Matematikai statisztika gyakorlat

Programtervező informatikus alapszak, A szakirány

2018/2019 tavaszi félév

Játékszabályok

- Az előadás és a gyakorlat számonkérése közös. Az előadásról és a hozzá tartozó konzultációról további információkat Arató Miklóstól lehet szerezni.
- A gyakorlatokról maximum 3-szor lehet hiányozni. Aki többször hiányzik, nem kaphat jegyet.
- 186 pontot lehet szerezni a félév során:
  - $2 \cdot 50$ pont: 90 perces "nagy" ZH-k: március 26. és május 14., 8:30-10:00 É 0.79 Jánossy Lajos terem
  - $3 \cdot 12$ pont: 15 perces röpZH-k: március 12., április 16. és május 7., 9:00-9:15 É 0.79 Jánossy Lajos terem
  - 50 pont: *önálló* beadandó feladat
- Mindkét "nagy" ZH-n minimálisan el kell érni 15 pontot.
- Az egyik "nagy" ZH pontszámán lehet javítani.
- A ZH-k az előadások és a gyakorlatok tananyagát egyaránt számon kérik.
- A beadandóról:
  - *önálló* statisztikai elemzés;
  - legalább 20 pontot el kell érni;
  - március 14-ig mindenki választ magának egy megfelelő adatbázist;
  - beadási határidő: május 20.;
  - ha előbb beküldöd, az oktató visszajelzése (pár nap) alapján javíthatsz az elemzésen és a végső határidőig újra beadhatod.

Infók a gyakvezetőről

Név	Varga László, *óraadó*
Munkahely	Morgan Stanley, Risk Management
Tanszék	Valószínűségelméleti és Statisztika Tanszék (ELTE TTK)
E-mail	`vargal4@cs.elte.hu`
Honlap	`vargal4.elte.hu`

Kötelező irodalom

- az előadás anyaga: `http://amiklos.web.elte.hu/Oktatas/2019_inf_stat/Matstat.htm`
- a gyakorlaton megoldott feladatok

Ajánlott irodalom

- Molnárné-Tóthné: Általános statisztika példatár I.
- Móri-Szeidl-Zempléni: Matematikai statisztikai feladatok

Az órán használt szoftver/programnyelv: **R**

- Statisztikai modellezésre, data science-re kiváló
- Nyílt forráskódú, minden fontos problémára van library/package
- Letöltési helye: `https://cran.r-project.org/`
- Szövegszerkesztésre ajánlott szoftver: RStudio; letöltési helye: `https://www.rstudio.com/products/rstudio/download/`

1.) A 2011. évi népszámlálás alapján a 20-24 év közötti népesség nemek szerinti megoszlása (Forrás: `http://www.ksh.hu/nepszamlalas/tablak_demografia`):

Nem	Népesség száma (fő)
Férfi	317 039
Nő	301 196
Összesen	618 235

a.) Add meg a táblázat adataiból számítható viszonyszámokat!

b.) A 2016-os Mikrocenzus szerint Magyarország népessége 9 803 837 fő. Számítsd ki a népsűrűséget! Ez milyen viszonyszám?

2.) Az euró eladási árfolyamának alakulása az K&H Banknál a következő volt:

Időpont	Árfolyam (Ft/euró)
2018. február 8.	318,33
2019. február 8.	327,80

Adj meg és értelmezz a táblázat adataiból számítható dinamikus viszonyszámot!

3.) Egy termelő vállalatnál a fizikai munkát végzők összesen 18000 db alkatrészt állítottak elő, amiből a nők teljesítménye 8500 db volt. A vállalatnak 950 férfi fizikai dolgozója van. A nőknél a termelékenység, azaz az egy főre jutó termelt mennyiség 17 db/fő.

a.) Milyen viszonyszám található a feladat szövegében és hogyan számoljuk?

b.) Szerkessz statisztikai táblát az adatokból és töltsd ki a hiányzó rubrikákat!

4.) Néhány információ az ELTE matematika alapszakjára 2016-ban jelentkezőkről: az állami finanszírozásos képzésre 348-an jelentkeztek, 36,494%-uk első helyen jelentkezett, végül 110-et vettek fel. A költségtérítéses képzési formára jelentkezők 10,227%-át, 9 főt vették fel. Összesen 141 ember jelölte be az ELTE matematika szakát első helyen.

a.) Milyen viszonyszám(ok) található(k) a feladat szövegében?

b.) Szerkessz statisztikai táblát az adatokból és töltsd ki a hiányzó rubrikákat!

5.) Egy vállalat négy részleggel rendelkezik, az ott dolgozók bruttó fizetéséről az alábbi adatok állnak rendelkezésünkre:

Részleg	Átlagkereset (e Ft/fő)	Dolgozók létszáma (fő)
Raktár	200	10
Összeszerelő	250	16
Tampóműhely	250	8
Irodaház	300	10
Összesen	...	...

a.) Milyen viszonyszám található a táblázatban és hogyan számoljuk?
b.) Számítsd ki a hiányzó pontozott értékeket!

6.) Egy szálloda 2016-os vendégforgalmáról az alábbiakat ismerjük:

Származási ország szerint a vendég	Vendégéjszakák száma (éj)	Egy vendégéjszakára jutó szállás díja (Ft/éj)	Egy vendégre jutó vendégéjszakák száma (éj/fő)
Belföldi	5000	16000	4
Külföldi	4000	12000	2
Összesen	9000	...	...

Határozd meg a teljes hotelre vonatkozóan az egy vendégéjszakára jutó szállás díjat és az egy vendégre jutó vendégéjszakák számát!

7.) Magyarország népességéről az alábbiakat ismerjük:

Település jellege	Népesség megoszlása 2012-ben (%)	Népesség változása 1990-ről 2012-re (%)
Budapest	17,4	-14,4
Többi város	51,9	-2,4
Községek	30,7	-0,8
Összesen	100,0	...

a.) 1990 és 2012 között évente átlagosan mennyivel változott a budapesti lakosság?
b.) Hány százalékkal változott a népesség száma 1990-ről 2012-re?
c.) Melyik településen élők részaránya csökkent?

8.) Egy szabályos dobókockával 4-szer dobtunk és a következőket kaptuk: 1, 3, 6, 1.
a.) Számold ki a mintaátlagot, tapasztalati szórást és korrigált tapasztalati szórást, a szórási együtthatót (a korrigált szórást használva), valamint a második tapasztalati momentumot!
b.) Számítsd ki és rajzold fel a tapasztalati eloszlásfüggvényt! Mennyi a tapasztalati eloszlásfüggvény értéke a $2, 3, 4$ helyeken?
c.) Mi a kockadobás elméleti eloszlásfüggvénye? Ábrázold ezt a függvényt!
d.) A `floor(runif(100, min = 1, max = 7))` utasítással generálj 100 kockadobást és ábrázold a tapasztalati eloszlásfüggvényét! Mit tapasztalsz?
e.) Tekintsük a kockadobás értékek 100-zal való eltolását: 101, 103, 106, 101. Mennyi lesz most a mintaátlag és a tapasztalati szórás?
f.) Az a.)-pontbeli adatokat szorozzuk meg -3-mal: $-3; -9; 0; -3$. Hogyan változik ekkor a mintaátlag és a tapasztalati szórás?

9.) Egy csoportban a hallgatók magassága (cm):
180 163 1500 157 165 165 174 191 172 165 1-68 186
a.) Nézzük át nagy vonalakban az adatokat, reálisak-e! Próbáljuk javítani az esetleges adathibákat!
b.) Határozd meg a rendezett mintát!
c.) Rajzold fel a tapasztalati eloszlásfüggvényt! Mennyi a tapasztalati eloszlásfüggvény értéke a 180 helyen? Értelmezd szövegesen!
d.) Elemezd a hallgatók testmagasságát alapstatisztikák: átlag, korrigált tapasztalati szórás, szórási együttható, kvartilisek, terjedelem, interkvartilis terjedelem, tapasztalati ferdeség, tapasztalati csúcsosság segítségével! Értelmezd szövegesen az eredményeket!
e.) Készíts boxplot ábrát!
f.) Készíts alkalmas osztályközös gyakorisági sort, majd abból hisztogramot!

10.) Elemezd az alábbi adatokat az előző feladat elemzési szempontjai alapján:
a.) A honlapomon található `Nyarhom.Rdata` nevű fájl a 2014. nyári napi maximum-hőmérsékleteket tartalmazza egy településen (°C)
b.) Minta futási időkből: mérd meg 1000 alkalommal, hogy az R milyen gyorsan generál és rendez egy 10^4 elemű standard normális mintát! Javasolt a `microbenchmark` package használata a futási idő mérésére.
A mintából készíts hisztogramokat különböző sávszélesség esetén! Melyiket tartod a "legjobbnak"?

11.) Legyen `adat=c(2,0,1,0,8,3,5,7,8,2,3,5,1,7,8,3,5,3,2,8)`. Mit számol az alábbi R program?
a.) `sum(adat<3)`
b.) `names(table(adat))[table(adat)==max(table(adat))]`
c.) `sd(adat)== sqrt(sum((adat-mean(adat))^ 2)/(length(adat)))`
TRUE vagy FALSE? Amennyiben hamis az állítás, hogyan lehet igazzá tenni?
d.)
```
rep=rep(c("A","B"),c(10,10))
df = cbind(as.data.frame(adat),as.data.frame(rep))
library(ggplot2)
ggplot(df, aes(x = rep, y = adat)) +
      geom_boxplot(fill = "gold") +
      scale_x_discrete(name = "A és B csoport")
```

12.) Határozzuk meg a mintateret a következő esetekben:
a.) Egy dobókocka háromszori feldobása.
b.) Egy diák felkelési időpontjait jegyzik fel 20 napon keresztül.
c.) Három pénzérmét n-szer dobunk fel.

13.) Legyen $X_1, \ldots, X_n$ független, azonos, abszolút folytonos eloszlású minta, a mintaelemek eloszlásfüggvényét jelölje $F(x)$, a sűrűségfüggvényét pedig $f(x)$. Mutasd meg, hogy a minimum és a maximum sűrűségfüggvénye a következő: $f_{X_1^*}(x) = n \cdot f(x) \cdot (1 - F(x))^{n-1}$ és $f_{X_n^*}(x) = n \cdot f(x) \cdot (F(x))^{n-1}$.

14.) Adjunk torzítatlan becslést at $E(0, \vartheta)$ eloszlás ismeretlen $\vartheta > 0$ paraméterére $T_1(\mathbf{X}) = \overline{X}$ $T_2(\mathbf{X}) = X_n^*$ $T_3(\mathbf{X}) = X_1^*$ statisztikák segítségével. Hasonlítsuk őket össze hatásosság szempontjából!

15.) Próbáljuk **R**-ben meghatározni az előző feladat becsléseit! Generáljunk 100000-

szer 6 elemű $[0,3]$ intervallumon egyenletes eloszlású mintát! Hasonlítsuk össze a becsléseket!

16.) Legyen $X_1, \ldots, X_n$ i.i.d. $\mathrm{Exp}(\lambda)$, $\lambda > 0$ eloszlásból. Torzítatlan becslése az ismeretlen λ paraméternek a $T(\mathbf{X}) = \frac{1}{\sqrt[n]{X_1 \cdot \ldots \cdot X_n}}$ statisztika?

Útmutatás: az integrál kiszámolásához használjuk az Euler-féle gamma-függvényt: $\Gamma(z) = \int\limits_0^\infty x^{z-1} e^{-x}\, dx$

17.) 10-szer választunk egy gép gyártmányai közül. Mindegyik gyártmányról megállapítjuk, hogy selejtes vagy sem. Minket a gépről kikerülő gyártmányok selejtaránya érdekel, amit nem ismerünk.

Modellezzük a problémát a következőképp: legyen $X_1, \ldots, X_{10}$ i.i.d. minta indikátor eloszlásból, ami azt mutatja meg, hogy az egyes gyártmányok selejtesek-e vagy nem. Az $X_1 = 1$ azt az eseményt jelentse, hogy az 1. gyártmány selejtes, ennek ismeretlen valószínűségét pedig jelölje p.

a.) Határozd meg a mintateret és a paraméterteret!

b.) A $T(\mathbf{X}) = \frac{1}{10}(X_1 + \ldots + X_{10})$ statisztika torzítatlanul becsüli a p paramétert?

c.) Keressünk elégséges statisztikát!

18.) Torzítatlan-e a tapasztalati közép reciproka az exponenciális eloszlás paraméterére? Ha nem, hogyan lehet torzítatlanná tenni?

19.) Keressünk elégséges statisztikát a következő eloszláscsaládokból vett n elemű minta esetén, és ahol tudjuk, írjuk fel a kapott elégséges statisztika eloszlását is.

a.) $\mathrm{Bin}(r, p)$, $r \geq 1$ egész ismert, $0 < p < 1$ paraméter,

b.) $\mathrm{Geo}(p)$, $0 < p < 1$ paraméter,

c.) diszkrét egyenletes az $\{1, 2, \ldots, N\}$ halmazon, $N \geq 1$ egész paraméter,

d.) $E(-\vartheta, \vartheta)$, $\vartheta > 0$ paraméter,

e.) $E(\vartheta, 2\vartheta)$, $\vartheta > 0$ paraméter.

20.) Legyen $X_1, \ldots, X_n$ i.i.d. minta $\mathrm{Poi}(\lambda)$, $\lambda > 0$ eloszlásból.

a.) Adjunk hatásos becslést a $g(\lambda) = e^{-\lambda}$ mennyiségre!

b.) Milyen más becsléseket alkalmaznál még?

c.) Szimuláljunk különböző elemszámú és paraméterű Poisson-mintákat, majd vizsgáljuk meg az egyes becslések viselkedését!

d.) Alkalmazd ezt a hatásos becslést a vízi halálos balesetek számára, forrás: `http://www.ksh.hu/docs/hun/xstadat/xstadat_eves/i_ods001.html`

21.) Legyen $X_1, \ldots, X_n$ i.i.d. $\mathrm{Exp}(\lambda)$, $\lambda > 0$ eloszlásból.

a.) Adjunk blackwellizálással jó minőségű torzítatlan becslést a $g(\lambda) = e^{-c\lambda}$ mennyiségre ($c > 0$ konstans)!

b.) Generáljunk **R**-ben 1 paraméterű 10 elemű exponenciális mintákat és próbáljuk megbecsülni a fenti mennyiségeket $c = 0,5; 1; 2; 3; 4$-re. Mekkorák lesznek a hibák?

22.) Tekintsünk egy n elemű i.i.d. Poisson eloszlású mintát.

a.) Adjunk maximum likelihood becslést az ismeretlen paraméterre!

b.) Tegyük fel, hogy a `http://www.ksh.hu/docs/hun/xstadat/xstadat_eves/i_ods001.html` linken található közúti baleseti halálesetek száma Poisson-eloszlást követ. Adjunk becslést az eloszlás paraméterére!

c.) Megfelelőnek tartod ezt az eljárást? Tegyük fel, hogy az i-edik év Poisson paramétere $\mu\rho^{i-1990}$ (tehát a mintaelemek nem azonos eloszlásúak). Becsüld meg μ-t és ρ-t maximum likelihood módszerrel!

d.) Próbálkozz más modellekkel is!

23.) Tekintsünk egy n elemű i.i.d. geometriai eloszlású mintát.

a.) Adjunk maximum likelihood becslést az ismeretlen paraméterre!

b.) Generálj **R**-ben 0.01 paraméterű 200 elemű geometriai eloszlású mintákat! Mit lehet mondani a fenti becslés viselkedéséről?

24.) Legyenek $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ független, nem egyforma paraméterű normális eloszlású minták. A mintákat nem tudjuk megfigyelni, csak az $\epsilon_{ij} = I(X_i < Y_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ indikátor változókat. Hogyan lehetne az eredeti X, Y változók paramétereit becsülni?

25.) A `http://reliawiki.com/index.php/Lognormal_Example_5_Data#Lognormal_Distribution_Examples` címen meghibásodási időket talál. Ezeket gyakran lognormális eloszlással közelítik.

a.) Adjunk a paraméterekre maximum likelihood becslést!

b.) Becsüljük a paramétereket momentum módszerrel is!

c.) Adjunk becslést annak a valószínűségére, hogy az első 700 órában nem történik meghibásodás!

26.) A `http://amiklos.web.elte.hu/Oktatas/2019_inf_stat/felkar.RData` file-ban a Settenkedő Sáskák Biztosító 83 db felelősségbiztosítási kárát láthatjuk millió forintban. A biztosító az ilyen típusú károkat Pareto-eloszlással modellezi. A Pareto-eloszlás sűrűségfüggvénye $\frac{\alpha\beta^\alpha}{(\beta+x)^{\alpha+1}} I(x > 0)$.

a.) Adjunk maximum likelihood becslést az α paraméterre, ha $\beta = 2.5$!

b.) Adjuk meg α momentum módszeres becslését és vessük össze az ML-becsléssel!

c.) Határozzuk meg az előző becsléseket, ha egyik paraméter sem ismert!

27.) A `http://amiklos.web.elte.hu/Oktatas/2019_inf_stat/postagalamb.RData` címen 90 postagalamb visszaérkezési időpontját (napban számolva) találjuk meg. Tegyük fel, hogy a visszaérkezési időpontok exponenciális eloszlást követnek.

a.) Határozzuk meg a paraméter maximum likelihood becslését! Mivel kellene ezt szorozni, hogy torzítatlan becslést kapjunk?

b.) Bizonyítsuk be, hogy az így kapott becslés hatásos!

c.) Határozzuk meg a Fisher-féle információmennyiséget!

d.) Határozzuk meg az információs határt, ha a paramétert becsüljük!

e.) Határozzuk meg az információs határt, ha a paraméter reciprokát becsüljük!

f.) Generáljunk 1000-szer 2 paraméterű 90 elemű mintát. Hasonlítsuk össze a kapott négyzetes hibákat az információs határokkal!

28.) Legyen $X_1 \dots, X_n$ i.i.d. minta $E(0, \vartheta)$ eloszlásból.

a.) Határozd meg a paraméter maximum likelihood becslését! Mivel kellene ezt szorozni, hogy torzítatlan becslést kapjunk?
b.) Határozd meg a Fisher-féle információmennyiséget!
c.) Határozd meg az információs határt, ha a paramétert becsüljük!
d.) Határozd meg az információs határt, ha a paraméter négyzetét becsüljük!
e.) 1000-szer generálj 10 paraméterű 100 elemű mintát. Hasonlítsd össze a kapott négyzetes hibákat az információs határokkal!

29.) Legyen $X_1, \dots, X_n \sim N(m, \sigma^2)$ i.i.d. minta, σ ismert, m ismeretlen. Adjunk m-re α megbízhatóságú szimmetrikus konfidenciaintervallumot !

30.) Tekintsük a 9. feladatban szereplő hallgatói magasságokat, amikről tegyük fel, hogy normális eloszlást követnek.

a.) Adjunk 95%-os megbízhatóságú konfidenciaintervallumot a hallgatók magasságának várható értékére, ha a magasságok szórása 10 cm!
b.) Hány elemű mintára van szükség, ha azt szeretnénk, hogy a konfidenciaintervallum legfeljebb 8 cm hosszúságú legyen?
c.) Adjunk 95%-os megbízhatóságú konfidenciaintervallumot a hallgatók magasságának várható értékére és szórására, ha a magasság szórása ismeretlen!

31.) Tekintsük a 27. feladatban szereplő postagalambos mintát, amiről tegyük fel, hogy elemei függetlenek és exponenciális eloszlásúak.

a.) Adjunk az ismeretlen paraméterre aszimptotikus intervallumbecslést a centrális határeloszlás-tétel segítségével!
b.) Adjunk pont- és intervallumbecslést annak a valószínűségére, hogy egy tubicának 2 óránál kevesebb időre van szüksége a visszaérkezéshez! Hasonlítsuk össze a naiv pontbecsléssel (relatív gyakoriság)!

32.) Tekintsük a 9. feladatban szereplő hallgatói magasságokat, amikről tegyük fel, hogy függetlenek és a $f_\vartheta(x) = \frac{2x}{3\vartheta^2} I(\vartheta < x < 2\vartheta)$ sűrűségfüggvényű eloszlásból származnak, ahol $\vartheta > 0$ ismeretlen valós paraméter.

a.) Adjunk 95%-os megbízhatóságú konfidenciaintervallumot ϑ-ra! Induljunk ki a ϑ ML-becsléséből, majd próbáljunk meg egy alkalmas transzformációval pivotal statisztikát előállítani, ennek segítségével pedig adjuk meg a legszűkebb, az ML-becslést tartalmazó konfidenciaintervallumot!
b.) Adjunk pont- és intervallumbecslést annak a valószínűségére, hogy egy hallgató magasabb 190 cm-nél. Vessük össze a relatív gyakorisággal!

33.) A butitizmus betegségnél a vér kitamin tartalma (ezrelékben) jól közelíthető $N(20; 4)$ eloszlással. A butitizmusban nem szenvedőknél ez az eloszlás $N(18; 1)$. Az orvost felkeresi egy beteg, az a feladatunk, hogy döntést hozzunk: butitizmusban szenved-e, avagy sem.

a.) Határozzunk meg egy 5%-os elsőfajú hibavalószínűségű próbát 1 elemű minta esetén!
b.) Határozzuk meg ennek a próbának a másodfajú hibavalószínűségét!
c.) Végezzünk 100 kísérletet butitista betegekkel! Hányszor döntünk helyesen?
d.) Végezzünk 100 kísérletet butitizmusban nem szenvedőkkel! Hányszor döntünk helyesen?
e.) Oldjuk meg úgy a feladatot, hogy n elemű minta alapján szeretnénk dönteni!

34.) 5-elemű $E(0, \vartheta)$ független mintánk van. A nullhipotézis $H_0 : 0 < \vartheta \leq 10$, az ellenhipotézis pedig $H_1 : \vartheta > 10$. Próbánk a következő: H_0 mellett döntünk, ha a legnagyobb megfigyelésünk kisebb 9-nél, különben az ellenhipotézist választjuk.

a.) Határozzuk meg a próba terjedelmét!
b.) Rajzoljuk fel a próba erőfüggvényét!
c.) 1000-szer generáljuk le a kísérletet $\vartheta = 9.8$ és $\vartheta = 11$ esetén. Mit tapasztalunk?

35.) 24 emberen végeznek emberkísérletet. 3 korsó sört kell meginniuk. 2 korsó Kukutyini APA sört és egy korsó Rézfalvai IPA sört. Mindenkinek rá kell mutatnia az eltérő sörre. Jelölje p annak a valószínűségét, hogy egy kísérleti alany a Rézfalvai APA sört választja ki. A nullhipotézis szerint a sörök megkülönbözhetetlenek, azaz $H_0 : p = \frac{1}{3}$, míg az ellenhipotézis szerint megkülönböztethetők, tehát $H_1 : p > \frac{1}{3}$. Próbánk a következő: elutasítjuk H_0-t, ha legalább y_c kísérleti alany helyesen választotta ki a Rézfalvai IPA sört.

a.) Rajzoljuk fel a helyesen válaszolók eloszlását $p = \frac{1}{3}$ és $p = 0.5$ esetén!
b.) Határozzuk meg a próba elsőfajú hibavalószínűségét $y_c = 12$ és $y_c = 13$ esetén!
c.) Rajzoljuk fel a próba erőfüggvényét a fenti paramétereknél!
d.) 1000-szer generáljuk le a kísérletet $p = \frac{1}{3}$ és $p = 0.5$ esetén. Mit tapasztalunk?

36.) Bublisztánban az ÖDSZ párt vezetőségi tagjainak havi keresete (millió bublikban) jól közelíthető $N(\mu_1, 2^2)$ eloszlással. A többi lakosnál a kereset $N(\mu_2, 4^2)$ eloszlással közelíthető. Rita Tora oknyomozó újságíró kiderítette néhány, a Nagy vezér stadionban szurkoló ember keresetét:

VIP páholyban ülők	20.47	21.10	18.67	16.67	18.00	20.40	22.17
	20.05	24.85	19.93	19.73	20.39		
Normál sorban ülők	4.56	6.67	4.10	11.91	3.89	5.48	3.89
	10.12	5.13	4.24	2.36	0.22		

a.) Amennyiben a VIP páholyban csak az ÖDSZ párt vezetőségi tagjai ülnek, akkor 5%-os elsőfajú hibavalószínűség mellett el tudjuk fogadni a $H_0 : \mu_1 = 20$ hipotézist kétoldali ellenhipotézissel szemben / értelmes egyoldali ellenhipotézissel szemben?
b.) Tekintsük a normál sorban ülőket. 5%-os elsőfajú hibavalószínűség mellett el tudjuk fogadni a $H_0 : \mu_2 = 8$ hipotézist a kétoldali ellenhipotézissel szemben?
c.) Mennyi a p-érték az előző részfeladatnál?

d.) El tudjuk fogadni a $H_0 : \mu_1 = \mu_2$ hipotézist?

A továbbiakban tegyük fel, hogy az ÖDSZ párt vezetőségi tagjainak havi keresete jól közelíthető $N(\mu_1, \sigma_1^2)$ eloszlással, a többi lakosé pedig $N(\mu_2, \sigma_2^2)$ eloszlással.

e.) El tudjuk fogadni a $H_0 : \sigma_1 = \sigma_2$ hipotézist a $H_1 : \sigma_1 < \sigma_2$ ellenhipotézissel szemben?

f.) El tudjuk fogadni a $H_0 : \mu_1 = \mu_2$ hipotézist?

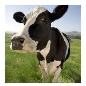
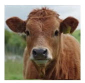
37.) A fogyasztóvédelmi hatóság többszöri lakossági bejelentést kapott, hogy a Portokall nevű, fél literes kiszerelésű narancsitalokban a flakonra írt 500 ml-nél jóval kevesebb üdítő van. Ez alapján vizsgálatot kezdtek, a fogyasztóvédelem munkatársa vásárolt a boltban 10 darabot, majd megnézte a benne lévő édes nedű térfogatát (ml): 483, 502, 498, 496, 502, 483, 494, 491, 505, 486.

a.) Ellenőrizzük Shapiro-Wilk próbával, hogy egy fél literes üdítős üvegbe töltött narancslé mennyisége normális eloszlást követ-e!

b.) Állíthatjuk-e, hogy a narancslé mennyiségének szórása 10?

c.) Állíthatjuk-e, hogy a Portokall gyártója át akarja verni a vevőket?

38.) Bálint gazdának 66 tehene van, teheneit reggel kitereli nagy birtokára, és egész nap ott legelésznek. Este összefut a helyi kocsmában a szomszéd gazdálkodóval, Máté gazdával, aki elmeséli, a tehenei tejének tejzsírszázaléka jelentősen megnőtt, mióta szilázzsal is eteti őket minden nap. Ezen felbuzdulva, Bálint gazda úgy dönt, hogy 6 kedvenc tehenén kipróbálja ezt a "diétát" – egy hónapon keresztül szilázzsal is etette őket, majd megnézte a tejük tejzsírszázalékát:

						
Mit ettek	Julcsa	Bogár	Riska	Csendes	Bimbó	Mula
Csak füvet	3,84	3,79	3,78	4,00	3,83	3,84
Szilázst is	3,90	4,05	3,8	4,01	3,81	3,9

Vizsgáljuk meg alkalmas statisztikai próbával, hogy a silázs növeli-e a tej tejzsírszázalékát!

39.) Egy gyártó megfigyelte, hogy 100, általa előállított SSD merevlemezen 5 év használat után hány hibás szektort talál az ezek felkutatására készített szoftver:

Hibás szektorok száma	0	1	2	3	4	5	7	Összesen
Gyakoriságok	45	35	12	5	1	1	1	100

Vizsgáljuk meg, hogy a szektorhibák száma Poisson-eloszlást követ-e!

40.) Rendelkezésünkre áll a következő minta: 0,55; 0,59; 0,34; 0,69; 0,95; 0,34; 0,53; 0,54; 0,03; 0,11; 0,15; 0,67; 0,48; 0,09; 0,55; 0,02; 0,37; 0,76; 0,83; 0,92.

A megoldás során alkalmazzunk diszkretizálást, azaz képezzünk alkalmas gyakorisági sort az adatokból.

a.) Elfogadhatjuk-e azt a hipotézist, hogy a minta (0,2) intervallumon egyenletes eloszlású? Vizsgáljuk meg Q-Q plot-tal is!

b.) Elfogadhatjuk-e azt a hipotézist, hogy a minta egyenletes eloszlású? Vizsgáljuk meg Q-Q plot-tal is!

c.) Elfogadhatjuk-e azt a hipotézist, hogy a minta exponenciális eloszlású? Vizsgáljuk meg Q-Q plot-tal is!

41.) Az Informatikai Kar III. évfolyamán 300-an tanulnak. Megszámolták, hogy a legutóbbi vizsgaidőszakban hányszor buktak az egyes hallgatók. Az eredményeket tartalmazza az alábbi táblázat.

Bukások száma	0	1	2	3	4
Hallgatók száma	80	113	77	27	3

a.) Elfogadhatjuk-e azt a hipotézist, hogy egy hallgató bukásszáma Bin(4; 0,25) eloszlású?

b.) és azt, hogy Bin(4;p) eloszlású?

42.) Az alábbi kontingencia-táblázat mutatja, hogy 100 évben a csapadék mennyisége és az átlaghőmérséklet hogyan alakult.

Hőmérséklet \ Csapadék	Kevés	Átlagos	Sok
Hűvös	15	10	5
Átlagos	10	10	20
Meleg	5	20	5

(A cellákban az egyes esetek gyakoriságai találhatóak.) Tekinthető-e a csapadékmennyiség és a hőmérséklet függetlennek?

43.) Egy webtervező azt gyanítja, hogy az általa létrehozott internetes vásárlás honlapján a vásárlások mértéke összefügg azzal, hogy milyen nap van a héten. Ennek a sejtésnek az ellenőrzésére egy héten keresztül adatokat gyűjt – összesen 3758 látogatót számlált meg:

Vásárlás	H	K	Sz	Cs	P	Sz	V	Össz.
Nem vásárolt	399	261	284	263	393	531	502	2633
1 vásárlás	119	72	97	51	143	145	150	777
Több vásárlás	39	50	20	15	41	97	86	348
Összesen	557	383	401	329	577	773	738	3758

Alkalmas statisztika próbával döntsünk arról, hogy helyes-e a webtervező sejtése!

44.) Legyenek adottak a következő (x, y) párok:

x_i	0	1	6	5	3
y_i	4	3	0	1	2

a.) Határozzuk meg és ábrázoljuk is az $aX + b$ alakú regressziós egyenest!

b.) Számoljuk ki a reziduálisokat és becsüljük meg a hiba-szórásnégyzetet!

c.) Adjunk előrejelzést $x = 10$-re a regressziós egyenes alapján!

45.) A január 8-án tartott Statisztika II. vizsgát 8 hallgató írta meg, akiktől megkérdeztük, mennyi órát készültek a vizsgára, hány pontot szereztek a tantárgy előfeltételének számító Statisztika I. tantárgyból a vizsgán és milyen magasok:

Statisztika II. pontszám	49	55	56	62	65	70	78	92
Hány órát készült (ó)	15	16	14	13	12	19	21	24
Statisztika I. pontszám	60	50	66	53	67	76	88	87
Testmagasság (cm)	160	174	178	182	173	168	191	167

a.) Vizsgáljuk meg lineáris regresszióval a tanulási idő hatását a Statisztika II. pontszámra! Ábrázoljuk a regressziós egyenest!

b.) Illesszünk négyzetes regressziós függvényt a Statisztika II. pontszámra, ha a magyarázó változó a tanulási idő! Ábrázoljuk a regressziós egyenest!

c.) Illesszünk lineáris regressziót a Statisztika II. pontszámára, ha a magyarázó változók a tanulási idő és a Statisztika I. pontszám!

d.) Illesszünk lineáris regressziót a Statisztika II. pontszámára, ha a magyarázó változók a tanulási idő, a Statisztika I. pontszám és a testmagasság!

e.) Vessük össze a modelleket!

f.) A Statisztika II. vizsga sikeres, ha a hallgató legalább 50 pontot elér. Juli 175 cm magas, a Statisztika I.-ből 60 pontot szerzett és a Statisztika II.-re 18 órát tervez tanulni. Várhatóan át fog menni a Statisztika II. vizsgán?

46.) Olvassuk be a `kerdoiv.txt` fájlt, ami egy 2017-es hallgató kérdőíves felmérés adatait tartalmazza. A következőkre válaszoltak: nem, testmagasság (cm), súly (kg), cipőméret, hányast szerzett valszámból a 2017-es vizsgán, hány percet utazik az egyetemre, szorgalmi időszakban átlagosan hány órát tanul egy héten.

a.) Nézzük meg pontdiagrammal néhány adatpár közti összefüggést (pl. magasság és súly, nem és cipőméret, stb.)!

b.) A továbbiakban célunk a testmagasság modellezése/magyarázása a többi változó segítségével. Tekintsük az alábbi regressziós modelleket:

I.) Testmagasság = Testsúly + Hiba , ami a
Testmagasság $= a_0 + a_1 \cdot$ Testsúly + Hiba modell rövidített változata

II.) Testsúly = Testmagasság + Hiba

III.) Testmagasság = Testsúly + Lábméret + Hiba

IV.) Testmagasság = Nem + Hiba

Vizsgáljuk meg a korrelációs mátrixot! Keressük meg a legjobban illeszkedő modellt!

c.) Adjunk előrejelzést a legjobbnak tűnő modell(ek) alapján egy olyan fiú hallgató testmagasságára, aki 70 kg-os, 45-ös a cipőmérete, 5-öse volt valszámból, 25 percet utazik az egyetemre és heti 12 órát tanul!