

▼ 17. Lineáris regressziós egyenes illesztése adatokra és statisztikai tesztelése.

Valószínűségszámítás és statisztika számítógép algebrai támogatással	Készült a PTE PMMIK-n a TÁMOP -4.1.2.A/1-11/1-2011-0098 támogatásával
Klincsik Mihály	

▼ 17.1 Bevezetés

Tegyük fel, hogy az $X = (X_1, X_2, \dots, X_n)$ és $Y = (Y_1, Y_2, \dots, Y_n)$ adatsorok pontjait felajzolva az (X, Y) síkon egy olyan pontfelhőt látunk, amely egy

$$Y = a + b \cdot X$$

alakú lineáris egyenes körül helyezkedik el. A legkisebb négyzetes elv alapján a

$$\sum_{k=1}^n (Y_k - (a + b \cdot X_k))^2$$

eltérések négyzetösszegét minimalizáljuk az "a" és "b" paraméterek meghatározása érdekében. A gyakorlati feladatok során általában teljesül az a feltétel, hogy egyértelműen meghatározott (a, b) paraméterpárt kapunk az adatokból. Az illesztéstől elvárt, hogy a

$$K = Y - a - b \cdot X$$

eltérés normál eloszlású változó legyen 0 várható értékkel $K \sim N(0, S)$. Ezt χ^2 – módszer segítségével tudjuk ellenőrizni.

Az illesztés erősségét az r korrelációs együtthatóval mérhetjük és a modell megfelelőségét a Q-Q plot segítségével ellenőrizhetjük. Igazoljuk, hogy a korrelációs együttható mindig $[-1, 1]$ intervallumba esik. Becslést adunk Student-T-eloszlás segítségével a paraméterekre. Valamint felrajzoljuk az egyenes 95%-os konfidencia intervallumát!

Célunk az alábbi fogalmak, összefüggések és eljárások megismerése, megértése	
Lineáris regressziós modell alkalmazhatósági kritériumai	A <i>Statistics</i> csomag <i>LinearFit</i> eljárásának hívása
Legkisebb négyzetes elv	A <i>Statistics</i> csomag <i>Correlation</i> eljárásának használata
Regressziós egyenes együtthatóinak számítása	A <i>Statistics</i> csomag <i>ScatterPlot</i> eljárásának használata
Korrelációs együttható számítási képlete és értelmezése	A <i>Statistics</i> csomag <i>ChiSquareSuitableModelTest</i> eljárásának használata
A korrelációs együttható mindig -1 és 1 közé esik.	

▼ 17.2. Egyváltozós lineáris regressziós modell

Lineáris Regressziós Modell

Tegyük fel, hogy az Y véletlen változó lineáris függvénye az x változónak. A lineáris kapcsolat alatt azt értjük, hogy az Y változó $E(Y)$ várható értéke az x független változó lineáris függvénye

$$E(Y) = \alpha + \beta x$$

A két konstans értékét, vagyis az α tengelymetszetet illetve a β meredekség értékét az Y változó egy n elemű $Y_1, Y_2, \dots, Y_n$ mintája alapján határozzuk meg, amelyek megfelelnek az $x_1, x_2, \dots, x_n$ értékeknek. Minden kísérlet során x felvesz valamilyen x_k értéket, melyhez tartozó Y_k valószínűségi változó várható értékére

$$E(Y_k) = \alpha + \beta x_k \quad (k = 1, 2, \dots, n)$$

teljesül.

Természetesen az Y véletlen változó maga is (nemcsak a várható értéke) függvénye x-nek! Defináljuk a K **különbség** valószínűségi változót a

$$K = Y - (\alpha + \beta x)$$

egyenlőséggel. Rendezés után kapjuk az

$$(*) \quad Y = \alpha + \beta x + K$$

relációt. Ebben feltesszük, hogy K várható értéke 0, azaz $E(K) = 0$ és $Var(K) = \sigma^2$ varianciája megegyezik Y varianciájával. A σ^2 értékét nem ismerjük ugyan, de feltesszük, hogy olyan konstans, amely nem függ x-től! A K véletlen változó értékét úgy is tekinthetjük, mint az x és Y lineáris kapcsolatához adódó **zaj**, vagy **mérési pontatlanság**.

A most megadott kapcsolatot **lineáris regressziós modell**nek nevezzük, ahol α és β a **regressziós együtthatók**.

A mintavétel ebben a modellben az

$$(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$$

rendezett párokból áll, ahol az x_k értékek megfelelnek x különböző értékeinek és az Y_k -k pedig független megfigyelések során kapott minta elemek, amelyek az Y populációból valók. Az (x_k, Y_k) $k=1,2,\dots,n$ párok tekinthetők úgy, mint egy **minta** az Y véletlen változóra, adott $x_1, x_2, \dots, x_n$ értékek mellett.

Ilyen módon az $x_1, x_2, \dots, x_n$ értékeknek nem kell feltétlenül különbözőknek lenni. Azonban ahhoz, hogy az α és β együtthatók egyértelműen meghatározhatók legyenek, ahhoz köztük legalább két különböző x értéknek elő kell fordulni.

▼ 17.3. Becslés legkisebb négyzetes elv alapján

A LEGKISEBB NÉGYZETES ELV

Az α és β regressziós paraméterek **becslésére** a **legkisebb négyzetes elv** olyan optimális α_{opt} és β_{opt} paramétereket javasol, melyekre **minimális** lesz a megfigyelt y_k értékek és a megfelelő $(\alpha_{opt} + \beta_{opt} x_k) = E(Y_k)$ várható értékek eltéréseinek **négyzetösszege**.

Jelölje tetszőleges α és β az **eltéréseket** vagy **maradékokat**

$$e_k = y_k - (\alpha + \beta \cdot x_k) \quad (k = 1, 2, \dots, n)$$

melyeket **reziduálisoknak** is nevezünk.

Így a reziduálisok négyzetösszege a

$$Q(\alpha, \beta) = \sum_{k=1}^n e_k^2 = \sum_{k=1}^n (y_k - (\alpha + \beta x_k))^2 \rightarrow \text{minimális}$$

legkisebb értékét felveszi az $\alpha = \alpha_{opt}$ és $\beta = \beta_{opt}$ értéknél.

Tehát keressük a $Q(\alpha, \beta)$ kétváltozós függvény minimumát az α és β változókra!

▼ 17.4. A regressziós egyenes együtthatóinak számítása

Nevezzük át az α és β paramétereket 'a'-ra illetve 'b'-re!

Felírjuk a $Q(a, b)$ eltérések négyzetösszegét, mint kétváltozós függvényt Maple-ben!

> *restart*; $Q := unapply\left(\sum_{k=1}^n (y_k - a - b x_k)^2, a, b\right) : 'Q'(a, b) = Q(a, b)$

$$Q(a, b) = \sum_{k=1}^n (y_k - a - b x_k)^2 \quad (1.4.1)$$

A $Q(a, b)$ kétváltozós függvény minimumát olyan 'a' és 'b' értékre veszi fel, ahol a parciális deriváltak eltűnnek. Deriváljuk 'a' és 'b' szerint parciálisan a $Q(a, b)$ függvényt!

> $da := \frac{\partial}{\partial a} Q(a, b) = \frac{\partial}{\partial a} Q(a, b); db := \frac{\partial}{\partial b} Q(a, b) = \frac{\partial}{\partial b} Q(a, b)$

$$da := \frac{\partial}{\partial a} \left(\sum_{k=1}^n (y_k - a - b x_k)^2 \right) = \sum_{k=1}^n (-2 y_k + 2 a + 2 b x_k)$$

$$db := \frac{\partial}{\partial b} \left(\sum_{k=1}^n (y_k - a - b x_k)^2 \right) = \sum_{k=1}^n (-2 (y_k - a - b x_k) x_k) \quad (1.4.2)$$

Írjuk fel a **szélsőérték létezésének** $\frac{\partial}{\partial a} Q(a, b) = 0$ és $\frac{\partial}{\partial b} Q(a, b) = 0$ szükséges feltételét!

> $rendszer := \left[\text{simplify}\left(\text{rhs}\left(\frac{da}{2}\right)\right) = 0, \text{simplify}\left(\text{rhs}\left(\frac{db}{2}\right)\right) = 0 \right];$

$$rendszer := \left[\sum_{k=1}^n (-y_k + a + b x_k) = 0, \sum_{k=1}^n (-y_k + a + b x_k) x_k = 0 \right] \quad (1.4.3)$$

Vezessük itt be az $S_x = \sum_{k=1}^n x_k = n x_{\text{átlag}}$ és $S_y = \sum_{k=1}^n y_k = n y_{\text{átlag}}$ jelöléseket, ahol értelemszerűen $x_{\text{átlag}}$ az x értékek átlaga és $y_{\text{átlag}}$ az y értékek átlaga! Az egyenletrendszert felírhatjuk az alábbi egyszerűbb alakban a rendezések után.

> $E_1 := n a + b S_x = S_y;$

> $E_2 := S_x a + b S_{xx} = S_{xy}$

$$E_1 := n a + b S_x = S_y$$

$$E_2 := S_x a + b S_{xx} = S_{xy} \quad (1.4.4)$$

ahol $S_{xx} = \sum_{k=1}^n x_k^2$ és $S_{xy} = \sum_{k=1}^n x_k y_k$.

Oldjuk meg **solve** eljárással az E_1 és E_2 egyenletrendszert az "a" és "b" ismeretlenekre!

> $egyutthatok := \text{solve}(\{E_1, E_2\}, \{a, b\});$

$$\text{egyutthatok} := \left\{ a = - \frac{S_x S_{xy} - S_y S_{xx}}{n S_{xx} - S_x^2}, b = \frac{n S_{xy} - S_x S_y}{n S_{xx} - S_x^2} \right\} \quad (1.4.5)$$

Bővebben kifejtve ez a következő!

> *solve(rendszer, {a, b});*

$$\left\{ a = - \frac{\left(\sum_{k=1}^n x_k y_k \right) \left(\sum_{k=1}^n x_k \right) - \left(\sum_{k=1}^n x_k^2 \right) \left(\sum_{k=1}^n y_k \right)}{- \left(\sum_{k=1}^n x_k \right)^2 + \left(\sum_{k=1}^n 1 \right) \left(\sum_{k=1}^n x_k^2 \right)}, b = \frac{- \left(\sum_{k=1}^n x_k \right) \left(\sum_{k=1}^n y_k \right) + \left(\sum_{k=1}^n 1 \right) \left(\sum_{k=1}^n x_k y_k \right)}{- \left(\sum_{k=1}^n x_k \right)^2 + \left(\sum_{k=1}^n 1 \right) \left(\sum_{k=1}^n x_k^2 \right)} \right\} \quad (1.4.6)$$

Alakítsuk át a 'b' meredekség képletét az átlagok bevezetésével

$$\begin{aligned} > b = \text{simplify} \left(\text{eval} \left(\text{eval}(b, \text{egyutthatok}), \left\{ S_x = n \cdot x_{\text{átlag}}, S_y = n \cdot y_{\text{átlag}}, S_{xx} = \sum_{k=1}^n x_k^2, S_{xy} = \sum_{k=1}^n x_k y_k \right\} \right) \right); \end{aligned}$$

$$b = \frac{\sum_{k=1}^n x_k y_k - n x_{\text{átlag}} y_{\text{átlag}}}{\sum_{k=1}^n x_k^2 - n x_{\text{átlag}}^2} \quad (1.4.7)$$

Tanulságos levezetni, hogy a fenti képlet $b = \frac{\sum_{k=1}^n (x_k - x_{\text{átlag}}) (y_k - y_{\text{átlag}})}{\sum_{k=1}^n (x_k - x_{\text{átlag}})^2}$ alakban is írható.

Először a számlálót alakítsuk át az (1.4.7) képlet számlálójára

$$\begin{aligned} \sum_{k=1}^n (x_k - x_{\text{átlag}}) (y_k - y_{\text{átlag}}) &= \sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k y_{\text{átlag}} - \sum_{k=1}^n x_{\text{átlag}} y_k + \sum_{k=1}^n x_{\text{átlag}} y_{\text{átlag}} = \\ &= \sum_{k=1}^n x_k y_k - y_{\text{átlag}} \cdot \sum_{k=1}^n x_k - x_{\text{átlag}} \cdot \sum_{k=1}^n y_k + n \cdot x_{\text{átlag}} y_{\text{átlag}} = \\ &= \sum_{k=1}^n x_k y_k - y_{\text{átlag}} \cdot n \cdot x_{\text{átlag}} - x_{\text{átlag}} \cdot n \cdot y_{\text{átlag}} + n \cdot x_{\text{átlag}} y_{\text{átlag}} = \sum_{k=1}^n x_k y_k - n \cdot x_{\text{átlag}} y_{\text{átlag}} \end{aligned}$$

Ezzel megkaptuk az (1.4.7) képlet számlálóját.

Most a b paraméter nevezőjét alakítsuk át az (1.4.7) képlet nevezőjére

$$\begin{aligned} \sum_{k=1}^n (x_k - \bar{x})^2 &= \sum_{k=1}^n (x_k^2 - 2 \cdot x_k \cdot \bar{x} + \bar{x}^2) = \sum_{k=1}^n x_k^2 - 2 \cdot \sum_{k=1}^n x_k \cdot \bar{x} + \sum_{k=1}^n \bar{x}^2 = \\ &= \sum_{k=1}^n x_k^2 - 2 \cdot \bar{x} \cdot \sum_{k=1}^n x_k + n \cdot \bar{x}^2 = \sum_{k=1}^n x_k^2 - 2 \cdot \bar{x} \cdot n \cdot \bar{x} + n \cdot \bar{x}^2 = \sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \end{aligned}$$

Tehát

$$\rightarrow B := b = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2};$$

$$B := b = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2} \quad (1.4.8)$$

17.4.1. DEFINÍCIÓ. Két statisztikai adatsor kovarianciája

Az $X = (x_1, x_2, \dots, x_n)$ és $Y = (y_1, y_2, \dots, y_n)$ adatok **kovarianciája** alatt a

$$\text{Cov}(X, Y) = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{n}$$

kifejezést értjük.

Ezzel a jelöléssel az (1.4.8) képlet számlálója $n \text{ Cov}(X, Y)$ és nevezője $(n - 1) \text{ Var}(X)$, ahol

$$\text{Var}(X) = \frac{1}{n - 1} \cdot \sum_{k=1}^n (x_k - \bar{x})^2$$

az X adatsor varianciája. Így a meredekség tömör alakban

$$b = \frac{n}{n - 1} \cdot \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

17.4.2. MEGJEGYZÉSEK

(i) A kovariancia képletében a nevező azért **nem** $(n - 1)$, hanem n , mert megmutatható, hogy így lesz a

$$\text{Cov}(X, Y) = E((X - E(X)) \cdot (Y - E(Y)))$$

valószínűségi változó **torzítatlan becslése**.

(ii) Statisztikai számításokra nem teljesül a valószínűségi változókra megszokott $\text{Cov}(X, X) = \text{Var}(X)$ képlet, hanem helyette

$$\text{Var}(X) = \frac{n}{n - 1} \cdot \text{Cov}(X, X).$$

Természetesen az $\frac{n}{n - 1}$ szorzó értéke közel van 1-hez, ha n értéke "nagy".

(iii) Könnyen igazolható, hogy ha

$$Y = m \cdot X \quad (m \neq 0),$$

vagyis az Y adatsor az X adatsor m -szerese, akkor a meredekség $b = m$. Ugyanis

$$b = \frac{\sum_{k=1}^n (x_k - x_{\text{átlag}}) (m \cdot x_k - m \cdot x_{\text{átlag}})}{\sum_{k=1}^n (x_k - x_{\text{átlag}})^2} = m \cdot \frac{\sum_{k=1}^n (x_k - x_{\text{átlag}})^2}{\sum_{k=1}^n (x_k - x_{\text{átlag}})^2} = m$$

(iv) A b meredekség nincs értelmezve, ha a nevező nulla, azaz

$$\sum_{k=1}^n (x_k - x_{\text{átlag}})^2 = 0.$$

Ez viszont csak akkor nulla, ha $x_1 = x_2 = \dots = x_n = x_{\text{átlag}}$, amely azt jelenti, hogy minden mérési adat egy x értékre vonatkozik. A gyakorlati problémáknál ez az eset szinte nem fordul elő.

Az 'a' paramétert már 'b' paraméter segítségével adhatjuk meg

$$a = y_{\text{átlag}} - b \cdot x_{\text{átlag}}$$

formában.

Igazoljuk ezt Maple egyszerűsítés segítségével. Írjuk be az $y_{\text{átlag}} = \frac{S_y}{n}$, $x_{\text{átlag}} = \frac{S_x}{n}$ képleteket és 'b' értékét az (1.4.5) képletből.

$$> \frac{S_y}{n} - \left(\frac{n \cdot S_{xy} - S_x \cdot S_y}{n S_{xx} - S_x^2} \right) \cdot \frac{S_x}{n}; a = \text{simplify}(\%)$$

Ezzel megkaptuk az 'a' paraméter értékét. (lásd az (1.4.5) képletet)

Ez azt is jelenti, hogy a regressziós egyenes átmegy az $(x_{\text{átlag}}; y_{\text{átlag}})$ ponton!

A $Q(a, b)$ függvény második parciális deriváltjainak előjel tesztje azt mutatja, hogy a kapott helyen Q függvénynek valóban minimuma van!

> $\text{Hesse} := \text{simplify}(\text{VectorCalculus:-Hessian}(Q(a, b), [a, b]));$
 $\text{determináns} := \text{LinearAlgebra:-Determinant}(\text{Hesse});$

A másodrendű parciális deriváltakból képezett Hesse-mátrix determinánsa nem-negatív, mert

$$(X, 1)^2 \leq \|X\|^2 \cdot \|1\|^2$$

ahol $X = (x_1, x_2, \dots, x_n)$ és $1 = (1, 1, \dots, 1)$ n -dimenziós vektorok skaláris szorzata $(X, 1) = \sum_{k=1}^n x_k$ és

$$\text{normáik négyzete } \|X\|^2 = \sum_{k=1}^n x_k^2 \text{ illetve } \|1\|^2 = \sum_{k=1}^n 1 = n$$

Igazoljuk ezt az

$$\|X + \lambda \cdot 1\|^2 = \|X\|^2 + 2 \cdot \lambda \cdot (X, 1) + \lambda^2 \cdot n \geq 0$$

nem-negatív értékű λ valós másodfokú polinomja vizsgálatával. A λ -ban másodfokú kifejezésnek akkor nincs valós két különböző zérushelye, ha a diszkriminánsa nem-pozitív

$$\text{diszkrimináns} = 2^2 \cdot (X, 1)^2 - 4 \cdot n \cdot \|X\|^2 \leq 0.$$

Ahonnakapjuk a bizonyítandó egyenlőtlenséget

$$(X, 1)^2 = \left(\sum_{k=1}^n x_k \right)^2 \leq n \cdot \|X\|^2 = n \cdot \left(\sum_{k=1}^n x_k^2 \right)$$

és az egyenlőség itt csak akkor áll fenn, ha $X = -\lambda \cdot 1 = (-\lambda, -\lambda, \dots, -\lambda)$ konstans x -mérési adatok szerepelnek.

Ennélfogva a Hesse-mátrix determinánsa pozitív, ezért $Q(a, b)$ függvénynek szélsőértéke van és

mivel a mátrix bal felső sarkában szereplő $2 \cdot n > 0$, ezért a szélsőérték minimum.
Rögzítsük a kapott eredményt tétel formában!

14.7.3. TÉTEL. Regressziós egyenes paramétereinek számítása

Az $X = (x_1, x_2, \dots, x_n)$ és $Y = (y_1, y_2, \dots, y_n)$ adatokra illeszthető lineáris

$$y = a + b \cdot x$$

alakú függvény paraméterei a legkisebb négyzetes elv alapján

$$a = y_{\text{átlag}} - b \cdot x_{\text{átlag}}, \quad b = \frac{\sum_{k=1}^n (x_k - x_{\text{átlag}}) (y_k - y_{\text{átlag}})}{\sum_{k=1}^n (x_k - x_{\text{átlag}})^2}$$

számítjuk, ahol $x_{\text{átlag}}$ és $y_{\text{átlag}}$ az X és Y adatok átlaga. Átrendezve a lineáris egyenletet

$$y - y_{\text{átlag}} = \frac{\sum_{k=1}^n (x_k - x_{\text{átlag}}) (y_k - y_{\text{átlag}})}{\sum_{k=1}^n (x_k - x_{\text{átlag}})^2} \cdot (x - x_{\text{átlag}})$$

alakban írhatjuk és ez értelmezhető $x_1 = x_2 = \dots = x_n$ eset kivételével.

▼ 17.5. Anyag koncentráció függése a hőmérséklettől. Példa.

17.5.1. PÉLDA

Egy kémiai folyamatban a keletkező anyag **koncentrációja** (amelyek az Y értékek) %-ban mérve közel lineáris mértékben növekszik a **hőmérséklet** (melyek az X értékek °C-ban mérve) növekedésével!

Hőmérsékletek (°C) = [45, 50, 55, 60, 65, 70, 75, 80, 85, 90],

Anyag mennyiségek (%) = [43, 45, 48, 51, 55, 57, 59, 63, 66, 68]

(a) A fenti 10 mérési adatsorból határozzuk meg $E(Y)$ -ra az $a + b \cdot x$ lineáris regressziós egyenes a és b együtthatóit!

(b) Ábrázoljuk a mérési pontokat, az egyenest és a reziduális eltéréseket azonos koordináta-rendszerben!

(c) Rajzoljuk fel az e_k **eltérések gyakoriság hisztogramját**!

(d) Mutassuk meg, hogy az eltérések 95%-os megbízhatósággal 0 várható értékű **normál eloszlást** adnak!

(e) Az ú.n. Q-Q rajz segítségével vizsgáljuk a lineáris modell megbízhatóságát!

Megoldás

Felvesszük a két adatsort az X és Y vektorokban!

> *restart : with(Statistics) :*

> $X, Y := [45, 50, 55, 60, 65, 70, 75, 80, 85, 90], [43, 45, 48, 51, 55, 57, 59, 63, 66, 68] :$

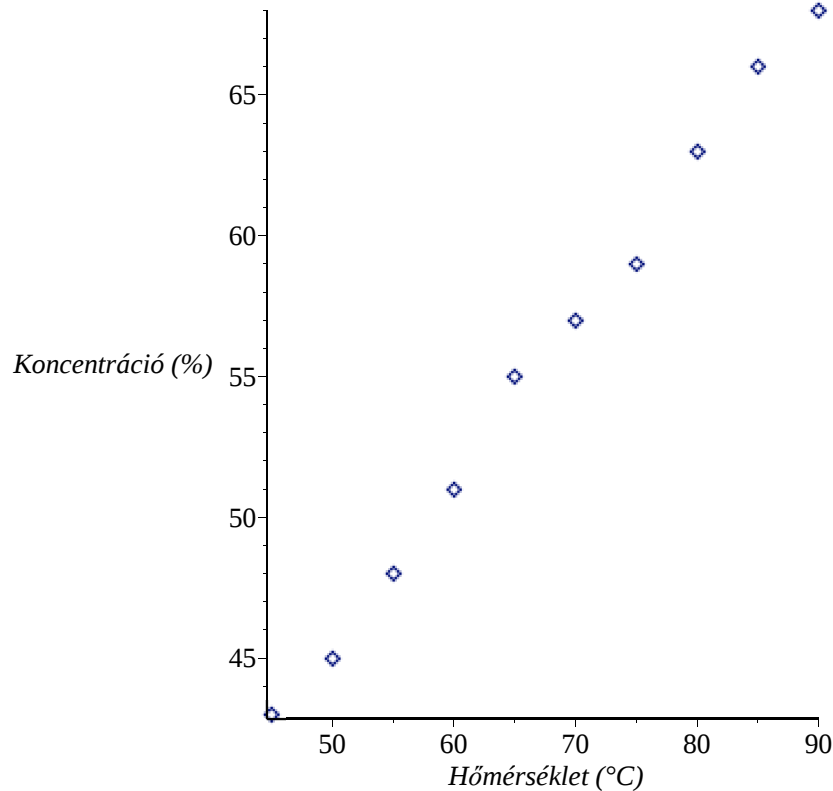
Rajzoljuk fel a pontokat a Statistics csomag **ScatterPlot** eljárását használva!

> *rajz1 := ScatterPlot(X, Y, symbolsize = 14, title*

= A hőmérséklet és a keletkezett anyag koncentrációjának kapcsolata, labels

= [Hőmérséklet (°C), Koncentráció (%)]) : rajz1;

A hőmérséklet és a keletkezett anyag koncentrációjának kapcsolata



(a) Számoljuk ki a **Statistics** csomag Count függvényével az adatszámot, *Covariance* eljárásával a kovarianciát, *Var(X)* eljárásával a *X* varianciáját és ez alapján az $\alpha_{opt} + \beta_{opt} \cdot x$ képletben a

$$\beta_{opt} = \frac{n \cdot \text{cov}(X, Y)}{(n - 1) \cdot \text{Var}(X)}$$

meredekséget!

A β_{opt} meredekség paraméter értékének számítása

```
> n := Count(X);
> cov := Covariance(X, Y);
> VarX := Variance(X);
>  $\beta := \frac{n \cdot cov}{(n - 1) \cdot VarX}$ 
```

$n := 10$

$cov := 118.250000000000$

$VarX := 229.166666666667$

$\beta := 0.573333333333333$

(1.5.1)

Az α_{opt} tengely metszet paraméter értékének számításához szükséges X és Y átlaga, mert az

$y = \alpha_{opt} + \beta_{opt} \cdot x$ regressziós egyenes átmegy az $(x_{\text{átlag}}, y_{\text{átlag}})$ ponton. Ebből a feltételből kifejezhető az α_{opt} paraméter!

> $\text{átlag}X, \text{átlag}Y := \text{Mean}(X), \text{Mean}(Y);$

> $\alpha := \text{átlag}Y - \beta \cdot \text{átlag}X$

$\text{átlag}X, \text{átlag}Y := 67.50000000000000, 55.50000000000000$

$\alpha := 16.80000000000000$

(1.5.2)

Így a regressziós egyenes

> $\text{linearis} := x \rightarrow \alpha + \beta \cdot x; \text{Regressziós egyenes} = \text{linearis}(x);$

$\text{linearis} := x \rightarrow \alpha + \beta x$

$\text{Regressziós egyenes} = 16.80000000000000 + 0.57333333333333 x$

(1.5.3)

A *Statistics* csomagban levő **LinearFit** egy lépésben kiszámolja a lineáris illesztés együtthatóit, melyek ugyanazok lesznek, mint amit mi is meghatároztunk!

> $\text{LinearFit}([1, x], X, Y, x);$

$16.80000000000000 + 0.57333333333333 x$

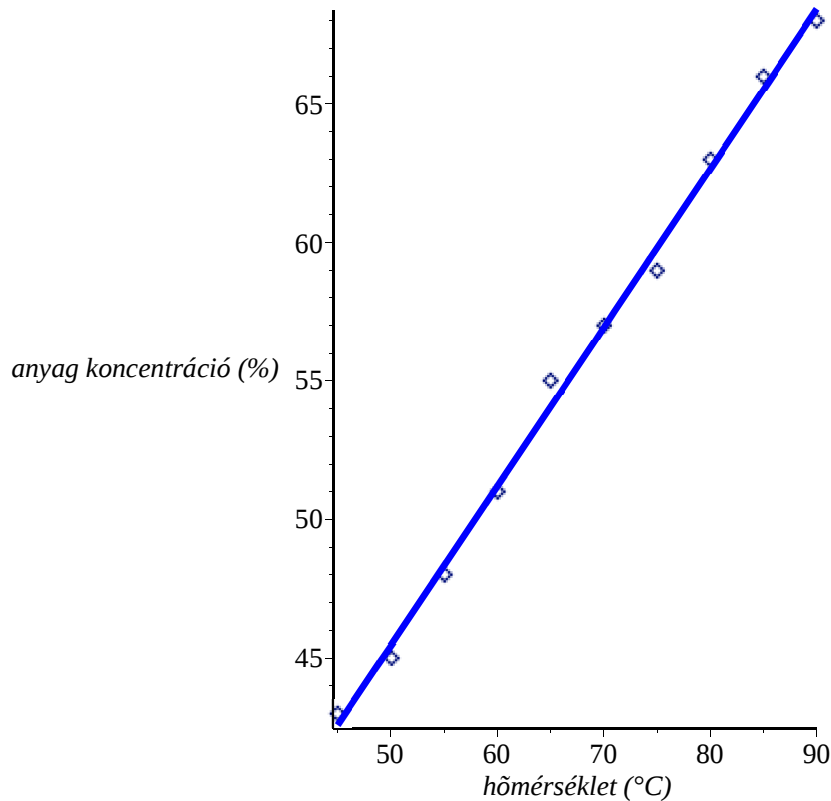
(1.5.4)

(b) Rajzoljuk be az egyenest a pontokhoz!

> $\text{rajz2} := \text{plot}(\text{linearis}(x), x = X[1]..X[n], \text{color} = \text{blue}, \text{thickness} = 3) :$

$\text{plots[display]}([\text{rajz1}, \text{rajz2}], \text{labels} = [\text{hőmérséklet } (^{\circ}\text{C}), \text{anyag koncentráció } (\%)],$
 $\text{title} = \text{Lineáris kapcsolat a hőmérséklet és az anyag koncentrációja között});$

Lineáris kapcsolat a hőmérséklet és az anyag koncentrációja között

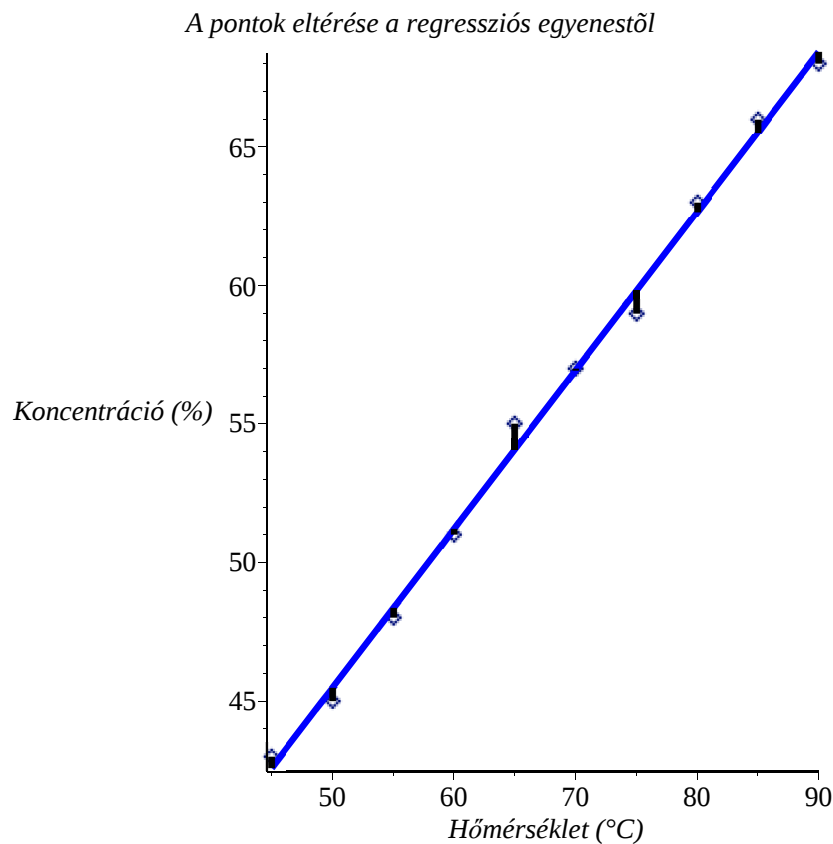


A rajzoláshoz képezzük a mérési pontok és az egyenesen levő pontok koordinátáinak listáját!

```
> elteresek_szakas := [seq([X[k], Y[k]], [X[k], linearis(X[k])], k = 1..n)];
elteresek_szakas := [[45, 43], [45, 42.6000000000000], [50, 45], [50,
45.4666666666667], [55, 48], [55, 48.3333333333333], [60, 51], [60,
51.2000000000000], [65, 55], [65, 54.0666666666667], [70, 57], [70,
56.9333333333333], [75, 59], [75, 59.8000000000000], [80, 63], [80,
62.6666666666667], [85, 66], [85, 65.5333333333333], [90, 68], [90,
68.4000000000000]]] (1.5.5)
```

Tegyük az ábrához az eltérések szakaszait!

```
> rajz3 := plot(elteresek_szakas, color = black, thickness = 3) :
plots[display]([rajz1, rajz2, rajz3], title = A pontok eltérése a regressziós egyenestől);
```



(c) Számoljuk ki a lineáris modellből az Y értékeket és vegyük ennek eltérését a mért Y értékektől!

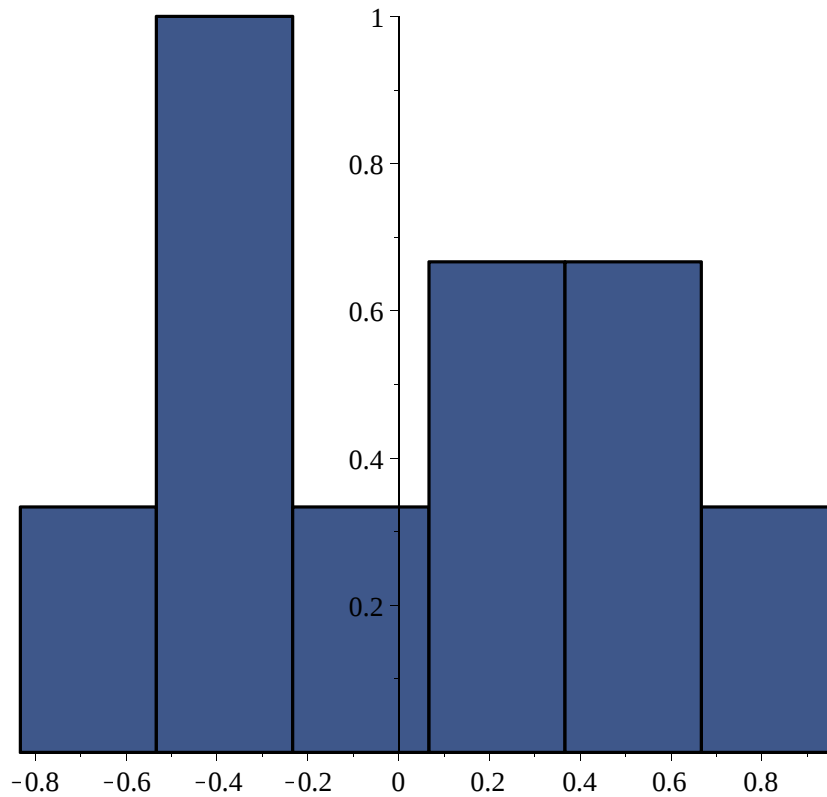
```
> modellY := [seq(evalf(linearis(X[k]), 9), k = 1..n)]
modellY := [42.6000000, 45.4666667, 48.3333333, 51.2000000, 54.0666667,
56.9333333, 59.8000000, 62.6666667, 65.5333333, 68.4000000] (1.5.6)
```

```
> eltérések := Y – modellY
eltérések := [0.4000000, -0.4666667, -0.3333333, -0.2000000, 0.9333333, 0.0666667,
-0.8000000, 0.3333333, 0.4666667, -0.4000000] (1.5.7)
```

Az eltérések hisztogramja normál eloszlásra emlékeztet!

```
> H := Histogram(eltérések, binwidth = 0.3, title
= "Az egyenestől vett eltérések gyakoriság hisztogramja") : H;
Histogram Type: default
Data Range: -.83333335 .. .96666665
Bin Width: .3
Number of Bins: 6
Frequency Scale: relative
```

Az egyenestől vett eltérések gyakoriság hisztogramja



Az eltérések várható értéke gyakorlatilag nulla! Számoljuk ki az eltérések szórását!

> *Az eltérések várható értéke = Mean(eltérések); S := StandardDeviation(eltérések);*

Az eltérések várható értéke = $-5.55111512312578 \cdot 10^{-18}$

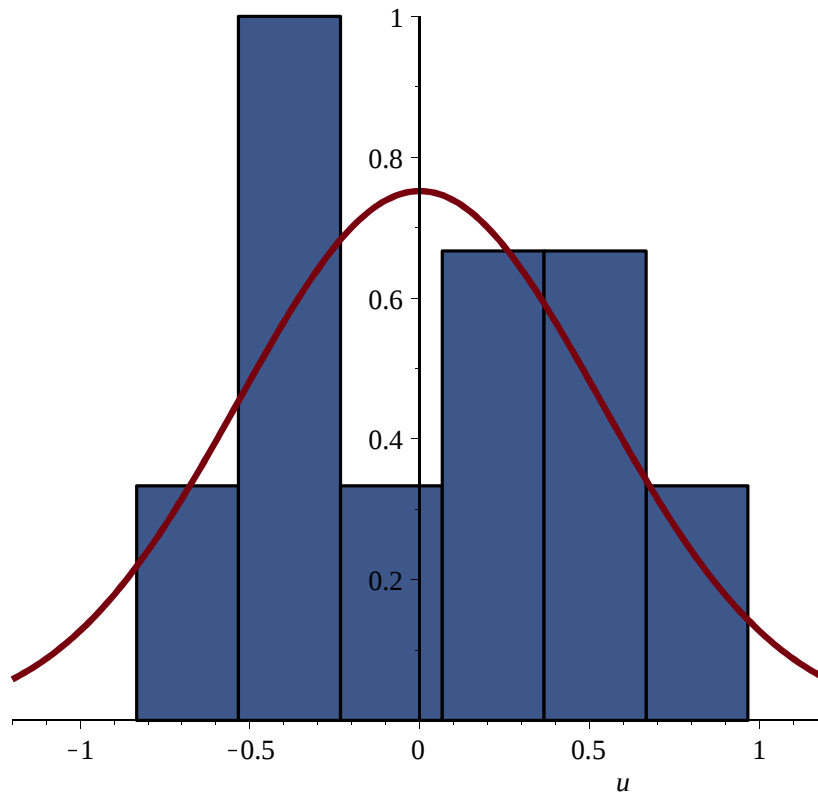
S := 0.530548279647609

(1.5.8)

Rajzoljuk az eltérések hisztogramjához az $N(0, S)$ normál eloszlás sűrűség függvényét!

> *suruseg := plot(PDF(Normal(0, S), u), u = -1.2 .. 1.2, thickness = 3) :*
plots[display]([H, suruseg]);

Az egyenestől vett eltérések gyakoriság hisztogramja



(d) Teszteljük azt a H_0 hipotézist, hogy az eltérések normál eloszlásúak, melynek 0 a várható értéke és szórása megegyezik az eltérésekből számolt szórással!

```
> inforevel[Statistics] := 1 : ChiSquareSuitableModelTest(eltérések, Normal(0, S), bins
= 10);
```

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: 4.70827
Computed pvalue: 0.858961
Critical value: 16.9189774487099
```

Result: [Accepted]

There is no statistical evidence against the null hypothesis

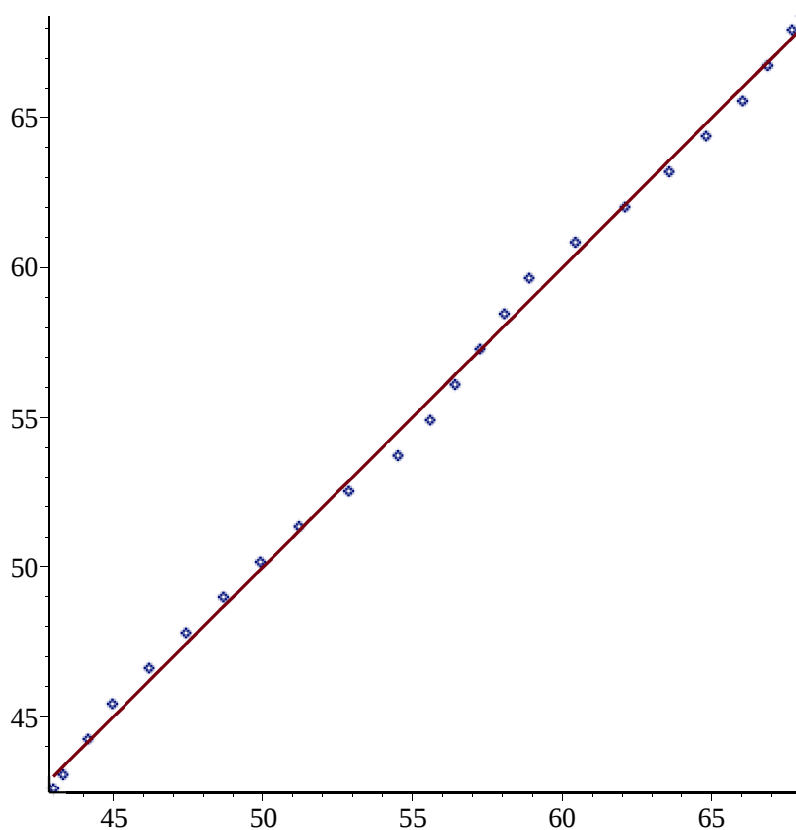
hypothesis = true, criticalvalue = 16.9189774487099, distribution = ChiSquare(9), pvalue (1.5.9)

= 0.858961484967997, statistic = 4.70827385401723

Tehát a reziduálisok normál eloszlást követnek. Ezért a lineáris modell alkalmazható. Hogy mennyire jó a lineáris illesztés, arról a Q-Q rajz ad szemléletes képet.

(e) A Q betű a Quantile =Kvantilis szó kezdőbetűje és a növekvőleg rendezett adatsor gyakoriságainak egyenlő részekre osztását jelenti. A Q-Q plot az Y_i mérési adatok ábrázolása a lineáris modellből számolt adatok ellenében. A modell akkor megfelelő, ha a pontok a berajzolt szögfelező egyenes közelében helyezkednek el.

> *QuantilePlot(Y, modelly, scaling = constrained);*



Mivel a pontok a szögfelező egyenes közelében helyezkednek el, ezért a példában a lineáris modell megfelelő.

▼ 17.6. Korreláció az X és Y értékek között. Az R^2 együtttható kétféle számítása.

Számoljuk ki a lineáris kapcsolat szorosságát mutató számot a **korrelációs együttthatót**! Milyen következtetést vonhatunk le még a korrelációs együtttható értékéből?

17.6.1. DEFINÍCIÓ. KORRELÁCIÓS EGYÜTTHATÓ

Az X és Y adatsorok kovarianciáját osztva a szórásainak szorzatával és szorozva $\frac{n}{n-1}$ – el egy dimenzió nélküli számot kapunk, amelyet az adatsorok **korrelációjának** nevezünk

$$r_{XY} = \frac{n}{n-1} \cdot \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

A korreláció méri az X és Y között meglévő lineáris kapcsolat erősségét és irányát. Ha r értéke közel van 0-hoz, akkor a kapcsolat gyenge, míg +1 vagy (-1)-hez közeli r érték szoros (szignifikáns) lineáris kapcsolatot jelez. Ha $r_{XY} > 0$, akkor X növekedése esetén Y értéke is növekszik, míg ha $r_{XY} < 0$, akkor X növekedésével Y értéke csökken.

17.6.2. PÉLDA. A 17.5.1. példa folytatása

Számoljuk ki a 17.5.1. példában szereplő hőmérséklet és anyag koncentráció adatsorok kapcsolatának szorosságát a Maple Statistics csomag *Correlation* eljárásával!

Megoldás

> $r := \text{Correlation}(X, Y);$

$$r := 0.998136887967885 \quad (1.6.1)$$

A korrelációs együttható értéke alapján kétféle következtetést vonhatunk le.

- (i) Mivel $r > 0$ pozitív, ezért az X hőmérséklet növekedésével az Y koncentráció is növekszik
- (ii) Mivel $r \approx 0.998$ nagyon közel van 1-hez, ezért a lineáris kapcsolat **szignifikáns** vagyis szoros.

Az együttható számítását az alábbi képlet alapján is megkaphatjuk.

$$> \frac{n}{n-1} \cdot \frac{\text{Covariance}(X, Y)}{\text{StandardDeviation}(X) \cdot \text{StandardDeviation}(Y)};$$

$$0.998136887967885 \quad (1.6.2)$$

Ellenőrizzük, hogy a korrelációs együttható számítását az adatokból közvetlenül az

$$r = \frac{\sum_{k=1}^n (x_k - \bar{x}) \cdot (y_k - \bar{y})}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2} \cdot \sqrt{\sum_{k=1}^n (y_k - \bar{y})^2}}$$

képlet adja meg!

$$> \frac{\sum_{k=1}^n (X_k - \text{átlag}X) \cdot (Y_k - \text{átlag}Y)}{\sqrt{\sum_{k=1}^n (X_k - \text{átlag}X)^2} \cdot \sqrt{\sum_{k=1}^n (Y_k - \text{átlag}Y)^2}}$$

$$0.9981368882 \quad (1.6.3)$$

Az X és Y adatok szorosságát mérő másik szám az R^2 érték.

17.6.3. DEFINÍCIÓ. R^2 értéke

Az

$$R^2 = 1 - \frac{\sum_{k=1}^n (y_k - y_{modell, k})^2}{\sum_{k=1}^n (y_k - y_{\text{átlag}})^2} = 1 - \frac{RSS}{TSS}$$

együttható méri az Y mérési adatok és a modell alapján számolt Y_{modell} adatok jóságát.

A számlálóban szereplő $\sum_{k=1}^n (y_k - y_{modell, k})^2 = RSS$ érték az **eltérések négyzetösszege** vagy

Residual Sum of Square. A nevezőben szereplő $\sum_{k=1}^n (y_k - y_{\text{átlag}})^2 = TSS$ értéke a **teljes négyzetösszeg** vagy Total Sum of Squares.

17.6.4. TÉTEL

Lineáris modell esetén R^2 értéke megegyezik az r korrelációs együttható négyzetével.

Bizonyítás a példa kapcsán

Számoljuk a Residual Sum of Square (RSS) és a Total Sum of Square (TSS) értékét

$$> RSS := \sum_{k=1}^n (Y[k] - \text{modell}Y[k])^2;$$

$$TSS := \sum_{k=1}^n (Y[k] - \text{Mean}(Y))^2$$

$$RSS := 2.533333294$$

$$TSS := 680.5000000 \quad (1.6.4)$$

A kerekítési hibákon belül $\left(1 - \frac{RSS}{TSS}\right)$ egyenlő az r^2 korrelációs együttható négyzetével!

$$> 1 - \frac{RSS}{TSS} = r^2$$

$$0.9962772472 = 0.996277247122214 \quad (1.6.5)$$

▼ 17.7. A korrelációs együttható értéke mindig -1 és 1 között van

17.7.1. TÉTEL

Tetszőleges X és Y valószínűségi változó R korrelációs együtthatója a $[-1, 1]$ zárt intervallumba esik, azaz

$$R = \left| \frac{E((X - E_X)(Y - E_Y))}{\sigma_X \sigma_Y} \right| = \frac{|E_{XY} - E_X E_Y|}{\sigma_X \sigma_Y} = |\leq 1|.$$

BIZONYÍTÁS

Tekintsük a

$$\Phi(t) = E\left(\left(t \cdot (X - \mu_X) + (Y - \mu_Y)\right)^2\right)$$

egyváltozós nem-negatív függvény minimalizálását a t változóra, ahol $\mu_X = E(X)$ és $\mu_Y = E(Y)$ jelöli az X illetve Y valószínűségi változók várható értékét!

Végezzük el a négyzetre emelést a várható értéken belül, miután bevezettük a $K_X = X - \mu_X$ és $K_Y = Y - \mu_Y$ jelöléseket a különbségekre

$$\Phi(t) = E\left((t \cdot K_X - K_Y)^2\right) = E\left(t^2 \cdot K_X^2 - 2 \cdot t \cdot K_X \cdot K_Y + K_Y^2\right)$$

Használjuk az E várható érték lineáris tulajdonságát

$$\begin{aligned}\Phi(t) &= t^2 \cdot E(K_X^2) - 2 \cdot t \cdot E(K_X \cdot K_Y) + E(K_Y^2) = \\ &= t^2 E\left((X - \mu_X)^2\right) - 2 t E\left((X - \mu_X) \cdot (Y - \mu_Y)\right) + E\left((Y - \mu_Y)^2\right)\end{aligned}$$

Írjuk be a varianciákat és a kovarianciát a megfelelő kifejezés helyére

$$\Phi(t) = t^2 \cdot \text{Var}_X - 2 \cdot t \cdot \text{Cov}(X, Y) + \text{Var}_Y$$

Tehát $\Phi(t)$ a t változó másodfokú polinomja és grafikonja egy olyan parabola, melynek minimuma van, mert t^2 együtthatója $\text{Var}_X \geq 0$. Továbbá $\Phi(t) \geq 0$ minden valós t esetén, hiszen egy teljesnégyzet várható értéke. Ezért a másodfokú polinom diszkriminánsa nem lehet pozitív

$$\text{diszkrimináns} = 4 \text{Cov}(X, Y)^2 - 4 \text{Var}_X \text{Var}_Y \leq 0$$

Rendezve az egyenlőtlenséget kapjuk a kívánt becslést a korrelációs együtthatóra

$$\text{Cov}(X, Y)^2 \leq \text{Var}_X \text{Var}_Y$$

$$R^2 = \frac{\text{Cov}(X, Y)^2}{\text{Var}_X \text{Var}_Y} \leq 1$$

Tehát a korrelációs együttható abszolút értéke 1-nél nem nagyobb

$$R = \sqrt{\frac{\text{Cov}(X, Y)^2}{\text{Var}_X \text{Var}_Y}} \leq 1$$

▼ 17.8. Az egyenes együtthatóinak konfidencia intervalluma T-eloszlással

17.8.1. PÉLDA. 17.5.1. példa 'a' és 'b' paramétereinek 95%-os konfidencia becslése

Adjunk az $y = a x + b$ egyenes a és b paraméterére 95%-os megbízhatósági szinten konfidencia intervallumot!

Megoldás

Az intervallumot megadja a **LinearFit** eljárás, ha az "**output=solutionmodule**" opcióval kérdezősködünk.

> **restart :**

with(Statistics) :

X, Y := Vector([45, 50, 55, 60, 65, 70, 75, 80, 85, 90]), Vector([43, 45, 48, 51, 55, 57, 59, 63, 66, 68]) :

f := LinearFit([x, 1], X, Y, x);

$$f := 0.573333333333334 x + 16.8000000000000 \quad (1.8.1)$$

> **h := LinearFit([x, 1], X, Y, x, output = solutionmodule);**

(h:-Settings) (); eredmények := (h:-Results) ();

[["svdtolerance" = 1.0 10⁻¹², "confidencelevel" = 0.95]]

eredmenyek := "residualmeansquare" = 0.3166666666666666, "residualsumofsquares" **(1.8.2)**

= 2.53333333333332, "residualstandarddeviation" = 0.562731433871137,

"degreesoffreedom" = 8, "parametervalues" = $\begin{bmatrix} 0.573333333333334 \\ 16.8000000000000 \end{bmatrix}$, "parametervector"

$$= \begin{bmatrix} 0.57333333333334 \\ 16.8000000000000 \end{bmatrix}, \text{"leastquaresfunction"} = 0.57333333333334 x$$

$$+ 16.8000000000000, \text{"standarderrors"} = \begin{bmatrix} 0.0123909383637944 & 0.855109420607748 \end{bmatrix},$$

$$\text{"confidenceintervals"} = \begin{bmatrix} 0.544759792876018 \text{..} 0.601906873790649 \\ 14.8281151509429 \text{..} 18.7718848490571 \end{bmatrix}, \text{"residuals"}$$

= [0.39999999999988, -0.46666666666665, -0.33333333333328, -0.19999999999998,

0.93333333333337, 0.066666666666675, -0.79999999999995, 0.33333333333341,

0.466666666666661, -0.400000000000009], "leverages" = [0.345454545454545,

0.248484848484848, 0.175757575757576, 0.127272727272727, 0.103030303030303,

0.103030303030303, 0.127272727272727, 0.175757575757576, 0.248484848484849,

0.345454545454546], "variancecovariancematrix"

$$= \begin{bmatrix} 0.000153535353535353 & -0.0103636363636363 \\ -0.0103636363636363 & 0.731212121212119 \end{bmatrix}, \text{"internallystandardizedresiduals"}$$

$$= \begin{bmatrix} 0.878595370213933 \\ -0.956614006031951 \\ -0.652454285316016 \\ -0.380442955126338 \\ 1.75124422340639 \\ 0.125088873100458 \\ -1.52177182050536 \\ 0.652454285316042 \\ 0.956614006031944 \\ -0.878595370213980 \end{bmatrix}, \text{"externallystandardizedresiduals"} = \begin{bmatrix} 0.864622932574618 \\ -0.950865762628046 \\ -0.627231353045424 \\ -0.359135384577473 \\ 2.08609390343129 \\ 0.117124524693578 \\ -1.68874268373006 \\ 0.627231353045450 \\ 0.950865762628039 \\ -0.864622932574669 \end{bmatrix},$$

$$\text{"CookDstatistic"} = \begin{bmatrix} 0.203703703703692 \\ 0.151288405717727 \\ 0.0453867692587858 \\ 0.0105537280701753 \\ 0.176137018953522 \\ 0.000898658259966961 \\ 0.168859649122806 \\ 0.0453867692587895 \\ 0.151288405717725 \\ 0.203703703703714 \end{bmatrix}, \text{"AtkinsonTstatistic"}$$

$$= \begin{bmatrix} 1.25626799578685 \\ -1.09352896682826 \\ -0.579278352705598 \\ -0.274294180760558 \\ 1.41402618727293 \\ 0.0793910306799715 \\ -1.28979852958265 \\ 0.579278352705623 \\ 1.09352896682825 \\ -1.25626799578692 \end{bmatrix}$$

Olyan hosszú adatsort kaptunk, hogy végig olvasni is nehéz! Bennünket most csak a 9-ik adat érdekel, amely az **a** és **b** együtthatók **konfidencia intervallumait** tartalmazza, az automatikusan beállított 95%-os szignifikancia szintre vonatkozóan!

> *h:-Results()*₉

$$\text{"confidenceintervals"} = \begin{bmatrix} 0.544759792876018 \dots 0.601906873790649 \\ 14.8281151509429 \dots 18.7718848490571 \end{bmatrix} \quad (1.8.3)$$

Az első konfidencia intervallum az első alapfüggvény együtthatójának konfidencia intervalluma. Ez azt jelenti, hogy 95%-os megbízhatósággal az $y = a x + b$ képletben az x együtthatója

$$a_1 = 0.54475 < a < 0.601906 = a_2$$

intervallumba kell, hogy essen.

Hasonlóan 95%-os megbízhatósággal a **b** együtthatóra a

$$b_1 = 14.828 < b < 18.7718 = b_2$$

intervallum adódik. Ez utóbbi észrevehetően elég nagy variabilitást enged meg b -re!

Megmutatjuk, hogyan jöttek ki ezek a konfidencia intervallumok az **adatok** és a **Student T - eloszlás** segítségével.

Az intervallum szimmetrikus lesz a kiszámolt 'a' illetve 'b' értékre nézve, amelyet 'A' illetve B nevű változóknak tárolunk.

> *A := coeff (f, x, 1); B := coeff (f, x, 0)*

$$A := 0.5733333333333334$$

$$B := 16.800000000000000 \quad (1.8.4)$$

Kezdjük az 'a' együttható intervallumának kiszámításával!

A számítás egyik fontos alappillére, a **StudentT - eloszlásból** és a **konfidencia szintből** számított T érték. A Student T-eloszláshoz szükséges paraméterét, a **szabadsági fokot** megtaláljuk az eredmények 4-ik elemeként.

> *szabadsagfok := eredmények[4];*

szint := (h:-Settings) () [2]; N := LinearAlgebra[Dimension](Y);

szabadsagfok := "degreesoffreedom" = 8

szint := "confidencelevel" = 0.95

$$N := 10 \quad (1.8.5)$$

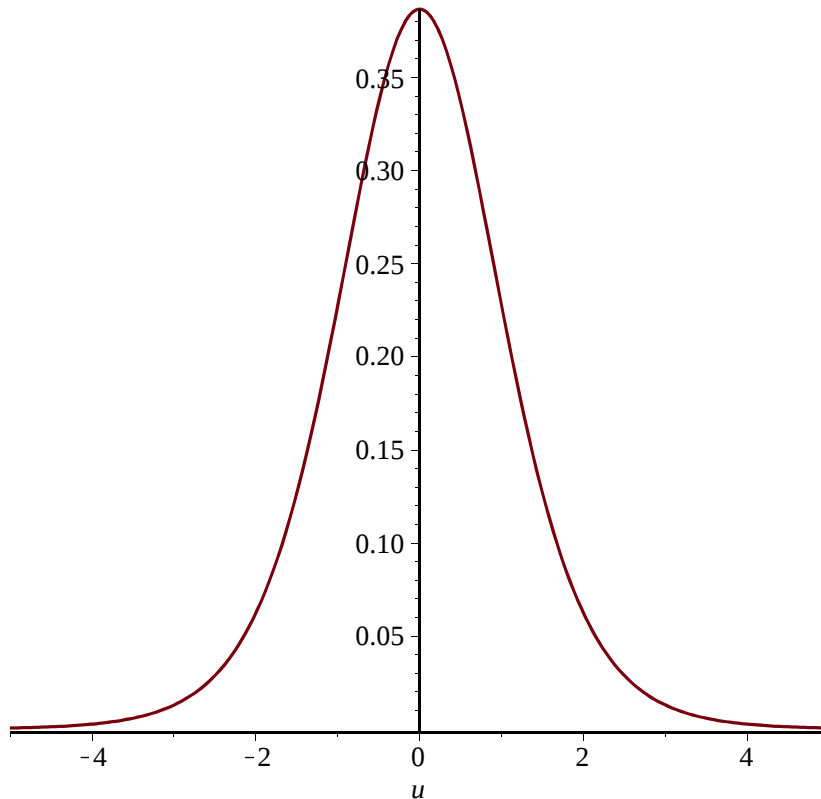
Láthatjuk, hogy a T-eloszlás szabadságfoka $(N - 2) = 8$, ahol N jelöli az adatok számát, jelen pillanatban $N=10$.

Nézzük csak az (N-2) szabadsági fokú **StudentT** eloszlás sűrűség függvényét!

```
> T3 := RandomVariable(StudentT(N-2)) :
surusegT := unapply(PDF(T3, u), u);
rajz3 := plot(surusegT(u), u = -5..5, title
= "A 8 szabadsági fokú T -eloszlás sűrűség függvénye") : rajz3;
```

$$surusegT := u \rightarrow \frac{35}{256} \frac{\sqrt{8}}{\left(1 + \frac{1}{8} u^2\right)^{9/2}}$$

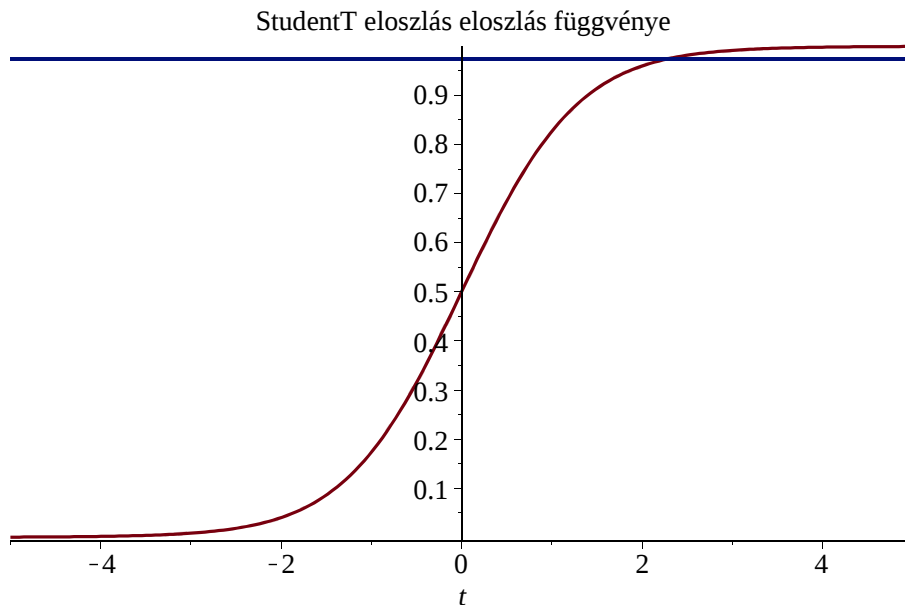
A 8 szabadsági fokú T -eloszlás sűrűség függvénye



Az eloszlás függvényt úgy kapjuk, ha integráljuk a sűrűség függvényt $(-\infty)$ -tól t -ig!

```
> alpha := 0.05;
eloszlasT := unapply(evalf(int(surusegT(u), u = -infinity..t)), t);
plot([eloszlasT(t), 1-alpha/2], t = -5..5, title = "StudentT eloszlás eloszlás függvénye");
alpha := 0.05
```

$$eloszlasT := t \rightarrow \frac{1}{(16. + 2. t^2)^{9/2}} \left(0.3535533905 \left(1.414213562 (16. + 2. t^2)^{9/2} \right. \right. \\ \left. \left. + 2.86720 \cdot 10^5 t + 1.07520 \cdot 10^5 t^3 + 16128. t^5 + 1152. t^7 + 32. t^9 \right) \right)$$



A T-eloszlás eloszlásfüggvénye és a 0,95 megbízhatósági szint segítségével meg kell határozni azt a **T** helyet az x- tengelyen, amelyre teljesül az

$$\text{eloszlásT}(T) = 1 - \frac{1 - 0,95}{2} = 1 - \frac{0,05}{2} = 0,975$$

egyenlőség. A számítás kivitelezésére az **fsolve** eljárás kiválóan alkalmas, amely az egyenlet közelítő megoldását adja meg!

> **T := fsolve(eloszlásT(t) = 0.975, t);**

$$T := 2.306004145 \quad (1.8.6)$$

A számítás **másik alappillére** az a legkisebb négyzetes eltérés, amelyet az adatok és az egyenes közötti értékek adnak. Ezt a Maple **reziduális standard eltérésnek** nevezi és ez az eredmények 3-ik eleme.

> **eredmenyek[3]; S := rhs(%);**

"residualstandarddeviation" = 0.562731433871137

$$S := 0.562731433871137 \quad (1.8.7)$$

Megmutatjuk, hogy ez valóban a legkisebb négyzetes eltérés (N-2)-vel osztott érték négyzetgyöke!

$$\begin{aligned} > \sqrt{\frac{\sum_{k=1}^N (Y_k - A X_k - B)^2}{N - 2}} = \sqrt{\frac{\text{add}((Y_k - A X_k - B)^2, k = 1 .. N)}{N - 2}} \\ & \quad \sqrt{\frac{\sum_{k=1}^N (Y_k - A X_k - B)^2}{N - 2}} = 0.562731433871138 \end{aligned} \quad (1.8.8)$$

A számításhoz szükséges még az X adatok m várható értéke!

> $m := \text{Mean}(X);$

$$m := 67.50000000000000 \quad (1.8.9)$$

Ezekkel az adatokkal az $y = a \cdot x + b$ egyenes 'a' paraméterére a konfidencia - intervallum hossza

$$a \text{ konfidencia-intervallum hossza } a\text{-ra} = \frac{T_{N-2, 1-\frac{\gamma}{2}} S}{\sqrt{\sum_{k=1}^N (X_k - m)^2}}$$

képlettel számolható.

Nézzük az adatok alapján számolt konfidencia intervallumot és hasonlítsuk össze a **LinearFit** eljárás által visszaadott értékkel.

$$\begin{aligned} > \text{'konfidencia intervallum } a\text{-ra'} = A - \frac{TS}{\sqrt{\text{add}((X_k - m)^2, k = 1..N)}} ..A \\ &+ \frac{TS}{\sqrt{\text{add}((X_k - m)^2, k = 1..N)}} \\ &\text{konfidencia intervallum } a\text{-ra} = 0.544759778105984 ..0.601906888560683 \end{aligned} \quad (1.8.10)$$

A 'b' együtttható konfidencia - intervallumának számításához már minden adott, mert ennek hossza a

$$a \text{ konfidencia intervallum hossza } b\text{-re} = T_{N-2, 1-\frac{\gamma}{2}} S \sqrt{\frac{1}{N} + \frac{m^2}{\sum_{k=1}^N (X_k - m)^2}}$$

képlet alapján számolható.

$$\begin{aligned} > \text{'konfidencia intervallum } b\text{-re'} = B - TS \sqrt{\frac{1}{N} + \frac{m^2}{\text{add}((X_i - m)^2, i = 1..N)}} ..B \\ &+ TS \sqrt{\frac{1}{N} + \frac{m^2}{\text{add}((X_i - m)^2, i = 1..N)}} \\ &\text{konfidencia intervallum } b\text{-re} = 14.8281141316500 ..18.7718858683500 \end{aligned} \quad (1.8.11)$$

Ez megegyezik a LinearFit által b-re adott megbízhatósági intervallummal.

Ezek a becslések nem jelentik azt, hogy az $y = a x + b$ egyenes az $y_1 = a_1 x + b_1$ alsó egyenes és az $y_2 = a_2 x + b_2$ felső egyenes között helyezkedik el 95%-os biztonsággal! Csak az **a** és **b** számításának pontosságát jelzik a fenti eredmények! Hogyan lehet akkor a regressziós egyenes konfidencia intervallumát megadni?

▼ 17.9. A regressziós egyenes konfidencia intervalluma

17.9.1. PÉLDA. A17.5.1. példa regressziós egyenesének 95%-os megbízhatóságú becslése
Rajzoljuk fel a lineáris közelítés körül a 95%-os megbízhatóságú alsó és felső becslést!

Megoldás

Az $f(x) = a x + b$ lineáris függvény alsó és felső becslése szimmetrikus az egyenesre nézve, de nem egyenes, hanem egy-egy négyzetgyökös függvény az alábbi képlettel

$$\text{Az egyenes becslésének hibája} = T_{N-2, 1-\frac{\gamma}{2}} \sqrt{\frac{1}{N} + \frac{(x-m)^2}{\sum_{k=1}^N (X_k - m)^2}} S$$

Ezt az eltérést felvesszük egy változóban és levonjuk illetve hozzáadjuk a kiszámolt 'f' egyenes képletéhez.

```
> elteres := T * sqrt( 1/N + (x-m)^2 / add( (X_k - m)^2, k=1..N) ) * S;
> also := evalf( f - elteres ); felso := evalf( f + elteres )
elteres :=
```

$$1.29766101902863 \sqrt{\frac{1}{10} + 0.000484848484848485 (x - 67.5000000000000)^2}$$

```
also := 0.573333333333334 x + 16.8000000000000
- 1.29766101902863
```

$$\sqrt{0.1000000000 + 0.000484848484848485 (x - 67.5000000000000)^2}$$

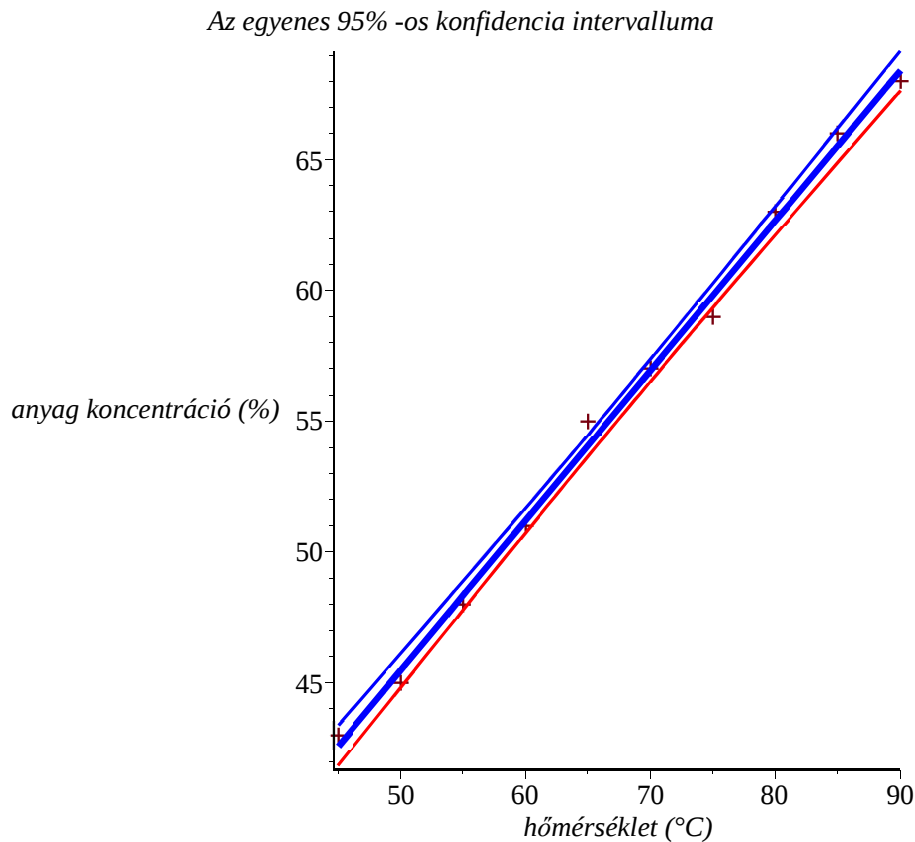
```
felso := 0.573333333333334 x + 16.8000000000000
+ 1.29766101902863
```

$$\sqrt{0.1000000000 + 0.000484848484848485 (x - 67.5000000000000)^2}$$

(1.9.1)

Nincs más hátra, mint felrajzolni a két görbét az egyenes körül.

```
> x0 := min( op( convert( X, list ) ) ); x1 := max( op( convert( X, list ) ) )
> pontok := zip( ( x, y ) -> [ x, y ], X, Y ) :
    rajz1 := plot( pontok, style = point, symbolsize = 14, symbol = cross ) : rajz2 := plot( f, x = x0
    .. x1, color = blue, thickness = 3 ) : rajz3 := plots[display]( [ rajz1, rajz2 ], labels
    = [ hőmérséklet (°C), anyag koncentráció (%) ], title
    = Lineáris kapcsolat a hőmérséklet és az anyag koncentrációja között ) :
> gorbek := plot( [ also, felso ], x = x0 .. x1, color = [ red, blue, green ] ) :
    plots[display]( [ rajz3, gorbek ], title = `Az egyenes 95% -os konfidencia intervalluma` )
    x0 := 45
    x1 := 90
```



Jól eső érzéssel nyugtázzuk, hogy "majdnem" minden adat az egyenes körüli 95%-os megbízhatósági tartományban van! Alaposabban megnézve az alsó és felső határt azt mondhatjuk, hogy az adatok közepe felé a sáv összeszűkül, míg a szélek felé kinyílik ollószerűen. Ez persze a négyzetgyökös képlet következménye!

▼ 17.10. Elméleti ellenőrző kérdések

17.10.1. Milyen összefüggések teljesülése esetén lesz az X és Y valószínűségi változók kapcsolata lineáris?

17.10.2. Milyen minimalizálási számítást követel meg a legkisebb négyzetes elv?

17.10.3. Hogyan számoljuk ki a lineáris regressziós egyenes együtthatóit az adatokból?

17.10.4. Milyen adatokra nem illeszthető lineáris görbe?

17.10.5. Mit lehet leolvasni a korrelációs együtthatók értékeiből?

17.10.6. Milyen eloszlást használunk az 'a' és 'b' paraméterek konfidencia intervallumának becsléséhez?

▼ 17.11. Gyakorló feladatok

▼ 17.11.1. feladat

Az ivóvízkészletek tisztán tartása érdekében a szennyező anyagok mennyiségét rendszeresen ellenőrzik mérésel. A mérési módszer indirekt, melyben felhasználják, hogy a kémiai

szennyezőanyagok elnyelik a fényt. Ezért fényt eresztenek a vízben keresztül és mérik a fény elnyelés mértékét. A mérés kalibrációja érdekében a laboratórium standard oldatokon méri a fényelnyelést és a szennyezés koncentrációját, majd regressziós illesztést alkalmaznak. Ez alapján egy ismeretlen szennyezésű víz mérése automatikussá válik. A következő táblázat nitráttal szennyezett víz (X, milligram/literben) fényelnyelési (Y) kalibrációs adatait tartalmazza

Nitrát tartalom (milligram/liter) X:=[50,50,100,200,400,800,1200,1600,2000,2000]
(milligram/liter)

Fényelnyelés Y:=[7.0,7.5,12.8,24.0,47.0,93.0,138.0,183.0,230.0,226.0] (dimenzió nélküli mennyiség)

(a) Ábrázolja az összetartozó nitrát tartalom (X) és fényelnyelés (Y) értékeket a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a nitrát tartalom és a fényelnyelési értékek között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Várhatóan 50 egység fényelnyeléshez mennyi milligram/literrel kell növelni a szennyezést?

(e) Számolja ki a fényelnyelési értékek és az illesztett lineáris modell alapján számolt eltéréseket, majd rajzolja fel ezek hisztogramját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Tesztelje, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.2. feladat

A levegő nitrogén-oxid tartalma (Y) hatással van az atmoszféra ózon tartalmára is és a belélegzett tiszta nitrogén mennyiségére is. Az alábbi táblázat az egyes években (X) mért levegő nitrogén-oxid érték (Y) mennyiségét mutatják.

A mérés évei X:=[1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,1994,1995,1996,1997,1998,1999]

A levegő nitrogén-oxid tartalma Y:= [0.73,2.55, 2.90, 3.83, 2.53, 2.77, 3.93, 2.03, 4.39, 3.04, 3.41, 5.07, 3.95, 3.14, 3.44, 3.63, 4.50, 3.95, 5.24, 3.30, 4.36, 3.33]

(a) Ábrázolja az évek és a levegő nitrogén-oxid tartalmának összetartozó értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható az évek múlása és a levegő nitrogén-oxid tartalma között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Átlagosan mennyi Δx évnek kell eltelni ahhoz, hogy levegő nitrogén-oxid tartalma $\Delta y = 1$ egységgel növekedjen?

(e) Számolja ki az eltéréseket a levegő nitrogén-oxid tartalma és az illesztett lineáris modell alapján számolt értékek között, majd rajzolja fel ezek hisztogramját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség

függvényét!

(g) Tesztelje, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ elsőfajú hiba szinten?

(h) Rajzolja fel az illesztett és a mért adatsorokat a **QuantilePlot** segítségével! Számolja ki R^2 értékét! Mennyire illeszkedik jól a lineáris modell?

17.11.3. feladat

A következő táblázat USA-ban élő családok éves jövedelmét (X, \$-ban) és kiadásait (Y,\$-ban) tartalmazza.

Jövedelem (\$-ban) X:=[41200,50100,52000,66000,44500,37700,73500,37500,56700,35600,37500,36900,48200,34400,29900,44750,60550,52000,67700,20600] (\$-ban)

Kiadások (\$-ban) Y:=[2400,2650,2350,4950,3100,2500,5106,3100,2900,1750,1450,2020,3750,1675,2400,2550,3880,3330,4050,1150] (\$-ban)

(a) Ábrázolja az összetartozó jövedelem (X) és kiadás (Y) értékeket a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a jövedelem és a kiadás értékek között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Mennyi jövedelem növekedés esetén várható $\Delta y=100$ \$ kiadás növekedés?

(e) Számolja ki a kiadási értékek és az illesztett lineáris modell alapján számolt értékek eltéréseit, majd rajzolja fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Tesztelje, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Rajzolja fel az illesztett és a mért adatsorokat a **QuantilePlot** segítségével! Számolja ki R^2 értékét! Mennyire jó a lineáris modell?

17.11.4. feladat

Egy rúd lehajlási értékeit (Y) mérték növekvő terhelési (X) értékek mellett. A mérési adatok a következők voltak

terhelési értékek X:=[0.11019,0.21956,0.32949, 0.43899, 0.54803,0.65694,0.76562, 0.87487,0.98292, 1.09146, 1.20001, 1.30822, 1.41599 ,1.52399, 1.63194, 1.73947, 1.84646, 1.95392,2.06128, 2.16844]

lehajlási értékek Y:=[18, 30, 44, 60, 75,90, 104, 120,136,150,165,178, 195, 210,227, 240, 253,275,285,300]

(a) Ábrázolja az egyes terhelésekhez tartozó lehajlási adatok összetartozó értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a terhelés és a lehajlás között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Várhatóan mennyivel kell növelni a terhelést, hogy a lehajlás 10 egységgel növekedjen?

(e) Számolja ki a lehajlási értékek és az illesztett lineáris modell alapján számolt eltéréseket, majd rajzolja fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Tesztelje, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Rajzolja fel az illesztett és a mért adatsorokat a **QuantilePlot** segítségével! Számolja ki R^2 értékét! Mennyire jó a lineáris modell?

17.11.5. feladat

Kutatók mérték a **jégcsapok növekedését** (Y) cm-ben az idő (X) múlásával azonos körülmények között. Egy 3 órás időtartam alatt 10 percenkénti méréssel az alábbi adatokat kapták.

Idő tartam X:=[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180] (percekben)

Jégcsap hossza Y:=[0.6, 1.8, 2.9, 4.0, 5.0, 6.1, 7.9, 10.1, 10.9 , 12.7, 14.4, 16.6, 18.1, 19.9, 21.0, 24.7, 27.8] (cm -ben)

(a) Ábrázolja az egyes időpontokhoz (X) tartozó jégcsap hosszúságok (Y) értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható az idő és a jégcsap hossza között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Várhatóan mennyi idő alatt növekszik a jégcsap 1 cm-et?

(e) Számolja ki a jégcsap hosszak és az illesztett lineáris modell alapján számolt eltéréseket, majd rajzolja fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Tesztelje, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Rajzolja fel az illesztett és a mért adatsorokat a **QuantilePlot** segítségével! Számolja ki R^2 értékét! Mennyire jó a lineáris modell?

17.11.6. feladat

Vajon a bor jó hatással van a szívre? Néhányan úgy vélik, hogy a mérsékelt bor fogyasztás megóvja az embert a szívrohamtól. Az alábbi táblázat 19 nemzet éves X bor fogyasztási átlagot tartalmazza (literben mérve személyenként) és az éves Y halálozási gyakoriságot szívroham miatt (100 000 emberre vonatkoztatva).

Nemzetek = [Ausztrália, Hollandia, Ausztria, Új Zéland, Belgium, Norvégia, Kanada, Spanyolország, Dánia, Svédország, Finnország, Svájc, Franciaország, Egyesült Királyság, Izland, Egyesült Államok, Írország, Németország, Olaszország]

bor fogyasztási adatok X:=[2.5, 1.8,3.9,1.9,2.9,0.8,2.4,6.5,2.9,1.6,0.8,5.8,9.1,1.3,0.8,1.2,0.7, 2.7,7.9]

halálozási adatok szívroham miatt Y:=[211,167,167,266,131,227,191,86,220,207,297,115, 71,285,211,199,300,172,107]

(a) Ábrázolja az egyes nemzetekhez tartozó borfogyasztást és a szívroham miatti halálozási adatok összetartozó értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a

korrelációs együttható a borfogyasztás és a szívroham miatti halálok száma között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Várhatóan mennyivel kell növelni a borfogyasztást, hogy a halál esetek száma 50-nel csökkenjen?

(e) Számolja ki a halál esetek számának és az illesztett lineáris görbe alapján számolt halálok számának eltéréseit, majd rajzolja fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Tesztelje, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.7. feladat

Vajon egy gyermek esetében a beszédképesség kialakulásának X időpontja (hónapokban) hatással van-e a gyermek későbbi szellemi fejlődésére. Feljegyezték 21 gyermek esetében azt a hónapot, amikor az első szót kimondta és később ún. Gesell-teszt Y pontszáma segítségével mérték a gyermek értelmi képességét. Az adatokat az alábbi táblázat mutatja.

a hónap, amikor beszélni kezdett: $X=[15, 26, 10, 9, 15, 20, 18, 11, 8, 20, 7, 9, 10, 11, 11, 10, 12, 42, 17, 11, 10]$

a Gesell-teszt pontszáma : $Y=[95, 71, 83, 91, 102, 87, 93, 100, 104, 94, 113, 96, 83, 94, 102, 100, 105, 57, 121, 86, 100]$

(a) Ábrázolja az egyes gyermekekhez tartozó beszéd időpont és Gesell-pontszám összetartozó értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a beszéd időpont és a kapott pontszám között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Az egyik gyermek pontszáma 10-zel kevesebb, mint a másiké. Vajon hány hónappal később szólalt meg ez a gyermek?

(e) Számolja ki a kapott pontszámok és az illesztett lineáris görbe alapján számolt pontok értékének eltéréseit, majd rajzolja fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Tesztelje, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.8. feladat

A Coloradó Állami Egyetemen oktató William Gray egy trópusi meteorológiai project vezetője, akinek éves hurrikán előrejelzéseit különösen nagy érdeklődés kíséri. Megadjuk Dr. Gray által jósolt atlanti trópusi viharok számát (X értékek) és a ténylegesen bekövetkezett viharok (Y értékek) számát 1984 és 2008 között:

Jósolt viharok száma:

$X=[10, 11, 8, 8, 11, 7, 11, 8, 8, 11, 9, 12, 10, 11, 10, 14, 12, 12, 11, 14, 14, 15, 17, 17, 15]$

Megvalósult viharok száma:

$Y=[12, 11, 6, 7, 12, 11, 14, 8, 6, 8, 7, 19, 13, 7, 14, 12, 14, 15, 12, 16, 14, 27, 9, 14, 16]$

(a) Ábrázolja az összetartozó jósolt és megvalósult éves vihar számait a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a

korrelációs együttható a jósolt és megvalósult viharok között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!

(d) Várhatóan mennyi vihart jósol William Gray abban a szezonban, amelyben a trópusi viharok száma 10?

(e) Számolja ki a megvalósult viharok és az illesztett lineáris görbe alapján számolt viharok eltéréseit, majd rajzolja fel ezek hisztogramját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Tesztelje, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha = 0.05$ szignifikancia szinten?

17.11.9. feladat

Vizsgálták a gépjárművek X súlyának és autópályán az Y benzin fogyasztás hatékonyságának (1 gallon benzinnel megtett mérföldek értéke) közötti kapcsolatot.

Az alábbi táblázat különböző típusú gépjárművekhez tartozó adatokat tartalmazza, az X súly értékei 1000 font-ban és az Y hatásfok értékei mérföld/gallon-ban adottak.

Súlyok: $X = [3.545, 2.6, 3.245, 3.93, 3.995, 3.115, 3.235, 3.225, 2.44, 3.24, 2.29, 2.5, 4.02]$

Hatásfok: $Y = [30, 32, 30, 24, 26, 30, 33, 27, 37, 32, 37, 34, 26]$

(a) Ábrázolja az összetartozó súly és hatásfok értékeket a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a súly és hatásfok között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Várhatóan hány ezer font a súlya annak az autónak, mely 38 mérföldet képes megtenni 1 gallon benzinnel?

(e) Számolja ki a hatékonysági adatok és az illesztett lineáris görbe alapján számolt eltéréseket, majd rajzolja fel ezek hisztogramját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét! Vajon a várható érték nullának tekinthető-e?

(g) Teszteljük, hogy vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha = 0.05$ szignifikancia szinten?

17.11.10. feladat

Egy vállalat mérte a havi befektetési költségeit (X, ezer \$-ban) és a termelést (Y, ezer darabban). A adatok az alábbiak:

Havi befektetési költségek (ezer \$ mértékegységben)

$X := [46.75, 42.18, 41.86, 43.29, 42.12, 41.78, 41.47, 42.21, 41.03, 39.84, 39.15, 39.20, 39.52, 38.05, 39.16, 38.59, 36.54, 37.03, 36.60, 37.58, 36.48, 38.25, 37.26, 38.59, 40.89, 37.66, 38.79, 38.78, 36.70, 35.10, 33.75, 34.29, 32.26, 30.97, 28.20, 24.58, 20.25, 17.09, 14.35, 13.11, 9.50, 9.74, 9.34, 7.51, 8.35, 6.25, 5.45, 3.79]$

Havi termelés (ezer darabban)

$Y := [92.64, 88.81, 86.44, 88.80, 86.38, 89.87, 88.53, 91.11, 81.22, 83.72, 84.54, 85.66, 85.87, 85.23, 87.75, 92.62, 91.56, 84.12, 81.22, 83.35, 82.29, 80.92, 76.92, 78.35, 74.57, 71.60, 65.64, 62.09, 61.66, 77.14, 75.47, 70.37, 66.71, 64.37, 56.09, 50.25, 43.65, 38.01, 31.40, 29.45, 29.02, 19.05, 20.36, 17.68, 19.23, 14.92, 11.44, 12.69]$

(a) Ábrázoljuk a költség és a termelés összetartozó értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a

korrelációs együttható a két adatsor között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) A lineáris regresszió alapján átlagosan mennyi befektetés szükséges, hogy a termelés 10 ezer darab termékkel növekedjen egy hónapban?

(e) Számolja ki a mért és az illesztett lineáris görbe alapján a termelések közötti eltéréseket, majd rajzolja fel ezek hisztogramját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Rajzolja fel az illesztett és a mért adatsorokat a QuantilePlot segítségével! Számolja ki az R^2 értékét! Mennyire jó a lineáris modell?

17.11.11. feladat

A faipari termékek fontos jellemzője a fa sűrűsége és keménysége. Továbbá a dokumentációkon fel kell tüntetni ezek kapcsolatát is. Egy műhelyben megmérték 30 fa minta X sűrűségét és Y keménységét. A adatok az alábbiak:

Sűrűség adatok (font/inch³ mértékegységben)

X:= [9.50, 9.80, 8.30, 8.60, 7.00, 17.40, 15.20, 16.70, 15.00, 14.80, 25.60, 24.40, 19.50, 22.80, 19.80, 8.40, 11.00, 9.90, 6.40, 8.20, 15.00, 16.40, 15.40, 14.50, 13.60, 23.40, 23.30, 21.20, 21.70, 21.30]

Keményiségi adatok (font/inch² mértékegységben)

Y:= [14814.00, 14007.00, 7573.00, 9714.00, 5304.00, 43243.00, 28028.00, 49499.00, 26222.00, 26751.00, 96305.00, 72594.00, 32207.00, 70453.00, 38138.00, 17502.00, 19443.00, 14191.00, 8076.00, 10728.00, 25319.00, 41792.00, 25312.00, 22148.00, 18036.00, 104170.00, 49512.00, 48218.00, 47661.00, 53045.00]

(a) Ábrázoljuk a sűrűség és a keménység összetartozó értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a két adatsor között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) A lineáris regresszió alapján átlagosan mennyi sűrűségi változás szükséges 10000 font/inch² keménység növekedéséhez?

(e) Számolja ki a mért keménységi értékek és az illesztett lineáris görbe közötti eltéréseket, majd rajzolja fel ezek hisztogramját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Rajzolja fel az illesztett és a mért adatsorokat a QuantilePlot segítségével! Számolja ki az R^2 értékét! Mennyire jó a lineáris modell?

17.11.12. feladat

Vizsgálták a lucerna termés mértékét (X változó) és az öntözés (Y változó) kapcsolatát! Az alábbi adatsorok 7 farmon mért mérési adatok

X=[12,18,24,30,36,42,48] az öntözéshez használt víz magassága inch-ben

Y=[5.27,5.68,6.25,7.21,8.02,8.71,8.42] a lucerna termés mennyisége tonnában 1 acre

($\sim 4047 \text{ m}^2$) területen

(a) Ábrázolja a víz magasság és a lucerna termés összetartozó értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a két adatsor között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) A lineáris regresszió alapján átlagosan mennyi víz szükséges 1 tonna lucerna termés növekedéséhez?

(e) Számolja ki az lucerna termés és az illesztett lineáris görbe alapján számolt termés közötti eltéréseket, majd rajzolja fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.13. feladat

Roberts és Griffiths (1937) az angol gyermekek értelmi képességét pontozták az ún. BINET illetve OTIS teszt alapján. Az alábbi táblázat tartalmazza a két teszt alapján ugyanazon személyek elért eredményeit!

X=[36,28,72,74,75,76,77,78,81,82,83,83,83,84,85,86,86,86,87,89] a BINET-teszt eredményei
Y= [36,28,34,28,48,50,62,22,82,84,64,77,82,92,91,65,75,76,68,80] az OTIS -teszt eredményei

(a) Ábrázolja a két teszt összetartozó eredményeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a két teszt között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) A lineáris regresszió alapján mennyi BINET pont növekedés szükséges 5 OTIS pont növekedéshez átlagosan?

(e) Számolja ki az Otis pontszámok és az illesztett lineáris görbe alapján számolt pontok közötti eltéréseket, majd rajzolja fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.14. feladat

Az Egyesült Államok Statisztikai folyórata a "Technometrics" 1989. évi novemberi számában közzétett 21 **könyv** oldalszámának és árának (USA \$-ban) adatait.

Az alábbi táblázat tartalmazza az oldal számok (X) értékeit és a megfelelő könyv (Y) árát.

Oldal számok (X) : 264, 427, 273, 546, 328, 296, 63, 405, 53,130, 243, 188, 595, 96, 307, 88, 163, 265, 370, 475, 393.

Ár \$-ban (Y) : 45, 45, 36, 51, 59.95, 47.50, 8.75, 45, 7.95, 39.95, 39.95, 39.95, 48.64, 9.95, 37.5, 25, 45,59,84.25, 55, 51.95

(a) Ábrázolja a könyvek összetartozó oldalszámok és árak pontjait a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a könyv mérete és ára között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be

az egyenest a pontokhoz!

(d) A lineáris regresszió alapján 100 oldalanként átlagosan mennyivel növekszik egy könyv ára? (1 pont)

(e) Van 35\$ pénzünk. Hány oldalas könyvet kapunk érte?

(f) Számolja ki a megadott árak és az illesztett lineáris görbe alapján számolt árak közötti eltéréseket, majd rajzoljuk fel ezek hisztogrammját!

(g) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(h) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.15. feladat

Az Egyesült Államokban a múlt század elejétől folyamatosan végeztek felméréseket a farmokon élő lakosság létszámáról. A következő táblázat azt mutatja, hogyan csökkent a farmokon élők száma 1935 és 1980 között 5 éves ugrásokkal millió főben számolva.

Évek (X): 1935, 1940, 1945, 1950, 1955, 1960, 1965, 1970, 1975, 1980

Népesség (Y) 32.1, 30.5, 24.4, 23.0, 19.1, 15.6, 12.4, 9.7, 8.9, 7.2

(a) Ábrázoljuk a pontokat a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számoljuk ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a farmokon élő népesség száma és az idő múlása között?

(c) Illesszünk lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!

(d) A lineáris regresszió alapján évente átlagosan mennyivel csökken a farmon élők száma?

(e) Vajon mennyi ember élt 1990-ben a farmokon, ha a tendencia nem változott meg? Reális-e ez az adat?

(f) Számoljuk ki a mérési adatok és az illesztett lineáris görbe közötti eltéréseket, majd rajzoljuk fel ezek hisztogrammját!

(g) Számoljuk ki az eltérések átlagát és szórását! Rajzoljuk be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(h) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.16. feladat

Véletlenszerűen kiválasztott 26 férfi súlyát (X) és szisztolés vérnyomását (Y) mérték meg. A méréseket az alábbi két lista tartalmazza

X:= [165, 167, 180, 155, 212, 175, 190, 210, 200, 149, 158, 169, 170, 172, 159, 168, 174, 183, 215, 195, 180, 143, 240, 235, 192, 187]

Y:= [130, 133, 150, 128, 151, 146, 150, 140, 148, 125, 133, 135, 150, 153, 128, 132, 149, 158, 150, 163, 156, 124, 170, 165, 160, 159]

(a) Számítsuk ki az X és Y adatsor lineáris kapcsolatának szorosságát mutató korrelációs együttható értékét!

(b) Rajzoljuk fel az Y adatsor váltakozását az X függvényében!

(c) Határozzuk meg a lineáris regressziós egyenest és rajzoljuk fel a mérési pontokkal azonos ábrába!

(d) Rajzoljuk fel az eltérések hisztogramját!

(e) Mutassuk meg, hogy az eltérések normál eloszlást határoznak meg nulla várható értékkel és a kiszámolt szórással, $\alpha=0.05$ szignifikancia szinten!!

17.11.17. feladat

Mérték a vas darabok sűrűségét (X) és vas tartalmát (Y). Az eredményeket az alábbi táblázat

mutatja

Sűrűség (X) : 2.8, 2.9, 3.0, 3.1 , 3.2, 3.2, 3.2 ,3.3 ,3.4

Vas tartalom (Y) 27, 23 , 30, 28, 30 , 32, 34 , 33, 30

- (a) Ábrázoljuk a pontokat a **Statistics** csomag **ScatterPlot** eljárásával!
- (b) Számoljuk ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a vas tartalom és a sűrűség között?
- (c) Illesszünk lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!
- (d) Vajon mennyi a vas tartalma annak a darabnak, melynek sűrűsége 4, ha a tendencia változatlan?
- (e) Számoljuk ki a mérési adatok és az illesztett lineáris görbe közötti eltéréseket, majd rajzoljuk fel ezek hisztogrammját!
- (f) Számoljuk ki az eltérések átlagát és szórását! Rajzoljuk be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!
- (g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.18. feladat

A "Tappi Journal" folyóirat 1986. év márciusi számában adatokat közöltek a papír készítéshez használt pépes zöld folyadék Na_2S koncentrációjáról (X , gramm/literben) és a naponta elkészült papír Y mennyiségéről (Y, tonna/nap). A feljegyzett adatokat az alábbi táblázat mutatja

Na_2S koncentráció (X gramm/liter): [825, 830, 890, 895, 890, 910, 915, 960, 990, 1010, 1012, 1030, 1050]

Az elkészült papír mennyiség (Y tonna/nap) : [40, 42, 49, 46, 44, 48, 46, 43, 53, 52, 54, 57, 58]

- (a) Ábrázoljuk az (X;Y) pontokat a **Statistics** csomag **ScatterPlot** eljárásával!
- (b) Számoljuk ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a Na_2S koncentráció és az elkészült papír mennyisége között?
- (c) Illesszünk lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!
- (d) Vajon hány tonna papírt gyártanak azon anapon, amikor a pépes folyadék Na_2S koncentrációja 950 gram/liter, ha a tendencia változatlan?
- (e) Számoljuk ki a mérési adatok és az illesztett lineáris görbe közötti eltéréseket, majd rajzoljuk fel ezek hisztogrammját!
- (f) Számoljuk ki az eltérések átlagát és szórását! Rajzoljuk be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!
- (g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.19. feladat

Egy utazó ügynök az év 10 hónapját a termékeinek eladásával töltötte.

Feljegyezte minden hónapban, hogy mennyi kilométert (X változó) tett meg és mennyi volt az üzletkötésből származó bevétele (Y változó) ezer Ft.-ban.

Az adatok a következők

X:=[1370, 1050, 980, 1770, 1340, 1560, 2110, 1540, 1480,1670]

Y:=[2240,1800, 1900, 2200, 2300, 2300, 2500, 2300, 2100, 2200]

- (a) Vizsgáljuk meg a lineáris korreláció szorosságát az utazott km-ek mennyisége és a bevétel mértéke között!

- (b) Számítsuk ki az $Y = a + b \cdot X$ lineáris trend 'a' és 'b' együtthatóit!
- (c) Rajzoljuk fel a pontokat és az egyenest azonos koordináta-rendszerben!
- (d) Vizsgáljuk meg, hogy a mért Y_k adatok és a lineáris illesztésből számolt Y adatok eltérése normál eloszlást követ-e nulla várható értékkel 95%-os megbízhatósági szinten! Rajzoljuk fel az eltérések hisztogramját!

17.11.20. feladat

Egy kísérlet során feltételezték, hogy az anyagból szárítás segítségével távozó anyag mennyisége (Y) gramokban összefügg a szárítási idővel (X, órában). Mérték 10 szárítással kapcsolatos anyag csökkenési tömeget gramokban és mérték a szárítási időt órákban. Az adatok a következők

X:=[2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0]

Y:=[4.3, 1.5, 1.8, 4.9, 4.2, 4.8, 5.8, 6.2, 7.0, 7.9]

- (a) Vizsgáljuk meg a lineáris korreláció szorosságát a szárítási idő és az eltávozott anyag mennyisége között!
- (b) Számítsuk ki az $Y = a + b \cdot X$ lineáris trend 'a' és 'b' együtthatóit!
- (c) Rajzoljuk fel a pontokat és az egyenest azonos koordináta-rendszerben!
- (d) Vizsgáljuk meg, hogy a mért Y_k adatok és a lineáris illesztésből számolt Y adatok eltérése normál eloszlást követ-e nulla várható értékkel 95%-os megbízhatósági szinten! Rajzoljuk fel az eltérések hisztogramját!

17.11.21. feladat

Az alábbi táblázat 26 mérési adatpárt tartalmaz a szezonális átlagos talajvíz szintre (X) [méterben] és a hozzátartozó átlagos termés hozamra (Y) [tonna/hektárban] vonatkozóan.

X = [0.15,0.20,0.20,0.25,0.25,0.32,0.40,0.40,0.40,0.45,0.45,0.47,0.50,0.50,0.60,0.65,0.65, 0.70,0.70,0.75,0.90,0.90,0.91,1.00,1.05,1.08]

Y = [4.0,4.5 ,3.0 , 4.0,3.7,3.5,5.0,4.5,4.5,4.8,4.5,5.5,5.2,4.0,4.5,6.0,4.5,5.7,5.0,5.3,5.5,4.7,5.0, 4.5,5.7,5.5]

- (a) Ábrázolja az összetartozó talajvíz szint és az átlagos termés hozam értékeit a síkon a Statistics csomag ScatterPlot eljárásával!
- (b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a talajvíz és a termés hozam között?
- (c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!
- (d) Mekkora talajvíz növekedés esetén nő a termés hozam 1 tonna/ha-ral?
- (e) Számolja ki a megadott termés hozamok és az illesztett lineáris görbe alapján számolt eltéréseket, majd rajzoljuk fel ezek hisztogramját!
- (f) Számolja ki az eltérések átlagát és szórását! Rajolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!
- (g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?
- (h) Számítsuk ki a termés hozamok (Y) legkisebb, legnagyobb értékét, terjedelmét, átlagát, korrigált szórását, móduszát és mediánját!
- (i) Rajolja fel a termés hozamok váltakozását az átlaggal és a kvartilisekkel együtt! (4 pont)
- (j) Rajzoljuk fel közös koordináta-rendszerben a termés hozamok gyakoriság hisztogramját és annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása megegyezik a megadott Y értékek átlagával és standard szórásával! (3 pont)
- (k) Vizsgáljuk meg, hogy a termés hozamok normál eloszlást követnek-e $\alpha=0.05$ szignifikancia szinten, ha 10 egyenlő részre osztjuk a terjedelmet!
- (l) Adjuk meg a termés hozam átlagára 95%-os megbízhatóságú kétoldalú konfidencia intervallumot!

Teszteljük a

H_0 : az átlagos terméshozam hektáronként 5 tonna hipotézist 95% megbízhatósággal teljesül-e?

(m) Hány mérést kellene végezni, ha az átlag körül (plusz és minusz irányban) $L=0.25$ tonna/ha (tehát összesen 0.5 t/ha) hosszúságú konfidencia intervallumot szeretnénk kapni $\alpha = 0.05$ hibaszinten.

17.11.22. feladat

Megmérték 19 iskolás korú (11 évtől 15 éves korú) gyermekek X magasságát (inchben) és az Y testsúlyát (poundban)! Az alábbi adatokat kapták.

$X = [69.0, 56.5, 65.3, 62.8, 63.5, 57.3, 59.8, 62.5, 62.5, 59.0, 51.3, 64.3, 56.3, 66.5, 72.0, 64.8, 67.0, 57.5, 66.5]$

$Y = [112.5, 84.0, 98.0, 102.5, 102.5, 83.0, 84.5, 112.5, 84.0, 99.5, 50.5, 90.0, 77.0, 112.0, 150.0, 128.0, 133.0, 85.0, 112.0]$

(a) Ábrázolja az összetartozó magasság és testsúly értékeket a síkon a Statistics csomag ScatterPlot eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a magasság és a testsúly között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!

(d) Mekkora magasság növekedés esetén nő a testsúly 10 poundot?

(e) Számolja ki a megadott testsúly és az illesztett lineáris görbe alapján számolt testsúlyok közötti eltéréseket, majd rajzolja fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Számítsuk ki a testsúly legkisebb, legnagyobb értékét, terjedelmét, átlagát, korrigált szórását, móduszát és mediánját!

(i) Rajzolja fel a testsúly váltakozását az átlaggal és a kvartilisekkel együtt!

(j) Rajzoljuk fel közös koordináta-rendszerben a testsúly gyakoriság hisztogramját és annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása megegyezik a megadott testsúly átlagával és standard szórásával!

(k) Vizsgáljuk meg, hogy a testsúlyok normál eloszlást követnek-e $\alpha=0.05$ szignifikancia szinten, ha 10 egyenlő részre osztjuk a terjedelmet!

(l) Adjuk meg a testsúly átlagára 95%-os megbízhatóságú kétoldalú konfidencia intervallumot!

Teszteljük, a

H_0 : az átlagos testsúly 110 (pound)

hipotézist 95% megbízhatósággal teljesül-e?

(m) Hány gyerek testsúlyát kellene megvizsgálni, ha az átlag körül (plusz és minusz irányban) $L=8$ pound (tehát összesen 16 pound) hosszúságú konfidencia intervallumot szeretnénk kapni $\alpha = 0.05$ hibaszinten.

17.11.23. feladat

Az alábbi táblázat 15 ember láb (X) és kéz (Y) méretét mutatja inch mértékegységben.

$X = 9.00, 8.50, 9.25, 9.75, 9.00, 10.00, 9.50, 9.00, 9.25, 9.50, 9.25, 10.00, 10.00, 9.75, 9.50$

$Y = 6.50, 6.25, 7.25, 7.00, 6.75, 7.00, 6.50, 7.00, 7.00, 7.00, 7.50, 7.25, 7.25, 7.25$

(a) Ábrázolja az összetartozó lábak és kezek méreteit a síkon a Statistics csomag ScatterPlot eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a lábak és kezek mérete között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be

az egyenest a pontokhoz!

(d) Mekkora láb növekedés esetén nő a kéz 1 inch -el?

(e) Számolja ki a megadott kéz és az illesztett lineáris görbe alapján számolt kéz méretei közötti eltéréseket, majd rajzoljuk fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Számítsuk ki a kéz méretek legkisebb, legnagyobb értékét, terjedelmét, átlagát, korrigált szórását, móduszát és mediánját!

(i) Rajzolja fel a kéz méretek váltakozását az átlaggal és a kvartilisekkel együtt!

(j) Rajzoljuk fel közös koordináta-rendszerben a kéz méretek gyakoriság hisztogramját és annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása megegyezik a megadott kéz méretek átlagával és standard szórásával!

(k) Vizsgáljuk meg, hogy az kéz méretek normál eloszlást követnek-e $\alpha=0.05$ szignifikancia szinten, ha 10 egyenlő részre osztjuk a terjedelmet!

(l) Adjuk meg a kéz méretek átlagára 95%-os megbízhatóságú kétoldali konfidencia intervallumot!

Teszteljük, a

H_0 : az átlagos kézméret 7 (inch)

hipotézist 95% megbízhatósággal teljesül-e?

(m) Hány ember kezét kellene megvizsgálni, ha az átlag körül (plusz és mínusz irányban) $L=0.15$ inch (tehát összesen 0.3 inch) hosszúságú konfidencia intervallumot szeretnénk kapni $\alpha = 0.05$ hibaszinten.

17.11.24. feladat

A "Journal of Sound and Vibration" folyóirat (Vol. 151, 1991, pp. 383–394) számában tanulmányozták, hogy a zaj (X) mértéke a füleken keresztül az (Y) magas vérnyomásra növekvőleg hat-e? Az alábbi táblázat 20 különböző ember mérési adatait rögzíti, melyben X a hanghatás mértéke decibel-ben mérve és Y a megfelelő vérnyomás növekedés a hang hatására higanymilliméterben mérve!

$X = [60, 63, 65, 70, 70, 70, 80, 90, 80, 80, 85, 89, 90, 90, 90, 90, 94, 100, 100, 100]$

$Y = [1, 0, 1, 2, 5, 1, 4, 6, 2, 3, 5, 4, 6, 8, 4, 5, 7, 9, 7, 6]$

(a) Ábrázolja az összetartozó zaj és vérnyomás növekedés értékeit a síkon a Statistics csomag ScatterPlot eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a zaj és a vérnyomás növekedés között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!

(d) Mekkora zaj növekedés esetén nő a vérnyomás értéke 1 higanymilliméterrel?

(e) Számolja ki a megadott vérnyomás növekedés és az illesztett lineáris görbe alapján számolt vérnyomás növekedés közötti eltéréseket, majd rajzoljuk fel ezek hisztogrammját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Számítsuk ki a vérnyomás növekedés legkisebb, legnagyobb értékét, terjedelmét, átlagát, korrigált szórását, móduszát és mediánját!

(i) Rajzolja fel a vérnyomás növekedés váltakozását az átlaggal és a kvartilisekkel együtt!

(j) Rajzoljuk fel közös koordináta-rendszerben a vérnyomás növekedés gyakoriság

hisztogramját és annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása megegyezik a megadott vérnyomás növekedés átlagával és standard szórásával!

(k) Vizsgáljuk meg, hogy a vérnyomás növekedések normál eloszlást követnek-e $\alpha=0.05$ szignifikancia szinten, ha 10 egyenlő részre osztjuk a terjedelmet!

(l) Adjuk meg a vérnyomás növekedés átlagára 95%-os megbízhatóságú kétoldalú konfidencia intervallumot!

Teszteljük, a

H_0 : az átlagos vérnyomás növekedés 3 (higanymilliméter)

hipotézist 95% megbízhatósággal teljesül-e?

(m) Hány ember vérnyomás növekedését kellene megvizsgálni, ha az átlag körül (plusz és mínusz irányban) $L=1$ higanymilliméter (tehát összesen 2 higanymilliméter) hosszúságú konfidencia intervallumot szeretnénk kapni $\alpha = 0.05$ hibaszinten.

17.11.25. feladat

A testsúly (X) és a szisztolés vérnyomás (Y) adatait mérték 26 véletlenszerűen választott férfi esetén, akiknek a kora 25 év 30 év között volt. Feltételezzük, hogy a testsúly és a vérnyomás normál eloszlású. Az eredményeket az alábbi táblázat mutatja

Test súly (X) kg-ban: 82.5, 83.5, 90, 77.5, 106, 87.5, 95, 105, 100, 74.5, 79, 84.5, 85, 86, 79.5, 84, 87, 91.5, 107.5, 97.5, 90, 71.5, 120, 117.5, 96, 93.5

Szisztolés vérnyomás (Y) Hg mm-ben 130, 133, 150, 128, 151, 146, 150, 140, 148, 125, 133, 135, 150, 153, 128, 132, 149, 158, 150, 163, 156, 124, 170, 165, 160, 159

(a) Ábrázolja a pontokat a Statistics csomag ScatterPlot eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a testsúly és a vérnyomás között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzolja be az egyenest a pontokhoz!

(d) Átlagosan mekkora testsúly növekedés esetén növekszik a vérnyomás 1 Hgmm-t? (1 pont)

(e) Számolja ki a mérési adatok és az illesztett lineáris görbe közötti eltéréseket, majd rajzolja fel ezek hisztogramját! (2 pont)

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét! (2 pont)

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $> \alpha; \alpha=0.05$ szignifikancia szinten?

17.11.26. feladat

Egy utazó ügynök az év 10 hónapját a termékeinek eladásával töltötte.

Feljegyezte minden hónapban, hogy mennyi kilométert (X változó) tett meg és mennyi volt az üzletkötésből származó bevétele (Y változó) ezer Ft.-ban.

Az adatok a következők

X:=[1370, 1050, 980, 1770, 1340, 1560, 2110, 1540, 1480, 1670]

Y:=[2240, 1800, 1900, 2200, 2300, 2300, 2500, 2300, 2100, 2200]

(a) Vizsgáljuk meg a lineáris korreláció szorosságát az utazott km-ek mennyisége és a bevétel mértéke között!

(b) Számítsuk ki az $Y = a + b \cdot X$ lineáris trend 'a' és 'b' együtthatóit!

(c) Rajzoljuk fel a pontokat és az egyenest azonos koordináta-rendszerben!

(c) Vizsgáljuk meg, hogy a mért $Y[k]$ adatok és a lineáris illesztésből számolt Y adatok eltérése normál eloszlást követ-e nulla várható értékkel 95%-os megbízhatósági szinten!

Rajzoljuk fel az eltérések hisztogramját!

17.11.27. feladat

Egy kísérlet során feltételezték, hogy az anyagból szárítás segítségével távozó anyag mennyisége (Y) gramokban összefügg a szárítási idővel (X, órában). Mérték 10 szárítással

kapcsolatos anyag csökkenési tömeget gramokban és mérték a szárítási időt órákban. Az adatok a következők

$X = [2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0]$

$Y = [4.3, 1.5, 1.8, 4.9, 4.2, 4.8, 5.8, 6.2, 7.0, 7.9]$

(a) Vizsgáljuk meg a lineáris korreláció szorosságát a szárítási idő és az eltávozott anyag mennyisége között!

(b) Számítsuk ki az $Y = a + b \cdot X$ lineáris trend 'a' és 'b' együtthatóit!

(c) Rajzoljuk fel a pontokat és az egyenest azonos koordinátarendszerben!

(c) Vizsgáljuk meg, hogy a mért $Y[k]$ adatok és a lineáris illesztésből számolt Y adatok eltérése normál eloszlást követ-e nulla várható értékkel 95%-os megbízhatósági szinten! Rajzoljuk fel az eltérések hisztogramját!

17.11.28. feladat

Az alábbi táblázat 24 ház eladását és az eladás után éves szinten befizetett adó mértékét tartalmazza ezer dollárban.

Eladási árak(ezer \$-ban) $X = [25.9, 29.5, 27.9, 25.9, 29.9, 29.9, 30.9, 28.9, 35.9, 31.5, 31.0, 30.9, 30.0, 36.9, 41.9, 40.5, 43.9, 37.5, 37.9, 44.5, 37.9, 38.9, 36.9, 45.8]$

Éves adó (ezer \$-ban) : $Y = [4.9176, 5.0208, 4.5429, 4.5573, 5.0597, 3.8910, 5.8980, 5.6039, 5.8282, 5.3003, 6.2712, 5.9592, 5.0500, 8.2464, 6.6969, 7.7841, 9.0384, 5.9894, 7.5422, 8.7951, 6.0831, 8.3607, 8.1400, 9.1416]$

a) Ábrázolja a házak eladási árát és a befolyt adó mértékének pontjait a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a házak ára és a befizetett adó között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!

(d) Mekkora ház esetén lesz az adó mértéke 7 ezer \$?

(e) Számolja ki a megadott adók és az illesztett lineáris görbe alapján számolt adók közötti eltéréseket, majd rajzoljuk fel ezek hisztogramját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Számítsuk ki az adó legkisebb, legnagyobb értékét, terjedelmét, átlagát, korrigált szórását, móduszát és mediánját!

(i) Számítsuk ki azt az adót, amelynél kevesebbet 6-an fizettek illetve melynél többet 6-an fizettek (az alsó és felső kvartilishoz, használja a Quartile eljárást)!

(j) Rajzolja fel a befizetett adók váltakozását az átlaggal és a kvartilisekkel együtt!

(k) Rajzoljuk fel közös koordinátarendszerben az adók gyakoriság hisztogramját és annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása megegyezik a megadott adók átlagával és standard szórásával!

(l) Vizsgáljuk meg, hogy az adófizetés normál eloszlást követ-e $\alpha=0.05$ szignifikancia szinten, ha 12 egyenlő részre osztjuk a terjedelmet!

(m) Adjuk meg az adófizetés átlagára 95%-os megbízhatóságú kétoldalú konfidencia intervallumot!

Teszteljük, az előírt

H_0 : várt adó 7 (ezer dollár)

hipotézist 95% megbízhatósággal teljesül-e?

(n) Mennyi adó fizetést kellene megvizsgálni, ha az átlag körül (plusz és minusz irányban) $L=$

0.5 ezer \$ (tehát összesen 1 ezer \$) hosszúságú konfidencia intervallumot szeretnénk kapni $\alpha = 0.05$ hibaszinten.

17.11.29. feladat

Az alábbi táblázat 26 mérési adatpárt tartalmaz a szezonális átlagos talajvíz szintre (X) [méterben] és a hozzá tartozó átlagos terméshozamra (Y) [tonna/hektárban] vonatkozóan.
 $X = [0.15, 0.20, 0.20, 0.25, 0.25, 0.32, 0.40, 0.40, 0.40, 0.45, 0.45, 0.47, 0.50, 0.50, 0.60, 0.65, 0.65, 0.70, 0.70, 0.75, 0.90, 0.90, 0.91, 1.00, 1.05, 1.08]$

$Y = [4.0, 4.5, 3.0, 4.0, 3.7, 3.5, 5.0, 4.5, 4.5, 4.8, 4.5, 5.5, 5.2, 4.0, 4.5, 6.0, 4.5, 5.7, 5.0, 5.3, 5.5, 4.7, 5.0, 4.5, 5.7, 5.5]$

(a) Ábrázolja az összetartozó talajvíz szint és az átlagos terméshozam értékeit a síkon a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számolja ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a talajvíz és a terméshozam között?

(c) Illesszen lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!

(d) Mekkora talajvíz növekedés esetén nő a terméshozam 1 tonna/ha-ral?

(e) Számolja ki a megadott terméshozamok és az illesztett lineáris görbe alapján számolt eltéréseket, majd rajzoljuk fel ezek hisztogramját!

(f) Számolja ki az eltérések átlagát és szórását! Rajzolja be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Számítsuk ki a terméshozamok (Y) legkisebb, legnagyobb értékét, terjedelmét, átlagát, korrigált szórását, móduszát és mediánját!

(i) Rajzolja fel a terméshozamok váltakozását az átlaggal és a kvartilisekkel együtt!

(j) Rajzoljuk fel közös koordináta-rendszerben a terméshozamok gyakoriság hisztogramját és annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása megegyezik a megadott Y értékek átlagával és standard szórásával!

(k) Vizsgáljuk meg, hogy a terméshozamok normál eloszlást követnek-e $\alpha=0.05$ szignifikancia szinten, ha 10 egyenlő részre osztjuk a terjedelmet!

(l) Adjuk meg a terméshozam átlagára 95%-os megbízhatóságú kétoldalú konfidencia intervallumot!

Teszteljük, a

H_0 : az átlagos terméshozam hektáronként 5 tonna

hipotézist 95% megbízhatósággal teljesül-e?

(m) Hány mérést kellene végezni, ha az átlag körül (plusz és minusz irányban) $L=0.25$

tonna/ha (tehát összesen 0.5 t/ha) hosszúságú konfidencia intervallumot szeretnénk kapni $\alpha = 0.05$ hibaszinten.

17.11.30. feladat

Mérték a vas darabok sűrűségét (X) és vas tartalmát (Y). Az eredményeket az alábbi táblázat mutatja

Sűrűség (X) : 2.8, 2.9, 3.0, 3.1, 3.2, 3.2, 3.2, 3.3, 3.4

Vas tartalom (Y) 27, 23, 30, 28, 30, 32, 34, 33, 30

(a) Ábrázoljuk a pontokat a **Statistics** csomag **ScatterPlot** eljárásával!

(b) Számoljuk ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a vas tartalom és a sűrűség között?

(c) Illesszünk lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!

(d) Vajon mennyi a vas tartalma annak a darabnak, melynek sűrűsége 4, ha a tendencia

változatlan?

(e) Számoljuk ki a mérési adatok és az illesztett lineáris görbe közötti eltéréseket, majd rajzoljuk fel ezek hisztogrammját!

(f) Számoljuk ki az eltérések átlagát és szórását! Rajzoljuk be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

17.11.31. feladat

Egy kémiai desztillációs folyamatban mérték a folyamat eredményeként kapott oxigén tisztaságát (Y, %-ban) és a desztillátorban jelenlevő szénhidrogén szintet (X). A megfelelő mérési eredményeket megadtuk az alábbi két listában.

X:=[0.99, 1.02, 1.15, 1.29, 1.46, 1.36, 0.87, 1.23, 1.55, 1.40, 1.19, 1.15, 0.98, 1.01, 1.11, 1.20, 1.26, 1.32, 1.43, 0.95]

Y:=[90.01, 89.05, 91.43, 93.74, 96.73, 94.45, 87.59, 91.77, 99.42, 93.65, 93.54, 92.52, 90.56, 89.54, 89.85, 90.39, 93.25, 93.41, 94.98, 87.33]

(a) Ábrázoljuk a pontokat a Statistics csomag ScatterPlot eljárásával!

(b) Számoljuk ki X és Y között a korrelációs együtthatót! Milyen kapcsolatot mutat a korrelációs együttható a szénhidrogén szint és a kapott oxigén tisztasága között?

(c) Illesszünk lineáris görbét a legkisebb négyzetek módszerével a pontokra, majd rajzoljuk be az egyenest a pontokhoz!

(d) Vajon Y=1% oxigén tisztaság növekedést átlagosan mennyi X szénhidrogén szint változás eredményez?

(e) Számoljuk ki a mérési adatok és az illesztett lineáris görbe közötti eltéréseket, majd rajzoljuk fel ezek hisztogrammját!

(f) Számoljuk ki az eltérések átlagát és szórását! Rajzoljuk be az eltérések hisztogramjának ábrájába a 0 várható értékű és az eltérésekkel azonos szórású normál eloszlás sűrűség függvényét!

(g) Vajon normál eloszlást követnek-e az eltérések nulla várható értékkel $\alpha=0.05$ szignifikancia szinten?

(h) Határozzunk meg az oxigén tisztaság átlag értékére 95%-os megbízhatósággal kétoldali konfidencia intervallumot! A feltételezett 93% -os (közepes) szint benne van-e ebben a megbízhatósági intervallumban?