12. Gyakorlat

Statisztika

Feladattípusok:

-Alapstatisztikák: átlag, szórás, medián, módusz, terjedelem, kvartilisek/kvantilisek
--Boxplot ábra
--Hisztogram, binwidth, bincount
--FrequencyTable, FrequencyPlot
--Count, Mode, Median, Mean, StandardDeviation, Range, Variance
--Empirikus eloszlásfüggvény LineChart-tal

-Diszkrét illeszkedésvizsgálat (khí-négyzet), regresszióra is
--infolevel[Statistics] := 1
--ChiSquareSuitableModelTest(data, distribution, bins, confidence level);
--kritikus érték, próbastatisztika számítása, sûrûségfüggvény alatti terület besatírozása
--illeszkedésvizsgálat gyakorisági táblázattal, elméleti gyakoriságok számolása
--ChiSquareGoodnessOfFitTest(empirical frequencies, theoretical frequencies , confidence level);

-konfidencia-intervallum (u-próba, t-próba)
--konfidencia-intervallum ábrázolása
--mintaelemszám adott konfidencia-intervallum hosszhoz, ellenôrzés adatok generálásával
--u-próba
--OneSampleZTest(data, mu, sigma, confidence level);
--t-próba számolása
--OneSampleTTest(data, mu, confidence level);
--Konfidencia-intervallum normál eloszlás szórására
--OneSampleChiSquareTest(data, sigma, confidence level);

-Lineáris regresszió, illesztés legkisebb négyzetek módszerével, ábrázolása
--ScatterPlot-os ábrázolás
--Korrelációs együttható számolása, R^2 számítása
--korrelációs együtthatók számítása
--regressziós függvény értékei
--reziduálisok
--reziduálisok ábrázolása hisztogramon, közelítés normál eloszlással
--LinearFit([1, x], X, Y, x);
--QuantilePlot-os ábrázolás
--Regressziós együtthatók konfidencia-intervalluma t-eloszlással
--LinearFit([x, 1], X, Y, x, output=solutionmodule);
--regressziós egyenes konfidencia-intervalluma

Tanácsok a ZH-hoz:
-A munkalapon név, Neptun, Neptun szerinti gyakorlatvezetô szerepeljen, feladatok szövege, sorszáma maradjon meg
-Feladatok elején restart;
-Feladatban megadott adatokat kigyûjteni
-Minden lépést 1-2 mondatban indokolni, leírni, hogy mi történik (kódot megmagyarázni)
-Számolás mehet a Maple beépített eszközeivel

-Elméletet, képleteket megtanulni
-Minden feladatrész végén 1-2 mondatos szöveges válasz
-Gondoljuk meg, hogy ésszerû eredményt kaptunk-e! Ha nem, akkor ellenôrizzünk!
-Ne csússzunk ki a leadási határidôbôl!

A 2. ZH-n nem kizárólag statisztika feladatok lesznek!!! A valószínûségi változóktól kezdôdôen minden témakörbôl lehetnek feladatok: valószínûségi változók, várható érték, szórás, nevezetes diszkrét és folytonos eloszlások, nagy számok törvénye, statisztika.

Kidolgozott feladatok

1. Feladat (Alapstatisztikák)

Egy egyetemi felvételi teszten az alábbi pontszámok születtek:
41, 45, 49, 59, 24, 81, 21, 11, 15, 23, 30, 41, 46, 17, 30, 76, 17, 52, 47, 47.
a) Hány adatpont van? Számítsuk ki a minta terjedelmét, átlagát, szórását, mediánját, móduszát, valamint 1. és 3. kvartiliseit!
b) Ábrázoljuk az adatok váltakozását az átlaggal és a kvartilisekkel együtt, valamint rajzoljuk föl a box-diagramot az átlag feltüntetésével!
c) Rendezzük a mintát és ábrázoljuk a tapasztalati (empirikus) eloszlásfüggvényt!
d) Osszuk fel a teljes terjedelmet 5 egyenlô részre! Adjuk meg táblázat formájában az egyes kategóriákhoz tartozó tapasztalati gyakoriságokat és relatív gyakoriságokat valamint ábrázoljuk azokat pontdiagramon!
e) Ábrázoljuk az elôzô feladatrészben kapott relatív gyakoriságokat hisztogramon! Illesszük a hisztogramra egy azonos várható értékû és szórású normál eloszlás sûrûségfüggvényét!

Megoldás

```
> restart;
```

a) Hány adatpont van? Számítsuk ki a minta terjedelmét, átlagát, szórását, mediánját, móduszát, valamint 1. és 3. kvartiliseit!
Vegyük fel egy listába az adatsort!

```
> PSZ := [41, 45, 49, 59, 24, 81, 21, 11, 15, 23, 30, 41,
  46, 17, 30, 76, 17, 52, 47, 47];
```

$$PSZ := [41, 45, 49, 59, 24, 81, 21, 11, 15, 23, 30, 41, 46, 17, 30, 76, 17, 52, 47, 47] \quad \textbf{(1.1.1.1.1)}$$

Ezután száoljuk ki a kért statisztikai mutatókat a Maple beépített eszközeivel! (Az elsô 4 mutatónál megjegyzésben meg van adva, hogyan lehetne meghatározni a Statistics csomag használata nélkül.)

```
> with(Statistics):
> n := Count(PSZ); # = nops(PSZ) = numelems(PSZ)
  terjedelem := trunc(Range(PSZ)); # = max(PSZ) - min(PSZ)
  m := Mean(PSZ); # = sum(PSZ[i], i = 1..n)/n
  sigma := StandardDeviation(PSZ); # = sqrt(sum((PSZ[i] -
  m)^2, i = 1..n)/(n - 1))
  M := Median(PSZ); # a medián a 2. kvartilis
  Mode(PSZ);
  Q1 := Quartile(PSZ, 1);
  Q3 := Quartile(PSZ, 3);
```

$$n := 20$$

$$terjedelem := 70$$
$$m := 38.6000000000000$$
$$\sigma := 19.6211486331512$$
$$M := 41.$$
$$17.$$
$$Q1 := 21.8333333333333$$
$$Q3 := 48.1666666666667 \qquad \textbf{(1.1.1.1.2)}$$

b) Ábrázoljuk az adatok váltakozását az átlaggal és a kvartilisekkel együtt, valamint rajzoljuk föl a box-diagramot az átlag feltüntetésével!

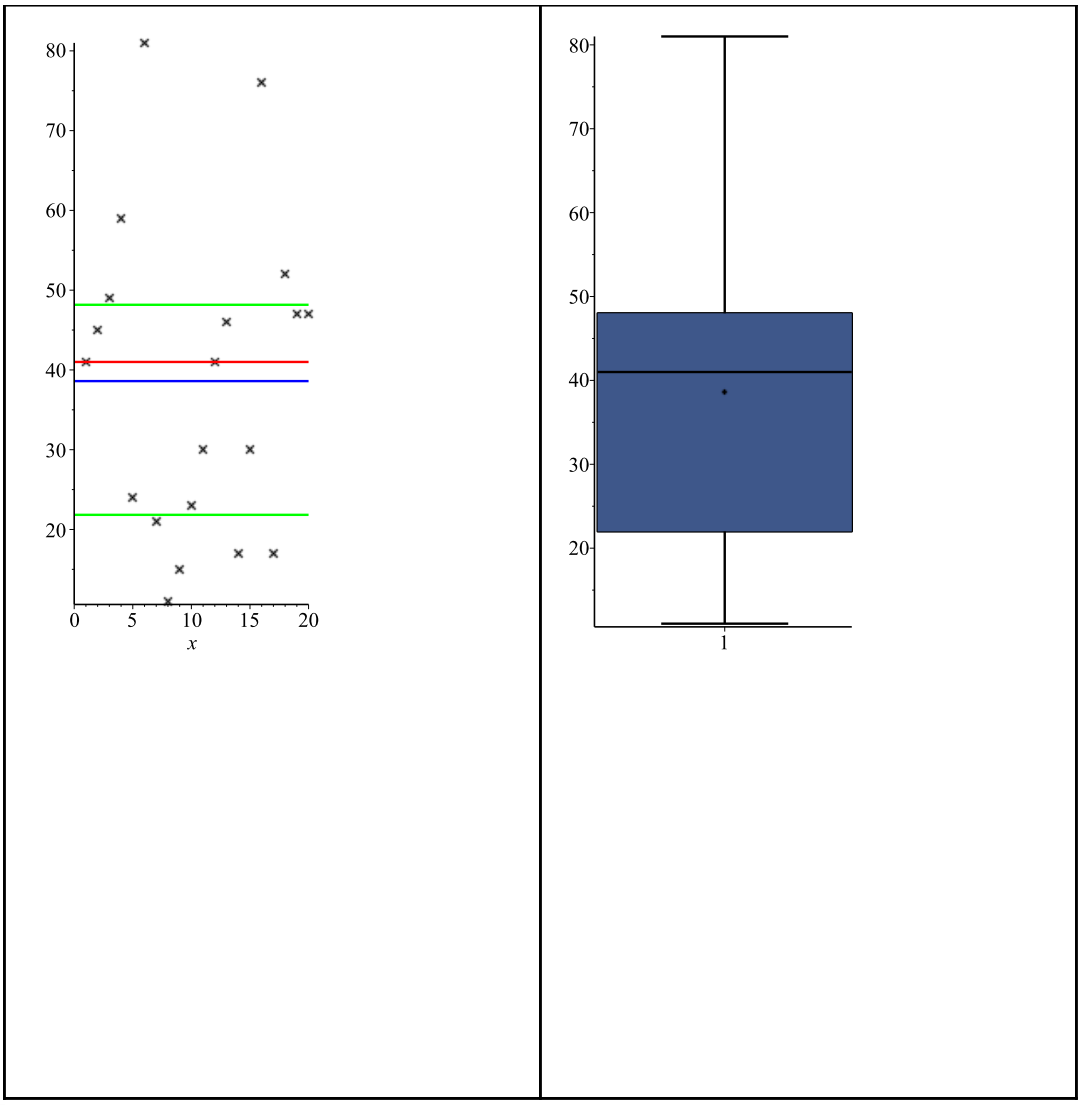
Az adatpontok koordinátái sorban:

```
> adatpontok := [seq([i, PSZ[i]], i = 1..n)];
```

$$adatpontok := [[1, 41], [2, 45], [3, 49], [4, 59], [5, 24], [6, 81], [7, 21], [8, 11], [9, 15], [10, 23], [11, 30], [12, 41], [13, 46], [14, 17], [15, 30], [16, 76], [17, 17], [18, 52], [19, 47], [20, 47]] \qquad \textbf{(1.1.1.1.3)}$$

Készítsük el a két ábrát és jelenítsük meg egymás mellett egy mátrixban:

```
> P1 := plot([adatpontok, m, Q1, M, Q3], x = 0..20, style
  = [point, line, line, line, line], color = [black, blue,
  green, red, green], symbol = diagonalcross, symbolsize =
  12):
  P2 := BoxPlot(PSZ):
  plots[display](Matrix(1, 2, [P1, P2]));
```



A box-diagram az adatsokaság vizualizálására szolgáló hasznos eszköz. Az alsó és felsô kvartilisek határolják a "dobozt", melyet középtájon a medián oszt ketté. Lefelé és felfelé a "doboz" határától az interkvartilis terjedelem ($Q_3 - Q_1$) 3/2-szerese van felmérve (vagy ameddig az adat terjedelme ér), az ezen kívül esô adatpontok kiugrónak tekinthetôk.

c) Rendezzük a mintát és ábrázoljuk a tapasztalati (empirikus) eloszlásfüggvényt!

Listát rendezni a *sort* eljárással lehet:

```
> PSZ_rend := sort(PSZ); # rendezett lista
```

PSZ_rend := [11, 15, 17, 17, 21, 23, 24, 30, 30, 41, 41, 45, 46, 47, 47, 49, 52, 59, 76, 81] **(1.1.1.1.4)**

A terjedelmet megkapnánk az utolsó és elsô szám különbségeként.

```
> # terjedelem := PSZ_rend[n] - PSZ_rend[1];
```

AFrequencyTable eljárással egyszerûen megkaphatjuk az értékek gyakoriságainak, relatív gyakoriságainak, halmozott gyakoriságainak és halmozott relatív gyakoriságainak listáját. Az utóbbiból gyárthatjuk le az empirikus eloszlásfüggvényt. Persze ehhez olyan felbontást kell választani, amely mellett az egyes adatpontok jól elkülönülnek. Ezt a *bins* opció terjedelemre való beállításával érhetjük el.

```
> gyak := FrequencyTable(PSZ, bins = terjedelem);
```

$$gyak := \begin{bmatrix} 1..70 \text{ x } 1..5 \text{ Array} \\ \text{Data Type: anything} \\ \text{Storage: rectangular} \\ \text{Order: Fortran\_order} \end{bmatrix} \quad \textbf{(1.1.1.1.5)}$$

Készítsünk átlátható táblázatot ebbôl a tömbbôl! Válasszuk ki a megfelelô oszlopokat lássuk el fejléccel.

```
fejlec:=[`Intervallumok`,`Gyakoriságok`,`Halmozott relatív
gyakoriságok`]:
> gyakorisag_ertekek:= matrix(terjedelem + 1, 3, [fejlec,
  seq([gyak[k,1], gyak[k,2], gyak[k,5]], k = 1..
  terjedelem)]);
```

gyakorisag_ertekek := **(1.1.1.1.6)**

Intervallumok	*Gyakoriságok*	*Halmozott relatív gyakoriságok*
11. ..12.	1.	5.000000000
12. ..13.	0.	5.000000000
13. ..14.	0.	5.000000000
14. ..15.	0.	5.000000000
15. ..16.	1.	10.00000000
16. ..17.	0.	10.00000000
17. ..18.	2.	20.00000000
18. ..19.	0.	20.00000000
19. ..20.	0.	20.00000000
20. ..21.	0.	20.00000000
21. ..22.	1.	25.00000000
22. ..23.	0.	25.00000000
23. ..24.	1.	30.00000000
24. ..25.	1.	35.00000000
25. ..26.	0.	35.00000000
26. ..27.	0.	35.00000000
27. ..28.	0.	35.00000000
28. ..29.	0.	35.00000000
29. ..30.	0.	35.00000000
30. ..31.	2.	45.00000000
31. ..32.	0.	45.00000000
32. ..33.	0.	45.00000000
33. ..34.	0.	45.00000000
34. ..35.	0.	45.00000000
35. ..36.	0.	45.00000000
36. ..37.	0.	45.00000000
37. ..38.	0.	45.00000000
38. ..39.	0.	45.00000000
39. ..40.	0.	45.00000000
40. ..41.	0.	45.00000000
41. ..42.	2.	55.00000000
42. ..43.	0.	55.00000000
43. ..44.	0.	55.00000000

A halmozott relatív gyakoriságok (százalékban megadva) a *gyak* tömb 5. oszlopában vannak.

```
> halmozott_gyakorisagok := gyak[1..terjedelem, 5];
```

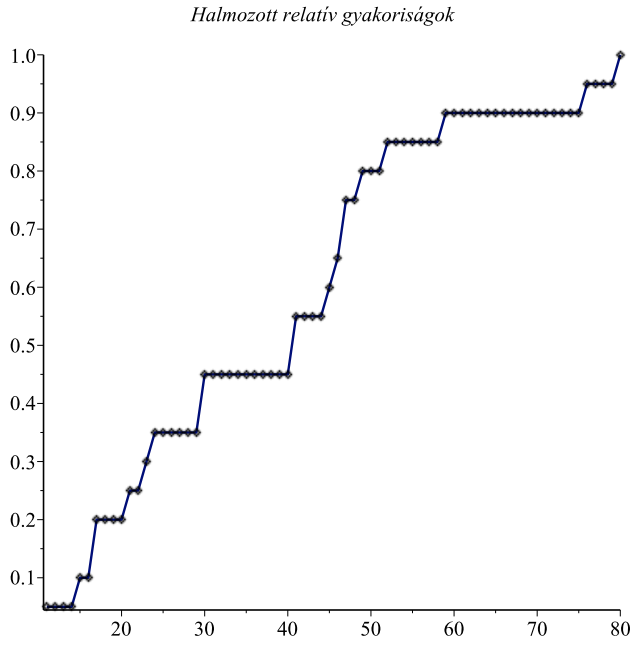
$$halmozott\_gyakorisagok := \begin{bmatrix} 1 \text{ .. } 70 \text{ Array} \\ \text{Data Type: anything} \\ \text{Storage: rectangular} \\ \text{Order: Fortran\_order} \end{bmatrix} \qquad \textbf{(1.1.1.1.7)}$$

Ezekbôl a LineChart eljárással kaphatunk az empirikus eloszlásfüggvényre hasonlító grafikont. Az ábrázoláshoz válasszuk az adatok értéktartományát ([A, B] intervallum)!

```
> A := PSZ_rend[1];
  B := PSZ_rend[n];
  LineChart(halmozott_gyakorisagok, xcoords=[seq(k, k = A.
  .B)], title=`Halmozott relatív gyakoriságok`, scale=
  relative);
```

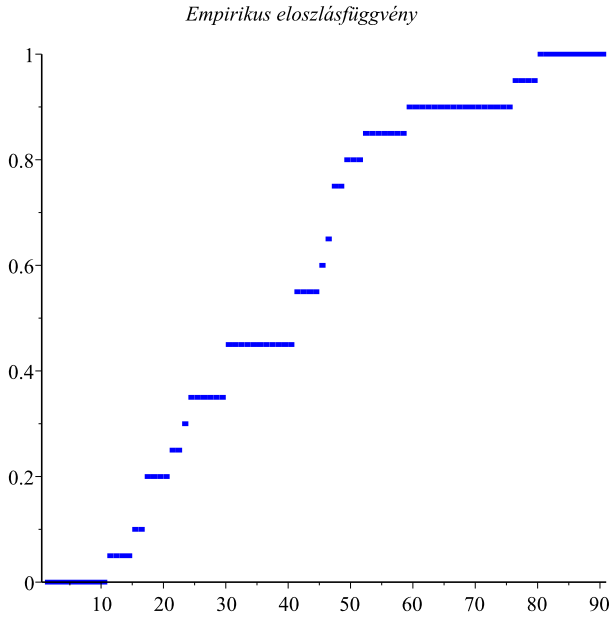
$$A := 11$$

$$B := 81$$



Az egyetlen (jobbára elhanyagolható) eltérés a tapasztalati eloszlásfüggvénytôl az, hogy itt a lépcsôk között tiszta ugrás helyett meredek ferde szakaszok vannak. Egy kicsit bonyolultabb ábrázoló eljárással ezen is segíthetünk:

```
> plot({[[A - 10, 0], [A, 0]], seq([[A + i - 1,
  halmozott_gyakorisagok[i]/100], [A + i,
  halmozott_gyakorisagok[i]/100]], i = 1..terjedelem), [
  [B, 1], [B+10, 1]]}, thickness = 3, color = blue, title=
  `Empirikus eloszlásfüggvény`);
```



d) Osszuk fel a teljes terjedelmet 5 egyenlô részre! Adjuk meg táblázat formájában az egyes kategóriákhoz tartozó tapasztalati gyakoriságokat és relatív gyakoriságokat valamint ábrázoljuk azokat pontdiagramon!

Az elôbbi *FrequencyTable*-hívás alkalmas módosításával megkapjuk a táblázatot. Mivel a terjedelmet 5 egyenlô részre kell osztani, ezért most 5 ládát használunk.

```
> gyak2 := FrequencyTable(PSZ, bins = 5)[1..5, 1..3];
```

$$gyak2 := \begin{bmatrix} 11..25. & 7. & 35.00000000 \\ 25..39. & 2. & 10.00000000 \\ 39..53. & 8. & 40.00000000 \\ 53..67. & 1. & 5.000000000 \\ 67..81. & 2. & 10.00000000 \end{bmatrix} \qquad \textbf{(1.1.1.1.8)}$$

Adjunk fejlécet a táblázatnak! (Figyeljük meg, hogy most a korábban látottól eltérôen sorvektorként adjuk meg a fejlécet! Ez egy alternatív megoldás, azonos végeredménnyel.)

```
> Matrix(6, 3, [[<`Intervallumok` | `Gyakoriságok` |
```

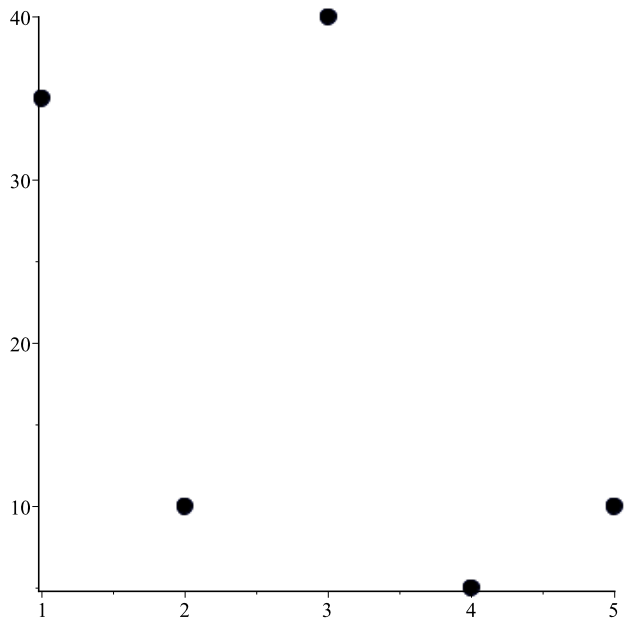
```
`Relatív gyakoriságok`>], [gyak2]]);
```

Intervallumok	*Gyakoriságok*	*Relatív gyakoriságok*
11..25.	7.	35.00000000
25..39.	2.	10.00000000
39..53.	8.	40.00000000
53..67.	1.	5.000000000
67..81.	2.	10.00000000

(1.1.1.1.9)

A pontdiagramot legegyszerûbben a LineChart eljárással készíthetjük el. A *style* opciót állítsuk *point*-ra. (Egyetlen szépséghibája a módszernek, hogy az *y*-tartományt nem lehet megadni, így 5-tôl indul a skála. Ez a különbség látszik az alábbi hisztogramon is.)

```
> LineChart(gyak2[1..5, 3], style = point, symbol =
  solidcircle, symbolsize = 20);
```

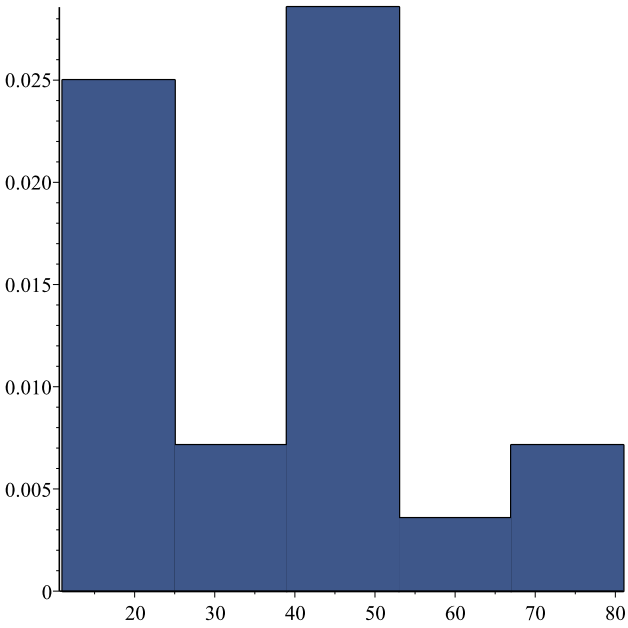


e) Ábrázoljuk az elôzô feladatrészben kapott relatív gyakoriságokat hisztogramon! Illesszük a hisztogramra egy azonos várható értékû és szórású normál eloszlás sûrûségfüggvényét!

Az elôbb elkészített gyakoriság adattal dolgozunk. Fontos, hogy mind az alaptartományt (*range*), mind az oszlopok szélességét (*binwidth*) megfelelôen állítsuk be!

```
> H := Histogram(PSZ, binwidth = terjedelem/5, range = A..
```

```
B): H;
```

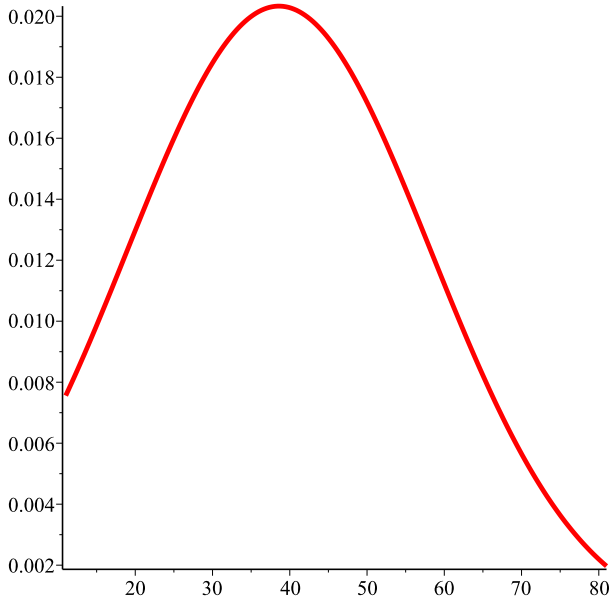


Az azonos várható értékû és szórású normál eloszlás sûrûségfüggvénye azonos tartományon:

```
> X := RandomVariable(Normal(m, sigma));
```

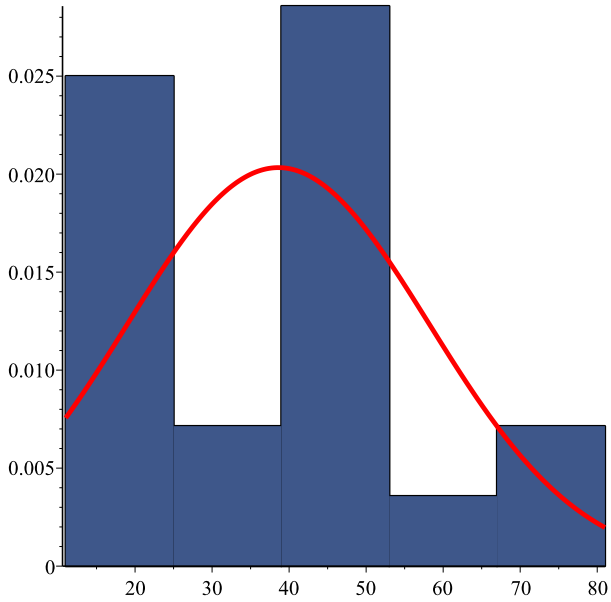
$$X := \_R \qquad \textbf{(1.1.1.1.10)}$$

```
> P3 := DensityPlot(X, range = A..B, color = red,
  thickness = 3): P3;
```



Illesszük ezt rá a hisztogramra!

```
> plots[display](H, P3);
```



Nem úgy tûnik, hogy a gyakoriságok normál eloszlást követnének, bár ezt 5 oszlopból nehéz elsô ránézésre megmondani. Többek között ilyen vitás kérdések eldöntésére szolgál a kövezkezô feladatban tárgyalt illeszkedésvizsgálat.

2. Feladat (Diszkrét illeszkedésvizsgálat)

Egy év minden napján feljegyezzük egy adott útkeresztezôdésben a piros lámpán szabálytalanul áthajtó gépjármûvek számát. A napi gyakoriságokat a kövezkezô táblázat mutatja:

Szabálytalanságok száma	**0**	**1**	**2**	**3**	**≥**
Gyakoriságok	**4**	**8**	**9**	**7**	**5**

a) Vizsgáljuk meg 1 %-os szignifikanciaszint mellett azt a nullhipotézist, hogy a piros lámpán egy nap alatt áthajtó gépjármûvek száma Poisson eloszlású $\lambda = 2$ paraméterrel!
b) Adjuk meg a kritikus értéket, a próbastatisztika értékét és a p-értéket! Magyarázzuk meg a döntést a p-érték alapján is!
c) Szemléltessük a kritikus tartományt és a döntést grafikonon a sûrûségfüggvény alatti terület beszínezésével!

Megoldás

```
> restart;
```

Korábbi feladatok megoldása során már sokszor végeztünk szimulációt a kapott elméleti eredmények empirikus igazolására. Ilyenkor kiindultunk egy ismert valószínûségi eloszlásból, melybôl mintát vettünk, majd megmértük bizonyos események relatív tapasztalati gyakoriságait.
A statisztikában gyakran találkozunk a fordított szituációval: adva van egy minta (mérési eredmények, kérdôív, stb.) és azt szeretnénk tudni, milyen *háttéreloszlás*ból (milyen eloszlású sokaságból) lett kisorsolva. Persze a minta alapján elkészíthetünk egy, azt tökéletesen leíró diszkrét eloszlást (melynek értékei a mintában elôforduló értékek, a hozzájuk tartozó valószínûségek pedig a mintában számolt relatív gyakoriságok), de ugyanakkor tisztában vagyunk azzal, hogy a mintát véletlenszerû események alakítják, más-más mintákra más-más értékeket/gyakoriságokat kapunk, így a konkrét mintánk nem írja le pontosan a háttérben meghúzódó eloszlást.
Ehelyett olyan valószínûségi szabályszerûséget keresünk, mely tetszôleges mintára igaz lehet és így az egész populációt jellemzi: feltételezéssel élünk a háttéreloszlás típusával kapcsolatban, mely elméleti eredményeken, méréseken, másik fél által vállalt garanciákon ill. a megszerzett tapasztalatainkon alapul. Ilyen például az, ha sok független, azonos eloszlású tényezô összegeként adódó értéket normál eloszlásúnak feltételezünk (lásd CHT), vagy valamely berendezés élettartamát exponenciális eloszlással írjuk le. A háttéreloszlás paraméterei néha adottak, néha pedig a minta alapján kell becsülnünk azokat (*pontbecslés*). Ehhez kapcsolódó statisztikai alpprobléma az ún. *illeszkedésvizsgálat*, melynek során döntést hozunk arról, hogy egy bizonyos *megbízhatósági szint* mellett az adott minta származhat-e a sejtett eloszlásból, azaz az eloszlás illeszkedik-e a mintára (ez az ún. *nullhipotézis* (H_0), melynek tagadása az *ellenhipotézis* (H_1)). A megbízhatósági szint (*konfidenciaszint*) azt jelenti, hogy mekkora valószínûséggel fogadjuk el a nullhipotézist, ha egyébként a háttéreloszlásra vonatkozó feltételezésünk igaz volt. A *szignifikanciaszint* a konfidenciaszintet 1-re kiegészítô valószínûség, melyet *elsôfajú hibá*nak is neveznek. Érdemes megjegyezni, hogy nagyobb konfidenciaszint megengedôbb/elfogadóbb tesztet eredményez.

a) Vizsgáljuk meg 1 %-os szignifikanciaszint mellett azt a nullhipotézist, hogy a piros lámpán egy nap alatt áthajtó gépjármûvek száma Poisson eloszlású $\lambda = 2$ paraméterrel!

Az adatok:

```
> alpha := 0.01; # szignifikanciaszint
  lambda := 2;
```

$$\alpha := 0.01$$

$$\lambda := 2 \qquad \textbf{(1.1.2.1.1)}$$

Vegyük fel egy-egy listában a feljegyzett szabálytalanság-számokat és a hozzájuk tartozó gyakoriságokat (hány napon történt pont annyi szabálytalanság):

```
> x := [0, 1, 2, 3, 4]; # értékek, az utolsó '>= 4'-nek
  értendô
  gyak := [49, 89, 97, 71, 59]; # mintagyakoriságok
```

$$x := [0, 1, 2, 3, 4]$$

$$gyak := [49, 89, 97, 71, 59] \qquad \textbf{(1.1.2.1.2)}$$

Az utolsó kategória egy gyûjtôkategória, mely a 'legalább 4 szabálatalanság'-hoz tartozik. A gyakoriságok összege éppen az év napjainak száma, mivel egy évig minden nap végeztünk megfigyelést.

```
> n := 5; # a kategóriák száma
```

```
N := sum(gyak[i], i = 1..n); # = az év napjainak száma,
ahogyan a feladat állítja
```

$$n := 5$$

$$N := 365 \qquad \textbf{(1.1.2.1.3)}$$

Az illeszkedésvizsgálat során a tapasztalati gyakoriságok összehasonlításra kerülnek az elméleti gyakoriságokkal. Utóbbiakat a valószínûségelméleti eszközökkel kiszámolt pontos valószínûségek mintamérettel vett szorzatai adják.

Hozzunk létre tehát egy $\lambda = 2$ paraméterû Poisson-eloszlású valószínûségi változót

```
> with(Statistics):
X := RandomVariable(Poisson(lambda));
```

$$X := \_R \qquad \textbf{(1.1.2.1.4)}$$

Számítsuk ki az elméleti gyakoriságokat! Ügyeljünk arra, hogy az utolsó kategóriába az összes "maradék" gyakoriság kerüljön, ne csak az $\{X = 4\}$ eseményhez tartozó!

```
> elm_gyak := evalf([seq(N*ProbabilityFunction(X, i), i =
0..n - 1)]): # elméleti gyakoriságok
elm_gyak[5] := N - sum(elm_gyak[i + 1], i = 0..n - 2): #
az utolsó gyakoriság legyen N*P(X>=4), vagyis az összes
maradék gyakoriság, a kategóriák választásával
összhangban
'elm_gyak' = elm_gyak;
```

$$elm\_gyak = [49.39737837, 98.79475674, 98.79475674, 65.86317116, 52.1499370] \qquad \textbf{(1.1.2.1.5)}$$

Ellenôrizhetjük, hogy az elméleti gyakoriságok összege is az év napjainak száma:

```
> sum(elm_gyak[i], i = 1..n);
```

$$365.0000000 \qquad \textbf{(1.1.2.1.6)}$$

Az illeszkedés megállapításához végezzünk χ^2-próbát! Ezt a Statistics csomag hosszú nevû *ChiSquareGoodnessOfFitTest* eljárása végzi. Paraméterei a tapasztalati és elméleti gyakoriságok listája, valamint a szignifikanciaszint. Állítsuk a Statistics csomag információs szintjét 1-re, hogy szép beszédes kimenetet kapjunk.

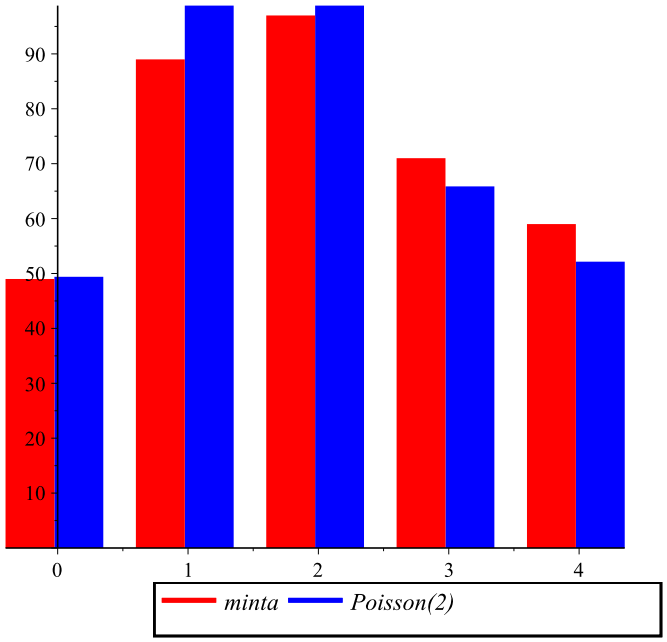
```
> infolevel[Statistics] := 1:
statisztika := ChiSquareGoodnessOfFitTest(gyak,
elm_gyak, level = alpha);
Chi-Square Test for Goodness-of-Fit
-----------------------------------
Null Hypothesis:
Observed sample does not differ from expected sample
Alt. Hypothesis:
Observed sample differs from expected sample
Categories:                5
Distribution:              ChiSquare(4)
Computed statistic:        2.30729
Computed pvalue:           0.679442
Critical value:            13.2767041359876
Result: [Accepted]
There is no statistical evidence against the null
hypothesis
```

$$statisztika := hypothesis = true, criticalvalue = 13.2767041359876, distribution = ChiSquare(4), pvalue = 0.679442271133119, statistic = 2.307289344 \qquad \textbf{(1.1.2.1.7)}$$

Az outputban a következôket látjuk: null és ellenhipotézis szövegesen, kategóriák száma (5), próbastatisztika eloszlása (

$(\chi^2)_4$), próbastatisztika értéke (2.307), p-érték (0.679), kritikus érték (13.277), döntés (elfogadás/elutasítás) szöveges magyarázattal.
Jelen esetben *elfogadjuk a nullhipotézist* az adott szignifikanciaszinten, azaz nem találtunk statisztikai bizonyítékot a nullhipotézis ellenében. Elfogadhatjuk azt a feltételezést, hogy a szabálytalankodó gépjármûvek száma $Poisson(2)$ eloszlást követ!
Rajzoljuk fel oszlopdiagramon az empirikus és elméleti gyakoriságokat, hogy vizualizáljuk az illeszkedés mértékét!

```
> ColumnGraph([gyak, elm_gyak], color = [red, blue],
  offset = -0.4, legend = [`minta`, `Poisson(2)`]);
```



Az illeszkedés az ábra alapján is meggyôzô!
b) Adjuk meg a kritikus értéket, a próbastatisztika értékét és a p-értéket! Magyarázzuk meg a döntést a p-érték alapján is!
A fenti próba eredményét elmentettük a *statisztika* változóba, ami egyenletek formájában tartalmazza a keresett értékeket. Olvassuk ki ezeket!

```
> kritikus_ertek := rhs(statisztika[2]); # csak az
  egyenletek jobb oldalán álló számot kérjük
  probastatisztika := rhs(statisztika[5]);
  p_ertek := rhs(statisztika[4]);
```

$$kritikus\_ertek := 13.2767041359876$$

$$probastatisztika := 2.307289344$$

(1.1.2.1.8)

$$p\_ertek := 0.679442271133119 \quad \textbf{(1.1.2.1.8)}$$

Lássuk röviden összefoglalva az illeszkedésvizsgálat döntési folyamatát!
Általánosságban véve minden statisztikai próbához tartozik egy jól meghatározott, mintából számolható *próbastatisztika*, melynek ismert az elméleti eloszlása és egy *kritikus tartomány*, mely a végsô döntést (elfogadás/elutasítás) szolgálja. A nullhipotézis visszautasítjuk, ha a próbastatisztika értéke a kritikus tartományba esik, egyébként pedig elfogadjuk (*elfogadási tartomány*). Itt fontos megjegyezni, hogy az elfogadás nem jelenti automatikusan a nullhipotézis igazságát, csupán azt, hogy nem nem vetjük el, mert nem találtunk ellene szóló statisztikai bizonyítékot az adott *megbízhatósági szint*en $(1-\alpha)$.
A kritikus tartományba esést úgy döntjük el, hogy a próbastatisztika kiszámolt értékét összehasonlítjuk egy ún. *kritikus érték*kel, mely a szignifikanciaszinttôl és esetleg a minta egyéb determinisztikus (rögzített) paramétereitôl függ (pl. *mintaméret*, *kategóriák száma*) és a *próbastatisztika* (ismert) *eloszlásfüggvényének inverzé*vel számolható. Ez utóbbira a Statistics csomag Quantile eljárását használhatjuk.
A fent leírt döntési folyamat alternatív múdon úgy is elvégezhetô, hogy kiszámítjuk az ún. *p-értéket*, vagyis hogy mekkora szignifikanciaszinten esne éppen egybe a kritikus érték a próbastatisztika értékével (még éppenhogy elfogadnánk H_0-t), és azt összehasonlítjuk a feladatban megadott szignifikanciaszinttel. Ha a p-érték nagyobb, mint a szignifikanciaszint, elfogadunk, ha kisebb, elutasítunk (egyenlôség esetén is elfogadunk). A p-értéket a próbastatisztika eloszlásfüggvényével, a próbastatisztika (mintán felvett) értékébôl számolhatjuk.

Az illeszkedésvizsgálatnál a fent leírt komponensek a következôk:

- Próbastatisztika: $d = \sum_{i=1}^{n} \frac{\nu_i^2}{N p_i} - N$, ahol ν_i-k a mintagyakoriságok, $N \cdot p_i$-k az elméleti gyakoriságok, N a mintaméret, n pedig a kategóriák száma
- Próbastatisztika eloszlása: $\left(\chi^2\right)_{n-1}$ ($n-1$ szabadsági fokú khí-négyzet eloszlás). Ez elméletileg igazolható, de itt nem tárgyaljuk részletekbe menôen.
- Kritikus érték: $\chi^2_{n-1,\alpha} = F^{-1}_{(\chi^2)_{n-1}}(1-\alpha)$ (a $\left(\chi^2\right)_{n-1}$ eloszlás $(1-\alpha)$-kvantilise), ahol α a szignifikanciaszint, $F^{-1}_{(\chi^2)_{n-1}}$ pedig a $\left(\chi^2\right)_{n-1}$ eloszlás eloszlásfüggvényének inverze.
- Kritikus tartomány: $\left\{d > \chi^2_{n-1,\alpha}\right\}$, azaz a kritikus tartomány a kritikus értéktôl jobbra lévô félegyenes; csak akkor utasítunk el, ha a próbastatisztika értéke nagyobb, mint a kritikus érték
- Elfogadási tartomány: $\left\{d \le \chi^2_{n-1,\alpha}\right\}$ (a kritikus tartomány komplementere)
- p-érték: $p_{érték} = 1 - F_{(\chi^2)_{n-1}}(d)$; ez alapján a kritikus tartomány ekvivalens megadása: $\left\{p_{érték} < \alpha\right\}$

```
> d := sum(gyak[i]^2/elm_gyak[i], i = 1..n) - N; #
  próbastatisztika értéke
```

$$d := 2.3072892 \quad \textbf{(1.1.2.1.9)}$$

```
> Y := RandomVariable(ChiSquare(n-1)); # khí-négyzet
  eloszlású n-1 szabadsági fokkal
```

$$Y := \_R0 \quad \textbf{(1.1.2.1.10)}$$

```
> kritikus_ertek := Quantile(Y, 1 - alpha); # kritikus
  érték
```

$$kritikus\_ertek := 13.2767041359876 \quad \textbf{(1.1.2.1.11)}$$

```
> p_ertek := 1 - CDF(Y, d); # p-érték
```

$$p\_ertek := 0.679442297338026 \quad \textbf{(1.1.2.1.12)}$$

Döntés a próbastatisztika alapján:

```
> # `elutasítás` = evalb(d > kritikus_ertek);
  `elfogadás` = evalb(d <= kritikus_ertek);
```

$$elfogadás = true \quad \textbf{(1.1.2.1.13)}$$

Döntés a p-érték alapján:

```
> # `elutasítás` = evalb(p_ertek < alpha);
  `elfogadás` = evalb(p_ertek >= alpha);
```

$$elfogadás = true \quad \textbf{(1.1.2.1.14)}$$

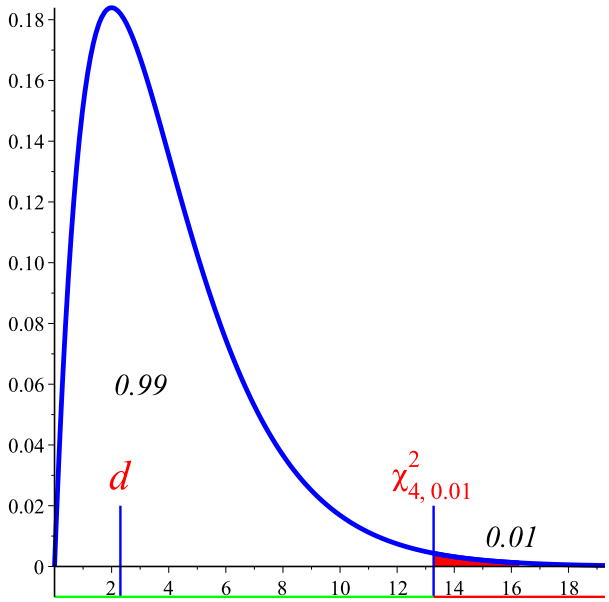
Mivel a két döntés ekvivalens, ugyanazt az eredményt kaptuk!

c) Szemléltessük a kritikus tartományt és a döntést grafikonon a sûrûségfüggvény alatti terület beszínezésével!

Elôször a $\left(\chi^2\right)_4$ eloszlású próbastatisztika sûrûségfüggvényét ábrázoljuk, majd beszínezzük a grafikon alatt, a kritikus értéktôl jobbra lévô (*kritikus tartomány* feletti) területet. Bejelöljük továbbá a kritikus értéket ($\chi^2_{4,\,0.01}$) és a próbastatisztika értékét (d), hogy össze tudjuk hasonlítani ôket. Végül berajzoljuk még az elfogadási és elutasítási tartományt, valamint felírjuk azok valószínûségét a nullhipotézis esetén.

```
> P1 := DensityPlot(Y, range = 0..kritikus_ertek + 6,
  thickness = 3, color = blue):
  P2 := plot(PDF(Y, t), t = kritikus_ertek..kritikus_ertek
  + 3, thickness = 3, color = blue, filled = [color = red,
  transparency = 0.5], labels = ["",""]):
  P3 := plot({[[probastatisztika, 0.02],
  [probastatisztika, -0.01]], [[kritikus_ertek, 0.02],
  [kritikus_ertek, -0.01]]}, thickness = 1, color = blue):
  P4 := plots[textplot]([[probastatisztika, 0.03, 'd',
  'font'=["times","roman",20]], [kritikus_ertek, 0.03,
  'chi[4, 0.01]^2', 'font'=["times","roman",15]]], color =
  red):
  P5 := plot({[[0, -0.01], [kritikus_ertek, -0.01]], [
  [kritikus_ertek, -0.01], [kritikus_ertek + 6, -0.01]]},
  color = [green, red]);
  P6 := plots[textplot]([kritikus_ertek/2, -0.02,
  `Elfogadás`, 'font'=["times","roman",15]], color =
  green):
  P7 := plots[textplot]([kritikus_ertek + 3, -0.02,
  `Elutasítás`, 'font'=["times","roman",15]], color = red)
  :
  P8 := plots[textplot]([[3, 0.06, `0.99`, 'font'=
  ["times","roman",15]], [16, 0.01, `0.01`, 'font'=
  ["times","roman",15]]]):
  plots[display](P1, P2, P3, P4, P5, P6, P7, P8);
```

$$P5 := PLOT(\ldots)$$



A rajzon szépen látszik, hogy a próbastatisztika jócskán a kritikus értéktôl balra, az elfogadási tartományba esik. Ez azt is jelenti, hogy a nullhipotézist (a háttéreloszlás 2 paraméterû Poisson) nagyon simán elfogadtuk! A nullhipotézist csak abban az esetben utasítanánk el, ha d a kritikus értéktôl jobbra, a pirossal jelölt elutasítási tartományba esne!

3. Feladat (Konfidencia-intervallum ismert szórással)

Méréseket végzünk egy élelmiszer szénhidrát-tartalmának meghatározására. 14 mérésbôl az $\overline{X} = 4.75$ % átlagot kaptuk. Tegyük fel, hogy a minta normál eloszlásból származik és ismert a szórás: $\sigma = 0.4$ %.

a) Teszteljük 99 %-os konfidenciaszint mellett, hogy $m_0 = 4.5$ % lehet-e a sokaság várható értéke! Adjuk meg a várható értékre vonatkozóan a 99 %-os szintnek megfelelô konfidencia-intervallumot!

b) Végezzük el a tesztet 95 %-os megbízhatóság mellett is! Szemléltessük a két döntés közötti különbséget a két konfidencia-intervallum felrajzolásával és a sûrûségfüggvény alatti megfelelô terület besatírozásával!

c) Hány elemû mintát kellene vennünk, hogy a konfidencia-intervallum hossza 5 %-os szignifikanciaszint mellett 0.2 % legyen? Ellenôrizzük a számítást véletlen adatok generálásával, feltételezve, hogy pont $\mu = \overline{X} = 4.75$ % az elméleti várható érték!

Megoldás

```
> restart;
```

Gyûltsük ki az adatokat!

```
> n := 14;
  atlag := 4.75;
  sigma := 0.4;
```

$$n := 14$$

$$atlag := 4.75$$

$$\sigma := 0.4 \qquad \textbf{(1.1.3.1.1)}$$

a) Teszteljük 99 %-os konfidenciaszint mellett, hogy $m_0 = 4.5$ % lehet-e a sokaság várható értéke! Adjuk meg a várható értékre vonatkozóan a 99 %-os szintnek megfelelô konfidencia-intervallumot!

A statisztikai próbák egy másik fontos típusa az ún. paraméteres próba. Ekkor feltételezzük, hogy a háttéreloszlás típusa ismert, de bizonyos paraméterei nem. Azt szeretnénk tesztelni a minta alpján, hogy egy ilyen paraméter értéke megegyezhet-e egy adott értékkel. A leggyakrabban elôforduló paraméteres próba a normál eloszlás várható értékére (μ) vonatkozik ismert szórás (σ) mellett. Ez az ún. (egymintás, kétoldali) u-próba.

A nullhipotézisünk "$H_0: \ \mu = m_0$", ahol az elôre adott m_0 érték az elvárt/sejtett várható érték.

A próba komponensei és menete megegyezik az illeszkedésvizsgálatnál leírtakkal, de más eloszlással és képletekkel kell számolni. A próbastatisztika itt standard normál eloszlású.

- Próbastatisztika: $u = \dfrac{\overline{X} - m_0}{\dfrac{\sigma}{\sqrt{n}}}$, ahol n a mintaméret, σ az *ismert* szórás, m_0 pedig a feltételezett várható érték, amit elfogadni/cáfolni szeretnénk
- Próbastatisztika eloszlása: $N(0, 1)$ (standard normál eloszlás).
- Kritikus érték: $u_{\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$ (a standard normál eloszlás $\left(1 - \dfrac{\alpha}{2}\right)$-kvantilise), ahol α a szignifikanciaszint, Φ pedig a standard normál eloszlás eloszlásfüggvénye.
- Kritikus tartomány: $\left\{|u| > u_{\frac{\alpha}{2}}\right\}$, azaz a kritikus tartomány két félegyenes uniója, az origótól a kritikus értéknél távolabb lévô pontok halmaza; csak akkor utasítunk el, ha a próbastatisztika abszolútértéke nagyobb, mint a kritikus érték
- Elfogadási tartomány: $\left\{|u| \le u_{\frac{\alpha}{2}}\right\}$ (a kritikus tartomány komplementere)
- p-érték: $p_{érték} = 2 \cdot (1 - \Phi(u))$; ez alapján a kritikus tartomány ekvivalens megadása: $\left\{p_{érték} < \alpha\right\}$

Az u-próbával egyenértékû a várható értékre vonatkozó, $1 - \alpha$ szintû konfidencia-intervallum megadása. Ez lényegében azt jelenti, hogy a minta $\overline{X}$ átlaga körül megadunk egy szimmetrikus intervallumot, melybe az elméleti várható érték $1 - \alpha$ valószínûséggel beleesik, feltéve, hogy a nullhipotézis ($\mu = m_0$) igaz. Ezután megnézzük, hogy m_0 beleesik-e ebbe az intervallumba és ez alapján döntünk – ha beesik, elfogadjuk H_0-t; ha nem, akkor pedig elutasítjuk.

- Konfidencia-intervallum: $I_K = \left[\overline{X} - u_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}, \overline{X} + u_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}\right]$

• Konfidencia-intervallum hossza: $L = 2\, u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ (ez nem függ a konkrét mintától, csak a mintaelemek számától)
• Elfogadás: ha $m_0 \in I_K$

Kezdjük hozzá a feladatrész megoldásához! Két új információt tudunk:

```
> alpha := 0.01; # szignifikanciaszint
  m0 := 4.5;
```

$$\alpha := 0.01$$

$$m0 := 4.5 \qquad \textbf{(1.1.3.1.2)}$$

A statisztikai sokaság normál eloszlásúnak van feltételezve ismert szórással. A nullhipotézis a várható értékre vonatkozik: $H_0 : \mu = m_0$. Az ellenhipotézis ennek tagadása: $H_1 : \mu \neq m_0$. Mivel ismert a szórás, ezért *u*-próbát végzünk. Erre szolgál a Statistics csomag *OneSampleZTest* eljárása. Érdemes megjegyezni, hogy az elsô paraméternek egy teljes mintát kellene megadnunk, de ha ez nem ismert (mint itt), akkor helyettesíthetô egy olyan *n* elemû listával, melyben minden adatpont megegyezik az átlaggal. Állítsuk a Statistics csomag információs szintjét 1-re, hogy beszédes kimenetet kapjunk!

```
> with(Statistics):
  infolevel[Statistics] := 1:
  statisztika := OneSampleZTest([seq(atlag, i = 1..n)],
  m0, sigma, confidence = 1 - alpha);
Standard Z-Test on One Sample
-----------------------------
Null Hypothesis:
Sample drawn from population with mean 4.5 and known
standard deviation 0.4
Alt. Hypothesis:
Sample drawn from population with mean not equal to
4.5 and known standard deviation 0.4
Sample size:              14
Sample mean:              4.75
Distribution:             Normal(0,1)
Computed statistic:       2.33854
Computed pvalue:          0.0193595
Confidence interval:      4.47463226446128 ..
5.02536773553872
                          (population mean)
Result: [Accepted]
There is no statistical evidence against the null
hypothesis
```

$$statisztika := hypothesis = true, confidenceinterval = 4.47463226446128 \qquad \textbf{(1.1.3.1.3)}$$
$$..5.02536773553872, distribution = Normal(0, 1), pvalue$$
$$= 0.0193594673670699, statistic = 2.33853586596743$$

Az outputban a következôket látjuk: null és ellenhipotézis szövegesen, mintaméret (14), mintaátlag (4.75), próbastatisztika eloszlása (N(0, 1)), próbastatisztika értéke (2.336), *p*-érték (0.019), konfidencia-intervallum a várható értékre([4.475, 5.025]), döntés (elfogadás/elutasítás) szöveges magyarázattal.
Jelen esetben *elfogadjuk a nullhipotézist* az adott szignifikanciaszinten, azaz nem találtunk statisztikai bizonyítékot a nullhipotézis ellen. Elfogadhatjuk azt a feltételezést, hogy a háttéreloszlás várható értéke $\mu = m_0 = 4.5$ %!

A próba eredményét elmentettük a *statisztika* változóba, amibôl kiolvashatjuk a konfidenciaintervallumot:

```
> konfidencia_intervallum := rhs(statisztika[2]); # az
  egyenlet jobb oldalát kell venni
```

$$konfidencia\_intervallum := 4.47463226446128..5.02536773553872 \qquad \textbf{(1.1.3.1.4)}$$

Lássuk, hogyan zajlik az *u*-próba lépésenként!
Elôször meghatározzuk a próbastatisztika értékét, mely az egyes mintaelemektôl nem, csak a mintaátlagtól függ.

```
> u := evalf((atlag - m0)/(sigma/sqrt(n))); #
  próbastatisztika értéke
```

$$u := 2.338535867 \qquad \textbf{(1.1.3.1.5)}$$

Ezután számoljuk ki a kritikus értéket, mely a standard normál eloszlás $1 - \frac{\alpha}{2}$ kvantilise!

```
> X := RandomVariable(Normal(0, 1));
  kritikus := Quantile(X, 1 - alpha/2);
```

$$X := \_R$$

$$kritikus := 2.57582930355009 \qquad \textbf{(1.1.3.1.6)}$$

Határozzuk meg a *p*-értéket a fent leírt képlettel! A *p*-érték annak a valószínûsége, hogy a próbastatisztika abszolút értéke nagyobb, mint az aktuális mintából kiszámolt *u* érték.

```
> p_ertek := 2*(1 - CDF(X, u));
```

$$p\_ertek := 0.0193594673135720 \qquad \textbf{(1.1.3.1.7)}$$

A próbastatisztika értéke beleesik az elfogadási tartományba, így elfogadjuk a nullhipotézist:

```
> elfogadas = evalb(abs(u) <= kritikus);
```

$$elfogadas = true \qquad \textbf{(1.1.3.1.8)}$$

Ugyanez a döntés a *p*-érték alapján:

```
> elfogadas = evalb(p_ertek >= alpha);
```

$$elfogadas = true \qquad \textbf{(1.1.3.1.9)}$$

b) Végezzük el a tesztet 95 %-os megbízhatóság mellett is! Szemléltessük a két döntés közötti különbséget a két konfidencia-intervallum felrajzolásával és a hibaszinteket a sûrûségfüggvény alatti megfelelô terület besatírozásával!

```
> alpha2 := 0.05;
```

$$\alpha 2 := 0.05 \qquad \textbf{(1.1.3.1.10)}$$

```
> statisztika2 := OneSampleZTest([seq(atlag, i = 1..n)],
  m0, sigma, confidence = 1 - alpha2);
Standard Z-Test on One Sample
-----------------------------
Null Hypothesis:
Sample drawn from population with mean 4.5 and known
standard deviation 0.4
Alt. Hypothesis:
Sample drawn from population with mean not equal to
4.5 and known standard deviation 0.4
Sample size:              14
Sample mean:              4.75
Distribution:             Normal(0,1)
Computed statistic:       2.33854
Computed pvalue:          0.0193595
```

```
Confidence interval:     4.54047103648669 ..
4.95952896351331
                                (population mean)
Result: [Rejected]
There exists statistical evidence against the null
hypothesis
```

$$statisztika2 := hypothesis = false,\ confidenceinterval = 4.54047103648669 \ldots 4.95952896351331,\ distribution = Normal(0,\ 1),\ pvalue = 0.0193594673670699,\ statistic = 2.33853586596743 \quad \textbf{(1.1.3.1.11)}$$

```
> konfidencia_intervallum2 := rhs(statisztika2[2]);
```

$$konfidencia\_intervallum2 := 4.54047103648669 \ldots 4.95952896351331 \quad \textbf{(1.1.3.1.12)}$$

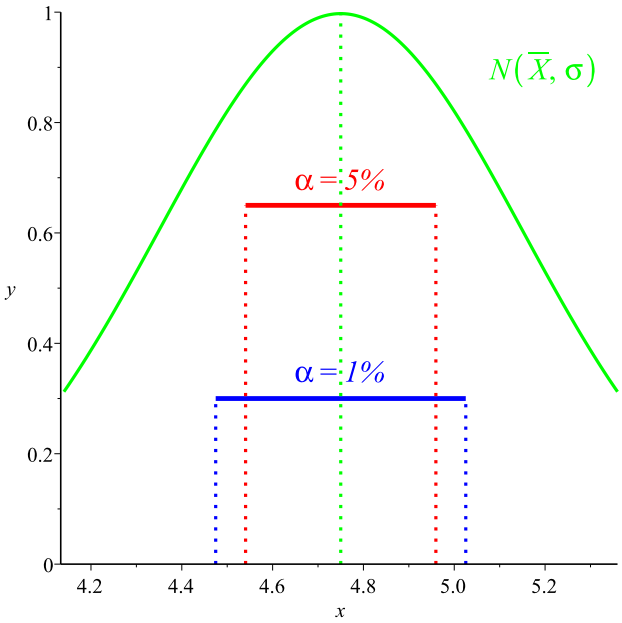
```
> A := lhs(konfidencia_intervallum);
  B := rhs(konfidencia_intervallum);
  A2 := lhs(konfidencia_intervallum2);
  B2 := rhs(konfidencia_intervallum2);
```

$$A := 4.47463226446128$$
$$B := 5.02536773553872$$
$$A2 := 4.54047103648669$$
$$B2 := 4.95952896351331 \quad \textbf{(1.1.3.1.13)}$$

```
> X := RandomVariable(Normal(atlag, sigma));
```

$$X := \_R0 \quad \textbf{(1.1.3.1.14)}$$

```
> P1 := plot(PDF(X, x), x = A2 - 0.4..B2 + 0.4, y = 0..1,
  color = green, thickness = 2):
  P2 := plot([[A, 0.3], [B, 0.3]], [[A2, 0.65], [B2,
  0.65]]], color = [blue, red], thickness = 3): # zh-ban
  ez a sor elég + az x és y tartomány beállítása
  P3 := plot([[[A, 0], [A, 0.3]], [[B, 0], [B, 0.3]], [
  [A2, 0], [A2, 0.65]], [[B2, 0], [B2, 0.65]], [[atlag,
  0], [atlag, 1]]], color = [blue, blue, red, red, green],
  thickness = 2, linestyle = dot):
  P4 := plots[textplot]({[atlag, 0.35, ''alpha''=`1%`,
  font=["times","roman",15], color = blue], [atlag, 0.7,
  ''alpha''=`5%`, font=["times","roman",15], color = red],
  [5.2, 0.9, ''N(conjugate(X), sigma)'', font=["times",
  "roman",15], color = green]}):
  plots[display](P1, P2, P3, P4);
```




```
> U := RandomVariable(Normal(0, 1));
```

$$U := \_R1 \qquad \textbf{(1.1.3.1.15)}$$

```
> kritikus := Quantile(U, 1 - alpha/2); # alpha/2-t kell
  kivonni, mert kétoldali nullhipotézisünk van
  kritikus2 := Quantile(U, 1 - alpha2/2);
  probastatisztika := rhs(statisztika[5]); # ez ugyanaz
  mindkét esetben
```

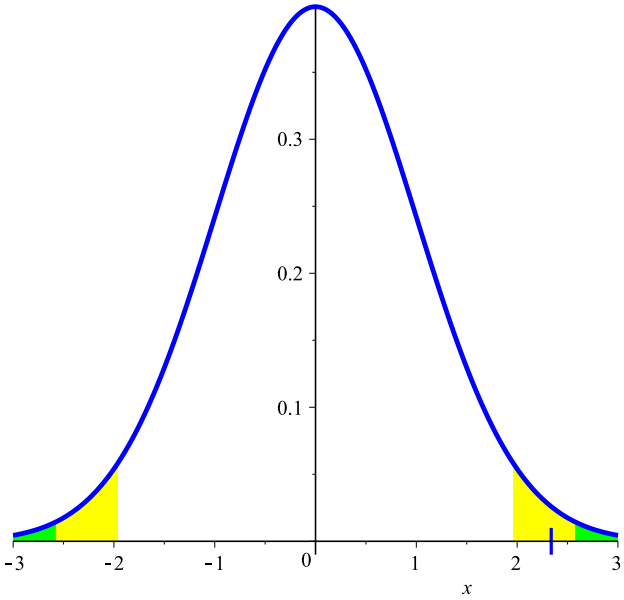
$$kritikus := 2.57582930355009$$

$$kritikus2 := 1.95996398453944$$

$$probastatisztika := 2.33853586596743 \qquad \textbf{(1.1.3.1.16)}$$

```
> P5 := plot(PDF(U, x), x = -3..3, color = blue, thickness
  = 3, caption = "zöld terület = alpha = 1%; sárga terület
  = alpha2 = 5%"):
  P_bal1 := plot(PDF(U, x), x = -3..-kritikus, color =
  blue, thickness = 2, filled = [color = green]):
  P_jobb1 := plot(PDF(U, x), x = kritikus..3, color =
  blue, thickness = 2, filled = [color = green]):
  P_bal2 := plot(PDF(U, x), x = -3..-kritikus2, color =
```

```
blue, thickness = 2, filled = [color = yellow]):
P_jobb2 := plot(PDF(U, x), x = kritikus2..3, color =
blue, thickness = 2, filled = [color = yellow]):
P6 := plots[textplot]([probastatisztika, -0.02, "u",
font = ["times","roman",15], color = blue]):
P7 := plot([[probastatisztika, -0.01],
[probastatisztika, 0.01]], color = blue, thickness = 2):
plots[display](P5, P_bal1, P_jobb1, P_bal2, P_jobb2, P6,
P7);
```



zöld terület = alpha = 1%; sárga terület = alpha2 = 5%

c) Hány elemû mintát kellene vennünk, hogy a konfidencia-intervallum hossza 5 %-os szignifikanciaszint mellett legfeljebb 0.2 % legyen? Ellenôrizzük a számítást véletlen adatok generálásával, feltételezve, hogy pont $\mu = \overline{X} = 4.75$ % az elméleti várható érték!

```
> L := 0.2;
```

$$L := 0.2 \qquad \textbf{(1.1.3.1.17)}$$

```
> evalf(2*kritikus*sigma/sqrt(n));
  evalf(2*kritikus2*sigma/sqrt(n));
```

$$0.550735470930242$$
$$0.419057926914626 \qquad \textbf{(1.1.3.1.18)}$$

```
> solve(2*kritikus2*sigma/sqrt(N) <= L, {N});
```

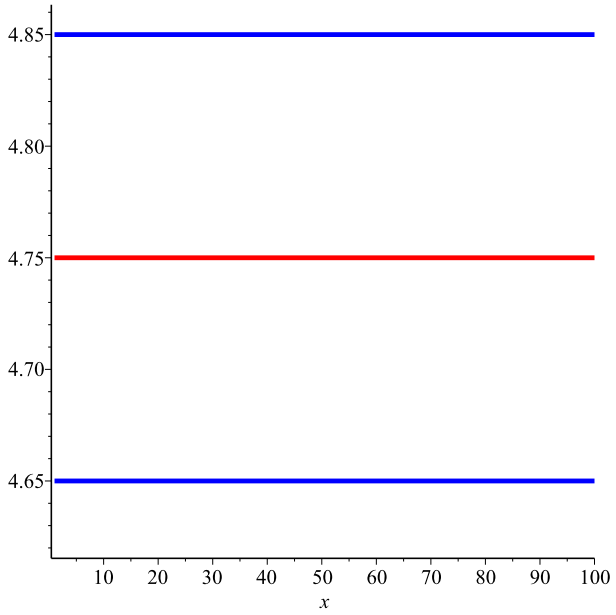
$$\{61.46334113 \le N\} \qquad \textbf{(1.1.3.1.19)}$$

```
> n2 := 62;
```

$$n2 := 62 \qquad \textbf{(1.1.3.1.20)}$$

```
> randomize():
  pontok:=[seq([i, Mean(Sample(X, n2))], i = 1..100)]:
```

```
> plot([pontok, atlag, atlag - L/2, atlag + L/2], x = 1.
  .100, style = [point, line, line, line], symbol = cross,
  symbolsize = 15, color = [black, red, blue, blue],
  thickness = 3);
```



4. Feladat (Lineáris regresszió)

Véletlenszerûen kiválasztott városokban vizsgálták a népesség száma és a közcsatorna-hálózatba bekapcsolt lakások aránya közötti összefüggést. Az adatokat az alábbi táblázatban foglalták össze, ahol *X* jelöli a népességet ezer fôben és *Y* jelöli a bekapcsolt

lakások arányát százalékban.

X	3	4	7	2	5	3	6	1	7	3
Y	5	5	8	4	7	5	8	3	8	4

a) Ábrázoljuk az összetartozó értékpárokat a síkon a *Statistics* csomag *ScatterPlot* eljárásával!
b) Számoljuk ki *X* és *Y* között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a népesség és a közcsatorna-hálózatba bekapcsolt lakások aránya között?
c) Határozzuk meg a regressziós egyenest, majd rajzoljuk fel azt az a) feladatrészben kapott ábrára!
d) A kapott lineáris összefüggés segítségével adjunk becslést a közcsatorna-hálózatba bekapcsolt lakások arányára egy 25 000, egy 61 000 és egy 75 000 lakosú városban!
e) Számoljuk ki a közcsatorna-hálózatba bekapcsolt lakások arányát az illesztett lineáris modell alapján, majd számoljuk ezek eltéréseit a mérési értékektôl. Rajzoljuk fel az eltérések (reziduálisok) hisztogrammját!
f) Gyôzôdjünk meg arról, hogy az eltérések átlaga 0! Rajzoljuk be az eltérések hisztogramjának ábrájába a 0 várható értékû és az eltérésekkel azonos szórású normál eloszlás sûrûségfüggvényét!
g) Teszteljük, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel és az eltérések szórásával $\alpha = 1\ \%$ szignifikanciaszinten, ha a terjedelmet 10 egyenlô részre osztjuk?
h) Rajzoljuk fel az illesztett és a mért adatsorokat a *QuantilePlot* eljárás segítségével! Számoljuk ki R^2 értékét! Mennyire jó a lineáris modell ez alapján?
i) Adjunk 99 %-os megbízhatóságú konfidencia-intervallumot az *a* és *b* regressziós együtthatók értékére! Becsüljük meg a reziduális szórását!

Megoldás

Mérnöki feladatok megoldása, gazdasági, fizikai, biológiai folyamatok vizsgálata során rendkívül hasznos, ha két véletlen változó között függvényszerû kapcsolatot tudunk feltárni és azt le is tudjuk írni. Ilyenkor az egyik változót (X) függetlennek, a másikat (Y) függônek tekintjük, azaz a függô mennyiséget mérjük a független mennyiség rögzített értékei mellett. A legegyszerûbb függvénykapcsolat a lineáris függés, melyet a *lineáris regresszió* segítségével adhatunk meg. Ha az adataink táblázatos formában

$$\begin{bmatrix} X & x_1 & x_2 & \dots & x_n \\ Y & y_1 & y_2 & \dots & y_n \end{bmatrix}$$

akkor a modellünk:

$$\widehat{y}_i = a + b \cdot x_i$$

Az eltérések (hibák, *reziduálisok*):

$$e_i = y_i - \widehat{y}_i.$$

Az a és b regressziós együtthatókat a *legkisebb négyzetek elve* alapján határozzuk meg, vagyis úgy, hogy az eltérések négyzetösszege, $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2$, minimális legyen.

Ennek az optimalizációs problémának a megoldása:

$$b = r\frac{S_y}{S_x} = \frac{n \cdot Cov(x,y)}{(n-1) \cdot S_x^2} \text{ és}$$

$$a = \overline{y} - b \cdot \overline{x}$$

ahol $r = Corr(x, y)$ a korrelációs együttható, $\overline{x}$ ill. $\overline{y}$ a két adatsor átlaga, S_x ill. S_y pedig a két adatsor (korrigált) szórása. Ezt visszahelyettesítve a modellbe kapjuk, hogy

$$\widehat{y}_i = \overline{y} - r\frac{S_y}{S_x}\overline{x} + r\frac{S_y}{S_x} \cdot x_i$$

Ez a képlet könnyebben megjegyezhetô a standardizált formában:

$$\frac{\widehat{y}_i - \overline{y}}{S_y} = r \cdot \frac{x_i - \overline{x}}{S_x}$$

Ebbôl az is jól látszik, hogy miért a korrelációs együttható adja meg a két változó közötti lineáris összefüggés erôsségét: r a két mennyiség standardizáltja közötti skálázási tényezô! A regressziós egyenes egyenlete a fentiek alapján:

$$y = bx + a = \\ = r\frac{S_y}{S_x}x + \overline{y} - r\frac{S_y}{S_x}\overline{x}$$

Az $\varepsilon = Y - a - b \cdot X$ elméleti reziduális szórására torzítatlan becslést adhatunk a minta alapján:

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - a - bx_i)^2}{n-2}}$$

Ez a *reziduális standard eltérés*, mely jól látható összefüggésben van a legkisebb négyzetek módszerével minimalizált reziduális négyzetösszeggel.

Reziduális négyzetösszeg (Residual Sum of Squares): $RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$

Teljes négyzetösszeg (Total Sum of Squares): $TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2$

Determinációs együttható: $R^2 = 1 - \frac{RSS}{TSS}$ - azt mondja meg, hogy a függô változó varianciájának hányadrésze magyarázható a lineáris modellel. Lineáris modell esetén $R^2 = r^2$ (de nem-lineáris esetben a kettô különbözhet).

```
> with(Statistics):
  n := 3;
  x := [0, 1, 2];
  y := [0, 1, 1];
```

$$n := 3$$
$$x := [0, 1, 2]$$
$$y := [0, 1, 1] \qquad \textbf{(1.1.4.1.1)}$$

```
> corrxy := Correlation(x, y);
  covxy := Covariance(x, y);
```

$$corrxy := 0.866025403784438$$
$$covxy := 0.500000000000000 \qquad \textbf{(1.1.4.1.2)}$$

```
> mx := Mean(x);
  my := Mean(y);
```

$$mx := 1.$$
$$my := 0.666666666666667 \qquad \textbf{(1.1.4.1.3)}$$

```
> sx := StandardDeviation(x);
  sy := StandardDeviation(y);
```

$$sx := 1.$$

$$sy := 0.577350269189626 \qquad \textbf{(1.1.4.1.4)}$$

```
> sqrt(sum((y[i]-my)^2, i = 1..nops(y))/2);
  sum((x[i]-mx)*(y[i]-my), i = 1..nops(x))/3;
```

$$0.5773502692$$

$$0.3333333333 \qquad \textbf{(1.1.4.1.5)}$$

```
> n*covxy/((n-1)*sx*sy); # így kapjuk a korrelációt Maple
  16-ban, de Maple 17-ben már covxy/(sx*sy)
```

$$0.866025403784438 \qquad \textbf{(1.1.4.1.6)}$$

```
> restart;
```

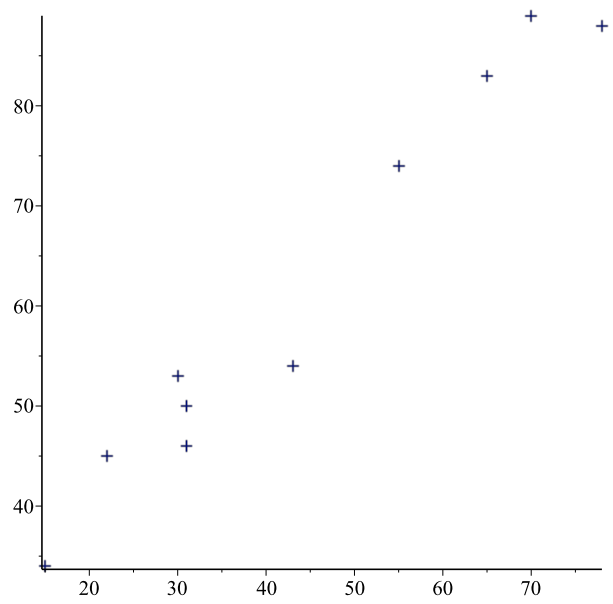
```
> minta_X := [30, 43, 70, 22, 55, 31, 65, 15, 78, 31];
  minta_Y := [53, 54, 89, 45, 74, 50, 83, 34, 88, 46];
```

$$minta\_X := [30, 43, 70, 22, 55, 31, 65, 15, 78, 31]$$

$$minta\_Y := [53, 54, 89, 45, 74, 50, 83, 34, 88, 46] \qquad \textbf{(1.1.4.1.7)}$$

a) Ábrázoljuk az összetartozó értékpárokat a síkon a *Statistics* csomag *ScatterPlot* eljárásával!

```
> with(Statistics):
  P1 := ScatterPlot(minta_X, minta_Y, symbol = cross,
  symbolsize = 15): P1;
```



b) Számoljuk ki *X* és *Y* között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a népesség és a közcsatorna-hálózatba bekapcsolt lakások aránya között?

```
> r := Correlation(minta_X, minta_Y);
```

$$r := 0.981148387646947 \qquad \textbf{(1.1.4.1.8)}$$

c) Határozzuk meg a regressziós egyenest, majd rajzoljuk fel azt az a) feladatrészben kapott ábrára!

```
> regresszios_egyenes := unapply(LinearFit([1, x],
  minta_X, minta_Y, x), x);
```

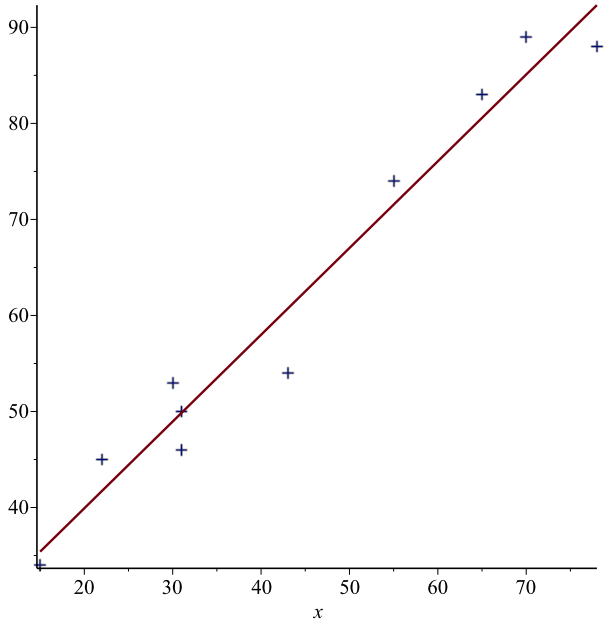
$$regresszios\_egyenes := x \rightarrow 21.851057827926645 + 0.9033850493653034\, x \qquad \textbf{(1.1.4.1.9)}$$

```
> a := coeff(regresszios_egyenes(x), x, 0);
  b := coeff(regresszios_egyenes(x), x, 1);
```

$$a := 21.8510578279266$$

$$b := 0.903385049365303 \qquad \textbf{(1.1.4.1.10)}$$

```
> P2 := plot(regresszios_egyenes(x), x = min(minta_X)..max
  (minta_X)):
  plots[display](P1, P2);
```



d) A kapott lineáris összefüggés segítségével adjunk becslést a közcsatorna-hálózatba bekapcsolt lakások arányára egy 25 000, egy 61 000 és egy 75 000 lakosú városban!

```
> regresszios_egyenes(25);
  regresszios_egyenes(61);
  regresszios_egyenes(75);
```

44.4356840620592
76.9575458392102
89.6049365303244 **(1.1.4.1.11)**

e) Számoljuk ki a közcsatorna-hálózatba bekapcsolt lakások arányát az illesztett lineáris modell alapján, majd számoljuk ezek eltéréseit a mérési értékektôl. Rajzoljuk fel az eltérések (reziduálisok) hisztogrammját!

```
> modell_Y := map(k -> regresszios_egyenes(k), minta_X);
```

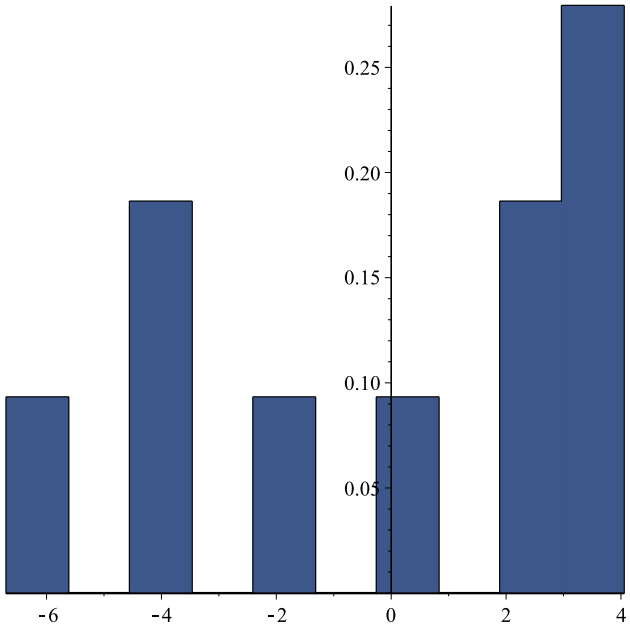
modell_Y := [48.9526093088857, 60.6966149506347, 85.0880112834979, **(1.1.4.1.12)**
41.7255289139633, 71.5372355430183, 49.8559943582511,

80.5710860366714, 35.4018335684062, 92.3150916784203,
49.8559943582511]

```
> rezidualisok := minta_Y - modell_Y;
```

rezidualisok := [4.04739069111425, -6.69661495063469, **(1.1.4.1.13)**
3.91198871650212, 3.27447108603668, 2.46276445698166,
0.144005641748947, 2.42891396332863, -1.40183356840620,
-4.31509167842032, -3.85599435825105]

```
> H := Histogram(rezidualisok, bincount = 10): H;
```



f) Gyôzôdjünk meg arról, hogy az eltérések átlaga 0! Rajzoljuk be az eltérések hisztogramjának ábrájába a 0 várható értékû és az eltérésekkel azonos szórású normál eloszlás sûrûségfüggvényét!

```
> Mean(rezidualisok);
```

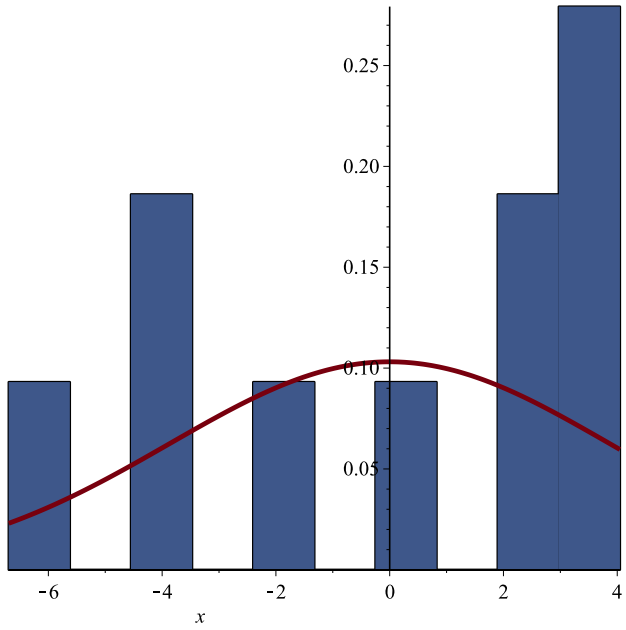
$$3.55271367880050\ 10^{-15}$$ **(1.1.4.1.14)**

```
> X := RandomVariable(Normal(0,StandardDeviation
```

```
(rezidualisok)));
```

$$X := \_R0 \qquad \textbf{(1.1.4.1.15)}$$

```
> P3 := plot(PDF(X, x), x = min(rezidualisok)..max
  (rezidualisok), thickness = 3):
  plots[display](H, P3);
```



g) Teszteljük, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel és az eltérések szórásával $\alpha = 1$ % szignifikanciaszinten, ha a terjedelmet 10 egyenlô részre osztjuk?

```
> infolevel[Statistics] :=1:
  ChiSquareSuitableModelTest(rezidualisok, X, bins = 10,
  level = 0.01);
Chi-Square Test for Suitable Probability Model
----------------------------------------------
Null Hypothesis:
Sample was drawn from specified probability
distribution
Alt. Hypothesis:
Sample was not drawn from specified probability
distribution
```

```
Bins:                    10
Distribution:            ChiSquare(9)
Computed statistic:      10.271
Computed pvalue:         0.328998
Critical value:          21.6659943178256
Result: [Accepted]
There is no statistical evidence against the null
hypothesis
```

$hypothesis = true, criticalvalue = 21.6659943178256, distribution = ChiSquare(9), pvalue = 0.328997701200926, statistic = 10.2709654505886$ **(1.1.4.1.16)**

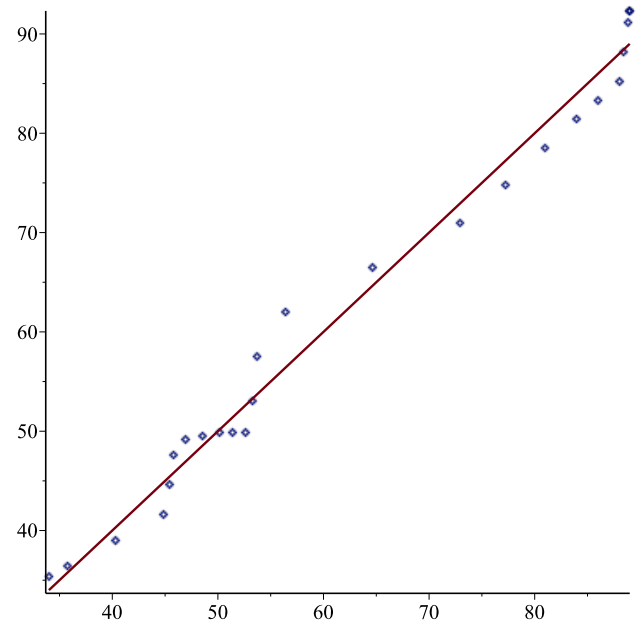
```
> FrequencyTable(rezidualisok, bins=10);
```

(1.1.4.1.17)

```
[[-6.69661495063469 ..-5.62221438645980, 1., 10.00000000, 1.,
    10.00000000],
    [-5.62221438645980 ..-4.54781382228490, 0., 0., 1., 10.00000000],
    [-4.54781382228490 ..-3.47341325811001, 2., 20.00000000, 3.,
    30.00000000],
    [-3.47341325811001 ..-2.39901269393511, 0., 0., 3., 30.00000000],
    [-2.39901269393511 ..-1.32461212976022, 1., 10.00000000, 4.,
    40.00000000],
    [-1.32461212976022 ..-0.250211565585325, 0., 0., 4., 40.00000000],
    [-0.250211565585325 ..0.824188998589570, 1., 10.00000000, 5.,
    50.00000000],
    [0.824188998589570 ..1.89858956276446, 0., 0., 5., 50.00000000],
    [1.89858956276446 ..2.97299012693936, 2., 20.00000000, 7.,
    70.00000000],
    [2.97299012693936 ..4.04739069111425, 3., 30.00000000, 10.,
    100.0000000]]
```

h) Rajzoljuk fel az illesztett és a mért adatsorokat a *QuantilePlot* eljárás segítségével! Számoljuk ki R^2 értékét! Mennyire jó a lineáris modell ez alapján?

```
> QuantilePlot(minta_Y, modell_Y);
```




```
> R2 := r^2;
```

$$R2 := 0.962652158582204 \qquad \textbf{(1.1.4.1.18)}$$

```
> N := numelems(minta_X);
  atlag := Mean(minta_Y);
```

$$N := 10$$

$$atlag := 61.6000000000000 \qquad \textbf{(1.1.4.1.19)}$$

```
> RSS := sum((minta_Y[i] - modell_Y [i])^2, i = 1..N);
  TSS := sum((minta_Y[i] - atlag)^2, i = 1..N);
```

$$RSS := 134.691255289140$$

$$TSS := 3606.400000 \qquad \textbf{(1.1.4.1.20)}$$

```
> R2 = 1 - RSS/TSS;
```

$$0.962652158582204 = 0.962652158582204 \qquad \textbf{(1.1.4.1.21)}$$

i) Adjunk 99 %-os megbízhatóságú konfidencia-intervallumot az *b* regressziós együtthatók értékére! Becsüljük meg a reziduális szórását!

```
> regresszio_output := LinearFit([1, x], minta_X, minta_Y,
  x, output = solutionmodule):
  eredmenyek := (regresszio_output:-Results)():
In LinearFit (container form)
```

```
> konf_int := eredmenyek[9];
```

$$konf\_int := \text{"confidenceintervals"} \qquad \textbf{(1.1.4.1.22)}$$

$$= \begin{bmatrix} 14.8013583713903 \,..28.9007572844630 \\ 0.758312306235144 \,..1.04845779249546 \end{bmatrix}$$

```
> konf_int_a := rhs(konf_int)[1];
  konf_int_b := rhs(konf_int)[2];
```

$$konf\_int\_a := 14.8013583713903 \,..28.9007572844630$$

$$konf\_int\_b := 0.758312306235144 \,..1.04845779249546 \qquad \textbf{(1.1.4.1.23)}$$

```
> rezidualis_szoras := rhs(eredmenyek[3]);
```

$$rezidualis\_szoras := 4.10321909129191 \qquad \textbf{(1.1.4.1.24)}$$

```
> rezidualis_szoras = sqrt(RSS/(N-2));
```

$$4.10321909129191 = 4.10321909129192 \qquad \textbf{(1.1.4.1.25)}$$

Gyakorló feladatok

Gy/1. Feladat (Konfidencia-intervallum ismeretlen szórással)

Egy adagológép által dobozba töltendô anyag tömegének várható értékére az elôírás 120 g, és feltételezzük, hogy a tömeg normál eloszlást követ.Véletlenszerû mintavétel során az alábbi értékeket kaptuk (grammban mérve):
120.4, 119.6, 119.8, 120.0, 120.1, 119.6, 119.3, 119.8, 119.5, 119.6, 119.9, 119.1, 119.3, 119.8, 120.3, 119.1

.

a) Határozzuk meg az adatok átlagát, szórását, terjedelmét, mediánját! Rajzoljuk fel az adat hisztogramját a terjedelem 5 egyenlô részre osztásával!
b) Teljesül-e a várható értékre a $m_0 = 120$ g elôírás 95 %-os megbízhatósági szinten?
Adjuk meg a konfidencia-intervallumot, a próbastatisztika értékét és a kritikus értéket!
c) Hány elemû mintát kellene vennünk, hogy a konfidencia-intervallum hossza 5 %-os szignifikanciaszint mellett 0.2 g legyen?
d) Adjuk meg a szórás konfidencia-intervallumát 5 %-os szignifikanciaszint esetén!

Megoldás

```
> restart;
```

Ha feltételezzük, hogy a vizsgált mennyiség normál eloszlást követ, de sem a várható értéke, sem a szórása nem ismert, akkor a μ paraméter becslésére ún. (egymintás, kétoldali) t-próbát alkalmazunk. A nullhipotézisünk "$H_0: \mu = m_0$", ahol az elôre adott m_0 érték az elvárt/sejtett várható érték. A próba komponensei és menete megegyezik az u-próbánál leírtakkal, de kicsit más eloszlással és képletekkel kell számolni. A próbastatisztika itt t-eloszlású.

• Próbastatisztika: $t = \dfrac{\overline{X} - m_0}{\dfrac{S}{\sqrt{n}}}$, ahol n a mintaméret, $S = \dfrac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$ a minta (korrigált) tapasztalati szórása, m_0 pedig a feltételezett várható érték, amit elfogadni/cáfolni szeretnénk

• Próbastatisztika eloszlása: T_{n-1} (t-eloszlás $n-1$ szabadsági fokkal).

• Kritikus érték: $t_{\frac{\alpha}{2}} = F^{-1}_{T_{n-1}}\left(1 - \dfrac{\alpha}{2}\right)$ (az $n-1$ szabadsági fokú t-eloszlás $\left(1 - \dfrac{\alpha}{2}\right)$-kvantilise), ahol α a szignifikanciaszint, $F^{-1}_{T_{n-1}}$ pedig az $n-1$ szabadsági fokú t-eloszlás eloszlásfüggvényéne.

• Kritikus tartomány: $\left\{|t| > t_{\frac{\alpha}{2}}\right\}$, azaz a kritikus tartomány két félegyenes uniója, az origótól a kritikus értéknél távolabb lévô pontok halmaza; csak akkor utasítunk el, ha a próbastatisztika abszolútértéke nagyobb, mint a kritikus érték

• Elfogadási tartomány: $\left\{|t| \le t_{\frac{\alpha}{2}}\right\}$ (a kritikus tartomány komplementere)

• p-érték: $p_{érték} = 2 \cdot \left(1 - F_{T_{n-1}}(t)\right)$; ez alapján a kritikus tartomány ekvivalens megadása: $\left\{p_{érték} < \alpha\right\}$

A t-próbával egyenértékû a várható értékre vonatkozó, $1-\alpha$ szintû konfidencia-intervallum megadása. Ez lényegében azt jelenti, hogy a minta $\overline{X}$ átlaga körül megadunk egy szimmetrikus intervallumot, melybe az elméleti várható érték $1-\alpha$ valószínûséggel beleesik, feltéve, hogy a nullhipotézis ($\mu = m_0$) igaz. Ezután megnézzük, hogy m_0 beleesik-e ebbe az intervallumba és ez alapján döntünk – ha beesik, elfogadjuk H_0-t; ha nem, akkor pedig elutasítjuk.

• Konfidencia-intervallum: $I_K = \left[\overline{X} - t_{\frac{\alpha}{2}} \cdot \dfrac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2}} \cdot \dfrac{S}{\sqrt{n}}\right]$

• Konfidencia-intervallum hossza: $L = 2\, t_{\frac{\alpha}{2}} \cdot \dfrac{S}{\sqrt{n}}$ (ez nem függ a konkrét mintától, csak a mintaelemek számától)

• Elfogadás: ha $m_0 \in I_K$

A korrigált tapasztalati szórást a *Statistics* csomag *StandardDeviation* eljárásával, a kritikus értéket pedig a *Quantile* eljárással lehet számolni.

Ha t-próbát kell alkalmaznunk, akkor szükség lehet arra is, hogy konfidenciaintervallumot adjunk a háttéreloszlás σ szórására. Ekkor a megfelelô próbastatisztika χ^2-eloszlású lesz $n-1$ paraméterrel.

• Konfidencia-intervallum varianciára: $I_{K,\,Var} = \left[\dfrac{n-1}{c_2}S^2, \dfrac{n-1}{c_1}S^2\right]$, ahol c_1 ill. c_2 az $n-1$ szabadsági fokú χ^2-eloszlás $\dfrac{\alpha}{2}$ és $1 - \dfrac{\alpha}{2}$ kvantilisei

• Konfidencia-intervallum szórásra: $I_{K,\sigma} = \left[\sqrt{\frac{n-1}{c_2}}\, S, \sqrt{\frac{n-1}{c_1}}\, S \right]$

a) Határozzuk meg az adatok átlagát, szórását, terjedelmét, mediánját! Rajzoljuk fel az adat hisztogramját a terjedelem 5 egyenlô részre osztásával!

```
> m0 := 120;
  minta := [120.4, 119.6, 119.8, 120.0, 120.1, 119.6,
  119.3, 119.8, 119.5, 119.6, 119.9, 119.1, 119.3, 119.8,
  120.3, 119.1];
```

$$m0 := 120$$

$$minta := [120.4, 119.6, 119.8, 120.0, 120.1, 119.6, 119.3, 119.8, 119.5, 119.6, 119.9, 119.1, 119.3, 119.8, 120.3, 119.1] \quad \textbf{(1.2.1.1.1)}$$

```
> with(Statistics):
  atlag := Mean(minta);
  szoras := StandardDeviation(minta);
  terjedelem := Range(minta);
  medián := Median(minta);
```

$$atlag := 119.700000000000$$

$$szoras := 0.389871773792362$$

$$terjedelem := 1.30000000000001$$

$$medián := 119.700000000000 \quad \textbf{(1.2.1.1.2)}$$

b) Teljesül-e a várható értékre a $m_0 = 120$ g elôírás 95 %-os megbízhatósági szinten? Adjuk meg a konfidencia-intervallumot, a próbastatisztika értékét és a kritikus értéket!

```
> with(Statistics):
  alpha := 0.05;
  infolevel[Statistics] := 1:
  statisztika := OneSampleZTest([seq(atlag, i = 1..nops
  (minta))], m0, szoras, confidence = 1 - alpha);
```

$$\alpha := 0.05$$

```
Standard Z-Test on One Sample
-----------------------------
Null Hypothesis:
Sample drawn from population with mean 120 and known
standard deviation 0.389872
Alt. Hypothesis:
Sample drawn from population with mean not equal to
120 and known standard deviation 0.389872
Sample size:                16
Sample mean:                119.7
Distribution:               Normal(0,1)
Computed statistic:         -3.07794
Computed pvalue:            0.0020844
Confidence interval:        119.508966341195 ..
119.891033658805
                            (population mean)
Result: [Rejected]
```

There exists statistical evidence against the null hypothesis

statisztika := *hypothesis* = *false*, *confidenceinterval* = 119.508966341195 ..119.891033658805, *distribution* = *Normal*(0, 1), *pvalue* = 0.00208440335338232, *statistic* = −3.07793505625540 **(1.2.1.1.3)**

c) Hány elemû mintát kellene vennünk, hogy a konfidencia-intervallum hossza 5 %-os szignifikanciaszint mellett 0.2 g legyen?

d) Adjuk meg a szórás konfidencia-intervallumát 5 %-os szignifikanciaszint esetén!

Gy/2. Feladat (Véletlenszám-generátor)

Egy véletlenszám-generátorral a következô számokat kaptuk:
0.554, 0.110, 0.168, 0.440, 0.037, 0.600, 0.683, 0.612, 0.992, 0.025, 0.683, 0.758, 1.000, 0.845, 0.997, 0.631, 0.278, 0.350, 0.801, 0.543

a) Ábrázoljuk az értékeket hisztogramon, a terjedelem 10 részre osztása mellett!
b) Teszteljük 95 %-os megbízhatósági szint mellett, hogy a véletlenszám-generátor valóban egyenletes eloszlással sorsol-e számokat a [0, 1] intervallumból! (Használjuk a *Statistics* csomag *ChiSquareSuitableModelTest* eljárását!)
c) Adjuk meg a kritikus értéket, a próbastatisztika értékét és a *p*-értéket! Reprodukáljuk a döntést a *p*-érték alapján is!
d) Szemléltessük a kritikus tartományt és a döntést grafikonon a sûrûségfüggvény alatti terület besatírozásával!

Megoldás

```
> restart;
```

Gy/3. Feladat (Fémrudak tömege)

Egy gyártósoron 1 m-es fémrudak készülnek. Véletlenszerû kiválasztással lemérik 6 db rúd hosszát és súlyát. Az eredményt a következô táblázat tartalmazza:

X (cm)	10\	10\	98\	99	97	10\
Y (dkg)	60\	62\	58\	59\	57\	60\

a) Számoljuk ki a rudak hosszának terjedelmét, átlagát, szórását, mediánját és kvartiliseit! Vizualizáljuk az adathalmazt a BoxPlot eljárás használatával!
b) Számítsuk ki az *X* és *Y* közötti a korrelációs együttható értékét, és ha feltételezhetô lineáris kapcsolat, akkor határozzuk meg a regressziós egyenes egyenletét!
c) Ábrázoljuk az összetartozó értékpárokat a síkon a *Statistics* csomag *ScatterPlot* eljárásával, és illesszük rá a regressziós egyenest a kapott ábrára!
d) Számoljuk ki a rudak tömegét az illesztett lineáris modell alapján, majd számoljuk ezek eltéréseit a mért tömeg értékektôl. Gyôzôdjünk meg arról, hogy az eltérések átlaga

0! Határozzuk meg az eltérések szórását!
e) Rajzoljuk fel az eltérések (reziduálisok) hisztogrammját a terjedelem 4 egyenlô részre osztásával! Illesszünk rá az eltérések hisztogramjára egy azonos várható értékû, ill. szórású haranggörbét!
f) Teszteljük, hogy vajon normál eloszlást követnek-e az eltérések (nulla várható értékkel és az eltérések szórásával) 95 %-os konfidenciaszinten, ha a terjedelmet 4 egyenlô részre osztjuk!
g) Rajzoljuk fel az illesztett és a mért adatsorokat a *QuantilePlot* eljárás segítségével!
h) Számoljuk ki az *RSS* (residual sum of squares) és a *TSS* (total sum of squares) értékeket, majd ezekbôl az R^2 determinációs együttható értékét! Mennyire jó ez alapján az illesztett lineáris modell?

Megoldás

```
> restart;
```