

Orvosbiológiai számítógépes gyakorlatok
(BMEVITMM203)
Mérési jegyzőkönyv

Biostatisztikai alapok

Készítették:

Jánosa Dávid Péter (FDSA7Y)

Mokánszki Béla (FA8YEZ)

Veres Dániel Sándor (GLZPT9)

2014. március 06.



M Ű E G Y E T E M 1 7 8 2

Tartalom

Bevezető	3
A mérés célja	3
A méréshez felhasznált eszközök	3
1. feladat: adatbázis importálása	4
2. feladat: változók általános deskriptív jellemzése.....	4
4. és 5. feladat: RBC és CRP mutatók hisztogramjának ábrázolása, eloszlásillesztés	10
A CRP értékek hisztogramjai:	13
6. feladat: RBC és CRP változók sűrűségfüggvényének magfüggvényes becslése.....	16
Az RBC érték becslője:	16
A CRP érték becslője:	17
7. feladat: RBC és CRP változók sodrófadiagramja (box plot).....	18
Az RBC értékek sodrófadiagramja:.....	18
A CRP értékek sodrófa diagramja:.....	19
8. feladat: változók eloszlásának véleményezése.....	21
9. feladat: változók eloszlásának ellenőrzése	22
Az RBC értékek QQ-plotja:.....	22
A CRP értékek QQ-plotja:.....	23
Analitikus vizsgálatok:.....	24
Következtetés:.....	25
10. feladat: eltérések vizsgálata az egészséges és beteg populáció között.....	28
11. feladat: az elhízásban megjelenő laborparaméter-változások fiziológiai okai	31
12. feladat: a változók összefüggéseinek vizsgálata főkomponens elemzés segítségével	32
13. feladat: az adott csoportban szereplő együttmozgó változók élettani tartalma	36

14. feladat: a változók összefüggéseinek vizsgálata főkomponens elemzés segítségével	37
Zárszó.....	37
Függelék: MATLAB forráskódok.....	38
Adatbázis betöltése	38
Változók általános deskriptív statisztikai jellemzése.....	38
RBC és CRP mutatók hisztogramjának ábrázolása, eloszlásillesztés.....	39
RBC és CRP változók sűrűségfüggvényének magfüggvényes becslése	40
RBC és CRP változók sodrófadiagramja (box plot).....	40
Változók eloszlásának ellenőrzése	40
Eloszlások normalitásának ellenőrzése	40
Eltérések vizsgálata az egészséges és beteg populáció között /Welch-próba/	41
A változók összefüggéseinek vizsgálata főkomponens elemzés segítségével	41

Bevezető

Modern orvosi kutatás elképzelhetetlen statisztika alkalmazása nélkül – az EBM (evidence based medicine) pedig kiemelten igényli ezt. A statisztikai módszerek egységes kialakításához vezető folyamat régen kezdődött, valamint a statisztika XX. század elején bekövetkezett robbanásszerű fejlődésének egyik elsődleges mozgatórugója éppen az orvosi, állatorvosi, alkalmazott biológiai feladatok során adódó problémák megoldásának igénye volt. Mára az alapvető eljárások letisztázódtak, és beépültek a mindennapi gyakorló orvosi, kutató orvosi gyakorlatba. (Kiemelnénk, hogy az orvosbiológiai cikkekben bizonyos statisztikai eszköztár alkalmazása ma már teljesen magától értetődő elvárás.) Jelen labormérés e módszerekre kíván példát mutatni, egy belgyógyászati (obezitológiai) esettanulmány felhasználásával.

A mérés célja

A mérés célja egy többváltozós obezitológiai adatbázis önálló biostatistikai elemzése. Az adatbázisban található adatokról, valamint a feladatok során használt biostatistikai megoldásokról, azok minimális elméleti háttéréről a feladatkiírás (https://www.iit.bme.hu/sites/default/files/biostat_2014_v03.pdf) melléklete részletes információkat ad, amelyeket a továbbiakban ismertnek feltételezünk, és a benne foglaltakra csak akkor hivatkozunk, amikor ez az eredmények értékelésének szerves része.

A méréshez felhasznált eszközök

A feladatokat jórészt MATLAB 2010a környezetben oldottuk meg, pár mellékszámításhoz pedig StatSoft – Statistica v8.0 programot használtunk. A jegyzőkönyv mellékleteként a teljes MATLAB forráskód dokumentálva megtalálható, ezért azt az eredmények tárgyalása során külön nem idézzük, gyakran felhasználjuk azonban a kód kimeneteként kapott táblázatokat és grafikonokat. A Statistica

számítások gyakorlatilag atomiak, és grafikus felületen keresztül történtek, így ezekhez forráskódot vagy dokumentációt nem adunk meg.

1. feladat: adatbázis importálása

A mérésvezető által közölt instrukcióknak megfelelően töltse le a felhasználandó adatfájlt (ObesityDatabase.csv) és olvastasság be a tartalmát MATLAB alá! (A fájl az útmutatóban említett adatbázis részlete, természetesen anonimizálva.) Az egyes változók tartalma nevük alapján általában egyértelmű, ezt a mérésvezető is ismerteti a mérés megkezdésekor.

Az adatokat a következő MATLAB kódokkal importáltuk:

```
db = importdata('ObesityDatabase.csv'); %importáljuk az adatokat%
colnames = db.colheaders;
col_source = find(strcmp(colnames, 'Source'));
col_rbc = find(strcmp(colnames, 'RBCCount'));
col_crp = find(strcmp(colnames, 'CRP'));
rbc_normal = group_normal(:, col_rbc); %csoportosítunk normál és obes szerint%
rbc_obese = group_obese(:, col_rbc);
crp_normal = group_normal(:, col_crp);
crp_obese = group_obese(:, col_crp);
```

2. feladat: változók általános deskriptív jellemzése

Valamennyi változóra (azaz laboreredményre) számítsa ki, mégpedig csoportonként bontva (egészséges/elhízott), a legfontosabb deskriptív mutatókat: az átlagot (mean), a mediánt (median) és a szórást (std). Figyeljen a jól kiértékelhető, célszerű elrendezésre (pl. táblázatos formátum, változók a sorokban, egészségesek és elhízottak egymás mellett stb.).

Az értékeket 40 normális, és 60 elhízott gyermek mérési adataiból származtattuk, amelyeket az „ObesityDatabase.csv” fájl tartalmaz.

A változók átlaga, mediánja és szórása a MATLAB kimenet alapján az 1. táblázatban található meg. A kapott statisztikai jellemzőket a kapott adatok számábrázolási pontosságának megfelelően (1 tizedes jeggyel többel) adtuk meg (a legutolsó tizedesekben lévő 0 értékeket nem tüntettük fel).

Alapvető leíró statisztikai elemek	Egészséges			Elhízott		
	Átlag	Medián	Szórás	Átlag	Medián	Szórás
AbsNeutrophilCount	3,428	3,35	0,921	3,834	3,75	0,986
AbsLymphocyteCount	2,207	2,175	0,43	2,733	2,72	0,572
AbsMonocyteCount	0,639	0,644	0,1428	0,6775	0,6165	0,2124
AbsEosinophilCount	0,1869	0,175	0,097	0,2574	0,1825	0,23
RBCCount	5,312	5,36	0,264	5,174	5,13	0,337
Haemoglobin	154,3	156	8,7	142,4	141	11,8
Haematocrit	0,4602	0,4635	0,0212	0,4223	0,4195	0,0298
MCV	86,71	86,85	2,79	81,62	82,3	3,16
MCH	29,06	29	1,02	27,53	27,6	1,5
MCHC	335,2	335,5	7,8	337,3	337	11
RDW	13,35	13,2	0,62	13,83	13,75	0,77
CRP	1,108	0,35	1,691	5,27	3,415	10,207
seSodium	139,3	139	1,4	138,3	138	2,1
seTotalProtein	77,87	77,95	4,23	75,86	76,1	4,62
seAlbumin	51,53	51,45	2,72	47,52	47,55	2,65
seCreatinine	78,9	76	9,4	60,8	60,5	9,5
Triglycerides	0,81	0,65	0,451	1,367	1,15	0,691
TotalCholesterol	3,622	3,595	0,682	4,28	4,125	0,953
HDLCholesterol	1,253	1,21	0,235	1,148	1,1	0,224
GOT	21,3	20	5,1	25,4	25	8,1
GPT	20,3	17,5	10,4	29,4	24,5	17,2
GGT	22,3	18	11,6	25,3	22	11

1. táblázat: A változók átlaga, mediánja és szórása

Az értékek mértékegységével a feladatkiírás és a segédanyag külön nem foglalkozik, bár a klinikai gyakorlatban az egyes laboratóriumok néha eltérő mértékegységeket használnak, illetve számos paraméter értéke relatív skálán értendő, amely az adott labor mérési módszereitől is függhet. Azonban a gyakorlat célja a statisztikai módszerek megismerése volt, így mi sem tartottuk szükségesnek, hogy az eredeti elemzésben utánajárjunk a mértékegységeknek.

A táblázatot megvizsgálva néhány érdekességet is észrevehetünk. Az MCV, CRP és szérum kreatinin értékeknél a két minta átlaga között az eltérés nagyobb, mint az adott változók (bármely csoportbeli) szórásának duplája. Önmagában ebből statisztikai következtetés természetesen nem vonható le, de ez további vizsgálódásra adhat okot, a későbbiekben erre vissza is térünk.

A CRP érték ráadásul olyan, amelynek mediánja az átlagtól főleg elhízásos esetben lényegesen eltér, azaz feltételezhető, hogy a CRP értékek outlierekkel terheltek, vagy

a változó eloszlása ferde (a későbbiekben ezt is tovább vizsgáljuk). A CRP szórása szintén jelentősen eltér a két mintában, de tekintve, hogy az átlaghoz hasonlóan a szórás sem robusztus mutató, az előzőek fényében ezt egy hibásan felvett, vagy kiszóró adat (pl. társult gyulladásos betegség) is okozhatja.

3. feladat: RBC és CRP mutatók részletes deskriptív jellemzése

Két változóra, a vörösvértestszámra (RBC) és a C-reaktív proteinre (CRP) számolja ki az összes fontos deskriptív mutatót: átlag (mean), 5%-os trimmelt átlag (trimmean), medián (median), alsó és felső kvartilis (quantile), minimum (min), maximum (max), terjedelem (range), korrigált szórás (std), interkvartilis terjedelem (iqr) és átskálázott medián abszolút eltérés (mad, de vigyázzon: ez alapbeállításban nem az általunk definiált MAD-et adja vissza). Összevetve egymással a különböző módon számolt, de hasonló tulajdonságot vizsgáló mutatók értékeit a két változónál, mit lát? (Gondoljon a robusztusságra! Ha kell, támaszkodjon a következő néhány feladat eredményére, és ide azok megoldása után térjen ismét vissza.)

A változók jellemzői MATLAB kimenetnek megfelelően (a számábrázolás a kapott értékeknek megfelelően – 1 tizedes jeggyel többel – történt):

RBC mutatók

'Mutató'	'Egészséges'	'Elhízott'
'Átlag'	[5.3120]	[5.1742]
'5% trimmelt átlag'	[5.3155]	[5.1717]
'Medián'	[5.3600]	[5.1300]
'Alsó kvartilis'	[5.1300]	[4.9500]
'Felső kvartilis'	[5.4850]	[5.3700]
'Minimum'	[4.6300]	[4.4900]
'Maximum'	[5.8600]	[6]
'Terjedelem'	[1.2300]	[1.5100]
'Korr. szórás'	[0.2635]	[0.3366]
'Interkvartilis terjedelem'	[0.3550]	[0.4200]
'Átskálázott MAD'	[0.2587]	[0.2957]

CRP mutatók

'Mutató'	'Egészséges'	'Elhízott'
'Átlag'	[1.1083]	[5.2697]
'5% trimmelt átlag'	[0.9292]	[4.0912]
'Medián'	[0.3500]	[3.4150]
'Alsó kvartilis'	[0.2150]	[1.3300]

'Felső kvartilis'	[1.2500]	[5.9400]
'Minimum'	[0.0300]	[0.1900]
'Maximum'	[8.9900]	[78.7000]
'Terjedelem'	[8.9600]	[78.5100]
'Korr. szórás'	[1.6906]	[10.2066]
'Interkvartilis terjedelem'	[1.0350]	[4.6100]
'Átskálázott MAD'	[0.3696]	[3.4523]

Az RBC értékekről általában elmondható, hogy az egészséges és az elhízott populáció deskriptív jellemzői nagyon hasonlóak, nagyobb arányú eltérés csupán a korrigált szórásban, és a terjedelemben található. Tekintve, hogy eközben sem a medián, sem az átlag nem mozdul el a szóráshoz képest komolyabban, feltételezhető, hogy az elhízott csoport RBC változó sűrűségfüggvénye a nem elhízott csoporthoz képest laposabb. Ellenőrzésképpen kiszámítottuk az RBC csúcsosságát Statistica programmal a két csoportban. (A MATLAB nem használja az ilyen számításoknál szokásos -3-as korrekciót, ezért egyértelműbb volt az említett szoftverrel a csúcsosság és ferdeség számítása.) A kontroll csoportban 0,015-ös, amíg az elhízottaknál valóban negatív, -0,351-es értéket kaptunk.

Mivel a trimmelt és a hagyományos átlag közti különbség viszonylag kicsi, nem valószínű, hogy a szórást pár véletlenszerű kiugró érték változtatta volna meg. Az RBC érték leírására a számítások alapján célszerű lehet a kevésbé robusztus, ellenben statisztikailag „erősebb” deskriptív mutatók használata (átlag, szórás).

Ezzel szemben a terjedelem és az interkvartilis terjedelem nagyságrendi különbsége valószínűsíti, hogy a CRP értékek nem robusztus mutatóinak (átlag, korrigált szórás) lényeges eltéréseért a jobbra kiugró érték(ek), illetve a ferde eloszlás lehet(nek) a felelősek. Azonban a robusztus mutatók (medián, kvartilisok, átskálázott MAD, trimmelt átlag) vizsgálata azt jelzi, hogy a változó eloszlása jelentősen lapult, valamint jobbra tolódott mind az egészséges (csúcsosság: 12,22; ferdeség: 3,21, Statistica alapján), és még inkább az elhízott csoportban (csúcsosság: 47,19; ferdeség: 6,55, Statistica alapján).

A feladat későbbi részeiben a CRP értékek viselkedéséről és eloszlásáról még fogunk szót ejteni, azonban ezen adatokból az tűnik célravezetőnek, hogy a CRP érték vizsgálataihoz inkább a robusztusabb, de statisztikailag „gyengébb” deskriptív mutatókat (medián, MAD, IQR) érdemes használnjuk.

Megjegyzendő az is, hogy a CRP szintnél a szórás nagyjából annyi, mint az átlag, (valamint a MAD is közel van a mediánhoz), az érték viszont fizikailag nem mehet 0

alá. Ez azt vetíti előre, hogy a CRP szint nem normális, hanem esetleg lognormális eloszlást követ. Ennek vizsgálatával később még foglalkozni fogunk.

A CRP érték megnövekedéséért elhízott emberekben feltehetőleg a bőr alatti zsírszövet által termelt IL-6 interleukin felelős, mely ebben a kontextusban a zsírok mobilizálását segíti. Az IL-6 és az általa mennyiségileg szabályozott CRP azonban szerepet játszik a szervezet gyulladásos reakcióiban, valamint hatással van az érfalak állapotára. A folyamatosan magas IL-6 és CRP koszorúér problémák, valamint ízületi betegségek (pl. rheumatoid arthritis) rizikófaktorai.

<http://atvb.ahajournals.org/content/19/4/972.long>

4. és 5. feladat: RBC és CRP mutatók hisztogramjának

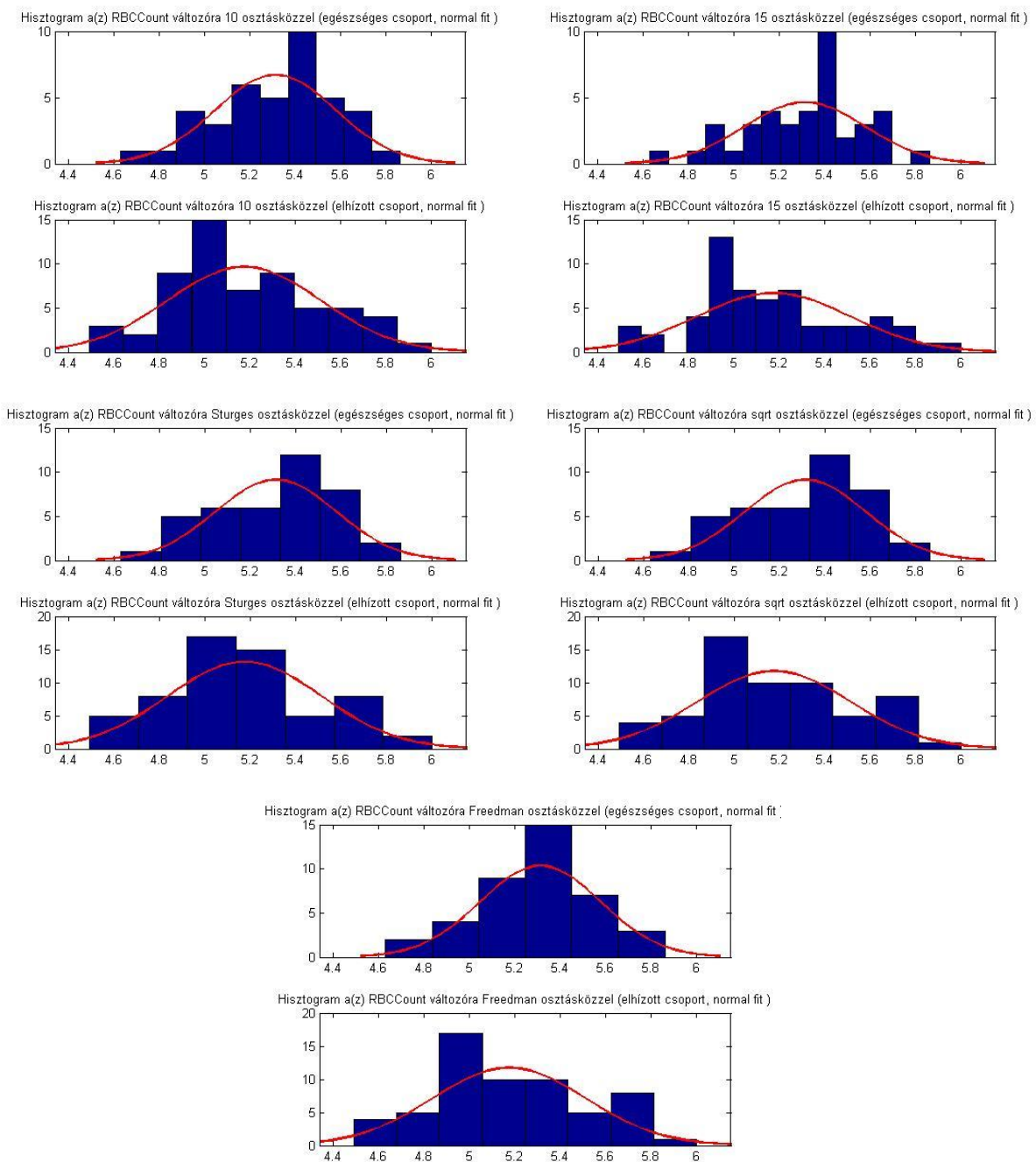
ábrázolása, eloszlásillesztés

A vörösvértestszám és a CRP eloszlását vizsgálja hisztogram létrehozásával (hist) külön az egészséges és elhízott csoportokra. Kísérletezzen az intervallum-hosszúság változtatásával, vizsgálja meg, milyen hatása van ennek a mutatott eloszlásra. Milyen tapasztalatokat tud megfogalmazni ez alapján az intervallumok számának választására vonatkozóan? Próbálja ki és vizsgálja meg az analitikus szabályok (Sturges, Freedman–Diaconis stb.) alkalmazását is. (A hist használatakor kényelmesebb lehet az intervallum-hosszúság helyett az intervallumok számának megadása; ilyenkor az egyes intervallumok értelemszerűen azonos hosszúságúak lesznek. Ha mégis kísérletezni szeretne az intervallumhosszok megadásával is (például mert nem azonos hosszúságú intervallumokat szeretne definiálni), akkor fontolja meg a histc alkalmazását is.) Úgy szerkessze meg és rendezze el az összetartozó (egy laboreredményt leíró) ábrákat, hogy világosan mutassák a változást a csoportok (elhízott, ill. egészséges) között, ha van ilyen. (Ehhez – itt és a további hasonló helyzetekben is! – célszerű és elegáns eszköz lehet subplot használata.)

A fent említett hisztogramokra rajzolja fel a rájuk legjobban illeszkedő normális eloszlás sűrűségfüggvényét is! Ítéld meg ez alapján a változók normalitását, illetve lognormalitást is. (Tipp: egy változó logaritmusát a log paranccsal képezhető tetszőleges eloszlást sűrűségfüggvényét a pdf, a normális eloszlás sűrűségfüggvényét egyszerűbben a normpdf utasítás tudja kiszámítani adott pontokban. Ennél az útnál könnyebb megoldás a histfit alkalmazása.)

A következőkben megvizsgáljuk, milyen képet mutatnak az RBC és a CRP értékek 10, 15, valamint a Sturges, négyzetgyök és Freedman-Diaconis szabályoknak megfelelő osztásközös hisztogramon ábrázolva. A hisztogramokra az ötödik feladatnak megfelelően egy becsülő normális eloszlás valószínűségi sűrűségfüggvényét is illesztjük.

Az RBC értékek histogramjai:



1. ábra Histogramok az RBCCount változóra különböző osztásközzel

Az RBC histogramokon jól látszik a módszer elsődleges problémája, nevezetesen hogy az osztályközök lényegesen befolyásolják a kapott képet, különösen kevés adat esetén. Ráadásul az is látszik, hogy a histogramok osztásközének megválasztását

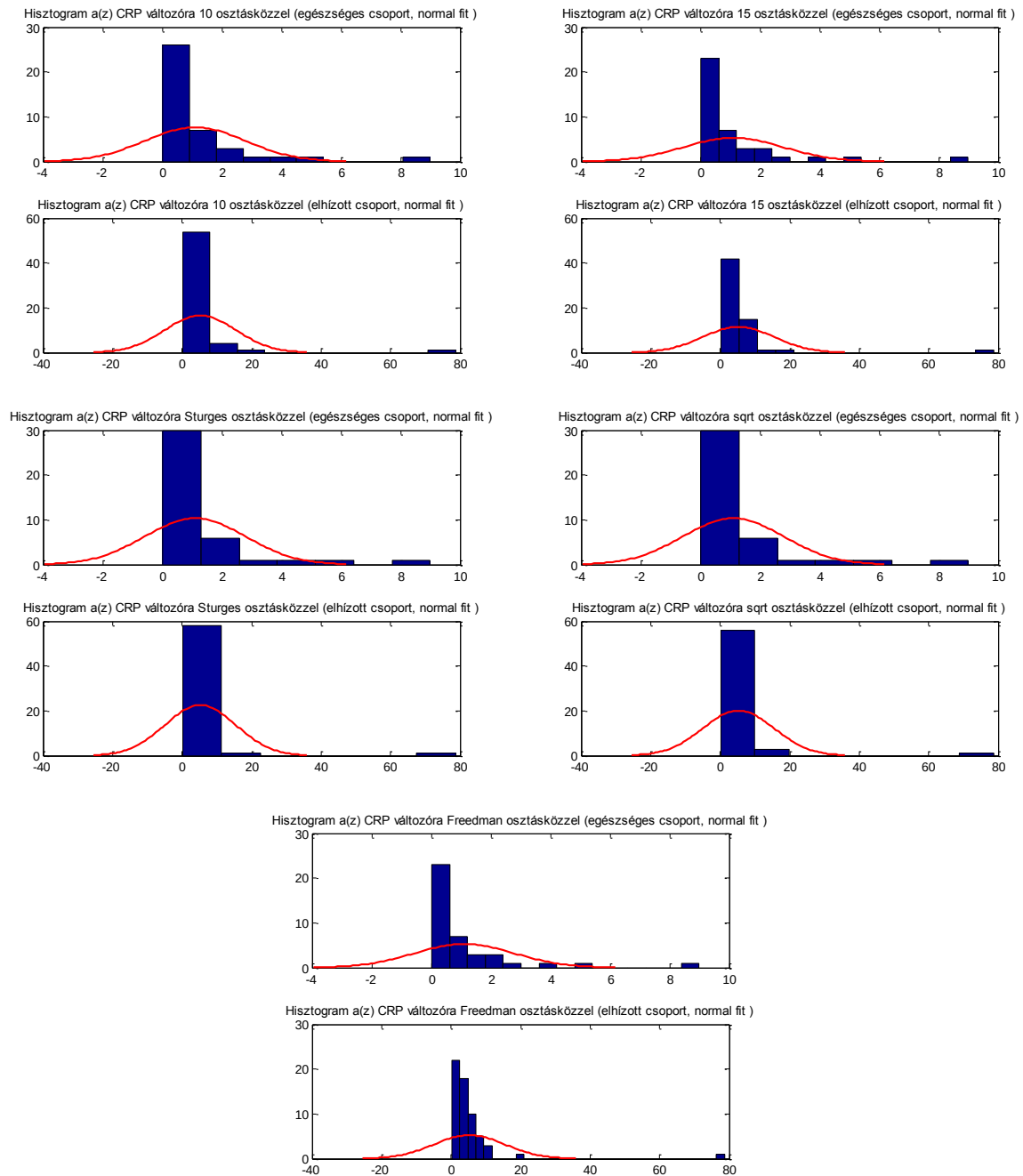
elméletben könnyítő ökölszabályok a valóságban eltérő eredményt produkálhatnak, például jelen esetben a Struges és az sqrt szabály használata esetén. Ez jól mutatja a módszer kritikusságát, és azt is, hogy nem szabad hagyni, hogy egyszerű empirikus szabályok hamis biztonságérzetet alakítsanak ki egy módszer alkalmazásakor.

Élettani háttérismereteinkből alapján normális eloszlású RBC értéket várunk, azonban ezt szemléletesen a nem elhízott minta esetén csak a Freedman-szabályt követő hisztogram adja vissza, ez elhízott populációból származó hisztogramok pedig laikus szemmel nem nagyon hasonlítanak normális eloszlásra. Érdekes megnézni, hogy a normális minta gyökös és Freedman-szabály alapján elkészített hisztogramja között csak egy osztályköz különbség van, ez azonban az ábrázolást sokkal árnyaltabbá teszi a Freedman javára.

Az egészséges és elhízott csoportokat összevetve a hisztogramok alapján bármely beosztást nézzük is, azt mondhatjuk, hogy amíg az egészségeseknél egy jobbra, addig az elhízottaknál beteg egy balra „ferdülő” (a csúcs ebbe az irányba tolódik) eloszlás sejthető.

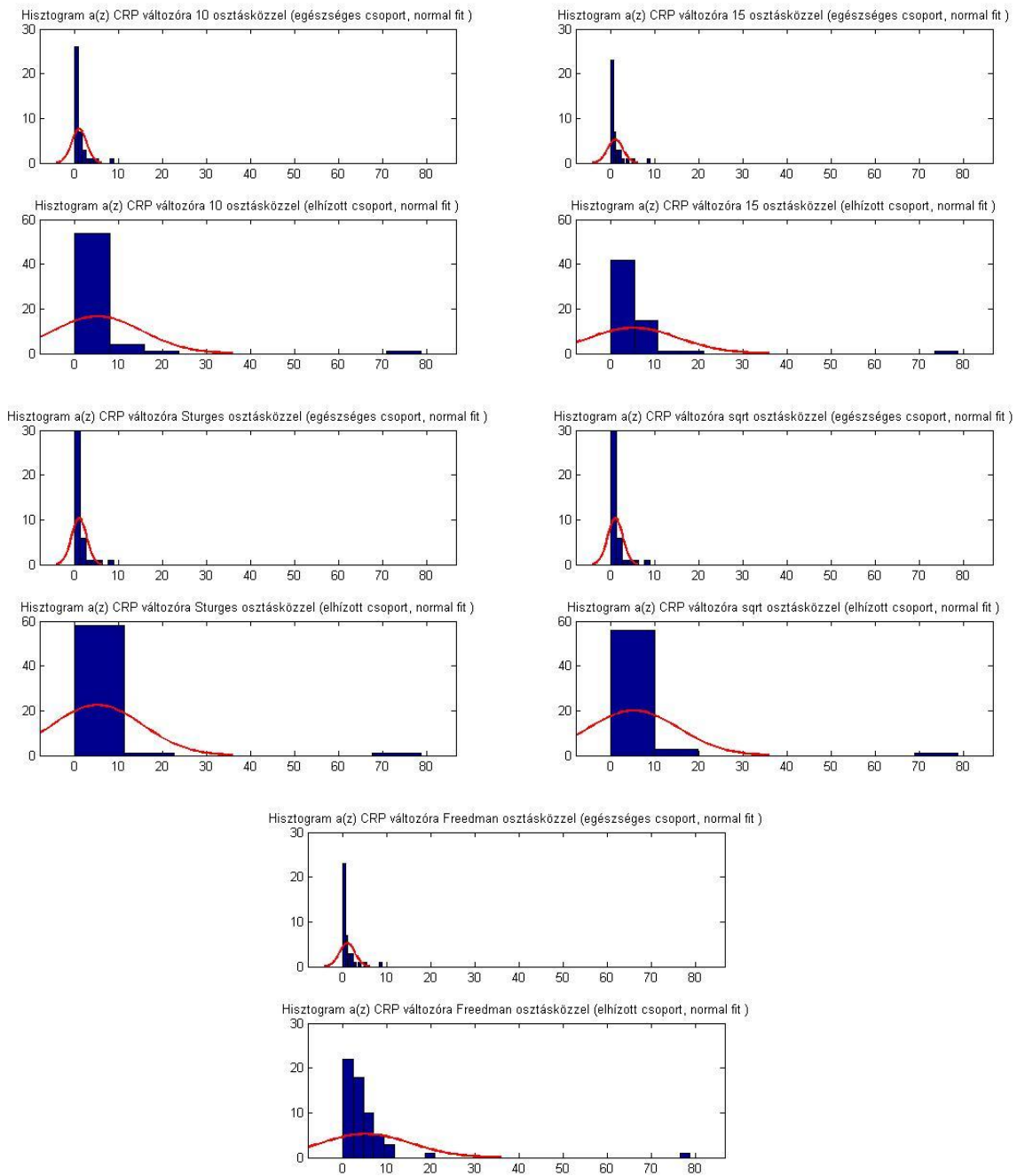
Az igazán szemléletes hisztogramok elkészítése roppant igényes munka, azonban az osztályközök minden kontroll nélküli „optimalizálása” magában rejti a veszélyt, hogy a végeredménybe a mérést értékelő személy olyan tendenciákat csempészhetsz bele, melyek az eredeti sokaságot valójában nem jellemzik, csupán az adott minta és illesztési szabály összefonódása által keltett illúzióról van szó. Ezért is roppant fontosak az a priori definiált hisztogramillesztési algoritmusok – de az előbbieket is csak megfelelő megfontolások után érdemes használnunk.

A CRP értékek hisztogramjai:



2. a. ábra Hisztogramok CRP-re különböző osztásközzel

A CRP értékek hisztogramjait két módon is ábrázoltuk a különböző összehasonlíthatóságok miatt: az első esetben az osztályközök méretének változását figyelhetjük meg jobban, a második ábraszorozatnál (2. b.) pedig könnyebb a csoportok összevetése az egységes x skála miatt.



2.b, ábra Hisztogramok CRP-re különböző osztásközzel

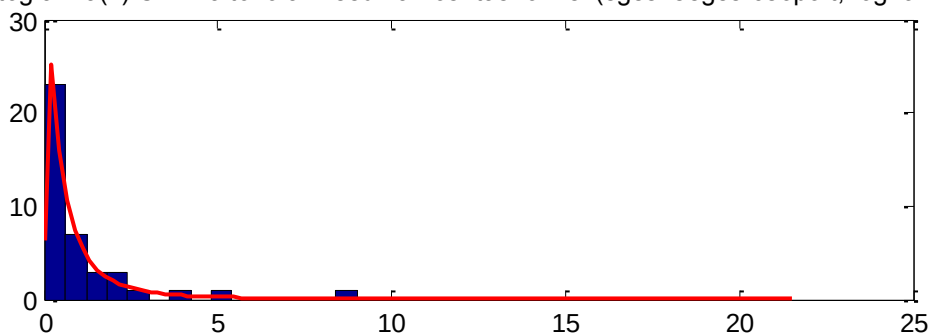
A CRP értékek hisztogramja esetében az ekvidisztális osztályközök visszaütnek. A távoli, kiugró értékek miatt több olyan osztályunk van, amiben nincs érték, a minta nagy része pedig két-három osztályba tömörül. Részben orvosolja a problémát a 10, 15 és az elhízott populáció mintájának esetén a Freedman osztályközös ábrázolás, ami még a nagyszámú kihasználatlan osztállyal együtt is képes jó felbontással reprezentálni a mintát, mivel ebben az esetben az eleve nagyszámú osztályból a hasznos területre is több jut. Ezzel együtt, ha egy igazán szép hisztogram lenne a cél,

mindenképp érdemes volna a kiugró értékeket 1-2 saját osztályba vonni, a tényleges mintát pedig az eddigiekhez hasonlóan felbontani.

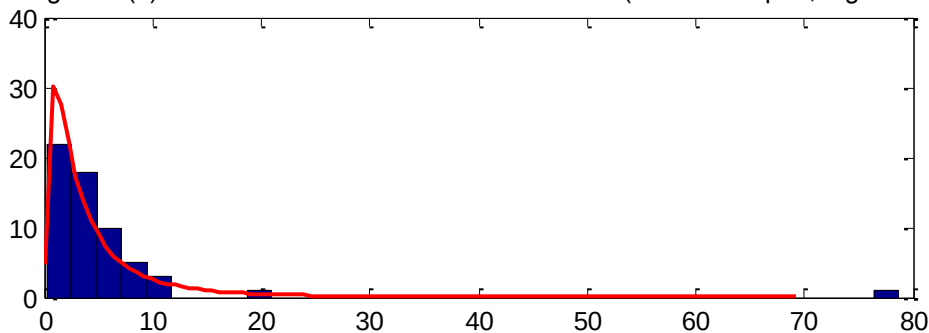
Másik probléma, hogy az illesztések alapján a normális eloszlás nem megfelelő leírója a CRP szintnek, ami a deskriptív mutatók vizsgálatakor is felmerülhetett volna, amikor a szórás és az átlag viszonyát vizsgáltuk. Természettudományos ismereteink alapján feltételezhetjük, hogy a CRP (és általában a sok komponensű, korlátos és a szóráshoz képest a korláthoz közel eső várható értékű valószínűségi változó) lognormális eloszlást követ.

Ennek bemutatására a 15 osztályközös hisztogramot lognormális eloszlás illesztésével ábrázoljuk:

Hisztogram a(z) CRP változóra Freedman osztásközzel (egészséges csoport, lognormal fit)



Hisztogram a(z) CRP változóra Freedman osztásközzel (elhízott csoport, lognormal fit)



3. ábra Hisztogramok a CRP változóra Freedman osztásközzel

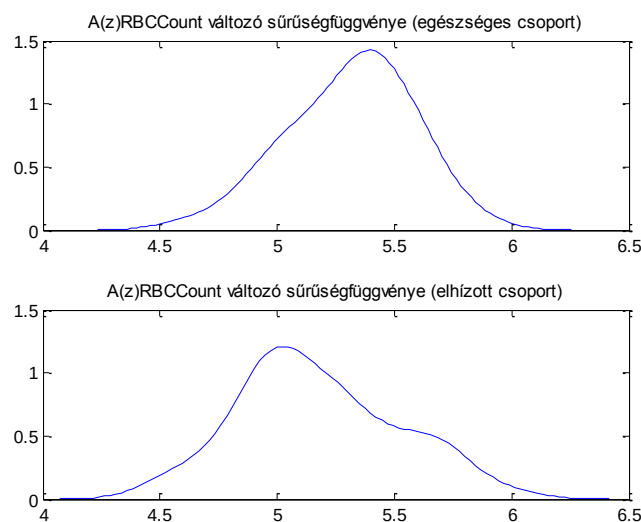
Látható, hogy a lognormális illesztés sokkal hűebben követi a hisztogram kontúrait.

6. feladat: RBC és CRP változók sűrűségfüggvényének

magfüggvényes becslése

Készítsen a fenti két változó sűrűségfüggvényéről (természetesen továbbra is egészséges és elhízott csoportok szerint bontva) magfüggvényes becslést is (*ksdensity*) elfogadva az alapértelmezett paramétereket! Úgy szerkessze meg és rendezze el az összetartozó (egy laboreredményt leíró) ábrákat, hogy világosan mutassák a változást a csoportok (elhízott, ill. egészséges) között, ha van ilyen.

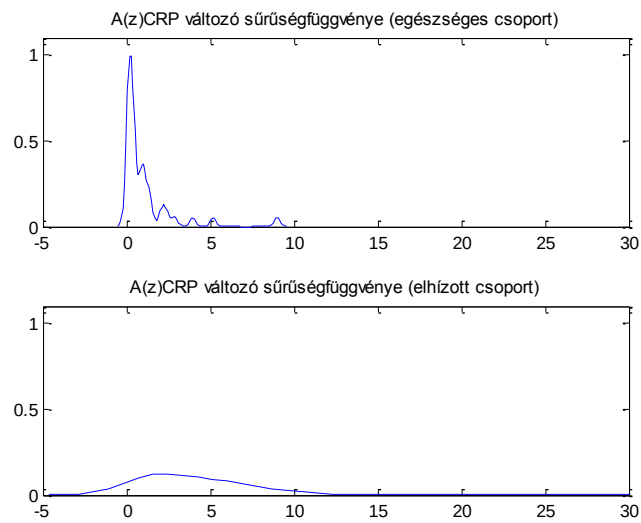
Az RBC érték becslője:



4. ábra Az RBCCount változó sűrűségfüggvényei a két csoportban

A becsült sűrűségfüggvény az elhízott gyerekeknél enyhén balra tolódik, valamint ellapul, ahogy arra a deskriptív mutatók vizsgálatakor is gondoltunk. A deskriptív mutatókból azonban nem látszott, hogy míg az egészséges mintát negatív, addig a túlsúlyos mintát pozitív ferdeségű eloszlás jellemzi. Ennek megfelelően a ferdeség értékek: -0,41 az egészségeseknél, amíg 0,33 az elhízottaknál.

A CRP érték becslője:



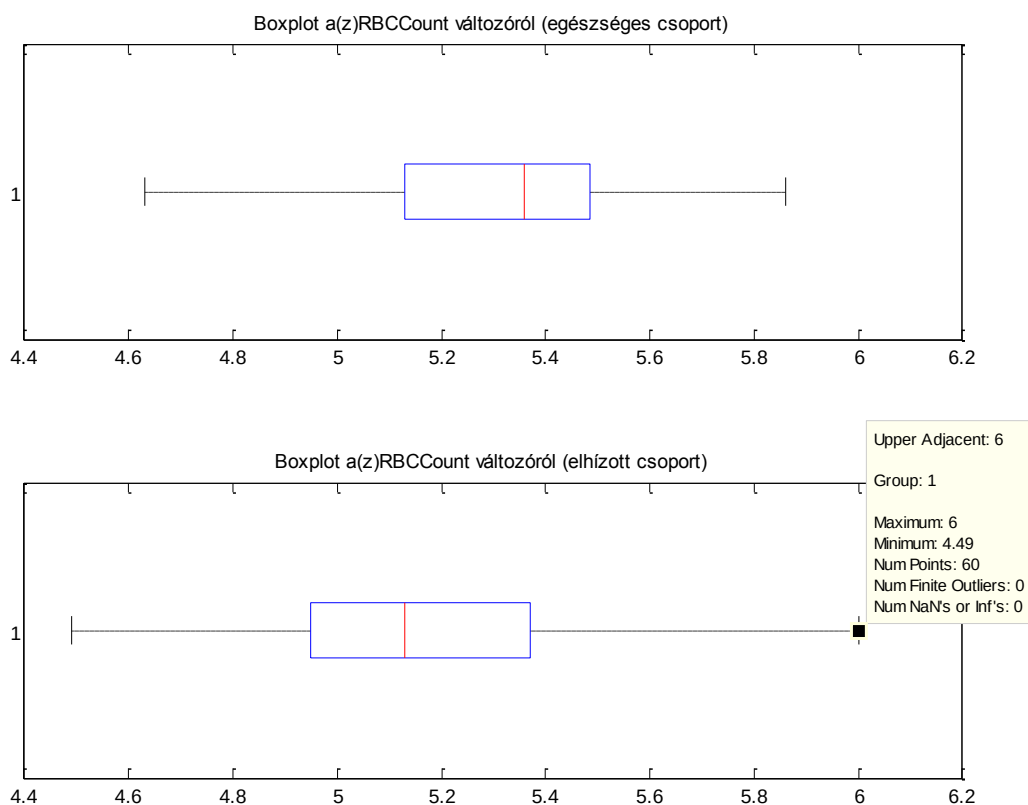
5. ábra A CRP változó sűrűségfüggvényei a két csoportban

Ebben az esetben nagyon szembeűnő az eloszlás ellaposodása, csökkenő ferdesége, valamint erős jobbra tolódása – ahogyan ezt korábban is jeleztük. Az eltérés szignifikánsságának meghatározása itt is a későbbi feladatokban fog megtörténni.

7. feladat: RBC és CRP változók sodrófadiagramja (box plot)

Készítsen csoportbontásban boxplotot (boxplot) a fenti két változóról! Lát outlier 1,5IQR szabály (tipp: ez az alapértelmezés) használatával? Ha igen, jellemezze elhelyezkedésüket, számukat stb. Mire tud következtetni ez alapján az eloszlásokra vonatkozóan? Úgy szerkessze meg és rendezze el az összetartozó (egy laboreredménytleíró) ábrákat, hogy világosan mutassák a változást a csoportok (elhízott, ill. egészséges) között, ha van ilyen.

Az RBC értékek sodrófadiagramja:



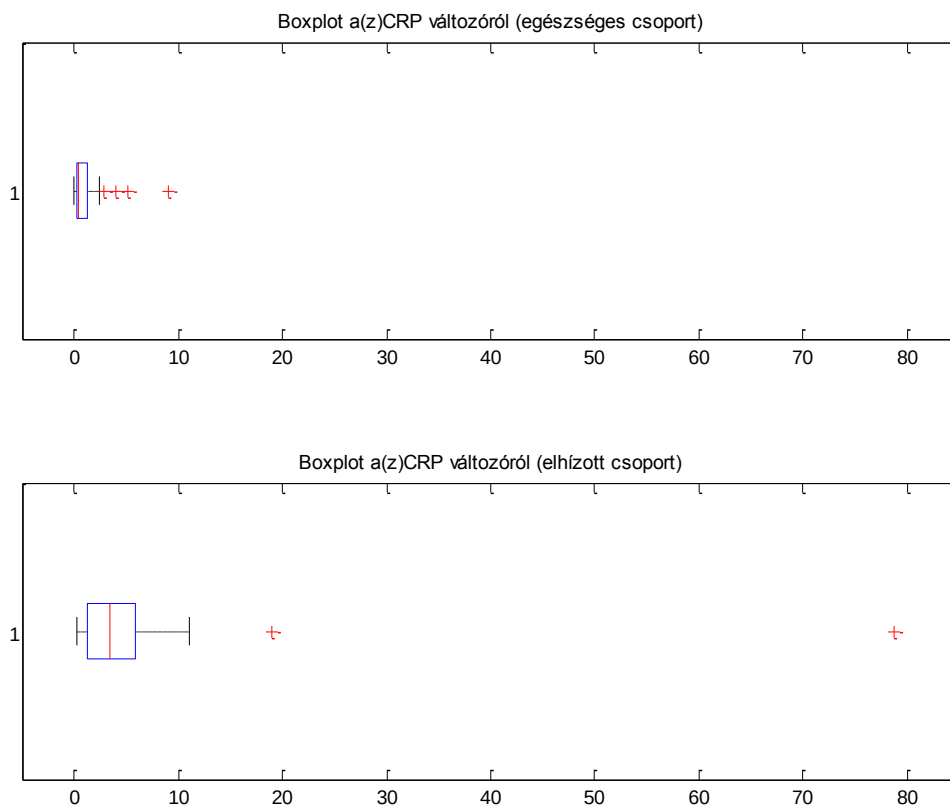
6. ábra Boxplot az RBCCount változóra a két csoportban

Az egészséges és az elhízott alanyok mintája is mentes az outlier értékektől, azonban az elhízottak között akad egy hatos érték, melynek távolsága a 3. kvartilistól éppen 1,5 IQR. Az outlier értékek a nem robusztus deskriptív mutatókat, így az egyébként balra tolódó átlagot elhúzhatják, azonban láttuk, hogy az átlag és a trimmelt átlag

közötti eltérés meglehetősen kicsi volt, így azt, hogy a medián az átlagnál jelentősebben változott, inkább a ferdeség változása okozhatja.

A medián, valamint a teljes eloszlás balra tolódása, valamint a ferdeség változása (a medián aszimmetrikus elhelyezkedése az IQR-en belül) itt is megfigyelhető.

A CRP értékek sodrófa diagramja:



7. ábra Boxplot a CRP változóról a két csoportban

Bár az egészségesek esetében az ábrán jobban megfigyelhető lenne a boxplot kisebb tartományú skálán, de a két csoport összevethetősége miatt jobbnak láttuk ugyanazon a skálán ábrázolni a két csoport sodrófáját. Az egészséges populációból származó mintában 4 outlier található, ezek távolsága a mediántól +2,53, +3,6, +4,82 és +8,64 IQR. Az obese populáció mintáját 2 két outlier terheli, ezek elhelyezkedése a mediánhoz képest +3,38 és +16,33 IQR. Különösen az utóbbi értékek hatvan körüli elemszámnál jelentősen elhúzzák az átlagot és a szórást is, tehát feltevésünk, hogy a

nem robusztus deskriptív jellemzők eltéréseért kiugró értékek is felelősek, igaznak tűnik.

A sodrófa diagramon megfigyelhetjük azt is, hogy a medián elhelyezkedése az IQR-en belül az elhízott populációból származó mintában jobbra tolódik és szimmetrikusabb, ami egybevág a becsült sűrűségfüggvény ferdeségének csökkenésével is. Szintén megfigyelhető az eloszlás ellapulása, az IQR növekedése.

Korábban említettük, hogy a CRP a gyulladásos immunfolyamatokban játszik szerepet. Egy ilyen outlier értéknél jogos a felvetés, hogy az értéknek nem annyira az elhízáshoz, mint inkább egy bakteriális fertőzéshez lehet köze, ekkor a CRP értéke 40 és 200 mg/l között szokott változni. Tekintve, hogy egy ilyen „rejtett változó” előfordulási valószínűsége főleg nagy mintánál jelentős, valóban érdemes lehet a CRP mérésénél a robusztus mutatókra támaszkodni.

8. feladat: változók eloszlásának véleményezése

A fentiek alapján véleményezze, hogy ez a két változó milyen eloszlást követ, természetesen csoportokon belül (csoportonként).

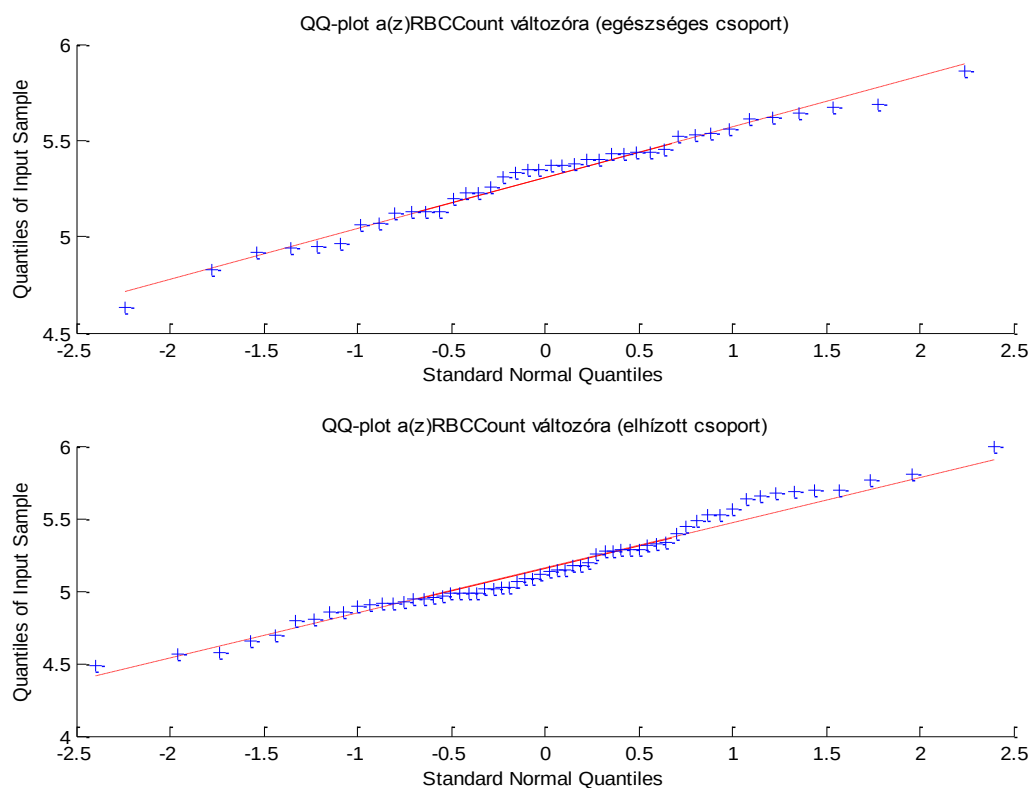
A mintánk eddigi vizsgálatait alapján feltételezhető, hogy amíg az RBC érték mind az egészséges, mint az elhízott populációban normális eloszlással írható le, addig a CRP az egészséges populációban semmiképp, de a beteg populációban sem nagyon adható meg normális eloszlással, mivel azt az érték alsó korlátja nagymértékben torzítja. Ezen változó esetében a lognormális eloszlás tűnik helytállónak az eddigi vizsgálatok alapján.

9. feladat: változók eloszlásának ellenőrzése

QQ-plot használatával (*qqplot*) grafikusan, valamint Lilliefors-teszt (*lillietest*) és Jarque–Bera-teszt (*jbtest*) használatával analitikusan ellenőrizze az előző pontbeli feltevések teljesülését. (Tipp: figyeljen oda, ha lognormalitást kíván tesztelni, hiszen ezek a próbák (a JB-teszt mindenképp, a Lilliefors-teszt alapbeállításában) normalitást vizsgál.) A hipotézisvizsgálatok eredményét – itt és a továbbiakban is! – *p*-érték formájában is közölje, ne pusztán a döntést írja le!

Az előző feladatban megfogalmazott állításokat az alábbiakban analitikus és grafikus eszközökkel igyekszünk alátámasztani. Elsőként a két változó normalitását vizsgáljuk QQ-plot segítségével. Ezen az ábrázoláson a normális eloszlásúnak jósolt RBC értékek egyenest kell, hogy kirajzoljanak, míg a lognormális CRP-től egy egyenestől elhajló ívet várunk.

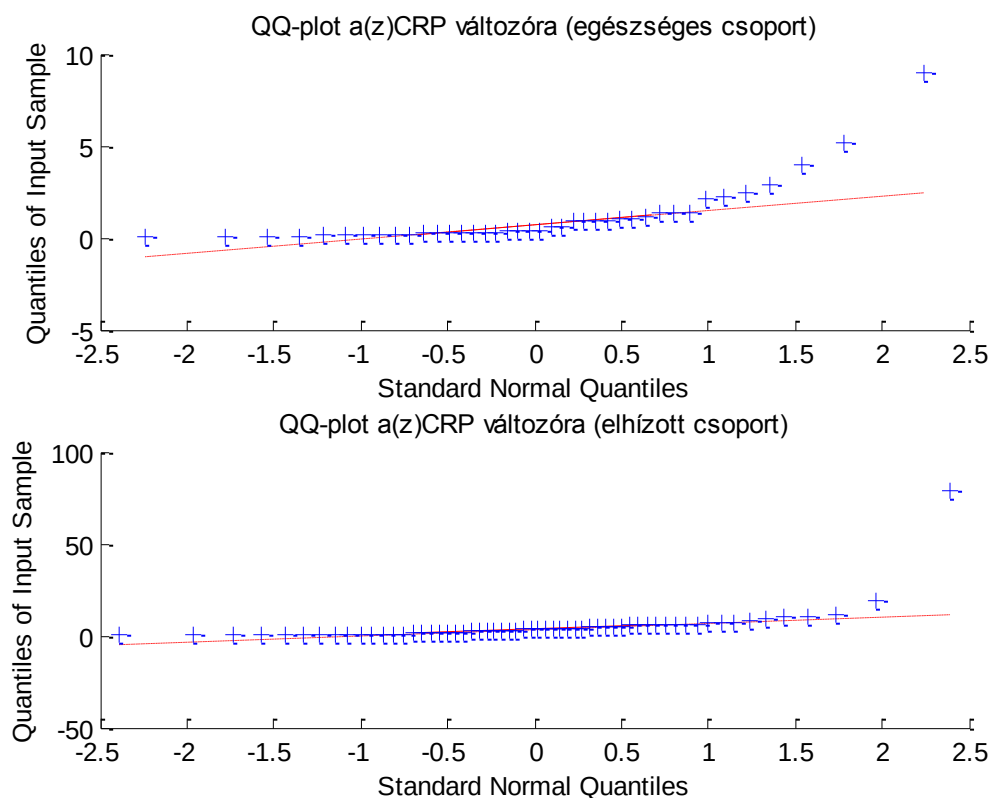
Az RBC értékek QQ-plotja:



8. ábra QQ-plot az RBCCount változóra a két csoportban

Mindkét minta szemre nagyjából jól illeszkedik egy egyenesre, ami alátámasztja feltevésünket arról, hogy az RBC változó mindkét csoportban (és feltételezhetően populációban) normál eloszlást követ. Bár az elhízott mintában a nagyobb értékek között több komolyabban eltér a normális eloszláshoz tartozó egyenestől, azonban a pontok nagy része jól illeszkedik, így feltehetően itt is normális eloszlást fognak mutatni az analitikus tesztek, bár előfordulhat, hogy kisebb bizonyossággal.

A CRP értékek QQ-plotja:



93. ábra QQ-plot a CRP változóra a két csoportban

A CRP esetében az egészséges minta QQ-plotja nem mutat az egyenesre illeszkedést, így az szinte biztosan nem normális eloszlást követ. Az elhízott minta azonban nem ilyen egyértelmű: a minta jó része szépen illeszkedik az egyenesre, az utolsó, kiugró érték azonban nagyon elüt ettől. Ezzel együtt nem valószínű, hogy az analitikus tesztek emiatt elvessék a minta normalitását.

Analitikus vizsgálatok:

Analitikus vizsgálatként Lilliefors és Jarque-Bera tesztekét végeztünk el az RBC változóra (mely sejtésünk szerint normális eloszlású), valamint a CRP változónál annak logaritmusára (mely sejtésünk szerint lognormális eloszlású) is.

A MATLAB ezeken vizsgálatokhoz külön függvényeket kínál, melyek az alábbi kimeneteket adják:

h: értéke 0, ha a nullhipotézist, azaz hogy a vizsgált változó normális eloszlású, 5%-os szignifikanciaszint mellett megtartjuk. *h* értéke 1, ha a nullhipotézist ugyanezen szignifikanciaszint mellett elvetjük.

p: a nullhipotézis szignifikancia szintje, azaz annak a valószínűsége, hogy a nullhipotézist elvetem, pedig az igaz volt.

Mivel a MATLAB a *p* értékeket egy előre beprogramozott táblázat alapján interpolációval becsli, ezért ezek az értékek nem teljesen pontosak, valamint nem tudnak az intervallum széléhez szélsőségesen közeli értékeket felvenni, így ezeket a MATLAB az intervallum megfelelő szélén 0,5-el, vagy 0,001-el helyettesíti. Ha ennél pontosabb meghatározásra van szükség, ahhoz Monte-Carlo szimuláció futtatása célszerű.

Az analitikus normalitásvizsgálatok eredménye:

Az RBC változó normális eloszlásának ellenőrzése (Lilliefors):

egészséges csoport: $[h, p] = [0, 0.2819]$

elhízott csoport: $[h, p] = [0, 0.1467]$

Az RBC változó normális eloszlásának ellenőrzése (Jarque-Bera):

egészséges csoport: $[h, p] = [0, 0.4790]$

elhízott csoport: $[h, p] = [0, 0.3716]$

Az CRP változó normális eloszlásának ellenőrzése (Lilliefors):

Warning: P is less than the smallest tabulated value, returning 0.001.

> In lillietest at 170

egészséges csoport: [h, p] = [1, 0.0010]

Warning: P is less than the smallest tabulated value, returning 0.001.

> In lillietest at 170

elhízott csoport: [h, p] = [1, 0.0010]

Az CRP változó normális eloszlásának ellenőrzése (Jarque-Bera) :

Warning: P is less than the smallest tabulated value, returning 0.001.

> In jbstest at 142

egészséges csoport: [h, p] = [1, 0.0010]

Warning: P is less than the smallest tabulated value, returning 0.001.

> In jbstest at 142

elhízott csoport: [h, p] = [1, 0.0010]

Az CRP változó lognormális eloszlásának ellenőrzése (Lilliefors) :

egészséges csoport: [h, p] = [1, 0.0257]

elhízott csoport: [h, p] = [0, 0.2757]

Az CRP változó lognormális eloszlásának ellenőrzése (Jarque-Bera) :

Warning: P is greater than the largest tabulated value, returning 0.5.

egészséges csoport: [h, p] = [0, 0.5000]

elhízott csoport: [h, p] = [0, 0.3482]

Következtetés:

Az különböző analitikus vizsgálatok a döntések lényegére tekintve ugyanazt az eredményt adták, kivéve az egészséges csoport lognormalitására vonatkozóan.

A normalitásvizsgálatok a korábbi sejtéseinket igazolták, azaz a következőket. A vörösvértetek számának eloszlása (RBC) mind az egészséges, mint az elhízott

populációban nem különböznek szignifikánsan a normális eloszlástól (ezt jelenti a nullhipotézisünk elfogadása), ezért normál eloszlásúnak tekinthetjük őket. Mindenképpen megjegyzendő azonban, hogy statisztikai próbáknál a nullhipotézis elfogadása a másodfajú hiba elkövetésének lehetőségét vonja maga után, amelyet az elsőfajú hibával ellentétben nem ismerünk – így bár a magas p értékeink azt mutatják, hogy a nullhipotézis elvetése nagy valószínűséggel okoz hibás döntést a populációra vonatkozóan, ezért ezt nem tesszük meg, azonban a nullhipotézis téves elfogadásának valószínűségéről nem tudunk biztosat mondani.

A C-reaktív protein (CRP) eloszlása sem az egészséges populációban sem a beteg populációban nem tekinthető normál eloszlásúnak (különbözik attól) mintáink alapján – ezt jelenti a nullhipotézisünk elvetése, amely döntésről jelen esetben tudjuk, hogy legfeljebb mekkora valószínűséggel hibás (p értékek).

Az elhízottak esetében korábbi feltevéseinkre (természetesen a néhány bekezdésben feljebb levő megjegyzést figyelembe véve) egyértelmű választ kaptunk tesztjeink alapján: a CRP eloszlása itt lognormális.

A legérdekesebb eredményt az egészséges populáció CRP lognormalitásának vizsgálatakor kaptuk: korábbi sejtésinkkel ellentétben a Lilliefors módszer eredménye alapján eloszlásunk szignifikánsan különbözik a lognormalistól. Ezzel ellentétben ráadásul a Jarque-Bera teszt magas p értéket ad. A programkódban nem találtunk hibát, így ezt a furcsaságot a Statistica-val is ellenőriztük – amely hasonló eredményt hozott. Ez utóbbi programcsomaggal másik két normalitásvizsgálatot is végeztünk a logaritmált CRP értékekre: mind a Kolmogorov-Smirnov (ennek egy módosítása a Lilliefors), mind a Shapiro-Wilk próbák eredménye a Jarque-Bera-hoz hasonló eredményt mutatott: (Az eredményeket a MATLAB kimeneteli formátumához hasonlóan tüntetjük fel az egyszerűbb összehasonlíthatóság kedvéért.):

Az CRP változó lognormális eloszlásának ellenőrzése az egészséges csoportban (Statistica):

Lilliefors: [h, p] = [1, <0.05]

Kolmogorov-Smirnov: [h, p] = [0, >0.2]

Shapiro-Wilk: [h, p] = [0, 0.5788]

Ezek alapján feltételezzük, hogy a Lilliefors jóval érzékenyebb a kiugró értékekre, mint a többi próba. Összevetve eddigi eredményeinket úgy gondoljuk, hogy a CRP eloszlása tekinthető lognormális eloszlásnak az egészségesek körében is.

Az itt kapott eltérés és a korábban a hisztogramoknál megfigyelték alapján a későbbiekben is fokozottan ügyelnünk kell az egészséges csoportban a CRP vizsgálatára megfelelő, kiugró adatokra kevésbé érzékeny tesztek alkalmazva, vagy ha ez indokolható (ha a kiugró értéknek metodikai, illetve fiziológiai okát megtaláljuk) a kiugró értékek elhagyásával érdemes vizsgálódnunk.

[További „érdekességre” bukkanhatnánk, ha vizsgálnánk a normál eloszlásúnak tekintett RBC változónk lognormalitását – azaz, hogyan tekinthető a normális eloszlás lognormál eloszlásnak – de ennek elemzése, okai nem tartoznak szorosan a jegyzőkönyvhöz, így eltekintettünk tőle.]

10. feladat: eltérések vizsgálata az egészséges és beteg

populáció között

Kétmintás Welch-próba (ttest2) használatával vizsgálja meg (a próba előfeltételeinek ellenőrzésével most nem kell törődnie), hogy mely változók (laboreredmények) esetében van szignifikáns eltérés az egészséges és elhízott csoportok között, valamennyi változót tekintve. (A „szignifikáns eltérést” most mint a sokasági várható értékek szignifikáns különbségét operacionalizálja.) A kitűzött szignifikanciaszint 5%, de ne feledkezzen meg ennek korrekciójáról (figyelemmel arra, hogy több összehasonlítást is végez). A Bonferroni-korrekció használata most elégséges. (Vigyázzon, a ttest2 alapértelmezésben, ha mást nem specifikál, feltételezi a szóráshomogenitást, azaz nem Welch-, hanem klasszikus t-próbát végez.)

Ebben a feladatban Welch-próbával fogjuk megvizsgálni, hogy vannak-e olyan változók, amelyek eltérést mutatnak az elhízott és az egészséges populációra a minta alapján. A vizsgálat szignifikancia szintje 5%, azonban tekintettel kell lennünk arra, hogy statisztikailag nem korrekt ugyanazt a sokaságot több szempontból is úgy vizsgálni, mintha minden vizsgálathoz külön minta állna rendelkezésre. Emiatt a szignifikancia szintet Bonferroni-korrekciónak vetjük alá. Ez jelen esetben $\alpha = 0,21\%$ -os szignifikanciát ír elő. Amennyiben minden vizsgálatban tartjuk magunkat ehhez a szignifikancia szinthez, úgy elmondhatjuk, hogy az összes vizsgálatra nézve a szignifikancia szintünk 5%.

A MATLAB az értékelés során az alábbi kimeneteket adja:

h: értéke 0, ha a nullhipotézist, azaz hogy a két változó eloszlása az adott szignifikanciaszint mellett nem eltérő, megtartjuk. Értéke egy, ha a nullhipotézist elvetjük, azaz az annak az esélye, hogy az eloszlások közti abszolút eltérés csupán a véletlennek köszönhető, kisebb, mint α .

p: a nullhipotézis szignifikancia szintje, azaz annak a valószínűsége, hogy a nullhipotézist elvetem, pedig az igaz volt.

A próbák eredményei:

Welch-próba

'Változó'	'h'	'p'
'AbsNeutrophilCount'	[0]	[0.0385]
'AbsLymphocyteCount'	[1]	[9.4771e-007]
'AbsMonocyteCount'	[0]	[0.2803]
'AbsEosinophilCount'	[0]	[0.0380]
'RBCCount'	[0]	[0.0243]
'Haemoglobin'	[1]	[8.0840e-008]
'Haematocrit'	[1]	[3.8656e-011]
'MCV'	[1]	[4.2921e-013]
'MCH'	[1]	[2.3610e-008]
'MCHC'	[0]	[0.2708]
'RDW'	[1]	[8.8194e-004]
'CRP'	[0]	[0.0029]
'seSodium'	[0]	[0.0062]
'seTotalProtein'	[0]	[0.0275]
'seAlbumin'	[1]	[1.7435e-010]
'seCreatinine'	[1]	[1.0112e-014]
'Triglycerides'	[1]	[4.2015e-006]
'TotalCholesterol'	[1]	[1.1385e-004]
'HDLCholesterol'	[0]	[0.0277]
'GOT'	[0]	[0.0026]
'GPT'	[1]	[0.0014]
'GGT'	[0]	[0.2051]

Szignifikáns tehát az eltérés az abszolút limfocitaszámban, a hemoglobin, a hematokrit, az MCV, MCH, RDW, szérum albumin, szérum kreatinin, triglicerid valamint a teljes koleszterin értékekben.

Ebből a kreatinin és az MCV semmiképp nem meglepetés, ezek eltérése már csupán a deskriptív jellemzőkre tekintve is jelentősnek tűnt. Érdekes azonban, hogy a próba szerint a CRP eltérés nem szignifikáns. Ez a (szakirodalom alapján melleleg hamis) következtetés több oknak is betudható. Először is a Bonferroni-korrektúra miatt sok olyan összevetés esetében is meg kellett tartanunk a nullhipotézist, amelyek esetében

egy metodikailag helyesen kivitelezett egyváltozós vizsgálat találhatott volna szignifikáns különbséget. Másrészt a feladat során nem foglalkoztunk a Welch-próba alkalmazhatósági feltételeivel. Ez ugyanis megköveteli, hogy az eloszlások normálisak legyenek, azonban a CRP eloszlásáról korábban beláttuk, hogy inkább lognormálisnak tekinthető, valamint a kiugró értékek is problémát jelentenek a teszt alkalmazhatóságát illetően. Emiatt érdemes lehet a Welch-próba helyett más statisztikai eszköz használata a kérdés eldöntéséhez. Érdekességképpen elvégeztünk egy Mann-Whitney-U próbát a CRP változón, és így a korrigálatlan p érték $<10^{-6}$ -nak adódott, így feltételezésünket igazoltuk.

11. feladat: az elhízásban megjelenő laborparaméter-

változások fiziológiai okai

Bónuszfeladat: van ötlete, hogy a fenti eltérések mivel magyarázhatóak kórélettani szempontból? (Több különböző jelenség van a háttérben, de már azt is értékeljük, ha csak egyre sikerül megfelelő választ adni!) Itt egyrészt arra kell tekintettel lenni, hogy mely változókban van egyáltalán szignifikáns eltérés (10. feladat), illetve, hogy az eltérés milyen irányú és nagyságú (2.-7. feladatok).

Az elhízás – a testmagassághoz képesti nagyobb testtömeg – az emberi szervezetben a zsírszövet felszaporodását jelenti, amely építőkövei a lipidek – trigliceridek – ezért a trigliceridszint emelkedése az elhízás egy kísérőjele. A trigliceridek metabolizmusában szerepet kap a koleszterin is, de a koleszterinek jelentős szerepet kapnak a lipidszállítás kapcsán is, valamint a zsírszövet hormontermelése is befolyásolja a koleszterin metabolizmust és így a koleszterinszintet. A trigliceridek szállításában kulcsszerepe van a szérumalbuminnak, amely a vér teljes fehérjetartalmának jelentős részét teszi ki – így e paraméterek csökkent volta az elhízottaknál ezen keresztül összetetten magyarázható.

Mint azt korábban is említettük (9. oldal) a krónikus gyulladás és az elhízás korrelációját már régóta ismerik, vizsgálják – e két kórkép korrelációjának oka összetett, nem közvetlen hatásokban keresendő. A gyulladás jelzői közé tartoznak a fehérvérsejtek, illetve az általuk fokozottan indukált CRP szint is.

A veseelégtelenség a modernkori kórképek közül kitűnik abban, hogy a túlsúlyosak esetében jobb prognózisú – ezt az ismereteink szerint még nem teljesen tisztázott tény mutathatja az eltérő kreatinin érték is.

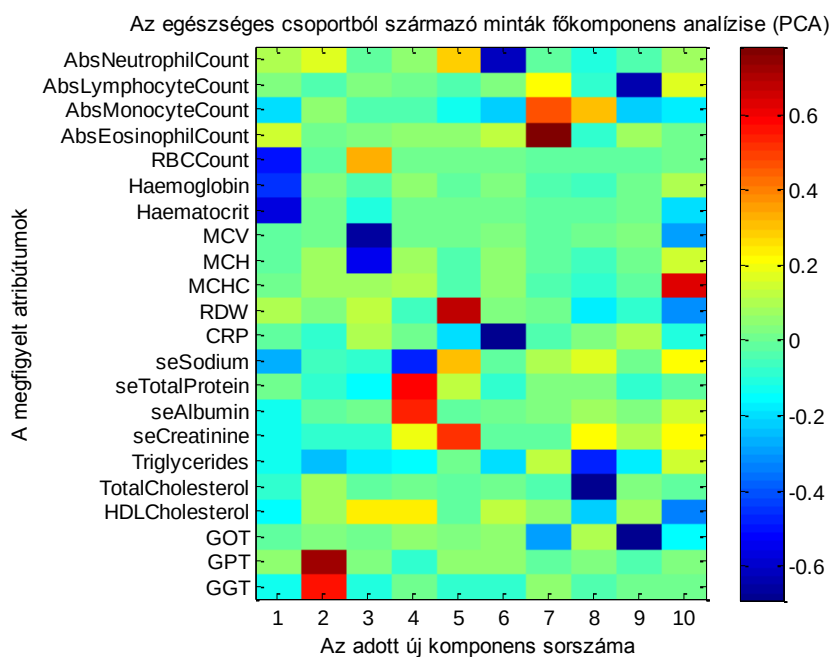
Az elhízás kísérője a nagyobb testtömeg, amely jelentősebb metabolikus igényt is jelent a szervezet számára, amely nagyobb oxigénfogyasztást jelent, ebből kifolyólag nagyobb RBC és ehhez kapcsolódó MCV, MCH, haemoglobin, haematokrit értékeket várnánk, de nem ezt kapjuk, ugyanis ezen értékeket relatívan, a teljes vérmennyiséghez mérten kell tekintenünk.

12. feladat: a változók összefüggéseinek vizsgálata

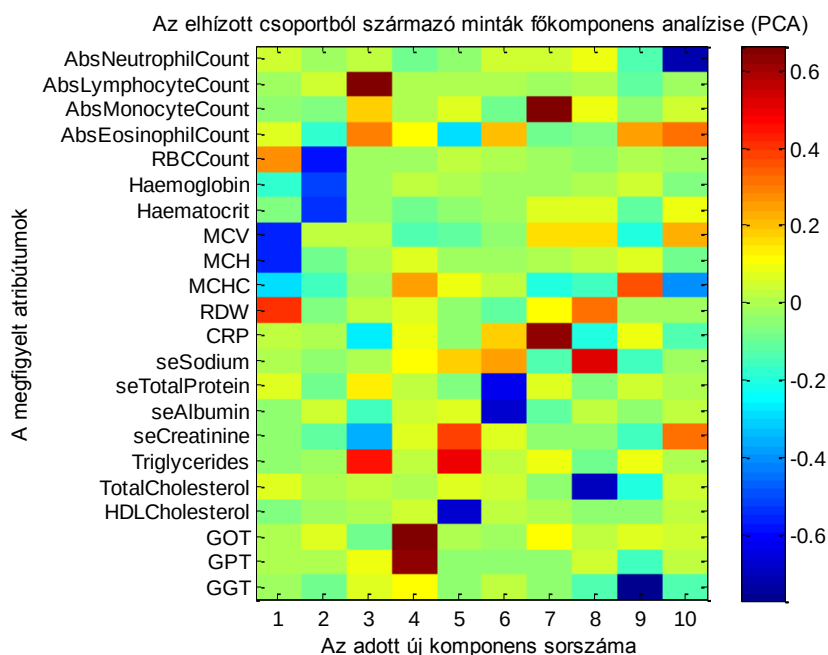
főkomponens elemzés segítségével

Az egyes változók összefüggéseinek vizsgálatára (természetesen ezúttal is az összes változót használva) végezzen főkomponens elemzést (princomp) a csoportokon (elhízott/ egészséges) belül! (Ne feledkezzen meg az adatmátrix standardizálásáról sem (zscore), ezt a princomp nem végzi el automatikusan!) Mindkét esetben 10 főkomponenst használjon, majd rotálja őket (rotatefactors) varimax-kritérium szerint. Milyen együttmozgó változó-csoportok azonosíthatóak (e vizsgálat alkalmas végeredményét felhasználva)? Az eredményt igyekezzon szemmel gyorsan áttekinthető, látványos módon vizualizálni! (Tipp: a heatmap alkalmazása (image) célravezető lehet. A színek jelentését a colorbar paranccsal rajzoltathatja ki.)

A főkomponens elemzést a feladatkiírásban szereplő javaslatoknak megfelelően végeztük el, és hő térképszerűen ábrázoltuk a kapott eredményeket.



10. ábra Egészséges csoportból származó minták főkomponens analízise (PCA)



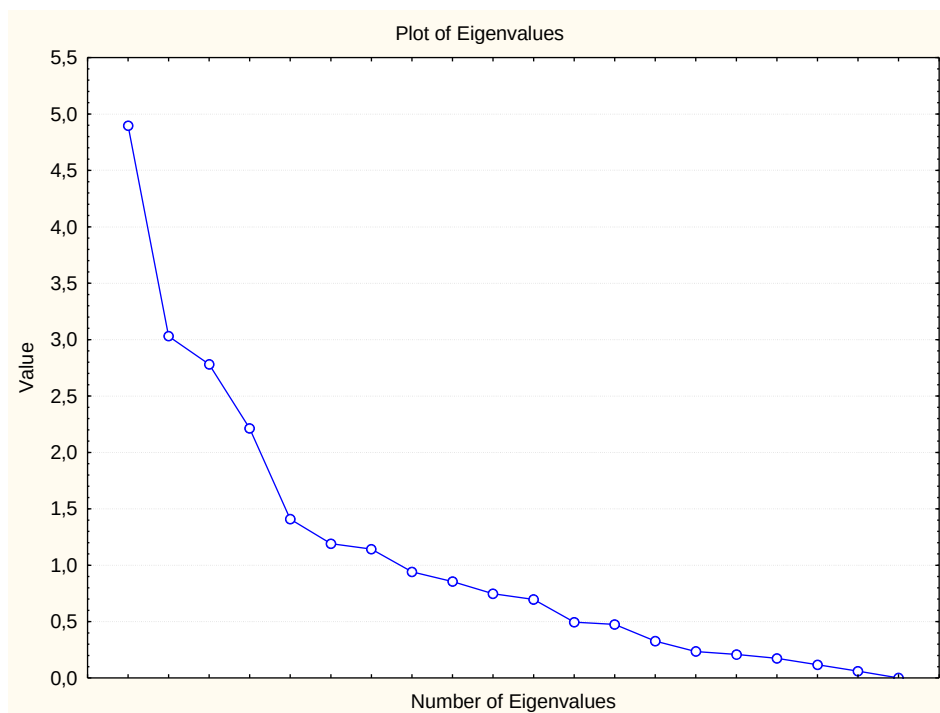
11. ábra Elhízott csoportból származó minták főkomponens analízise (PCA)

Ezen kívül további vizsgálatokat végeztünk a Statistica programmal a komponensek vizsgálatára felhasználva az órai jegyzet néhány pontját. Vizsgáltuk például, hogy hány komponenst érdemes használni – mennyire csökkentjük az adathalmaz dimenzióját úgy, hogy a mintabeli „változékonyságot”, reprodukálhatóságot megtartsuk – a Kaiser-kritériumnak megfelelően a következő komponensek adnak 1-nél nagyobb sajátértéket (eigenvalue) és felelnek a mintában található „változékonyság” adott mértékéért. Ábrázoltuk továbbá a scree-plotot is.

Az egészséges csoportban kapott eredmények:

Egészséges csoport PCA főkomponenseinek sajátértékei				
	Eigenvalue	% Total - variance	Cumulative - Eigenvalue	Cumulative - %
1	4,896546	22,25703	4,89655	22,25703
2	3,029042	13,76837	7,92559	36,02540
3	2,780473	12,63851	10,70606	48,66391
4	2,213007	10,05912	12,91907	58,72303
5	1,408907	6,40412	14,32798	65,12716
6	1,191457	5,41571	15,51943	70,54287
7	1,142684	5,19402	16,66212	75,73689

2. táblázat Egészséges csoport PCA főkomponenseinek sajátértékei



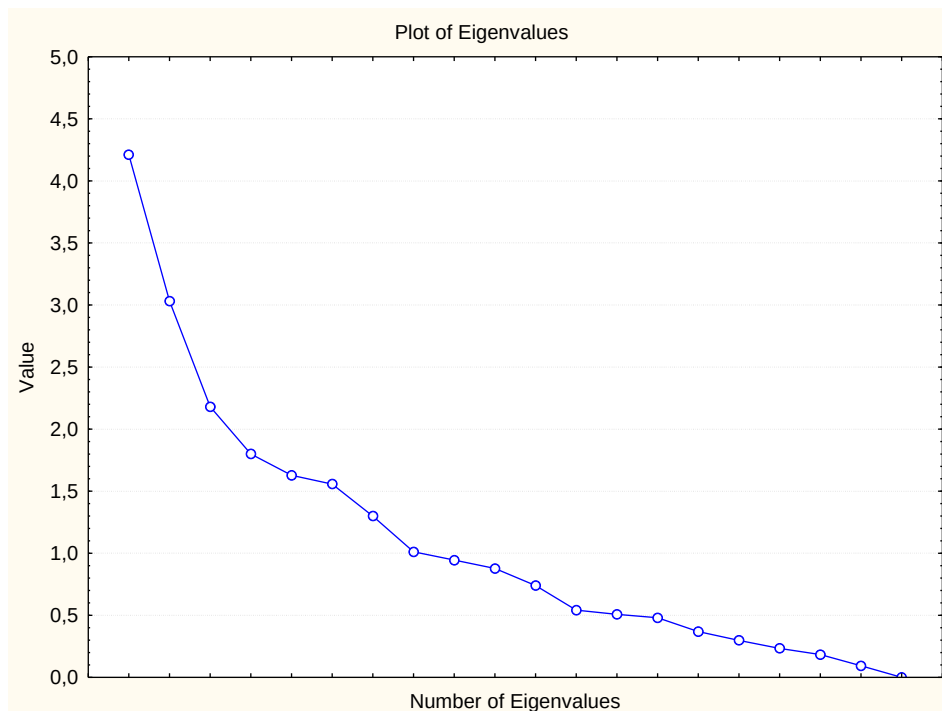
12. ábra Egészséges csoport főkomponenseinek sajátértékei

A Kaiser-kritérium alapján 7 főkomponenst érdemes használnunk, a scree-ploton is itt kapunk egy törést, illetve ez alapján megfontolandó akár csak 5 komponens használata is.

Az elhízott csoportban kapott eredmények:

Elhízott csoport PCA főkomponenseinek sajátértékei				
	Eigenvalue	% Total - variance	Cumulative - Eigenvalue	Cumulative - %
1	4,212782	19,14901	4,21278	19,14901
2	3,032320	13,78327	7,24510	32,93228
3	2,180659	9,91209	9,42576	42,84437
4	1,800302	8,18319	11,22606	51,02756
5	1,628392	7,40178	12,85446	58,42935
6	1,557162	7,07801	14,41162	65,50736
7	1,299021	5,90464	15,71064	71,41200
8	1,012784	4,60356	16,72342	76,01556

3. táblázat Elhízott csoport PCA főkomponenseinek sajátértékei



13. ábra Elhízott csoport főkomponenseinek sajátértékei

A Kaiser-kritérium alapján 8 főkomponenst érdemes használnunk az elhízottak esetében, a scree-ploton is kapunk itt egy törést, de 4 és 12 komponensnél is.

Az egészséges populációban az első főkomponensben az RBC szám, a Haemoglobin és a Haematokrit kapja a fő szerepet, a második komponensben a GOT, GPT, GGT és az abszolút limfocitaszám. A harmadikban az MCV, MCH, a negyedikben a szérumszfehérje és a szérumalbumin, az ötödikben az abszolút neutrofilszám és a CRP (és valamennyire az abszolút monocitaszám), a hatodikban az RDW, amíg a hetedikben pedig az abszolút eosinofilszám. Ezeket a kapcsolatokat láthatjuk akkor is, ha elkészítjük a változók korrelációs mátrixát. (Ennek mérete nagyobb annál, hogy dokumentumunkba beillesszük, így ezt egy excel fájlként csatoljuk a két csoport korrelációs mátrixait.)

Az elhízott populációban a hasonlóan kerülnek egy komponensbe a változók, de a komponensek ilyen sorrendje kissé eltér, amint az látható a hőtésképeken. Ezen kívül egy lényeges különbség akad, mégpedig az, hogy a GOT és GPT mellett nem jelenik meg a GGT – ezt a korrelációs mátrixot megtekintve is láthatjuk.

13. feladat: az adott csoportban szereplő együttmozgó

változók élettani tartalma

Bónuszfeladat: milyen élettani tartalommal tölthetők meg a fenti csoportok, mi magyarázza együttmozgásukat?

1. RBC szám, Haematokrit és Haemoglobin

A felsorolt változók korrelációja élettani szempontból nem meglepő, hiszen a haematokrit a vörösvértestek százalékos térfogata a vérben: nagyobb vörösvértestszám nagyobb térfogatot jelent; a haemoglobin pedig a vér vastartalmát mutatja: nagyobb vörösvértestszám nagyobb vaskoncentrációt is jelent.

2. GOT, GPT, GGT és az abszolút limfocitaszám

A transzaminázok (GOT, GPT) és a transzferáz (GGT) mind a májfunkciót jellemzik (ún. májfunkciós enzimek), de a GGT szint ezen kívül epeúti elzáródásokban, gyulladásos kórképekben más mértékben változik – ez előbbi okozhatja, hogy az elhízottak esetében ennek korreláció kisebb a transzaminázokkal. A máj, mint a legjelentősebb tagja a reticuloendotheliális-rendszernek (RES), felel részben a limfociták számáért, ezért nem meglepő a májfunkciós enzimekkel való korrelációja.

3. MCV, MCH

Az MCV az átlagos vörösvértest-térfogatot, az MCH a vörösvértestek haemoglobin tartalmát jelzi, a korreláció ezért nem meglepő: több haemoglobin nagyobb térfogatot eredményez a vörösvértestekben.

4. CRP, abszolút neutrofilek száma, monociták száma

Ez a két laborparaméter jelzi leginkább a szervezetben létrejövő gyulladást. A CRP szintézisének fokozódása a májban nagyrészt fagocitális indukció hatására valósul meg – a fagociták egyik fajtája pedig a neutrofil granulocita, illetve a monocita.

14. feladat: a változók összefüggéseinek vizsgálata

főkomponens elemzés segítségével

Értelmezésünk szerint a 12. és a 14. feladat azonos, ezért ez utóbbit nem végeztük el. Amennyiben tévedtünk, és a 14. feladat eltér a korábbtól, várjuk a visszajelzést.

Zárszó

A házi feladat során elsajátítottuk a MATLAB alapvető statisztikai szolgáltatásainak használatát, valamint mélyítettük ismereteinket az általános MATLAB programozásban.

A kapott adatbázis mintái azonban nem csak erre adtak lehetőséget, hanem köszönhetően valós voltak, általuk több biostatistikai probléma is körüljárható volt (pl. outlierok kezelése), amelyek a tankönyvi példákból gyakran kimaradnak. Ráadásul egy valódi adatbázis feldolgozásakor sokkal nagyobb a motiváció az adatsorban kirajzolódó mintázatok háttérének feltárására, hisz várható, hogy ezen keresztül tényleges ismeretek nyerhetők. Ennek köszönhetően nem csak statisztikai, de élettani ismereteinket is bővíthettük.

Függelék: MATLAB forráskódok

Adatbázis betöltése

```
% import database file
db = importdata('ObesityDatabase.csv');

% precalc column indices that we're interested in
colnames = db.colheaders;
col_source = find(strcmp(colnames, 'Source'));
col_rbc = find(strcmp(colnames, 'RBCCount'));
col_crp = find(strcmp(colnames, 'CRP'));

% split data into normal/obese categories
group_normal = db.data((db.data(:,col_source) == 0),:);
group_obese = db.data((db.data(:,col_source) == 1),:);

% extract frequently needed data
rbc_normal = group_normal(:, col_rbc);
rbc_obese = group_obese(:, col_rbc);
crp_normal = group_normal(:, col_crp);
crp_obese = group_obese(:, col_crp);
```

Változók általános deskriptív statisztikai jellemzése

```
header = cellstr(char('Laboreredmény', 'Egészséges', 'Elhízott'))';
a = mean(group_normal);
b = mean(group_obese);
disp('Átlag');
disp('-----');
disp([header; colnames(2:end)' num2cell(a(2:end)) num2cell(b(2:end))]);

a = median(group_normal);
b = median(group_obese);
disp('Medián');
disp('-----');
disp([header; colnames(2:end)' num2cell(a(2:end)) num2cell(b(2:end))]);

a = std(group_normal);
b = std(group_obese);
disp('Szórás');
disp('-----');
disp([header; colnames(2:end)' num2cell(a(2:end)) num2cell(b(2:end))]);
```

RBC és CRP mutatók részletes deskriptív jellemzése

```
header = cellstr(char('Mutató', 'Egészséges', 'Elhízott'))';
rownames = cellstr(char('Átlag', '5% trimmelt átlag', 'Medián', 'Alsó kvartilis',
'Felső kvartilis', 'Minimum', 'Maximum', 'Terjedelem', 'Korr. szórás',
'Interkvartilis terjedelem', 'Átskálázott MAD'));

disp('RBC mutatók');
disp('-----');
data1 = rbc_normal;
```

```

data2 = rbc_obese;
measures = [
    mean(data1) mean(data2);
    trimmean(data1, 5) trimmean(data2, 5);
    median(data1) median(data2);
    quantile(data1, 0.25) quantile(data2, 0.25);
    quantile(data1, 0.75) quantile(data2, 0.75);
    min(data1) min(data2);
    max(data1) max(data2);
    range(data1) range(data2);
    std(data1) std(data2);
    iqr(data1) iqr(data2);
    1.4785 * mad(data1, 1) 1.4785 * mad(data2, 1);
];
disp([header; rownames num2cell(measures)]);

disp('CRP mutatók');
disp('-----');
data1 = crp_normal;
data2 = crp_obese;
measures = [
    mean(data1) mean(data2);
    trimmean(data1, 5) trimmean(data2, 5);
    median(data1) median(data2);
    quantile(data1, 0.25) quantile(data2, 0.25);
    quantile(data1, 0.75) quantile(data2, 0.75);
    min(data1) min(data2);
    max(data1) max(data2);
    range(data1) range(data2);
    std(data1) std(data2);
    iqr(data1) iqr(data2);
    1.4785 * mad(data1, 1) 1.4785 * mad(data2, 1);
];
disp([header; rownames num2cell(measures)]);

```

RBC és CRP mutatók hisztogramjának ábrázolása,

elosztásillesztés

```

str_normal = 'egészséges';
str_obese = 'elhízott';

% histograms
Histogram('RBCCount', rbc_normal, str_normal, rbc_obese, str_obese, 10, 'normal');
Histogram('RBCCount', rbc_normal, str_normal, rbc_obese, str_obese, 15, 'normal');
Histogram('RBCCount', rbc_normal, str_normal, rbc_obese, str_obese, 'Sturges',
'normal');
Histogram('RBCCount', rbc_normal, str_normal, rbc_obese, str_obese, 'sqrt',
'normal');
Histogram('RBCCount', rbc_normal, str_normal, rbc_obese, str_obese, 'Freedman',
'normal');

Histogram('CRP', crp_normal, str_normal, crp_obese, str_obese, 10, 'normal');
Histogram('CRP', crp_normal, str_normal, crp_obese, str_obese, 15, 'normal');
Histogram('CRP', crp_normal, str_normal, crp_obese, str_obese, 'Sturges',
'normal');
Histogram('CRP', crp_normal, str_normal, crp_obese, str_obese, 'sqrt', 'normal');

```



```
Histogram('CRP', crp_normal, str_normal, crp_obese, str_obese, 'Freedman',
'normal');
```

```
Histogram('CRP', crp_normal, str_normal, crp_obese, str_obese, 'Freedman',
'lognormal');
```

RBC és CRP változók sűrűségfüggvényének magfüggvényes

becslése

```
% kernel density plots
KSDensityPlot('RBCCount', rbc_normal, str_normal, rbc_obese, str_obese);
KSDensityPlot('CRP', crp_normal, str_normal, crp_obese, str_obese, [-5 30 0 1.1]);
% manual adjustment of axes based on the results
```

RBC és CRP változók sodrófadiagramja (box plot)

```
% boxplots
BoxPlot('RBCCount', rbc_normal, str_normal, rbc_obese, str_obese, [4.4 6.2 0.6
1.4]); % manual adjustment of axes based on the results
BoxPlot('CRP', crp_normal, str_normal, crp_obese, str_obese, [-5 85 0.6 1.4]);
% manual adjustment of axes based on the results
```

Változók eloszlásának ellenőrzése

```
% QQ-plots
QQPlot('RBCCount', rbc_normal, str_normal, rbc_obese, str_obese);
QQPlot('CRP', crp_normal, str_normal, crp_obese, str_obese);
```

Eloszlások normalitásának ellenőrzése

```
% Lilliefors` test of RBC for normality
disp('Az RBC változó normális eloszlásának ellenőrzése (Lilliefors):');
[h, p] = lillietest(rbc_normal);
fprintf(1, ' egészséges csoport: [h, p] = [%i, %.4f]\n', h, p);
[h, p] = lillietest(rbc_obese);
fprintf(1, ' elhízott csoport: [h, p] = [%i, %.4f]\n\n', h, p);

% Jarque-Bera test of RBC for normality
disp('Az RBC változó normális eloszlásának ellenőrzése (Jarque-Bera):');
[h, p] = jbstest(rbc_normal);
fprintf(1, ' egészséges csoport: [h, p] = [%i, %.4f]\n', h, p);
[h, p] = jbstest(rbc_obese);
fprintf(1, ' elhízott csoport: [h, p] = [%i, %.4f]\n\n', h, p);
```

```

% Lilliefors` test of CRP for lognormality
disp('Az CRP változó lognormális eloszlásának ellenőrzése (Lilliefors):');
[h, p] = lillietest(log(crp_normal));
fprintf(1, ' egészséges csoport: [h, p] = [%i, %.4f]\n', h, p);
[h, p] = lillietest(log(crp_obese));
fprintf(1, ' elhízott csoport: [h, p] = [%i, %.4f]\n\n', h, p);

% Jarque-Bera test of CRP for lognormality
disp('Az CRP változó lognormális eloszlásának ellenőrzése (Jarque-Bera):');
[h, p] = jbtest(log(crp_normal));
fprintf(1, ' egészséges csoport: [h, p] = [%i, %.4f]\n', h, p);
[h, p] = jbtest(log(crp_obese));
fprintf(1, ' elhízott csoport: [h, p] = [%i, %.4f]\n\n', h, p);

% Lilliefors` test of CRP for normality
disp('Az CRP változó normális eloszlásának ellenőrzése (Lilliefors):');
[h, p] = lillietest(crp_normal);
fprintf(1, ' egészséges csoport: [h, p] = [%i, %.4f]\n', h, p);
[h, p] = lillietest(crp_obese);
fprintf(1, ' elhízott csoport: [h, p] = [%i, %.4f]\n\n', h, p);

% Jarque-Bera test of CRP for normality
disp('Az CRP változó normális eloszlásának ellenőrzése (Jarque-Bera):');
[h, p] = jbtest(crp_normal);
fprintf(1, ' egészséges csoport: [h, p] = [%i, %.4f]\n', h, p);
[h, p] = jbtest(crp_obese);
fprintf(1, ' elhízott csoport: [h, p] = [%i, %.4f]\n\n', h, p);

```

Eltérések vizsgálata az egészséges és beteg populáció

között /Welch-próba/

```

% apply Bonferroni correction on the 5% significance level
% we won't compare the source and number columns
alpha = 0.05 / (size(group_normal, 2))

% perform a Welch's test (two-sample t-test with unequal variances)
[h p] = ttest2(group_normal(:, 3:end), group_obese(:, 3:end), alpha, 'both',
'unequal');

disp('Welch-próba');
disp('-----');
header = cellstr(char('Változó', 'h', 'p'))';
disp([header; colnames(3:end)' num2cell(h') num2cell(p')]);

```

A változók összefüggéseinek vizsgálata főkomponens

elemzés segítségével

```

% standardise the data matrix, then run PCA
[coeff_normal, score_normal] = princomp(zscore(group_normal(:, 3:end)));

```

```

[coeff_obese,score_obese] = princomp(zscore(group_obese(:, 3:end)));

rot_normal = rotatefactors(coeff_normal(:,1:10));
rot_obese = rotatefactors(coeff_obese(:,1:10));

figure();
imagesc(rot_normal);
colorbar;
xlabel('Az adott új komponens sorszáma');
ylabel('A megfigyelt atribútumok');
title('Az egészséges csoportból származó minták főkomponens analízise (PCA)');
set(gca,'Ytick', 1:(size(rot_normal,1)));
set(gca,'Xtick', 1:(size(rot_normal,2)));
set(gca,'YTickLabel',colnames(3:end));

figure();
imagesc(rot_obese);
colorbar;
xlabel('Az adott új komponens sorszáma');
ylabel('A megfigyelt atribútumok');
title('Az elhízott csoportból származó minták főkomponens analízise (PCA)');
%set(gca,'Ytick', 1:(size(rot_obese,1)));
%set(gca,'Xtick', 1:(size(rot_obese,2)));
%set(gca,'YTickLabel',colnames(3:end));

```