Matematikai statisztika gyakorlat

Programtervező informatikus alapszak, A szakirány

2021/2022 őszi félév

Játékszabályok

- Az előadás és a gyakorlat számonkérése közös. Az előadásról és a hozzá tartozó konzultációról további információkat az előadótól, Zempléni Andrástól lehet szerezni (`zempleni.elte.hu`).
- A gyakorlatokról maximum 4-szer lehet hiányozni. Aki többször hiányzik, nem kaphat jegyet. A jelenlét ellenőrzése: egy egyszerű feladat végrehajtása és a végeredmény leírása a Teams-előadás ablakában.
- 200 pontot lehet szerezni a félév során, ebből 150 (+20) pontot a szorgalmi időszakban:
  - $2 \cdot 50 = 100$ pont: ZH-k
    - idő: október 14. és november 25., 8:15–9:15 $\rightsquigarrow$ 60 perces
    - hely: online, Canvas-ban
    - mindkét ZH-n minimálisan el kell érni 30%-ot, azaz 15 pontot
    - 60% kifejtős feladatok, 20% elméletibb jellegű összefüggések számonkérése, 20% R nyelvű számítógépes output-ok kiértékelése
  - 50 pont: önálló statisztikai elemzés
    - *önálló* statisztikai elemzés, részletes feltételek és a feladat/adatbázis kihirdetése később
    - legalább 40%-ot, azaz 20 pontot el kell érni
    - beadási határidő: december 2., 23:59
  - legfeljebb +20 pont: szorgalmi feladatok (SZ jelű példák, fix határidőig)

Infók a gyakvezetőről

Név	Varga László, *óraadó*
Munkahely	Morgan Stanley, Risk Management
Tanszék	Valószínűségelméleti és Statisztika Tanszék (ELTE TTK)
E-mail	`vargala4@gmail.com`
Honlap	`vargal4.elte.hu`

Kötelező irodalom

- az előadás anyaga: `https://zempleni.elte.hu/stata_20_2.html`
- a gyakorlaton feldolgozott elmélet és 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/`
- Kódszerkesztésre ajánlott szoftver: RStudio; letöltési helye: `https://www.rstudio.com/products/rstudio/download/`
- R MarkDown: egyszerű szövegszerkesztő és kódot is futtató package, rövid bevezető: `https://rmarkdown.rstudio.com/lesson-1.html`

1.) 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?

2.) 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!

3.) 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"?

4.) 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")
```

SZ1.) **[IX.23.-ig]** Egy magyarkártya-csomagból visszatevéssel húzunk 4 lapot. R-es szimulációval számold ki, hogy milyen eséllyel húzunk pontosan két zöld színű lapot! Legalább mennyi ismétlésszámot ajánlanál, hogy a valódi valószínűséget legalább 0,5%-os pontossággal közelítsük? Válaszodat a konvergencia sebességét bemutató alkalmas ábrával támaszd alá! (2p)

5.) 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.

6.) 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 mintaelemek minimumának és maximumának sűrűségfüggvénye a következő: $f_{X_1^*}(x) = n \cdot f(x) \cdot (1 - F(x))^{n-1}$ és $f_{X_n^*}(x) = n \cdot f(x) \cdot (F(x))^{n-1}$.

7.) 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.

8.) Próbáljuk **R**-ben meghatározni az előző feladat becsléseit! Generáljunk 100000-szer 6 elemű $[0, 3]$ intervallumon egyenletes eloszlású mintát! Hasonlítsuk össze a becsléseket!

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

10.) Generáljunk R-ben 100000-szer 10 elemű 0,5 paraméterű exponenciális mintát és határozzuk meg, majd vizsgáljuk meg az előző feladatban meghatározott becslést!

11.) [Momentum becsléshez elméleti eredmény] Tegyük fel, hogy a minta kétparaméteres eloszláscsaládból származik, a paraméterek a és b.
Ekkor mutassuk meg, hogy az $\begin{cases} E_{a,b}X &= m_1 \\ E_{a,b}X^2 &= m_2 \end{cases}$ egyenletrendszer megoldása megegyezik az $\begin{cases} E_{a,b}X &= m_1 \\ D^2_{a,b}X &= s_n^2 \end{cases}$ egyenletrendszer megoldásával.

12.) Határozzuk meg az ismeretlen paraméter(ek) maximum likelihood és momentum becslését, ha a minta i.i.d.

a.) Poi(λ) eloszlású;

b.) Exp(λ) eloszlású;

c.) $E(0; \vartheta)$ eloszlású, ahol $0 < \vartheta$ valós paraméter;

d.) $N(m; \sigma^2)$ eloszlású, ahol m valós paraméter, σ ismert.

13.) Generáljunk $n = 10, 50, 100$ elemű mintákat Poi(λ) eloszlásból különböző λ paraméterértékek esetén, majd számoljuk ki a maximum likelihood becsléseket az R egyik beépített optimalizáló rutinja segítségével! Sok-sok szimuláció alapján vizsgáljuk meg, vajon a becslések torzítatlanok-e! Határozzuk meg a becslések szórását szimulációkkal!

14.) [Előadás] Legyen X_1 Bin$(2; p)$ eloszlású (egyelemű) minta, ahol $p \in (0; 1)$ ismeretlen valós paraméter. Adj X_1 segítségével torzítatlan becslést $g(p) = \frac{1}{p}$-re!

15.) [Előadás] Minden nap a Mester utca megállónál szállok fel a 4-es/6-os villamosok valamelyikére. E hét munkanapjain az alábbi várakozási időket mértem (perc):
1,2 2 1,5 3 2,1
A várakozási időről tegyük fel, hogy exponenciális eloszlású.

a.) Adjuk meg a mintateret és a paraméterteret!

b.) Határozzuk meg az ismeretlen paraméter ML-becslését!

c.) Határozzuk meg az ismeretlen paraméter momentum-becslését!

d.) Szimulációval vizsgáljuk meg, hogy 10, 20, 50 és 100 elemű exponenciális mintából számolt ML-becslés torzítatlanul becsüli-e az ismeretlen paramétert!

e.) Torzítatlan, illetve konzisztens az ML-becslés? Amennyiben nem torzítatlan, tegyük azzá!

f.) Mutassuk meg, hogy az $S(\mathbf{X}) = n \cdot X_1^*$ statisztika torzítatlan, de nem konzisztens becslése $g(\vartheta) = \frac{1}{\lambda}$-nak!

16.) Definiáljuk az alábbi diszkrét eloszlást: $P(X = -1) = c$, $P(X = 1) = 3c$, $P(X = 2) = 1 - 4c$, ahol $0 < c < 1/4$ az ismeretlen paraméter. A -1, 1 és 2-es értékekből rendre a következő mennyiségűt kaptuk egy 20 elemű mintában: 4, 10, 6. Határozzuk meg c ML és momentum módszeres becslését!

SZ2.) **[IX.30.-ig]** Legyen $X_1, \ldots, X_n$ i.i.d. Exp(λ) , $\lambda > 0$ eloszlásból. Torzítatlan becslése az ismeretlen λ paraméternek a $T(\mathbf{X}) = \frac{1}{\sqrt[n]{X_1 \cdot \ldots \cdot X_n}}$ statisztika? (2p)
Útmutatás: az integrál kiszámolásához használjuk az Euler-féle gamma-függvényt:
$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

SZ3.) **[X.7.-ig]** Határozzuk meg az ismeretlen paraméter(ek) maximum likelihood és momentum becslését, ha a minta i.i.d. $f_\vartheta(x) = \frac{2x}{\vartheta^2} I(0 \le x \le \vartheta)$ sűrűségfüggvénnyel, ahol ϑ valós paraméter! (2p)

17.) Keressünk elégséges statisztikát a következő eloszláscsaládokból vett n elemű minta esetén:

a.) $E(0,\vartheta)$, $\vartheta > 0$ paraméter;
b.) $Exp(\lambda)$, $\lambda > 0$ paraméter;
c.) Geo(p), $0 < p < 1$ paraméter.

18.) Vegyünk n elemű i.i.d. mintát az Ind(p) eloszlásból.
a.) Határozzuk meg a mintában lévő Fisher-információ értékét!
b.) Számold ki az információs határt! Mutassuk meg, hogy a mintaátlag hatásos becslése p-nek!

19.) Vegyünk n elemű i.i.d. mintát $N(m,\sigma^2)$ eloszlásból, ahol σ ismert. Határozzuk meg a mintában lévő Fisher-információ értékét!

SZ4.) **[X.14.-ig]** Számítsuk ki az n elemű mintában rejlő Fisher-információt, ha a mintaelemek közös sűrűségfüggvénye a következő:
$f_\vartheta(x) = \frac{x}{\vartheta^2} e^{-\frac{x^2}{2\vartheta^2}} I(x>0)$, ahol $\vartheta > 0$ valós paraméter! (2p)

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

21.) Tekintsük a 2. feladatban szereplő hallgatói magasságokat, amikről tegyük fel, hogy normális eloszlást követnek.
a.) Adjunk 95%-os megbízhatóságú szimmetrikus konfidenciaintervallumot a hallgatók magasságának várható értékére, ha a magasságok szórása 10 cm!
b.) Ha a magasságok szórása 10 cm, akkor hány elemű mintára van szükség, ha azt szeretnénk, hogy a szimmetrikus konfidenciaintervallum legfeljebb 8 cm hosszúságú legyen?
c.) Adjunk 95%-os megbízhatóságú szimmetrikus konfidenciaintervallumot a hallgatók magasságának várható értékére és szórására, ha a magasság szórása ismeretlen!

22.) Internetes felmérésen megkérdezték, havonta mennyit keresnek a 2019-ben végzett "A" szakirányos informatikus hallgatók a végzés után 5 évvel (e Ft): 1100, 2000, 500, 900, 300, 1200, 1200, 1100, 400, 500, 300, 200, 5300, 4600, 600, 1800, 2500, 2500, 2100, 1000. Tegyük fel, hogy a fizetések exponenciális eloszlást követnek [később ennek a feltevésnek az ellenőrzésére is tanulunk statisztikai módszert].
a.) Mutassuk meg, hogy az n elemű i.i.d. $Exp(\lambda)$ mintában lévő Fisher-információ $\frac{n}{\lambda^2}$!
b.) Adjunk pont-, majd a Fisher-információ segítségével intervallumbecslést az exponenciális eloszlás ismeretlen paraméterére!
c.) Adjunk pont-, és intervallumbecslés arra, hogy ezen minta alapján vajon egy véletlenszerűen választott informatikus hallgató várhatóan mennyit fog keresni a végzés után 5 évvel!
d.) Adjunk pont-, és intervallumbecslés arra, hogy ezen minta alapján vajon egy véletlenszerűen választott informatikus hallgató milyen eséllyel fog 2 millió felett keresni a végzés után 5 évvel!

SZ5.) **[X.21.-ig]** Adjunk intervallumbecslést az ismeretlen $\vartheta > 0$ paraméterre n elemű i.i.d. minta alapján, ha a mintaelemek közös sűrűségfüggvénye
$f(x) = \begin{cases} \frac{2x}{3\vartheta^2} & \text{ha } \vartheta \le x \le 2\vartheta \\ 0 & \text{egyébként} \end{cases}$.
Útmutatás: érdemes tanulmányozni az E11. (előadás)példa megoldásának menetét, de most induljunk ki az elégséges statisztika helyett az ML-becslésből. (3p)

23.) Valaki azt állítja, hogy a klíma változik, és ezt azzal véli bizonyítottnak, hogy az elmúlt 10 évben 2-szer is volt jégeső május 1-jén, pedig korábban az egyes évekre a jégeső valószínűsége a hivatalos adatok alapján csupán $p = 0.1$ volt május 1-jére. Írjuk fel a hipotéziseket, a próbát és állapítsuk meg az elsőfajú hiba valószínűségét, valamint az erőfüggvényt a $p = 0.2$ pontban!

24.) Az alábbi minta 4 év október 18-án Budapesten mért napi középhőmérséklet adatait tartalmazza (°C fok): 14.8, 12.2, 16.8, 11.1. Tegyük fel, hogy az adatok normális eloszlásból származnak valamint azt, hogy a napi középhőmérséklet szórása $\sigma = 2$. Azt szeretnénk vizsgálni, hogy a napi középhőmérséklet október 18-án Budapesten 15 °C alatt volt-e.
a.) Írjuk fel a nullhipotézist és az alternatív hipotézist!
b.) Teszteljük a hipotéziseket $\alpha = 0.05$ terjedelem mellett! Adjuk meg a kritikus tartományt és a p-értéket! Mi a döntés?
c.) Milyen hipotézist írjunk fel, ha azt szeretnénk vizsgálni, hogy a napi középhőmérséklet október 18-án Budapesten 15 °C-tól különböző volt-e? Teszteljük a fenti adatok segítségével!
d.) Teszteljük az eredeti hipotézist úgy is, hogy nem használjuk a szórásra vonatkozó előzetes információt!

SZ6.) **[XI.11.-ig]** A butitizmus betegségnél a vér kitamin tartalma (ezrelékben) jól közelíthető $N(18;2^2)$ eloszlással. A butitizmusban nem szenvedőknél ez az eloszlás $N(16;1)$. 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? (1+1+1+1=4p)

25.) Az alábbi minták két különböző gyáregységben tapasztalt selejtarányra vonatkoznak (ezrelékben). Tegyük fel, hogy a minták normális eloszlásúak és függetlenek.
A: 12.1, 13.0, 12.9, 12.2, 12.7, 12.6, 12.6, 12.8, 13.0, 13.1
B: 11.9, 12.1, 12.8, 12.2, 12.5, 11.9, 12.5, 11.8, 12.4, 12.9
a.) Melyik állítást van értelme vizsgálni a minták alapján az alábbiak közül? Vé-

gezz hipotézisvizsgálatot!
I.) Az 'A' gyáregység jobban dolgozott.
II.) A 'B' gyáregység jobban dolgozott.

b.) Végezzünk alkalmas hipotézisvizsgálatot arra vonatkozóan, hogy az 'A' gyáregységben a selejtarányok szórása vajon nagyobb-e 0.3-nál!

26.) Egy üzemben az 500 ml-es ásványvíz palackozása két gyártósoron folyik immáron 5 éve. A sok-sok éves tapasztalatok alapján az üzemmérnök megállapította, hogy a palackokba töltött ásványvíz mennyiségének szórása 20 ml-nek tekinthető mindkét gyártósoron, a töltött mennyiségek pedig normális eloszlásúak.
Az utóbbi időben a vevőktől nagyon sok panasz érkezett amiatt, hogy a palackokban lévő ásványvíz túl kevés, ezért a cég vezetősége belső vizsgálatot rendelt el: mindkét gyártósorról véletlenszerűen kiválasztottak egy-egy mintát és megmérték, mennyi a töltött mennyiség az egyes palackokban. Az 1. gyártósorról 30 palackot vizsgáltak meg, az átlagos töltőmennyiségre 490 ml adódott; míg a 2. gyártósorról 20 palackot vettek le, ezeknél az átlagos ásványvíztartalom 480 ml volt.

a.) Állíthatjuk-e, hogy a két gyártósor működésében nincs eltérés?
b.) Mi a következtetés – jogos a vevők észrevétele?

27.) Az alábbi két minta 10 forgalmas csomópont levegőjében található szennyezőanyag-koncentrációra vonatkozó két adatsort tartalmaz, a méréseket mindkét napon ugyanabban az időpontban végezték:
november 15.: 20.9, 17.1, 15.8, 18.8, 20.1, 15.6, 14.8, 24.1, 18.9, 12.5
november 29.: 21.4, 16.7, 16.4, 19.2, 19.9, 16.6, 15.0, 24.0, 19.2, 13.2
Szignifikánsan változott-e a légszennyezettség?

SZ7.) [XI.19.-ig] Két ország fővárosában a következő testtömegeket mérték a férfiak körében kg-ban:

1. város: 85, 80, 75, 90, 79, 101
2. város: 70, 68, 82, 78, 72, 81

Tegyük fel, hogy a testtömeg normális eloszlásúak és véletlenszerűen választották ki az alanyokat. Vizsgáljuk meg a következő állításokat hipotézisvizsgálattal:

a.) Az 1. városban nem nagyobb az átlagos testtömeg, mint a másodikban.
b.) A két város férfi testtömegértékei átlagosan megegyeznek.
c.) Az 1. városban a testtömegek *relatív* szórása nagyobb 10%-nál. (1+1+1=3p)

28.) Legyen X egyelemű minta, tekintsük a következő hipotéziseket:
H_0: $X \sim E(0;2)$, $\quad H_1$: $X \sim Exp(1)$

a.) Adjunk meg α terjedelemhez egyenletesen legerősebb próbát! Számoljuk ki a másodfajú hiba valószínűségét!
b.) Generáljunk mintákat a null- és az ellenhipotézisnek megfelelő eloszlásból és végezzük el a próbát!

29.) Tegyük fel, hogy kételemű mintánk van a Bin(4;p) eloszlásból. Adjuk meg a legerősebb 0.05 terjedelmű próbát az alábbi hipotézisekre:

a.) $H_0 : p = 0.5, \quad H_1 : p = 0.25$
b.) $H_0 : p \geq 0.5, \quad H_1 : p < 0.5$

30.) Tegyük fel, hogy n elemű mintánk van $N(m, \sigma^2)$ eloszlásból, σ ismert. Adjuk meg a legerősebb α terjedelmű próbát az alábbi hipotézisekre:

a.) $H_0 : m = m_0, \quad H_1 : m > m_0$
b.) $H_0 : m \leq m_0, \quad H_1 : m > m_0$

SZ8.) [XI.25.-ig] Legyen X_1 minta az $f(x)$ sűrűségfüggvényű eloszlásból. Tekintsük a következő hipotéziseket:
H_0: $f_0(x) = 2(1-x) \cdot I(0 < x < 1)$
H_1: $f_1(x) = 2x \cdot I(0 < x < 1)$
Adjunk meg α terjedelemhez egyenletesen legerősebb próbát! (2p)

31.) Egy gyárban egy termék minőségét 4 elemű mintákat véve ellenőrzik, havonta 300 mintavétellel. Megszámolták, hogy a legutóbbi hónapban hányszor volt selejtes a minta, melynek eredményeit az alábbi táblázat tartalmazza:

Selejtesek száma	0	1	2	3	4
Gyakoriságok	80	113	77	27	3

Modellezhető a mintákban levő selejtesek száma
a.) Bin(4; 0.25)
b.) Bin(4;p) paraméterű binomiális eloszlással?
Oldjuk meg a feladatot úgy is, hogy nem vonjuk össze a csoportokat és szimuláljuk a p-értéket!

32.) Generáljunk 2000 darab 1000 elemű mintát a $Bin(5;p)$ eloszlásból $p = 0.45, 0.46, \ldots, 0.55$ értékek esetén, majd vizsgáljuk meg illeszkedésvizsgálattal a $H_0 : X \sim Bin(5; 0.5)$ nullhipotézist. Számítsuk ki, hogy a 2000 ismétlés során milyen arányban vetettük el a nullhipotézist!

33.) Rendelkezésünkre áll a következő minta: 0.21, 2.02, 0.76, 0.70, 0.17, 0.23, 1.57, 3.52, 0.24, 0.85, 1.06, 1.60. Elvethető-e az a hipotézis, hogy a minta
a.) E(0; 4);
b.) Exp(1);
c.) N(2; 1) eloszlású?

34.) Generáljunk egy 200 elemű mintát 2 paraméterű exponenciális eloszlásból, majd diszkretizálás után χ^2-próbával vizsgáljuk meg, hogy a generált minta származhat-e 1, 1.1, ..., 3 paraméterű exponenciális eloszlásból! Nézzük meg Kolmogorov-Szmirnov próbával is!

35.) Az alábbi kontingencia-táblázat mutatja, hogy 100 évben a csapadék mennyisége és az átlaghőmérséklet hogyan alakult, a cellákban az egyes esetek gyakoriságai találhatóak.

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

Tekinthető-e a csapadékmennyiség és a hőmérséklet függetlennek?

36.) 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!

37.) Két dobókockával dobva az alábbi gyakoriságokat figyeltük meg:

Dobások	1	2	3	4	5	6
1. kocka	27	24	26	23	18	32
2. kocka	18	12	15	21	14	20

Döntsünk $\alpha = 0.05$ mellett arról, hogy a két eloszlás azonosnak tekinthető-e!

SZ9.) **[XII.2.-ig]** Tegyük fel, hogy a villamosmegállóban állva minden nap feljegyeztük, hány villamos ment el az ellenkező irányba, míg a mienk befutott. 90 nap megfigyelései alapján az alábbi gyakorisági tábla adódott:

Villamosok száma	0	1	2
Gyakoriság	30	40	20

Vizsgáljuk meg alkalmas statisztikai próbával, hogy a szembejövő villamosok száma egyenletes eloszlású-e a {0, 1, 2} számhalmazon! (2p)

SZ10.) **[XII.9.-ig]** Generálj egy 1000 elemű mintát az $E(0; 2)$ eloszlásból, majd diszkretizálással és Kolmogorov-Szmirnov próbával vizsgáld meg, hogy származhate az $f(x) = \frac{x}{2} I(0 < x < 2)$ sűrűségfüggvényű eloszlásból! (2p)

38.) Egy kisbolt tulajdonosa sokéves tapasztalata alapján azt mondja, hogy a vevők fele 1500 Ft alatti, másik fele 1500 Ft feletti összeget költ egy-egy reggeli bevásárlása során. Állításának alátámasztására feljegyezte a 7:00 és 7:10 között betérő vásárlók költéseit (Ft): 1410, 655, 5500, 640, 2300, 1730, 250, 1370.
Döntsünk 90%-os megbízhatósággal, statisztikailag megállja-e a helyét a bolttulajdonos állítása!

39.) Adott a következő 7 megfigyeléspár egy biztosítótársaság kárkifizetéseire régiónként két évre (M Ft-ban).

Régiók	1	2	3	4	5	6	7
2018	2767	234	262	223	1718	326	658
2019	1845	127	195	212	1486	320	634

Vajon szignifikánsan változott-e a kárkifizetés? Ellenőrizzük, hogy mit kapnánk a t-próbával. Miért nem használható erre az esetre?

40.) Tegyük fel, hogy két gyáregységben az alábbiak szerint alakultak az éves jövedelmek (M Ft-ban):
A: 3.2, 4.3, 5.8, 5.9, 7.2, 11.3, 25.6, 31.2
B: 2.6, 3.1, 4.1, 4.4, 4.8, 5.2, 5.4, 6.6
Van-e szignifikáns eltérés a gyáregységekben elérhető jövedelmek között?

SZ11.) **[XII.12.-ig]** Kétfajta instant kávé oldódási idejét tesztelték, melyekből minden alkalommal azonos mennyiséget tettek 1 dl forrásban lévő vízbe. A kísérletek eredményeit az alábbi táblázat tartalmazza (oldódási idő, mp):
Mokka Makka: 11.2, 5.0, 6.8, 6.7, 5.8, 7.3, 6.4, 5.8
Koffe In: 5.1, 4.3, 3.4, 3.7, 9.1, 4.7
Alkalmas statisztikai próbával vizsgáljuk meg, hogy van-e különbség az oldódási idők között! (2p)

41.) 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, valamint a becsléseink szórásnégyzetét!
c.) Adjunk előrejelzést $x = 10$-re a regressziós egyenes alapján!
d.) Szignifikáns a lineáris összefüggés a változók között?

42.) Az mtcars adatbázison vizsgáljuk meg R-ben lineáris regresszióval az alábbiakat $\alpha = 0.05$ mellett. Milyen hatással van
a.) a váltó típusa (am) a fogyasztásra (mpg)?
b.) a váltó típusa (am) és az autó tömege (wt) a fogyasztásra (mpg)?
c.) a hengerek száma (cyl) a fogyasztásra (mpg)?
d.) a teljesítmény (hp) a fogyasztásra (mpg)? Nézzük meg kovarianciával és korrelációval is, majd vessük össze a lineáris regresszióra kapott eredményekkel!
e.) a teljesítménynek (hp) és a tömegnek (wt) fogyasztásra (mpg)? Melyik magyarázóváltozó hatása erősebb?

43.) Február 17-én Budapesten az elmúlt 10 évben az alábbi középhőmérsékleteket mérték: 2, 2.5, 1.6, -4.5, 5.3, 7.9, 1.5, -1.6, -2.2, 1.6.
Készítsük el a Parzen-Rosenblatt-féle sűrűségfüggvény-becslést Gauss-magfüggvény esetén különböző sávszélességekre (R segítségével)! Vessük össze a hisztogrammal!

44.) 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 fenti regressziós modelleket!

c.) Számítsuk ki és elemezzük a korrelációs mátrixot! Keressük meg a legjobban illeszkedő modellt, ha az eredményváltozó a testmagasság!

d.) 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!

45.) Tekintsük a `https://stats.idre.ucla.edu/stat/data/binary.csv` linken elérhető amerikai egyetemi felvételi adatokat és próbáljuk meg modellezni a felvétel eredményét (valószínűségét) a rendelkezésre álló magyarázó változókkal! Értelmezzük a kapott eredményeket!

SZ12.) **[XII.12.-ig]** Egy baráti társaságban a következők ismertek: Peti 180 cm magas, Réka 165 cm, Juli 172 cm, Gábor 178 cm. Vizsgáld meg, hogy ezt a baráti társaságot alapul véve, a nem szignifikáns hatással van-e a testmagasságra! (1p)

SZ13.) **[XII.12.-ig]** Az mtcars adatbázison vizsgáljuk meg R-ben lineáris regresszióval, hogy milyen hatással van a fogyasztás (mpg) és a hengerek száma (cyl) a teljesítményre (hp) $\alpha = 0.05$ mellett! Interpretáljuk is az eredményeket! (1p)

46.)[3. ZH minta, 3. feladat] Frissen gyártott gépet tesztelünk, 10 ugyanolyan gombja van. H_0: a gép gombjai működnek, H_1: van hibás gomb. A következő próbát alkalmazzuk: véletlenszerűen választunk két különböző gombot és azokat megnyomjuk. Amennyiben mindkettő jó, H_0-t fogadjuk el, különben elvetjük.

a.) Mennyi az elsőfajú hiba valószínűsége?
b.) Mennyi a próba ereje, ha 3 hibás gomb van?

47.) Egy ügyvédi vállalkozásnak két irodája van: egy Budapesten és egy Miskolcon. 2018-ban Budapesten 4-en dolgoztak, a bruttó fizetésük 500 e Ft, 600 e Ft, 550 e Ft, 510 e Ft volt. A miskolci telephely dolgozói bruttó 450 e, 350 e, 400 e Ft-ot vittek haza.

a.) A telephely hány %-ban magyarázza a fizetések változékonyságát?
b.) A telephely szignifikáns hatással van a fizetésekre?
c.) Hogyan változik az előző két kérdésre adott válasz, ha Győrben is van egy telephely, az ott dolgozók fizetése pedig 450 e Ft és 550 e Ft?

48.) A távolsági autóbusz-közlekedés néhány jellemző adata egy év során:

Járat	Szállított utasok száma (fő)	Utazási távolság (km/fő) átlaga	szórása
Menetrend szerinti	450	17	5
Szerződéses	30	23	15
Külön	3	151	70
Összesen/együtt	...	...	...

a.) Számítsd ki a táblázat kipontozott celláit!
b.) A járat fajtája szignifikáns hatással van az utazási távolságra?

49.) A következő táblázat a 2016. 1. félévben tanár szakos BSc-s hallgatóknak tartott 4 valszám gyakorlat év végi, 100-ra skálázott végső pontszámait tartalmazza:

Gyakvezér	Pontszámok											
Cs. V.	98	87	102	92	52	46	95	60	81	55	60	94
	81	58	80	93	70	66	49	94	50	88	74	
W. G.	77	46	54	57	50	45	39	63	26	107	75	
	66	52	109	91	35	65						
B. Á.	86	94	54	61	42	59	88	81	81	80	102	72
	88	96	58	90	110	58	80	90	84	80	94	
V. L.	66	60	72	49	52	54	80	56	36	91	68	
	60	51	40	38	54	62						

a.) Vizsgáljuk meg, az év végi pontszám függ-e attól, hogy a hallgató melyik csoportba jár! Hány %-ban magyarázza a pontszámok változékonyságát az, hogy a hallgatók melyik csoportba járnak?
b.) Adjunk intervallumbecslést az egyes csoportok várható pontszámára!
c.) Állíthatjuk-e, hogy Cs. V. és B. Á. csoportjának átlagpontszámai (statisztikailag) egyenlők?