

# Biostatisztika feladatok

1	Diszkrét valószínűségi változók	1
2	Folytonos valószínűségi változók	1
3	A normális eloszlás	2
4	Statisztikai becslések	3
5	Statisztikai grafikonok	4
6	Konfidencia intervallum a várható értékre, t-próba	6
7	Az egymintás és a páros t-próba	6
8	Az egyszempontos ANOVA és a Levene-teszt	8
9	Lineáris és nemlineáris regresszió	9
10	Korrelációs együtthatók és függetlenségvizsgálat	10
11	Illeszkedésvizsgálat, normalitás tesztelése	10
12	Valószínűségek becslése és tesztelése	11
13	Diszkrét változók együttes eloszlása és függetlensége	13
14	Többszörös lineáris regresszió és többszempontos ANOVA	14
15	Paraméteres próbák és rangpróbák	15
16	Diszkriminanciaanalízis	17
	Megoldások	20

## 1. Diszkrét valószínűségi változók

- 1.1. A Pick Szeged férfi kézilabda csapatában az átlövők testmagassága 193, 198, 199, 200, 203 és 203 centiméter. Véletlenszerűen kiválasztva egy játékost mi az esélye annak, hogy az ő testmagassága legalább 200 cm? Mennyi a testmagasság várható értéke és szórása?
- 1.2. A biológiai kutatások egyik új és fontos területe a sárkányok vizsgálata. A tudósok eddig 1, 3, 7 és 12 fejű sárkányokat figyeltek meg, ezek aránya a populáción belül 10, 40, 30 illetve 20 százalék. Véletlenszerűen kiválasztunk egy egyedet a populációból, és jelölje  $\xi$  a fejek számát a választott egyednél! Adjuk meg a  $\xi$  változó értékészletét és valószínűségeloszlását! A valószínűségeloszlást ábrázoljuk grafikonon is! Határozzuk meg a változó móduszát, várható értékét és szórását! Mi az utolsó három mutatószám szemléletes jelentése a populációra nézve?
- 1.3. Biológusok azt vizsgálták, hogy egy nemzeti parkban hány egyed él egy ritka fafajból. Felosztották a park területét 1 hektár területű négyzetekre, és felmérték, hogy az egyes négyzetekben hány egyed található ebből a fajból. Egy egyedet sem találtak a négyzetek 40 százalékán, 1 egyedet találtak a négyzetek 30 százalékán, 2 egyedet találtak a négyzetek 20 százalékán, és végül 3 egyedet találtak a négyzetek 10 százalékán. Három egyednél többet sehol sem találtak. Legyen  $\xi$  az egyedek száma egy véletlenszerűen kiválasztott négyzetben!
- Adjuk meg a  $\xi$  változó értékészletét és valószínűségeloszlását! Mennyi az esélye, hogy a kiválasztott négyzeten 1-nél több egyed található a fafajból? Mennyi a  $\xi$  módusza, várható értéke illetve szórása? Mi a jelentése az utolsó három mutatószámnak a teljes nemzeti parkra nézve?
- 1.4. Egy szerencsejátékban a játékos 1000, 2000, 3000 vagy 5000 forintot nyerhet, ezen nyeremények esélye 50, 30, 15 illetve 5 százalék. Egyszer játsszuk ezt a játékot, jelölje  $\xi$  a nyeremény nagyságát! Adjuk meg a  $\xi$  változó értékészletét, valószínűségeloszlását, móduszát, várható értékét és szórását! Mennyi az esélye annak, hogy legfeljebb 2000 forintot nyerünk?

## 2. Folytonos valószínűségi változók

- 2.1. Jelölje  $\xi$  a napi középhőmérsékletet Celsiusban egy januári napon. A  $\xi$  egy folytonos valószínűségi változó, melynek sűrűségfüggvénye  $f(x) = 1/20$  ha  $-15 \leq x \leq 5$ , és  $f(x) = 0$  minden más  $x$  esetén.
- Vázlatosan rajzoljuk fel a sűrűségfüggvény grafikonját, és adjuk meg a  $\xi$  változó értékészletét!
  - Mennyi annak az esélye, hogy a napi középhőmérséklet  $-10^\circ\text{C}$  és  $10^\circ\text{C}$  közé esik? Mekkora valószínűséggel lesz a napi középhőmérséklet legalább  $0^\circ\text{C}$ ?

- c. Határozzuk meg a  $\xi$  változó eloszlásfüggvényét! Ezek után válaszoljunk az előző pont kérdéseire az eloszlásfüggvény alkalmazásával!
- d. Adjuk meg a  $\xi$  valószínűségi változó 80%-os kvantilisét, várható értékét és szórását! Mi ezeknek a mutatószámoknak a szemléletes jelentése?

**2.2.** Egy erdőben a fák törzsének méterben kifejezett átmérője a következő sűrűségfüggvénnyel írható le:  $f(x) = 3\sqrt{x}/2$  ha  $0 \leq x \leq 1$ , és  $f(x) = 0$  minden más  $x$  esetén. Véletlenszerűen kiválasztunk egy fát, és legyen  $\xi$  ezen egyed átmérője!

- a. Vázlatosan rajzoljuk fel a sűrűségfüggvény grafikonját, és adjuk meg a  $\xi$  változó értékkészletét!
- b. Mennyi a  $P(0.5 \leq \xi \leq 1.5)$  és  $P(\xi \leq 0.8)$  valószínűségek értéke? Mi a jelentése ezeknek az értékeknek az erdő szempontjából?
- c. Határozzuk meg a  $\xi$  változó eloszlásfüggvényét! Ezek után válaszoljunk az előző pont kérdéseire az eloszlásfüggvény alkalmazásával!
- d. Adjuk meg a várható értéket, a szórást, a mediánt illetve az alsó és a felső kvartilist! Mi a szemléletes jelentése ezeknek a mutatószámoknak?

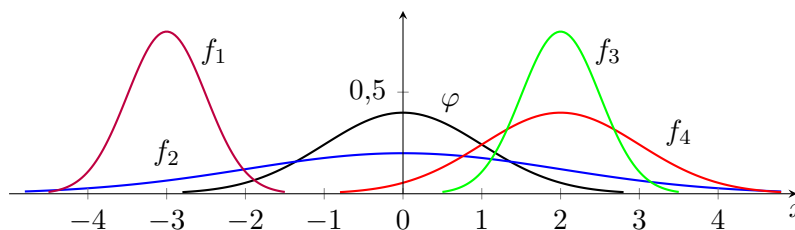
**2.3.** Egy állatpopulációban az egyedek testhossza a következő sűrűségfüggvénnyel írható le:  $f(x) = 8/(3x^3)$  ha  $1 \leq x \leq 2$ , és  $f(x) = 0$  minden más  $x$  valós számra.

- a. Rajzoljuk fel a sűrűségfüggvényt, és adjuk meg a testhossz értékkészletét!
- b. A populációban az egyedek mekkora hányadának esik a testhossza 0.5 és 1.5 közé? Az egyedek hány százaléka éri el az 1.8 hosszúságot?
- c. Határozzuk meg a  $\xi$  változó eloszlásfüggvényét! Ezek után válaszoljunk az előző pont kérdéseire az eloszlásfüggvény alkalmazásával!
- d. Határozzuk meg a testhossz várható értékét és szórását! Adjunk meg három intervallumot a testhosszra olyan módon, hogy mindegyikbe az egyedek harmada essen!

### 3. A normális eloszlás

**3.1.** Az alábbi ábrán  $\varphi$  a standard normális eloszlás sűrűségfüggvénye. Határozzuk meg, hogy az  $f_1, f_2, f_3, f_4$  sűrűségfüggvények közül melyik tartozik az alábbi  $\mu$  várható értékkel és  $\sigma$  szórással definiált normális eloszlásokhoz. Adjuk meg a kimaradt sűrűségfüggvényhez tartozó várható értéket és szórást is.

- a.  $\mu = 2, \sigma = 0,5$
- b.  $\mu = 2, \sigma = 1$
- c.  $\mu = 0, \sigma = 2$



- 3.2.** Az IQ tesztek úgy állítják össze, hogy az eredmény a felnőtt népességen belül normális eloszlást kövessen 100 pont várható értékkel és 15 pont szórással. A felnőtt népesség mekkora hányadának esik az IQ pontszáma 90 és 120 közé? A Mensa egy nemzetközi egyesület, ahol a belépés feltétele a legalább 131 pontos IQ. A népesség hány százaléka felel meg ennek a követelménynek? Adjunk meg egy olyan intervallumot, melyre teljesül, hogy az emberek 95 százalékának ebbe az intervallumba esik az IQ pontszáma.
- 3.3.** Biológusok azt vizsgálták, hogy a szavannán élő majmok reggelente milyen eloszlás szerint ébrednek fel, és másznak le a fáról. A megfigyelések alapján az ébredési idő egy normális eloszlású valószínűségi változó. A majmok átlagosan reggel 7 órakor kelnek fel, a szórással 0.75 óra. A majmok mekkora hányada kel fel 6 és 7 óra között? Mekkora hányad ébred 8 óra után? Adjunk meg egy olyan időintervallumot, melyre teljesül, hogy a majmok 90 százaléka ebben az időintervallumban mászik le a fáról!
- 3.4.** Szegeden az éves csapadékmennyiség egy olyan  $\xi$  valószínűségi változó, mely normális eloszlást követ 500 ml várható értékkel és 50 ml szórással. Mennyi az esélye annak, hogy egy adott évben a csapadék mennyisége 460 ml és 525 ml közé esik? Adjunk meg egy olyan intervallumot, mely 95% valószínűséggel tartalmazza a  $\xi$  változót!

## 4. Statisztikai becslések

- 4.1.** A 'carData' csomagban található 'Davis' adatsor egy pszichológia felmérés eredményét tartalmazza. A változók:  
sex: nem (F=nő, M=férfi)  
weight: testsúly (kg)  
height: testmagasság (cm)  
repwt: az alany mekkorának gondolja saját testsúlyát (kg)  
reph: az alany mekkorának gondolja saját testmagasságát (cm)
- Adjunk becslést a 'repwt' változó teljes sokaságban vett várható értékére és szóráására! Nevezzük meg, hogy mely statisztikai becsléseket alkalmaztuk!
  - Hány megfigyelés van a 'repwt' változóra, és mennyi a hiányzó adatok száma?
  - Adjunk becslést a 'repwt' változónak a teljes sokaságban mért mediánjára, alsó kvartilisére és felső kvartilisére!
  - Határozzuk meg a 'repwt' változó esetében a következő mutatószámok értékét: minimum, maximum, terjedelem, IQR! Mi ezeknek a mutatószámoknak a jelentése a mintára nézve?
  - Adjunk meg a 'repwt' változóra három olyan intervallumot, hogy mindegyikbe a mintaelemek harmada essen!
- 4.2.** A 'carData' csomagban található 'Mroz' adatsor egy amerikai felmérés eredménye, az alanyok férjezett nők. Az 'age' változó az alanyok életkorát tartalmazza.

- a. Adjunk becslést az ‘age’ változó várható értékére, mediánjára és szórására! Nevezzük meg pontosan, hogy mely statisztikai mutatószámokat alkalmaztuk!
- b. Hány megfigyelés van az ‘age’ változóra, és mennyi a hiányzó adatok száma?
- c. Határozzuk meg a ‘age’ változó esetében a következő mutatószámok értékét: minimum, maximum, alsó és felső kvartilis, terjedelem, IQR! Mi ezeknek a mutatószámoknak a jelentése a mintára nézve?
- d. Adjunk meg az ‘age’ változóra öt olyan intervallumot, hogy mindegyikbe a mintaelemek ötöde essen!

**4.3.** Régészek radiokarbonos módszerrel szeretnék meghatározni egy lelőhely igazi korát. Sajnos a radiokarbonos módszer egy adott ásatáson nem pontosan ugyanazt a kort adja minden lelet esetében. Az egyes leletek radiokarbonos kora egy  $\xi$  valószínűségi változó, és a lelőhely igazi kora ennek a változónak a várható értéke.

Egy ásatáson a radiokarbonos módszert hét leleten alkalmazva a következő korokat kapták: 1180, 1220, 1230, 1250, 1270, 1290 és 1340 év.

- a. Határozzuk meg a mintaméretet, a mintaátlagot, valamint a korrigálatlan és a korrigált empirikus szórást! Ezek alapján milyen becslés adható a lelőhely igazi korára? A két empirikus szórás közül melyikkel érdemes becsülni a  $\xi$  változó igazi szórását?
- b. Adjuk meg a minta mediánját és terjedelmét!

## 5. Statisztikai grafikonok

**5.1.** A ‘carData’ csomagban található ‘Davis’ adatsor egy pszichológia felmérés eredményét tartalmazza. A változók:

sex: nem (F=nő, M=férfi)

weight: testsúly (kg)

height: testmagasság (cm)

repwt: az alany mekkorának gondolja saját testsúlyát (kg)

repht: az alany mekkorának gondolja saját testmagasságát (cm)

- a. Az adatsorban mely változók diszkrét és melyek folytonosak?
- b. Ábrázoljuk a ‘sex’ változót oszlopdiagram és kördiagram segítségével! Hány nő és hány férfi található a mintában?
- c. Ábrázoljuk a ‘repwt’ változó hisztogramját, majd adjunk grafikus becslést a sűrűségfüggvényre! Mennyi a minta ferdesége? Ezek alapján mit állíthatunk a hisztogramról: jobbra ferde, balra ferde vagy közel szimmetrikus?
- d. Ábrázoljuk a ‘repwt’ változó boxplotját! Hány outlier érték van a mintában? Nevezzük meg, hogy mely statisztikai mutatószámok jelennek meg a boxploton, és adjuk meg ezen mutatószámok pontos értékét!

- e. A fentiek alapján mit állíthatunk a 'repwt' változó eloszlásáról: közel normális vagy nagy mértékben különbözik a normálistól?
- f. Ábrázoljuk a 'weight' változó boxplotját! Vegyük észre, hogy az egyik outlier érték nagyon kilóg! Keressük ki ezt a megfigyelést az adatsorból, és adjunk magyarázatot a jelenségre!

**5.2.** A 'carData' csomagban található 'Mroz' adatsor egy amerikai felmérés eredménye, az alanyok férjezett nők. A fontosabb változók:

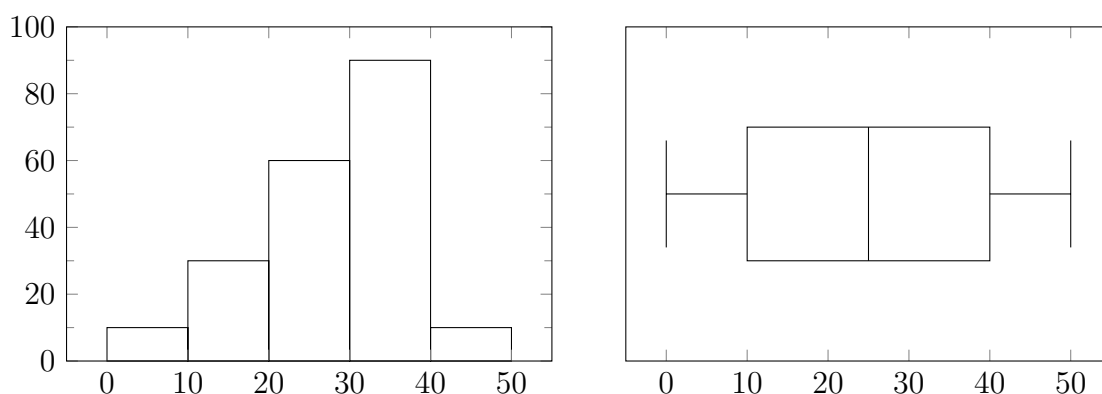
age: életkor (év)

wc: rendelkezik-e főiskolai vagy egyetemi végzettséggel (yes=igen, no=nem)

k5: a legfeljebb 5 éves gyerekek száma a családban

- a. A fenti változók közül melyek diszkrét és melyek folytonosak?
- b. Ábrázoljuk a 'wc' változót oszlopdiagramon és kördiagramon! Az alanyok közül hányan rendelkeznek, és hányan nem rendelkeznek diplomával?
- c. Ábrázoljuk az 'age' változó hisztogramját, majd adjunk grafikus becslést a sűrűségfüggvényre! Mennyi a minta ferdesége? Ezek alapján mit állíthatunk a hisztogramról: jobbra ferde, balra ferde vagy közel szimmetrikus?
- d. Ábrázoljuk az 'age' változó boxplotját! Hány outlier érték van a mintában? Nevezzük meg, hogy mely statisztikai mutatószámok jelennek meg a boxploton, és adjuk meg ezen mutatószámok pontos értékét!
- e. A fentiek alapján mit állíthatunk az 'age' változó eloszlásáról: közel normális vagy nagy mértékben különbözik a normálistól?

**5.3.** Az alábbi ábrán egy hisztogram és egy boxplot látható.



- a. A hisztogram alapján körülbelül mennyi a minta elemszáma? Hozzávetőlegesen mekkora a legkisebb illetve a legnagyobb elem? Milyen előjelű a ferdeség? Hány módot látunk az ábrán és mely értékeknél?
- b. Nevezzük meg, hogy mely statisztikai mutatószámok jelennek meg a boxploton, majd olvassuk le ezek értékét a grafikonról! Van outlier érték?
- c. A hisztogram és a boxplot ugyanazon statisztikai mintához tartozik? A választ indokoljuk is!

## 6. Konfidencia intervallum a várható értékre, t-próba

- 6.1. Régészek radiokarbonos kormeghatározással szeretnék meghatározni egy lelőhely korát. Sajnos a radiokarbonos módszer az adott ásatáson nem pontosan ugyanazt a kort adja minden lelet esetében. Az egyes leletek radiokarbonos kora egy normális eloszlású  $\xi$  valószínűségi változó, és a lelőhely igazi kora ennek a változónak a várható értéke.

Egy ásatáson a radiokarbonos módszert öt leleten alkalmazva a következő korokat kapták: 1180, 1220, 1230, 1250 és 1270 év.

- Adjunk becslést a lelőhely igazi korára, tehát a  $\xi$  változó várható értékére! Számoljuk ki és értelmezzük a standard hibát is!
  - Írjunk fel egy 95% megbízhatóságú konfidencia intervallumot a lelőhely igazi korára!
  - A t-próba alkalmazásával teszteljük le 5%-os szignifikancia szinten azt a nullhipotézist, hogy a lelőhely igazi kora 1200 év!
  - Mennyi az elsőfajú illetve a másodfajú hiba nagysága ebben a feladatban?
- 6.2. Bejelentés érkezik a fogyasztóvédelemhez, hogy az egyik tejgyár 1 literes kiszerezésű dobozos teje a névleges tartalomnál kevesebbet tartalmaz. Tudni kell, hogy a töltőberendezések véletlen nagyságú hibával dolgoznak, így ténylegesen egyik dobozban sincs pontosan 1 liter tej. Feltehető, hogy a dobozokba töltött mennyiség egy  $\xi$  normális eloszlású valószínűségi változó, melynek 1 liter a várható értéke, ha a gép jól van beállítva. A fogyasztóvédelem emberei beszereznek hat doboz tejet, és azt találják, hogy ezek 975, 980, 985, 995, 1000, 1005 ml tejet tartalmaznak.
- Adjunk becslést a  $\xi$  változó várható értékére! Számoljuk ki és értelmezzük a standard hibát is!
  - Adjunk meg egy 90% megbízhatóságú konfidencia intervallumot a  $\xi$  változó várható értékére!
  - A t-próba alkalmazásával teszteljük le 10%-os szignifikancia szinten azt, hogy a gép jól van beállítva, tehát a tejesdobozokba átlagosan 1000 ml tej kerül!

## 7. Az egymintás és a páros t-próba

- 7.1. A 'vernyomas.xlsx' fájlban található adatsor egy orvosi kísérlet eredményét tartalmazza. A kísérlet keretei között két új vérnyomáscsökkentő gyógyszert vizsgáltak. Véletlenszerűen kiválasztottak 150 magas vérnyomású páciens, és három 50 fős csoportba sorolták őket. A 'kiserleti1' és a 'kiserleti2' csoport az 1. illetve a 2. kísérleti gyógyszert szedte néhány héten át. A 'kontroll' csoport a hagyományos kezelést kapta. A változók:

CSOPNEV: betegcsoport neve  
CSOPKOD: betegcsoport kódja  
SYS1: kezelés előtti szisztolés vérnyomás  
SYS2: kezelés utáni szisztolés vérnyomás

- a. Adjunk becslést a 'SYS1' változó teljes sokaságban mért átlagos értékére és szórására! Mennyire pontos a sokaság átlagára kapott becslés?
- b. Ábrázoljuk a 'SYS1' változó hisztogramját, és kérdezzük le a ferdeséget is! Mit állíthatunk a 'SYS1' változó eloszlásáról: közel normális vagy nagy mértékben különbözik a normálistól?
- c. Teszteljük le azt a nullhipotézist, hogy a 'SYS1' változó teljes sokaságban mért átlagos értéke 160 Hgmm! Teszteljük le a 165 Hgmm-es értéket is! Adjunk meg egy 95% megbízhatóságú konfidencia intervallumot a sokaság átlagára! Hogyan értelmezhető ez a konfidencia intervallum?
- d. Válogassuk le a 'kiserleti1' betegcsoport tagjait, majd adjunk becslést a 'SYS1' és a 'SYS2' változó várható értékére ebben a betegcsoportban! Ábrázoljuk a két változó hisztogramját is! Mit állíthatunk a két változó eloszlásáról?
- e. Teszteljük le 1%-os szignifikancia szinten azt a nullhipotézist, hogy a 'kiserleti1' betegcsoportban azonos a 'SYS1' és 'SYS2' változók várható értéke! Adjunk meg egy 99% megbízhatóságú konfidencia intervallumot a várható értékek különbségére!
- f. Ismételjük meg az utolsó két pont elemzését a 'kiserleti2' betegcsoportra!

**7.2.** Az 'iris.xlsx' állomány három Kanadában honos írisz (nőszírom) fajról tartalmaz adatokat, fajonként 50 növényről. A változók:

faj: faj megnevezése

fajkod: lásd faj

cseszehossz: szészelevél hossza (cm)

cseszszel: csészelevél szélessége (cm)

szíromhossz: szíromlevél hossza (cm)

szíromszel: szíromlevél szélessége (cm)

- a. Ábrázoljuk a 'szíromszel' változó hisztogramját! Hány módusza van ennek az eloszlásnak? Mi lehet ennek az oka? Mi a szokásos eljárás, ha a statisztikában ilyen adatsorral találkozunk?
- b. Válogassuk le a 'virginica' fajhoz tartozó növényeket, és ábrázoljuk a 'szíromszel' változó hisztogramját csak erre a fajra! Mit állíthatunk a 'szíromszel' változó eloszlásáról: közel normális vagy nagy mértékben különbözik a normálistól?
- c. Adjunk becslést a 'virginica' fajhoz tartozó növényeknél a 'szíromszel' változó várható értékére és szórására! Teszteljük le 5% szignifikancia szinten azt a nullhipotézist, hogy a várható érték 2 cm! Adjunk meg egy 95% megbízhatóságú konfidencia intervallumot is erre a várható értékre!



- d. Adjunk becslést a ‘virginica’ fajnál a csészelevél átlagos hosszúságára is! Teszteljük le 10% szignifikancia szinten azt a nullhipotézist, hogy a ‘virginica’ fajnál a szíromlevél átlagos szélessége azonos a csészelevél átlagos szélességével! Adjunk meg egy 90% megbízhatóságú konfidenciai intervallumot arra, hogy a szíromlevél átlagosan mennyivel szélesebb, mint a csészelevél!
- e. Végezzük el az előző pontok elemzését a másik két faj egyedeire is!

## 8. Az egyszempontos ANOVA és a Levene-teszt

8.1. Olvassuk be az ‘vernyomas.xlsx’ fájlban található statisztika adatsort, a leírásért lásd a 7.1. feladatot!

- a. Adjunk becslést a ‘SYS1’ változó várható értékére és szórására betegcsoportonkénti bontásban! Ábrázoljuk a változó boxplotját is, szintén betegcsoportonkénti bontásban! Látunk jelentős eltérést a három csoport között?
- b. Ábrázoljuk a ‘SYS1’ változó hisztogramját és kérdezzük le a változó ferdeségét betegcsoportonkénti bontásban! Mit állíthatunk a ‘SYS1’ változó eloszlásáról a csoportokon belül: közel normális vagy nagyon különbözik a normálistól?
- c. Teszteljük le 5% szignifikancia szinten azt a nullhipotézist, hogy a ‘SYS1’ változónak azonos a szórása a három betegcsoportban! Teszteljük le a várható értékek egyenlőségét is!
- d. Végezzük el az előző feladatrészek elemzését a ‘SYS2’ változóra is! Amennyiben szignifikáns eltérést tapasztalunk a várható értékek között, akkor adjunk becslést és 95% megbízhatósági szintű konfidencia intervallumot a csoportonkénti várható értékek közötti különbségre!

8.2. Olvassuk be az ‘iris.xlsx’ fájl tartalmát! Az adatsor leírása megtalálható a 7.2. feladatban!

- a. Adjunk becslést a ‘szíromszel’ változó várható értékére és szórására fajonkénti bontásban! Ábrázoljuk a változó boxplotját is, szintén fajonkénti bontásban!
- b. Ábrázoljuk a ‘szíromszel’ változó hisztogramját és adjuk meg a változó ferdeségét betegcsoportonkénti bontásban. Mit állíthatunk a ‘szíromszel’ változóról normalitás szempontjából?
- c. Teszteljük le azt a nullhipotézist, hogy a ‘szíromszel’ változó esetében a csoportonkénti szórások azonosak. A szignifikancia szint 5%.
- d. Teszteljük le a csoportonkénti várható értékek egyenlőségét is. Ha szignifikáns eltérés tapasztalható a várható értékek között, akkor adjunk becslést és 95% megbízhatósági szintű konfidencia intervallumot a várható értékek közötti különbségekre!
- e. Ismételjük meg a fenti elemzést a ‘csészeszel’ változóra!

## 9. Lineáris és nemlineáris regresszió

**9.1.** A 'UScars.txt' adatsorban a '80-as években az amerikai piacon forgalmazott néhány autótípus fontosabb műszaki paraméterei szerepelnek. A változók:

MODEL: a modell neve

COUNTRY: hol gyártották

VOL: utastér térfogata (köbláb)

HP: teljesítmény (lóerő)

MPG: hány mérföldet lehet megtenni 1 gallon üzemanyaggal (mérőföld/gallon)

SP: végsebesség (mérőföld/óra)

WT: teljes tömeg (100 font)

- a. Ábrázoljuk az 'SP' változót a 'HP' változó függvényeként! Végezzünk lineáris regressziót a változókon, és adjuk meg a regressziós egyenes egyenletét! Mennyire jól illeszkedik a regressziós egyenes a megfigyelt értékekhez? Ezek alapján milyen becslést adhatunk egy 150 lóerős autó végsebességére?
- b. Végezzük lineáris regressziót az 'SP' és a 'VOL' változóra is, fejezzük ki az 'SP' változót a 'VOL' függvényeként! Adjuk meg a regressziós egyenes egyenletét! Mennyire jól illeszkedik az egyenes az adatokhoz? A gyakorlati alkalmazások szempontjából ez egy jó becslés?
- c. Ábrázoljuk az 'MPG' változót a 'HP' függvényeként! Végezzünk lineáris és nemlineáris regressziót a két változóra az alábbi módszerekkel:  
Lineáris regresszió:  $MPG \approx aHP + b$   
Reciprokos regresszió:  $MPG \approx a/HP + b$   
Exponenciális regresszió:  $MPG \approx \exp(aHP + b)$   
Melyik módszer biztosítja a legjobb becslést az 'MPG' változóra?

**9.2.** A 'carData' csomag 'States' adatsora azt vizsgálja, hogy az Egyesült Államok egyes tagállamai mennyit költöttek a középiskolás oktatásra a 90'-es évek elején, és ennek hatására milyenek lettek az egyetemi felvételi eredmények. Olvassuk be az adatsort illetve kérdezzük le az adatsor leírását.

- a. Ábrázoljuk az 'SATV' változót az 'SATM' változó függvényeként! Végezzünk lineáris regressziót a változókon, és adjuk meg a regressziós egyenes egyenletét! Mennyire jól illeszkedik a regressziós egyenes a megfigyelt értékekhez? Ezek alapján milyen becslést adhatunk egy az 'SATV' változóra egy olyan tagállamban, ahol az 'SATM' értéke 500?
- b. Végezzük lineáris regressziót a 'pop' és 'dollars' változókra, fejezzük ki a 'dollars' változót a 'pop' függvényeként! Adjuk meg a regressziós egyenes egyenletét! Mennyire jól illeszkedik az egyenes az adatokhoz? A gyakorlati alkalmazások szempontjából ez egy jó becslés?

- c. Ábrázoljuk az ‘SATV’ változót a ‘percent’ változó függvényeként! Végezzünk reciprokos regressziót a két változóra az alábbi formulákkal:

$$\text{SATV} \approx a \frac{1}{\text{percent}} + b, \quad \text{SATV} \approx \frac{1}{a \cdot \text{percent} + b}.$$

Melyik formula biztosítja a jobb becslést az ‘SATV’ változóra?

## 10. Korrelációs együtthatók és függetlenségvizsgálat

- 10.1. Olvassuk be a ‘UScars.txt’ fájlban található statisztika adatsort, a leírásért lásd a 9.1. feladatot!

- Adjuk meg az ‘SP’ és ‘HP’ változók Pearson- illetve Spearman-féle korrelációs együtthatóját! Teszteljük le a két változó függetlenségét a kapcsolatos korrelációs tesztekkel! Értelmezzük is az eredményt!
- Végezzük el az előző pont elemzését az ‘SP’ és ‘VOL’ változókon is!
- Végezzük el az előző pont elemzését az ‘MPG’ és ‘HP’ változókon is!

- 10.2. A ‘carData’ csomag ‘States’ adatsora azt vizsgálja, hogy az Egyesült Államok egyes tagállamai mennyit költöttek a középsikolás oktatásra a 90’-es évek elején, és ennek hatására milyenek lettek az egyetemi felvételi eredmények. Olvassuk be az adatsort illetve kérdezzük le az adatsor leírását.

- Adjuk meg az ‘SATM’ és ‘SATV’ változók Pearson- illetve Spearman-féle korrelációs együtthatóját! Teszteljük le a két változó függetlenségét a kapcsolatos korrelációs tesztekkel! Értelmezzük is az eredményt!
- Végezzük el az előző pont elemzését a ‘pop’ és ‘dollars’ változókon is!
- Végezzük el az előző pont elemzését az ‘SATV’ és ‘percent’ változókon is!

## 11. Illeszkedésvizsgálat, normalitás tesztelése

- 11.1. Olvassuk be az ‘vernyomas.xlsx’ fájlban található statisztika adatsort, a leírásért lásd a 7.1. feladatot!

- Ábrázoljuk a ‘SYS1’ és a ‘SYS2’ változó hisztogramját, és kérdezzük le a ferdeségeket is! Ezek alapján melyik változó tűnik normális eloszlásúnak?
- Ábrázoljuk a ‘SYS1’ és a ‘SYS2’ változó QQ-plotját is. Mit olvashatunk le ezekről?
- Teszteljük le 5% szignifikancia szinten azt a nullhipotézist, hogy a ‘SYS1’ változó normális eloszlást követ a teljes sokaságban! Teszteljük le a ‘SYS2’ változó normalitását is!

- d. Teszteljük le 5% szignifikancia szinten azt a nullhipotézist, hogy a 'SYS2' változó normális eloszlást követ a 'CSOPNEV' változó által meghatározott csoportokban!

**11.2.** Olvassuk be az 'iris.xlsx' fájlban található statisztika adatsort, a leírásért lásd a **7.2.** feladatot!

- a. Ábrázoljuk a 'cseeszszel' és a 'sziromszel' változó hisztogramját, és kérdezzük le a ferdeségeket is! Ezek alapján melyik változó tűnik normális eloszlásúnak?
- b. Ábrázoljuk a két változó QQ-plotját is. Mit olvashatunk le ezekről?
- c. Teszteljük le 10% szignifikancia szinten azt a nullhipotézist, hogy a 'cseeszszel' változó normális eloszlást követ a teljes populációban! Teszteljük le a 'sziromszel' változó normalitását is!
- d. Teszteljük le a két változó normalitását fajonkénti bontásban! A szignifikancia szint 10%.

## 12. Valószínűségek becslése és tesztelése

- 12.1.**
- a. Feldobunk egy nem feltétlenül szabályos dobókockát 100 alkalommal. A dobások során 15 egyest, 15 kettést, 15 hármast, 15 négyest, 20 ötöst és 20 hatost kaptunk. Mekkora a minta elemszáma? A minta alapján adjunk pontbecslést az egyes értékek dobásának a valószínűségére.
  - b. Teszteljük le 5%-os szignifikancia szinten azt a nullhipotézist, hogy a dobókocka szabályos, tehát minden értéknek 1/6 az esélye. Teszteljük le külön azt a nullhipotézist is, hogy hatosdobás valószínűsége 1/6.
  - c. Ugyanezt a dobókockát most 1000 alkalommal dobjuk fel, melyből 150 egyest, 150 kettést, 150 hármast, 150 négyest, 200 ötöst és 200 hatost kapunk. Oldjuk meg az a. és b. feladatrészeket ezzel a módosítással.

**12.2.** A 'catsdogs.txt' adatsor véletlenszerűen kiválasztott amerikai egyetemistákról tartalmaz információt. A változók:

Gender: a hallgató neme (M=férfi, F=nő)

Smokes: dohányzik-e (yes=igen, no=nem)

Dogs: van-e kutyája (yes=igen, no=nem)

CatsDogs: van-e kutyája vagy macskája (both=mindkettő, cats=macska, dogs=kutya, none=egyik sem)

- a. Ábrázoljuk a 'CatsDogs' változót kördiagram segítségével! Adjuk meg az egyes értékek tapasztalati gyakoriságát és relatív gyakoriságát! Ezek alapján milyen becslést adhatunk az egyes értékek teljes sokaságban mért arányára?
- b. Teszteljük le 1% szignifikancia szinten azt a nullhipotézist, hogy az amerikai egyetemisták körében a 'CatsDogs' változó rendre 5%, 15%, 20% és 60% arányban veszi fel a 'both', 'cats', 'dogs' illetve 'none' értéket! Írjuk fel a várt gyako-

riságokat is! Hol tapasztalunk jelentős eltérést a tapasztalati és a várt gyakoriságok között?

- c. Ábrázoljuk a 'Dogs' változót oszlopdiagram segítségével, és adjuk meg az egyes értékek tapasztalati gyakoriságát és relatív gyakoriságát! Teszteljük le 1% szignifikancia szinten azt a nullhipotézist, hogy az amerikai egyetemisták 20% arányban tartanak kutyát!

- 12.3.** Augusto Pinochet 1973 és 1990 között volt Chile teljhatalmú vezetője. Tevékenysége nagyban megosztotta az ország lakosságát. Irányítása alatt a gazdaság fejlődött és az életszínvonal emelkedett, de a politikájával szembeszegülőket bebörtönöztette és kivégeztette. Mivel az őt kritizáló hangok az országon belül és kívül is felerősödtek, 1988-ban a hatalmát egy dél-amerikai mércével mérve szabad választáson próbálta megerősíteni. A 'carData' csomag 'Chile' adatsora egy közvélemény-kutatás eredménye, melyet néhány héttel a választások előtt tartottak. Az alanyok véletlenszerűen lettek kiválasztva a szavazásra jogosult lakosok közül. A fontosabb változók:
- region: melyik tartományban él  
sex: nem (F=nő, M=férfi)  
education: iskolai végzettség (P=alapfokú, S=középfokú, PS=felsőfokú)  
vote: hogyan fog szavazni a választáson (A = távol marad, N = Pinochet ellen,  
U = bizonytalan, Y = Pinochet mellett)

- a. Ábrázoljuk a 'sex' változót oszlopdiagram segítségével, továbbá adjuk meg az egyes értékek tapasztalati gyakoriságát és relatív gyakoriságát! Teszteljük le 5% szignifikancia szinten azt a nullhipotézist, hogy a választásra jogosult lakosok sokaságán belül 50% a nők aránya!

Miért tesztelünk le valamit, amiről tudjuk, hogy igaz? Mert kíváncsiak vagyunk arra, hogy a minta tényleg reprezentálja-e a sokaságot. Ha esetleg elvetjük a nyilvánvalóan igaz nullhipotézist, akkor olyan szintű ellenmondás van a minta és a valóság között, hogy a minta nem lehet reprezentatív.

- b. Ábrázoljuk az 'education' változót kördiagram segítségével! Adjunk becslést arra, hogy a választásra jogosult lakosságban belül az emberek milyen arányban rendelkeznek alapfokú, középfokú illetve felsőfokú végzettséggel! Teszteljük le 5% szignifikancia szinten azt a nullhipotézist, hogy 40% az alapfokú, 40% a középfokú és 20% a felsőfokú végzettség aránya!

### 13. Diszkrét változók együttes eloszlása és függetlensége

13.1. Egy ország lakosságát vizsgáljuk haj- és szemszín szempontjából, az alábbi táblázat a lehetséges kombinációk teljes sokaságon belüli arányait tartalmazza.

	fekete haj	barna haj	szőke haj
sötét szem	25%	30%	5%
világos szem	5%	20%	15%

- Adjuk meg a hajszín illetve a szemszín marginális eloszlását! Mennyi az oddsza a sötét szemnek illetve a szőke hajnak?
- Véletlenszerűen kiválasztunk egy embert az országból. Mennyi az esélye annak, hogy a kiválasztott ember fekete hajú, ha tudom, hogy sötét a szeme? Mennyi az a esélye annak, hogy világos szemű, ha tudom, hogy szőke alanyt választottam. Hogyan értelmezhetőek ezek a valószínűségek arányszámként?
- Független egymástól a szem és a haj színe a sokaságon belül? Milyen arányban fordulnának elő a lehetséges kombinációk, ha a haj és a szem színe független lenne egymástól?
- Adjuk meg a világos szem relatív kockázatát a szőke hajú emberek körében a fekete hajú alanyokhoz viszonyítva! Hogyan értelmezhető a relatív kockázat ebben a feladatban?

13.2. A 'catsdogs.txt' adatsor véletlenszerűen kiválasztott amerikai egyetemistákról tartalmaz információt, az adatsor leírása megtalálható a **12.2.** feladatban.

- Kérdezzük le a 'Gender' változó értékeinek tapasztalati és relatív gyakoriságait! Ennek alapján adjunk becslést a férfiak arányára és oddszára a sokaságon belül! Mi az oddsz szemléletes jelentése ebben a feladatban?
- Írjuk fel a 'Gender' és 'CatsDogs' változók kontingenciatáblázatát! Adjunk becslést arra, hogy a teljes sokaságban a nők illetve a férfiak milyen arányban tartanak macskát! Ezek alapján milyen becslés adható a 'cats' érték relatív kockázatára a nők körében a férfiakhoz viszonyítva?
- Teszteljük le 1% szignifikancia szinten azt a nullhipotézist, hogy a két változó független egymástól! Ha elvetjük a nullhipotézist, akkor hol érhető tetten, hogy a két változó nem független egymástól?
- Adjuk meg a 'Gender' és 'Smokes' változók kontingenciatáblázatát! Adjunk becslést arra, hogy a teljes sokaságban a nők illetve a férfiak milyen arányban dohányoznak! Adjunk becslést a dohányzás relatív kockázatára a férfiak körében a nőkhöz viszonyítva?
- Teszteljük le a 'Gender' és 'Smokes' változók függetlenségét!

13.3. Olvassuk be a 'carData' csomag 'Chile' adatsorát. A változók leírása megtalálható a **12.3.** feladatban.

- a. Írjuk fel a ‘region’ és ‘sex’ változók kontingenciatáblázatát! Teszteljük le 5% szignifikancia szinten azt a nullhipotézist, hogy a változók függetlenek egymástól, tehát az egyes régiókban azonos a férfiak és a nők egymáshoz viszonyított aránya.  
Miért tesztelünk olyan dolgot, amiről tudjuk, hogy igaz? Azt akarjuk ellenőrizni, hogy a minta reprezentatív, nem mond ellent a nyilvánvaló tényeknek.
- b. Határozzuk meg az ‘education’ és ‘vote’ változók kontingenciatáblázatát! Teszteljük le 5% szignifikancia szinten azt a nullhipotézist, hogy a két változó független egymástól, tehát az alanyok iskolai végzettsége nem befolyásolja Pinochet megítélését! Írjuk fel a várt gyakoriságok táblázatát is!
- c. Ha az előző pontban elvetettük a változók függetlenségét, akkor vizsgáljuk meg a khi-négyzet tagokat is! Magyarázzuk el, hogy az ‘education’ változó által definiált három csoport hogyan viszonyul Pinochet újraválasztásához!

## 14. Többszörös lineáris regresszió és többszemponos ANOVA

14.1. Az  $\xi$  és az  $\eta$  változóra az alábbi megfigyeléseket kaptunk:

$\xi$	1	2	4	5
$\eta$	2	0	8	6

- a. Határozzuk meg a Pearson-féle korrelációs együtthatót! Ezek alapján milyen irányú, milyen erősségű és milyen típusú kapcsolat van a két változó között?
- b. Végezzünk lineáris regressziót, adjunk becslést a  $\xi$  változóra az  $\eta$  alapján? Mennyire pontos ez a becslés?
- c. Milyen becslés adható a  $\xi$  változóra abban az esetben, amikor  $\eta = 5$ ?
- d. Ábrázoljuk a mintaelemeket és a regressziós egyenest koordináta-rendszerben!
- 14.2. Olvassuk be a ‘UScars.txt’ fájlban található statisztika adatsort, a leírásért lásd a 9.1. feladatot!

- a. Modellezzük az ‘MPG’ változót a következő formában:

$$\text{MPG} \approx a_1 \text{HP} + a_2 \text{VOL} + a_3 \text{WT} + b$$

Adjunk becslést az ismeretlen együtthatókra! Mely együtthatók különböznek szignifikáns módon a 0 értéktől? Mennyire pontos a kapott előrejelzés az MPG változóra?

- b. Vegyük ki az a. feladatrészben felírt modelltől azokat a tagokat, melyek nem voltak szignifikánsak, és végezzünk lineáris regressziót a megmaradt változókon. Milyen becslést adhatunk az együtthatókra, és mennyire pontos az előrejelzés az ‘MPG’ változóra?

- c. Bővítsük a **b.** feladatrészt modelljét, vegyük be az 1/HP magyarázó változót is! Milyen becslést adhatunk az ismeretlen együtthatókra, és mennyire pontos az előrejelzés? Mindegyik együttható különbözik szignifikáns módon nullától?

### 14.3.

- 14.4. A 'vernyomas2.xlsx' táblázat nagyrészt azonos a **7.1.** feladatban megismert adattal. Mindössze annyi a változás, hogy most az alanyok nemét is ismerjük, ez a 'NEM' változó. Értékei: F=férfi, N=nő.

- a. Végezzünk kétszemponos varianciaanalízist a SYS1 változón, és a modellbe vegyük be az interakciót is:  $E(\text{SYS1} \mid i, j) = m + a_i + b_j + c_{ij}$

Teszteljük le az alábbi nullhipotéziseket:

$$H_0 : a_1 = a_2 = a_3 = 0 \quad (\text{a 'CSOPNEV' szempontnál nincsen csoporthatás})$$

$$H_0 : b_1 = b_2 = 0 \quad (\text{a 'NEM' szempont esetében nincsen csoporthatás})$$

$$H_0 : c_{ij} = 0 \quad \text{minden } i \text{ és } j \text{ esetén} \quad (\text{nincsen interakciós hatás})$$

Befolyásolja a csoportbesorolás és a nem a SYS1 várható értékét?

- b. Teszteljük le a SYS1 változó esetében a cellánkénti szórások egyenlőségét is!  
 c. Ismételjük meg az a. feladatrészt elemzését a SYS2 változóra!  
 d. Adjunk becslést a SYS2 változó esetében a csoporthatásokra és az interakciós hatásokra!

## 15. Paraméteres próbák és rangpróbák

- 15.1. A  $\xi$  és  $\eta$  változókra az alábbi összetartozó mintákat kaptuk:

$\xi$	10	50	70	100	80
$\eta$	10	20	30	80	100

- a. Határozzuk meg a rangszámokat mindkét változó esetében! Ábrázoljuk az eredeti mintát illetve a rangszámokat koordináta-rendszerben!  
 b. Határozzuk meg a  $\xi$  változó rangszámainak  $\bar{R}_\xi$  mintaátlagát és  $D_n^*(R_\xi)$  korigált empirikus szórását!  
 c. Le lehet vezetni, hogy  $n$  elemű minta esetén a rangszámok mintaátlagja illetve szórása mindig az alábbi:

$$\bar{R}_\xi = \bar{R}_\eta = \frac{n+1}{2}, \quad D_n^*(R_\xi) = D_n^*(R_\eta) = \sqrt{\frac{n(n+1)}{12}}.$$

Behelyettesítéssel ellenőrizzük le, hogy az előző feladatrésztben tényleg ezeket az értékeket kaptuk meg!

- d. Adjuk meg a  $\xi$  és  $\eta$  változók Spearman-féle korrelációs együtthatóját! Ezek alapján a két változó között milyen kapcsolat tapasztalható?



**15.2.** Olvassuk be az ‘vernyomas.xlsx’ fájlban található statisztika adatsort, a leírásért lásd a **7.1.** feladatot! A szignifikancia szint a feladatban végig 5%.

- a. Ábrázoljuk a ‘SYS1’ változó hisztogramját és kérdezzük le a ferdeséget! A változó szimmetrikus eloszlásúnak tűnik?
- b. Egymintás t-próbával illetve Wilcoxon-féle előjeles rangpróbával teszteljük le azt a nullhipotézist, hogy a ‘SYS1’ várható értéke 160! Teszteljük le a 159-es értéket is! Hasonlítsuk össze a két próba által adott eredményeket!
- c. Válogassuk le a ‘kiserleti’ csoport tagjait és hozzuk létre a ‘SYS1’–‘SYS2’ változót. Ábrázoljuk az új változó hisztogramját és kérdezzük le a ferdeséget! A különbségváltozó szimmetrikus eloszlásúnak tűnik?
- d. Páros t-próba illetve páros Wilcoxon-próba alkalmazásával teszteljük le azt a nullhipotézist, hogy a ‘SYS1’ és ‘SYS2’ változóknak azonos a várható értéke a ‘kiserleti’ csoportban! Teszteljük le ugyanezt a nullhipotézist egymintás t-próba illetve Wilcoxon-féle előjeles rangpróba segítségével is!
- e. Maradjunk a ‘kiserleti’ csoport tagjainál! Páros t-próba illetve páros Wilcoxon-próba segítségével teszteljük le azt, hogy a ‘SYS1’ változó várható értéke 10 higanymilliméterrel magasabb, mint a ‘SYS2’ várható értéke!
- f. Térjünk vissza a teljes adatsorhoz, majd válogassuk le a ‘kiserleti2’ és a ‘kontroll’ betegcsoport tagjait! A kapott adatsoron ábrázoljuk a ‘SYS2’ változó hisztogramját csoportonkénti bontásban! A két sűrűségfüggvény egymás eltoltságának tűnik? Teszteljük le a szórások egyenlőségét is!
- g. Kétmintás t-próba és Mann–Whitney-féle U-próba alkalmazásával teszteljük le azt a nullhipotézist, hogy a ‘SYS2’ változónak azonos a várható értéke a ‘kiserleti2’ és a ‘kontroll’ csoportban! Teszteljük le azt is, hogy a ‘kiserleti2’ csoportban a ‘SYS2’ várható értéke 12 higanymilliméterrel magasabb, mint a ‘kontroll’ csoportban!

**15.3.** Cyril Burt egy pszichológus volt, aki azt vizsgálta, hogy az IQ pontszám kialakulásában a genetikai vagy a környezeti tényezőknek van nagyobb hatása. Olyan egyiptei ikreket hasonlított össze statisztikai módszerekkel, akikkel különböző családokban neveltek. Az ötlet egyszerű: ha a testvérek IQ pontszámai között erős kapcsolat van, akkor az intelligenciahányadost nagyrészt a genetikai tényezők határozzák meg; ha pedig a kapcsolat nem erős, akkor a környezeti tényezőknek is nagy szerepük van. A statisztikai elemzésből kiderült, hogy alapvetően a genetikai tényezőknek van jelentősége.

Cyrl Burt úttörő eredményeket ért el ezen a területen. Utólag sajnos kiderült, hogy a statisztikai adatokat egyszerűen meghamisította azért, hogy az elemzés az általa előzetesen várt eredményeket szolgáltatassa. A ‘carData’ csomag ‘Burt’ adatsora egy ilyen módosított(!) mintát tartalmaz. A változók:

IQbio: a biológiai szülők által nevelt testvér IQ pontszáma

IQfoster: a nevelőszülők által nevelt testvér IQ pontszáma

class: társadalmi osztály (értékei: high, medium, low)

- a. Hány megfigyelés (ikerpár) áll rendelkezésre az egyes társadalmi osztályokból? Ezek alapján bátran használhatjuk a paraméteres próbákat, vagy óvatosabban kell eljárunk?
- b. Ábrázoljuk az 'IQbio' változó hisztogramját a 'class' változó által definiált csoportok szerinti bontásban. Mit állíthatunk az egyes csoportokról szimmetria szempontjából?
- c. Teszteljük le azt a nullhipotézist, hogy a 'high' csoportban az 'IQbio' változó várható értéke 105! Teszteljük le ugyanezt a nullhipotézist a 'medium' csoportban is!
- d. Térjünk vissza a teljes adatsorhoz, és hozzunk létre egy új változót a következő formulával:  $IQdiff = IQfoster - IQbio$ . Ábrázoljuk az 'IQdiff' változó hisztogramját is csoportonkénti bontásban!
- e. Teszteljük le azt a nullhipotézist, hogy a 'high' csoportban a biológiai szülők illetve a nevelőszülők által nevelt testvéreknél azonos az IQ pontszám várható értéke!
- f. Teszteljük le azt a nullhipotézist, hogy a 'medium' csoportban a nevelőszülők által nevelt testvér IQ pontszáma átlagosan 10-zel magasabb, mint a biológiai szülők által nevelt testvért IQ pontszáma!
- g. Térjünk vissza a teljes adatsorhoz, majd válogassuk le azokat az alanyokat, akik nem a 'low' csoportba tartoznak! Teszteljük le azt a nullhipotézist, hogy a megmaradt két csoportban egyenlő az 'IQbio' változó szórása! Ezek után teszteljük le a várható értékek azonosságát is!

## 16. Diszkriminanciaanalízis

**16.1.** Az 'iris.xlsx' állomány három Kanadában honos írisz (nőszírom) fajról tartalmaz adatokat. Az adatsor leírása megtalálható a **7.2.** feladatban.

- a. Nyissuk meg az 'iris.xlsx' állományt Excelben, töröljük a 'setosa' növényeket, majd mentjük el a táblázatot tabulátorral tagolt szöveggé! Ezek után olvassuk be a szöveges állományt az R-be!
- b. Ábrázoljuk az adatsort koordináta rendszerben, jelöljük a két fajt két különböző színnel. A grafikonok alapján mely változók segítségével lehet a legjobban elkülöníteni a 'versicolor' illetve a 'virginica' fajt?
- c. Hozzunk létre egy 0–1 értékű változót, ami a növényfajt kódolja! Végezzünk ezen a változón logisztikus regressziót a 'szíromhossz' változó segítségével! Adjuk meg a regressziós függvényt, majd ábrázoljuk is a regressziós görbét koordináta-rendszerben! Határozzuk meg a 'szíromhossz' változó azon értékét,

ami elvágja egymástól a két csoportot! Ezen vágópont segítségével fogalmazzunk meg egy egyszerű szabályt arra, hogy mikor soroljuk a növényeket az egyes fajokba.

- d. Készítsünk egy összefoglaló táblázatot arról, hogy a logisztikus regresszió a tanulóminta elemei közül hányra ad helyes előrejelzést! Mennyire hatékony az előrejelzés a mintán?
- e. Az alábbi táblázat két olyan írisz növényt tartalmaz, melyekről nem tudjuk, hogy melyik fajhoz tartoznak. Adjunk előrejelzést a fajra, és adjuk meg az előrejelzések megbízhatóságát is!

cseszehossz	cseszessel	szíromhossz	szíromszel
6.3	3.2	4.5	1.1
7.5	2.8	4.9	2

- f. Végezzünk logisztikus regressziót a ‘cseszehossz’ változóval is, majd válaszoljunk a fenti kérdésekre!

**16.2.** A ‘UScars.txt’ adatsorban a ‘80-as években az amerikai piacon forgalmazott néhány autótípus fontosabb műszaki paraméterei szerepelnek. A részletes leírásért lásd a 9.1. feladatot!

- a. Olvassuk be az adatsort az R-be az ‘adat’ nevű táblázatba, majd válogassuk le azokat az autókat, melyek nem az Egyesült Államokban készültek. A leválogatást megtehetjük a következő parancs alkalmazásával:  
`adat = subset( adat, COUNTRY!="U.S." )`
- b. Ábrázoljuk az adatsort koordináta rendszerben, jelöljük az Európai illetve a Japán autókat különböző színekkel. A grafikonok alapján mely változók segítségével lehet a legjobban elkülöníteni a két országból származó járműveket?
- c. Hozzunk létre egy 0–1 értékű változót, ami az országokat kódolja, jelöljük a japán autókat 1-es kóddal! Végezzünk ezen a változón logisztikus regressziót az ‘MPG’ változó segítségével! Adjuk meg a regressziós függvényt, majd ábrázoljuk is a regressziós görbét koordináta-rendszerben! Határozzuk meg az ‘MPG’ változó azon értékét, ami elvágja egymástól a két csoportot! Ezen vágópont segítségével fogalmazzunk meg egy egyszerű szabályt arra, hogy az ‘MPG’ változó alapján milyen előrejelzést tehetünk a gyártási országra!
- d. Készítsünk egy összefoglaló táblázatot arról, hogy a logisztikus regresszió a tanulóminta elemei közül hányra ad helyes előrejelzést! Mennyire hatékony az előrejelzés a mintán?
- e. Az alábbi táblázat egy olyan autó adatait tartalmazza, melyről nem tudjuk, hogy melyik országban készült. Adjunk előrejelzést a származási országra, és adjuk meg az előrejelzés megbízhatóságát is!

VOL	HP	MPG	SP	WT
6.3	3.2	24	1.1	36

- f. Végezzünk logisztikus regressziót a ‘WT’ változóval is, majd válaszoljunk a fenti kérdésekre!

**16.3.** Az ‘iris.xlsx’ állomány három Kanadában honos írisz (nőszírom) fajról tartalmaz adatokat. Az adatsor leírása megtalálható a **7.2.** feladatban.

- a. Nyissuk meg az ‘iris.xlsx’ állományt Excelben majd mentjük el a táblázatot tabulátorral tagolt szöveggént! Olvassuk be a szöveges állományt az R-be!
- b. Ábrázoljuk az adatsort koordináta rendszerben, jelöljük a három fajt három különböző színnel. A grafikonok alapján melyik két változó segítségével lehet a legjobban elkülöníteni a három fajt?
- c. Végezzünk lineáris diszkriminanciaanalízist az adatsoron: adjunk előrejelzést a fajra a ‘szíromhossz’ és ‘szíromszel’ változók segítségével! Készítsünk egy összefoglaló táblázatot arról, hogy az algoritmust a tanulómintára alkalmazva hány esetben kapunk helyes előrejelzést! Mennyire hatékony az előrejelzés a mintán?
- d. Az alábbi táblázat két olyan írisz növényt tartalmaz, melyekről nem tudjuk, hogy melyik fajhoz tartoznak. Adjunk előrejelzést a fajra, és adjuk meg az előrejelzések megbízhatóságát is!

cseszehossz	cseszessel	szíromhossz	szíromszel
6.3	3.2	4.5	1.1
7.5	2.8	4.9	2

- e. Oldjuk meg az előző két feladatrészt olyan módon is, hogy az előrejelzés során mind a négy változót felhasználjuk!

**16.4.** A ‘UScars.txt’ adatsorban a ‘80-as években az amerikai piacon forgalmazott néhány autótípus fontosabb műszaki paraméterei szerepelnek. A részletes leírásért lásd a **9.1.** feladatot!

- a. Olvassuk be az adatsort az R-be! Ábrázoljuk az adatsort koordináta rendszerben, jelöljük a három fajt három különböző színnel. A grafikonok alapján melyik két változó segítségével lehet a legjobban elkülöníteni a három fajt?
- b. Végezzünk lineáris diszkriminanciaanalízist az adatsoron: adjunk előrejelzést a fajra a gyártási országra a ‘HP’ és ‘WT’ változók segítségével! Készítsünk egy összefoglaló táblázatot arról, hogy az algoritmust a tanulómintára alkalmazva hány esetben kapunk helyes előrejelzést! Mennyire hatékony az előrejelzés a mintán?
- c. Az alábbi táblázat egy olyan autó adatait tartalmazza, melyről nem tudjuk, hogy melyik országban készült. Adjunk előrejelzést a származási országra, és adjuk meg az előrejelzés megbízhatóságát is!

VOL	HP	MPG	SP	WT
120	110	24	100	36

- d. Oldjuk meg az előző két feladatrészt olyan módon is, hogy az előrejelzés során mind az öt numerikus változót felhasználjuk!

## Megoldások

1.1.  $\xi =$  a kiválasztott játékos testmagassága

$$P(\xi \geq 200) = 50\% = 0.5, \quad E(\xi) = 199.33, \quad D(\xi) = 3.4$$

1.2.  $R_\xi = \{1, 3, 7, 12\}$

$$P(\xi = 1) = 0.1, \quad P(\xi = 3) = 0.4, \quad P(\xi = 7) = 0.3, \quad P(\xi = 12) = 0.2$$

módusz = 3, jelentése: a legnagyobb arányban előforduló érték

$E(\xi) = 5.8$ , jelentése: a fejek számának átlagos értéke

$D(\xi) = 3.7$ , jelentése: a várható értéktől vett átlagos eltérés

1.3.  $R_\xi = \{0, 1, 2, 3\}$

$$P(\xi = 0) = 0.4, \quad P(\xi = 1) = 0.3, \quad P(\xi = 2) = 0.2, \quad P(\xi = 3) = 0.1$$

$$P(\xi > 1) = 0.3 = 30\%$$

módusz = 0, jelentése: a legnagyobb arányban előforduló érték

$E(\xi) = 1$ , jelentése: átlagosan ennyi fa található az 1 hektáros négyzetekben

$D(\xi) = 1$ , jelentése: a várható értéktől vett átlagos eltérés

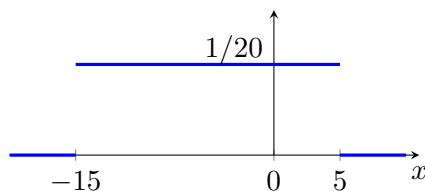
1.4.  $R_\xi = \{1000, 2000, 3000, 5000\}$

$$P(\xi = 1000) = 0.5, \quad P(\xi = 2000) = 0.3, \quad P(\xi = 3000) = 0.15, \quad P(\xi = 5000) = 0.05$$

$$\text{módusz} = 1000, \quad E(\xi) = 1800, \quad D(\xi) = 1030$$

$$P(\xi \leq 2000) = 0.8$$

2.1.a.  $R_\xi = [-15, +5]$



b.  $P(-10 \leq \xi \leq +10) = \int_{-10}^{+10} f_\xi(x) dx = 0.75$

$$P(\xi \geq 0) = \int_0^{+5} f_\xi(x) dx = 0.25$$

c.

$$F_{\xi}(t) = \begin{cases} 0, & t < -15, \\ \frac{t+15}{20}, & -15 \leq t \leq +5, \\ 1, & +5 < t, \end{cases}$$

$$P(-10 \leq \xi \leq +10) = F_{\xi}(+10) - F_{\xi}(-10) = 0.75$$

$$P(\xi \geq 0) = P(0 \leq \xi \leq +5) = F_{\xi}(+5) - F_{\xi}(0) = 0.25$$

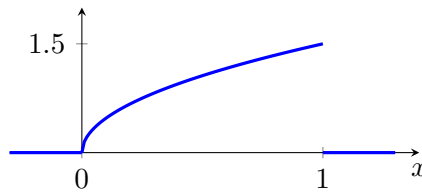
d.  $q_{\alpha} = 20\alpha - 15$ ,  $q_{80\%} = +1$

Jelentése: a napi középhőmérséklet 80% eséllyel lesz +1 Celsiusnál alacsonyabb

$E(\xi) = -5$ , jelentése: a napi középhőmérséklet átlagos értéke januárban

$D(\xi) = 5.77$ , jelentése: a napi középhőmérséklet átlagosan ennyivel tér el a  $-5$  fokos várható értéktől

2.2.a.  $R_{\xi} = [0, 1]$



b.  $P(0.5 \leq \xi \leq 1.5) = \int_{0.5}^{1.5} f_{\xi}(x) dx = 0.65$ , a fák 65 százaléka esik 0.5 és 1.5 közé

$P(\xi \leq 0.8) = \int_0^{0.8} f_{\xi}(x) dx = 0.72$ , a fák 72 százaléka legfeljebb 0.8 átmérőjű

c.

$$F_{\xi}(t) = \begin{cases} 0, & t < 0, \\ t^{3/2}, & 0 \leq t \leq 1, \\ 1, & 1 < t, \end{cases}$$

$$P(0.5 \leq \xi \leq 1.5) = F_{\xi}(1.5) - F_{\xi}(0.5) = 0.65$$

$$P(\xi \leq 0.8) = P(0 \leq \xi \leq 0.8) = F_{\xi}(0.8) - F_{\xi}(0) = 0.72$$

d.  $E(\xi) = 0.6$ , jelentése: a törzs átlagos átmérője az erdőben

$D(\xi) = 0.26$ , jelentése: a törzs átmérője átlagosan ennyivel tér el a várható értéktől

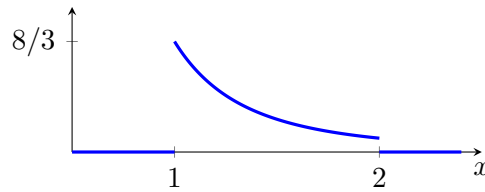
$$q_{\alpha} = \alpha^{2/3}, \quad q_{25\%} = 0.4, \quad q_{50\%} = 0.63, \quad q_{75\%} = 0.83$$

Jelentésük: az alábbi intervallumok mindegyikébe a fák negyede esik:

$[0, 0.4]$ ,  $[0.4, 0.63]$ ,  $[0.63, 0.83]$ ,  $[0.83, 1]$

2.3.  $\xi =$  egy véletlenszerűen kiválasztott egyed testhossza

a.  $R_{\xi} = [1, 2]$



b.  $P(0.5 \leq \xi \leq 1.5) = \int_{0.5}^{1.5} f_{\xi}(x) dx = 0.74$

$P(\xi \geq 1.8) = \int_{1.8}^2 f_{\xi}(x) dx = 0.08$

c.

$$F_{\xi}(t) = \begin{cases} 0, & t < 1, \\ \frac{4}{3} - \frac{4}{3t^2}, & 1 \leq t \leq 2, \\ 1, & 2 < t, \end{cases}$$

$P(0.5 \leq \xi \leq 1.5) = F_{\xi}(1.5) - F_{\xi}(0.5) = 0.74$

$P(\xi \geq 1.8) = P(1.8 \leq \xi \leq 2) = F_{\xi}(2) - F_{\xi}(1.8) = 0.08$

d.  $E(\xi) = 1.33$ , jelentése: átlagos érték a populációban

$D(\xi) = 0.27$ , jelentése: a várható értéktől vett átlagos eltérés

$q_{\alpha} = \sqrt{\frac{4}{4-3\alpha}}$ ,  $q_{33.3\%} = 1.15$ ,  $q_{66.6\%} = 1.41$

Az intervallumok:  $[1, 1.15]$ ,  $[1.15, 1.41]$ ,  $[1.41, 2]$

3.1. a.  $f_3$ ; b.  $f_4$ ; c.  $f_2$ . Kimaradt sűrűségfüggvény ( $f_1$ ):  $\mu = -3$ ,  $\sigma = 0,5$ .

3.2. 66%; 2%;  $[70.6, 129.4]$

3.3. 41%; 9%;  $[5.77, 8.23]$

3.4. 47%;  $[402, 598]$

4.1.a.  $E(\text{repwt}) \approx \overline{\text{repwt}} = 65.62$  (empirikus várható érték, mintaátlag)

$D(\text{repwt}) \approx D_n^*(\text{repwt}) = 13.78$  (korigált empirikus szórás)

b. Mintaméret:  $n = 183$ , hiányzó adatok száma: 17

c.  $q_{25\%} \approx \hat{q}_{25\%} = 55$ ,  $q_{50\%} \approx \hat{q}_{50\%} = 63$ ,  $q_{75\%} \approx \hat{q}_{75\%} = 73.5$

d. Minimum = legkisebb mintaelem = 41

Maximum = legnagyobb mintaelem = 124

Terjedelem = minimum – maximum = 83

Jelentése: ilyen hosszúságú intervallumon helyezkedik el a teljes minta.

$IQR = \hat{q}_{75\%} - \hat{q}_{25\%} = 18.5$

Jelentése: ilyen hosszúságú intervallumon helyezkedik el a minta középső 50%-a.

e.  $\hat{q}_{33.3\%} = 57$ ,  $\hat{q}_{66.6\%} = 69.2$

Az intervallumok:  $[41, 57]$ ,  $[57, 69.2]$ ,  $[69.2, 124]$

4.2.a.  $E(\text{age}) \approx \overline{\text{age}} = 42.54$  (empirikus várható érték, mintaátlag)

$D(\text{age}) \approx D_n^*(\text{age}) = 8.07$  (korrigált empirikus szórás)

$q_{50\%} \approx \hat{q}_{50\%} = 43$  (empirikus medián)

b.  $n = 753$ , nincs hiányzó adat

c. Minimum = legkisebb mintaelem = 30

Maximum = legnagyobb mintaelem = 60

$\hat{q}_{25\%} = 36$ , jelentése: a minta alsó negyedelőpontja

$\hat{q}_{75\%} = 49$ , jelentése: a minta felső negyedelőpontja

Terjedelem = minimum – maximum = 30

Jelentése: ilyen hosszúságú intervallumon helyezkedik el a teljes minta.

IQR =  $\hat{q}_{75\%} - \hat{q}_{25\%} = 13$

Jelentése: ilyen hosszúságú intervallumon helyezkedik el a minta középső 50%-a.

d.  $\hat{q}_{20\%} = 34$ ,  $\hat{q}_{40\%} = 40$ ,  $\hat{q}_{60\%} = 45$ ,  $\hat{q}_{80\%} = 50$

Az intervallumok:  $[30, 34]$ ,  $[34, 40]$ ,  $[40, 45]$ ,  $[45, 50]$ ,  $[50, 60]$

4.3.a.  $n = 7$ ,  $\bar{\xi} = 1254.3$ ,  $D_n(\xi) = 48.1$ ,  $D_n^*(\xi) = 51.9$

A lelőhely igazi kora =  $E(\xi) \approx \bar{\xi} = 1254.3$

$D(\xi) \approx D_n^*(\xi) = 51.9$

A kis mintaméret miatt a korrigált empirikus szórás pontosabb becslés.

b. Empirikus medián = középső mintaelem = 1250

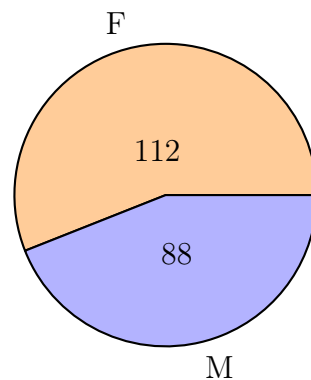
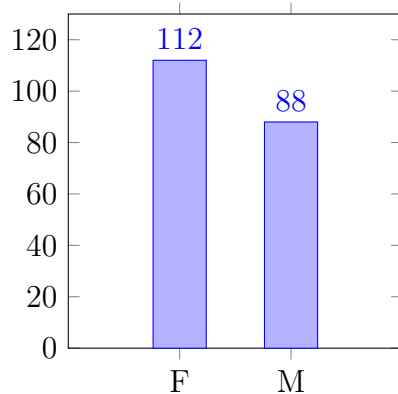
Terjedelem = maximum – minimum = 160

5.1.a. Diszkrét változók: sex.

Folytonos változók: weight, height, repwt, repht.

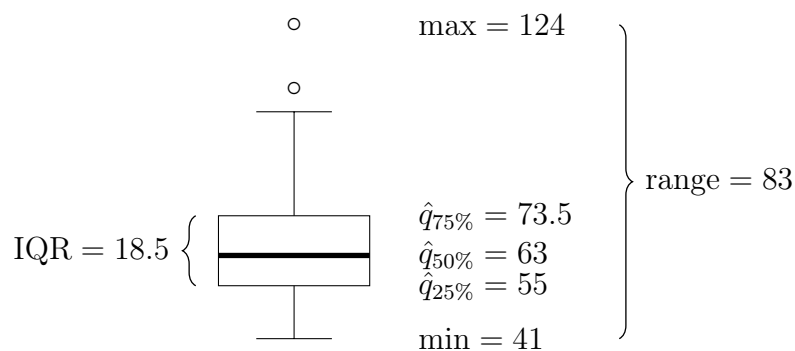
b. A mintában 112 nő és 88 férfi szerepel.





c. skewness = 1.04, a hisztogram jobbra ferde.

d. A mintában 2 outlier érték található.

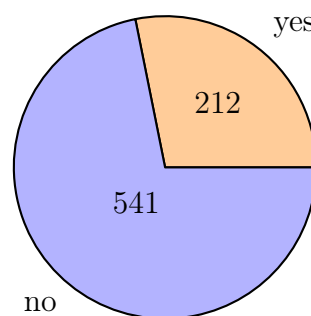
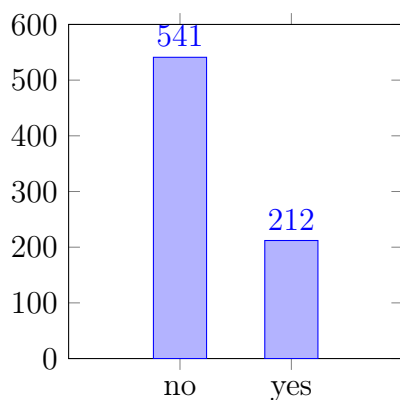


e. A változó enyhén jobbra ferde. A minta valószínűleg nem normális eloszlásból származik, de az eloszlás nem különbözik nagy mértékben a normálistól.

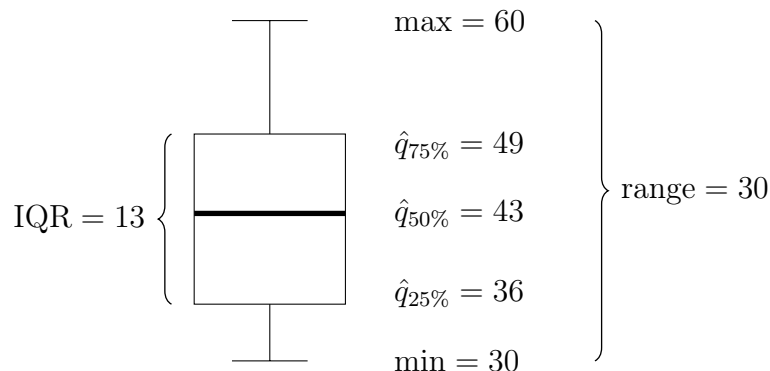
f. A 12. alanynál van egy elírás, felcserélték a weight illetve a height változó értékét.

5.2.a. Diszkrét változók: wc, k5. Folytonos változó: age.

b. 212 alany rendelkezik, 541 nem rendelkezik diplomával.



- c. A hisztogram ránézésre jobbra ferde. Viszont a skewness értéke 0.15, tehát a minta igazából közel szimmetrikus.
- d. A mintában nem található outlier érték.



- e. A boxplot és a skewness alapján a változó közel normális eloszlásúnak tűnik. Ezzel szemben a hisztogramra egyáltalán nem illeszkedik a normális eloszlás sűrűségfüggvénye. Emiatt a végső konklúzió az, hogy változó eloszlása nagyban különbözik a normálistól.
- 5.3.a. A megfigyelések száma körülbelül 200. A legkisebb elem valahol 0 és 10 között, a legnagyobb 40 és 50 között található. A ferdeség negatív előjelű. A grafikonon egy móduszt látunk valahol 30 és 40 között.

b. Minimum: 0, alsó kvartilis: 10, medián: 25, felső kvartilis: 40, maximum: 50. A terjedelem 50, az IQR 30. Nincs outlier érték.

c. A két ábrán azonosnak tűnik a minimum és a maximum értéke, de ez önmagában semmit sem jelent.

A hisztogram ferde, a boxplot ezzel szemben szimmetrikus. Viszont a boxploton nem mindig jelennek meg a részletek, tehát ez sem bizonyító erejű.

A boxploton az elemek negyede esik 0 és 10 közé, és megint csak a negyedük 40 és 50 közé. Ezzel szemben a hisztogramon ezen intervallumokba csak körülbelül 10–10 elem esik, ami a 200 elemű mintának jóval kisebb hányada. Tehát a két grafikon nem azonos mintához tartozik.

6.1.a.  $E(\xi) \approx \bar{\xi} = 1230$ ;  $SE = 15.17$

A standard hiba jelentése: egy  $n = 5$  elemű minta alapján átlagosan ekkora hibával lehet megbecsülni a várható értéket.

b. Most  $\alpha = 0.05$  és  $n = 5$ , ezért  $c_\alpha = \Phi_4^{-1}(0.975) = 2.776$ .

Konfidencia intervallum:  $[\bar{\xi} - c_\alpha SE, \bar{\xi} + c_\alpha SE] = [1187.89, 1272.11]$ .

c. Nullhipotézis:  $H_0 : E(\xi) = 1200$ .

Hipotetikus várható érték:  $\mu_0 = 1200$ .

Próba statisztika:  $t = (\bar{\xi} - \mu_0)/SE = 1.978$ .

Döntés:  $|t| \leq c_\alpha$ , ezért a nullhipotézist elfogadjuk. A minta alapján hihető, hogy a lelőhely igazi kora 1200 év.

d. Elsőfajú hiba: 5%.

Másodfajú hiba: nem ismerjük a nagyságát.

**6.2.a.**  $E(\xi) \approx \bar{\xi} = 990$ ;  $SE = 4.83$

A standard hiba jelentése: egy  $n = 6$  elemű minta alapján átlagosan ekkora hibával lehet megbecsülni a várható értéket.

b. Most  $\alpha = 0.1$  és  $n = 6$ , ezért  $c_\alpha = \Phi_5^{-1}(0.95) = 2.015$ .

Konfidencia intervallum:  $[\bar{\xi} - c_\alpha SE, \bar{\xi} + c_\alpha SE] = [980.27, 999.73]$ .

c. Nullhipotézis:  $H_0 : E(\xi) = 1000$ .

Hipotetikus várható érték:  $\mu_0 = 1000$ .

Próba statisztika:  $t = (\bar{\xi} - \mu_0)/SE = -2.07$

Döntés:  $|t| > c_\alpha$ , ezért a nullhipotézist elvetjük. A minta alapján nem hihető, hogy a töltőberendezés jól van beállítva.

**7.1.a.**  $E(\text{SYS1}) \approx \overline{\text{SYS1}} = 160.2$ ,  $D(\text{SYS1}) \approx D_n^*(\text{SYS1}) = 5.7$ .

A becslés várható hibája:  $SE = 0.46$ .

b. A hisztogram közel szimmetrikus, skewness = 0.06, a normális eloszlás sűrűségfüggvénye jól illeszkedik. Normális vagy közel normális eloszlásról van szó.

c.  $H_0 : E(\text{SYS1}) = 160$ , egymintás t-próba: p-érték=0.626, elfogadjuk.

$H_0 : E(\text{SYS1}) = 165$ , egymintás t-próba: p-érték=0.000, a nullhipotézist elvetjük.

Konfidencia intervallum: [159.31, 161.14]. Ezek a „hihető” várható értékek, a teszt az intervallumba eső értékeket fogadja el igazi várható értékeknek.

d. A ‘kiserleti1’ csoportban:  $E(\text{SYS1}) \approx 160.02$ ,  $E(\text{SYS2}) \approx 150.5$ .

A hisztogramok alapján mindkét változó közel normális eloszlású.

e.  $H_0 : E(\text{SYS1}) = E(\text{SYS2})$ , páros t-próba, p-érték=0.000, a nullhipotézist elvetjük.

Konfidencia intervallum az  $E(\text{SYS1}) - E(\text{SYS2})$  különbségre: [6.98, 12.06].

f. A ‘kiserleti2’ csoportban:  $E(\text{SYS1}) \approx 159.36$ ,  $E(\text{SYS2}) \approx 160.86$ .

$H_0 : E(\text{SYS1}) = E(\text{SYS2})$ , páros t-próba, p-érték=0.11, a nullhipotézist elfogadjuk.

Konfidencia intervallum az  $E(\text{SYS1}) - E(\text{SYS2})$  különbségre: [-3.98, 0.98].

**7.2.a.** Legalább 2, de talán 3 módusz is van. Ennek az az oka, hogy az adatsor több különböző fajról tartalmaz adatokat. Érdemesebb az elemzéseket nem a teljes adatsoron, hanem inkább fajonkénti bontásban elvégezni.

**b.** A hisztogram alapján a ‘sziromszel’ változó nem tűnik normális eloszlásúnak, de a szimmetria miatt ez még közel normális.

**c.** A ‘virginica’ növényeknél:  $E(\text{sziromszel}) \approx 2.03$ ,  $D(\text{sziromszel}) \approx 0.27$ .

$H_0 : E(\text{sziromszel}) = 2$ , egymintás t-próba: p-érték=0.51, elfogadjuk.

Konfidencia intervallum a várható értékre:  $[1.95, 2.10]$ .

**d.** A ‘virginica’ növényeknél:  $E(\text{cseszszel}) \approx 2.97$

$H_0 : E(\text{sziromszel}) = E(\text{cseszszel})$ , páros t-próba, p-érték=0, elvetjük.

Konfidencia intervallum az  $E(\text{sziromszel}) - E(\text{cseszszel})$  különbségre:  $[-1.02, -0.88]$ . ■

**e.** Azonos módszerekkel, mint az előző pontokban.

**8.1.a.** A becslések csoportonkénti bontásban:

	mean	sd
kiserleti1	160.02	6.19
kiserleti2	159.36	5.11
kontroll	161.30	5.64

Nem látható jelentős eltérés a mintaátlagok és a korrigált empirikus szórások között, illetve a boxplotok is nagyon hasonlóak. Emiatt megfogalmazhatjuk azt a sejtést, hogy a három csoportban azonos az elméleti várható érték és az elméleti szórás.

**b.** A ferdeségek és a hisztogramok alapján a ‘SYS1’ változó mindhárom csoportban normális vagy közel normális.

	skewness
kiserleti1	0.10
kiserleti2	-0.23
kontroll	0.17

**c.** Nullhipotézis: a csoportonkénti szórások azonosak. Formálisan:

$$H_0 : D(\text{SYS1} \mid \text{kiserleti1}) = D(\text{SYS1} \mid \text{kiserleti2}) = D(\text{SYS1} \mid \text{kontroll})$$

Levene-teszt, p-érték=0.24. A nullhipotézist elfogadjuk, nincs szignifikáns eltérés a szórások között.

Nullhipotézis: a csoportonkénti várható értékek azonosak. Formálisan:

$$H_0 : E(\text{SYS1} \mid \text{kiserleti1}) = E(\text{SYS1} \mid \text{kiserleti2}) = E(\text{SYS1} \mid \text{kontroll})$$

Egyszempontos ANOVA, p-érték=0.22. A nullhipotézist elfogadjuk, nem találtunk szignifikáns eltérést a várható értékek között.

d. A becslések csoportonkénti bontásban:

	mean	sd	skewness
kiserleti1	150.50	2.45	0.20
kiserleti2	160.86	5.30	-0.27
kontroll	149.44	6.19	0.22

A becslések és a boxpotok alapján a várható értékek és a szórások között is van különbség. A ferdeségek és a hisztogramok alapján a csoportonkénti normalitás rendben van.

$$H_0 : D(\text{SYS2} \mid \text{kiserleti1}) = D(\text{SYS2} \mid \text{kiserleti2}) = D(\text{SYS2} \mid \text{kontroll})$$

Levene-teszt, p-érték=0.000. Elvetjük a szórások egyenlőségét.

$$H_0 : E(\text{SYS2} \mid \text{kiserleti1}) = E(\text{SYS2} \mid \text{kiserleti2}) = E(\text{SYS2} \mid \text{kontroll})$$

Welch-féle F-próba, p-érték=0.000. Elvetjük a várható értékek egyenlőségét.

A páronkénti összehasonlítás alapján a 'kiserleti1' és a 'kontroll' csoport között nincs szignifikáns eltérés, de a 'kiserleti2' csoport már szignifikáns módon különbözik. Becslés és konfidencia intervallum a csoportonkénti várható értékek különbségére:

	becslés	konf. int.
kiserleti2 – kiserleti1	10.36	[8.03, 12.69]
kontroll – kiserleti1	-1.06	[-3.39, 1.27]
kontroll – kiserleti2	-11.42	[-13.75, -9.09]

8.2.a. A becslések csoportonkénti bontásban:

	mean	sd
setosa	0.246	0.105
versicolor	1.326	0.198
virginica	2.026	0.275

A mintaátlagok között látványos az eltérés, és valószínűleg a szórások sem lesznek egyenlőek. Ugyanez jelenik meg a boxploton is.

b. A 'setosa' fajon belül a változó jobbra ferde, de ez még tekinthető közel szimmetrikusnak. A másik két fajnál a változó közel szimmetrikus.

	skewness
setosa	1.25
versicolor	-0.03
virginica	-0.13

c.  $H_0 : D(\text{sziromszel} \mid \text{setosa}) = D(\text{sziromszel} \mid \text{versicolor}) = D(\text{sziromszel} \mid \text{virginica})$

Levene-teszt: p-érték=0.000, elvetjük a nullhipotézist.

d.  $H_0 : E(\text{sziromszel} \mid \text{setosa}) = E(\text{sziromszel} \mid \text{versicolor}) = E(\text{sziromszel} \mid \text{virginica})$

Welch-féle F-próba: p-érték=0.000, elvetjük a nullhipotézist.

A páronkénti összehasonlítás alapján szignifikáns különbség van mindhárom várható érték között. Becslés és konfidencia intervallum a várható értékek különbségére:

	becslés	konf. int.
versicolor – setosa	1.08	[0.98, 1.18]
virginica – setosa	1.78	[1.68, 1.88]
virginica – versicolor	0.70	[0.60, 0.80]

- e. A ‘cseeszszel’ változó esetében a Levene-teszt elfogadja a csoportonkénti szórások azonosságát. Emiatt alkalmazhatjuk az ANOVA tesztet, ami elveti a várható értékek egyenlőségét.

**9.1.a.**  $SP \approx 0.24 \cdot HP + 84.45$

$R^2 = 0.93$ , jó az illeszkedés a regressziós egyeneshez, a becslés pontos

Ha  $HP = 150$ , akkor  $SP \approx 0.24 \cdot 150 + 84.45 = 120.45$

**b.**  $SP \approx -0.03 \cdot VOL + 115.11$

$R^2 = 0.002$ , nagyon rossz az illeszkedés a regressziós egyeneshez, a becslés pontatlan, a gyakorlatban nem alkalmazható

**c.** Lineáris regresszió:  $MPG \approx -0.14 \cdot HP + 50.07$ ,  $R^2 = 0.62$

Reciprokos regresszió:

Új változó:  $repHP = 1/HP$

$MPG \approx 2373.11 \cdot repHP + 9.73 = 2373.11/HP + 9.73$ ,  $R^2 = 0.84$

Exponenciális regresszió:

Új változó:  $\log MPG = \log(MPG)$

$\log MPG \approx -0.0046 \cdot HP + 4.01$ ,  $R^2 = 0.73$

$MPG \approx \exp(-0.0046 \cdot HP + 4.01)$

A három módszer közül a reciprokos regresszió adja a legjobb becslést.

**9.2.a.**  $SATV \approx 0.86 \cdot SATM + 21.53$

$R^2 = 0.93$ , jó az illeszkedés a regressziós egyeneshez, a becslés pontos

Ha  $SATM = 500$ , akkor  $SATV \approx 0.86 \cdot 500 + 21.53 = 451.53$

**b.**  $dollars \approx 0.00004 \cdot pop + 4.998$

$R^2 = 0.02$ , nagyon rossz az illeszkedés a regressziós egyeneshez, a becslés a gyakorlatban nem alkalmazható

**c.** Első formula:

Új változó:  $repPercent = 1/percent$

$SATV \approx 428.1 \cdot repPercent + 421$ ,  $R^2 = 0.69$

$$\text{SATV} \approx 428.1/\text{percent} + 421$$

Második formula:

$$\text{Új változó: repSATV} = 1/\text{SATV}$$

$$\text{repSATV} \approx 0.0000054 \cdot \text{percent} + 0.002, \quad R^2 = 0.74$$

$$\text{SATV} \approx 1/(0.0000054 \cdot \text{percent} + 0.002)$$

A második formula egy kicsivel jobb illeszkedést biztosít, de a két becslés közel azonos pontosságú.

**10.1.a.**  $H_0$  : ‘SP’ és ‘HP’ független változók

Pearson-féle korrelációs teszt: p-érték = 0.000, a nullhipotézist elvetjük

$$\text{Pearson-féle korrelációs együttható: } r_n(\text{SP}, \text{HP}) = 0.97$$

A teszt alapján a két változó között lineáris kapcsolat tapasztalható. A korrelációs együttható alapján a kapcsolat erős és pozitív irányú.

Spearman-féle korrelációs teszt: p-érték = 0.000, a nullhipotézist elvetjük

$$\text{Spearman-féle korrelációs együttható: } \rho_n(\text{SP}, \text{HP}) = 0.88$$

A teszt alapján a két változó között rendezési kapcsolat tapasztalható. A korrelációs együttható alapján a kapcsolat erős és pozitív irányú.

**b.**  $H_0$  : ‘SP’ és ‘VOL’ független változók

Pearson-féle korrelációs teszt: p-érték = 0.7, a nullhipotézist elfogadjuk

$$\text{Pearson-féle korrelációs együttható: } r_n(\text{SP}, \text{VOL}) = -0.04$$

A teszt alapján a két változó között nem tapasztalható lineáris kapcsolat, ezért elfogadjuk a függetlenséget.

Spearman-féle korrelációs teszt: p-érték = 0.005, a nullhipotézist elvetjük

$$\text{Spearman-féle korrelációs együttható: } \rho_n(\text{SP}, \text{VOL}) = 0.31$$

A teszt alapján a két változó között rendezési kapcsolat tapasztalható, ezért elvetjük a függetlenséget. A kapcsolat pozitív irányú, de nagyon gyenge.

**c.**  $H_0$  : ‘MPG’ és ‘HP’ független változók

Pearson-féle korrelációs teszt: p-érték = 0.000, a nullhipotézist elvetjük

$$\text{Pearson-féle korrelációs együttható: } r_n(\text{MPG}, \text{HP}) = -0.79$$

A teszt alapján a két változó között lineáris kapcsolat tapasztalható. A korrelációs együttható alapján a kapcsolat közepesen erős és negatív irányú.

Spearman-féle korrelációs teszt: p-érték = 0.000, a nullhipotézist elvetjük

$$\text{Spearman-féle korrelációs együttható: } \rho_n(\text{MPG}, \text{HP}) = -0.91$$

A teszt alapján a két változó között rendezési kapcsolat tapasztalható. A korrelációs együttható alapján a kapcsolat erős és negatív irányú.

**10.2.a.**  $H_0$  : ‘SATM’ és ‘SATV’ független változók

Pearson-féle korrelációs teszt: p-érték = 0.000, a nullhipotézist elvetjük

Pearson-féle korrelációs együttható:  $r_n(\text{SATM}, \text{SATV}) = 0.96$

A teszt alapján a két változó között lineáris kapcsolat tapasztalható. A korrelációs együttható alapján a kapcsolat erős és pozitív irányú.

Spearman-féle korrelációs teszt: p-érték = 0.000, a nullhipotézist elvetjük

Spearman-féle korrelációs együttható:  $\rho_n(\text{SATM}, \text{SATV}) = 0.95$

A teszt alapján a két változó között rendezési kapcsolat tapasztalható. A korrelációs együttható alapján a kapcsolat erős és pozitív irányú.

**b.**  $H_0$  : ‘pop’ és ‘dollars’ független változók

Pearson-féle korrelációs teszt: p-érték = 0.31, a nullhipotézist elfogadjuk

Pearson-féle korrelációs együttható:  $r_n(\text{pop}, \text{dollars}) = 0.14$

A teszt alapján a két változó között nem tapasztalható lineáris kapcsolat, ezért elfogadjuk a függetlenséget.

Spearman-féle korrelációs teszt: p-érték = 0.54, a nullhipotézist elvetjük

Spearman-féle korrelációs együttható:  $\rho_n(\text{pop}, \text{dollars}) = 0.09$

A teszt alapján a két változó között nem tapasztalható rendezési kapcsolat, ezért elfogadjuk a függetlenséget.

**c.**  $H_0$  : ‘percent’ és ‘SATV’ független változók

Pearson-féle korrelációs teszt: p-érték = 0.000, a nullhipotézist elvetjük

Pearson-féle korrelációs együttható:  $r_n(\text{percent}, \text{SATV}) = -0.86$

A teszt alapján a két változó között lineáris kapcsolat tapasztalható. A korrelációs együttható alapján a kapcsolat közepesen erős és negatív irányú.

Spearman-féle korrelációs teszt: p-érték = 0.000, a nullhipotézist elvetjük

Spearman-féle korrelációs együttható:  $\rho_n(\text{percent}, \text{SATV}) = -0.85$

A teszt alapján a két változó között rendezési kapcsolat tapasztalható. A korrelációs együttható alapján a kapcsolat erős és negatív irányú.

**11.1.a.** A hisztogram és a ferdeség alapján a ‘SYS1’ változó normális eloszlásúnak tűnik. A ‘SYS2’ változó enyhén jobbra ferde, inkább tűnik nem normálisnak.

**b.** A QQ-ploton mindkét változó esetében közel helyezkednek el a megfigyelések az egyeneshez. Ezek alapján mindkét változó normális eloszlásúnak tűnik.

**c.**  $H_0$  : a ‘SYS1’ változó normális eloszlású a teljes populációban

Shapiro–Wilk-próba, p-érték=0.55, a nullhipotézist elfogadjuk.

$H_0$  : a ‘SYS2’ változó normális eloszlású a teljes populációban



Shapiro–Wilk-próba, p-érték=0.006, a nullhipotézist elvetjük.

- d. A normalitást a Shapiro–Wilk-próba alkalmazásával tudjuk tesztelni csoportonként. Minden csoportban elfogadjuk a nullhipotézist, ugyanis a p-értékek:

csoport	p-érték
kiserleti1	0.437
kiserleti2	0.759
kontroll	0.230

- 11.2.a. A ‘cseeszszel’ változó enyhén jobbra ferde. Potenciálisan normális, de biztosan közel normális eloszlású. A ‘szziromszel’ változó többmódusú, valószínűleg nem normális eloszlású.

- b. A ‘cseeszszel’ változó esetében nem rossz az ábra, bár a széleken nem illeszkednek a megfigyelések az egyeneshez. A ‘szziromszel’ változó esetében nagyon rossz a helyzet.

- c.  $H_0$  : a ‘cseeszszel’ változó normális eloszlású a teljes populációban

Shapiro–Wilk-próba, p-érték=0.101, a nullhipotézist elfogadjuk.

$H_0$  : a ‘szziromszel’ változó normális eloszlású a teljes populációban

Shapiro–Wilk-próba, p-érték=0.000, a nullhipotézist elvetjük.

- d. A normalitást fajonkénti bontásban is a Shapiro–Wilk-próbával tudjuk tesztelni. Az alábbi táblázat tartalmazza a p-értékeket. A ‘cseeszszel’ változó normalitását minden csoportban elfogadjuk, míg a ‘szziromszel’ változó normalitását minden csoportban elvetjük.

faj	cseeszszel	szziromszel
setosa	0.271	0.000
versicolor	0.338	0.027
virginica	0.181	0.087

- 12.1.a. A statisztikai minta a dobott számok sorozata, a minta elemszáma  $n = 100$ . Az ismeretlen valószínűségeket a relatív gyakoriságokkal becsülhetjük:

dobás ( $x_i$ )	1	2	3	4	5	6	össz.
$k_n(x_i)$	15	15	15	15	20	20	100
$r_n(x_i)$	0.15	0.15	0.15	0.15	0.2	0.2	1

- b. A szabályosság tesztelése:

$H_0$  :  $P(\text{dobás} = 1) = \dots = P(\text{dobás} = 6) = 1/6$ .

dobás ( $x_i$ )	1	2	3	4	5	6	össz.
$p_i$	1/6	1/6	1/6	1/6	1/6	1/6	1
$np_i$	16.67	16.67	16.67	16.67	16.67	16.67	100

Khi-négyzet próba valószínűségek tesztelésére:  $\chi^2 = 2$ ,  $c_\alpha = 11.07$ , elfogadjuk.

A hatosdobás tesztelése:

$H_0 : P(\text{dobás} = 6) = 1/6, P(\text{dobás} \neq 6) = 5/6.$

dobás ( $x_i$ )	6	nem 6	össz.
$k_n(x_i)$	20	80	100
$r_n(x_i)$	0.2	0.8	1
$p_i$	1/6	5/6	1
$np_i$	16.67	83.33	100

Khi-négyzet próba valószínűségek tesztelésére:  $\chi^2 = 0.8, c_\alpha = 3.841,$  elfogadjuk.

c. Az elemszám  $n = 1000,$  a relatív gyakoriságok azonosak az előző feladatrésszel.

A szabályosság tesztelése:

$H_0 : P(\text{dobás} = 1) = \dots = P(\text{dobás} = 6) = 1/6.$

Khi-négyzet próba valószínűségek tesztelésére:  $\chi^2 = 20, c_\alpha = 11.07,$  elvetjük.

A hatosdobás tesztelése:

$H_0 : P(\text{dobás} = 6) = 1/6, P(\text{dobás} \neq 6) = 5/6.$

Khi-négyzet próba valószínűségek tesztelésére:  $\chi^2 = 8, c_\alpha = 3.841,$  elvetjük.

**12.2.a.** A teljes sokaságban mért arányokat a relatív gyakoriságokkal becsülhetjük.

$x_i$	both	cats	dogs	none	össz.
$k_n(x_i)$	24	156	200	620	1000
$r_n(x_i)$	0.024	0.156	0.2	0.62	1

$P(\text{CatsDogs} = \text{both}) \approx 0.024, P(\text{CatsDogs} = \text{cats}) \approx 0.156,$

$P(\text{CatsDogs} = \text{dogs}) \approx 0.2, P(\text{CatsDogs} = \text{noness}) \approx 0.62$

b.  $H_0 : P(\text{CatsDogs} = \text{both}) = 0.05, P(\text{CatsDogs} = \text{cats}) = 0.15,$

$P(\text{CatsDogs} = \text{dogs}) = 0.2, P(\text{CatsDogs} = \text{noness}) = 0.6$

Khi-négyzet próba valószínűségek tesztelésére: p-érték=0.002, elvetjük.

$x_i$	both	cats	dogs	none	össz.
$p_i$	0.05	0.15	0.2	0.6	1
$np_i$	50	150	200	600	1000

A 'both' értéknél van a legnagyobb eltérés a tapasztalati és a várt gyakoriság között.

c. A teljes sokaságban mért arányokat a relatív gyakoriságokkal becsülhetjük.

$x_i$	no	yes	össz.
$k_n(x_i)$	776	224	1000
$r_n(x_i)$	0.776	0.224	1

$P(\text{Dogs} = \text{yes}) \approx 0.224, P(\text{Dogs} = \text{no}) \approx 0.776$

$H_0 : P(\text{Dogs} = \text{yes}) = 0.2, P(\text{Dogs} = \text{no}) = 0.8$

Khi-négyzet próba valószínűségek tesztelésére: p-érték=0.058, elfogadjuk.

**12.3.a.** A gyakoriságok és relatív gyakoriságok:

$x_i$	F	M
$k_n(x_i)$	1379	1321
$r_n(x_i)$	0.51	0.49

$$H_0 : P(\text{sex} = F) = 0.5, P(\text{sex} = M) = 0.5$$

Khi-négyzet próba valószínűségek tesztelésére: p-érték=0.264, elfogadjuk.

**b.** A teljes sokaságban mért arányokat a relatív gyakoriságokkal becsülhetjük.

$x_i$	P	PS	S
$k_n(x_i)$	1107	462	1120
$r_n(x_i)$	0.41	0.17	0.42

$$H_0 : P(\text{education} = P) = 0.4, P(\text{education} = PS) = 0.2, P(\text{education} = S) = 0.4$$

Khi-négyzet próba valószínűségek tesztelésére: p-érték=0.001, elvetjük.

**13.1.a.** A marginális eloszlások:

	fekeete haj	barna haj	szőke haj	összesen
sötét szem	25%	30%	5%	60%
világos szem	5%	20%	15%	40%
összesen	30%	50%	20%	100%

$$\text{oddsz}(\text{sötét szem}) = 1.5$$

A sötét szemű emberek másfélszer annyian vannak, mint a nem sötét szeműek.

$$\text{oddsz}(\text{szőke haj}) = 1/4$$

A szőke hajú emberek negyedannyian vannak, mint a nem szőke hajúak.

**b.**  $P(\text{fekeete haj} \mid \text{sötét szem}) = 41.7\%$

Jelentése: a fekete haj aránya a sötét szemű emberek körében.

$$P(\text{világos szem} \mid \text{szőke haj}) = 75\%$$

Jelentése: a világos szem aránya a szőke hajú emberek körében.

**c.** A hajszín és a szemszín nem független egymástól. Például:

$$P(\text{fekeete haj} \mid \text{sötét szem}) = 41.7\% \neq P(\text{fekeete haj}) = 30\%$$

A sötét szem megnöveli a fekete haj valószínűségét.

$$P(\text{világos szem} \mid \text{szőke haj}) = 75\% \neq P(\text{világos szem}) = 40\%$$

A szőke haj megnöveli a világos szem valószínűségét.

Függetlenség esetén az alábbi arányokat látnánk a sokaságban:

	fekeete haj	barna haj	szőke haj	összesen
sötét szem	18%	30%	12%	60%
világos szem	12%	20%	8%	40%
összesen	30%	50%	20%	100%

d.  $P(\text{világos szem} \mid \text{szőke haj}) = 75\% = 0.75$

$P(\text{világos szem} \mid \text{fekete haj}) = 1/6$

RR = 4.5. A szőke hajú emberek körében a világos szem 4.5-szer gyakrabban fordul elő, mint a fekete hajú emberek körében.

13.2.a. Tapasztalati és relatív gyakoriságok:

$x_i$	F	M
$k_n(x_i)$	416	584
$r_n(x_i)$	0.416	0.584

$P(M) \approx r_n(M) = 58.4\%$ ,  $\text{oddsz}(M) = \frac{P(M)}{1-P(M)} \approx \frac{0.584}{1-0.584} = 1.4$

Odds jelentése: a sokaságban 1.4-szer annyi férfi van, mint nem férfi (tehát nő).

b. A tapasztalati gyakoriságok:

	both	cats	dogs	none
F	15	97	98	206
M	9	59	102	414

$P(\text{cats} \mid \text{female}) \approx 23.3\%$ ,  $P(\text{cats} \mid \text{male}) \approx 10.1\%$

RR = 2.3. A nők körében 2.3-szer gyakoribb a macskatartás, mint a férfiak körében.

c. Függelenség esetén a várt gyakoriságok:

	both	cats	dogs	none
F	10.0	64.9	83.2	257.9
M	14.0	91.1	116.8	362.0

$H_0$  : a 'Gender' és a 'CatsDogs' változó független egymástól

Khi-négyzet próba függetlenség tesztelésére: p-érték=0.000, elvetjük.

A khi-négyzet komponensek alapján a tapasztalati és a várt gyakoriságok a 'cats' oszlopban térnek el a legnagyobb mértékben egymástól. Emellett jelentős eltérések vannak a 'none' oszlopban is.

d. Tapasztalati gyakoriságok:

	no	yes
F	345	71
M	462	122

$P(\text{Smokes} = \text{yes} \mid \text{Gender} = \text{M}) \approx 0.209$ ,  $P(\text{Smokes} = \text{yes} \mid \text{Gender} = \text{F}) \approx 0.171$

RR = 1.22. A férfiaknál a dohányosok aránya 1.22-ször nagyobb, mint nők körében.

e.  $H_0$  : a 'Gender' és a 'Smokes' változó független egymástól

Khi-négyzet próba függetlenség tesztelésére: p-érték=0.13, elfogadjuk.

**13.3.a.** Együttes gyakoriságok:

	C	M	N	S	SA
F	300	51	168	362	498
M	300	49	154	356	462

$H_0$  : a 'region' és 'sex' változók független egymástól

Khi-négyzet próba függetlenség tesztelésére: p-érték=0.938, elfogadjuk.

**b.** Bal oldalon a tapasztalati gyakoriságok, jobb oldalon a várt gyakoriságok:

	A	N	U	Y		A	N	U	Y
P	52	266	296	422	P	76.8	364.4	240.3	354.5
PS	32	224	52	130	PS	32.5	154.0	101.6	149.9
S	103	397	237	311	S	77.7	368.6	243.1	358.6

$H_0$  : az 'education' és 'vote' változók független egymástól

Khi-négyzet próba függetlenség tesztelésére: p-érték=0.000, elvetjük.

**c.** A khi-négyzet komponensek táblázata:

	A	N	U	Y
P	8.02	26.56	12.91	12.85
PS	0.01	31.77	24.21	2.64
S	8.23	2.19	0.15	6.32

Középfokú végzettség (S): A khi-négyzet komponensek relatíve alacsonyok, tehát sehol sincs nagy mértékű eltérés a tapasztalati és a várt gyakoriság között. Ez azt jelenti, hogy ezen a csoporton belül a választók hasonló arányban támogatják illetve ellenzik Pinochet hatalmát, mint a teljes lakosságon belül.

Felsőfokú végzettség (PS): Az 'A' és az 'Y' oszlopban a khi-négyzet komponens értéke alacsony, tehát a tapasztalati gyakoriság közel van a várt gyakorisághoz. Viszont az 'N' és az 'U' oszlopban magas értékeket találunk, itt jelentős eltérések vannak az országos arányokhoz viszonyítva. A gyakorisági táblázatok alapján a teljes népességhez viszonyítva a diplomások körében jóval alacsonyabb a bizonytalanok aránya, és jóval magasabb az ellenzékiek aránya.

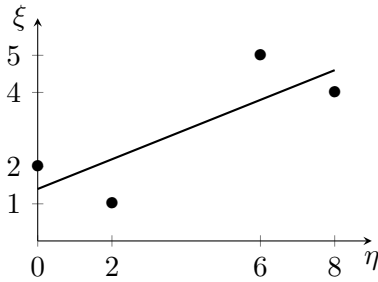
Alapfokú végzettség (P): Az 'N', az 'U' és az 'Y' oszlopban is magas a khi-négyzet komponens értéke, jelentős az eltérés a tapasztalati és a várt gyakoriság között. A teljes népességhez viszonyítva az alapfokú végzettséggel rendelkezők körében magasabb Pinochet támogatóinak az aránya és alacsonyabb az ellenzékiek aránya. Emellett náluk relatív sok a bizonytalan is.

**14.1.a.**  $\bar{\xi} = 3$ ,  $\bar{\eta} = 4$ ,  $D_n^*(\xi) = 1.83$ ,  $D_n^*(\eta) = 3.65$ ,  $r_n(\xi, \eta) = 0.8$

**b.**  $\xi \approx 1.6\eta - 0.8$ ,  $R^2 = 0.64$

**c.**  $\xi \approx 0.4 \cdot 5 + 1.4 = 3.4$

d.



14.2.a.  $a_1 \approx -0.031$ ,  $a_2 \approx -0.034$ ,  $a_3 \approx -0.894$ ,  $b \approx 68.44$

$$\text{MPG} \approx -0.031 \cdot \text{HP} - 0.034 \cdot \text{VOL} - 0.894 \cdot \text{WT} + 68.44, \quad R^2 = 0.83$$

A HP és a VOL változó együtthatója nem különbözik szignifikáns módon a 0 értéktől, ezeket ki lehet hagyni a modelltől.

b.  $\text{MPG} \approx a\text{WT} + b$ ,

$$a \approx -1.11, \quad b \approx 68.17$$

$$\text{MPG} \approx -1.11 \cdot \text{WT} + 68.17, \quad R^2 = 0.82$$

c.  $\text{MPG} \approx a_1/\text{HP} + a_2\text{WT} + b$

$$a_1 \approx 1387.18, \quad a_2 \approx -0.54, \quad b \approx 36.54$$

$$\text{MPG} \approx 1387.18/\text{HP} - 0.54 \cdot \text{WT} + 36.54, \quad R^2 = 0.89$$

Mindegyik együttható szignifikáns módon különbözik nullától.

14.4.a. Kétszemponos ANOVA interakcióval:

$H_0 : a_1 = a_2 = a_3 = 0$ , p-érték=0.228, elfogadjuk.

$H_0 : b_1 = b_2 = 0$ , p-érték=0.318, elfogadjuk.

$H_0 : c_{ij} = 0$  minden  $i$  és  $j$  esetén, p-érték=0.986, elfogadjuk.

Nincsen sem csoportthatás, sem interakciós hatás, a SYS1 változó várható értékét nem befolyásolják a vizsgált szempontok.

b.  $H_0 : a$  SYS1 változó szórása azonos minden cellában

Levene-próba: p-érték=0.115, elfogadjuk.

c. Kétszemponos ANOVA interakcióval:

$$E(\text{SYS2} \mid i, j) = m + a_i + b_j + c_{ij}$$

$H_0 : a_1 = a_2 = a_3 = 0$ , p-érték=0.000, elvetjük.

$H_0 : b_1 = b_2 = 0$ , p-érték=0.431, elfogadjuk.

$H_0 : c_{ij} = 0$  minden  $i$  és  $j$  esetén, p-érték=0.003, elvetjük.

Az első szempont (CSOPNEV) szerinti csoportthatás és az interakció szignifikánsan különbözik nullától. A második szempont (NEM) szerint nincs csoportthatás.

d. Az  $\overline{\text{SYS2}}_{ij}$  cellánkénti mintaátlagokat a program megadja:

	F	N
kiserleti1	149.36	151.64
kiserleti2	162.96	158.76
kontroll	149.40	149.48

A bázisérték becslése az  $\overline{\text{SYS2}}_{ij}$  értékek számtani átlaga:  $m \approx \hat{m} = 153.6$

Az  $a_1, a_2, a_3$  hatások szignifikáns módon különböznek egymástól, ezeket becsülni kell. Ehhez először ki kell számolni az  $\overline{\text{SYS2}}_{ij}$  értékek soronkénti számtani átlagát:

$i$	CSOPNEV	sorátlag	$\hat{a}_i = \text{sorátlag} - \hat{m}$
1	kiserleti1	150.50	-3.10
2	kiserleti2	160.86	+7.26
3	kontroll	149.44	-4.16

A  $b_1, b_2$  hatások nem különböznek szignifikáns módon nullától:  $b_1 = b_2 = 0$ .

Viszont a későbbiek miatt ezeket is meg kell becsülni. Ehhez először kiszámoljuk az  $\overline{\text{SYS2}}_{ij}$  értékek oszloponkénti számtani átlagát:

$j$	NEM	oszlopátlag	$\hat{b}_j = \text{oszlopátlag} - \hat{m}$
1	F	153.91	+0.32
2	N	153.29	-0.31

Megjegyzés: a csoporthatások összege valóban nulla mindkét szempont esetében.

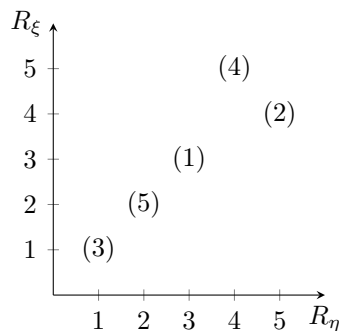
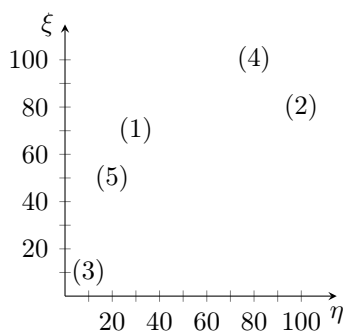
Most van interakció, az egyes cellák hatásának a becslése:  $\hat{c}_{ij} = \overline{\text{SYS2}}_{ij} - \hat{m} - \hat{a}_i - \hat{b}_j$ . Ez található az alábbi táblázatban:

	F	N
kiserleti1	-1.44	+1.44
kiserleti2	+1.79	-1.79
kontroll	-0.35	+0.35

Megjegyzés: a cellák hatásának valóban nulla az összege minden sorban és minden oszlopban.

15.1.a. A rangszámok:

$R_\xi$	1	2	3	5	4
$R_\eta$	1	2	3	4	5



b.  $\bar{R}_\xi = 3$ ,  $D_n^*(R_\xi) = 1.58$

c. Ugyanazt kapjuk.

d.  $C_n(R_\xi, R_\eta) = 2.25$ ,  $\rho_n(\xi, \eta) = r_n(R_\xi, R_\eta) = 0.9$

Erős pozitív irányú rendezési kapcsolat.

**15.2.a.**  $\text{skenewss} = 0.06$ , a változó szimmetrikus eloszlásúnak tűnik.

b.  $H_0 : E(\text{SYS1}) = 160$

Egymintás t-próba: p-érték=0.626, elfogadjuk.

Wilcoxon-féle előjeles rangpróba: p-érték=0.733, elfogadjuk.

c.  $\text{skenewss} = 0.3$ , a változó közelítőleg szimmetrikus eloszlásúnak tűnik.

d.  $H_0 : E(\text{SYS1} \mid \text{kiserleti1}) = E(\text{SYS2} \mid \text{kiserleti1})$

Páros t-próba:  $t=10.064$ , p-érték=0.000, elvetjük.

Páros Wilcoxon-féle előjeles rangpróba:  $V=1164.5$ , p-value=0.000, elvetjük.

$H_0 : E(\text{SYS1} - \text{SYS2} \mid \text{kiserleti1}) = 0$

Egymintás t-próba:  $t=10.064$ , p-érték=0.000, elvetjük.

Pontosan ugyanezt kaptuk a páros t-próbával is.

Wilcoxon-féle előjeles rangpróba:  $V=1164.5$ , p-value=0.000, elvetjük.

Pontosan ugyanezt kaptuk a páros Wilcoxon-próbával is.

e. Új változó:  $\text{SYS2plusz10} = \text{SYS2} + 10$

$H_0 : E(\text{SYS1} \mid \text{kiserleti1}) = E(\text{SYS2} \mid \text{kiserleti1}) + 10$

Ezzel ekvivalens:  $H_0 : E(\text{SYS1} \mid \text{kiserleti1}) = E(\text{SYS2plusz10} \mid \text{kiserleti1})$

Páros t-próba: p-érték=0.614, elfogadjuk.

Páros Wilcoxon-féle előjeles rangpróba: p-érték= 0.573, elfogadjuk.

f. A hisztogramok alapján a két sűrűségfüggvény egymás eltoltjának tűnik.

$H_0 : D(\text{SYS2} \mid \text{kiserleti2}) = D(\text{SYS2} \mid \text{kontroll})$

Levene-teszt: p-érték=0.333, elfogadjuk.

g.  $H_0 : E(\text{SYS2} \mid \text{kiserleti2}) = E(\text{SYS2} \mid \text{kontroll})$

Kétmintás t-próba: p-érték=0.000, elvetjük.

Mann-Whitney-féle U-próba: p-érték=0.000, elvetjük.

$H_0 : E(\text{SYS2} \mid \text{kiserleti2}) = E(\text{SYS2} + 12 \mid \text{kontroll})$

Új változó:  $\text{SYS2plusz} = \text{SYS2} + (2 - \text{CSOPKOD}) \cdot 6$

A nullhipotézis:  $H_0 : E(\text{SYS2plusz} \mid \text{kiserleti2}) = E(\text{SYS2plusz} \mid \text{kontroll})$



Kétmintás t-próba: p-érték=0.616, elfogadjuk.

Mann–Whitney-féle U-próba: p-érték=0.79, elfogadjuk.

**15.3.a.** A ‘class’ változó értékeinek gyakorisága a mintában:

Érték	high	low	medium
Gyakoriság	7	14	6

Az alacsony mintaméretetek miatt a paraméteres próbákkal óvatosan kell bánni, szerencsésebb rangpróbákat alkalmazni.

**b.** A hisztogram minden csoportban közel szimmetrikus, a ‘low’ csoportban talán még a normalitás is rendben van.

**c.**  $H_0 : E(\text{IQbio} \mid \text{high}) = 105$

Wilcoxon-féle előjeles rangpróba: p-érték=0.938, elfogadjuk.

$H_0 : E(\text{IQbio} \mid \text{medium}) = 105$

Wilcoxon-féle előjeles rangpróba: p-érték=0.063, elfogadjuk, de eléggé határeset.

**d.** A szimmetria nagyrészt rendben van. Ez azt jelenti, hogy a páros Wilcoxon-próba alkalmazható.

**e.** 1. megoldás:  $H_0 : E(\text{IQbio} \mid \text{high}) = E(\text{IQfoster} \mid \text{high})$

Páros Wilcoxon-féle előjeles rangpróba: p-érték=0.141, elfogadjuk.

2. megoldás:  $H_0 : E(\text{IQdiff} \mid \text{high}) = 0$

Wilcoxon-féle előjeles rangpróba: p-érték=0.141, elfogadjuk.

**f.** 1. megoldás:  $\text{IQplusz10} = \text{IQbio} + 10$

$H_0 : E(\text{IQfoster} \mid \text{medium}) = E(\text{IQbioplusz10} \mid \text{medium})$

Páros Wilcoxon-féle előjeles rangpróba: p-érték=0.063, elfogadjuk.

2. megoldás:  $H_0 : E(\text{IQdiff} \mid \text{medium}) = 10$

Wilcoxon-féle előjeles rangpróba: p-érték=0.063, elfogadjuk.

**g.**  $H_0 : D(\text{IQbio} \mid \text{high}) = D(\text{IQbio} \mid \text{medium})$

Levene-teszt: p-érték=0.194, elfogadjuk.

$H_0 : E(\text{IQbio} \mid \text{high}) = E(\text{IQbio} \mid \text{medium})$

Mann–Whitney-féle U-próba: p-érték=0.252, elfogadjuk.

**16.1.b.** A ‘szíromhossz’ illetve a ‘szíromszel’ változó alkalmazásával számíthatunk a legjobb elkülönítésre.

c. Regressziós függvény:

$$p(x) = P(\text{a növény a 'virginica' fajhoz tartozik} \mid \text{szíromhossz} = x) = \frac{1}{1 + e^{-9x+43.78}}$$

Vágópont: 4.86. Ha szíromhossz  $> 4.86$ , akkor 'virginica' az előrejelzés; ha pedig szíromhossz  $< 4.86$ , akkor 'versicolor'.

d. A teljes mintán 100 növényből 93 jól lett besorolva, ez 93 százalékos hatékonyság.

igazi faj	előrejelzés	
	versicolor	virginica
virginica	46	4
versicolor	3	47

e. Első növény: szíromhossz = 4.5  $< 4.86$ , ezért a 'versicolor' fajba soroljuk.

Megbízhatóság:

$$P(\text{a növény a 'versicolor' fajhoz tartozik} \mid \text{szíromhossz} = 4.5) = 1 - p(4.5) = 96.3\%$$

Második növény: szíromhossz = 4.9  $> 4.86$ , ezért a 'virginica' fajba soroljuk.

Megbízhatóság:

$$P(\text{a növény a 'virginica' fajhoz tartozik} \mid \text{szíromhossz} = 4.9) = p(4.9) = 58\%$$

f. Regressziós függvény:

$$p(x) = P(\text{a növény a 'virginica' fajhoz tartozik} \mid \text{cseszehossz} = x) = \frac{1}{1 + e^{-2.01x+12.57}}$$

Vágópont: 6.25. Ha cseszehossz  $> 6.25$ , akkor 'virginica' az előrejelzés; ha pedig cseszehossz  $< 6.25$ , akkor 'versicolor'.

A teljes mintán 100 növényből 73 jól lett besorolva, ez 73 százalékos hatékonyság.

igazi faj	előrejelzés	
	versicolor	virginica
virginica	36	14
versicolor	13	37

Első növény: cseszehossz = 6.3  $> 6.25$ , ezért a 'virginica' fajba soroljuk.

Megbízhatóság:

$$P(\text{a növény a 'virginica' fajhoz tartozik} \mid \text{cseszehossz} = 6.3) = p(6.3) = 52\%$$

Második növény: cseszehossz = 7.5  $> 6.25$ , ezért ezt is a 'virginica' fajba soroljuk.

Megbízhatóság:

$$P(\text{a növény a 'virginica' fajhoz tartozik} \mid \text{cseszehossz} = 7.5) = p(7.5) = 92\%$$

**16.2.b.** Az 'MPG' és a 'WT' változó alkalmazásával számíthatunk a legjobb elkülönítésre.

c. Regressziós függvény:

$$p(x) = P(\text{az autó Japánból származik} \mid \text{MPG} = x) = \frac{1}{1 + e^{-0.25x+6.07}}$$

Vágópont: 24.28. Ha  $\text{MPG} > 24.28$ , akkor 'Japan' az előrejelzés; ha  $\text{MPG} < 24.28$ , akkor pedig 'Europe'.

d. A teljes mintán 49 autóból 44 jól lett besorolva, ez 89.8 százalékos hatékonyság.

igazi ország	előrejelzés	
	Europe	Japan
Europe	7	3
Japan	2	37

e.  $\text{MPG} = 24 < 24.28$ , ezért 'Europe' az előrejelzés.

Megbízhatóság:  $P(\text{az autó Európában készült} \mid \text{MPG} = 24) = 1 - p(24) = 51.7\%$

f. Regressziós függvény:

$$p(x) = P(\text{az autó Japánban készült} \mid \text{WT} = x) = \frac{1}{1 + e^{0.32x-12.05}}$$

Vágópont: 37.66. Ha  $\text{WT} < 37.66$ , akkor 'Japan' az előrejelzés; ha  $\text{WT} > 37.66$ , akkor pedig 'Europe'.

Ismét 44 autó lett jól besorolva, tehát 89.8 százalékos a hatékonyság.

igazi ország	előrejelzés	
	Europe	Japan
Europe	7	3
Japan	2	37

Ismeretlen autó:  $\text{WT} = 36 > 37.66$ , ezért a 'Japan' az előrejelzés.

Megbízhatóság:  $P(\text{az autó Japánban készült} \mid \text{WT} = 36) = p(36) = 62.9\%$

**16.3.b.** A 'szíromhossz' és 'szíromszel' változókkal lehet a legjobban megkülönböztetni a három fajt.

c. Az algoritmus a 150 növényből 144 növényt sorol be helyesen, ez 96 százalékos hatékonyság.

igazi faj	előrejelzés		
	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	4	46

d. Első növény: a ‘versicololor’ fajba soroljuk. Megbízhatóság:

$$P(\text{a 'versicolor' fajhoz tartozik} \mid \text{szíromhossz} = 4.5, \text{szíromszel} = 1.1) = 99.99\%$$

Második növény: a ‘virginica’ fajba soroljuk. Megbízhatóság:

$$P(\text{a 'virginica' fajhoz tartozik} \mid \text{szíromhossz} = 4.9, \text{szíromszel} = 2) = 98.28\%$$

e. Négy magyarázó változó esetén a 150 növényből 147 lesz jól besorolva, ami 98 százalékos hatékonyság.

igazi faj	előrejelzés		
	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

Első növény: a ‘versicololor’ fajba soroljuk. Megbízhatóság:

$$P(\text{'versicolor' } \mid \text{a magyarázó változók értéke } 6.3, 3.2, 4.5, 1.1) = 100\%$$

Második növény: a ‘virginica’ fajba soroljuk. Megbízhatóság:

$$P(\text{'virginica' } \mid \text{a magyarázó változók értéke } 7.5, 2.8, 4.9, 2) = 71.44\%$$

**16.3.a.** Egyik grafikon sem biztató, talán a ‘HP’ és ‘WT’ változókkal különül el legjobban a három csoport.

b. A 82 autóból 52 autót soroltunk be helyesen, ez 63.4 százalékos hatékonyság.

igazi ország	előrejelzés		
	Europe	Japan	U.S.
Europe	5	0	5
Japan	2	30	7
U.S.	0	16	17

c. A modell szerint az autó az Egyesült Államokban készült. Megbízhatóság:

$$P(\text{U.S.} \mid \text{HP} = 110, \text{WT} = 36) = 70.5\%$$

d. Öt magyarázó változó esetén a 82 autóból 56 lesz jól besorolva, ami 68.3 százalékos hatékonyság.

igazi ország	előrejelzés		
	Europe	Japan	U.S.
Europe	5	0	5
Japan	2	31	6
U.S.	0	13	20

A modell szerint az extra autó Európából származik. Megbízhatóság:

$$P(\text{Europe} \mid \text{a magyarázó változók értéke } 120, 110, 24, 100, 36) = 95.7\%$$