Statisztika mérnököknek

1. Véges sokaságok és 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 ξ a fejek számát a választott egyednél! Adjuk meg a ξ változó értékkészletét 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 ξ az egyedek száma egy véletlenszerűen kiválasztott négyzetben!

Adjuk meg a ξ változó értékkészletét és valószínűségeloszlását! Mennyi az esélye, hogy a kiválasztott négyzeten 1-nél több egyed található a fafajból? Mennyi a ξ 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 ξ a nyeremény nagyságát! Adjuk meg a ξ változó értékkészletét, valószínűségeloszlását, móduszát, várható értékét és szórását! Mennyi az esélye annak, hogy legfeljebb 2000 forintot nyerünk?

2. Folytonos valószínűségi változók

2.1. Jelölje ξ a napi középhőmérsékletet Celsiusban egy januári napon. A ξ egy folytonos valószínűségi változó, melynek sűrűségfüggvénye $f(x) = 1/20$ ha $-15 \leq x \leq 5$, és $f(x) = 0$ minden más x esetén.

a. Vázlatosan rajzoljuk fel a sűrűségfüggvény grafikonját, és adjuk meg a ξ változó értékkészletét!

b. Mennyi annak az esélye, hogy a napi középhőmérséklet $-10°$C és $10°$C közé esik? Mekkora valószínűséggel lesz a napi középhőmérséklet legalább $0°$C?

c. Határozzuk meg a ξ változó eloszlásfüggvényét! Ezek után válaszoljunk az előző pont kérdéseire az eloszlásfüggvény alkalmazásával!

d. Adjuk meg a napi középhőmérséklet várható értékét és szórását! Mi ezeknek a mutatószámoknak a szemléletes jelentése?

e. Adjuk meg a ξ valószínűségi változó 80%-os kvantilisét!

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 ξ ezen egyed átmérője!

a. Vázlatosan rajzoljuk fel a sűrűségfüggvény grafikonját, és adjuk meg a ξ változó értékkészletét!

b. Mennyi a $\mathrm{P}(0.5 \leq \xi \leq 1.5)$ és $\mathrm{P}(\xi \leq 0.8)$ valószínűségek értéke? Mi a jelentése ezeknek az értékeknek az erdő szempontjából?

c. Határozzuk meg a ξ változó eloszlásfüggvényét! Ezek után válaszoljunk az előző pont kérdéseire az eloszlásfüggvény alkalmazásával!

d. Adjuk meg a törzs átmérőjének várható értékét és a szórását! Mi a szemléletes jelentése ezeknek a mutatószámnak?

e. Határozzuk meg a ξ változó mediánját illetve az alsó és a felső kvartilisét! Hogyan bontja fel ez a három osztáspont a teljes sokaságot?

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 ξ változó eloszlásfüggvényét! Ezek után válaszoljunk az előző pont kérdéseire az eloszlásfüggvény alkalmazásával!

d. Határozzuk meg a testhossz várható értéké és szórását!

e. Adjunk meg három intervallumot a testhosszra olyan módon, hogy mindegyikbe az egyedek harmada essen!

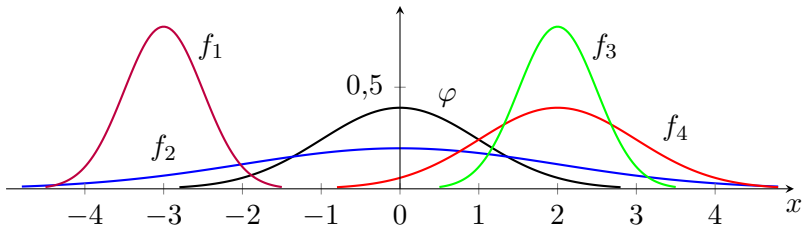
3. A normális eloszlás

3.1. Az alábbi ábrán φ a standard normális eloszlás sűrűségfüggvénye. Határozzuk meg, hogy az f_1, f_2, f_3, f_4 sűrűségfüggvények közül melyik tartozik az alábbi μ várható értékkel és σ 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 teszteket úgy állítják össze, hogy az eredmény a felnőtt népességen belül normális eloszlást kövessen 100 pont várható értékkel és 15 pont szórással.

a. A felnőtt népesség mekkora hányadának esik az IQ pontszáma 90 és 120 közé?

b. A Mensa egy nemzetközi egyesület, ahol a belépés feltétele a legalább 131 pontos IQ. A népesség hány százaléka felel meg ennek a követelménynek?

c. 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 szórás 0.75 óra.

a. A majmok mekkora hányada kel fel 6 és 7 óra között?

b. Mekkora hányad ébred 8 óra után?

c. Adjunk meg egy olyan időintervallumot, amelyre teljesül, hogy a majmok 90 százaléka ebben az időintervallumban mászik le a fáról!

3.4. Szegeden az éves csapadékmennyiség egy olyan ξ 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, amely 95% valószínűséggel tartalmazza a ξ változót!

4. Statisztikai becslések

4.1. Régészek radiokarbonos kormeghatározással szeretnék meghatározni egy lelőhely korát. Sajnos a radiokarbonos módszer egy adott ásatáson nem pontosan ugyanazt a kort adja minden lelet esetében. Az egyes leletek radiokarbonos kora egy ξ 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.

a. Határozzuk meg a mintaméretet, a mintaátlagot, valamint a korrigálatlan és a korrigált empirikus varianciát és szórást! Ezek alapján milyen becslés adható a lelőhely igazi korára? A két empirikus szórás közül melyikkel érdemes becsülni a ξ változó igazi szórását?

b. Adjuk meg a standard hiba értékét! Mi ennek a szemléletes jelentése?

c. Adjuk meg a minta mediánját és terjedelmét!

4.2. Bejelentés érkezik a fogyasztóvédelemhez, hogy az egyik tejgyár 1 literes kiszerelésű 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 ξ 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.

a. Határozzuk meg a mintaméretet, a mintaátlagot, valamint a korrigálatlan és a korrigált empirikus varianciát és szórást! Ezek alapján milyen becslés adható a ξ változó igazi várható értékére és szórására?

b. Mennyire pontos a várható értékre adott becslés?

c. Adjuk meg a minta mediánját és terjedelmét!

4.3. 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. Adjunk becslést a 'repwt' változó teljes sokaságban vett várható értékére és szórására! Nevezzük meg, hogy mely statisztikai becsléseket alkalmaztuk!

b. Hány megfigyelés van a 'repwt' változóra, és mennyi a hiányzó adatok száma?

c. Adjuk meg és értelmezzük a standard hibát!

d. Határozzuk meg a 'repwt' változó esetében a következő mutatószámok értékét: minimum, maximum, alsó és felső kvartilis, terjedelem, IQR! Mi ezeknek a mutatószámoknak a jelentése a mintára nézve?

e. Adjunk meg a 'repwt' változóra három olyan intervallumot, hogy mindegyikbe a mintaelemek harmada essen!

4.4. 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 és szórására! Nevezzük meg pontosan, hogy mely statisztikai mutatószámokat alkalmaztuk!

b. Hány megfigyelés van az 'age' változóra, és mennyi a hiányzó adatok száma?

c. Mennyire pontos a várható értékre adott becslés?

d. Határozzuk meg a 'age' változó esetében a következő mutatószámok értékét: minimum, maximum, medián, alsó és felső kvartilis, terjedelem, IQR! Mi ezeknek a mutatószámoknak a jelentése a mintára nézve?

e. Adjunk meg az 'age' változóra öt olyan intervallumot, hogy mindegyikbe a mintaelemek ötöde essen!

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 diszkrétek 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 ourlier 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 diszkrétek és melyek folytonosak?

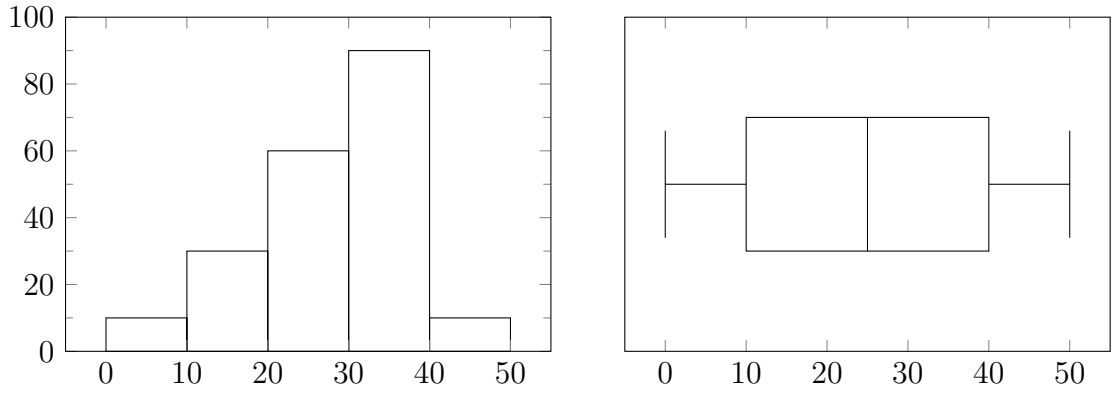
b. Ábrázoljuk a 'wc' változót oszlopdiagramon és kördiagramon! Az alanyok közül hányan rendelkeznek, és hányan nem rendelkeznek diplomával?

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 ourlier 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 móduszt 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?

6. Paraméterbecslési módszerek

6.1. A Poisson-eloszlás egy diszkrét eloszláscsalád, amely a $\lambda > 0$ (lambda) paraméterrel van indexezve. A várható érték és a valószínűségeloszlás:

$$\mathrm{E}(\xi) = \lambda, \qquad \mathrm{P}(\xi = x) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad R_\xi = \{0, 1, 2, \dots\}$$

Rendelkezésre áll egy mintarealizáció a ξ változóra: 1, 0, 4, 3, 0.

a. Végezzünk maximum likelihood becslést a likelihood függvény ábrázolásával. A szükséges R függvények: `factorial(x)` $= x!$, `exp(x)` $= e^x$.

b. Adjunk formulákat a momentum módszer és a maximum likelihood módszer révén kapott becslésekre egy általános $x_1, \dots, x_n$ mintarealizáció alapján.

6.2. Az exponenciális eloszlás egy folytonos eloszláscsalád, amely a $\lambda > 0$ paraméterrel van indexezve. A várható érték és a sűrűségfüggvény:

$$\mathrm{E}(\xi) = \frac{1}{\lambda}, \qquad f_\xi(x) = \lambda e^{-\lambda x}, \qquad R_\xi = (0, \infty)$$

Egy statisztikai minta a ξ változóra: 0.07, 0.21, 0.22, 0.35.

a. Végezzünk maximum likelihood becslést a likelihood függvény ábrázolásával. A szükséges R függvény: `exp(x)` $= e^x$.

b. Adjunk formulákat a momentum módszer és a maximum likelihood módszer révén kapott becslésekre egy általános $x_1, \ldots, x_n$ mintarealizáció alapján.

6.3. A Pareto-eloszlás egy folytonos eloszláscsalád, amely az $\alpha > 0$ paraméterrel van indexezve. A várható érték és a sűrűségfüggvény:

$$\mathrm{E}(\xi) = \frac{\alpha}{\alpha - 1}, \qquad f_\xi(x) = \frac{\alpha}{x^{\alpha+1}}, \qquad R_\xi = (1, \infty)$$

Egy statisztikai minta az ξ változóra: 1.39, 1.02, 1.23, 1.03, 1.28.

a. Végezzünk maximum likelihood becslést a likelihood függvény ábrázolásával.

b. Adjunk formulákat a momentum módszer és a maximum likelihood módszer révén kapott becslésekre egy általános $x_1, \ldots, x_n$ mintarealizáció alapján.

7. Konfidencia intervallum a várható értékre, t-próba

7.1. Régészek radiokarbonos kormeghatározással szeretnék meghatározni egy lelőhely korát. Sajnos a radiokarbonos módszer egy adott ásatáson nem pontosan ugyanazt a kort adja minden lelet esetében. Az egyes leletek radiokarbonos kora egy normális eloszlású ξ 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. A **4.1.** feladatban már kiszámoltuk, hogy a mintaátlag 1230, a korrigált empirikus szórás 51.9 és a standard hiba 15.17.

a. Adjunk meg egy 95% megbízhatósági szintű konfidencia intervallumot a lelőhely igazi korára!

b. A t-próba alkalmazásával teszteljük le 5%-os szignifikancia szinten azt a nullhipotézist, hogy a lelőhely igazi kora 1200 év!

c. Mennyi az elsőfajú illetve a másodfajú hiba nagysága ebben a feladatban?

7.2. Bejelentés érkezik a fogyasztóvédelemhez, hogy az egyik tejgyár 1 literes kiszerelésű 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 ξ 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. A **4.2.** feladatban már láttuk, hogy a mintaátlag 990, a korrigált empirikus szórás 33.91, a standard hiba 4.83.

a. Adjunk meg egy 90% megbízhatóságú konfidencia intervallumot a ξ változó várható értékére!

b. 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!

8. Az egymintás és a páros t-próba

8.1. 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 pácienst, 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!

8.2. Az ‘iris.xlsx’ állomány három Kanadában honos írisz (nőszirom) fajról tartalmaz adatokat, fajonként 50 növényről. A változók:

faj: faj megnevezése
fajkod: lásd faj
cseszehossz: szészelevél hossza (cm)
cseszeszel: csészelevél szélessége (cm)
sziromhossz: sziromlevél hossza (cm)
sziromszel: sziromlevél szélessége (cm)

a. Ábrázoljuk a ‘sziromszel’ 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 ‘sziromszel’ változó hisztogramját erre a fajra! Mit állíthatunk a ‘sziromszel’ változó eloszlásáról: közel normális vagy nagy mértékben különbözik a normálistól?

c. Adjunk becslést a ‘virginica’ fajhoz tartozó növényeknél a ‘sziromszel’ 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 sziromlevél átlagos szélessége azonos a csészelevél átlagos szélességével! Adjunk meg egy 90% megbízhatóságú konfidenciai intervallumot arra, hogy a sziromlevél átlagosan mennyivel szélesebb, mint a csészelevél!

e. Végezzük el az előző pontok elemzését a másik két faj egyedeire is!

9. Az egyszempontos ANOVA és a Levene-teszt

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

9.2. Olvassuk be az 'iris.xlsx' fájl tartalmát! Az adatsor leírása megtalálható a **8.2.** feladatban!

a. Adjunk becslést a 'sziromszel' változó várható értékére és szórására fajonkénti bontásban! Ábrázoljuk a változó boxplotját is, szintén fajonkénti bontásban!

b. Ábrázoljuk a 'sziromszel' változó hisztogramját, és adjuk meg a ferdeséget fajonkénti bontásban. Mit állíthatunk a 'sziromszel' változóról normalitás szempontjából?

c. Teszteljük le azt a nullhipotézist, hogy a 'sziromszel' 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% megbízhatóságú konfidencia intervallumot a várható értékek közötti különbségekre!

e. Ismételjük meg a fenti elemzést a 'cseszeszel' változóra!

10. Lineáris és nemlineáris regresszió

10.1. 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 (mérföld/gallon)
SP: végsebesség (mérföld/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: $\text{MPG} \approx a\text{HP} + b$

Reciprokos regresszió: $\text{MPG} \approx a/\text{HP} + b$

Exponenciális regresszió: $\text{MPG} \approx \exp(a\text{HP} + b)$

Melyik módszer biztosítja a legjobb becslést az 'MPG' változóra?

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.

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:

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

Melyik formula biztosítja a jobb becslést az 'SATV' változóra?

11. Korrelációs együtthatók és függetlenségvizsgálat

11.1. Olvassuk be a 'UScars.txt' fájlban található statisztika adatsort, a leírásért lásd a **10.1.** feladatot!

a. 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!

b. Végezzük el az előző pont elemzését az 'SP' és 'VOL' változókon is!

c. Végezzük el az előző pont elemzését az 'MPG' és 'HP' változókon is!

11.2. 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. 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!

b. Végezzük el az előző pont elemzését a ‘pop’ és ‘dollars’ változókon is!

c. Végezzük el az előző pont elemzését az ‘SATV’ és ‘percent’ változókon is!

Megoldások

1.1. $\xi =$ a kiválasztott játékos testmagassága

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

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

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

módusz $= 3$, jelentése: a legnagyobb arányban előforduló érték

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

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

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

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

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

módusz $= 0$, jelentése: a legnagyobb arányban előforduló érték

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

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

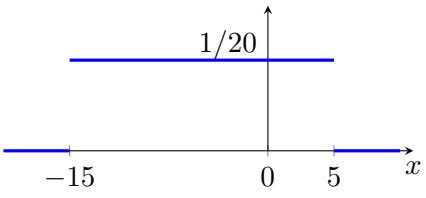
1.4. $R_\xi = \{1000, 2000, 3000, 5000\}$

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

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

$\mathrm{P}(\xi \leq 2000) = 0.8$

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



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

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

c.

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

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

$\mathrm{P}(\xi \ge 0) = \mathrm{P}(0 \le \xi \le +5) = F_\xi(+5) - F_\xi(0) = 0.25$

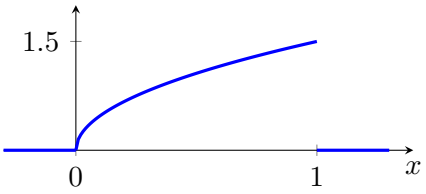
d. $\mathrm{E}(\xi) = -5$, jelentése: a napi középhőmérséklet átlagos értéke januárban

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

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

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

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



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

$\mathrm{P}(\xi \le 0.8) = \int_{0}^{0.8} f_\xi(x)\,dx = 0.72$, a fák 72 százaléka legfeljebb 0.8 átmérőjű

c.

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

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

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

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

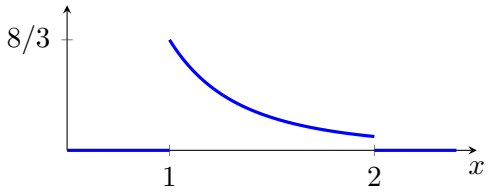
$\mathrm{D}(\xi) = 0.26$, jelentése: a törzs átmérője átlagosan ennyivel tér el a várható értéktől

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

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. ξ = egy véletlenszerűen kiválasztott egyed testhossza

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



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

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

c.

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

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

$\mathrm{P}(\xi \ge 1.8) = \mathrm{P}(1.8 \le \xi \le 2) = F_\xi(2) - F_\xi(1.8) = 0.08$

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

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

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

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. **a.** 66% **b.** 2% **c.** $[70.6, 129.4]$

3.3. **a.** 41% **b.** 9% **c.** $[5.77, 8.23]$

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

4.1.a. $n = 5$, $\overline{\xi} = 1230$, $\mathrm{D}_n(\xi) = 30.33$, $\mathrm{D}_n^*(\xi) = 33.91$

A lelőhely igazi kora $= \mathrm{E}(\xi) \approx \overline{\xi} = 1230$

$\mathrm{D}(\xi) \approx \mathrm{D}_n^*(\xi) = 51.9$
A kis mintaméret miatt a korrigált empirikus szórás pontosabb becslés.

b. SE $= 15.17$ Jelentése: a várható értékre adott becslés átlagos hibája.

c. Empirikus medián = középső mintaelem = 1230

Terjedelem = maximum − minimum = 90

4.2.a. $n = 6$, $\overline{\xi} = 990$, $\mathrm{Var}_n(\xi) = 116.67$, $\mathrm{D}_n(\xi) = 10.8$, $\mathrm{Var}_n^*(\xi) = 140$ $\mathrm{D}_n^*(\xi) = 33.91$
$\mathrm{E}(\xi) \approx \overline{\xi} = 1230$, $\mathrm{D}(\xi) \approx \mathrm{D}_n^*(\xi) = 33.91$

b. SE = 4.83 Jelentése: a várható értékre adott becslés átlagos hibája.

c. Empirikus medián = két középső átlaga = 990
Terjedelem = maximum − minimum = 30

4.3.a. $\mathrm{E}(\mathrm{repwt}) \approx \overline{\mathrm{repwt}} = 65.62$ (empirikus várható érték, mintaátlag)
$\mathrm{D}(\mathrm{repwt}) \approx \mathrm{D}_n^*(\mathrm{repwt}) = 13.78$ (korrigált empirikus szórás)

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

c. SE = 1.02 Jelentése: az $\mathrm{E}(\mathrm{repwt}) \approx \overline{\mathrm{repwt}}$ becslés várható hibája.

d. Minimum = legkisebb mintaelem = 41
$\hat{q}_{25\%}$ = alsó negyedelőpont = 55
$\hat{q}_{50\%}$ = minta felezéspontja (két középső átlaga) = 63
$\hat{q}_{75\%}$ = felső negyedelőpont = 73.5
Maximum = legnagyobb mintaelem = 124
Terjedelem = minimum − maximum = 83
Jelentése: ilyen hosszúságú intervallumon helyezkedik el a teljes minta.
$\mathrm{IQR} = \hat{q}_{75\%} - \hat{q}_{25\%} = 18.5$
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: $[41, 57]$, $[57, 69.2]$, $[69.2, 124]$

4.4.a. $\mathrm{E}(\mathrm{age}) \approx \overline{\mathrm{age}} = 42.54$ (empirikus várható érték, mintaátlag)
$\mathrm{D}(\mathrm{age}) \approx \mathrm{D}_n^*(\mathrm{age}) = 8.07$ (korrigált empirikus szórás)

b. $n = 753$, nincs hiányzó adat

c. SE = 0.29 Jelentése: az $\mathrm{E}(\mathrm{age}) \approx \overline{\mathrm{age}}$ becslés átlagos hibája.

d. Minimum = legkisebb mintaelem = 30
Maximum = legnagyobb mintaelem = 60
$\hat{q}_{50\%} = 43$, jelentése: középső érték a mintában
$\hat{q}_{25\%} = 36$, jelentése: a minta alsó negyedelőpontja
$\hat{q}_{50\%} = 43$, jelentése: középső mintaelem
$\hat{q}_{75\%} = 49$, jelentése: a minta felső negyedelőpontja

Terjedelem = minimum − maximum = 30
Jelentése: ilyen hosszúságú intervallumon helyezkedik el a teljes minta.

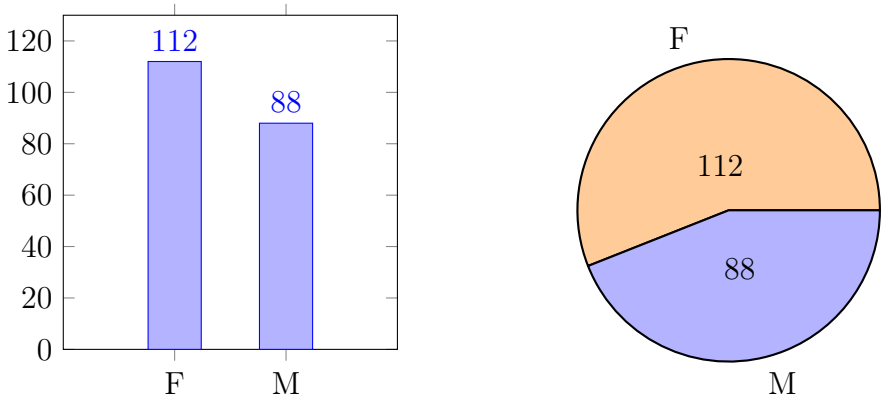
IQR = $\hat{q}_{75\%} - \hat{q}_{25\%} = 13$
Jelentése: ilyen hosszúságú intervallumon helyezkedik el a minta középső 50%-a.

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

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

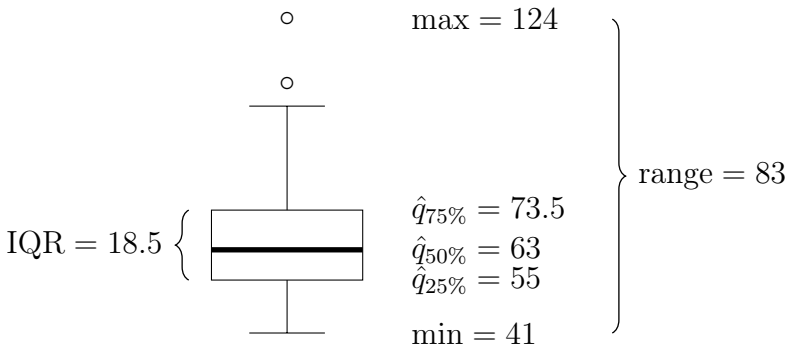
5.1.a. Diszkrét változók: sex.
Folytonos változók: weight, height, repwt, repht.

b. A mintában 112 nő és 88 férfi szerepel.



c. skewness = 1.04, a hisztogram jobbra ferde.

d. A mintában 2 outlier érték található.

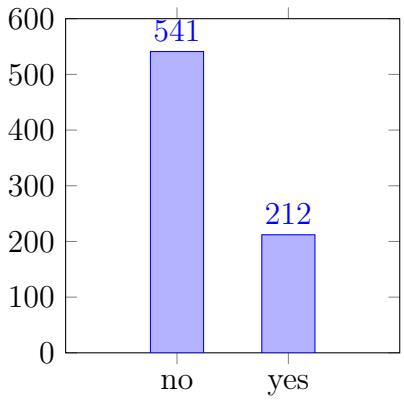


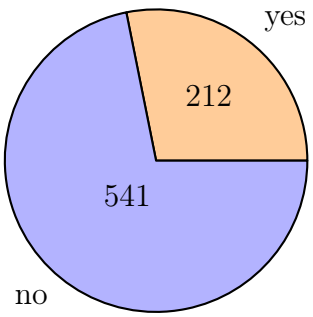
e. A változó enyhén jobbra ferde. A minta valószínűleg nem normális eloszlásból származik, de az eloszlás nem különbözik nagy mértékben a normálistól.

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.

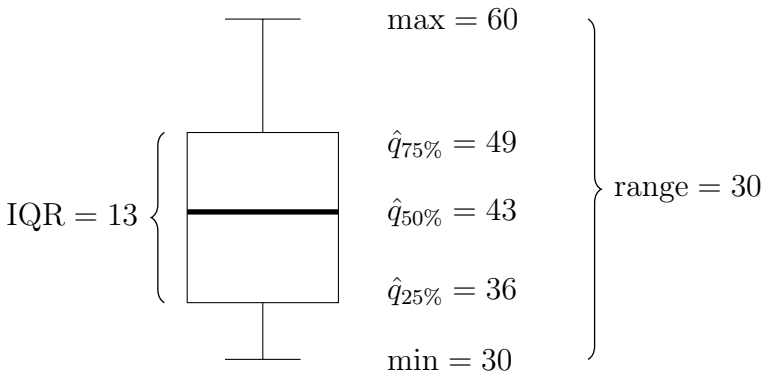
b. 212 alany rendelkezik, 541 nem rendelkezik diplomával.





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.

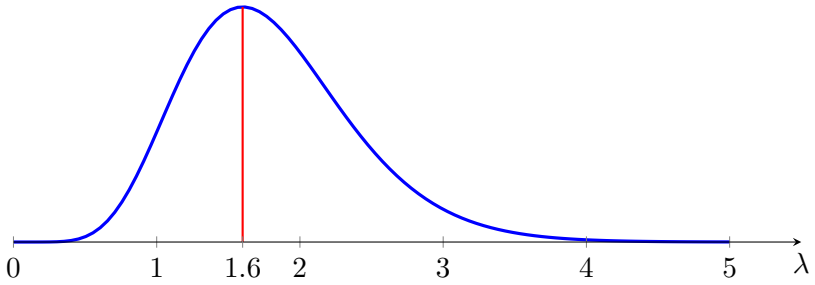


e. A boxplot és a skewness alapján a változó közel normális eloszlásúnak tűnik. Ezzel szemben a hisztogramra egyáltalán nem illeszkedik a normális eloszlás sűrűségfüggvénye. Emiatt a végső konklúzió az, hogy változó eloszlása nagyban különbözik a normálistól.

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.

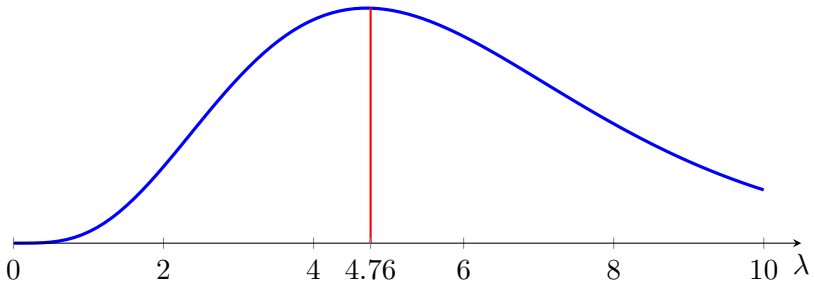
b. Minimum: 0, alsó kvartilis: 10, medián: 25, felső kvartilis: 40, maximum: 50. A terjedelem 50, az IQR 30. Nincs outlier érték.

6.1.a. Maximum likelihood becslés: $\lambda \approx 1.6$



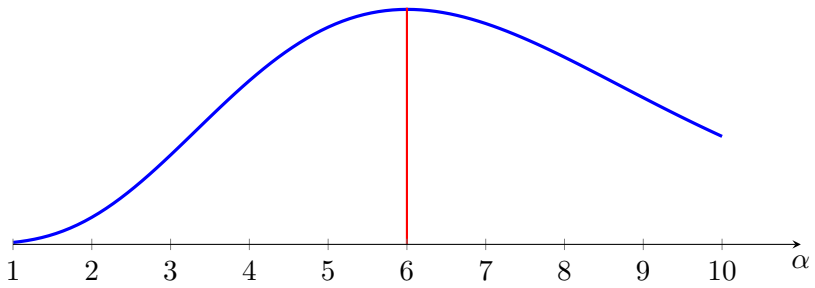
b. Mindkét becslés esetében: $\lambda \approx \overline{\xi}$.

6.2.a. Maximum likelihood becslés: $\lambda \approx 4.76$



b. Mindkét becslés esetében: $\lambda \approx 1/\overline{\xi}$

6.3.a. Maximum likelihood becslés: $\alpha \approx 6$



b. Momentum módszer: $\alpha \approx \overline{\xi}/(\overline{\xi} - 1)$

Maximum likelihood becslés: $\alpha \approx n/\ln(x_1 \dots x_n)$

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

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

b. Nullhipotézis: $\mathrm{H}_0 : \mathrm{E}(\xi) = 1200$.

Hipotetikus várható érték: $\mu_0 = 1200$.

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

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

c. Elsőfajú hiba: 5%.

Másodfajú hiba: nem ismerjük a nagyságát.

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

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

b. Nullhipotézis: $\mathrm{H}_0 : \mathrm{E}(\xi) = 1000$.

Hipotetikus várható érték: $\mu_0 = 1000$.

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

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

8.1.a. $\mathrm{E(SYS1)} \approx \overline{\mathrm{SYS1}} = 160.2$, $\mathrm{D(SYS1)} \approx \mathrm{D}_n^*(\mathrm{SYS1}) = 5.7$.

A becslés várható hibája: $\mathrm{SE} = 0.46$.

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

c. $\mathrm{H}_0 : \mathrm{E(SYS1)} = 160$, egymintás t-próba: p-érték=0.626, elfogadjuk.

$\mathrm{H}_0 : \mathrm{E(SYS1)} = 165$, 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 értéknek.

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

A hisztogramok alapján mindkét változó közel normális eloszlású.

e. $\mathrm{H}_0 : \mathrm{E(SYS1)} = \mathrm{E(SYS2)}$, páros t-próba, p-érték=0.000, a nullhipotézist elvetjük.

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

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

$\mathrm{H}_0 : \mathrm{E(SYS1)} = \mathrm{E(SYS2)}$, páros t-próba, p-érték=0.11, a nullhipotézist elfogadjuk.

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

8.2.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: $\mathrm{E(sziromszel)} \approx 2.03$, $\mathrm{D(sziromszel)} \approx 0.27$.

$\mathrm{H}_0 : \mathrm{E(sziromszel)} = 2$, egymintás t-próba: p-érték=0.51, elfogadjuk.

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

d. A ‘virginica’ növényeknél: E(cseszeszel) ≈ 2.97

H_0 : E(sziromszel) = E(cseszeszel), páros t-próba, p-érték=0, elvetjük.

Konfidencia intervallum az E(sziromszel)−E(cseszeszel) különbségre: $[-1.02, -0.88]$.

e. Azonos módszerekkel, mint az előző pontokban.

9.1.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 hisztogrammok 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(SYS1 | kiserleti1) = D(SYS1 | kiserleti2) = D(SYS1 | 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(SYS1 | kiserleti1) = E(SYS1 | kiserleti2) = E(SYS1 | 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(SYS2 | kiserleti1) = D(SYS2 | kiserleti2) = D(SYS2 | kontroll)

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

H_0 : E(SYS2 | kiserleti1) = E(SYS2 | kiserleti2) = E(SYS2 | 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	$[-13.75, -9.09]$

9.2.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(sziromszel | setosa) = D(sziromszel | versicolor) = D(sziromszel | virginica)

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

d. H_0 : E(sziromszel | setosa) = E(sziromszel | versicolor) = E(sziromszel | 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 'cseszeszel' változó esetében a Levene-teszt elfogadja a csoportonkénti szórások azonosságát. Emiatt alkalmazhatjuk az ANOVA tesztet, ami elveti a várható értékek egyenlőségét.

10.1.a. $\text{SP} \approx 0.24 \cdot \text{HP} + 84.45$

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

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

b. $\text{SP} \approx -0.03 \cdot \text{VOL} + 115.11$

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

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

Reciprokos regresszió:

Új változó: $\text{repHP} = 1/\text{HP}$

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

Exponenciális regresszió:

Új változó: $\text{logMPG} = \log(\text{MPG})$

$\text{logMPG} \approx -0.0046 \cdot \text{HP} + 4.01, \quad R^2 = 0.73$

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

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

10.2.a. $\text{SATV} \approx 0.86 \cdot \text{SATM} + 21.53$

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

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

b. $\text{dollars} \approx 0.00004 \cdot \text{pop} + 4.998$

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

c. Első formula:

Új változó: $\text{repPercent} = 1/\text{percent}$

$\text{SATV} \approx 428.1 \cdot \text{repPercent} + 421, \quad R^2 = 0.69$

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

Második formula:

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

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

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

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

11.1.a. H_0 : 'SP' és 'HP' független változók

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

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

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

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

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

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

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.

11.2.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.