$$n_j = n\frac{N_j}{\sum_{k=1}^{M} N_k} n\frac{N_j}{N}.$$

Egyszerű, a mintában ugyanazok a súlyarányok, mint a sokaságban.

c) Neyman-féle optimális elosztás. A nagyobb szóródású rétegekből nagyobb mintákat veszünk:

$$n_j = n\frac{N_j\sigma_j}{\sum_{k=1}^{M} N_k\sigma_k}.$$

A főátlag becslésénél a mintavételi hiba minimális. Problémás a megvalósítás, mert a σ_j szórások általában nem ismertek.

Csoportos minták

Egylépcsős (1L) minta

Az összes sokasági elem nem áll rendelkezésünkre (vagy csak nagyon drágán), de nagyobb összetartozó csoportokról van listánk.

Egylépcsős (csoportos) mintavétel esetén a csoportok halmazából választunk EV mintát, majd az így kiválasztott csoportokat teljeskörűen megfigyeljük.

Minél homogénebbek az egyes csoportok, annál kevésbé hatékony az eljárás.

Többlépcsős (TL) minta

Minden egyes lépcsőben a korábban kiválasztott csoportokból veszünk újabb mintát. Például a kétlépcsős mintavétel is kevesebb redundáns mintaelemet tartalmaz, mind az egylépcsős.

Becslőfüggvény

$y_1, y_2, \ldots, y_n$: minta (FAE, ritkábban EV).

Statisztika: a mintaelemek tetszőleges függvénye. Valószínűségi változó.

A becslőfüggvény olyan statisztika, ami valamely sokasági jellemző mintából történő közelítő meghatározására szolgál.

θ: becsülni kívánt sokasági jellemző. Becslőfüggvénye:

$$\hat{\theta}(y_1, y_2, \ldots, y_n) = \hat{\theta}(n) = \hat{\theta}.$$

Példa. $\theta = \sigma^2$, azaz a sokasági szórásnégyzetet becsüljük.

$$\hat{\theta}_1 = \hat{\theta}_1(n) = \hat{\theta}_1(y_1, y_2, \ldots, y_n) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = s^{*2},$$

$$\hat{\theta}_2 = \hat{\theta}_2(n) = \hat{\theta}_2(y_1, y_2, \ldots, y_n) = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = s^2.$$

s^2: korrigált tapasztalati (empirikus) szórásnégyzet.

A becslőfüggvény tulajdonságai. Torzítatlanság

Egy becslőfüggvényt torzítatlannak nevezünk, ha annak várható értéke megegyezik a becsülni kívánt sokasági jellemzővel, azaz

$$\mathsf{E}\big(\hat{\theta}\big) = \theta.$$

Példa. $\theta = \mu$, azaz a sokasági várható értéket becsüljük.

$$\hat{\theta}_1 = \hat{\mu}_1 = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Mind FAE, mind EV minta esetén: $\mathsf{E}\big(\hat{\mu}_1\big) = \mathsf{E}\big(\overline{y}\big) = \mu$.

A mintaátlag a sokasági várható érték torzítatlan becslése.

$$\hat{\theta}_2 = \hat{\mu}_2 = \big(y_1 + y_n\big)/2.$$

Mind FAE, mind EV minta esetén: $\mathsf{E}\big(\hat{\mu}_2\big) = \mu$ (torzítatlan).

Torzítás (bias):

$$\mathsf{Bs}\left(\hat{\theta}\right) = \mathsf{E}\big(\hat{\theta}\big) - \theta.$$

Példa

$\theta = \sigma^2$, azaz a sokasági szórásnégyzetet becsüljük.

$$\mathrm{E}\left(s^{*2}\right) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2.$$

Torzítás:

$$\mathrm{Bs}\left(s^{*2}\right) = \left(\sigma^2 - \frac{\sigma^2}{n}\right) - \sigma^2 = -\frac{\sigma^2}{n}.$$

Korrigált empirikus szórásnégyzet:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \frac{n}{n-1}s^{*2}.$$

$$\mathrm{E}\left(s^2\right) = \mathrm{E}\left(\frac{n}{n-1}s^{*2}\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2.$$

A korrigált empirikus szórásnégyzet a sokasági szórásnégyzet torzítatlan becslése.

Mintavételi szórásnégyzet

$\hat{\theta}$: a θ torzítatlan becslése, azaz $\mathsf{E}(\hat{\theta}) = \theta$.

A becslőfüggvény szórásnégyzetét mintavételi szórásnégyzetnek, ennek négyzetgyökét pedig a becslőfüggvény, illetve a becslés standard hibájának (standard error) nevezzük.

$$\mathsf{Se}\left(\hat{\theta}\right) = \sqrt{\mathsf{Var}\left(\hat{\theta}\right)}.$$

Példa. $\theta = \mu$, azaz a sokasági várható értéket becsüljük. FAE minta σ^2 sokasági szórásnégyzettel.

$$\hat{\theta}_1 = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \hat{\theta}_2 = \hat{\mu}_2 = \left(y_1 + y_n\right)/2.$$

Mindkét becslés torzítatlan.

$$\mathsf{Var}\left(\hat{\theta}_1\right) = \sigma^2/n, \qquad \mathsf{Var}\left(\hat{\theta}_2\right) = \sigma^2/2,$$
$$\mathsf{Se}\left(\hat{\theta}_1\right) = \sigma/\sqrt{n}, \qquad \mathsf{Se}\left(\hat{\theta}_2\right) = \sigma/\sqrt{2}.$$

A becslőfüggvény tulajdonságai. Hatásosság

θ egy olyan $\hat{\theta}_0$ torzítatlan becslőfüggvényét, melynek szórásnégyzete θ tetszőleges torzítatlan becslőfüggvénye szórásnégyzeténél nem nagyobb, θ minimális szórásnégyzetű torzítatlan becslőfüggvényének (MVUE, Minimum Variance Unbiased Estimator) nevezzük.

$\hat{\theta}_1$, $\hat{\theta}_2$: a θ torzítatlan becslőfüggvényei. A $\hat{\theta}_1$ becslőfüggvénynek a $\hat{\theta}_2$-re vonatkozó relatív hatásfoka:

$$\mathrm{Ef}_r = \frac{\mathrm{Var}\left(\hat{\theta}_1\right)}{\mathrm{Var}\left(\hat{\theta}_2\right)}.$$

$\mathrm{Ef}_r > 1$: $\hat{\theta}_2$ hatásosabb, mint $\hat{\theta}_1$.

Ha létezik $\hat{\theta}_0$ MVUE, akkor $\hat{\theta}_1$ abszolút hatásfoka:

$$\mathrm{Ef}_a = \frac{\mathrm{Var}\left(\hat{\theta}_1\right)}{\mathrm{Var}\left(\hat{\theta}_0\right)} \geq 1.$$

Példa

$\theta = \mu$, azaz a sokasági várható értéket becsüljük. FAE minta σ^2 sokasági szórásnégyzettel.

$$\hat{\theta}_1 = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \hat{\theta}_2 = \hat{\mu}_2 = \big(y_1 + y_n\big)/2.$$

Mindkét becslés torzítatlan.

$$\mathrm{Var}\left(\hat{\theta}_1\right) = \sigma^2/n, \qquad \mathrm{Var}\left(\hat{\theta}_2\right) = \sigma^2/2.$$

Relatív hatásfok:

$$\mathrm{Ef}_r = \frac{\mathrm{Var}\left(\hat{\theta}_2\right)}{\mathrm{Var}\left(\hat{\theta}_1\right)} = \frac{\sigma^2/2}{\sigma^2/n} = \frac{n}{2} > 1, \quad \text{ha } n > 2.$$

MSE kritérium. Aszimptotikus torzítatlanság

Ha két nem feltétlenül torzítatlan becslés összehasonlítására az átlagos négyzetes hiba (MSE, Mean Squared Error) szolgál.

$$\mathrm{Mse}\left(\hat{\theta}\right) = \mathrm{E}\left(\hat{\theta} - \theta\right)^2 = \mathrm{Var}\left(\hat{\theta}\right) + \mathrm{Bs}^2\left(\hat{\theta}\right).$$

Az a becslőfüggvény a ,,jobb", amelyiknek az *átlagos négyzetes hibája* kisebb.

$\hat{\theta} = \hat{\theta}(n)$ aszimptotikusan torzítatlan, ha

$$\lim_{n\to\infty} \mathrm{Bs}\left(\hat{\theta}(n)\right) = 0.$$

Példa

$$\mathrm{Bs}\left(s^{*2}\right) = -\frac{\sigma^2}{n} \to 0, \qquad \text{ha } n \to \infty.$$

s^{*2} torzított, de aszimptotikusan torzítatlan.

A becslőfüggvény tulajdonságai. Konzisztencia

A θ paraméter $\hat{\theta} = \hat{\theta}(n)$ becslőfüggvénye konzisztens, ha $n \to \infty$ esetén várható értéke tart a valódi paraméterértékhez, szórásnégyzete pedig tart nullához, azaz

$$\lim_{n\to\infty} \mathrm{E}\big(\hat{\theta}(n)\big) = \theta \qquad \text{és} \qquad \lim_{n\to\infty} \mathrm{Var}\,\big(\hat{\theta}(n)\big) = 0.$$

Példa. $\theta = \mu$, FAE minta σ^2 sokasági szórásnégyzettel.

$$\mathrm{E}\big(\overline{y}\big) = \mu, \qquad \mathrm{Var}\,\big(\overline{y}\big) = \frac{\sigma^2}{n} \to 0, \quad \text{ha } n \to \infty.$$

FAE minta esetén a mintaátlag a sokasági várható érték konzisztens (és torzítatlan) becslése.

Becslési módszerek. Momentumok módszere

$y_1, y_2, \ldots, y_n$: FAE minta egy θ paraméterű Y valószínűségi változóra.

A körüli r-edik elméleti momentum:

$$\mathcal{M}_r(A) = \mathrm{E}(Y - A)^r.$$

A körüli r-edik empirikus momentum:

$$M_r(A) = \frac{1}{n} \sum_{i=1}^{n} (y_i - A)^r.$$

Ismert típusú eloszlás esetén a momentumok az eloszlás paramétereinek függvényei. Ezekbe a függvényekbe behelyettesítve az empirikus momentumokat megkapjuk a paraméterek becsléseit.

Exponenciális eloszlás $\lambda > 0$ paraméterrel

$Y \sim \mathrm{Exp}(\lambda)$. Y sűrűségfüggvénye:

$$f(y) = \begin{cases} \lambda \mathrm{e}^{-\lambda y}, & \text{ha } y > 0; \\ 0, & \text{ha } y \leq 0. \end{cases}$$

$$\mathcal{M}_r(0) = \mathrm{E}\big(Y^r\big) = \frac{r!}{\lambda^r}, \quad \text{azaz} \quad \mathcal{M}_1(0) = \mathrm{E}(Y) = \frac{1}{\lambda}.$$

$y_1, y_2, \ldots, y_n$: FAE minta Y-ra, $M_1(0) = \overline{y}$.

λ becslése az $\mathcal{M}_1(0) = M_1(0)$, azaz az $1/\lambda = \overline{y}$ egyenlet megoldása:

$$\hat{\lambda} = \frac{1}{\overline{y}}.$$

Egyenletes eloszlás az $[a, b]$ intervallumon

$Y \sim U(a, b)$. Y sűrűségfüggvénye:

$$f(y) = \begin{cases} \frac{1}{b-a}, & \text{ha } y \in [a, b]; \\ 0, & \text{ha } y \notin [a, b]. \end{cases}$$

$$\mathcal{M}_1(0) = \mathrm{E}(Y) = (a+b)/2, \qquad \mathcal{M}_2\big(\mathrm{E}(Y)\big) = \mathrm{Var}(Y) = (b-a)^2/12.$$

$y_1, y_2, \ldots, y_n$: FAE minta Y-ra.

$$M_1(0) = \overline{y}, \qquad M_2(\overline{y}) = \frac{1}{N}\sum_{i=1}^{n}(y_i - \overline{y})^2 = s^{*2}.$$

(a, b) becslése $(a < b)$ az $\mathcal{M}_1(0) = M_1(0)$, $\mathcal{M}_2\big(\mathrm{E}(Y)\big) = M_2(\overline{y})$, azaz az

$$(a+b)/2 = \overline{y}, \qquad (b-a)^2/12 = s^{*2}$$

egyenletrendszer megoldása:

$$\hat{a} = \overline{y} - \sqrt{3s^{*2}}, \qquad \hat{b} = \overline{y} + \sqrt{3s^{*2}}.$$

Normális eloszlás μ, σ^2 paraméterekkel

$Y \sim \mathcal{N}(\mu, \sigma^2)$. Y sűrűségfüggvénye:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

$$\mathcal{M}_1(0) = \mathrm{E}(Y) = \mu, \qquad \mathcal{M}_2\big(\mathrm{E}(Y)\big) = \mathrm{Var}(Y) = \sigma^2.$$

$y_1, y_2, \ldots, y_n$: FAE minta Y-ra.

$$M_1(0) = \overline{y}, \qquad M_2(\overline{y}) = \frac{1}{N}\sum_{i=1}^{n}(y_i - \overline{y})^2 = s^{*2}.$$

(μ, σ^2) becslése az $\mathcal{M}_1(0) = M_1(0)$, $\mathcal{M}_2\big(\mathrm{E}(Y)\big) = M_2(\overline{y})$, azaz az

$$\mu = \overline{y}, \qquad \sigma^2 = s^{*2}$$

egyenletrendszer megoldása:

$$\hat{\mu} = \overline{y}, \qquad \widehat{\sigma^2} = s^{*2}.$$

Becslési módszerek. Maximum likelihood (ML) módszer

$y_1, y_2, \ldots, y_n$: FAE minta egy θ paraméterű Y valószínűségi változóra.

A minta $L(\theta; y_1, \ldots, y_n)$ likelihood függvénye diszkrét esetben a minta *együttes eloszlása*, folytonos esetben a minta *együttes sűrűségfüggvénye*.

Log-likelihood függvény: $\ell(\theta; y_1, \ldots, y_n) = \log L(\theta; y_1, \ldots, y_n)$.

A θ paraméter $\hat{\theta}$ ML becslése az

$$L(\theta; y_1, \ldots, y_n)$$

likelihood vagy az

$$\ell(\theta; y_1, \ldots, y_n)$$

log-likelihood függvény maximum helye.

Sokasági arány becslése

20-szor feldobunk egy érmét, y_i az i-edik dobás kimenetele.

$y_i = 1$, ha fejet dobunk és $y_i = 0$, ha írást. A minta realizációja:

$$0,\ 1,\ 0,\ 0,\ 1,\ 1,\ 1,\ 0,\ 0,\ 1,\ 1,\ 1,\ 1,\ 1,\ 0,\ 1,\ 1,\ 1,\ 0,\ 1.$$

Becsülendő a fej dobás p valószínűsége.

$$\begin{aligned} L(p; y_1, \dots, y_{20}) &= \mathrm{P}(y_1 = 0, y_2 = 1, \dots, y_{20} = 1 | p) = \mathrm{P}(y_1 = 0|p) \cdot \mathrm{P}(y_2 = 1|p) \cdot \ldots \cdot \mathrm{P}(y_{20} = 1|p) \\ &= (1-p)p(1-p)^2p^3(1-p)^2p^5(1-p)p^3(1-p)p = p^{13}(1-p)^7, \end{aligned}$$

$$\ell(p; y_1, \dots, y_{20}) = 13 \log p + 7 \log(1-p).$$

$$\frac{\partial \ell}{\partial p} = \frac{13}{p} - \frac{7}{1-p} = 0, \quad \text{azaz} \quad p = \frac{13}{20}.$$

$$\frac{\partial^2 \ell}{\partial p^2} = \frac{13}{p^2} - \frac{7}{(1-p)^2} < 0, \quad \text{azaz} \quad p = \frac{13}{20} \quad \text{maximumhely.}$$

$\hat{p} = 13/20$: a fej dobások *relatív gyakorisága*.

Exponenciális eloszlás $\lambda > 0$ paraméterrel

$Y \sim \mathrm{Exp}(\lambda)$. Y sűrűségfüggvénye:

$$f(y) = \begin{cases} \lambda \mathrm{e}^{-\lambda y}, & \text{ha } y > 0; \\ 0, & \text{ha } y \le 0. \end{cases}$$

$y_1, y_2, \ldots, y_n$: FAE minta Y-ra.

$$L(\lambda; y_1, \ldots, y_n) = \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda y_i} = \lambda^n \mathrm{e}^{-\lambda \sum_{i=1}^{n} y_i},$$

$$\ell(\lambda; y_1, \ldots, y_n) = n \log \lambda - \lambda \sum_{i=1}^{n} y_i = n \log \lambda - \lambda n \overline{y}.$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - n\overline{y} = 0, \quad \text{azaz} \quad \lambda = \frac{1}{\overline{y}}.$$

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0, \quad \text{azaz} \quad \lambda = \frac{1}{\overline{y}} \quad \text{maximumhely.}$$

További példák

Normális eloszlás μ, σ^2 paraméterekkel

Sűrűségfüggvénye:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

$y_1, y_2, \ldots, y_n$: FAE minta.

$$\hat{\mu} = \overline{y}, \qquad \widehat{\sigma^2} = s^{*2}.$$

Egyenletes eloszlás az $[a, b]$ intervallumon

Sűrűségfüggvénye:

$$f(y) = \begin{cases} \frac{1}{b-a}, & \text{ha } y \in [a, b]; \\ 0, & \text{ha } y \notin [a, b]. \end{cases}$$

$y_1, y_2, \ldots, y_n$: FAE minta.

$$\hat{a} = \min\{y_1, y_2, \ldots, y_n\}, \qquad \hat{b} = \max\{y_1, y_2, \ldots, y_n\}.$$

Egyenlőtlenségek

Markov egyenlőtlenség: Legyen $Y \geq 0$ egy valószínűségi változó, aminek létezik $\mathsf{E}(Y)$ várható értéke. Ekkor bármely $\delta > 0$ esetén

$$\mathsf{P}\big(Y \geq \delta\big) \leq \frac{\mathsf{E}(Y)}{\delta}.$$

Csebisev egyenlőtlenség: Legyen Y egy valószínűségi változó, aminek létezik $\mathsf{E}(Y)$ várható értéke. Ekkor bármely $\epsilon > 0$ esetén

$$\mathsf{P}\Big(\big|Y - \mathsf{E}(Y)\big| \geq \varepsilon\Big) \leq \frac{\mathsf{Var}(Y)}{\varepsilon^2}.$$

Példa. Hányszor kell egy szabályos kockát feldobnunk, hogy a hatos dobás valószínűségét az esemény relatív gyakorisága legalább 0.8 valószínűséggel 0.1-nél kisebb hibával megközelítse? Mi a helyzet, ha nem tudjuk, hogy a kocka szabályos-e?

Nagy számok gyenge törvénye

Azt mondjuk, hogy valószínűségi változók egy $Y_1, Y_2, \ldots, Y_n, \ldots$ sorozata sztochasztikusan konvergál egy Y valószínűségi változóhoz, ha bármely $\varepsilon > 0$ esetén

$$\lim_{n\to\infty} \mathrm{P}\big(|Y_n - Y| \geq \varepsilon\big) = 0.$$

Ha $Y \equiv c$ (konstans), akkor elegendő:

$$\lim_{n\to\infty} \mathrm{E}(Y_n) = c, \qquad \lim_{n\to\infty} \mathrm{Var}(Y_n) = 0.$$

Nagy számok gyenge törvénye: Legyenek $Y_1, Y_2, \ldots, Y_n, \ldots$ páronként független, azonos eloszlású valószínűségi változók, legyen $\mathrm{E}(Y_1) = \mu$ és $\mathrm{Var}(Y_1) < \infty$, továbbá legyen

$$S_n = Y_1 + Y_2 + \ldots + Y_n.$$

Ekkor $\overline{Y} = S_n/n$ sztochasztikusan konvergál a μ várható értékhez.

Az átlag a várható érték *konzisztens* becslése.

Központi határeloszlás tétel

Központi határeloszlás tétel: Legyenek $Y_1, Y_2, \ldots, Y_n, \ldots$ független, azonos eloszlású valószínűségi változók, $\mathsf{E}(Y_1) = \mu$ és $0 < \mathsf{Var}(Y_1) = \sigma^2 < \infty$, továbbá legyen

$$S_n = Y_1 + \ldots + Y_n.$$

Ekkor $\mathsf{E}(S_n) = n \cdot \mu$, $\mathsf{Var}(S_n) = n \cdot \sigma^2$ és

$$\lim_{n\to\infty} \mathsf{P}\left(\frac{S_n - n\cdot\mu}{\sqrt{n}\cdot\sigma} < x\right) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\mathrm{d}t, \quad x \in \mathbb{R}.$$

Nagy n esetén $\overline{Y} = S_n/n$ eloszlása hozzávetőlegesen normális μ várható értékkel és σ^2/n szórásnégyzettel.

Normálisból származtatható eloszlások. Khi-négyzet eloszlás

$X_1, X_2, \ldots, X_n$: független standard normális valószínűségi változók.

$$Y = X_1^2 + X_2^2 + \ldots + X_n^2 \geq 0.$$

Y eloszlása n szabadsági fokú khi-négyzet (chi-square) eloszlás. Jelölés: $Y \sim \mathcal{X}_n^2$.

Várható értéke: n; szórásnégyzete: $2n$.

p-kvantilis: $\mathcal{X}_p^2(n)$. Ha $Y \sim \mathcal{X}_n^2$, akkor $\mathrm{P}\big(Y < \mathcal{X}_p^2(n)\big) = p$. Táblázatból kiolvasható.

Ha $y_1, y_2, \ldots, y_n$ FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból, akkor

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{és} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

függetlenek, valamint

$$\overline{y} \sim \mathcal{N}\big(\mu, \sigma^2/n\big) \qquad \text{és} \qquad \frac{(n-1)s^2}{\sigma^2} \sim \mathcal{X}_{n-1}^2.$$

Normálisból származtatható eloszlások. t-eloszlás

$X_0, X_1, \ldots, X_n$: független standard normális valószínűségi változók.

$$Y = \frac{\sqrt{n} X_0}{\sqrt{X_1^2 + X_2^2 + \ldots + X_n^2}}.$$

Y eloszlása n szabadsági fokú t-eloszlás (Student-eloszlás). Jelölés: $Y \sim t_n$.

Várható értéke: 0, ha $n > 1$; szórásnégyzete: $n/(n-2)$, ha $n > 2$.

p-kvantilis: $t_p(n)$. Ha $Y \sim t_n$, akkor $\mathrm{P}\big(Y < t_p(n)\big) = p$. Táblázatból kiolvasható.

Ha $n \to \infty$, akkor $t_n \to \mathcal{N}(0,1)$ (standard normális). $n = \infty$ eset.

Ha $n \to \infty$, akkor $t_p(n) \searrow z_p$, ahol z_p a standard normális eloszlás p-kvatilise.

Ha $y_1, y_2, \ldots, y_n$ FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból, akkor

$$\frac{\overline{y} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Normálisból származtatható eloszlások. F-eloszlás

X_1, X_2: független khi-négyzet eloszlású valószínűségi változók n és m szabadsági fokkal.

$$Y = \frac{X_1/n}{X_2/m} \geq 0.$$

Y eloszlása n és m szabadsági fokú F-eloszlás. Jelölés: $Y \sim F_{n,m}$.

Várható értéke: $\frac{m}{m-2}$, ha $m > 2$; szórásnégyzete: $\frac{2m(n+m-2)}{n(m-2)^2(m-4)}$, ha $m > 4$.

p-kvantilis: $F_p(n; m)$. Ha $Y \sim F_{n,m}$, akkor $\mathrm{P}\big(Y < F_p(n; m)\big) = p$. Táblázatból kiolvasható.

Ha $x_1, x_2, \ldots, x_n$ és $y_1, y_2, \ldots, y_m$ FAE minták $\mathcal{N}(\mu_x, \sigma^2)$ és $\mathcal{N}(\mu_y, \sigma^2)$ eloszlásból, valamint

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{és} \qquad s_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_j - \overline{y})^2,$$

akkor

$$s_x^2/s_y^2 \sim F_{n-1,m-1}.$$

Az intervallumbecslés alapjai

$y_1, y_2, \ldots, y_n$: minta

θ: becsülni kívánt sokasági jellemző.

Intervallumbecslés esetén a minta alapján olyan intervallumot határozunk meg, amely előre megadott (nagy) valószínűséggel tartalmazza az ismeretlen jellemzőt. Ezt az intervallumot konfidencia intervallumnak nevezzük.

$0 < \alpha < 1$: adott érték (jellemzően $\alpha \leq 0.2$).

Keresünk olyan $\hat{\theta}_{a(\alpha)}$ és $\hat{\theta}_{f(\alpha)}$ becslőfüggvényeket, melyekre

$$\mathrm{P}\big(\hat{\theta}_{a(\alpha)} < \theta < \hat{\theta}_{f(\alpha)}\big) = 1 - \alpha.$$

$\hat{\theta}_{a(\alpha)}$, $\hat{\theta}_{f(\alpha)}$: az $(1-\alpha)\cdot 100\%$-os megbízhatóságú konfidencia intervallum alsó és felső határai. Például $\alpha = 0.05$: 95%-os megbízhatóságú konfidencia intervallum.

A minta egy konkrét *realizációját* behelyettesítve a $\hat{\theta}_{a(\alpha)}$ és $\hat{\theta}_{f(\alpha)}$ becslőfüggvénybe egy ,,konkrét" intervalumot kapunk.

Normális eloszlás, ismert szórás

$y_1, y_2, \ldots, y_n$: FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból, σ ismert.

$\theta = \mu$: becsülni kívánt sokasági jellemző.

$\overline{y} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$, ezért

$$Z = \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Ha (z_1, z_2) egy intervallum, és $\Phi(z)$ a standard normális eloszlás eloszlásfüggvénye, akkor

$$\mathsf{P}\left(z_1 < \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} < z_2\right) = \Phi(z_2) - \Phi(z_1).$$

Olyan intervallumot keresünk, hogy az intervallumon kívül esés valószínűsége mindkét oldalon egyenlő $(\alpha/2)$ legyen.

Z eloszlása szimmetrikus, $\hat{\theta}_{a(\alpha)}$ és $\hat{\theta}_{f(\alpha)}$ a mintaátlagra nézve szimmetrikusan helyezkednek el.

Normális eloszlás, ismert szórás

Adott $z \in \mathbb{R}$ esetén a $(-z, z)$ intervallumba esés valószínűsége:

$$\mathrm{P}\left(-z < \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} < z\right) = \Phi(z) - \Phi(-z) = \Phi(z) - \big(1 - \Phi(z)\big) = 2\Phi(z) - 1.$$

Adott $0 < \alpha < 1$ esetén keressük azt a z értéket, melyre

$$\mathrm{P}\left(-z < \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} < z\right) = 2\Phi(z) - 1 = 1 - \alpha \iff \Phi(z) = 1 - \alpha/2.$$

Megoldás: $z = z_{1-\alpha/2}$, a standard normális eloszlás $p = 1 - \alpha/2$ rendű kvantilise. Táblázatból meghatározható.

Például: $\alpha = 0.05,\ 1 - \alpha/2 = 0.975,\ z_{0.975} = 1.9600,$
$\alpha = 0.10,\ 1 - \alpha/2 = 0.950,\ z_{0.950} = 1.6449.$

$$\mathrm{P}\left(\overline{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Normális eloszlás, ismert szórás

Alsó határ: $\hat{\theta}_{a(\alpha)} = \overline{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ (valószínűségi változó).

Felső határ: $\hat{\theta}_{f(\alpha)} = \overline{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ (valószínűségi változó).

Hibahatár: $\Delta_{\overline{y}} = \Delta = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Ismételt mintavétel esetén az esetek átlagosan $(1-\alpha)\cdot 100$ százalékában igaz, hogy a $\left(\hat{\theta}_{a(\alpha)}, \hat{\theta}_{f(\alpha)}\right)$ intervallum lefedi (tartalmazza) a keresett sokasági jellemzőt.

- Minél nagyobb α értéke, annál kisebb a megbízhatóság. Kisebb megbízhatóság keskenyebb konfidencia intervallumot eredményez.
- A mintaelemszám növelése csökkenti a hibahatárt, azaz rövidebb intervallumot eredményez.

A mintavételezés után a számegyenes egy konkrét intervallumát kapjuk. Itt már nem beszélhetünk arról, hogy ez $1-\alpha$ valószínűséggel lefedi a keresett sokasági jellemzőt.

Példa

Egy teherautórakománnyi félliteres üdítőitalból 10 palackot véletlenszerűen kiválasztva és lemérve azok űrtartalmát az alábbi, milliliterben kifejezett értékeket kaptuk:

$$499,\ 525,\ 498,\ 503,\ 501,\ 497,\ 493,\ 496,\ 500,\ 495.$$

Ismert, hogy a palackokba töltött üdítőital mennyisége normális eloszlású 3 ml szórással. Adjon 95%-os megbízhatóságú konfidencia intervallumot az átlagos töltőtömegre.

$$n = 10,\ \sigma = 3,\ \alpha = 0.05,\ z_{1-\alpha/2} = z_{0.975} = 1.96,\ \overline{y} = 500.7.$$

A keresett konfidencia intervallum:

$$\overline{y} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 500.7 \pm 1.96\frac{3}{\sqrt{10}} = 500.7 \pm 1.8594.$$

Határok: $\hat{\theta}_a = 498.8406,\ \hat{\theta}_f = 502.5594$; hibahatár: $\Delta = 1.8594$.

Jelölés: $\mathrm{Int}_{0.95}(\mu) = (498.8406, 502.5594)$.

Példa

$$n = 10,\ \sigma = 3,\ \Delta = 1.8594,\ \mathrm{Int}_{0.95}(\mu) = (498.8406, 502.5594).$$

Mekkora mintaelemszám szükséges egy kétszer ilyen pontos 95%-os megbízhatóságú konfidencia intervallum meghatározásához?

Új hibahatár: $\widetilde{\Delta} = \Delta/2$.

Új mintanagyság: $\widetilde{n}$.

$$\widetilde{\Delta} = z_{0.975}\frac{\sigma}{\sqrt{\widetilde{n}}} = 1.96\frac{3}{\sqrt{\widetilde{n}}} = 1.96\frac{3}{2\sqrt{10}} = z_{0.975}\frac{\sigma}{2\sqrt{n}} = \Delta/2.$$

Megoldás: $\widetilde{n} = 4n = 40$, azaz négyszeres mintanagyság szükséges.

Példa

$$n = 10,\ \sigma = 3,\ \Delta = 1.8594,\ \text{Int}_{0.95}(\mu) = (498.8406, 502.5594).$$

Mekkora mintaelemszám szükséges egy fele ilyen pontos 90%-os megbízhatóságú konfidencia intervallum meghatározásához?

Új hibahatár: $\widetilde{\Delta} = 2\Delta$.

Új mintanagyság: $\widetilde{n}$.

Új megbízhatóság: $\widetilde{\alpha} = 0.1,\ z_{1-\widetilde{\alpha}/2} = z_{0.95} = 1.6449.$

$$\widetilde{\Delta} = z_{0.95}\frac{\sigma}{\sqrt{\widetilde{n}}} = 1.6449\frac{3}{\sqrt{\widetilde{n}}} = 2 \cdot 1.96\frac{3}{\sqrt{10}} = 2 \cdot z_{0.975}\frac{\sigma}{2\sqrt{n}} = 2\Delta.$$

$$\frac{1.6449}{\sqrt{\widetilde{n}}} = \frac{2 \cdot 1.96}{\sqrt{10}} \iff \sqrt{\widetilde{n}} = \frac{1.6449 \cdot \sqrt{10}}{2 \cdot 1.96} = 1.3269 \iff \widetilde{n} = 1.7608.$$

Megoldás: legalább 2 elemű minta szükséges.

Egyoldali konfidencia intervallumok

$y_1, y_2, \ldots, y_n$: minta. θ: becsülni kívánt sokasági jellemző.

Adott α esetén keresünk olyan $\hat{\theta}_{a(\alpha)}$ és $\hat{\theta}_{f(\alpha)}$ becslőfüggvényeket, melyekre

$$\mathsf{P}\big(\theta < \hat{\theta}_{f(\alpha)}\big) = 1 - \alpha \qquad \text{(baloldali konfidencia intervallum);}$$
$$\mathsf{P}\big(\hat{\theta}_{a(\alpha)} < \theta\big) = 1 - \alpha \qquad \text{(jobboldali konfidencia intervallum).}$$

Normális eloszlás, ismert szórás

$y_1, y_2, \ldots, y_n$: FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból, σ ismert.

$\theta = \mu$: becsülni kívánt sokasági jellemző.

Baloldali konfidencia intervallum:

$$\mathsf{P}\left(\mu < \overline{y} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \quad \text{azaz} \quad \left(-\infty, \overline{y} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right).$$

Jobboldali konfidencia intervallum:

$$\mathsf{P}\left(\overline{y} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} < \mu\right) = 1 - \alpha, \quad \text{azaz} \quad \left(\overline{y} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right).$$

Normális eloszlás, ismeretlen szórás

$y_1, y_2, \ldots, y_n$: FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból, σ nem ismert.

$\theta = \mu$: becsülni kívánt sokasági jellemző.

σ^2 becslése: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$.

$$T = \frac{\overline{y} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

ezért adott α esetén

$$\mathrm{P}\left(\overline{y} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \overline{y} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) = 1 - \alpha.$$

$t_{1-\alpha/2}(n-1)$: az $n-1$ szabadsági fokú t-eloszlás $p = 1 - \alpha/2$ rendű kvantilise. Táblázatból meghatározható.

Az $(1-\alpha) \cdot 100\%$-os megbízhatóságú konfidencia intervallum

alsó határa: $\overline{y} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$; felső határa: $\overline{y} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$.

Példa

Egy gabonaraktárban 60 kg-os kiszerelésben búzát csomagolnak. A havi minőségellenőrzés során lemértek tíz darab véletlenül kiválasztott zsákot. Eredményül a következőket kapták:

$$60.2,\ 63.4,\ 58.8,\ 63.6,\ 64.7,\ 62.5,\ 66.0,\ 59.1,\ 65.1,\ 62.0.$$

Feltételezve, hogy a töltőtömeg normális eloszlású, adjon 95%-os megbízhatóságú konfidencia intervallumot a zsákokba lévő búzamennyiség várható értékére.

$$n = 10,\ \alpha = 0.05,\ t_{1-\alpha/2}(n-1) = t_{0.975}(9) = 2.2622,\ \overline{y} = 62.54,\ s^2 = 6.2938,\ s = 2.5087.$$

A keresett konfidencia intervallum:

$$\overline{y} \pm t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} = 62.54 \pm 2.2622\frac{2.5087}{\sqrt{10}} = 62.54 \pm 1.7946,$$

azaz

$$\mathrm{Int}_{0.95}(\mu) = (60.7454, 64.3346)\ \text{kg}.$$

Sokasági variancia becslése

$y_1, y_2, \ldots, y_n$: FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból.

$\theta = \sigma^2$: becsülni kívánt sokasági jellemző.

σ^2 becslése: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$.

$$\mathcal{X}^2 = \frac{(n-1)s^2}{\sigma^2} \sim \mathcal{X}^2_{n-1},$$

ezért adott α esetén

$$\mathrm{P}\left(\frac{(n-1)s^2}{\mathcal{X}^2_{1-\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\mathcal{X}^2_{\alpha/2}(n-1)} \right) = 1 - \alpha.$$

$\mathcal{X}^2_{\alpha/2}(n-1)$ és $\mathcal{X}^2_{1-\alpha/2}(n-1)$: az $n-1$ szabadsági fokú $\mathcal{X}^2$-eloszlás $\alpha/2$, illetve $1-\alpha/2$ rendű kvantilisei. Táblázatból meghatározhatóak.

A σ^2-re adott $(1-\alpha)\cdot 100\%$-os megbízhatóságú konfidencia intervallum

alsó határa: $\frac{(n-1)s^2}{\mathcal{X}^2_{1-\alpha/2}(n-1)}$; felső határa: $\frac{(n-1)s^2}{\mathcal{X}^2_{\alpha/2}(n-1)}$.

Példa

Tekintsük az előző példa normális eloszlásúnak feltételezett mintáját:

$$60.2,\ 63.4,\ 58.8,\ 63.6,\ 64.7,\ 62.5,\ 66.0,\ 59.1,\ 65.1,\ 62.0.$$

Adjunk 95%-os megbízhatóságú konfidencia intervallumot a szórásra.

$$n = 10,\ \alpha = 0.05,\ \mathcal{X}^2_{\alpha/2}(n-1) = \mathcal{X}^2_{0.025}(9) = 2.7004,$$
$$\mathcal{X}^2_{1-\alpha/2}(n-1) = \mathcal{X}^2_{0.975}(9) = 19.0228,\ s^2 = 6.2938.$$

A σ^2-re vett 95%-os megbízhatóságú konfidencia intervallum

alsó határa $$\frac{(n-1)s^2}{\mathcal{X}^2_{1-\alpha/2}(n-1)} = \frac{9 \cdot 6.2938}{19.0228} = 2.9777,$$

felső határa $$\frac{(n-1)s^2}{\mathcal{X}^2_{\alpha/2}(n-1)} = \frac{9 \cdot 6.2938}{2.7004} = 20.9762.$$

A σ-ra vett 95%-os megbízhatóságú konfidencia intervallum:

$$\mathrm{Int}_{0.95}(\sigma) = \left(\sqrt{2.9777}, \sqrt{20.9762}\right) = (1.7256, 4.5800).$$

Sokasági arány becslése

Legyen adott egy esemény, aminek a valószínűsége P. Például feldobunk egy érmét és fejet dobunk, egy véletlenszerűen kiválasztott hallgató lány, stb.

n elemű minta: n darab független kísérlet az adott eseményre.

$\hat{P} = p = \frac{k}{n}$: P torzítatlan és konzisztens becslőfüggvénye, k a vizsgált esemény bekövetkezéseinek száma.

k eloszlása binomiális n és p paraméterekkel:

$$\mathrm{E}(p) = P, \qquad \mathrm{Var}(p) = \sigma_p^2 = \frac{P(1-P)}{n}, \quad \text{becslése} \quad s_p^2 = \frac{p(1-p)}{n}.$$

Ha a mintaelemszám nagy, azaz $\min\{np, n(1-p)\} \geq 10$, akkor

$$Z = \frac{p - P}{s_p} \quad \text{eloszlása közel } \mathcal{N}(0,1).$$

Adott α esetén

$$\mathrm{P}\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < P < p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha.$$

Mintanagyság

Az P arányra vett $(1-\alpha)\cdot 100\%$-os megbízhatóságú konfidencia intervallum

alsó határa: $p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$; felső határa: $p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$.

Adott Δ pontossághoz szükséges mintanagyság:

$$n = \frac{z_{1-\alpha/2}^2 \cdot P \cdot (1-P)}{\Delta^2}.$$

P nem ismert, de $P\cdot(1-P) \le 1/4$, azaz egy felső becslés a mintanagyságra:

$$n = \frac{z_{1-\alpha/2}^2}{4\cdot\Delta^2}.$$

Példa

A Medián közvéleménykutató 2017 október végi 1200 fős reprezentatív mintán alapuló felmérése alapján az összes megkérdezett 30%-a vallotta magát bizonytalan szavazónak, vagy nem válaszolt a kérdezőbiztosnak (HVG, 44. szám, 2017. november 2). Adjon 95%-os megbízhatóságú konfidenciaintervallumot a bizonytalan/nem válaszoló szavazók az arányára az összes választópolgár között.

$$n = 1200,\ p = 0.3,\ \alpha = 0.05,\ z_{1-\alpha/2} = z_{0.975} = 1.96,\ s_p^2 = 0.000175.$$

A keresett konfidencia intervallum:

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.3 \pm 1.96\sqrt{\frac{0.3 \cdot 0.7}{1200}} = 0.3 \pm 0.0259,$$

azaz

$$\mathrm{Int}_{0.95}(\mathrm{P}) = (0.2741, 0.3259).$$

Példa

Az előző példában 90%-os megbízhatóság mellett hány elemű minta kell az 1%-os pontosság eléréséhez?

$$\alpha = 0.1,\ z_{1-\alpha/2} = z_{0.95} = 1.6449,\ \Delta = 0.01.$$

A szükséges mintaelemszám egy felső becslése:

$$n = \frac{z_{1-\alpha/2}^2}{4 \cdot \Delta^2} = \left(\frac{1.6449}{2 \cdot 0.01}\right)^2 = 6763.9.$$

6764 elemű minta már biztosan teljesíti a kívánt feltételeket.

Értékösszeg becslése

$Y_1, Y_2, \ldots, Y_N$: N elemű sokaság.

$\mu = \overline{Y}$: sokasági várható érték.

$Y' = N\overline{Y} = N\mu$: értékösszeg.

Ha a μ várható értékre adott egy az $y_1, y_2, \ldots, y_n$ minta alapján készült konfidencia intervallum, akkor az értékösszegre vett konfidencia intervallum ennek az intervallumnak az N-szerese.

Példa. Tíz véletlenszerűen kiválasztott zsák búza töltőtömege alapján az átlagos töltőtömegre vett 95%-os megbízhatóságú konfidencia intervallum:

$$\mathrm{Int}_{0.95}(\mu) = (60.7454, 64.3346) \text{ kg}.$$

Tegyük fel, hogy a raktárban egy hét alatt 10000 zsákot töltenek meg. Ekkor az összes töltőmennyiségre vett 95%-os megbízhatóságú konfidencia intervallum:

$$\mathrm{Int}_{0.95}(Y') = 10000(60.7454, 64.3346) \text{ kg} = (607.454, 643.346) \text{ tonna}.$$

Várható értékek különbségének becslése. Ismert szórás

$x_1, x_2, \ldots, x_{n_X}$: FAE minta $\mathcal{N}(\mu_X, \sigma_X^2)$ eloszlásból.

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta $\mathcal{N}(\mu_Y, \sigma_Y^2)$ eloszlásból.

A két minta **független** és a σ_X^2 és σ_Y^2 varianciák **ismertek**.

$\delta = \mu_Y - \mu_X$: becsülendő sokasági jellemző.

$\overline{d} = \overline{y} - \overline{x}$: becslőfüggvény, normális eloszlású,

$$\mathrm{E}(\overline{d}) = \delta \ \text{(torzítatlan)}, \quad \mathrm{Var}(\overline{d}) = \mathrm{Var}(\overline{y}) + \mathrm{Var}(\overline{x}) = \sigma_{\overline{d}}^2 = \frac{\sigma_Y^2}{n_Y} + \frac{\sigma_X^2}{n_X}.$$

Adott α esetén

$$\mathrm{P}\left(\overline{d} - z_{1-\alpha/2}\sigma_{\overline{d}} < \delta < \overline{d} + z_{1-\alpha/2}\sigma_{\overline{d}}\right) = 1 - \alpha.$$

Az $(1-\alpha)\cdot 100\%$-os megbízhatóságú konfidencia intervallum

alsó határa: $\overline{y} - \overline{x} - z_{1-\alpha/2}\sqrt{\frac{\sigma_Y^2}{n_Y} + \frac{\sigma_X^2}{n_X}}$; felső határa: $\overline{y} - \overline{x} + z_{1-\alpha/2}\sqrt{\frac{\sigma_Y^2}{n_Y} + \frac{\sigma_X^2}{n_X}}$.

Várható értékek különbségének becslése. Ismeretlen szórás

$x_1, x_2, \ldots, x_{n_X}$: FAE minta $\mathcal{N}(\mu_X, \sigma_X^2)$ eloszlásból.

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta $\mathcal{N}(\mu_Y, \sigma_Y^2)$ eloszlásból.

A két minta **független** és a σ_X^2 és σ_Y^2 varianciák **nem ismertek**, de **egyenlőek**, azaz $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

$\delta = \mu_Y - \mu_X$: becsülendő sokasági jellemző.

σ_X^2 becslése: $s_X^2 = \frac{1}{n_X - 1} \sum_{i=1}^{n_X} (x_i - \overline{x})^2$; σ_Y^2 becslése: $s_Y^2 = \frac{1}{n_Y - 1} \sum_{j=1}^{n_Y} (y_i - \overline{y})^2$.

$\sigma^2 = \sigma_X^2 = \sigma_Y^2$ kombinált becslése:

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

$\overline{d} = \overline{y} - \overline{x}$ becslőfüggvény *standard hibája* és *becsült standard hibája*

$$\sigma_{\overline{d}} = \sqrt{\frac{\sigma_Y^2}{n_Y} + \frac{\sigma_X^2}{n_X}}, \qquad s_{\overline{d}} = s_c \sqrt{\frac{1}{n_Y} + \frac{1}{n_X}}.$$

Várható értékek különbségének becslése. Ismeretlen szórás

$\overline{d} = \overline{y} - \overline{x}$ becslőfüggvény eloszlása $\mathcal{N}(\delta, \sigma_{\overline{d}}^2)$, valamint

$$T = \frac{\overline{d} - \delta}{s_{\overline{d}}} \sim t_\nu, \qquad \text{ahol} \quad \nu = n_X + n_Y - 2.$$

Adott α esetén

$$\mathrm{P}\left(\overline{d} - t_{1-\alpha/2}(\nu) s_{\overline{d}} < \delta < \overline{d} + t_{1-\alpha/2}(\nu) s_{\overline{d}}\right) = 1 - \alpha.$$

$t_{1-\alpha/2}(\nu)$: a $\nu = n_X + n_Y - 2$ szabadsági fokú t-eloszlás $p = 1 - \alpha/2$ rendű kvantilise. Táblázatból meghatározható.

Az $(1-\alpha)\cdot 100\%$-os megbízhatóságú konfidencia intervallum

alsó határa: $\overline{y} - \overline{x} - t_{1-\alpha/2}(n_X + n_Y - 2) s_c \sqrt{\frac{1}{n_Y} + \frac{1}{n_X}}$;

felső határa: $\overline{y} - \overline{x} + t_{1-\alpha/2}(n_x + n_Y - 2) s_c \sqrt{\frac{1}{n_Y} + \frac{1}{n_X}}$.

Példa

Kétfajta instant kávé oldódási idejét tesztelték, melyekből minden alkalommal azonos mennyiséget tettek 1 dl forrásban lévő vízbe. A kísérletek eredményeit az alábbi táblázat tartalmazza:

Kávé	Oldódási idő (másodperc)							
Mokka Makka (Y)	8.2	5.0	6.8	6.7	5.8	7.3	6.4	7.8
Koffe In (X)	5.1	4.3	3.4	3.7	6.1	4.7		

Az oldódási időket normálisnak, a szórásokat pedig egyenlőnek tételezve fel adjon 95%-os megbízhatóságú konfidencia intervallumot az átlagos oldódási idők különbségére.

$$n_Y = 8, \quad n_X = 6, \quad \nu = 12, \quad \alpha = 0.05, \quad t_{1-\alpha/2}(\nu) = 2.1788, \overline{y} = 6.75, \quad \overline{x} = 4.55, \quad s_Y^2 = 1.0857, \quad s_X^2 = 0.9670.$$

A variancia kombinált becslése:

$$s_c^2 = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X + n_Y - 2} = \frac{7 \cdot 1.0857 + 5 \cdot 0.9670}{12} = 1.0362, \quad s_c = 1.0180.$$

A keresett konfidencia intervallum:

$$\overline{y} - \overline{x} \pm t_{1-\alpha/2}(n_X + n_Y - 2)s_c\sqrt{\frac{1}{n_Y} + \frac{1}{n_X}} = 6.75 - 4.55 \pm 2.1788 \cdot 1.0180\sqrt{\frac{1}{8} + \frac{1}{6}}$$

$$\mathrm{Int}_{0.95}(\delta) = \mathrm{Int}_{0.95}(\mu_Y - \mu_X) = (1.0022, 3.3978).$$

Páros minta

$\begin{pmatrix} y_1 \\ x_1 \end{pmatrix}, \begin{pmatrix} y_2 \\ x_2 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ x_n \end{pmatrix}$: FAE minta $\begin{pmatrix} Y \\ X \end{pmatrix}$ vektorra. A két ismérv **nem feltétlenül független**!

$d_i = y_i - x_i$ normális eloszlású $(i = 1, 2, \ldots, n)$, $\mathrm{E}(Y) = \mu_Y$, $\mathrm{E}(X) = \mu_X$.

$\delta = \mu_Y - \mu_X$: becsülendő sokasági jellemző.

$d_1, d_2, \ldots, d_n$: új minta $\mathcal{N}(\delta, \sigma_d^2)$ eloszlásból. Ezzel a mintával készítünk konfidencia intervallumot δ-ra.

σ_d^2 becslése: $s_d^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(d_i - \overline{d}\right)^2$.

$$T = \frac{\overline{d} - \delta}{s_d/\sqrt{n}} \sim t_{n-1}.$$

Adott α esetén

$$\mathrm{P}\left(\overline{d} - t_{1-\alpha/2}(n-1)s_d/\sqrt{n} < \delta < \overline{d} + t_{1-\alpha/2}(n-1)s_d/\sqrt{n}\right) = 1 - \alpha.$$

Az $(1-\alpha)\cdot 100\%$-os megbízhatóságú konfidencia intervallum

alsó határa: $\overline{d} - t_{1-\alpha/2}(n-1)s_d/\sqrt{n}$; felső határa: $\overline{d} + t_{1-\alpha/2}(n-1)s_d/\sqrt{n}$.

Példa

A Mindent Tudás Egyeteme másodéves gazdaságinformatikus hallgatói két zárthelyi dolgozatot írtak statisztikából. Az alábbi táblázat tíz véletlenszerűen kiválasztott hallgató eredményeit tartalmazza:

Hallgató	A	B	C	D	E	F	G	H	I	J
I. dolgozat (Y)	57	63	67	82	45	65	53	32	51	27
II. dolgozat (X)	53	62	63	80	46	64	44	28	50	29

A dolgozateredmények eltérését normális eloszlásúnak tételezve fel adjon 98%-os megbízhatóságú konfidencia intervallumot az átlagos pontszámok különbségére.

Új minta ($d_i = y_i - x_i$): $4,\ 1,\ 4,\ 2,\ -1,\ 1,\ 9,\ 4,\ 1,\ -2$.

$$n = 10,\ \alpha = 0.02,\ t_{1-\alpha/2}(n-1) = t_{0.99}(9) = 2.8214,\ \overline{d} = 2.3,\ s_d^2 = 9.7889,\ s_d = 3.1287.$$

A keresett konfidencia intervallum:

$$\overline{d} \pm t_{1-\alpha/2}(n-1)\frac{s_d}{\sqrt{n}} = 2.3 \pm 2.8214\frac{3.1287}{\sqrt{10}} = 2.3 \pm 2.7915,$$

azaz

$$\mathrm{Int}_{0.98}(\delta) = \mathrm{Int}_{0.98}(\mu_Y - \mu_X) = (-0.4915, 5.0915).$$

Hányadosbecslés páros minta esetén

$\binom{y_1}{x_1}, \binom{y_2}{x_2}, \ldots, \binom{y_n}{x_n}$: FAE minta $\binom{Y}{X}$ vektorra.

$$\mathrm{E}(Y) = \mu_Y, \quad \mathrm{E}(X) = \mu_X, \quad \mathrm{Var}(Y) = \sigma_Y^2, \quad \mathrm{Var}(X) = \sigma_X^2.$$

$H = \mu_Y/\mu_X$: becsülendő sokasági jellemző.

$h = \overline{y}/\overline{x}$: becslőfüggvény, eloszlása nem ismert, de nagy minta esetén közel normális.

$$\mathrm{E}(h) \approx H + \frac{H}{n}\left(V_X^2 - r(X,Y)V_X V_Y\right).$$

$V_X = \sigma_X/\mu_X$, $V_Y = \sigma_Y/\mu_Y$: X és Y *relatív szórása*.

$r(X,Y)$: X és Y *korrelációja*.

$$\mathrm{Var}(h) \approx \frac{H^2}{n}\left(V_X^2 + V_Y^2 - 2r(X,Y)V_X V_Y\right).$$

h a H hányados *torzított*, de *aszimptotikusan torzítatlan* és *konzisztens* becslése.

Külső információs becslés

Külső információ: például ismert μ_X.

μ_Y becslése: $\hat{\mu}_Y = \mu_X \cdot \frac{\overline{y}}{\overline{x}} = \mu_X \cdot h$.

$$\mathrm{Var}(\hat{\mu}_Y) = \mu_X^2 \,\mathrm{Var}(h) \approx \frac{1}{n}\big(\sigma_Y^2 + H^2\sigma_X^2 - 2Hr(X,Y)\sigma_X\sigma_Y\big).$$

Hagyományos becslés varianciája: $\mathrm{Var}(\overline{y}) = \frac{\sigma_Y^2}{n}$.

$$\mathrm{Var}(\hat{\mu}_Y) \le \mathrm{Var}(\overline{y}), \qquad \text{ha} \qquad \frac{\sigma_Y^2}{n} \ge \frac{1}{n}\big(\sigma_Y^2 + H^2\sigma_X^2 - 2Hr(X,Y)\sigma_X\sigma_Y\big),$$

azaz

$$r(X,Y) \ge \frac{1}{2} \cdot \frac{V_X}{V_Y}.$$

Ha az X és Y korrelációja elég nagy, a külső információt használó becslés hatékonyabb, mint a mintaátlag.

Becslés EV mintából

EV minta: N elemű sokaságból választunk n elemet visszatevés nélkül.

Különbségek a FAE mintához képest.

- A minta fontos jellemzője az alapsokaság N nagysága.
- Az egyes mintaelemek nem függetlenek.
- A mintajellemzők eloszlásának meghatározása jóval bonyolultabb, mint FAE minta esetén. Nagy minták esetén a mintaátlag, az értékösszeg és a sokasági arány közelítőleg normális eloszlású.

$$\mathrm{E}(\overline{y}_{EV}) = \mu, \qquad \mathrm{Var}(\overline{y}_{EV}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right).$$

EV mintából a várható érték becslése pontosabb, mint ugyanakkora FAE mintából.

Sokasági átlag becslése, nagy minta

$y_1, y_2, \ldots, y_n$: EV minta egy N elemű sokaságból, $n \geq 30$.

$\mu = \overline{Y}$: becsülni kívánt sokasági jellemző.

Ha σ értéke ismert, adott α esetén

$$\mathrm{P}\left(\overline{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}} < \overline{Y} < \overline{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right) = 1-\alpha.$$

Az $(1-\alpha)\cdot 100\%$-os megbízhatóságú konfidencia intervallum

alsó határa: $\overline{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$; felső határa: $\overline{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$.

Értékösszeg becslése: FAE mintához hasonlóan

Ha σ nem ismert, akkor az s empirikus szórással helyettesíthetjük.

Adott Δ pontossághoz szükséges mintaelemszám:

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{z^2_{1-\alpha/2}\sigma^2/N + \Delta^2}.$$

Sokasági arány becslése EV mintából

Legyen P valamilyen tulajdonságú elemek aránya az N elemű sokaságban.

Minta: n darab visszatevés nélkül kiválasztott sokasági elem.

$\hat{P} = p = \frac{k}{n}$: P torzítatlan és konzisztens becslőfüggvénye, k a vizsgált tulajdonságú elemek száma a mintában.

$$\mathrm{E}(p) = P, \qquad \mathrm{Var}(p) = \frac{P(1-P)}{n}\left(1 - \frac{n}{N}\right) \approx s_p^2 = \frac{p(1-p)}{n}\left(1 - \frac{n}{N}\right).$$

Ha a mintaelemszám nagy, akkor

$$Z = \frac{p - P}{s_p} \qquad \text{eloszlása közel } \mathcal{N}(0,1).$$

Adott α esetén az $(1-\alpha)\cdot 100\%$-os megbízhatóságú konfidencia intervallum

alsó határa: $p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{1-\frac{n}{N}}$; felső határa: $p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{1-\frac{n}{N}}$.

A hipotézisvizsgálat általános kérdései

A sokaságokra vonatkozó különféle feltevéseket hipotéziseknek, az azok helyességének mintavételi eredményekre alapozott vizsgálatát pedig hipotézisvizsgálatnak nevezzük. A hipotézisek a vizsgált sokaság(ok) eloszlására vagy az adott eloszláso(ok) egy vagy több paraméterére vonatkozhatnak. A különféle hipotézisek vizsgálatára szolgáló eljárásokat statisztikai próbáknak nevezzük.

Eredményül nem azt kapjuk, hogy egy hipotézis igaz-e, vagy sem, hanem hogy az adott körülmények között elfogadjuk-e.

Példa

1. Tudva, hogy egy üdítőitalt gyártó gépsorról lekerülő palackokban a folyadékmennyiség normális eloszlású 3 ml szórással, egy 10 elemű minta alapján vizsgáljuk meg, az átlagos töltőmennyiség 500 ml-e.
2. Egy 15 elemű minta alapján vizsgáljuk meg, a gyártósorról lekerülő palackokban a folyadékmennyiség normális eloszlású-e.
3. 500 ember haj-, illetve szemszínét megvizsgálva döntsünk, a hajszín független-e a szemszíntől.

Hipotézisek megfogalmazása

Két egymásnak ellentmondó feltevést – hipotézist – fogalmazunk meg.

Az egyik: nullhipotézis, jelölése H_0, erről hozunk döntést.

A másik: alternatív hipotézis, vagy ellenhipotézis, jelölése H_1.

Példa

1. Jelölje μ a palackokba töltött folyadékmennyiség várható értékét.

$$H_0: \ \mu = 500 \text{ ml}; \qquad H_1: \ \mu \neq 500 \text{ ml}.$$

2. H_0: a folyadékmennyiség normális eloszlású;
   H_1: a folyadékmennyiség **nem** normális eloszlású.
3. H_0: a hajszín és a szemszín függetlenek egymástól;
   H_1: a hajszín és a szemszín **nem** függetlenek egymástól.

Egyszerű nullhipotézis: fennállása esetén a sokaság eloszlása egyértelműen meghatározott.

A próbafüggvény meghatározása

Próbafüggvény: az $y_1, y_2, \ldots, y_n$ minta egy olyan $T(y_1, y_2, \ldots, y_n)$ függvénye, melynek eloszlása H_0 teljesülése esetén ismert.

Példa. Az üdítőitalt gyártó gépsorról lekerülő palackokban a folyadékmennyiség normális eloszlású $\sigma = 3$ ml szórással és ha H_0 igaz, akkor 500 ml várható értékkel. $n = 10$ esetén $\overline{y} \sim \mathcal{N}(500, 3^2/10)$, azaz

$$z = \frac{\overline{y} - 500}{\sigma/\sqrt{n}} = \frac{\overline{y} - 500}{3/\sqrt{10}} \sim \mathcal{N}(0, 1).$$

A hipotézisek vizsgálatára *próbafüggvényeket* használunk, amik a becslőfüggvényekhez hasonlóan valószínűségi változók. A próbafüggvényt úgy kell megválasztani, hogy

- a sokaságra tett bizonyos kikötések teljesülése (például normális eloszlás, ismert szórás),
- a mintavétel adott módja és a minta adott nagysága (például FAE minta vagy $n \geq 100$),
- az ellenőrzendő H_0 helyességének feltételezése

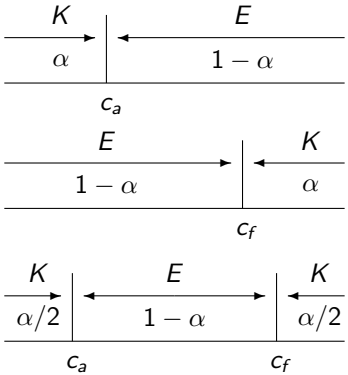
mellett annak eloszlása pontosan ismert legyen. Ehhez H_0-nak egyszerű hipotézisnek kell lennie.

Szignifikanciaszint, kritikus tartomány

A próbafüggvény értékkészletét két diszjunkt részre bontjuk, egy elfogadási (E) és egy kritikus (K) tartományra.

A határok megadása: ha H_0 teljesül, akkor a próbafüggvény egy előre megadott nagy $1-\alpha$ valószínűséggel az *elfogadási tartományba* esik, ahol α kicsi (például 0.1, 0.05, 0.01).

Szignifikanciaszint: $\alpha \cdot 100\%$ (például 10%, 5%, 1%)



Bal oldali kritikus tartomány.

c_a: a próbafüggvény eloszlásának $p = \alpha$ rendű kvantilise.

Jobb oldali kritikus tartomány.

c_f: a próbafüggvény eloszlásának $p = 1-\alpha$ rendű kvantilise.

Kétoldali kritikus tartomány.

c_a, c_f: a próbafüggvény eloszlásának $p = \alpha/2$, illetve $1-\alpha/2$ rendű kvantilise.

Egy- és kétoldali kritikus tartományok

A valóságnak a nullhipotézisben rögzített állapottól való meghatározott irányú eltérései egyoldali alternatív hipotézisként írhatók fel. Ha az egyoldali alternatív hipotézis fennállása esetén a próbafüggvény kisebb értéket vesz fel, mint H_0 fennállásakor, bal oldali, ellenkező esetben pedig jobb oldali alternatív hipotézisről beszélünk. A bal oldali alternatív hipotéziseket ezután H_1^b-vel, a jobb oldaliakat pedig H_1^j-vel jelöljük.

A valóságnak a nullhipotézisben rögzített állapottól való tetszőleges irányú eltérései kétoldali alternatív hipotézisként fogalmazhatók meg. A kétoldali alternatív hipotézis fennállása esetén a próbafüggvény értéke akár kisebb, akár nagyobb lehet, mint H_0 fennállásakor. A kétoldali alternatív hipotéziseket ezután H_1-gyel jelöljük.

Példa

$$H_0 : \mu = 500 \text{ ml};$$
$$H_1^b : \mu < 500 \text{ ml}; \quad H_1^j : \mu > 500 \text{ ml}; \quad H_1 : \mu \neq 500 \text{ ml}.$$

c_a, c_f: kritikus értékek, hozzátartoznak a kritikus tartományhoz.

Példa

Tudva, hogy egy üdítőitalt gyártó gépsorról lekerülő palackokban a folyadékmennyiség normális eloszlású 3 ml szórással, egy 10 elemű minta alapján 5%-os szinten vizsgáljuk meg, az átlagos töltőmennyiség 500 ml-e. Írjuk fel az egyoldali és a kétoldali kritikus tartományokat.

Szignifikanciaszint: 5%, azaz $\alpha = 0.05$.

Próbafüggvény: $z = \dfrac{\overline{y} - 500}{3/\sqrt{10}}$. Ha H_0 teljesül, eloszlása standard normális.

Kvantilisek:

Rend: p	$\alpha/2 = 0.025$	$\alpha = 0.05$	$1-\alpha = 0.95$	$1-\alpha/2 = 0.975$
Kvantilis: z_p	-1.9600	-1.6449	1.6449	1.9600

$$H_0 : \mu = 500 \text{ ml}; \qquad H_1 : \mu \neq 500 \text{ ml}.$$

Kritikus tartomány: $z \leq c_a = z_{\alpha/2} = z_{0.025} = -1.96$, vagy
$z \geq c_f = z_{1-\alpha/2} = z_{0.975} = 1.96$, azaz $|z| \geq 1.96$.

$$H_0 : \mu = 500 \text{ ml}; \qquad H_1^b : \mu < 500 \text{ ml}.$$

Kritikus tartomány: $z \leq c_a = z_\alpha = z_{0.05} = -1.6449$.

$$H_0 : \mu = 500 \text{ ml}; \qquad H_1^j : \mu > 500 \text{ ml}.$$

Kritikus tartomány: $z \geq c_f = z_{1-\alpha} = z_{0.95} = 1.6449$.

Mintavétel és döntés

A minta adataiból kiszámítjuk a próbafüggvény értékét. Ha az a kritikus tartományba esik, a megadott szinten elvetjük H_0-t, ellenkező esetben elfogadjuk.

Példa. Egy teherautórakománnyi félliteres üdítőitalból 10 palackot véletlenszerűen kiválasztva és lemérve azok űrtartalmát az alábbi, milliliterben kifejezett értékeket kaptuk:

$$499,\ 525,\ 498,\ 503,\ 501,\ 497,\ 493,\ 496,\ 500,\ 495.$$

Ismert, hogy a palackokba töltött üdítőital mennyisége normális eloszlású 3 ml szórással. 5%-os döntési szintet használva vizsgálja meg a gyártó azon állítását, hogy a palackokba átlagosan fél liter üdítőitalt töltöttek.

$$H_0 : \mu = 500 \text{ ml}; \qquad H_1 : \mu \neq 500 \text{ ml} \quad \text{(kétoldali ellenhipotézis)}.$$

$n = 10,\ \alpha = 0.05,\ \sigma = 3,\ \overline{y} = 500.7.$ Kritikus tartomány: $|z| \geq 1.96$.

Próbafüggvény értéke: $z = \dfrac{500.7 - 500}{3/\sqrt{10}} = 0.7379 < 1.96.$

5%-os szinten **elfogadjuk** H_0-t.

Hibák

Elsőfajú hiba: a H_0 hipotézist elvetjük, pedig igaz. Valószínűsége megegyezik az α szignifikanciaszinttel.

Másodfajú hiba: a H_0 hipotézist elfogadjuk, pedig nem igaz. Ennek a β valószínűségét csak akkor számszerűsíthetjük, ha pontosan tudjuk, a H_0 helyett a valóságban milyen egyszerű alternatíva áll fenn.

H_0	igaz	nem igaz
elvetjük	elsőfajú hiba (α)	helyes döntés ($1-\beta$)
elfogadjuk	helyes döntés ($1-\alpha$)	másodfajú hiba (β)

Adott mintanagyság és egyszerű alternatíva mellett az elsőfajú és a másodfajú hiba elkövetési valószínűsége egymással ellentétes irányba mozog.

p-érték

A p-érték az a legkisebb szignifikanciaszint, amin H_0 már éppen elvethető H_1-el szemben. A p-érték a T próbafüggvénynek a hipotézisvizsgálathoz használt mintából nyert értéke alapján határozható meg.

Egyoldali alternatív hipotézis esetén a p-érték úgy határozható meg, hogy a próbafüggvény mintából nyert értékét H_1 irányának megfelelően alsó vagy felső kritikus értéknek tekintjük, majd megállapítjuk, vagy megbecsüljük a hozzá tartozó szignifikanciaszintet.

Kétoldali alternatív hipotézis esetén a próbafüggvény mintából nyert értékét előjelétől – egyes esetekben nagyságától – függően alsó vagy felső kritikus értéknek tekintjük, majd a hozzátartozó szignifikanciaszint kétszeresét vesszük.

Adott α szint esetén:

$p \leq \alpha$: elvetjük H_0-t; $p > \alpha$: elfogadjuk H_0-t.

Példa

Minta:

$$499,\ 525,\ 498,\ 503,\ 501,\ 497,\ 493,\ 496,\ 500,\ 495.$$

Hipotézisek:

$$H_0 : \mu = 500 \text{ ml}; \qquad H_1 : \mu \neq 500 \text{ ml} \quad \text{(kétoldali ellenhipotézis)}.$$

$n = 10,\ \sigma = 3,\ \overline{y} = 500.7.$

Próbafüggvény értéke: $z = \dfrac{500.7 - 500}{3/\sqrt{10}} = 0.7379 > 0.$

Felső kritikus értékként kezeljük:

$$\mathrm{P}(Z \geq 0.7379) = 1 - \mathrm{P}(Z \leq 0.7379) = 1 - \Phi(0.7379) = 1 - 0.7697 = 0.2303.$$

p-érték: $2 \cdot 0.2303 = 0.4606.$

$\alpha \geq 0.4606$: elvetjük H_0-t; $\alpha < 0.4606$: elfogadjuk H_0-t.

Összetett nullhipotézisek

Példa

$$H_0 : \mu \geq 500 \text{ ml}; \qquad H_1 : \mu < 500 \text{ ml}.$$

Az összetett nullhipotézis helyett az egyoldali alternatív hipotézisnek legkevésbé ellentmondó egyszerű hipotézist választjuk. A példában: $\mu = 500$ ml.

A H_1^b vagy H_1^j egyoldali alternatív hipotézisnek legkevésbé ellentmondó egyszerű hipotézist technikai nullhipotézisnek nevezzük és H_0^T-vel jelöljük. A példában:

$$H_0^T : \mu = 500 \text{ ml}; \qquad H_1 : \mu < 500 \text{ ml}.$$

Ha H_0^T elvethető valamely egyoldali alternatív hipotézissel szemben, akkor vele együtt elvethető az adott egyoldali alternatív hipotézisnek H_0^T-nél jobban ellentmondó minden egyszerű hipotézis is. Ha H_0^T nem vethető el valamely egyoldali alternatív hipotézissel szemben, akkor csak annyi állítható, hogy a vizsgált alternatív hipotézissel szemben legalább egy egyszerű hipotézis nem utasítható vissza.