STATISZTIKA1 FELADATOK

Bolla Marianna és Mala József

BME Matematika Intézet, Sztochasztika Tanszék

2021. március 30.

1. feladatsor

Alapstatisztikák és többdimenziós adatrendszerek

1. A Kandó Kálmán Műszaki Főiskolára a jelentkező 300 fiúból 200, míg az 50 lányból 40 került be. Ugyanakkor a Színművészeti Főiskolára a jelentkező 100 fiúból 5, míg a 200 lányból 20 került be. Láthatjuk, hogy mindkét főiskolára a lányoknak nagyobb %-a jutott be, mint a fiúknak, ugyanakkor az egyesített mintában éppen az ellenkezőjét tapasztaljuk: a két főiskolára összességében a lányoknál rosszabb a felvételi arány, mint a fiúknál. Hogyan lehetséges ez?

Megoldás: Az itt leírt paradoxont, amely abban nyilvánul meg, hogy különböző részcsoportok keverékénél a részekkel ellentétes következtetésre jutunk, *Simpson-paradoxon*nak is nevezik (*Simpson, E. H., The interpretation of interaction in contingency tables, J. Royal. Statist. Soc. Ser. B 13 (1951), 238-241*). Azon alapul, hogy különböző súlyokkal szerepelnek az egyes intézmények a fiúknál ill. a lányoknál. Mondhatnánk, hogy a lányok jobban teljesítettek, a fiúk viszont jobban választottak. Megfigyeléseink valójában 3 kategorikus (esetünkben bináris) változóra vonatkoznak: NEM (fiú/lány), INTÉZMÉNY (Kandó Kálmán Műszaki Főiskola/Színművészeti Főiskola), FELVÉTEL (igen/nem). A $2 \times 2 \times 2$ térbeli kocka alkú *kontingenciatábla* adatai (8 cellával) az alábbi síkbeli táblázatokba sűríthetők össze (a 8 értéket cella-gyakoriságnak nevezzük).

A NEM versus FELVÉTEL 2×2-es marginális tábla a következő:

	felvettek	felvételizők
fiú	205	400
lány	60	250

A fiúk felvételi aránya 51.3%, a lányoké 24%. Ha azonban az intézményenkénti rétegeket nézzük, akkor azokban a két táblázat a fentivel ellentétes arányokat mutat:

intézmény	nem	felvettek	felvételizők
Kandó	fiú	200	300
	lány	40	50
Színm.	fiú	5	100
	lány	20	200

A Kandóra a lányok felvételi aránya 80%, míg a fiúké 66.6%, a Színművészetire a lányok felvételi aránya 10%, míg a fiúké 5%. Ugyanakkor, összességében a fiúk több mint fele bekerült (ebbe a két) felsőoktatási intézménybe, míg a lányoknak még negyede sem.

Még pontosabban, itt 3 bináris változó (NEM, FELVÉTEL, INTÉZMÉNY) együttes eloszlásáról van szó, ami megadható $2 \times 2 \times 2$-es kontingenciatáblával szeletekben. A tábla NEM versus FElVÉTEL szeletei az INTÉZMÉNY kétféle értékére a következők.

Kandó:

	felvett	nem felvett
fiú	200	100
lány	40	10

Színművészeti:

	felvett	nem felvett
fiú	5	95
lány	20	180

A paradoxon abból adódik, hogy a NEM versus FELVÉTEL marginális tábla (az INTÉZMÉNYekre összegezve) nem a fenti arányokat mutatja.

Összesen:

	felvett	nem felvett
fiú	205	195
lány	60	190

A 2. feladatban megadunk egy algebrai feltételt, amely mellett a paradoxon egyáltalán előfordulhat alkalmas cella-gyakoriságokkal. A 3. és 6. feladatban inkább azt vizsgáljuk, milyen feltételek mellett nem fordul elő a paradoxon. Ezek feltételes függetlenségekkel fogalmazhatók meg és a grafikus modellek irányába mutatnak.

2. Általában, ha $p > q$ a felvett lányok aránya ill. felvett fiúk aránya a Kandóra, továbbá $r > s$ fejezi ki ugyanezt a Színművészetire, akkor igazolja, hogy a Simpson paradoxon pontosan akkor állhat elő (valamely jelentkezési arányokkal), ha $q > r$ vagy $s > p$!

3. **Floridai gyilkosságok:** A következő két táblázat floridai gyilkosságok miatti vádemelések számát tartalmazza a vádlott bőrszínének ill. annak a függvényében, hogy halálbüntetést kiszabtak-e vagy sem. Az első táblázat - amelyben az áldozat bőrszíne nem szerepel - azt mutatja, hogy fehéreket nagyobb arányban ítéltek halálra, mint feketéket (3.2% ill. 2.3%). Ha azonban az áldozat bőrszíne szerint csoportosítjuk az adatokat, a kép egészen más: mind a fekete, mind a fehér áldozatok esetében a feketék nagyobb arányban kaptak halálbüntetést. Mindkét táblázat ugyanannak a helyzetnek bizonyos szempontból pontos leírása; azonban a példa mutatja, hogy egy gondatlan vizsgálat milyen hamar megalapozatlan eredményre vezethet.

vádlott	halálraítéltek	elítéltek
fekete	59	2606
fehér	72	2257

A táblázatból látható, hogy a fekete elítéltek között a halálraítéltek aránya 2.3%, míg ugyanez a szám a fehérek között 3.2%.

Az áldozat bőrszíne szerinti adatok a következők voltak:

áldozat	vádlott	halálbüntetés	egyéb
fekete	fekete	11	2320
	fehér	0	111
fehér	fekete	48	286
	fehér	72	2146

Ekkor a fekete áldozatok között a fekete-fehér elkövetők halálraítélési arányai $0.5-0\%$ és ugyanezek az arányok a fehér áldozatok között $16.8-3.3\%$.

Az adatok Floridában 1973-78 közt gyűjtött adatokon alapulnak (Range, P. R. (1979), Will he be the first? The New York Times Magazine, 11th March, 72-82.)

Itt is 3 binális változó (GYILKOS (bőrszíne), ÍTÉLET, ÁLDOZAT (bőrszíne)) együttes eloszlásáról van szó, ami megadható $2 \times 2 \times 2$-es kontingenciatáblával. A tábla GYILKOS versus ÍTÉLET szeletei az ÁLDOZAT kétféle értékére a következők.

Az ÁLDOZAT fekete:

	halálos ítélet	nem halálos ítélet
fekete gyilkos	11	2320
fehér gyilkos	0	111

Az ÁLDOZAT fehér:

	halálos ítélet	nem halálos ítélet
fekete gyilkos	48	286
fehér gyilkos	72	2146

A paradoxon abból adódik, hogy a GYILKOS versus ÍTÉLET marginális tábla (az ÁLDOZATokra összegezve) nem a fenti arányokat mutatja.

Összesen:

	halálos ítélet	nem halálos ítélet
fekete gyilkos	59	2606
fehér gyilkos	72	2257

4. Ha a változók kettőnél több értéket is felvehetnek, akkor is hasonló helyzet adódhat. A következő példa egy képzeletbeli referendumról szól; a szavazók számát, illetve a szavazás eredményét tartalmazza az alábbi két táblázat három városban (A, B, C) és három korosztály (fiatalok (F), középkorúak (K) és nyugdíjasok (N)) szerint.

szavazók száma:

	A	B	C
F	15000	15000	70000
K	25000	30000	45000
N	45000	35000	20000

igennel szavazók száma:

	A	B	C
F	12000	9000	24500
K	17000	16500	13500
N	27000	17500	5000

igennel szavazók százaléka városok szerint:

	A	B	C
F	80	60	35
K	68	55	30
N	60	50	25

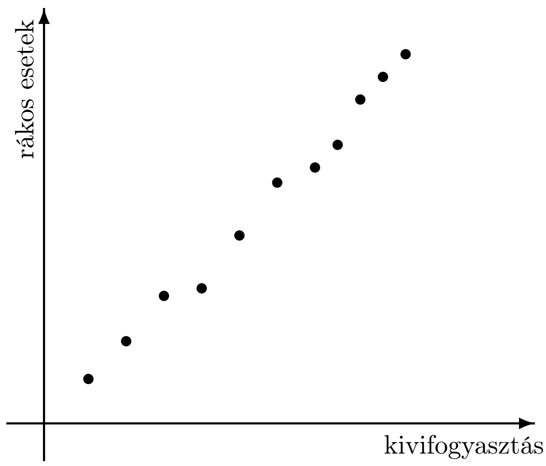
igennel szavazók százaléka összesen:

	szav.	igen szav.	százalék
F	100000	45500	45.5
K	100000	47000	47
N	100000	49500	49.5

Látható, hogy mindhárom városban a fiatalok között legnagyobb a referendum támogatottsága és a nyugdíjasok között legkisebb; azonban a korosztályok szerinti összesített adatok mást mutatnak: a sorrend éppen a fordítottja a fentinek; összességében a nyugdíjasok között a legnagyobb a támogatottság (49.5%), utána következnek a középkorúak (47%) és a fiatalok között a legkisebb (45.5); a fiatalok sokat veszítenek C-ben, a nyugdíjasok viszont sokat nyernek A-ban. (Szintén vegyük észre, hogy a népes C-ben viszonylag kevés a nyugdíjas. A városokra összegezve így a KOR versus TÁMOGATJA kereszttáblában a N korosztályban lesz a legkisebb a nem-támogatók, és így legnagyobb a támogatók aránya.) Egyébként itt az eltérések nem lényegesek (a szignifikancia fogalmával a hipotézisvizsgálatoknál fogunk megismerkedni).

5. Folytonos eloszlású változók esetén is előfordulhat, hogy más következtetésre jutunk, ha csak kettőt ragadunk ki a sok változó közül. Például télen azt tapasztaljuk, minél magasabb a gázszámlánk, annál hidegebb van a lakásban, ha nem vesszük figyelembe a külső hőmérsékletet és a gáz összetételének romlását csúcsfogyasztás esetén. A problémát regresszióanalízis címszó alatt tárgyaljuk, és a parciális korrelációval van kapcsolatban. Ez két változó kapcsolatának „szorosságát” méri, megtisztítva a többi változó hatásától. Előfordulhat ugyanis hogy ha csupán két változót ragadunk ki több közül, azok negatívan korrelálnak, mert nem vettük figyelembe a zavaró hatásokat.

A következő ábra az Egyesült Államokban megfigyelt rákos esetek számát mutatja a kivifogyasztás függvényében. Mivel 1970 és 1980 között mindkét mennyiség növekedett, ezek évente megfigyelt értékei *korreláltak.* Jóllehet ez matematikai bizonyosság, mégsem állíthatjuk, hogy a rákos esetek számának növekedését az *okozta,* hogy az emberek több kivit ettek. A ténylegesen talált (és statisztikailag bizonyított) korrelációt csak akkor szabad ok-okozati kapcsolatnak tekinteni, ha erre *elméleti indok* van.



Hasonlóan, egy országban az egy lakosra jutó laptopok száma pozitívan korrelál az életszínvonallal. De senki sem gondolja, hogy pl. Rhodesiában az életszínvonal megnő, ha teherhajókkal laptopot küldenek oda.

Ilyen példákat az élet legkülönbözőbb területein találhatunk. Például határozottan pozitív korreláció van a Duna vízállása és a BME-n tartózkodó hallgatók száma között, hiszen késő tavasszal és késő ősszel magas a Duna vízállása és éppen ezek az időszakok előzik meg a vizsgaidőszakot. Itt és kivis példában is az idő a közvetítő, míg a laptopos példában a pénz.

6. Folytonos eloszlásokra a következő tétel alkalmazható, ld. *Cox, D. R. and Wermuth, N., Avoiding effect reversal after marginalization, J. Royal. Stat. Assoc. B. 65 (2003), 937–941.* A tétel kiterjeszthető diszkrét eloszlásokra is.

Tétel: Legyen X, Y, W nornális eloszlású, X prediktor, Y célváltozó (válasz), W pedig háttérváltozó. Cox and Wermuth megvizsgálták, hogy a közönséges regressziós együttható β_{YX} mikor egyezik meg a $\beta_{YX|W}$ parciális regressziós együtthatóval, azaz mikor nincsen W-nek befolyása X és Y egymás közti viszonyára.

A későbbi regressziós tananyag jelöléseit használjuk. Tegyük fel, hogy változóinkat már standardizáltuk (0 várható érték, 1 szórás). Akkor $\beta_{YX} = r_{YX}$ (Pearson korreláció), $\beta_{YX|W} = r_{YX|W}$ pedig a parciális korrelációs együttható X és Y közt. A Cochran formula szerint

$$r_{YX} = r_{YX|W} + \beta_{YW|X} \cdot r_{XW}. \tag{1}$$

Így $r_{YX} = r_{YX|W}$ akkor és csak akkor teljesül, ha vagy $r_{XW} = 0$ vagy $\beta_{YW|X} = 0$. Tehát W-nek pontosan akkor nincs befolyása X és Y egymás közti viszonyára, ha vagy X and W (marginálisan) függetlenek, vagy Y és W feltételesen függetlenek X adott értéke mellett. Jelöléssel:

$$X \perp\!\!\!\perp W \quad \text{vagy} \quad Y \perp\!\!\!\perp W | X. \tag{2}$$

Most példát látunk arra, hogy nem következik be a paradoxon.

Gyerekverő dán nők: A példát *Lauritzen, S. L., Graphical Models, Oxfor Univ. Press (1995)* könyvéből vettük, a Gallup Intézet adatai alapján.

237 dán nőnél vizsgálták a következőket: használ-e fizikai fenyítést gyerekeinél (U=Usage), politikai pártállás (A=Affiliation)), gyerekkorában tapasztalt-e fizikai fenyítést (E=Experience). Az alábbi táblázat ezt mutatja (angolul):

1. táblázat. Frequencies (gyakoriságok).

Childhood experience and use of physical punishment(PP)		Political Affiliation: Left	Soc.Dem.	Right
Has experienced PP	Uses PP	12	27	58
	Does not use PP	7	28	30
Has not experienced PP	Uses PP	9	5	9
	Does not use PP	19	15	18

U (usage of physical punishment y/n), E (childhood experience of physical punishment y/n), A (party affiliation, L/S/R).

2. táblázat. Tables describing marginal associations between childhood experience (E) of physical punishment, use (U) of physical punishment and political affiliation (A)

	U	
E	yes	no
yes	97	65
no	23	52

	A		
E	l	s	r
yes	19	55	88
no	28	20	27

	A		
U	l	s	r
yes	21	32	67
no	26	43	48

Ellenőrizhető, hogy $U \perp\!\!\!\perp A | E$, ami részletesebben azt jelenti, hogy

$$\mathbb{P}(U = u, A = a \,|\, E = e) = \mathbb{P}(U = u \,|\, E = e) \cdot \mathbb{P}(A = a \,|\, E = e)$$

vagy ami ezzel ekvivalens,

$$\mathbb{P}(U = u \,|\, A = a,\, E = e) = \mathbb{P}(U = u \,|\, E = e)$$

az összes lehetséges a, e, u értékhármasra. Ez statisztiksan értendő (a később tanulandó statisztikai próbák, pl. χ^2 ismeretében). Pl. a marginálisok szorzataként előálló 2×3-as U versus A kereszt-tábla mind az $E = y$, mind az $E = n$ szeletekben jól közelítette az eredetit, ami az E adott értéke melletti függetlenségét sugallja. Ezt a χ^2 próbák is alátámasztják:

3. táblázat. Results from the χ^2 tests for independence

Null Hypothesis (H_0)	$U \perp\!\!\!\perp A$	$U \perp\!\!\!\perp A \| E = \mathrm{y}$	$U \perp\!\!\!\perp A \| E = \mathrm{n}$
Degree of Freedom	2	2	2
χ^2 Value	5.25	4.08	0.42
Significance	0.072	0.130	0.810
Conclusion (7.5%)	Reject	Do not reject	Do not reject

Következtetés: Alkalmazzuk a Tételt az eddigi példákban!

- A dán nők esetében:

$$Y = U, \quad X = E, \quad W = A.$$

  A fentiek alapján $Y \perp\!\!\!\perp W | X$ teljesül (2)-ban, így A-nak nincs befolyása U and E kapcsolatára, nem áll elő a paradoxon.
  Erre az $A \leftarrow E \rightarrow U$ jelölés is használatos, ami azt is jelenti, hogy A és U feltételesen független E adott értéke mellett, de marginálisan nem függetlenek. Ezen az alapon log-lineáris modell konstruálható:

$$\mathbb{P}(U = u, A = a, E = e) = \mathbb{P}(E = e) \cdot \mathbb{P}(U = u \,|\, E = e) \cdot \mathbb{P}(A = a \,|\, E = e)$$

  vagy logaritmálással

$$\ln p_{uae} = \ln f_{ue} + \ln f_{ae}$$

  bizonyos ún. másodrendű interakciókkal.

- Az 1. felvételi példában:

$$Y = \text{FELVÉTEL}, \quad X = \text{NEM}, \quad W = \text{INTÉZMÉNY}.$$

Itt X és W nem függetlenek és Y, W sem feltételesen függetlenek X-re. Ezért (2) alapján W-nek hatása van Y és X kapcsolatára, ami miatt a paradoxon bekövetkezik.

- A 3. floridai gyilkossági példában:

$$Y = \text{ÍTÉLET}, \quad X = \text{GYILKOS}, \quad W = \text{ÁLDOZAT}.$$

Itt (2)-beli egyik feltétel sem teljesül. Ezért W-nek hatása van Y és X kapcsolatára, ami miatt a paradoxon bekövetkezik.

- A következő példa *Wainer, H., Book review of Social Indicators III: selected data on social conditions and trends in the United States, J. Amer. Stat. Assoc. 78 (1980), 492–496* és *Wermuth, N., Graphical Markov models, unifying results and their interpretation, In: (Balakrishnan, N. et al. eds) Wiley StatsRef: Statistics Reference Online (2015), ArXiv: 1505.02456* cikkeiből való.

  Ellentétben a dán nők példájával, itt egy harmadik változó inkább feltételes függőséget indukál két független között. American Banks (1980): Y jövedelem, X iskolázottság (években), W nem. Azt tapasztalták, hogy Y növekedett X-el mindkét nem esetében, így Y and X nem feltételesen függetlenek W-re. Ugyanakkor $X \perp\!\!\!\perp W$ marginálisan független (férfiak és nők nemüktől függetlenül részesülhettek az oktatásban). De Y és W már nem feltételesen függetlenek X-re. Így van paradoxon. Hasonlóan, X és W nem feltételesen függetlenek Y-ra. Valóban, adott fizetésnél a nőknek magasabb volt az iskolázottsága, mint a férfiaknak.

  Ezt a feltételes függőség generálást úgy is jelölhetjük, hogy
  $X \to Y \leftarrow W$ (itt Y időben később következik).

7. A feladat a Matematikai statisztika példatárból (ELTE Eötvös Kiadó, 1997) való. Egy gyáregységben az alkalmazottak havi fizetése (ezer Ft-ban): 25, 26, 18, 20, 51, 45, 80, 20, 30, 25, 250, 46, 38, 142, 40, 21, 25, 26, 52, 90, 44, 21, 24, 40, 51. A rendezett minta alapján ábrázoljuk az adatokat sűrűség-hisztogrammal, és készítsük el az empirikus eloszlásfüggvényt is. Gyanakodhatunk-e Pareto eloszlásra? Számoljuk ki a minta középértékeit (átlag, medián)! Melyik középértéket mikor használnánk? Keressük meg a kvartilis értékeket is!

8. Egy zeneműboltban kiválasztottunk hét különböző kiadású CD-t, melyeken Beethoven 9. szimfóniája szerepel. A mű időtartamai az egyes CD-ken (percben mérve): 66.9, 66.2, 71.0, 68.6, 65.4, 68.4, 71.9.

Keressük meg

(a) a tapasztalati mediánt;

(b) a tapasztalati átlagot;

(c) a tapasztalati szórást.

9. A kalciumhiány komoly probléma az idős nőknél. Hogy a hiányt megállapítsa, egy kutató idős nőkben megmérte az alkarban található kalciummennyiséget, majd ugyanezt elvégezte egy év múltán. A veszteségeket tartalmazza a következő táblázat:

8	7	13	3	6
4	8	6	3	4
0	1	11	7	1
8	6	12	13	10
9	11	3	2	9
7	1	16	3	2
10	15	2	5	8
17	8	2	5	5

a) Számolja ki $\bar{x}$ - et és s - et, ahol (és a továbbiakban is) s a korrigált tapasztalati szórást jelöli!

b) Számolja ki azon esetek arányát, melyek az $\bar{x} \pm s, \bar{x} \pm 2s$ ill. $\bar{x} \pm 3s$ intervallumokba esnek;

c) Hasonlítsa össze a kapott számokat az $(\bar{x}, s)$ - paraméterű normális eloszlás megfelelő értékeivel!

10. A következő táblázat az eddigi amerikai elnökök életkorát tartalmazza a beiktatáskor:

1.	Washington	57	16.	Lincoln	52	31.	Hoover	54
2.	J.Adams	61	17.	A.Johnson	56	32.	F.D. Roosevelt	51
3.	Jefferson	57	18.	Grant	46	33.	Truman	60
4.	Madison	57	19.	Hayes	54	34.	Eisenhower	62
5.	Monroe	58	20.	Garfield	49	35.	Kennedy	43
6.	J.Q.Adams	57	21.	Arthur	50	36.	L.Johnson	55
7.	Jackson	61	22.	Cleveland	47	37.	Nixon	56
8.	Van Buren	54	23.	B.Harrison	55	38.	Ford	61
9.	W.H.Harrison	68	24.	Cleveland	55	39.	Carter	52
10.	Tyler	51	25.	McKinley	54	40.	Reagan	69
11.	Polk	49	26.	T.Roosevelt	42	41.	G.Bush	64
12.	Taylor	64	27.	Taft	51	42.	Clinton	46
13.	Fillmore	50	28.	Wilson	56	43.	G.W.Bush	54
14.	Pierce	48	29.	Harding	55	44.	Obama	47
15.	Buchanan	65	30.	Coolidge	51	45.	D.Trump	70
						46.	J.Biden	78

Adjuk meg a mediánt és a kvartiliseket, valamint a minta maximális és minimális értéket, a mintaterjedelmet és a mintaátlagot! Készítsünk sűrűséghisztogramot 6 osztóponttal! A 6 osztópont felhasználásával végezzen normalitás vizsgálatot!

11. a) Számolja ki $\bar{x}$-t and s-t a $6, 8, 4, 9, 8$ adatokra!

b) Tekintsük az $106, 108, 104, 109, 108$, adatokat, melyeket az előzőekből 100 - zal való eltolással kaptunk. Mennyi lesz most az $\bar{x}$ és s értéke?

c) Az a)-pontbeli adatokat megszorozzuk -3-mal: $-18, -24, -12, -27, -24$.
Most mi lesz $\bar{x}$ és s értéke?

2. feladatsor

Elégséges statisztikák és maximum likelihood becslések

12. Legyen $X_1, \ldots, X_n$ fae. minta az $f_\theta(x) = 2\theta x(1-x^2)^{\theta-1}$ (ha $0 < x < 1$, különben 0) sfv.-el definiált eloszlásból. Keressünk elégséges statisztikát és ML-becslést az ismeretlen $\theta > 0$ paraméterre!

Megoldás: a likelihood fv.:

$$L_\theta(\mathbf{x}) = L_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i) = [2^n\theta^n(\prod_{i=1}^n (1-x_i)^2)^{\theta-1})] \cdot [\prod_{i=1}^n x_i],$$

ahol az első kapcsos zárójelben álló fv. $g_\theta(\mathbf{x})$, míg a másodikban álló nem függ a paramétertől, ez $h(\mathbf{x})$. (Megjegyezzük, hogy $h(\mathbf{x})$-be az $I(0 \le x_1^* \le \cdots \le x_n^* \le 1)$ indikátorfv-t is bevehettük volna, ez sem függ a paramétertől.) Így a $\prod_{i=1}^n (1-X_i^2)$ statisztika elégséges lesz, de $\sum_{i=1}^n \log(1-X_i^2)$ is az lenne.

A log-likelihood fv. θ szerint deriválható, a derivált gyökhelye lehet az ML-becslés:

$$\frac{\partial}{\partial\theta} \ln L_\theta(\mathbf{x}) = \frac{\partial}{\partial\theta} \ln g_\theta(\mathbf{x}) = \frac{n}{\theta} + \sum_{i=1}^n \log(1-X_i^2) = 0,$$

ahonnan

$$\hat\theta = -\frac{n}{\sum_{i=1}^n \log(1-X_i^2)}.$$

Könnyen látható, hogy $\hat\theta$ 1 val.séggel pozitív, és ez az egyetlen lokális, sőt globális maximum; továbbá $\hat\theta$ egy elégséges statisztika fv-e.

13. Legyen $X_1, \ldots, X_n$ fae. minta az $f_\theta(x) = \frac{1.5x^2}{\theta^3}$ (ha $-\theta \le x \le \theta$, különben 0) sfv.-el definiált eloszlásból. Keressünk elégséges statisztikát és ML-becslést az ismeretlen $\theta > 0$ pareméterre!

Megoldás: a likelihood fv.:

$$L_\theta(\mathbf{x}) = \prod_{i=1}^n \frac{1.5x_i^2}{\theta^3} I(-\theta \le x_i \le \theta) = \left[\frac{1}{\theta^{3n}} I(\max_i |x_i| \le \theta)\right] \cdot \left[\prod_{i=1}^n 1.5x_i^2\right],$$

ahol az első kapcsos zárójelben álló fv. $g_\theta(\mathbf{x})$, míg a másodikban álló nem függ a paramétertől, ez $h(\mathbf{x})$. Így $\max_i |X_i|$ elégséges.

A likelihood fv. most nem deriválható θ szerint (ez általában így van, ha a sfv. tartója függ a paramétertől), viszont látható, hogy θ csökkenésével monoton nő mindaddig, amíg $\theta \ge \max_i |x_i|$. Így $\hat\theta = \max_i |X_i|$ ML-becslés.

14. Egy halastóból kifogtunk j db halat, megjelöltük és visszadobtuk őket. Ezután visszatevés nélkül kifogtunk n db halat melyek közül x jelölt és $n-x$ jelöletlen. Adjunk maximum likelihood becslést a halastóban található jelöletlen halak k számára.

Megoldás: a likelihood fv. itt a diszkrét k paramétertől függ, megfigyelésünket pedig x-ben sűrítettük össze. Itt ugyan nincsenek független mintaelemek (visszatevés nélkül történik a mintavételezés), mégis maximalizálhatjuk az adott szituáció valószínűségét:

$$L_k(x) = \frac{\binom{j}{x}\binom{k}{n-x}}{\binom{j+k}{n}} \to \max. \quad k - ban.$$

Diszkrét maximalizálásról lévén szó, az

$$\frac{L_k(x)}{L_{k-1}(x)} = \frac{k(j+k-n)}{(k-n+x)(j+k)}$$

hányados növekedési viszonyainak vizsgálatával keressük a maximumot. Látható, hogy ez $\frac{nj}{x} - j$ alsó egész része, ha ez a szám nem egész, különben $\frac{nj}{x} - j$ és $\frac{nj}{x} - j - 1$ is ML-becslések. Előbbi esetben az összes hal számára $\frac{nj}{x} - j + j = \frac{nj}{x}$ alsó egész része lesz az ML-becslés. A józan ész is ezt diktálja: extrapolálunk az összes jelölt és a kifogott mintában található jelöltek számának hányadosával.

15. A következő két kísérletet végezték el annak érdekében, hogy megállapítsák egy adott párt népszerűségét:

(a) addig kérdezték véletlenszerűen kiválasztva az embereket, amíg 10 olyat nem találtak, aki az adott pártra szavazna. Azt tapasztalták, hogy ehhez 1000 embert kellett megkérdezni;

(b) Véletlenszerűen megkérdeztek 1000 embert és azt találták, hogy közülük 10 választaná az adott pártot.

Mutassuk meg, hogy mind a két kísérlet ugyanahhoz a maximum likelihood becsléshez vezet!

16. Tegyük fel, hogy egy almáskertben egymástól függetlenül találhatók fertőzött fák. Tíz egyforma nagy, egyenként három sorból álló ültetvényben rendre $0, 3, 0, 1, 0, 0, 2, 1, 1, 2$ beteg fát találtak. Adjunk maximum likelihood becslést az egy sorban található fertőzött fák számának várható értékére!

17. Egy alkatrész élettartama exponenciális eloszlású ϑ/t várható értékkel, ha t hőmérsékleten működtetjük. Tegyük fel, hogy az n megfigyelést a különböző $t_1, t_2, \ldots, t_n$ hőmérsékleteken végeztük és $x_1, x_2, \ldots, x_n$ élettartamokat figyeltünk meg. Adjunk maximum likelihood becslést ϑ - ra!

Megoldás: Itt a megfigyelések függetlenek, de nem azonos eloszlásúak. Az i. mintaelem eloszlása t_i/θ paraméterű exponenciális. A maximalizálandó likelihood függvény: $\frac{1}{\vartheta^n} \exp\left(-\frac{1}{\vartheta}\sum_{i=1}^n t_i x_i\right)$. A loglikelihood függvény deriváltja: $-\frac{n}{\vartheta} + \frac{1}{\vartheta^2}\sum_{i=1}^n t_i x_i$, ennek zérushelye: $\hat{\vartheta} = \frac{1}{n}\sum_{i=1}^n t_i x_i$. A megoldás nem meglepő, olyan, mint a paraméter reciprokának ML becslése a $t_i x_i$ realizáltakból $(i = 1, \ldots, n)$.

18. T ideig vizsgálták, hogy n páciens meddig élt még a kemoterápiás kezelés után. Az adatok: $x_1, \ldots, *, \ldots$, ahol $*$ - ot írtunk az olyan esetekben, ahol a páciens élt a T időpillanatban. Feltéve, hogy a sorozatban k db x_i található, adjunk maximum likelihood becslést a páciensek túlélési időtartamára, ha az exponenciális eloszlású!

Megoldás: Ha a várható érték ϑ, akkor a likelihood függvény: $\prod_{i=1}^{k} \frac{1}{\vartheta}\exp(-\frac{x_i}{\vartheta})\cdot \prod_{i=1}^{n-k}\exp(-\frac{T}{\vartheta})$. A loglikelihood függvény deriváltja: $-\frac{k}{\vartheta}+\frac{\sum_{i=1}^{k} x_i}{\vartheta^2}+\frac{(n-k)T}{\vartheta^2}$, és ennek zérushelye: $\hat{\vartheta}=\frac{\sum_{i=1}^{k} x_i+(n-k)T}{k}$. Ez biztosan maximumhely, mert a likelihood függvény limesze 0-ban és ∞ - ben is 0.

Cenzorált megfigyelśek: a gyakorlatban sokszor előfordul, hogy a kísérlet lezárultáig nem következik be a válasz, a páciensek esetében pl. csak azt tudjuk, hogy a megfigyelés végéig túléltek, a valódi túlélési időről pedig csak annyit tudunk, hogy minimum annyi, mint a megfigyelési idő. Ezt cenzorált mintának nevezzük és általánosan a Kaplan–Meyer algoritmussal kezeljük.

Általánosabb példa: n villanyégő élettartamát figyeltük meg. Közülük k db. elromlott a vizsgálati idő alatt (élettartamuk $x_1,\dots,x_k$), a többi $n-k$ túlélte a vizsgálatára szánt t_j időt ($j=1,\dots,n-k$) (ezek a cenzoráltak). Feltéve, hogy az égők valódi élettartama exponenciális eloszlású λ paraméterrel, adjunk ML becslást λ-ra.

Megoldás: A cenzorált adatok esetében a $\mathbb{P}(X_j>t_j)=e^{-\lambda t_j}$ exponenciális val.ség lép a likelihood fv-be, ami

$$L_\theta(\mathbf{x})=\prod_{i=1}^{k}\lambda e^{-\lambda x_i}\cdot\prod_{j=1}^{n-k}e^{-\lambda t_j},$$

$$\ln L_\theta(\mathbf{x})=k\ln\lambda-\sum_{i=1}^{k}\lambda x_i-\sum_{j=1}^{n-k}\lambda t_j,$$

melynek deriválásával

$$\hat{\lambda}=\frac{k}{\sum_{i=1}^{k}x_i+\sum_{j=1}^{n-k}t_j},$$

ha $k>0$.

19. Egy város energiafogyasztása (megfelelő mértékegységben) 100 várható értékű és egységnyi szórású normális eloszlású valószínűségi változó. ϑ nap után egy konstans, egységnyi fogyasztású üzem is megkezdte a működését. Adjuk meg a likelihood függvényt $x_1,\dots,x_n$ megfigyelés alapján. Adjunk maximum likelihood becslést az alábbi adatsor esetén: 99.2, 101.4, 99.7, 100.2, 102.4, 100.1, 101.6, 99.8, 102.4, 100.5.

20. Egy laborban a mérést általában az ismert (σ) szórású műszeren végzik. n ilyen mérés elvégzése után (a független, azonos $\mathcal{N}(\mu,\sigma^2)$ eloszlású adatok: $x_1,\dots,x_n$) elromlott a készülék és csak a régi $k\sigma$ (szintén ismert) szórású műszert lehetett használni. Ezzel a műszerrel az $y_1,\dots,y_n$ adatokhoz jutottunk (μ változatlan). Adjunk maximum likelihood becslést μ - re!

21. Legyen $X_1, \dots, X_n$ n megfigyelés az $f(x) = 0.5e^{-|x-\vartheta|}$ sűrűségfüggvényű eloszlásból. Adjunk maximum likelihood becslést ϑ-ra!

Megoldás: A maximalizálandó likelihood függvény $-\sum_{i=1}^n |\vartheta - x_i|$, melynek maximumhelye $\hat{\vartheta} = m_n$, ahol m_n jelöli az $x_1, \dots, x_n$ számsokaság mediánját (Ez $n = 2k$ esetén nem egyértelmű: ekkor az $x_k^{(n)}, x_{k+1}^{(n)}$ intervallum tetszőleges pontja megoldása a szélsőérték feladatnak).

Az eredmény például abból az okoskodásból adódik, hogy függvényünk folytonos és szakaszonként deriválható; a derivált konstans minden $(x_m^{(n)}, x_{m+1}^{(n)})$ intervallumon és a konstansok fogyó sorozatot alkotnak. A 0 értéket éppen a fentiekben megadott helyeken veszi fel (illetve a páratlan esetben "ugorja át" a függvény).

Másik megoldás, hogy ha ϑ egyik oldalán több x_i van, mint a másikon, akkor az $\sum_{i=1}^n |\vartheta - x_i|$ távolságösszeg nőni fog, ha a medián(ok) felé mozdulunk el.

22. A Smarties cukorka k színben készül. Tegyük fel, hogy nem ismerjük k-t. Kiveszünk 3 szemet a zacskóból, a színeik rendre piros, zöld, piros.

(a) Adjunk maximum likelihood becslést k-ra!

Kiveszünk egy negyediket is, ez sárga.

(b) Adjunk maximum likelihood becslést k-ra!

23. Legyen az X változó sűrűségfüggvénye $x \in (0, 1)$ esetén $\theta x^{\theta-1}$, egyébként 0, $\theta > 0$ ismeretlen. Az $X_1, \dots, X_n$ minta esetén adjunk elégséges statisztikát és ML-becslést θ-ra!

24. Az ún. α modell (l. pl. Csiszár,V., Hussami,P., Komlós,J., Móri,T.F., Rejtő,L. and Tusnády,G. (2011), When the degree sequence is a sufficient statistic, Acta Math. Hung. 134, 45-53) a következő. Adott egy véletlen gráf n csúccsal, melynek szomszédsági mátrixa $\mathbf{A} = (A_{ij})$. $\mathbf{A}$ diagonálisa zéró, a diagonalis feletti A_{ij}-k pedig függetlenek, és A_{ij} Bernoulli eloszlású $p_{ij} = \mathbb{P}(A_{ij} = 1)$ paraméterrel, különben $\mathbf{A}$ szimmetrikus. A modell szerint a $\frac{p_{ij}}{1-p_{ij}}$ ún. esélyhányadosokra

$$\frac{p_{ij}}{1-p_{ij}} = \alpha_i \alpha_j \quad (1 \le i < j \le n)$$

teljesül, ahol $\alpha_1, \dots, \alpha_n$ pozitiv valós paraméterek. Könnyen látható, hogy

$$p_{ij} = \frac{\alpha_i \alpha_j}{1 + \alpha_i \alpha_j} \quad \text{és} \quad 1 - p_{ij} = \frac{1}{1 + \alpha_i \alpha_j}.$$

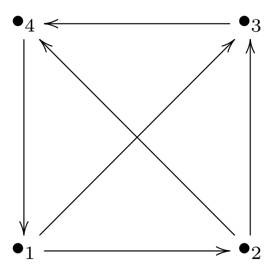
Úgy tűnhet, hogy egyelemű mintánk van, azonban itt az élek a független mintaelemek (n rögzített). Jelölje $\underline{D} = (D_1, \dots, D_n)$ a fokszámsorozatot, ahol $D_i = \sum_{j=1}^n A_{ij}$ $(i = 1, \dots n)$. Belátjuk, hogy ez elégséges az $\underline{\alpha} = (\alpha_1, \dots, \alpha_n)$ paramétervektorra. $\mathbf{A}$ szimmetriája és a $0^0 = 1$ konvenció miatt

$$
\begin{aligned}
L_{\underline{\alpha}}(\mathbf{A}) &= \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p_{ij}^{A_{ij}} (1-p_{ij})^{1-A_{ij}} = \left\{ \prod_{i=1}^{n} \prod_{j=1}^{n} p_{ij}^{A_{ij}} (1-p_{ij})^{1-A_{ij}} \right\}^{1/2} \\
&= \left\{ \prod_{i=1}^{n} \prod_{j=1}^{n} \left(\frac{p_{ij}}{1-p_{ij}} \right)^{A_{ij}} \prod_{i=1}^{n} \prod_{j=1}^{n} (1-p_{ij}) \right\}^{1/2} \\
&= \left\{ \prod_{i=1}^{n} \alpha_i^{\sum_{j=1}^{n} A_{ij}} \prod_{j=1}^{n} \alpha_j^{\sum_{i=1}^{n} A_{ij}} \prod_{i \neq j} (1-p_{ij}) \right\}^{1/2} \\
&= \left\{ \prod_{i \neq j} \frac{1}{1+\alpha_i \alpha_j} \right\}^{1/2} \left\{ \prod_{i=1}^{n} \alpha_i^{D_i} \prod_{j=1}^{n} \alpha_j^{D_j} \right\}^{1/2} \\
&= \left\{ \prod_{i<j} \frac{1}{1+\alpha_i \alpha_j} \right\} \left\{ \prod_{i=1}^{n} \alpha_i^{D_i} \right\} = C_{\underline{\alpha}} \times \prod_{i=1}^{n} \alpha_i^{D_i}.
\end{aligned}
$$

Itt kihasználtuk, hogy $A_{ij} = A_{ji}$, $p_{ij} = p_{ji}$ $(i<j)$ és $A_{ii} = 0$, $p_{ii} = 0$ $(i = 1, \ldots, n)$. $C_{\underline{\alpha}} = \prod_{i<j} \frac{1}{1+\alpha_i \alpha_j}$ csak az $\underline{\alpha}$ paramétertől függ, és a likelihood fv. az A_{ij} mintaelemektől csak D_i-ken keresztül függ. Így $\underline{D}$ elégséges. A csak mintaelemektől függő tényező itt 1, amiből következik, hogy a mátrixelemek együttes eloszlasa, feltéve a fokszámsorozatot, egyenletes. Azaz adott fokszámsorozat esetén ezzel a modellel generálhatunk véletlen gráfokat.

Megjegyezzük, hogy az Erdős–Rényi véletlen gráf ennek az a specialis esete, melyben az összes α, és így p_{ij} megegyezik. Véletlen, téglalap alakú 0-1 mátrixokra is általánosítható a módszer, l. Rasch modell és Bolla, M., Elbanna, A. (2015), Estimating parameters of a probabilistic heterogeneous block model via the EM algorithm, Journal of Probability and Statistics. Article 657965.

25. Az alábbi ábrán egy bajnokság végeredménye látható, ahol i - ből j - be pontosan akkor fut él, ha i legyőzte j - t. Ha az i játékos játékerejét $w_i > 0$ "méri" oly módon, hogy az i játékos $\frac{w_i}{w_i+w_j}$ valószínűséggel győzi le j - t, továbbá a mérközések eredményei függetlenek egymástól, akkor adjon maximum likelihood becslést az egyes játékosok játékerejére, feltéve, hogy $\sum_{i=1}^{4} w_i = 1$!



Megoldás. $\left(\frac{3}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{8}\right)$. $\square$

3. feladatsor

Becsléselmélet: torzítatlan és aszimptotikusan torzítatlan becslések

26. Ha $X \sim \mathcal{E}xp(\lambda)$, akkor az $X_1, \ldots, X_n$ fae. minta alapján $\overline{X}_n$ torzítatlan becsése $\frac{1}{\lambda}$-nak (hisz az a várható érték). Nem várhatjuk azonban, hogy $\frac{1}{\overline{X}_n}$ torzítatlan becslése legyen λ-nak. Ezt mindjárt be is látjuk az $n > 1$ esetben. (Az $n = 1$ esetben nem létezik $\frac{1}{X_1}$ várható értéke.) Azonban belátjuk, hogy $\frac{1}{\overline{X}_n}$ aszimptotikusan torzítatlan becslése λ-nak, ha $n > 1$.

Megoldás. Kihasználjuk, hogy $\sum_{i=1}^n X_i \sim \Gamma_n(\lambda)$:

$$\mathbb{E}_\lambda(\frac{1}{\overline{X}_n}) = \mathbb{E}_\lambda\left(\frac{n}{\sum_{i=1}^n X_i}\right) = \int_0^\infty \frac{n}{x}\frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}\,dx =$$
$$= \frac{n}{n-1}\lambda \int_0^\infty \frac{\lambda^{n-1} x^{n-2} e^{-\lambda x}}{(n-2)!}\,dx = \frac{n}{n-1}\lambda = \left(1 + \frac{1}{n-1}\right)\lambda \neq \lambda,$$

azonban aszimptotikusan torzítatlan, hiszen

$$\lim_{n\to\infty} \mathbb{E}_\lambda(\frac{1}{\overline{X}_n}) = \lambda.$$

□

27. Bizonyítsa be, hogy az empirikus variancia aszimptotikusan torzítatlan, míg a korrigált empirikus variancia torzítatlan becslése az alapeloszlás valódi varianciájának, amennyiben az létezik.

Megoldás. Jelölje $\mu = \mathbb{E}_\theta(X_1)$ és $\sigma^2 = \mathrm{Var}_\theta(X_1)$ az alapeloszlás várható értékét és varianciáját (ezek függenek θ-tól, de nem jelöljük külön). Ekkor $\mathbb{E}_\theta(X_1^2) = \sigma^2 + \mu^2$. Ezért

$$\mathbb{E}_\theta(\sum_{i=1}^n (X_i - \bar{X})^2 = \mathbb{E}_\theta(\sum_{i=1}^n X_i^2 - n\bar{X}^2) =$$
$$= \sum_{i=1}^n \mathbb{E}_\theta(X_i^2) - n\mathbb{E}_\theta(\bar{X}^2) = n\left(\sigma^2 + \mu^2\right) - n(\frac{\sigma^2}{n} + \mu^2) = (n-1)\sigma^2.$$

Így

$$\mathbb{E}_\theta({S_n}^2) = \frac{n-1}{n}\sigma^2 \to \sigma^2, \quad n \to \infty.$$

A korrekció után

$$\mathbb{E}_\theta({S_n^*}^2) = \sigma^2, \qquad \forall\theta \in \Theta,$$

azaz torzítatlan becslést kapunk. □

4. feladatsor

Becslések további tulajdonságai: hatásosság, konzisztencia

28. Legyen $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$ független azonos eloszlású. Tudjuk, hogy $S = X_1$ torzítatlan becslés és $T = \sum_{i=1}^n X_i$ elégséges statisztika λ-ra. Keressük meg az $U = \mathbb{E}_\lambda(S|T)$ blackwellizáltat!

Megoldás. S feltételes eloszlása a $\sum_{i=1}^n X_i = t$ feltétel mellett:

$$\begin{aligned}\mathbb{P}_\lambda(X_1 = k|\sum_{i=1}^n X_i = t) = &= \frac{\mathbb{P}_\lambda(X_1 = k, \sum_{i=1}^n X_i = t)}{\mathbb{P}_\lambda(\sum_{i=1}^n X_i = t)} = \\ &= \frac{\mathbb{P}_\lambda(X_1 = k) \cdot \mathbb{P}_\lambda(\sum_{i=2}^n X_i = t-k)}{\mathbb{P}_\lambda(\sum_{i=1}^n X_i = t)} = \\ &= \frac{\frac{\lambda^k}{k!}e^{-\lambda} \cdot \frac{((n-1)\lambda)^{t-k}}{(t-k)!}e^{-(n-1)\lambda}}{\frac{(n\lambda)^t}{t!}e^{-n\lambda}} = \\ &= \binom{t}{k}(\frac{1}{n})^k(1-\frac{1}{n})^{t-k},\end{aligned}$$

$(k = 0, \dots, t)$, ahol kihasználtuk, hogy $X_1 + \dots + X_n \sim \mathcal{P}(n\lambda)$. Ebből látható, hogy S feltételes eloszlása a $\sum_{i=1}^n X_i = t$ feltétel mellett $\mathcal{B}_t(1/n)$ (nyilván nem függ λ-tól, mert T elégséges), feltételes várható értéke pedig

$$\mathbb{E}_\lambda(S|\sum_{i=1}^n X_i = t) = t \cdot \frac{1}{n},$$

ahonnan

$$\mathbb{E}_\lambda(S|T) = T \cdot \frac{1}{n} = \bar{X}.$$

□

Megjegyezzük, hogy T egy másik tulajdonsága miatt (teljesség), bármilyen torzítatlan S-t blackwellizálunk vele, ugyanarra az eredményre jutunk. Így $\bar{X}$ hatásos becslés λ-ra.

A Rao–Blackwell–Kolmogorov tétel üzenete: a hatásos becslést sok esetben az elégséges statisztika torzítatlanná tételével kapjuk.

29. Legyen $X \sim \mathcal{N}(\mu, \sigma_0^2)$. Mutassa meg, hogy $\overline{X}$ a μ paraméter hatásos becslése (σ_0^2 adott)!

30. Legyen $(X_1, \dots, X_n)$ indikátorminta a p paraméterű Bernoulli eloszlásból ($n > 2$).

(a) Adjunk X_1 és X_2 függvényeként torzítatlan becslést $p(1-p)$-re;

(b) Adjunk elégségs statisztikát p-re;

(c) Az (a)-ban megadottnál konstruáljunk hatásosabb becslést a Rao-Blackwell-Kolmogorov tétel segítségével.

31. Legyen $X_1, \ldots, X_n$ fae. minta az alábbi diszkrét eloszlásból ($0 < \alpha < 1$ paraméter):

x	1	2
$p_\alpha(x)$	α	$1-\alpha$

(a) Adjunk X_1 függvényeként torzítatlan becslést α-ra;

(b) Adjunk meg elégséges statisztikát;

(c) Az (a)-ban megadottnál konstruáljunk hatásosabb becslést a Rao-Blackwell-Kolmogorov tétel segítségével.

32. Legyen $X_1, \ldots, X_n$ független azonos eloszlású minta egy tetszőleges olyan eloszlásból, amelynek második momentuma létezik. Akkor S_n^2 erősen konzisztens becslése a varianciának.

Megoldás. A nagy számok erős törvényét alkalmazzuk az X háttérváltozó és X^2 várható értékére: $n \to \infty$ esetén

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \to \mathbb{E}_\theta(X^2) - (\mathbb{E}_\theta(X))^2 = \sigma^2(\theta)$$

1 valószínűséggel ($\mathbb{P}_\theta$-m.m.), $\forall\theta \in \Theta$. Itt a 27. feladat megoldásának gondolatmenetét követtük. □

Könnyen látható, hogy S_n^{*2} is erősen konzisztens becslése a varianciának, és persze gyengén konzisztens is. A négyzetes középben való konzisztenciához az alapeloszlás negyedik momentumának létezését is fel kell tenni.

5. feladatsor

Becslési módszerek

33. Adjunk momentum becslést az $\mathcal{U}[a,b]$ eloszlás $a < b$ valós paramétereire egy $X_1, \dots, X_n$ fae. minta alapján!

Megoldás. Mivel két becsülendő paraméterúnk van, az eloszlás első két momentumára lesz szükségúnk:

$$m_1 = \frac{a+b}{2}, \quad m_2 = \sigma^2 + \mu^2 = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4}.$$

Könnyen kiszámolható, hogy az $(a,b) \to (m_1, m_2)$ leképezés Jacobi-determinánsa nem 0, az inverz leképezés pedig:

$$a = m_1 - \sqrt{3(m_2 - m_1^2)}, \qquad b = m_1 + \sqrt{3(m_2 - m_1^2)}.$$

Végül

$$\hat{m}_1 = \bar{X}, \qquad \hat{m}_2 - \hat{m}_1^2 = S_n^2$$

figyelembevételével a paraméterek momentum becslésére

$$\hat{a} = \bar{X} - \sqrt{3}S_n, \qquad \hat{b} = \bar{X} + \sqrt{3}S_n$$

adódik, ami eltér az ML-becsléstől. □

Megjegyezzük, hogy ún. exponenciális eloszláscsaládban (pl. Poisson, Bernoulli, exponenciális, normális eloszlások) a paraméterek momentum becslése ugyanaz, mint az ML-becslése. Jelen eloszlás azonban nem tartozik ide, mivel a sűrűségfüggvény tartója függ a paramétertől. Láttuk, hogy az elégséges statisztika is rendezett mintás.

34. Legyen $X_1, \dots, X_n$ fae. minta az alábbi eloszlásból:

$$f(x) = \frac{2x}{\theta^2}, \qquad 0 \le x \le \theta.$$

(a) Adjon momentum becslést θ-ra!

(a) Adjon ML-becslést θ-ra!

(c) Adjunk meg θ-ra elégséges statisztikát!

35. Konstruáljon 95%-os konfidenciaintervallumot az exponenciális eloszlás λ paraméterére az $X_1, \dots, X_n$ fae. minta alapján, ahol n „nagy" ($n \ge 30$).

Megoldás. A CHT miatt „nagy" n-re $\bar{X}$ közel $\mathcal{N}(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$ eloszlású. Így

$$0.95 = \mathbb{P}_\lambda \left(-1.96 < \frac{\bar{X} - \frac{1}{\lambda}}{\frac{1}{\sqrt{n}\lambda}} < 1.96 \right) =$$

$$= \mathbb{P}_\lambda \left(\frac{1}{\bar{X}} - \frac{1.96}{\sqrt{n}\bar{X}} < \lambda < \frac{1}{\bar{X}} + \frac{1.96}{\sqrt{n}\bar{X}} \right).$$

□

36. a. Legyen $x_1, \ldots, x_{20}$ egy fae. normális eloszlású minta realizációja, ahol a μ várható érték ismeretlen, míg a szórás ismert: 0.05. Az $\bar{x} = 1.5$ mintaátlag alapján konstruáljon 98%-os konfidenciaintervallumot μ-re!

b. Legyen $x_1, \ldots, x_{20}$ egy fae. normális eloszlású minta realizációja, ahol a μ várható érték ismeretlen, a szórásra pedig az $s = 0.05$ becslés adódik (korrigált empirikus szórás) . Az $\bar{x} = 1.5$ mintaátlag alapján konstruáljon 98%-os konfidenciaintervallumot μ-re!

37. Legyen X_1, X_2, X_3 fae. minta λ paraméterű Poisson eloszlásból. Az alábbi statisztikák melyike ad torzítatlan becslést λ-ra? Válaszát csak indoklással fogadjuk el.

(a) $\overline{X}$

(b) $\frac{1}{2}X_1 + \frac{1}{3}X_2$

(c) $\frac{1}{2}X_1 + \frac{1}{3}X_2 + \frac{1}{6}X_3$

(d) $\sum_{i=1}^{3}(X_i - \overline{X})^2$

(e) $\frac{1}{2}\sum_{i=1}^{3}(X_i - \overline{X})^2$

(f) $\frac{1}{3}\sum_{i=1}^{3}(X_i - \overline{X})^2$

6. feladatsor

Hipotézisvizsgálati alapfogalmak, szignifikancia, diszkrét és folytonos alapeloszlás, egy- és kétoldali alternatíva

38. *Folytonos alapeloszlás, kétoldali alternatíva (egymintás, kétoldali z-próba).*

Vásárlói panaszok érkeznek, hogy egy élelmiszerboltban az 1 kg-os feliratú kenyér súlya kevesebb. Szeretnénk korrekt módon kivizsgálni az ügyet. Kiszállunk az üzletbe, megmérünk n véletlenszerűen kiválasztott kenyeret, $X_1, \dots, X_n$ a minta. Legyen $n = 25$, és a realizációban azt találjuk, hogy átlaguk 0.98 kg. Mit tegyünk? Az eltérést okozhatja a véletlen is, hiszen az 1 kg várható értékű, normális eloszlású mintaelemek eltérhetnek a várható értéktől. A következőképpen gondolkozunk: az ártatlanság vélelme alapján tegyük fel, hogy nem csalnak, vagyis a normális eloszlású háttérváltozó várható értéke valóban 1 kg. Szerkesszünk például 95%-os szintű konfidenciaintervallumot a várható értékre a minta alapján! Amennyiben az 1 kg hipotetikus várható érték nincsen benne ebben az intervallumban, akkor két eset lehetséges:

- Mivel az esetek 95%-ában a várható érték benne van ebben az intervallumban, a véletlen folytán lehet, hogy mégiscsak bekövetkezett az az 5% valószínűségű esemény, hogy nincsen benne.
- Nem igaz eredeti elképzelésünk, hogy 1 kg a várható érték.

Nagyon kis okunk van azt hinni, hogy bekövetkezett egy 0.05 valószínűségű esemény, inkább az utóbbi mellett voksolunk, hogy nem 1 kg a várható érték. Azaz 95%-os biztonsággal (0.05 szignifikanciával) úgy döntünk, hogy csaltak. Ellenkező esetben, ha az 1 kg benne van a konfidenciaintervallunban, viszont úgy döntünk, hogy nem csaltak. Lehet, hogy hibásan döntöttünk. Úgy is dönthettünk hibásan, hogy felmentettük a boltot a vád alól, holott az igaz volt. Vizsgáljuk meg a hibás döntések valószínűségét!

Fogalmazzuk meg a feladatot a következőképpen: a H_0 ún. *null-hipotézis* és a H_1 *alternatív hipotézis (ellen-hipotézis)* közt szeretnénk dönteni. Esetünkben az $X \sim \mathcal{N}(\mu, \sigma_0^2)$ háttérváltozó ismeretlen μ várható értékére vonatkoznak a hipotézisek (a σ_0 szórást most ismertnek vesszük).

$$H_0 : \mu = \mu_0 (= 1 \text{ kg}) \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

(Valójában itt a $H_1 : \mu < \mu_0$ alternatívát kellene inkább vizsgálni, ezt egyoldali ellen-hipotézisnek nevezzük, és a 39. példában tárgyaljuk is.)

A döntést az $X_1, \dots, X_n$ független azonos eloszlású minta, illetve az ebből számolt

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0} \sqrt{n}$$

statisztika alapján hozzuk. Ettől függetlenül választunk egy α szignifikanciát (esetünkben $\alpha = 0.05$), és ehhez meghatározzuk a

$$z_{\alpha/2} = \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)$$

ún *kritikus érték*et. A konfidencia-intervallum szerkesztésénél láttuk, hogy

$$\mathbb{P}_{\mu_0}\left(\mu_0 \in \left(\bar{X} - \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}\right)\right) = \mathbb{P}_{\mu_0}\left(|Z| < z_{\alpha/2}\right) = 1 - \alpha\,.$$

Tehát H_0 fennállása esetén μ_0 benne van $1-\alpha$ valószínűséggel a fenti, $\bar{X}$ körüli, szimmetrikus konfidenciaintervallumban. Ezzel ekvivalens, hogy $\bar{X}$ standardizáltjának, a Z valószínűségi változónak az abszolút értéke kisebb, mint a $z_{\alpha/2}$ kritikus érték. Ezért az ún. *z-próba* (*u-próba*) a következő lépésekből áll:

- A mintából kiszámoljuk a z próbastatisztikát.
- Az adott α *szignifikanciához* meghatározzuk az $z_{\alpha/2}$ ún. kritikus értéket.
- Döntünk: ha $|z| < z_{\alpha/2}$, akkor α szignifikanciával elfogadjuk H_0-t, a $|z| \geq z_{\alpha/2}$ esetben pedig elutasítjuk azt. Ezzel ekvivalens, hogy az $\mathbf{x}$ minta-realizáció benne van az

  $$\mathcal{X}_k = \{\mathbf{x} \,:\, |z(\mathbf{x})| \geq z_{\alpha/2}\}$$

  ún. *kritikus tartomány*-ban (elutasítási tartományban). Ha ez megtörténik, azt mondjuk, hogy a kenyerek súlya (α szignifikanciával) szignifikánsan eltér az 1 kg-tól.

Példánkban $\bar{x} = 0.98$, $\mu_0 = 1$, $n = 25$ és legyen $\sigma_0 = 0.05$. Így $z = -2$. $\alpha = 0.05$ mellett $z_{\alpha/2} = 1.96$, ezért 0.05 szignifikanciával el kell utasítanunk a null-hipotézist, azaz megállapítjuk, hogy csaltak. $\alpha = 0.01$ mellett ezt már nem tudjuk megtenni, ugyanis akkor $z_{\alpha/2} = 2.58$, ezért 0.01 szignifikanciával el kell fogadnunk a null-hipotézist. Ez nem meglepő, hiszen a konfidencia-intervallum szerkesztésénél megállapítottuk, hogy az $1-\alpha$ szint növelése (α csökkentése) növeli a konfidenciaintervallum szélességét (a mintaelemszám növelése viszont csökkenti azt). Azt mondhatjuk tehát, hogy 0.05 szignifikanciával állíthatjuk, hogy csaltak, de 0.01-el már nem állíthatjuk ugyanezt. (Azaz a boltot „elsőfokon" elítélik, de egy szigorúbb bíróság „másodfokon" felmenti a vád alól. A szigorúság a vádlott érdekeit képviseli: minél kisebbé akarják tenni annak valószínűségét – másodfokon ez 0.01 –, hogy ártatlanul elítéljék.)

A standard normális eloszlásfüggvény táblázatából kikereshető, hogy $\alpha = 0.0456$ esetén lenne $z_{\alpha/2} = 2$, azaz ez lenne az a legkisebb α, ami mellett már, illetve 95.44% lenne az a legnagyobb biztonság, ami mellett még el tudnánk utasítani a null-hipotézist. Ezt a határ α-t nevezzük *p-érték*nek.

Döntésünkkor kétfajta hiba is felléphet:

- *I. fajú hiba*: H_0 fennáll, mégis elutasítjuk.
- *II. fajú hiba*: H_0 nem áll fenn, mégis elfogadjuk.

(A fenti példában I. fajú hibát követünk el, ha elítéljük az ártatlant, és II. fajút, ha felmentjük a bűnöst.)

Jelölje p_1 illetve p_2 az I. illetve II. fajú hiba valószínűségét. Nyilván

$$p_1 = \mathbb{P}_{\mu_0}\left(|Z| \geq z_{\alpha/2}\right) = \alpha,$$

így ezt a fajta hibát uralni tudom a α megválasztásával. A másodfajú hiba azonban függ a valódi $\mu \neq \mu_0$ paraméterértéktől:

$$p_2 = \mathbb{P}_\mu\left(|Z| < u_{\alpha/2}\right),$$

továbbá függ α-tól és a mintaelemszámtól is. Hogy a μ-től való függés mikéntjét megnézzük, vezessük be a

$$\beta(\mu) = p_2 = \mathbb{P}_\mu\left(|Z| < z_{\alpha/2}\right)$$

másodfajú hibavalószínűséget, ill. a

$$\gamma(\mu) = 1 - p_2 = \mathbb{P}_\mu\left(|Z| \geq z_{\alpha/2}\right)$$

*erőfüggvény*t, melyet a következő alakban írunk fel:

$$\begin{aligned}
\gamma(\mu) &= 1 - \mathbb{P}_\mu\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma_0}\sqrt{n} < z_{\alpha/2}\right) = \\
&= 1 - \mathbb{P}_\mu\left(-z_{\alpha/2} - \Delta_n < \frac{\bar{X} - \mu}{\sigma_0}\sqrt{n} < a_{\alpha/2} - \Delta_n\right) = \\
&= 1 - \Phi(z_{\alpha/2} - \Delta_n) + \Phi(-z_{\alpha/2} - \Delta_n) = \\
&= 2 - \Phi(z_{\alpha/2} - \Delta_n) - \Phi(z_{\alpha/2} + \Delta_n),
\end{aligned}$$

ahol

$$\Delta_n = \frac{\mu - \mu_0}{\sigma_0}\sqrt{n},$$

és $\frac{\bar{X}-\mu}{\sigma_0}\sqrt{n} \sim \mathcal{N}(0,1)$, ha μ a valódi várható érték.

Nézzük meg, mi történik, mikor n-t és α-t rögzítjük, és μ-t változtatjuk. Ekkor

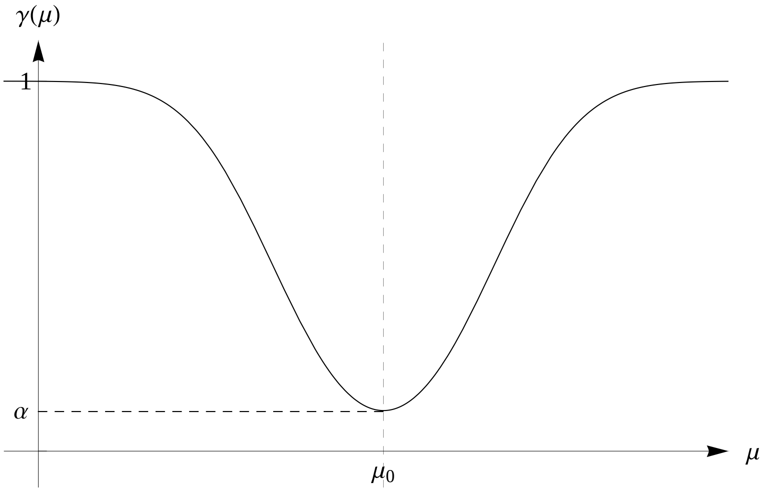
$$\begin{aligned}
&\frac{\partial\gamma(\mu)}{\partial\mu} = \phi(z_{\alpha/2} - \Delta_n)\frac{\sqrt{n}}{\sigma_0} - \phi(z_{\alpha/2} + \Delta_n)\frac{\sqrt{n}}{\sigma_0} = \\
&= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma_0}\left[\exp\left(\frac{-(z_{\alpha/2} - \Delta_n)^2}{2}\right) - \exp\left(\frac{-(z_{\alpha/2} + \Delta_n)^2}{2}\right)\right] \begin{cases} < 0, & \text{ha } \mu < \mu_0, \\ = 0, & \text{ha } \mu = \mu_0, \\ > 0, & \text{ha } \mu > \mu_0. \end{cases}
\end{aligned}$$

Ezért az erőfüggvény $(-\infty, \mu_0)$-on fogy, μ_0-ban minimuma lenne (ha ott értelmezve lenne), a minimum értéke α lenne, $(\mu_0, +\infty)$-n pedig nő (ezt úgy is mondják, hogy a z-próba *torzítatlan*: mindig nagyobb valószínűséggel ítélik el a bűnöst, mint az ártatlant). Az erőfüggvény fenti alakjából az is látható, hogy

$$\lim_{\mu\to-\infty}\gamma(\mu) = \lim_{\mu\to+\infty}\gamma(\mu) = 1.$$

Tehát az erőfüggvény a legkisebb μ_0 közelében, és egyre nagyobb (1-hez tart), ha μ távolodik μ_0-tól, mint azt az 1. ábra mutatja. Ekvivalens módon, a másodfajú hiba a legnagyobb μ_0 közelében (nehéz ekkor különbséget tenni), és egyre kisebb, ha μ távolodik μ_0-tól.

Megjegyzzük, hogy rögzített μ és α esetén – az előbbi gondolatmenettel – az erőfüggvény szintén 1-hez tart, ha $n \to \infty$ (ezt úgy mondják, hogy



1. ábra. Egymintás, kétoldali z-próba erőfüggvénye

a z-próba konzisztens), azaz az 1. ábrán látható erőfüggvény sokkal meredekebb „nagy” n-re, mint „kicsire”. Ugyanakkor rögzített μ és n esetén az erőfüggvény annál nagyobb, minél nagyobb az α. Ez következik abból, hogy az erőfüggvények nem metszik egymást, mint azt átni fogjuk a 41.példában. Ekvivalens módon, rögzített μ és n esetén, a másodfajú hiba annál nagyobb, minél kisebb az elsőfajú, és megfordítva.

Mivel csak az elsőfajú hiba „uralható”, a másodfajú változása pedig vele ellentétes, előbbit nem érdemes túlságosan leszorítani. Az is egy megoldás, hogy a H_0, H_1 szereposztást választjuk meg úgy, hogy a másodfajú hiba elkövetése ne legyen fatális, az első fajú hibáé legyen a súlyosabb vétség, ennek valószínűségét viszont tetszőlegesen kicsivé tudjuk tenni kellőképpen „kis” szignifikancia választásával.

Például gyógyszer-hatásvizsgálatnál legyen

$$H_0: \text{ a gyógyszer hatástalan vagy káros,} \quad H_1: \text{ a gyógyszer hatásos.}$$

Ilyenkor az uralhatatlan másodfajú hiba azt jelenti, hogy egy hatásos gyógyszert nem vezetnek be, mert hatástalannak vagy károsnak minősítjük, ami azért nem okoz fatális problémákat. Az elsőfajú hiba – hogy hatásosnak minősítünk és bevezetünk egy hatástalan, netán káros készítményt – valószínűsége viszont kellően kicsivé tehető, például legyen $\alpha = 0.001$, így ennek bekövetkezése nagyon valószínűtlen. Általában is, az orvosi gyakorlatban a null-hipotézis gyakran a pejoratív verziót tartalmazza: nincsen hatása egy kezelésnek, egy klinikai mérésnek nincs diagnosztizáló hatása, stb., tehát örülünk, ha ezt el tudjuk utasítani minél kisebb I. fajú hibával. Megjegyezzük, hogy a z-próba egy gyógyszer hatástalanságát úgy jellemzi, hogy a gyógyszert szedése hatására a paraméter nem változik (pl. ugyanaz a vérnyomás, mint a gyógyszer szedése előtt volt), ami egyenlőséggel fejezhető ki.

Más szituációban viszont inkább nagynak választjuk az elsőfajú hibát. Például egy szigorúan rögzített méretű alkatrész gyártásakor gyakran előfordul, hogy a gyártóberendezés kopása miatt a várható érték megváltozik (a szórás kicsi). Minőségellenőrzést végzünk arra vonatkozóan, hogy az alkatrésze megfelel-e a szabványnak. Ekkor a

$$H_0:\ \text{a várható érték megegyezik a szabvány mérettel},\quad H_1:\ \text{nem egyezik meg}$$

hipotézisek közötti választásnál viszonylag nagy α-t kell választanunk, ha szigorúak akarunk lenni: vállaljuk, hogy selejtnek minősítünk egy jó alkatrészt is, semmint véletlenül rosszat építsünk be.

Elterjedt az a gyakorlat, hogy nem adjuk meg előre α-t, hanem nézzük, hogy mi az a legkisebb α, amelyre α szignifikanciával már el tudjuk utasítani a null-hipotézist. A felhasználó aztán eldönti, elég-e neki ekkora szignifikancia. a programcsomagok ezt a küszöb-α-t írják ki, amit *p-érték*nek neveznek. Amúgy, ha egy próba konzisztens, „kellően nagy” mintaelemszám esetén a másodfajú hiba tetszőlegesen kicsivé tehető, így ilyenkor nyugodtan kicsinek választhatjuk α-t.

39. *Folytonos alapeloszlás, egyoldali alternatíva (egymintás, egyoldali z-próba).*

Az előző példabeli vádat most átfogalmazzuk egyoldalira:

$$H_0:\ \mu \geq \mu_0 \quad \text{versus} \quad H_1:\ \mu < \mu_0$$

(azt tekintjük csak csalásnak, ha kimondottan kisebb a kenyerek súlyának várható értéke, mint $\mu_0 = 1$ kg). Az elutasítási (kritikus) tartomány ekkor

$$\mathcal{X}_k = \{\mathbf{x}\ :\ z(\mathbf{x}) \leq -z_\alpha\},$$

a Tankönyv IV. fejezet, 2. paragrafus alapján.

Ekkor 0.05 és 0.025 szignifikanciával biztonsággal elutasítjuk a null-hipotézist (azaz állítjuk, hogy csaltak). A $-z_\alpha = -2$ egyenletből adódik, hogy a legkisebb szignifikancia, ami mellett még el tudjuk utasítani H_0-t, $\alpha = 0.0228$. (A 38.példa kétoldali esetében ez éppen a duplája, 0.0456 volt.) Azaz az átfogalmazott vád alapján nagyobb biztonsággal (kisebb p-értékkel) állíthatjuk, hogy csaltak ugyanazon evidencia alapján. Mégegyszer, a p-érték annak a valószínűsége, hogy alaptalanul ítéljük el a boltot, ha az ártatlan.

Az erőfüggvény (rögzített n és α mellett)

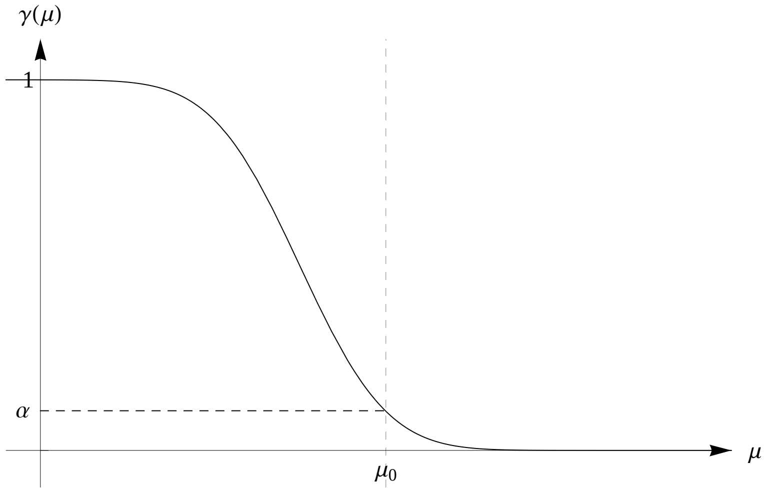
$$\gamma(\mu) = \mathbb{P}_\mu(Z \leq -z_\alpha) = \mathbb{P}_\mu(\frac{\bar{X} - \mu}{\sigma_0}\sqrt{n} + \Delta_n \leq -z_\alpha) = \Phi(-z_\alpha - \Delta_n).$$

Látható, hogy itt $\gamma(\mu)$ szigorúan monoton csökken μ-ben, mint azt a 2. ábra mutatja.

40. *Diszkrét alapeloszlás, egyoldali alternatíva.*

Egy adott betegségből való felgyógyulás 60%-os a szokásos gyógyszerrel. Egy új gyógyszer gyártói be akarják bizonyítani, hogy ez az arány szignifikánsan nagyobb az új gyógyszerrel. 20 páciensen próbálják ki a gyógyszert, és közülük a felgyógyultak számát jelölje X. Kérdés az, hogy milyen szignfikanciával lehet az új gyógyszer hatásosságát bizonyítani az alábbi kritikus tartományok mellett:

$$\mathrm{A}:\ \mathcal{X}_k = \{X \geq 15\},\quad \mathrm{B}:\ \mathcal{X}_k = \{X \geq 18\},\quad \mathrm{C}:\ \mathcal{X}_k = \{X \geq 14\}.$$



2. ábra. Egymintás, egyoldali z-próba erőfüggvénye

Megoldás. Itt $X_1, \ldots, X_{20}$ fae. Bernoulli minta p paraméterrel ($0 < p < 1$), ahol $X_i = 1$, ha az i-edik páciens felgyógyul az új gyógyszertől, és 0, ha nem. A teszt-statisztika $X = \sum_{i=1}^{20} X_i \sim \mathcal{B}_{20}(p)$, melynek alapján döntünk a

$$H_0 : p \leq 0.6 \quad \text{versus} \quad H_1 : p > 0.6$$

alternatíváról, és örülünk, ha el tudjuk utasítani (ez jelenti azt, hogy az új gyógyszer hatásosabb, mint a régi).

A kritikus tartományok három döntési stratégiának felelnek meg, melyek mellett vizsgáljuk a γ-függvényt. Pl. (A) esetben ez a következő:

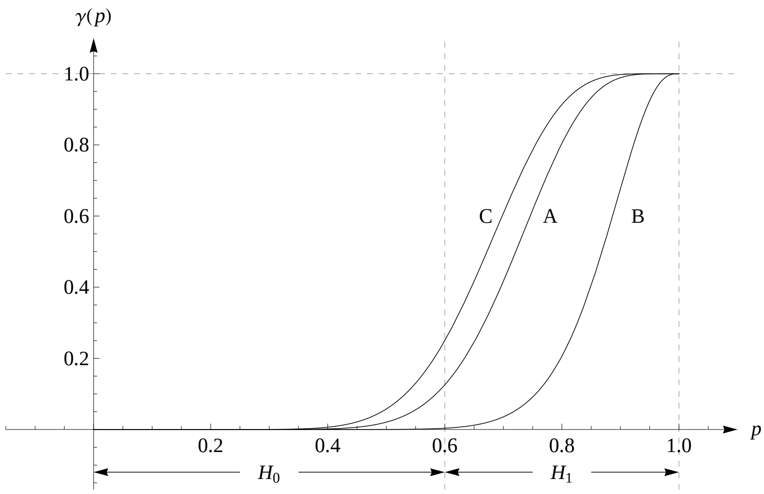
$$\gamma(p) = \mathbb{P}(X \geq 15) = 1 - \sum_{k=0}^{14} \binom{20}{k} p^k (1-p)^{20-k} = 1 - F_p^+(14),$$

ahol F_p^+ a $\mathcal{B}_{20}(p)$-eloszlás jobbról folytonos eloszlásfüggvénye (értékei táblázatokból kikereshetők). Könnyen látható, hogy a $\gamma(p)$ függvény szigorúan monoton nő p-ben, és a $p = 0.6$ helyen felvett értéke (a fenti képletbe való behelyettesítéssel) az (A) esetben 0.126. Az 4. táblázat mutatja az erőfüggvény néhány értékét a három stratégia mellett.

p	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$(A)\ \gamma(p) = P(X \geq 15)$	0.000	0.002	0.021	0.126	0.416	0.804	0.989
$(B)\ \gamma(p) = P(X \geq 18)$	0.000	0.000	0.000	0.004	0.035	0.206	0.677
$(C)\ \gamma(p) = P(X \geq 14)$	0.000	0.006	0.058	0.250	0.608	0.913	0.998

4. táblázat. Az erőfüggvény néhány értéke a három stratégia mellett

Mivel H_0 most összetett, többféle I. fajú hiba is fellép, ezek maximuma, a $\max_{p \leq 0.6} \gamma(p)$ érték azonban a γ-fügvény monotonitása miatt $\gamma(0.6) =$



3. ábra. Az erőfüggvény ábrázolása a három stratégia mellett a gyógyszer hatásvizsgálatnál

0.126 az (A) esetben. Tehát 0.126 az a p-érték, ami mellett az (A)-stratégiával még el tudjuk utasítani H_o-t. A $p > 0.6$ értékekre az erőfüggvény monoton nő és 1, ha $p = 1$.

Az erőfüggvények hasonló viselkedést mutatnak a szigorúbb (B) és az enyhébb (C) stratégiák mellett is (mint azt a 3. ábra mutatja), és a velük adódó szignifikanciák 0.004 ill. 0.250 lesznek. Tehát legbiztonságosabban a (B) stratégiával tudjuk elutasítani H_0-t (bizonyítani az új gyógyszer hatékonyságát), azonban ez túl szigorú evidenciát követel (legalább 18 embernek fel kell gyógyulnia a 20-ból, hogy ezt állíthassuk). A (C) stratégia szignifikanciája viszont túlságosan nagy, 0.250 a valószínűsége, hogy alaptalanul állítjuk az új gyógyszer hatásosabb voltát. Így végülis maradhatunk az (A) stratégiánál, melynek szignifikanciája még elfogadható.

Megjegyezzük, hogy a $p > 0.6$ értékekre lesz valójában a γ-függvény az erőfüggvény, és a három stratégia mellett, akárcsak a szignifikanciáknál, a B stratégia erőfüggvénye kisebb étréket vesz fel, mint A-é, az pedig kisebbet, mint B-é ($\forall p > 0.6$). A másodfajú hibavalószínűségeknél, mivel $\beta(p) = 1 - \gamma(p)$, ha $p > 0.6$, ez a reláció megfordul. A legszigorúbb (B) stratégia dolgozik a legnagyobb másodfajú hibával, azaz itt lesz a legnagyobb a valószínűsége annak, hogy hatástalannak minősítik a hatásosabb gyógyszert.

□

41. *Diszkrét alapeloszlás, kétoldali alternatíva.*

Egy macskaeledel gyártó meg akarja állapítani, hogy az A és B termékek különböző módon keltik-e fel a macskák érdeklődését. Így két azonos méretű tálat megtölt A ill. B eledellel, és odahív 15 macskát. Azt látja, hogy közülük X eszik A-ból, a többi B-ből. Ennek alapján milyen szignfikanciával állíthatjuk azt, hogy az A és B eledel nem egyformán vonzza a macskákat az alábbi kritikus tartományok mellett:

$$\mathrm{a}: \mathcal{X}_k = \{X \le 4 \quad \text{vagy} \quad X \ge 11\}, \quad \mathrm{b}: \mathcal{X}_k = \{X \le 3 \quad \text{vagy} \quad X \ge 12\},$$

ill. amellett az evidencia mellett, hogy azt látjuk: 5 macska eszik A-ból, 10 pedig B-ből.

Megjegyezzük, hogy a kétoldali alternatíva miatt (A jobban vonzza a macskákat, mint B, vagy megfordítva) itt a kritikus tatományok is szimmetrikusak a [0,15] intervallum felezőpontjára. Az evidenciához szimmetrizálva a

$$\mathrm{c}: \mathcal{X}_k = \{X \le 5 \quad \text{vagy} \quad X \ge 10\}$$

kritikus tartományt konstruálhatjuk, mint a legszűkebbet, ami még (a határán) tartalmazza az 5-öt.

Megoldás. Most $X_1, \dots, X_{15}$ a fae. Bernoulli minta p paraméterrel ($0 < p < 1$), ahol $X_i = 1$, ha az i-edik macska A-ból eszik, és 0, ha B-ből. A teszt-statisztika $X = \sum_{i=1}^{15} X_i \sim \mathcal{B}_{15}(p)$, melynek alapján döntünk a

$$H_0: p = 0.5 \quad \text{versus} \quad H_1: p \ne 0.5$$

alternatíváról, és ha el tudjuk utasítani H_0-t, az azt jelenti, hogy szignifikáns különbség van a kétféle macskaeledel közt.

A kritikus tartományok három döntési stratégiának felelnek meg, melyek mellett vizsgáljuk az erőfüggvényt. Pl. (a) esetben ez a következő:

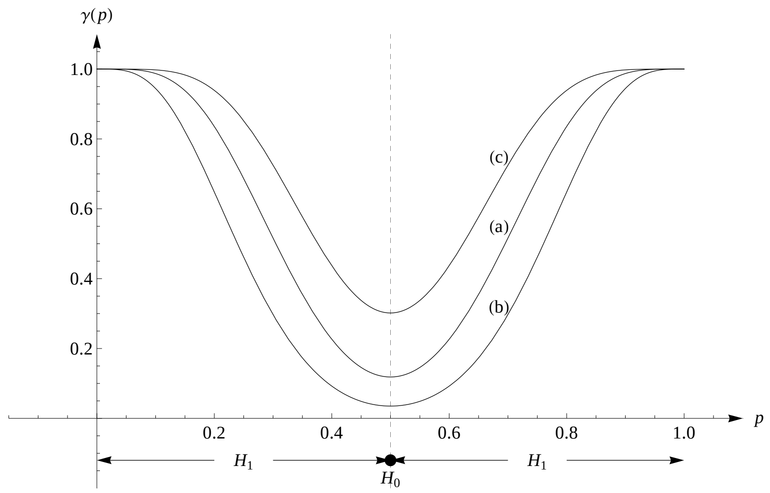
$$\gamma(p) = \mathbb{P}(X \le 4) + \mathbb{P}(X \ge 11) = F_p^+(4) + 1 - F_p^+(10).$$

Könnyen látható, hogy a $\gamma(p)$ függvény szigorúan monoton fogy p-ben $p = 0.5$-ig, majd szigorúan monoton nő (1-ig), és a $p = 0.5$ helyen felvett értéke (a fenti képletbe való behelyettesítéssel) 0.118 az (a) esetben.

Az erőfüggvények hasonló viselkedést mutatnak a szigorúbb (b) és a valóságot tükröző enyhébb (c) stratégiák mellett is (ezt a 4. ábra mutatja), és a velük adódó szignifikanciák 0.036 ill. 0.302 lesznek. Tehát legbiztonságosabban a (b) stratégiával tudjuk elutasítani H_0-t, azonban ez túl szigorú evidenciát követel (legalább 3:12 arányban kell a macskáknak megoszlania). A valóságot tükröző (c) stratégia szignifikanciája viszont túlságosan nagy, 0.302 a valószínűsége, hogy alaptalanul állítjuk a két eledel különbözőségét.

Megjegyezzük, hogy az erőfüggvények a kölönböző (véges sok) stratégia mellett továbbra sem metszik egymást, a stratégiákkal csak véges sok szignifikancia érhető el, és hogy (b) ereje kisebb, mint (a)-é, viszont nagyobb a másodfajú hibavalószínűség (b)-vel, mint (a)-val.

Érdekes, hogy a macskák 5:10 arányú megoszlása nem elég erős bizonyíték. Ez azért van így, mert akárcsak az előző feladatban, kis mintáink vannak (nem akarunk egy új gyógyszert túl sok emberen kipróbálni ill. nem akarunk macskaeledellel túl sok macskát megkínálni pusztán kísérleti célokból). A továbbiakban látunk módszert arra, hogy „nagy” minták esetén ($n \ge 30$) a populációs arányra z-statisztikával konstruálhatunk próbát a CHT alapján. Ha pl. 150 macska oszlik meg 50:100 arányban, az ott már igen erős bizonyíték lesz az A és B termékek különbözőségére, mint azt a 42. példa mutatja. □



4. ábra. Az erőfüggvény ábrázolása a három stratégia mellett a macskaeledel példában

42. *Populációs arány nagy mintára, kétoldali alternatíva.*

A feladat ugyanaz, mint az előbb, csak most $n = 150$ macska lakmározik. Határozzuk meg, a macskák milyen arányú megoszlása esetén lesz a p-érték 0.05 ill. 0.01!

Gondolatébresztő. Most is $X_1, \ldots, X_n$ a fae. Bernoulli minta p paraméterrel ($0 < p < 1$). Mivel n „nagy”, a Moivre–Laplace tétel értelmében $X = \sum_{i=1}^n X_i \sim \mathcal{B}_n(p)$ eloszlása, és így $\bar{X}$ eloszlása is közelíthető normálissal, utóbbié $\mathcal{N}(p, \frac{p(1-p)}{n})$-eloszlással, ahol a paraméterekre az ún. *populációs arány* $r = \bar{X}$ és $\frac{r(1-r)}{n}$ „jó” becslések. Így a

$$H_0 : p = 0.5 \quad \text{versus} \quad H_1 : p \neq 0.5$$

alternatíva vizsgálatára konstruált próbastatisztika

$$Z = \frac{r - 0.5}{\sqrt{r(1-r)}} \sqrt{n},$$

ami H_0 fennállása esetén közelítően standard normális („nagy” n-re). A kritikus tartomány ugyanaz, mint a kétoldali z-próbáé:

$$\mathcal{X}_k = \{|z| \geq z_{\alpha/2}\},$$

de hangsúlyozzuk, hogy itt nem z-próbáról, hanem populációs arány vizsgálatáról van szó z-statisztikával. Fontos, hogy olyan próbastatisztikát tudjunk konstruálni, melynek pontos, vagy („nagy” n esetén) aszimptotikus eloszlását ismerjük H_0 fennállásakor.

Ha 150 macskából 50 eszik az A, 100 pedig a B eledelből, akkor $r = \frac{1}{3}$ és a $z = -4.33$ érték adódik. Ez annak a kritikus tartománynak van a határán, melyre $z_{\alpha/2} = |-4.33|$, így a p-értéket adó α gyakorlatilag 0 (sok-sok 0 tizedesjeggyel). Ez azt jelenti, hogy az 50:100 evidencia alapján nagyon nagy biztonsággal elutasíthatjuk H_0-t, és még inkább megtehetjük ezt a 40:110 vagy a 30:120 evidenciák alapján.

Megjegyezzük, hogy itt a p-re konstruált $1-\alpha$ szintű konfidenciaintervallum a következő:

$$r \pm z_{\alpha/2}\sqrt{\frac{r(1-r)}{n}}. \tag{3}$$

Ezért a

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0$$

alternatívára H_0-t „nagy” n esetén kézenfekvő elfogadni, ha p_0 benne van a (3) konfidenciaintervallumban. Ez ekvivalens azzal, hogy a

$$Z = \frac{r - p_0}{\sqrt{r(1-r)}}\sqrt{n}$$

próbastatisztikára, ami H_0 fennállása esetén aszimptotikusan standard normális, $|Z| < z_{\alpha/2}$ teljesül.

Ugyancsak megjegyezzük, hogy H_0 fennállása esetén a

$$Z' = \frac{r - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$$

próbastatisztika is aszimptotikusan standard normális, így használhatnánk ezt is. Azonban, ha n „nagy”, akkor Z-ben és Z'-ben is $\sqrt{n}$ dominál, így a döntésben nincs nagy különbség.

□

Ezek után a feladat az, hogy határozzuk meg, a macskák milyen arányú megoszlása esetén lesz a p-érték 0.05 ill. 0.01!

43. *Populációs arány nagy mintára, egyoldali alternatíva.*

A feladat ugyanaz, mint a 40. példában, csak most $n = 200$ betegen próbálják ki az új gyógyszert. Határozuuk meg, legalább hánynak kell felgyógyulnia, hogy a p-érték 0.05 ill. 0.01 legyen!

Gondolatébresztő. Most is $X_1, \ldots, X_n$ a fae. Bernoulli minta p paraméterrel ($0 < p < 1$). Mivel n „nagy”, a Moivre–Laplace tétel értelmében $X = \sum_{i=1}^{n} X_i \sim \mathcal{B}_n(p)$ eloszlása, és így $\bar{X}$ eloszlása is közelíthető normálissal, utúbbié $\mathcal{N}(p, \frac{p(1-p)}{n})$-eloszlással, ahol a paraméterekre $r = \bar{X}$ és $\frac{r(1-r)}{n}$ „jó” becslések. Így a

$$H_0 : p \leq 0.6 \quad \text{versus} \quad H_1 : p > 0.6$$

alternatíva vizsgálatára konstruált próbastatisztika

$$Z = \frac{r - 0.6}{\sqrt{r(1-r)}}\sqrt{n},$$

ami $p = 0.6$ esetén esetén standard normális. A kritikus tartomány ugyanaz, mint az egyoldali z-próbáé:

$$\mathcal{X}_k = \{z \geq z_\alpha\}.$$

Az erőfüggvény monotonitása miatt ez összetett H_0-ra is megfelel.

Ha 200 betegből 140 gyógyul fel, akkor $r = \frac{140}{200}$ és $z = 5.09$ adódik. Ez annak a kritikus tartománynak van a határán, melyre $z_\alpha = 5.09$, így a p-értéket adó α gyakorlatilag 0 (sok-sok 0 tizedesjeggyel). Ez azt jelenti, hogy már a 140:60 evidencia alapján is nagyon nagy biztonsággal elutasíthatjuk H_0-t, és még inkább a 150:50 vagy a 180:20 evidenciák alapján, azaz nyugodtan állíthatjuk, hogy az új gyáogyszer hatásosabb, mint a régi. $\square$

Ezek után a feladat az, hogy határozuuk meg, legalább hány betegnek kell felgyógyulnia, hogy a p-érték 0.05 ill. 0.01 legyen!

44. *Populációs arányok összehasonlítása nagy mintákban (kétmintás eset).*

Azonosan valószínű-e a hypertónia átlagsúlyú és elhízott lakosságnál? A null-hipotézis: igen, s a felvetett kérdés alapján kétoldali az alternatíva. Tudjuk, hogy 4200 átlagos testsúlyú felnőtt közül 792, míg 1000 elhízott felnőtt közül 249 szenved hypertóniában egy bizonyos társadalmi rétegben. Adataink: $n_1 = 4200$, $r_1 = 792/4200 = 18.9\%$, $n_2 = 1000$, $r_2 = 249/1000 = 24.9\%$. Ezekből

$$Z = \frac{r_1 - r_2}{\sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}}},$$

amely H_0 fennállása esetén közelítően $\mathcal{N}(0,1)$-eloszlású. Esetünkben

$$z = \frac{-0.06}{0.0149} = -4.2$$

adódik, s mivel ez abszolút értékben még $\alpha = 0.01$ szinten is nagyobb, mint a $z_{\alpha/2}$ kritikus érték, azt mondhatjuk, hogy a hypertónia valószínűsége 0.01 szignifikanciával különbözik átlagos és elhízott lakosságnál. Azaz az elhízás befolyásolja a hypertónia kialakulását.

Ha azt akarjuk tudni, hogy pozitív irányban befolyásolja-e, akkor egyoldali alternatívát állítunk fel. A két minta indexelését most felcseréljük, és ellen-hipotézisünk az, hogy az elhízottak körében nagyobb a hypertónia valószínűsége, mint átlagos testsúlyú lakosságnál (amit bizonyítani szeretnénk). A felcserélés miatt most $z = 4.02$, amely még $\alpha = 0.01$ mellett is nagyobb, mint a z_α kritikus érték. Így azt is elmondhatjuk, hogy a hypertónia valószínűsége 0.01 szignifikanciával szignifikánsan nagyobb az elhízott lakosságnál.

Megjegyezzük még, hogy ezt a próbát használhatjuk két közvéleménykutatás eredményének összehasonlítására is, bővebben a Tankönyv IV. fejezet 2. paragrafusában, Paraméteres próbák alatt olvashatunk erről is.

Szintén megjegyezzük, hogy a kétmintás, kétoldali esetben, mikor a

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 \neq p_2$$

alternatívára a $Z = \frac{r_1 - r_2}{\sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}}}$, próbastatisztikát használjuk, akkor H_0 elfogadása ekvivalens azzal, hogy $p_1 - p_2$ benne van az

$$r_1 - r_2 \pm z_{\alpha/2} \sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}}$$

konfidenciaintervallumban.

Itt H_0 fennállása esetén $r_1 - r_2$ becsült szórása , $\sqrt{\frac{\hat{r}(1-\hat{r})}{n}}$ is használható, ahol $\hat{r} = \frac{n_1 r_1 + n_2 r_2}{n}$ és $n = n_1 + n_2$. Ezzel az ún. összevolt szórással a próbastatisztika

$$Z' = \frac{r_1 - r_2}{\sqrt{\hat{r}(1-\hat{r})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

alakú, de aszimptotikusan hasonlóan viselkedik, mint Z.

7. feladatsor

További feladatok diszkrét eloszlásokra és z-próbára

45. Azt szeretnénk bebizonyítani, hogy egy gyártmányban a selejtarány (θ) a megengedhető 5%-ot meghaladja. Hogy a

$$H_0: \ \theta \leq 0.05 \quad \text{versus} \quad H_0: \ \theta > 0.05$$

alternatívát teszteljük, 25 elemű mintát veszünk. Jelölje X a selejtesek számát a 25-ből. Adja meg a próba szignifikanciáját és vázolja az erőfüggvényt a következő kritikus tartományok esetén:

$$a. \ \mathcal{X}_k = \{X \geq 2\} \qquad b. \ \mathcal{X}_k = \{X \geq 3\} \qquad c. \ \mathcal{X}_k = \{X \geq 4\}.$$

46. A "biológiai oxigénigény" (BOD)a szennyeződés mérőszáma, amelyet példánkban a papírgyárak kibocsátott szennyvizéből határoznak meg naponta. Egy adott papírgyárnál a tavaszi és nyári időszakok folyamán az átlag 3246-nak, a korrigált tapasztalati szórás pedig 757-nek adódott. A gyár célul tűzte ki, hogy az átlagos napi kibocsátás 3000 lesz. Alátámasztják-e az adatok 0.05 szignifikanciával, hogy a gyár nem teljesítette a célt?

47. Tekintsük a

$$H_0: \ \mu \leq 10 \quad \text{versus} \quad H_1: \ \mu > 10$$

alternatívát az $n = 64$, $\sigma = 2$ (ismert) és $\alpha = 0.025$ feltételekkel. Az elutasítási tartomány ekkor a $Z = \frac{\bar{X}-10}{2}\sqrt{64}$ próba-statisztika alapján:

$$\mathcal{X}_k = \{Z \geq 1.96\} = \{\bar{X} \geq 10 + 1.96\frac{2}{\sqrt{64}} = 10.49\}.$$

Ha $\gamma(\mu)$ jelöli az erőfüggvényt, akkor ennek a tesznek az ereje a $\mu_1 = 11$ alternatívára:

$$\gamma(11) = \mathbb{P}_{11}(Z \geq 1.96) = \mathbb{P}_{11}(V \geq \frac{10.49 - 11}{2}\sqrt{64}) = 0.9793,$$

hiszen ekkor $V = \frac{\bar{X}-11}{2}\sqrt{64}$ lesz standard normális eloszlású.

(a) Számítsuk ki a fenti teszt erejét, amikor $\mu_1 = 10.5$!

(b) Számítsuk ki a fenti teszt erejét, amikor $\mu_1 = 10.8$!

(c) Vázolja az erőfüggvényt és állapítsa meg aszimptotikus viselkedését, ha $\mu \to \infty$!

(d) Mit tud mondani a különböző $\mu > 10$ értékek mellett fellépő másodfajú hibákról?

48. Tekintsük a

$$H_0: \ \mu = 77 \quad \text{versus} \quad H_1: \mu \neq 77$$

alternatívát az $\alpha = 0.05$ szinten, melyhez az alábbi számítógépes eredmény tertozik:

```
Test of mu = 77 versus mu not = 77

Variable                     N     Mean     StDev
malt extract                40   77.458     1.101

Variable       95.0% Conf.  Int.        Z  P-value
malt extract   ( 77.116, 77.799)     2.63    0.009
```

(a) Rövidebb vagy hosszabb lesz a 98%-os konfidenciaintervallum, mint a számítógépes eredményben szereplő? Számítsa is ki a 98%-os konfidenciaintervallum végpontjait!

(b) A számítógépes eredményben szereplő Z érték segítségével döntsünk 0.05 szignifikanciával a

$$H_0: \ \mu \leq 77 \quad \text{versus} \quad H_1: \mu > 77$$

alternatíváról (mindenféle számolás nélkül, csak a számítógépes eredményre hagyatkozva)!

Az adatok egyébként a 5. táblázatban láthatók.

75.3	77.9	77.6	76.6	78.3	77.9	77.5	77.6	77.1	78.0
77.9	76.3	75.7	77.4	77.4	76.9	77.9	77.4	78.1	77.4
76.4	79.1	80.0	76.9	78.5	78.4	77.8	80.4	75.9	77.0
79.2	76.2	77.0	75.9	77.9	78.4	76.7	76.4	76.6	77.4

5. táblázat. Egy maláta eszencia %-os koncentrációja különböző mérésekkor

49. Egy irodalmkritikus bizonyítani akarja, hogy egy novellában a mondatonkénti szavak átlagos száma nem 9.1. Egy, a novellából vett 36 mondatos minta átlaga 8.6-nak, korrigált tapasztalati szórása pedig 1.2-nek adódott.

(a) Határozza meg a null- ill. az alternatív hipotézist!

(b) Adja meg a próba-statisztikát!

(c) Mi lesz a kritikus (elutasítási) tartomány?

(d) Mi a teszt konklúziója $\alpha = 0.1$ mellett?

(e) Mekkora másodfajú hibát követünk el, ha $\alpha = 0.1$ és a mondatonkénti szavak átlagos száma 10?

(f) Határozza meg a p-értéket a fenti 8.6-os evidencia mellett!

8. feladatsor

Kis minták, t-próba

50. R.A. Fisher két altató hatását vizsgálta ugyanazon a 10 páciensen. A többlet alvás, amit az A ill. B altató okozott (órákban) a következő volt:

No.	A	B	B-A
1.	$+0.7$	$+1.9$	$+1.2$
2.	-1.6	$+0.8$	$+2.4$
3.	-0.2	$+1.1$	$+1.3$
4.	-1.2	$+0.1$	$+1.3$
5.	-0.1	-0.1	0.0
6.	$+3.4$	$+4.4$	$+1.0$
7.	$+3.7$	$+5.5$	$+1.8$
8.	$+0.8$	$+1.6$	$+0.8$
9.	0.0	$+4.6$	$+4.6$
10.	$+2.0$	$+3.4$	$+1.4$

Jelölje X az A, Y pedig a B altató által okozott alvástöbbletet! Ugyanazon a 10 emberen próbálták ki mindkét gyógyszert. Az $X_1, \dots, X_{10}$ és $Y_1, \dots, Y_{10}$ minták tehát messze nem függetlenek egymástól, így a $V_i = Y_i - X_i$ mintaelemekkel számolunk, melyek a $V = Y - X$ alvástöbblet-különbségre vonatkoznak. Erre nézve vizsgáljuk a

$$H_0 : \mathbb{E}(V) = 0 \quad versus \quad H_1 : \mathbb{E}(V) \neq 0$$

alternatívát. Ezt nevezzük *páros mintás t-próbának.*

A fenti táblázat B-A oszlopával számolva $n = 10$, $\bar{v} = 1.58$, $s^*_{10} = 1.23$, így a próbastatisztika értéke:

$$t = \frac{1.58 - 0}{1.23} \cdot \sqrt{10} = 4.06,$$

ezt vetjük össze a $t(9)$-eloszlás $(1 - \alpha/2)$-kvantilisével. A t-eloszlás táblázatából látható, hogy a null-hipotézist 0.05 és 0.01 szignifikanciával is elutasítjuk, 0.001 szignifikanciával azonban már nem tudjuk elutasítani. Fisher ezután megnézi azt, mi történne, ha méréseinket két különböző csoporton végrehajtottnak tekintenénk (például 10 fős inszomniás- és 10 fős kontroll-csoport). Ekkor 2-mintás t-próbával már 0.01 szignifikanciával sem tudnánk elutasítani H_0-t. Ezután megvizsgálhatnánk külön-külön 1-mintás t-próbával, hogy az A és B altató hatásos-e (az alvastóbblet szignifikánsan eltér-e 0-tól).

Azt is megkérdezhetnénk, hogy a B altató szignifikansan jobb-e, mint az A? Határozzuk ezt meg a fenti adatok alapján, különböző szignifikanciák mellett, páros mintás t-próbával!

51. 10-10 szeget gyártanak le két különböző gépen. Az átlagos méretek és a korrigált empirikus szórások (cm-ben):

$$\bar{x} = 0.625, \quad \bar{y} = 0.471, \quad s^*_x = 0.754, \quad s^*_y = 1.269.$$

Hasonlítsa össze a varianciákat F-próbával, majd vizsgálja a null-hipotézist, hogy nincsen különbség a méretek tekintetében a két gép közt. Használjon $\alpha = 0.10$ szignifikanciát! Hogyan vizsgálná ki ugyanezt a hipotézist akkor, ha 100-100 szeget gyártanának ugyanezekkel az empirikus jellemzőkkel?

52. A Dunában megmérték a víz által sodort szilárd anyag mennyiségét 14 hétfői reggelen. Azt találták, hogy $\bar{x} = 47, s = 9.4$. Adjon

(a) 95%-os ill. 99%-os biztonságú konfidenciaintervallumot

(b) 95%- os ill. 99%-os biztonságú alsó korlátot

a folyóban sodort szilárd anyagmennyiség átlagára!

(c) A víz minősége elfogadható, ha a fenti anyagmennyiség 49 alatt van. Állítsa fel a megfelelő hipotéziseket és végezze el a hipotézisvizsgálatot 5%-os szignifikancia-szinten!

Megoldás.

(a) A háttérváltozót normális eloszlásúnak feltételezve n megfigyelésből az $1 - \alpha$ megbízhatósági szintű konfidenciaintervallum végpontjai:

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}.$$

A feladatban $\alpha = 0.05$ ill. $\alpha = 0.01$ és $n = 14$. A t-eloszlás táblázatából $t_{0.025}(13) = 2.160$ ill. $t_{0.005}(13) = 3.012$, és így a 95%-os ill. 99%-os konfidenciaintervallumok: $(41.37, 52.63)$ ill. $(39.15, 54.85)$ lesznek.

(b) Az alsó korlát most:

$$\bar{x} - t_{\alpha}(n-1)\frac{s}{\sqrt{n}},$$

ahonnan a $t_{0.05}(13) = 1.771$ ill. $t_{0.01}(13) = 2.650$ értékekkel a 95%-os ill. 99%-os alsó korlátok 42.38 ill. 40.09 lesznek.

Megjegyezzük, hogy az egy- és kétoldali t-értékekre vonatkozó táblázat is ugyanezt mutatja, csak más felállásban.

(c) A hipotézisek: $H_0 : \mu \geq 49$ versus $H_1 : \mu < 49$. A számolt tesztstatisztika: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{47-49}{9.4/\sqrt{13}} = -0.767$, ezért nem tudunk elutasítan, azaz 0.05 szignifikanciával bizonyítani, hogy $\mu < 49$, a 47 átlag evidencia alapján. Ui. $-0.767 > t_{0.05}(13) = -1.771$.

$\square$

53. Egy zoológus 20 gyíkot gyűjtött össze és megmérte a hosszukat:

179	157	169	146	143	131	159	142	141	130
142	116	130	140	138	137	134	114	90	114

Adjunk 95%-os konfidenciaintervallumot a gyíkok átlagos hosszára!

54. Egy genetikai modell szerint egy bizonyos – két különböző fajtájú magról keresztezett – növény 80%-ban alacsony lesz. 200 ilyen keresztezést elvégezve azt találták, hogy közülük 136 alacsony növésű lett.

(a) Ellentmondanak-e ezek az adatok a modell állításának?

(b) Szerkesszen 95%-os konfidenciaintervallumot az alacsony növények arányára!

55. Egy előző széleskörű felmérés alapján állíthatjuk, hogy az 1-2 gyerekes családok heti élelmiszer kiadása átlagosan 148$, a szórás pedig 25$. Egy évvel később ismét felmérést terveznek.

(a) Hány családot kérdezzenek meg, ha azt szeretnék, hogy a becslési hiba 95% valószínűséggel 2$-nál kisebb legyen?

(b) Végül 100 családot megkérdeztek és a kapott minta átlaga 155$ lett, a szórás pedig 22$. Konstruáljon 98%-os konfidenciaintervallumot a családok költségeire!

56. Könyvelési tanulmányaikat kezdő tanulók tesztet töltöttek ki, amely a komputerrel kapcsolatos aggodalmaikat mérte. A következő pontszámokat kapták 15 tanuló esetén:

2.90	1.00	1.90	2.37	3.32	3.79	3.26	1.90
1.84	2.58	1.58	2.90	2.42	3.42	2.53	

(a) Adjon becslést a hallgatók pontszámának szórására!

(b) Szerkesszen 95%-os konfidenciaintervallumot az ismeretlen szórásra!

(c) Ellenőrizze, hogy az (a)-ban kapott becslés a felezőpontja-e a (b)-ben kapott konfidenciaintervallumnak!

57. Megmérték a $30-40$ és a $60-70$ éves korosztályból 250-250 ember alvásidejét. Az alábbi adatok születtek:

	alvásidő		
életkor	< 8	≥ 8	összesen
30-40	173	77	250
60-70	120	130	250
összesen	293	207	500

(a) Alátámasztják-e ezek az adatok, hogy a $30-40$ közöttiek közül többen alszanak 8 óránál kevesebbet átlagban, mint a $60-70$ közöttiek? Az indokláshoz használja a p-értéket!

(b) Legyen p_1 ill. p_2 azoknak az aránya a két korcsoporton belül, akik 8 óránál kevesebbet alszanak. Szerkesszen 95%-os konfidenciaintervallumot $p_1 - p_2$-re!

58. Hogy kiértékeljék egy egyszerűsített adóbevallási ív két változatát, 40-40 embert kértek meg, hogy töltse ki az A1 és A2 típusú íveket. Az adatok összesítése során az alábbiakat kapták a kitöltési időkre:

$$\begin{aligned} A1: &\quad \bar{x} = 12.2, \quad s_1 = 1.1 \\ A2: &\quad \bar{y} = 7.2, \quad\ \ s_2 = 2.1 \end{aligned}$$

Szerkesszen 95%-os konfidenciaintervallumot a két ív kitöltésének átlagidejére!

59. A **58**. feladatban tekinthetők-e egyenlőknek a szórások 0.05 szignifikanciával?

60. Tegyük most fel, hogy a **58**. feladatban nem 40-40 ember töltötte ki az A1-es ill. A2-es ívet, hanem $n_1 = 8$ ill. $n_2 = 7$.

(a) Most is elutasítaná-e 0.05-ös szinten, hogy a két szórás egyenlő?

(b) Adjon 95%-os konfidenciaintervallumot az A2 és A1 ív kitöltési ideje közti különbségre! Ismertesse, milyen feltevésekkel élt!

Megoldás. (a) Nem, mert $\frac{s_2^2}{s_1^2} = 3.645$, viszont az $F(7,8)$ eloszlás felső 0.025-ös kvantilise 4.53.

(b) **60**.a alapján tekintsük egyenlőnek a szórásokat. Ekkor az $1-\alpha$ megbízhatóságú konfidenciaintervallum végpontjai: $\bar{x} - \bar{y} \pm t_{\alpha/2}(n_1 + n_2 - 2)s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, ahol $s_p = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ az ú.n. összevont szórás. Az adatokkal $t_{\alpha/2}(n_1 + n_2 - 2) = 2.160, s_p = 1.639$. Tehát a keresett intervallum $(3.168, 6.832)$.

□

61. 10.13. Két rokon lepkefaj szárnyméretét mérték meg (cm-ben) és ezt kapták:

1. fajta	6	4	7	3
2. fajta	6	9	6	

(a) Határozza meg az összevont tapasztalati szórásnégyzetet!

(b) Adjon becslést a két lepkefaj szárnyméretének közös szórására!

(c) Adja meg a két szárnyméret egyenlőségének teszteléséhez szükséges t-statisztikát!

62. 10.37. Egy táplálkozástudománnyal foglalkozó kutató szeretné tudni, hogy van-e különbség a előtenyésztett baktériumot (psychrotrops, röviden PC). tartalmazó és nem tartalmazó lefölözött tej között. Hét különböző, tejtermékekkel foglalkozó gazdaságból vettek egy-egy mintát a lefölözött tejből. Ezután mindegyik minta egyik felét beoltották PC-vel, a másik felét nem. Az elkészült joghurt keménységét megmérték és az alábbi adatokat kapták:

	tejgazdaság						
joghurt keménység	A	B	C	D	E	F	E
PC-vel	68	75	62	86	52	46	72
PC nélkül	61	69	64	76	52	38	68

(a) Alátámasztják-e ezek az adatok, hogy a PC-vel beoltott (lefölözött) tej nagyobb keménységű, mint a a nem beoltott?

(b) adjon meg egy 90% - os intervallumot az átlagos keménységnövekedésre a PC kezelés után!

63. 10.46. Egy összehasonlító tanulmányban az A és B gyógyszereket próbálták ki 120 ill. 150 betegen és az alábbi adatokat kapták:

	A	B
kezelt	55	88
nem kezelt	70	62
összes	120	150

(a) alátámasztják-e ezek az adatok, hogy a B gyógyszer eredményesebb, mint az A? Hasznnáljon 0.05-ös szignifikanciát.

(b) Szerkesszen 95%-os konfidenciaintervallumot a gyógyultak arányának különbségére a B ill. az A gyógyszer tekintetében!

64. Egy közgazdász egy pénzügyi botrány hatásait szeretné felmérni egy bizonyos részvény árának változékonyságában. Az illető adatokat gyűjt a botrány előtti és utáni árakról és él a feltételezéssel, hogy az árak normális eloszlásúak voltak a botrány előtt és után is, továbbá, hogy a botrány előtti és utáni árak függetlennek tekinthetők. A pénzügyi elméletek szerint az árak közel normális eloszlásúak. A közgazdász tesztelni szeretné, hogy a botrány növelte vagy csökkentette az árak varianciáját. Az esemény előtti 25 elemű realizált minta korrigált tapasztalati szórásnégyzete $s_1^2 = 9.3$ (dollár^2) és az esemény utáni 24 elemű minta korrigált tapasztalati szórásnégyzete $s_2^2 = 3.0$ (dollár^2). Végezzük el a próbát $\alpha = 0.05$ mellett!

Megoldás. Egyolali F-próbát végzünk.

$$H_0 : \sigma_1 = \sigma_2 \quad \text{versus} \quad H_1 : \sigma_1 > \sigma_2$$

tesztelésére itt is az

$$F = \frac{s_1^2}{s_2^2} = \frac{9.3}{3} = 3.1$$

próbastatisztikát használjuk, de a kritikus tartomány

$$\mathcal{X}_k = \{F \geq F_{0.05}(24, 23) = 2.01\}.$$

Tehát elutasítjuk H_0-t 0.05 szignifikanciával.

□

65. Egy közgazdász ellenőrizni szeretné, hogy a nyersolaj ára befolyásolja-e a fogyasztói árindexet (CPI). Kétféle adatot gyűjt össze; az egyik a CPI 14 havi százalékos növekedéseit tartalmazza, amikor a nyersolaj ára 66$ volt, a másik a CPI 9 havi százalékos növekedéseit tartalmazza, amikor a nyersolaj ára 58$ volt. Az adatok: $\bar{x}_1 = 0.317\%, s_1 = 0.12\%, n_1 = 14; \bar{x}_2 = 0.210\%, s_2 = 0.11\%, n_2 = 9$. Teszteljük a két mintához tartozó várható értékek egyenlőségét, feltéve, hogy a CPI varianciák megegyeznek!

9. feladatsor

A mintaelemszám választása, konfidenciaintervallum a szórásra és egyéb tesztek a korrelációra

66. Hány elemű mintát kell vennünk ahhoz, hogy ε pontosságú becslést adjunk a háttéreloszlás ismeretlen (de létező) μ várható értékére $1-\alpha$ biztonsággal? (Tegyük fel, hogy az eloszlás szórás σ_0 adott.)

Megoldás. Ha mintánk normális eloszlásból származik, vagy n „nagy", akkor az $\bar{X}$ körüli ε sugarú konfidenciainervallum sugarára $\frac{z_{\alpha/2}\sigma_0}{\sqrt{n}} \leq \varepsilon$ kell, hogy teljesüljön. Innen $n \geq (\frac{z_{\alpha/2}\sigma_0}{\varepsilon})^2$.

Megjegyezzük, hogy ez az alsó korlát n-re sokkal kisebb, mint az, amit a Csebisev-egyenlőtlenséggel kapnánk:

$$\mathbb{P}(|\bar{X}-\mu| \leq \varepsilon) \geq 1 - \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = 1 - \frac{\sigma_0^2}{n\varepsilon^2} \geq 1-\alpha,$$

ahonnan $n \geq \frac{\sigma_0^2}{\alpha\varepsilon^2}$.

Nagy mintaelemszám esetén a becsült szórással dolgozhatunk. □

67. Szerkesszünk $1-\alpha$ szintű konfidenciaintervallumot a normális háttéreloszlás ismeretlen (de létező) σ szórására!

Megoldás. Lukács tételéből tudjuk, hogy $\frac{(n-1)S_n^{*2}}{\sigma^2} \sim \chi^2(n-1)$, így

$$\mathbb{P}\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S_n^{*2}}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha,$$

ahol $\chi^2_\alpha(n)$ az n szabadsági fokú χ^2-eloszlás $(1-\alpha)$-kvantilise. Innen

$$\mathbb{P}\left(\frac{(n-1)S_n^{*2}}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S_n^{*2}}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

és

$$\mathbb{P}\left(\sqrt{\frac{(n-1)S_n^{*2}}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S_n^{*2}}{\chi^2_{1-\alpha/2}}}\right) = 1-\alpha,$$

ahonnan a határok kiolvashatók. □

68. Az $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$ fae. minta alapján teszteljük a

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad \sigma^2 \neq \sigma_0^2$$

alternatívát!

Megoldás. H_0 elfogadása ekvivalens azzal, hogy a hipotetikus σ_0^2 benne van a **67**. feladatban konstruált $1-\alpha$ szintű konfidenciaintervallumban. Így az α szignifikanciájú próba próbastatisztikája $d = \frac{(n-1){S_n^*}^2}{\sigma_0^2}$, kritikus tartománya pedig:

$$\mathcal{X}_k = \{d \leq \chi^2_{1-\alpha/2} \quad \text{vagy} \quad d \geq \chi^2_{\alpha/2}\}.$$

□

69. Az $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ és $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$ egymástól is független minták alapján teszteljük a

$$H_0 \,:\, \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad \sigma_1^2 \neq \sigma_2^2$$

alternatívát α szignifikanciával! Erre ismerjük az F-próbát, de magyarázzuk meg a feleződést!

Megoldás. Tudjuk, hogy H_0 fennállásakor az

$$F = \frac{{S_X^*}^2}{{S_Y^*}^2}$$

statisztika $\mathcal{F}(n-1, m-1)$ Fisher-eloszlást követ. Ahhoz, hogy α I. fajú hibát produkáljunk, F-et ezen eloszlás $(1-\alpha)$-kvantiliséhez kellene viszonyítani. Azonban a táblázat csak 1-nél nagyobb F-értékeket tartalmaz. Ezért az X–Y szereposztást úgy választjuk meg, hogy $\frac{{S_X^*}^2}{{S_Y^*}^2} \geq 1$ legyen. Így valójában az $F^* = \max\{\frac{{S_X^*}^2}{{S_Y^*}^2}, \frac{{S_Y^*}^2}{{S_X^*}^2}\}$ próbastatisztikát használjuk.

Ha $F^* = F$, akkor $\mathbb{P}(F \geq F_{\alpha/2}(n-1, m-1)) = \alpha/2$. Ha $F^* = \frac{1}{F}$, akkor

$$\begin{aligned}\mathbb{P}(\frac{1}{F} \geq F_{\alpha/2}(m-1, n-1)) &= \mathbb{P}(F \leq \frac{1}{F_{\alpha/2}(m-1, n-1)}) \\ &= \mathbb{P}(F \leq F_{1-\alpha/2}(n-1, m-1)) = \alpha/2.\end{aligned}$$

Ezért a kritikus tartomány:

$$\mathcal{X}_k = \{F^* \geq F_{\alpha/2}(f_1 - 1, f_2 - 1)\},$$

ahol f_1 a nagyobbik, f_2 pedig a kisebbik empirikus varianciájú minta mintaelemszáma.

F-próbát a független mintás t-próba elvégzse elött kell végrehajtani. Ha elutasítjuk H_0-t, akkor a szokásos kétmintás t-próba helyett Welch-próbát kell használni (l. tankönyv). □

70. A korrelációs együtthatóra vonatkozó hipotézisvizsgálat alapja, hogy tetszőleges második momentummal rendelkező eloszlásra igazak az alábbi közelítő összefüggések:

$$\mathbb{E}(R_n) = r + \mathcal{O}(1/n) \quad \text{és} \quad \mathrm{Var}(R_n) = \frac{(1+r^2)^2}{n} + \mathcal{O}(\frac{1}{n^{3/2}})\,,$$

ahol r, illetve R_n jelöli a valódi, illetve az empirikus korrelációs együtthatót, és $\mathcal{O}(\cdot)$ jelentése nagy ordó. Tehát R_n aszimptotikusan torzítatlan és konzisztens becslése r-nek.

Fisher ennél többet bizonyított, nevezetesen azt, hogy

$$Z = \frac{1}{2}\ln\frac{1+R_n}{1-R_n} \to \mathcal{N}(\frac{1}{2}\ln\frac{1+r}{1-r} + \frac{r}{2(n-1)}, \frac{1}{n-3})$$

eloszlásban, azaz az ún. Fisher-féle Z statisztika $n \to \infty$ esetén aszimptotikusan normális eloszlású a fenti paraméterekkel. Így a

$$H_0 : r = r_0 \quad \text{versus} \quad r \neq r_0$$

alternatíva vizsgálatára a

$$\frac{Z - \frac{1}{2}\ln\frac{1+r_0}{1-r_0} + \frac{r_0}{2(n-1)}}{\sqrt{\frac{1}{n-3}}}$$

statisztika segítségével z-próba végezhető.

Be lehet látni, hogy 2-dimenziós normális háttéreloszlás esetén a

$$t = \sqrt{n-2}\frac{R_n}{1-R_n^2}$$

statisztika aszimptotikusan $n-2$ szabadsági fokú Student-eloszlású, ha $n \to \infty$ és $r = 0$. Így a

$$H_0 : r = 0 \quad \text{versus} \quad r \neq 0$$

alternatíva vizsgálata történhet a t-próbáéhoz hasonló módszerrel is. Itt H_0 elfogadása egyben a függetlenség hipotézisének elfogadását is jelenti.

71. Nem-normális háttéreloszlás esetén szokták az ún. Spearman-féle rangkorrelációs együtthatóval tesztelni két valószínűségi változó függetlenségét. Annak a null-hipotézisnek a tesztelésére, hogy a háttérváltozó két komponense független, nagy n értékekre az

$$\sqrt{n-1} \cdot r_{sp}$$

statisztikát használjuk, mely a null-hipotézis fennállása esetén aszimptotikusan standard normális eloszlást követ ($n \to \infty$), s mellyel z-próbát hajthatunk végre (r_{sp} definícióját l. az előjel-próbáknál).

10. feladatsor

χ^2-próba

72. *Illeszkedésvizsgálat* Szabályosnak tekinthető-e az alábbi dobókocka $\alpha = 0.05$ szignifikanciával?

(a) Feldobjuk 1200-szor és az egyes oldalak kijövetelének gyakorisága: $\nu_1 = 184$, $\nu_2 = 212$, $\nu_3 = 190$, $\nu_4 = 208$, $\nu_5 = 212$, $\nu_6 = 194$.

(b) Feldobjuk 12000-szer és az egyes oldalak kijövetelének gyakorisága: $\nu_1 = 1840$, $\nu_2 = 2120$, $\nu_3 = 1900$, $\nu_4 = 2080$, $\nu_5 = 2120$, $\nu_6 = 1940$.

Megoldás. (a) A kiszámolt χ^2-statisztika értéke:

$$\chi^2 = \sum_{i=1}^{6} \frac{(\nu_i - 200)^2}{200} = \frac{16^2 + 12^2 + 10^2 + 8^2 + 12^2 + 6^2}{200} = 3.72,$$

amely alapján $\alpha = 0.05$ szignifikanciával elfogadjuk H_0-t, hiszen $\chi^2_{0.05}(5) = 11.1$, sőt $\alpha = 0.01$ szignifikanciával is elfogadnánk.

(b) A kiszámolt χ^2-statisztika értéke most 37.2, amely alapján elutasítjuk H_0-t $\alpha = 0.05$, sőt még kisebb szignifikanciával is. A tanulság, hogy nagyobb mintaelemszám mellett ugyanazok a relatív gyakoriságok már sokkal kevésbé bizonyító erejűek, mint kisebb mintaelemszámnál (a CHT-nek köszönhetően).

□

73. *Homogenitásvizsgálat* 80 ill. 70 gyermek egészségi állapotát vizsgálták két különböző étrend mellett:

	Kiváló	Átlagos	Gyenge	összesen
A étrend	37	24	19	80
B étrend	17	33	20	70
összes	54	57	39	150

Van-e szignifikáns különbség a két étrend közt a gyermekek egészségi állapotát illetően?

Megoldás. Homogenitásvizsgálatot végzünk, ahol

H_0 : az egészségi állapot eloszlása ugyanolyan a kétféle étrend mellett.

Ha ez igaz, akkor az oszlopösszegek alapján becsülhetjük a három egészségi állapot valószínűségét:

$$\hat{p}_1 = \frac{54}{150}, \quad \hat{p}_2 = \frac{57}{150}, \quad \hat{p}_3 = \frac{39}{150}.$$

Így a H_0 mellett várt cellagyakoriságok: $\hat{p}_i n$ ill. $\hat{p}_i m$, $(i = 1, \ldots, r)$, ahol $r = 3$, $n = 80$, $m = 70$ a képletgyűjtemény jelöléseivel. Ezekkel képezve a

$$\sum \frac{(\text{megfigyelt} - \text{várt})^2}{\text{várt}}$$

összeget, a következőt kapjuk:

$$\sum_{i=1}^{r} \frac{(\nu_i - \hat{p}_i n)^2}{\hat{p}_i n} + \sum_{i=1}^{r} \frac{(\mu_i - \hat{p}_i m)^2}{\hat{p}_i m} = \cdots = nm \sum_{i=1}^{r} \frac{(\frac{\nu_i}{n} - \frac{\mu_i}{m})^2}{\nu_i + \mu_i}.$$

Itt az átalakítások során kihasználtuk, hogy $\hat{p}_i = \frac{\nu_i + \mu_i}{n+m}$.

A χ^2-statisztika H_0 fennállása esetén aszimptotikusan (n, m „nagy”) χ^2-eloszlást követ $r - 1$ szabadsági fokkal, ui. a $2(r - 1)$ szabad cellaszámból levonjuk a becsült paraméterek számát: ez $r - 1$, ui. $\sum_{i=1}^{r} \hat{p}_i = 1$. Esetünkben a statisztika értéke 8.224, ami nagyobb, mint $\chi^2_{0.025}(2)$, így 0.025 szignifikanciával elutasítjuk H_0-t, azaz alapos okunk van feltételezni, hogy az étrendválasztás szignifikánsan befolyásolja a gyermekek egészségi állapotát. □

Megjegyezzük, hogy itt a vizsgálat PROSPEKTÍV volt, azaz előre tervezték a kísérletet. Elhatározták, hogy 80 ill. 70 gyermeket állítanak rá az A ill. B étrendekre, amivel a kontingenciatábla egyik peremét rögzítették. Ezért a homogenitásvizsgálatot az EGYIK MARGINÁLIS FIXen tartásával végzett vizsgálatnak is nevezik.

74. *Függetlenségvizsgálat* Az USA-ban megkérdeztek 500 embert politikai pártállásáról és arról, hogy támogatna-e egy adóreformot. A következőt kapták:

	Támogatja	Közönbös	Ellenzi	összes
Demokrata	138	83	64	285
Rupublikánus	64	67	84	215
összes	202	150	148	500

Van-e ósszefüggés a politikai pártállás és az adóreformhoz való hozzáállás közt?

Megoldás. Most függetlenségvizsgálatot végzünk, ahol

H_0 : a politikai pártállás és az adóreformhoz való hozzáállás függetlenek.

Ha ez igaz, akkor a képletgyűjtemény alapján gyártott χ^2-statisztika aszimptotikusan ($n = 500$ „nagy”) χ^2-eloszlást követ $rs - 1 - (r - 1) - (s - 1) = (r - 1)(s - 1)$ szabadsági fokkal, ahol $r = 2$ és $s = 3$ a sorok ill. oszlopok száma. Esetünkben a statisztika értéke 22.153, ami nagyobb, mint $\chi^2_{0.005}(2)$, így 0.005 szignifikanciával elutasítjuk H_0-t, azaz alapos okunk van feltételezni, hogy a politikai pártállás és az adóreformhoz való hozzáállás messze nem függetlenek. □

Megjegyezzük, hogy itt a vizsgálat RETROSPEKTÍV, azaz a kísérlet elvégzése után tudjuk csak kiszámolni a marginálisokat, melyek a véletlentől függenek. Ezért a függetlenségvizsgálatot az EGYIK MARGINÁLIS SEM FIX feltétellel is szokás jellemezni.

Ugyanakkor, ha a fenti 2×3-as táblázaton homogenitás- és függetlenségvizsgálatot is végrehajtunk, formálisan ugyanazt a χ^2 értéket kapjuk és a szabadsági fok is megegyezik, így döntésünk ugyanaz. Mégis, a két hipotézis különbözik. Itt homogenitásvizsgálatnak akkor lenne értelme, ha pl. a képviselőházban kérdeznénk meg a már ismert pártállású képviselőket az adóeformról.

A biológiai kísérletek nagyrészt retrospektívek: pl. ha azt szeretnék bizonyítani, hogy a dohányzás növeli a tüdőrák kockázatát (a tüdőrák kialakulása nem független a dohányzástól), akkor ez úgy történik, hogy visszakövetik tüdőrákosok és kontroll személyek életútját, és megnézik, hogy dohányoztak-e; nem pedig azt teszik, hogy kiválasztanak egy csoport dohányost és egy másik csoport nemdohányost, majd figyelik, hogy kialakul-e náluk tüdőrák.

75. Egy klinikán 460 személyt vizsgáltak hipertónia és túlsúlyosság szempontjából. Azt találták, hogy köztük 416 hipertóniás volt, aki túlsúlyos, és 5 személy sem túlsúlyos, sem hipertóniás nem volt. Ugyanakkor 16 nem hipertóniás, de túlsúlyos személyt találtak köztük, míg 23 hipertóniás nem volt túlsúlyos. 0.05 szignifikanciával döntse el, hogy a hipertónia és a túlsúlyosság függetlenek-e egymástól?

Megjegyezzük, hogy amennyiben $2{\times}2$-es kontingencatáblában a χ^2-statisztika kiszámolása a

$$\chi^2 = n\frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1.}\nu_{2.}\nu_{.1}\nu_{.2}}$$

képletre egyszerűsíthető, ahol

$$\frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1.}\nu_{2.}\nu_{.1}\nu_{.2}}$$

a két bináris változó (igen=1, nem-0) közti korreláció empirikus megfelelőjének az n-szerese. Ennek alapján külübözó mérőszámok (Pearson, Cramer) konstruálhatók a függetlenség mérésére és az ún. korrespondanciaanalízis szekvenciálisan vizsgálja ezeket a korrelációkat tetszőleges véges értékkészletű diszkrét véltozók esetén.

76. A 6. táblázat egy külföldi egészgégbiztossítási társaság adatait tartalmazza. 200 kétgyerekes család körében nézték, hogy az elmúlt periódusban hány térítési igényt jelentettek be. Az igények száma követhet-e Poisson eloszlást?

Igények száma	0	1	2	3	4	5	6	7	teljes
Gyakoriság	22	53	58	39	20	5	2	1	200

6. táblázat. Biztosítóhoz benyújtott igények száma gyakorisággal

Megoldás. Illeszkedésvizsgálatot végzünk, ahol

$$H_0: \text{ az igények száma Poisson eloszlású.}$$

Az alábbi táblázatban mellékeljük a számolásokat, melyekkel illeszkedésvizsgálatot hajtunk végre. Megjegyezzük, hogy egy Poisson eloszlású valószínűségi változó értékkészlete a nemnegatív egészek halmaza, azonban ebben a véges mintában az előforduló legnagyobb érték a 7. Ráadásul a 6 és 7 értékek gyakorisága olyan alacsony, hogy az utolsó három kategóriát össze kell vonnunk: ide a legalább 5 értékek tartoznak. Fontos, hogy teljes eseményrendszerrel dolgozzunk, így igaz csak, hogy a megfigyelt és várt cellagyakoriságok összege megegyezik ($n = 200$).

Igények száma	0	1	2	3	4	legalább 5	teljes
ν_i	22	53	58	39	20	8	200
np_i	27.0	54.2	54.2	36.0	18.0	10.6	200

Itt a p_i valószínűségeket Poisson eloszlás szerint számoltuk, melynek paraméterét a mintaátlaggal becsültük:

$$\hat{\lambda} = \frac{0 \times 22 + 1 \times 53 + \cdots + 7 \times 1}{200} = 2.05.$$

A próba-statisztika értéke $\chi^2 = 2.33$ és a szabadsági fok $df = 6 - 1 - 1 = 4$. Mivel a 2.05 kisebb, mint a $\chi^2_{0.5}(4)$ érték, csak 0.5-nél nagyobb szignifikanciával tudnánk elutasítani H_0-t, azaz ilyen nagy lenne a téves elutasítás valószínűsége. Így természetesen elfogadjuk H_0-t. □

77. 500 ötgyermekes családban vizsgálták a fúk számát. 20 családban nem volt fiú, 75-ben 1, 145-ben 2, 140-ben 3, 85-ben 4, és 35-ben 5 fiú volt. Döntse el, hogy ebben a populációban egy 5-gyerekes családban a fiúk száma binomiális eloszlát követ-e ($\alpha = 0.05$)?

78. 0.01 szignifikanciával vizsgálja meg, hogy az alábbi 100 elemű minta származhat-e

(a) 3 paraméterű Poisson eloszlásból? A mintában a 0,1,2,3,4 értékek fordultak elő a következő gyakoriságokkal:

$$12, 32, 25, 21, 10.$$

(Használja a Poisson- és χ^2-eloszlások táblázatát!)

(b) Egyáltalán származhat-e a minta Poisson eloszlásból?

79. Megkérdeztünk 1000 embert, hogy a kávét vagy a teát szeretik-e jobban. Egy háromfokozatú skálán választhattak: szereti, közömbös, nem szereti. Azt tapasztalták, hogy 300 ember válaszolt ugyanúgy a két kérdésre: 150-en egyiket sem szeretik, 100-an minkettőt szeretik, és 50-nek mindkettő közömbös. A kávét összesen 500-an szeretik, akik közül 200-an nem szeretik a teát. 50 válaszadó nem szereti a teát és közömbös számára a kávé. A teát szeretők kávéra vonatkozó válaszai egyenlő arányban oszlanak meg a három lehetőség közt. Döntse el ennek alapján, hogy a kávé és a tea szeretete független-e ($\alpha = 0.05$)?

80. A Raymond Weil cég új karórával akar megjelenni a piacon és szeretné tudni a fogyasztók preferenciáit az óraszíj színét illetően. Négy szín került szóba és kérdés, hogy a négy színt egyformán értékelik-e a vásárlók. 80 potenciális vásárlót megkérdeztek és az alábbi eredményeket kapták:

sárgásbarna	barna	gesztenyebarna	fekete	Teljes
12	40	8	20	80

Vizsgáljuk meg azt a null-hipotézist, hogy az óraszíj preferenciák azonosak! Az alternatíva ennek a tagadása (nem szoktuk külön feltüntetni), csak azt, hogy

$$H_0 : p_1 = p_2 = p_3 = p_4 = 0.25.$$

Megoldás. A számolt próba-statisztika:

$$\chi^2 = \sum_{i=1}^{4} \frac{(\nu_i - 20)^2}{20} = \frac{(-8)^2 + 20^2 + (-12)^2 + 0^2}{20} = 30.4.$$

A kritikus tartomány: $\mathcal{X}_k = \{\chi^2 \geq \chi^2_{0.01}(3) = 11.3\}$. Tehát elutasítjuk H_0-t 0.01 szignifikanciával. □

81. Egy áruházlánc elemzője tesztelni szeretné, hogy az egyes fogyasztók vásárlásra fordított összege normális eloszlású-e. Ez fontos neki, mert - ha a feltételezése beigazolódik - varianciaanalízist végezhet arra vonatkozóan, hogy a fogyasztók egyforma mértékben költenek a lánc különböző áruházaiban. 100 vásárló megkérdezése után az átlagra \$125 és a korrigált tapasztalati szórásra \$40 adódott.

(0 - 84.99)	(85.00 - 107.39)	(107.4 - 124.99)	(125 - 142.59)	(142.6 - 164.99)	(165,∞)
14	20	16	19	16	15

7. táblázat. Vásárlási összegek és vásárlók száma kategóriák szerint

Megoldás. Illeszkedésvizsgálatot végzünk.

$$H_0 : \text{Az elköltött összeg normális eloszlású.}$$

Ha H_0 fennáll, akkor a 7. táblázatbeli adatokat standardizálva standard normális eloszlásból származó adatokat kapunk. Ezért a megfigyelt és várt cellagyakoriságok a standard normális eloszlás szerinti kategóriákban a következők (itt p_i-k a standard normális eloszlásfüggvény szerinti értékek és táblázatból kikereshetők).

Kategóriák	(-3.125,-1)	(-1,-0.44)	(-0.44,0)	(0,0.44)	(0.44,1)	(1,∞)
ν_i	14	20	16	19	16	15
np_i	15.87	17.13	17.00	17.00	17.13	15.87

A számolt próba-statisztika:

$$\chi^2 = \frac{(14-15.87)^2}{15.87} + \frac{(20-17.13)^2}{17.13} + \frac{(16-17.00)^2}{17.00} +$$

$$+ \frac{(19-17.00)^2}{17.00} + \frac{(16-17.13)^2}{17.13} + \frac{(15-15.87)^2}{15.87} = 1.12.$$

A számolt statisztika nem esik az elutasítási tartományba semmilyen α mellett a χ^2 táblázatban (a szabadsági fok 6-1-2=3, hiszen a kategóriák száma 6 és most két paramétert - a várható értéket és a szórást - becsültünk). Tehát elfogadjuk, hogy az adatok normális eloszlásúak. □

82. A cégek nyereségeinek és veszteségeinek tanulványozása végett 100 céget megvizsgálunk. Azt is figyelembe vesszük, hogy az adott cég a szolgáltatói szférába esik-e. Az adatok az 8. kontingenciatáblázatban találhatók. Kérdés, hogy a "veszteségesnek lenni" és a "szolgáltató iparba tartozni" tulajdonságok függetlenek-e?

	szolgáltató	nem szolgáltató	összesen
nyereséges	42	18	60
veszteséges	6	34	40
összesen	48	52	100

8. táblázat

Megoldás. Függetlenságvizsgálatot végzünk. A számolt próba-statisztika:

$$\chi^2 = \frac{(42-28.8)^2}{28.8} + \frac{(18-31.2)^2}{31.2} + \frac{(6-19.2)^2}{19.2} + \frac{(34-20.8)^2}{20.8} = 29.09,$$

ami nagyobb, mint a $\chi^2_\alpha(1)$ kritikus érték (a szabadsági fok most $(2-1)\times(2-1)=1$) minden szokásos szignifikanciával (még $\alpha = 0.005$ mellett is). Így elutasítjuk a függetlenséget. □

83. Mendel egyik nevezetes kísérletében 556 kerek sárga magvú borsót keresztezett ráncos zöld magvúval. Mendel elmélete szerint négyféle borsó keletkezhetett: kerek sárga, kerek zöld, ráncos sárga és ráncos zöld magvú, továbbá ezek arányai az elmélet szerint kb.: 9 : 3 : 3 : 1.

Megoldás. Illeszkedésvizsgálatot végzünk.

A számolt próba-statisztika a táblázat alapján:

$$\chi^2 = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(102-104.25)^2}{104.25} + \frac{(31-34.75)^2}{34.75} = 0.618.$$

típus	megfigyelt esetszám (ν_i)	jósolt arány (p_i)	várható esetszám (np_i)
kerek sárga	315	9/16	312.75
kerek zöld	108	3/16	104.25
ráncos sárga	102	3/16	104.25
ráncos zöld	31	1/16	34.75

9. táblázat. Mendel elméletének jóslatai és a megfelelő várt értékek

Ezt kell összevetnünk a $\chi^2(3)$ eloszlás felső kvantiliseivel. Mivel $\chi^2_{0.1}(3) = 6.251, \ldots, \chi^2_{0.75}(3) = 1.20$ felette vannak 0.618-nak, és $\chi^2_{0.9}(3) = 0.584$ az első a táblázatban, ami alatta van, csak a drasztikusan nagy $\alpha = 0.9$ mellett tudnánk elutasítani H_0-t. Az adatok meglepően jó egyezést mutatnak a jóslattal. Valójában azt várnánk, hogy a jegyzett adatok az esetek 90%-ában nagyobb eltérést mutatnak. Az általános vélemény ezért az, hogy Mendel elmélete helyes, de valaki "megmasszírozta" az adatokat az elmélet kedvéért. Fisher szerint ez a valaki Mendel kertésze lehetett. □

11. feladatsor

Előjel-próba és Wilcoxon-próba

84. *Kétmintás előjel-próba.*

Szeretnénk tudni, hogy egy napvédő krém hatása megváltozik-e egy új alkotórśz hozzáadásával. Ezért 7 független személy hátának egyik felét az agyik, másik felét pedig a másik krémmel kenték be. A leégés fokát egy folytonos skálán mérve a következőt kapták.

Személy sorszáma	1	2	3	4	5	6	7
Régi készítmény (X_i)	42	51	31	61	44	55	48
Új készítmény (Y_i)	38	53	36	52	33	49	36
Különbség ($D_i = X_i - Y_i$)	4	-2	-5	9	11	6	12

Ez egy páros mintás teszt, ahol

$$H_0 : \; X - Y \text{ mediánja} = 0,$$

azaz nincs különbség a két fényvédő krém hatása közt.

H_0 fennállásakor $\mathbb{P}(X \geq Y) = \mathbb{P}(D \geq 0) = \frac{1}{2}$, azaz a $\nu_7 = |\{i : D_i \geq 0\}$ valószínűségi változó binomiális eloszlást követ 7 és $\frac{1}{2}$ paraméterekkel. Egyoldali alternatívával szemben, ha

$$H_1 : \; X - Y \text{ mediánja} > 0,$$

a legszűkebb kritikus tartomány, amibe a $\nu_7 = 5$ beletartozik, $\mathcal{X}_k = \{\nu_7 \geq 5\}$, melynek valószínűsége H_0 fennállása esetén a $7, \frac{1}{2}$ paraméterű binomiális eloszlás táblázat szerint 0.2266. Ekkora p-érték nem elég bizonyíték az elutasításra. Még kevésbé tudjuk elutasítani H_0-t a kétoldali alternatívával szemben:

$$H_1 : \; X - Y \text{ mediánja} \neq 0,$$

ahol a legszűkebb kétoldali kritikus tartomány, amibe a $\nu_7 = 5$ beletartozik,

$$\mathcal{X}_k = \{\nu_7 \leq 2 \quad \text{vagy} \quad \nu_7 \geq 5\}.$$

Ennek valószínűsége H_0 fennállása esetén a $7, \frac{1}{2}$ paraméterű binomiális eloszlás táblázat szerint 2×0.2266. Ekkora p-érték mellett nem tudunk elutasítani.

Vegyük észre, hogy itt a párba állított mérések konkrét értéke nem, csak különbségük előjele számított.

Megjegyezzük, hogy más lenne a helyzet, hogy ha pl. 7 személy hátát az egyik, másik 7 vagy akár 10 személy hátát pedig a mśik krémmel kenik be, de egyszerre napoznak. Ebben az esetben két független mintát hasonlítunk össze és a Wilcoxon-próbát használjuk.

Azt is megjegyezzük, hogy amennyiben n „nagy”, a binomiális eloszlást normálissal közelíthetjük, és a H_0 melletti $\frac{n}{2}$ várható értékkel és $\sqrt{\frac{n}{4}}$ szórással standardizálva z-statisztika használható.

85. Páciensek viselkedését vizsgálták teliholdkor, azzal a feltevéssel, hogy ilyenkor agresszívabbakká válnak. Azt találták, hogy 15 páciensből 14-nél valóban megnőtt az agresszivitás teliholdkor. Mennyiben támasztják ezek az adatok alá a feltételezést?

Megoldás. Ha nullhipotézisünk az, hogy a telihold megléte nem növeli meg az agresszívabb páciensek átlagos számát, akkor az agresszívek ν_{15} száma binomiális eloszlású 15 és $\frac{1}{2}$ paraméterrel. Emellett a legszűkebb kritikus tartomány, amibe 14 beletartozik: $\mathcal{X}_k = \{\nu_{15} \geq 14\}$, melynek valószínűsége H_0 fennállásakor a p-érték:

$$p = \mathbb{P}(\nu_{15} \geq 14) = \binom{15}{14}\frac{1}{2^{15}} + \frac{1}{2^{15}} = 0.00049.$$

Ez nagyon kicsiny, így elvetjük a nullhipotézist (hisz ilyen kicsi a valószínűsége, hogy alaptalanul vetnénk el). □

86. Egy kutatási projekt keretében a magzatiból az újszülötti korba történő átmenet során vizsgálták keringési rendszert. A projekt részeként megmérték 19 újszülöttnek a légzésfunkcióját 15 napos kora előtt és 25 napos kora után is. Az adatok a 10. táblázatban találhatók.

sorszám	1	2	3	4	5	6	7	8	9	10
15 nap előtt	48	38	43	48	35	52	48	27	26	62
25 nap után	47	40	46	42	42	44	36	40	40	46

sorszám	11	12	13	14	15	16	17	18	19
15 nap előtt	68	67	80	88	84	75	67	45	70
25 nap után	45	31	42	48	45	38	44	45	35

10. táblázat. Újszülöttek légzésfunkciója két időszakban

(a) Végezzen χ^2-próbát a két változó különbségének normalitására vonatkozóan!

(b) Végezzen Mann-Whitney-féle próbát a különbség mediánjára $\alpha = 0.05$ szignifikanciával!

87. *Wilcoxon-féle rangösszeg próba.*

Legyen $X_1, \ldots, X_n$ fae. és $Y_1, \ldots, Y_m$ fae. minták, melyek egymást közt függetlenek, és eloszlásuk abszolút folytonos. Ebben a kétmintás próbában azt szeretnénk vizsgálni, hogy az X és Y háttérváltozók azonos eloszlásúak-e. Ez a próba eltolásparaméteres eloszláscsaládoknál használható, ha tudjuk valahonnan, hogy $F_X(x) = F_Y(x + \theta)$, $\forall x \in \mathbb{R}$, ahol θ eltolási paraméter, azonban nem ennek konkrét értékét, hanem csak előjelét teszteljük. Ezért a

$$H_0 : F_X = F_Y \quad (X \text{ és } Y \quad \text{azonos eloszlású}) \tag{4}$$

null-hipotézis megfelel $\theta = 0$-nak és annak, hogy X és Y mediánja megegyezik, azaz $\mathbb{P}(X > Y) = \frac{1}{2}$.

Ha az egyoldali alternatíva

$$H_1 : \theta > 0 \tag{5}$$

alakú, az azt jelenti, hogy X eloszlása pozitív eltoltja Y-énak, azaz X mediánja nagyobb, mint Y-é: $\mathbb{P}(X > Y) > \frac{1}{2}$. Ha az egyoldali alternatíva

$$H_1 : \theta < 0$$

alakú, az azt jelenti, hogy X eloszlása negatív eltoltja Y-énak, azaz X mediánja kisebb, mint Y-é: $\mathbb{P}(X > Y) < \frac{1}{2}$. A kétoldali alternatíva

$$H_1 : \theta \neq 0$$

lenne.

A konkrét feladatban $x_1 = 31.8$ és $x_2 = 39.1$ két új hibrid növény virágjának a kerülete, míg a régi fajtájúakból három kerület $y_1 = 21.3$, $y_2 = 27.6$, $y_3 = 35.5$ volt. A kérdés az, hogy a két típus virágjának kerülete azonos-e. A biometriai szóhasználattal élve, a kétféle kezelést (A és B) szeretnénk összehasonlítani. Itt is csak a nagyságrend számít. Vizsgáljuk meg az (4) null-hipotézist a (5) alternatívával szemben (ez fejezi ki azt, hogy az új hibridek nagyobb virágot növesztenek).

Tekintsük az egyesített rendezett mintát (előtte külön-külön akár rendezhetjük is őket). Legyen

$$\begin{aligned} r_i &:= \text{rang}(X_i) = X_i \quad \text{sorszáma} \quad \text{az egyesített rendezett mintában,} \\ s_j &:= \text{rang}(Y_j) = Y_j \quad \text{sorszáma} \quad \text{az egyesített rendezett mintában.} \end{aligned}$$

Nyilván $r_1 < r_2 < \cdots < r_n$ és $s_1 < s_2 < \cdots < s_m$ különböző egészek, amik együtt kiteszik az $\{1, \ldots, n+m\}$ halmazt (az egyenlőség valószínűsége abszolút folytonos eloszlások esetén 0, de tört rangszámokkal ezt az esetet is kezelni tdjuk).

H_0 fennállása esetén az összes sorrend egyformán valószínű, és ez a valószínűség $\frac{1}{\binom{n+m}{n}}$.

A próbastatisztika $W_A = \sum_{i=1}^{n} r_i$. Ennek párja a $W_B = \sum_{j=1}^{n} s_j$ statisztika, azonban a fent mondottak miatt köztük a $W_A + W_B = \sum_{i=1}^{n+m} = \frac{1}{2}(n+m)(n+m+1)$ összefüggés áll fenn, így elég csak W_A-t tekinteni (amelyikhez a kisebb mintaelemszám tartozik). Esetünkben $r_1 = 3$, $r_2 = 5$, ezért $W_A = 8$.

Mivel az A-kezelések rangjai az összes $\binom{5}{2} = 10$ lehetőségen egyenletes eloszlásúak, a következőt kapjuk:

A-rangok	W_A	valószínűség
1,2	3	0.1
1,3	4	0.1
1,4	5	0.1
1,5	6	0.1
2,3	5	0.1
2,4	6	0.1
2,5	7	0.1
3,4	7	0.1
3,5	8	0.1
4,5	9	0.1

A

$$H_1 : \mathbb{P}(X > Y) > \frac{1}{2}$$

alternatívával szemben a legszűkebb egyoldali kritikus tartomány, mely a $W_A = 8$ evidenciát tartalmazza, az utolsó két sor, melynek valószínűsége H_0 fennállása esetén 0.2. Ilyen p-érték mellett nehéz elutasítani H_0-t. A

$$H_1 : \mathbb{P}(X > Y) \neq \frac{1}{2}$$

kétoldali alternatívával szemben a legszűkebb kétoldali kritikus tartomány, mely a $W_A = 8$ evidenciát tartalmazza, az első kettő és az utolsó kettő sor, melyek valószínűségének összege H_0 fennállása esetén 0.4. Ilyen p-érték mellett még kevésbé tudjuk elutasítani H_0-t a kétoldali alternatívával szemben. Azaz ez az 5 mérés nem bizonyítja azt, hogy az új hibridek nagyobbak vagy különbözőek, mint a régiek.

Diszkrét egyenletes eloszláson alapuló táblázatok találhatók a Wilcoxon-próba kritikus értékeire „kis” n és m esetén. Amennyiben n és m „nagy”, a Mann–Whitney próbát használhatjuk, mely W_A aszimptotikus normalitásán alapul, így standardizálás után z-statisztika használható a szokásos egy- és kétoldali kritikus tartományokkal (l. tankönyv, 187-188. old).

88. Hasonló fizikai állapotú felnőttek napi fajlagos kalóriabevitelét (kcal/testsúly) jegyezték fel 23 bulimiás és 15 egészséges esetben (11. táblázat).

bulimiások			egészségesek	
15.9	19.6	25.6	20.7	33.2
16	21.5	28	22.4	33.7
16.5	21.6	28.7	23.1	36.6
17	22.9	29.2	23.8	37.1
17.6	23.6	30.9	24.5	37.4
18.1	24.1		25.3	40.8
18.4	24.5		25.7	
18.9	25.1		30.6	
18.9	25.2		30.6	

11. táblázat

Feltéve, hogy a bulimiás betegek és az egészségesek kalóriabevitelének eloszlásai egymás eltoltjai, végezzen Wilcoxon-féle rangpróbát a

$$H_0 : \mathbb{P}(X > Y) = \frac{1}{2} \quad \text{versus} \quad \mathbb{P}(X > Y) < \frac{1}{2}$$

alternatívára, ahol X : egy véletlenszerű bulimiás (B) kalóriabevitele, Y : egy véletlenszerű egészséges (E) kalóriabevitele.

Megoldás. Összefésülve az adatokat és beírva a rangokat (az egyenlő megfigyelések közt megosztjuk a rangot) a 12. táblázatot kapjuk:

bevitel	rang
15.9	1B
16	2B
16.5	3B
17	4B
17.6	5B
18.1	6B
18.4	7B
18.9	8.5B
18.9	8.5B
19.6	10B
20.7	11E
21.5	12B
21.6	13B
22.4	14E
22.9	15B
23.1	16E
23.6	17B
23.8	18E
24.1	19B
24.5	20.5B
24.5	20.5E
25.1	22B
25.2	23B
25.3	24E
25.6	25B
25.7	26E
28	27B
28.7	28B
29.2	29B
30.6	30.5E
30.6	30.5E
30.9	32B
33.2	33E
33.7	34E
36.6	35E
37.1	36E
37.4	37E
40.8	38E

12. táblázat. rangszámok

Ebbből az X-nek megfelelő rangösszeg: $W_A = 337.5$ (mivel aszimptotikus normalitással dolgozunk, nem baj, hogy a bulimiásokhoz tartozik a nagyobb mintaelemszám, és ekkora n, m elég nagynak számít). A számolt Mann–Whitney -statisztika:

$$z = \frac{337.5 - \frac{23\cdot(23+15+1)}{2}}{\sqrt{\frac{23\cdot 15\cdot(23+15+1)}{12}}} = -3.315,$$

mely annak az $\mathcal{X}_k = \{z \leq -z_\alpha\}$ kritikus tartománynak van a határán, melyre $\alpha = 0.0005$. Ezért elvetjük H_0-t, s döntésünk: $\mathbb{P}(X > Y) < \frac{1}{2}$, azaz a bulimiások kalóriabevitele szignifikánsan kisebb, mint az egészségeseké. □

89. Egy kutató azt vizsgálja, hogy egy adott városra hogyan hat az ENSO jelenség (az El Nino South Oscillation a déli féltekén hat és az északi félteke Golf-áramlatához hasonlít). Ezért a kutató megfigyelt 5 pozitív és 5 negatív ENSO időszakot, és feljegyezte az átlagos téli hőmérsékleteket (lásd 13. táblázat).

Alátámasztják-e ezek az adatok, hogy a pozitív ENSO időszakokhoz magasabb átlaghőmérséklet tartozik?

poz.	neg.
0.23	-1.2
0.18	1.2
0.86	-0.018
0.72	-0.25
1.4	-0.36

13. táblázat. Átlagos téli hőmérsékletek pozitív és negatív ENSO időszakokban

sorsz.	mat.	műv.
1	22	53
2	37	68
3	36	42
4	38	49
5	42	51
6	58	65
7	58	51

sorsz.	mat.	műv.
8	60	71
9	62	55
10	65	74
11	66	68
12	56	64
13	66	67
14	67	73
15	62	65

14. táblázat. Matematikai és művészeti teszt eredmények

90. Megmérték 15 véletlenszerűen kiválasztott diák matematikai és művészeti képességeit. A kapott pontszámokat a 14. táblázat tartalmazza.

A Spearman-féle rangkorreláció segítségével döntsön a kapcsolat szorosságáról!

Megoldás. A megfelelő rangokat a 15. táblázat tartalmazza (1-15 közt, tört rangok is vannak):

mat. rang	műv. rang
1	5
3	11.5
2	1
4	2
5	3.5
7.5	8.5
7.5	3.5

mat. rang	műv. rang
9	13
10.5	6
12	15
13	11.5
6	7
14	10
15	14
10.5	8.5

15. táblázat. Matematikai és művészeti teszt rangszámai

A Spearman-féle rangkorreláció:

$$r_{sp} = \frac{\sum_{i=1}^{n} \left(r_i - \frac{n+1}{2}\right)\left(s_i - \frac{n+1}{2}\right)}{n(n^2-1)/12} = 0.697,$$

ami elég magas, mutatja a pozitív kapcsolatot. A tankönyv végén található táblázat alapján ez nagyobb, mint az $n = 15$ sorban található

$\alpha = 0.01$-hez tartozó kritikus érték, ami 0.6500. Így a matematikai és művészeti képességek közti kapcsolat szignifikáns $\alpha = 0.01$ szinnifikanciával. A rangkorreláció pozitív előjele azt is mutatja, hogy magasabb matematikai képességekhez magasabb művészeti képességek társulnak, és megfordítva. □

91. Megmérték tíz személy IQ-ját és megkérdezték tőlük, hogy hány órát ülnek a tv előtt hetente. Az alábbi eredményeket kapták (16. táblázat):

IQ	óra
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

16. táblázat. IQ és tv-nézés időtartama

A Spearman-féle rangkorreláció segítségével döntsön a kapcsolat szorosságáról!

Megoldás. Mivel egyik változó értékei között sincsenek egyezések, ezért a valamivel egyszerűbb

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

képletet is használhatjuk, ahol $d_i = r_i - s_i$. A következő (17.) táblázat tartalmazza a változók értékeinek rangját, azok különbségét, valamint azok négyzetét is:

IQ	óra	IQ rang	óra rang	diff	diff^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

17. táblázat. IQ és tv-nézés időtartama rangokkal

Kapjuk, hogy $r_{sp} = -0.18$, ami nagyon alacsony, de negatív korrelációra utal.

□

92. Az alábbi pontszámok 10 véletlenszerűen kiválasztott főiskolás hallgató kézügyességének és agresszivitásának szintjét mérik:

Kézügyesség	23	29	45	36	49	41	30	15	42	38
Agresszivitás	45	48	16	28	38	21	36	18	31	37

Ennek alapján számolja ki a kézügyesség és agresszivitás Spearman-féle rangkorrelációs együtthatóját és vonjon le következtetéseket!

12. feladatsor

Lineáris regresszió

93. Egy allergiaellenes gyógyszer egy kutatási fázisában azt vizsgálták, hogyan függ a szer beadott mennyiségétől a tünetmentes időszak hossza. Tíz pácienst kértek meg, hogy a szer bevétele után jelezzék, ha a tünetek kezdenek visszatérni. Az adatokat a 18. táblázat tartalmazza. Határozzuk meg a regressziós egyenest és becsüljük meg a szórásnégyezetet!

Adag (mg)	Tünetmentes időszak (óra)
x	y
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22

18. táblázat. Allergiaellenes gyógyszer adagja és a tünetek enyhülése 10 páciens esetén

Megoldás. Az

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n}$$
$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

összefüggések miatt elegendő a $\sum_i x_i, \sum_i y_i, \sum_i x_i^2, \sum_i y_i^2, \sum_i x_i y_i$ mennyiségeket kiszámolni. A 19. táblázat alapján: $S_{xx} = 389 - \frac{59^2}{10} = 40.9$, $S_{yy} = 2651 - \frac{151^2}{10} = 370.9$, $S_{xy} = 1003 - \frac{59 \cdot 51}{10} = 112.1$.
Az $Y = ax + b + \varepsilon$ lineáris modellben, ahol $\mathrm{Var}(\varepsilon) = \sigma^2$, az

$$SST = SSR + SSE$$

szórásfelbontást alkalmazzuk, ahol $SST = S_{yy}$, $SSR = \frac{S_{xy}^2}{S_{xx}}$, $SSE =$

	x_i	y_i	x_i^2	y_i^2	x_iy_i
	3	9	9	81	27
	3	5	9	25	15
	4	12	16	144	48
	5	9	25	81	45
	6	14	36	194	84
	6	16	36	256	96
	7	22	49	484	154
	8	18	64	324	144
	8	24	64	576	192
	9	22	81	484	198
Σ_i	59	151	389	2651	1003

19. táblázat

$S_{yy} - \frac{S_{xy}^2}{S_{xx}} = \sum_{i=1}^n (Y_i - \hat{a}x_i - \hat{b})^2$ és a paraméterek torzítatlan becslései:

$$\hat{a} = \frac{S_{xy}}{S_{xx}}, \quad \hat{b} = \bar{Y} - \hat{a}\bar{x}, \quad \hat{\sigma}^2 = \frac{SSE}{n-2} =: s^2.$$

A feladatban $SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 370.9 - \frac{112.1^2}{40.9} = 63.6528$, s innen $\hat{a} = \frac{S_{xy}}{S_{xx}} = \frac{112.1}{40ds.9} = 2.74$, $\hat{b} = \bar{y} - \hat{a}\bar{x} = 15.1 - 2.74 \cdot 5.9 = -1.07$, $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{63.6528}{8} = 7.9566$. □

94. Egy hallgató adatokat gyűjtött a futballmeccsek alatt elfogyasztott nagy kerek pizzák y számáról és azt is feljegyezte, hogy ilyenkor hány hallgató volt jelen. A következőket kapta (20. táblázat).

nézők száma	pizzák száma
2	1
5	6
6	10
3	3
4	5

20. táblázat

(a) Ábrázolja az adatokat az x–y koordináta rendszerben!

(b) Számolja ki a és b legkisebb négyzetes becslését!

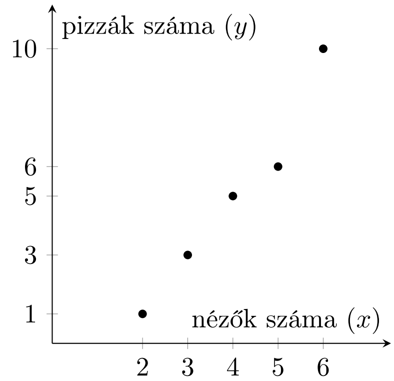
(c) Rajzolja be az egyenest a (**94**.a) pontban kapott ábrába!

(d) Ellenőrizze, hogy a maradéktagok összege 0!

(e) Mennyi lesz σ^2 becsült értéke?

Megoldás. (a) L. ay 5. ábrát.

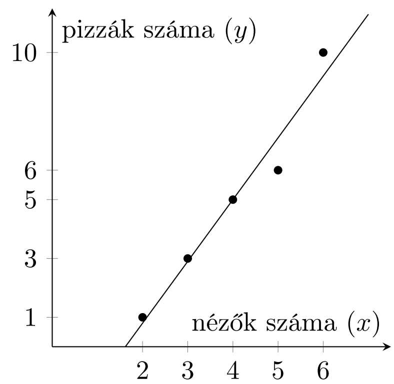
(b) A 21. táblázat alapján: $S_{xx} = 90 - 80 = 10$, $S_{yy} = 171 - 125 = 46$, $S_{xy} = 121 - 100 = 21$, ezért $\hat{a} = 2.1$, $\hat{b} = -3.4$.



5. ábra. Pizza adatok plotja

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	$y_i - \hat{y}_i$
	2	1	4	1	2	0.2
	5	6	25	36	30	-1.1
	6	10	36	100	60	0.8
	3	3	9	9	9	0.1
	4	5	16	25	20	0
Σ_i	20	25	90	171	121	0

21. táblázat



6. ábra. Regressziós egyenes a pizza adatokhoz

(c) L. a 6. ábrát.

(d) A 21. táblázatból látszik, hogy az utolsó oszlopbeli összeg 0.

(e) $SSE = 46 - 21^2/10 = 1.9$, $\hat{\sigma}^2 = \frac{\text{SSE}}{n-2} = \frac{1.9}{3} = 0.633$.

□

95. Egy ország fejlettségének egyik mérőszáma az ú.n. Emberi Fejlettségi Index (EFI) vagy angolul a Human Development Index (HDI). Várható élettartam, írni-olvasni tudás, iskolázottság, az egy főre eső bruttó hazai termék számértékeiből kapható a fenti EFI, ami egy 0 és 1 közötti szám, ahol 1 jelenti a legmagasabb fejlettséget. Az ENSZ Fejlesztési programja 177 ország EFI-jét közli. Véletlenszerűen kiválasztottunk ezek közül 15-öt (az első 25-öt nem számítva), az ezekhez tartozó EFI-t az 22. táblázat tartalmazza. Az x prediktor változó legyen a száz főre eső internet használók száma.

Ország	Internet/100	EFI
Bahrain	21.3	0.866
Lengyelország	26.2	0.870
Uruguay	14.3	0.852
Bulgária	20.6	0.824
Brazília	19.5	0.800
Ukrajna	9.7	0.788
Dominikai Köztársaság	16.9	0.799
Moldovai Köztársaság	9.6	0.708
India	5.5	0.619
Madagaszkár	0.5	0.533
Nepát	0.4	0,534
Tanzánia	0.9	0.467
Uganda	1.7	0.505
Zambia	2	0.434
Etiópia	0.2	0.406

22. táblázat

A fenti adatokból adódnak a következő értékek: $\bar{x} = 9.953$, $\bar{y} = 0.6670$, $S_{xx} = 1173.46$, $S_{yy} = 0.41772$, $S_{xy} = 20.471$.

(a) Határozza meg a lineáris kapcsolat szorosságát!

(b) Adjon 95%-os konfidenciaintervallumot a-ra!

(c) Döntsön 0.05 szignifikanciával a

$$H_0 : a = 0 \quad \text{versus} \quad H_1 : a \neq 0$$

alternatíváról!

(d) Az olyan országok körében, ahol az internet használóinak száma átlagosan $x^* = 22$, adjon 95%-os konfidenciaintervallumot az EFI-re!

(e) Feltételezve, hogy egy ország internet használóinak száma $x^* = 22$, adjon 95%-os konfidenciaintervallumot az EFI-re!

Megoldás. $\hat{a} = 0.017$, $hatb = 0.493$, $SSE = 0.061$.

(a) $R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.925$.

(b) A konfidenciaintervallum szerkesztését a következő eloszlások teszik lehetővé. Tudjuk, hogy

$$\mathbb{E}(\hat{a}) = a, \quad \mathrm{Var}(\hat{a}) = \frac{\sigma^2}{S_{xx}}$$

és

$$\mathbb{E}(\hat{b}) = b, \quad \mathrm{Var}(\hat{b}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right).$$

Mivel

$$\frac{SSE}{\sigma^2} = \frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$$

és (normális alapeloszlás esetén) független $\hat{a}$-tól és $\hat{b}$-tól, ezért a következő Student-eloszlású statisztikák konstruálhatók:

$$\frac{\frac{\hat{a}-a}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{\sqrt{\frac{(n-2)s^2}{\sigma^2}/(n-2)}} = \frac{(\hat{a}-a)\sqrt{S_{xx}}}{s} \sim t(n-2)$$

és

$$\frac{\frac{\hat{b}-b}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}}{\sqrt{\frac{(n-2)s^2}{\sigma^2}/(n-2)}} = \frac{\hat{b}-b}{s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} \sim t(n-2).$$

Így $1-\alpha$ szintő konfidenciaintervallum a-ra:

$$\left(\hat{a} - t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}, \hat{a} + t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}\right) = (0.013, 0.022).$$

(c) Mivel a 0 nincs az **95**.b feladat megoldásában kapott konfidenciaintervalumban, ezért elvetjük H_0 t $\alpha = 0.05$ szignifikanciával.

(d) Az $1-\alpha$ szintű konfidenciaintervallum a jósolt $ax^* + b$ válaszra a független változó átlagos x^* értéke esetén:

$$\hat{a}x^* + \hat{b} \pm t_{\alpha/2}s\sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}},$$

tehát az adatokkal: $(0.716, 1.038)$.

(e) Az $1-\alpha$ szintű konfidenciaintervallum a jósolt $ax^* + b$ válaszra a független változó egyetlen x^* értéke esetén:

$$\hat{a}x^* + \hat{b} \pm t_{\alpha/2}s\sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}},$$

tehát az adatokkal:$(0.8127, 0.9415)$.

□

96. Egy mérnök azt találta, hogy ha a feltölthető elemekhez a gyártás folyamán egy bizonyos anyagot hozzáad, akkor az megnöveli azok élettartamát. A mérnök megmérte, hogy mennyi adalékanyag mennyivel növeli meg az elemek élettartamát, ha azokat laptopban használják (23. táblázat).

(a) Adja meg a a, b, σ^2 legkisebb négyzetes becslését!

Adag	Élettartam (óra)
x	y
0	1.9
1	2.0
2	2.5
3	2.6
4	3

23. táblázat. Elemek élettartama

(b) Tesztelje a $H_0: \ a = 1$ versus $H_1: \ a \neq 1$ alternatívát 0.05 szignifikanciával!

(c) Adjon előrejelzést y-ra $x^* = 3.5$ mellett, és konstruáljon rá 95%-os konfidenciaintervallumot!

97. Hét eladott lakás esetében feljegyezték, hogy mennyi volt az előre becsült ár és a tényleges eladási ár (1000 USD-ban), l. a 24. táblázatot.

Becsült ár	Eladási ár
x	y
283.5	288.0
290	291.2
270.5	276.2
300.8	307.0
310.2	311.0
294.6	299.0
320.0	318.0

24. táblázat. Lakáspiac

(a) Ábrázolja a pontokat az x–y koordinátarendszerben!

(b) Határozza meg az regressziós egyenes egyenletét és rajzolja be a fenti ábrába!

(c) Adjon 95%-os konfidenciaintervallumot az egyenes meredekségére!

(d) Határozza meg a kapcsolat szorosságát!

(e) Adjon előrejelzést és konfidenciaintervallumot az eladási árra egy olyan lakásnál, ahol az előzetes értékbecslés $x^* = 290$ volt!

98. Hogyan tudná visszavezetni lineáris regresszióra a következő görbeillesztési feladatokat az (x_i, y_i), $i = 1, \ldots, n$ adatok alapján?

(a) $y = bx^a$

(b) $y = 2^{ax+b}$

(c) $y = \frac{1}{ax+b}$

13. feladatsor

Egyszempontos varianciaanalízis

99. Hogy egy kompakt lemez minőségét javítsák, négy különböző bevonat (A, B, C, D) hatását vizsgálják a lejátszás minőségére. A következő adatokat kapták:

$$\begin{aligned} A: &\quad 10, 15, 8, 12, 15 \\ B: &\quad 14, 18, 21, 15 \\ C: &\quad 17, 16, 14, 15, 17, 15, 18 \\ D: &\quad 12, 15, 17, 15, 16, 15 \end{aligned}$$

Kérdés: szignifikáns-e a különbség a négyféle kezelés közt a lejátszás minőségének tekintetében $(\alpha = 0.05)$?

Megoldás. Az egyszempontos ANOVA táblázat jelöléseivel: $k = 4$, $n_1 = 5$, $n_2 = 4$, $n_3 = 7$, $n_4 = 6$, $n = 22$; $\bar{x}_{1.} = 12$, $\bar{x}_{2.} = 17$, $\bar{x}_{3.} = 16$, $\bar{x}_{4.} = 15$ és $\bar{x}_{..} = 15$. Továbbá $Q_a = 68$, $Q_e = 94$. Ezért a próbastatisztika:

$$F = \frac{68/3}{94/18} = 4.34.$$

Mivel $4.34 > 3.16 = F_{0.05}(3, 18)$, ezért 0.05 szignifikanciával elutasítjuk H_0-t, azaz szignifikáns a különbség a négyféle kezelés közt. □

100. *Sandgrund et al., Americal Journal of Mental Deficiency 79(3) (1974), 327-330* cikkében azt vizsgálta, hogy a gyerekkori bántalmazás ill. elhanyagoltság befolyásolja-e a gyerekek IQ-ját. Ebből a célból 3 csoportot tekintettek: az A ill. B csoportba olyan szociális gondozásban részesülő családok gyerekei tartoztak, akiket bizonyíthatóan bántalmaztak ill. elhanyagoltak. A C kontroll csoportba normál családi kórúlmények között élő gyerekek tartoztak, akik nem részesültek szociális gondozásban. Az IQ-kra a következő adatokat kapták:

$$\begin{aligned} A: &\quad n_1 = 32, \quad \bar{x}_{1.} = 81.06, \quad s_1 = 17.05 \\ B: &\quad n_2 = 16, \quad \bar{x}_{2.} = 78.56, \quad s_2 = 15.43 \\ C: &\quad n_3 = 16, \quad \bar{x}_{3.} = 87.81, \quad s_3 = 14.36, \end{aligned}$$

ahol

$$s_i = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}{n_i - 1}.$$

Kérdés: szignifikáns-e a különbség a három csoport közt az IQ tekintetében $(\alpha = 0.05)$?