

▼ 14. Statisztikai minta. Empirikus eloszlásfüggvény. Pontbecslések várható értékre és szórásra.

▼ 14.1. Bevezetés

A statisztika a mérési adatok ismeretében következtetéseket fogalmaz meg a vizsgált folyamatról, amely a véletlent is magában foglalja. Például ha el kell döntenünk egy pénz érméről, hogy szabályos-e, akkor feldobjuk néhányszor, majd teszteljük a $H_0 : p = \frac{1}{2}$ hipotézist a $H_1 : p \neq \frac{1}{2}$

alternatív hipotézissel szemben $\alpha = 0.05$ szignifikancia szinten. De használhatjuk a nagy számok törvényének Bernoulli-alakját is a kérdés eldöntésére. Vagy felrajzoljuk a relatív gyakoriságok grafikonját és a szemlélet alapján döntünk. Látható, hogy a statisztika használja azokat az elméleti eredményeket és modelleket, amelyek a valószínűségszámítás témakörébe tartoznak. A fordított irányú probléma gyakrabban fordul elő. Nevezetesen a statisztika olyan kérdéseket tesz fel, amelyek megválaszolása a valószínűségszámítástól új eszközök kifejlesztését igényli. Tehát a valószínűségelmélet fejlődését a statisztikai kérdések elméleti tisztázása hajtja előre.

Mivel a statisztikai adatainkban a véletlenek megjelennek, ezért a statisztika a döntéseivel felvállalja azt a kellemetlenséget, hogy tévedhet. Különösen veszélyes lehet ez a tévedés akkor, ha az adatokat "manipuláljuk".

Ebben a fejezetben leíró statisztika alapjait tárgyaljuk. Ezek a minta átlaga, korrigált szórása, módusza, mediánja és kvartilisei. Olyan grafikus lehetőségeket ad, mint a gyakoriság hisztogram, empirikus eloszlás függvény, box diagramja. Ezek a fogalmak együttesen jellemzik adatainkat. Az adatokat részletes statisztikai analízis alá kell vetni, hogy a jellemző paraméterei alapján jellemezni tudjuk. A hisztogram alapján az adatok relatív gyakoriságait közelítjük valószínűségi változókkal vagy diszkréttel vagy folytonossal. El kell döntenünk, hogy az adatok diszkrét valószínűségi változókkal modellezhetők vagy folytonossal. Nagy elemszámú minta esetén elfordulhat, hogy diszkrét értékeket felvevő változót mégis folytonossal közelítünk.

Célunk az alábbi fogalmak, összefüggések és eljárások megismerése, megértése	
A statisztikai sokaság és a minta fogalma	Módusz meghatározása
A statisztika fogalma	Medián meghatározása és minimum tulajdonsága
A várható érték torzítatlan és konzisztens becslése	Kvartilisek és a box plot diagram
A várható érték minimum tulajdonsága	Korrigált szórás négyzet, mint a variancia torzítatlan becslése
Gyakoriság hisztogram	Az adatok eloszlásának modell tesztje
Empirikus eloszlás függvény	A Statistics csomag <i>Range</i> , <i>FrequencyTable</i> , <i>FrequencyPlot</i> , <i>ChiSquareSuitableModelTest</i>

▼ 14.2. A valószínűségszámítás és a statisztika kapcsolata

A valószínűségszámítás a matematika azon ága, amely a véletlen jelenségek **absztrakt**

modellezésével és a megfelelő **matematikai következtetések** levezetésével foglalkozik, és teszi mindezt **axiomatikus alapon**. Érdeklődésünk homlokterében az eddigi vizsgálatok során az állt, hogyan válasszuk ki a megfelelő valószínűségi modellt (vagy eloszlást), amely leírja a vizsgált fizikai vagy természeti folyamatot és a leírás segítségével megpróbáltuk feltárni a folyamat valószínűségi törvényeit!

Ezzel ellentétben illetve összhangban a statisztika a fenti elmélet **valós problémákra** való alkalmazásával foglalkozik, melynek kiindulási pontja a **megfigyelt vagy mért adatsor**.

A statisztika a megfigyelési vagy **mérési adatok** ismeretében foglalkozik

- (1) a megfelelő **valószínűségi modell** megválasztásával;
- (2) a modellben szereplő **paraméterek becslésével** és a becslés **megbízhatóságával**;
- (3) statisztikai **következtetések** levonásával.

▼ 14.3. A statisztikai minta

14.3.1. DEFINÍCIÓ. SOKASÁG

Populációnak vagy **sokaságnak** nevezzük homogén egyedek nagy számú együttesét, melyek mérhetők, azaz jellemezhetők valós számmal!

14.3.2. PÉLDA

Csapágyak gyártása során olyan **azonos** belső és külső átmérőkkel rendelkező csapágyak alkothatnak sokaságot, amelyet pl. egy műszak alatt készítettek.

14.3.3. DEFINÍCIÓ. MINTA

Mintának nevezzük a sokaságból **véletlenszerűen** kivett néhány egyedet.

A véletlenszerű kiválasztás azt jelenti, hogy nincs összefüggés vagy kapcsolat az egyik egyed és a másik egyed kiválasztása között.

14.3.4. PÉLDA

A 14.3.2. példa esetén a leggyártott sok csapágy közül kivesszünk $n = 50$ csapágyat, amelyek alkotják a mintát!

14.3.5. DEFINÍCIÓ. A MINTA ÉRTÉKEI

A kiválasztott **minta** elemeihez tartozó **valós számokat**, amellyel a sokaságot jellemezzük, nevezzük a **minta értékeinek**.

14.3.6. PÉLDA

A kiválasztott minta csapágyak belső és külső átmérőit megmérjük és ezek a csapágy minta értékei. Ha másik mintát veszünk ki, akkor a minta értékek megváltozhatnak. Ezért minden kiválasztás egy változó!

Az X populációból vett **mintát** $X_1, X_2, \dots, X_n$ valószínűségi változók véges sorozatával tudjuk megadni, amelyekre az alábbiak teljesülnek

- (1) az $X_1, X_2, \dots, X_n$ **független** valószínűségi változók: azaz

$$f_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_n}(x_n) : \text{az együttes eloszlás az egyes változók eloszlásainak szorzata} ;$$

- (2) az $X_1, X_2, \dots, X_n$ azonos **eloszlású** valószínűségi változók:

$$f_{X_k}(x) = f_X(x) , \quad (k = 1, 2, \dots, n) \text{ mindegyik eloszlás a sokaság eloszlásával azonos;}$$

ahol

$f_{X_k}(x)$ az X_k valószínűségi változó sűrűség függvénye folytonos esetben illetve eloszlása diszkrét esetben!

14.3.7. DEFINÍCIÓ. STATISZTIKA

Az $X_1, X_2, \dots, X_n$ **minta** tetszőleges

$$Y_n = h_n(X_1, X_2, \dots, X_n)$$

függvényét statisztikának nevezzük. Az Y egy új valószínűségi változót jelöl, mely a számítás eredménye.

A statisztika értéke kiszámítható a minta értékei alapján és nem szerepelnek benne a populáció paraméterei!

Ha a populációból egy másik mintát veszünk, akkor általában az Y statisztika értéke is megváltozik.

▼ 14.4. A sokaság várható értéke és a minta átlaga. Torzítatlanság és konzisztencia.

Jelöljön X egy sokaságot, melynek eloszlásfüggvénye $F(x)$ és eloszlása (diszkrét esetben) vagy sűrűségfüggvénye (folytonos esetben) $p(x)$. Az X valószínűségi változó alábbi jellemzőivel foglalkozunk illetve ezek statisztikai becslésével.

Várható érték= μ	Szórás= σ	Medián= m	Módusz= a
$\mu = E(X)$	$\sigma^2 = E((X - \mu)^2)$	$P(X \leq m)$ $= \frac{1}{2}$ $P(X \geq m)$ $= \frac{1}{2}$	$p(a) = \max p(x)$ $p(x) =$ diszkrét eloszlás vagy sűrűség függvény
Súlyozott átlag	Négyzetes eltérések átlaga	Közép érték	Legnagyobb valószínűségi érték

14.4.1. DEFINÍCIÓ. MINTAKÖZÉP vagy ÁTLAG vagy VÁRHATÓ ÉRTÉK

Az X sokaságból vett $X_1, X_2, \dots, X_n$ véletlen minta

$$\bar{X} = X_n^{\text{átlag}} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^n X_i = h_n(X_1, X_2, \dots, X_n)$$

statisztikát a **minta átlagának** nevezzük.

Számítsuk ki a **minta átlagának**, mint valószínűségi változónak a **várható értékét** és **varianciáját**, ha $\mu = E(X_k)$ és $\sigma^2 = \text{Var}(X_k)$ minden $k=1,2,\dots,n$ mellett!

Mivel a várható érték lineáris függvény, ezért

$$E(\bar{X}) = E(X_n^{\text{átlag}}) = \frac{1}{n} \sum_{k=1}^n E(X_k) = \frac{1}{n} \sum_{k=1}^n \mu = \frac{n\mu}{n} = \mu.$$

14.4.2. DEFINÍCIÓ. Torzítatlan becslés. Torzítás mértéke.

Azt mondjuk, hogy a $\widehat{\theta}_n = h_n(X_1, X_2, \dots, X_n)$ statisztika a θ paraméter **torzítatlan becslése**, ha

$$E(\widehat{\theta}_n) = E(h_n(X_1, X_2, \dots, X_n)) = \theta$$

teljesül. Ellenkező esetben a $h_n(X_1, X_2, \dots, X_n)$ statisztikai becslés **torzított**.

Egy torzított $\widehat{\theta}_n$ becslés esetén a

$$\text{torzítás}(\widehat{\theta}_n) = E(h_n(X_1, X_2, \dots, X_n) - \theta)$$

eltérés várható értéke a **torzítás mértéke**.

A fenti levezetés értelmében $E(X_n^{\text{átlag}}) = \mu$, ezért igaz az alábbi tétel.

14.4.3. TÉTEL. Az átlag a várható érték torzítatlan becslése

Az $X_n^{\text{átlag}}$ változó a μ várható érték torzítatlan becslése

$$E(X_n^{\text{átlag}}) = \mu$$

tetszőleges $n = 1, 2, 3, \dots$ esetén.

Torzított becslést bizonyos esetben korrigálni tudunk.

14.4.4. PÉLDA. Lineáris torzítás korrekciója

Ha a torzítás mértéke $\text{torzítás}(\widehat{\theta}_n) = c \cdot \theta$ lineáris, akkor a $\widetilde{\theta}_n = \frac{\widehat{\theta}_n}{c+1}$ **korrigált becslés** már torzítatlan a θ paraméterre.

Bizonyítás

A várható érték linearitása miatt

$$E(\widetilde{\theta}_n) = E\left(\frac{\widehat{\theta}_n}{c+1}\right) = \frac{1}{c+1} \cdot E(\widehat{\theta}_n) = \frac{1}{c+1} \cdot (E(\widehat{\theta}_n - \theta) + \theta) = \frac{1}{c+1} \cdot (c \cdot \theta + \theta) = \theta.$$

Tehát a $\widetilde{\theta}_n$ statisztika a θ paraméter torzítatlan becslése. *Q.e.d.*

Ilyen esetben a $\widetilde{\theta}_n$ becslést a $\widehat{\theta}_n$ becslés **torzítást-korrigáló** verziójának nevezzük.

Azt mondjuk, hogy egy $h_n(X_1, \dots, X_n)$ becslés torzítatlan a $g(\theta)$ függvényre nézve, ha

$$E(h_n(X_1, X_2, \dots, X_n)) = g(\theta).$$

A fentiekből következik, hogy ha $\widehat{\theta}_n$ torzítatlan becslése a θ paraméternek, akkor általában nem biztos, hogy $g(\widehat{\theta}_n)$ torzítatlan becslése a $g(\theta)$ paraméternek, még akkor sem, ha g lineáris függvény.

Két torzítatlan becslés közül az a jobbik, amelynek a varianciája vagy szórása kisebb.

Mutassuk meg, hogy az $X_{n+1}^{\text{átlag}}$ becslés jobb, mint a $X_n^{\text{átlag}}$ becslés a μ várható értékre.

14.4.5. PÉLDA.

Az $X_{n+1}^{\text{átlag}}$ becslés jobb, mint a $X_n^{\text{átlag}}$ becslés a μ várható értékre.

Bizonyítás

Mivel az $X_1, X_2, \dots, X_n$ valószínűségi változók függetlenek, ezért a varianciánál tanultak alapján

$$\text{Var}(X_n^{\text{átlag}}) = \text{Var}\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) = \frac{1}{n^2}(\text{Var}(X_1) + \dots + \text{Var}(X_n)) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Tehát

$$\text{Var}(X_{n+1}^{\text{átlag}}) = \frac{\sigma^2}{n+1} < \text{Var}(X_n^{\text{átlag}}) = \frac{\sigma^2}{n}. \text{ Q.e.d.}$$

A fentiekből következik, hogy

$$\lim_{n \rightarrow \infty} \text{Var}(X_n^{\text{átlag}}) = \lim_{n \rightarrow \infty} \frac{\sigma^2}{n} = 0.$$

14.4.6. DEFINÍCIÓ. Konzisztens becslés sorozat

A $\hat{\theta}_n = h_n(X_1, X_2, \dots, X_n)$ becslések sorozatát a θ paraméter **konzisztens becslésének nevezzük**, ha $\hat{\theta}_n$ valószínűségi mértékben konvergál a θ paraméterhez, azaz bármely $\epsilon > 0$ esetén

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

A Csebisev-egyenlőtlenség alapján kapjuk, hogy az átlag konzisztens becslése a várható értéknek.

14.4.7. TÉTEL. Az átlag a várható érték konzisztens becslése

Az $X_n^{\text{átlag}}$ változó a μ várható érték konzisztens becslése

$$\lim_{n \rightarrow \infty} P(|X_n^{\text{átlag}} - \mu| > \epsilon) = 0.$$

Bizonyítás

A Csebisev-egyenlőtlenség alapján

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\text{Var}(X_n^{\text{átlag}})}{\epsilon^2} = \frac{\sigma^2}{n \cdot \epsilon^2} \xrightarrow{n \rightarrow \infty} 0 \quad \text{Q.e.d.}$$

14.4.8. TÉTEL. A VÁRHATÓ ÉRTÉK MINIMUM TULAJDONSÁGA

Az X valószínűségi változó $\mu = E(X)$ várható értéke a c -re vonatkozó $\min(E((X - c)^2))$ minimum feladat egyetlen megoldása, tehát

$$E((X - \mu)^2) = \min_{c \in R} (E((X - c)^2))$$

Bizonyítás A jobb oldali kifejezésen belül vonjunk ki μ -t és adjunk is hozzá!

$$\begin{aligned} E((X - c)^2) &= E((X - \mu + \mu - c)^2) = / \text{emeljük négyzetre az összeget}/ \\ &= E((X - \mu)^2 + 2(X - \mu)(\mu - c) + (\mu - c)^2) = / \text{használjuk a várható érték lineáris tulajdonságát}/ \\ &= E((X - \mu)^2) + 2(\mu - c)E(X - \mu) + (\mu - c)^2 = / \text{mivel } E(X) = \mu, \text{ ezért } E(X - \mu) = E(X) - \mu = \\ &\mu - \mu = 0/ = E((X - \mu)^2) + (\mu - c)^2 >= E((X - \mu)^2). \quad / \text{és egyenlőség csak } c = \mu \text{ esetén áll}/ \end{aligned}$$

Tehát $E((X - c)^2)$ legkisebb értékét c -ben a $c = \mu$ helyen veszi fel! Q.e.d.

▼ 14.5. Példa: a kohó csapolások adatsora

14.5.1. PÉLDA. Kohó csapolási adatok leíró statisztikája

Egy kohóban feljegyezték egy adott hónap alatt csapolt acél mennyiségeket tonnában és a következő táblázatot vagy adatsort kapták.

[144, 143, 139, 161, 134, 135, 157, 158, 144, 138, 155, 162, 151, 134, 141, 121, 137, 162, 133, 147, 125, 154, 154, 150, 145, 157, 165, 157, 152, 164, 145, 187, 153, 144, 175, 150, 162, 158, 174, 150, 161, 165, 163, 124,

142, 152, 183, 141, 146, 172, 148, 133, 151, 164, 153, 176, 160, 148, 136, 160, 154, 148, 161, 149, 147, 161, 158, 155, 175, 137, 147, 132, 129, 154, 161, 132, 168, 147, 159, 118, 179, 162, 143, 151, 158, 157, 156, 146, 146, 157, 136, 161, 166, 152, 167, 160, 137, 156, 145, 184, 162, 153, 146, 144, 155, 139, 173, 151, 145, 166, 141, 148, 161, 155, 142, 150, 159, 166, 173, 171, 127, 164, 163, 159, 148, 143, 148, 149, 155, 131, 139, 142, 156, 149, 154, 163, 149, 164, 142, 136, 146, 153, 154, 144, 172, 150, 143, 150, 169, 157, 164, 152, 165, 116, 177, 169, 168, 138, 143, 160, 137, 140, 164, 142, 132, 176, 138, 165, 167, 158, 153, 162, 134, 153, 156, 137, 169, 161, 172, 143, 171, 158, 125, 170, 153, 170, 152, 145, 154, 134, 156, 153, 146, 138, 173, 179, 156, 154, 147, 180, 145, 165, 170, 127, 135, 156, 157, 174, 167, 169, 174, 150, 152, 137, 166, 142]

(a) Számoljuk ki a terjedelmet és az átlagot!

(b) Rajzoljuk fel az adatsort!

Megoldás

> *restart* :

> *csapolas* := [144, 143, 139, 161, 134, 135, 157, 158, 144, 138, 155, 162, 151, 134, 141, 121, 137, 162, 133, 147, 125, 154, 154, 150, 145, 157, 165, 157, 152, 164, 145, 187, 153, 144, 175, 150, 162, 158, 174, 150, 161, 165, 163, 124, 142, 152, 183, 141, 146, 172, 148, 133, 151, 164, 153, 176, 160, 148, 136, 160, 154, 148, 161, 149, 147, 161, 158, 155, 175, 137, 147, 132, 129, 154, 161, 132, 168, 147, 159, 118, 179, 162, 143, 151, 158, 157, 156, 146, 146, 157, 136, 161, 166, 152, 167, 160, 137, 156, 145, 184, 162, 153, 146, 144, 155, 139, 173, 151, 145, 166, 141, 148, 161, 155, 142, 150, 159, 166, 173, 171, 127, 164, 163, 159, 148, 143, 148, 149, 155, 131, 139, 142, 156, 149, 154, 163, 149, 164, 142, 136, 146, 153, 154, 144, 172, 150, 143, 150, 169, 157, 164, 152, 165, 116, 177, 169, 168, 138, 143, 160, 137, 140, 164, 142, 132, 176, 138, 165, 167, 158, 153, 162, 134, 153, 156, 137, 169, 161, 172, 143, 171, 158, 125, 170, 153, 170, 152, 145, 154, 134, 156, 153, 146, 138, 173, 179, 156, 154, 147, 180, 145, 165, 170, 127, 135, 156, 157, 174, 167, 169, 174, 150, 152, 137, 166, 142] :

Számoljuk meg a **minta** **N** **elemszámát** a **Statistics** csomag **Count** eljárásával!

> *with(Statistics)* :

> *N := Count(csapolas)*

N := 216

(1.5.1)

A vizsgált hónapban összesen 216-szor csapoltak. Számítsuk ki a csapolt mennyiségek **minimális** és **maximális** értékét. A kettő különbségét nevezzük a **minta terjedelmének**, melyet a **Statistics** csomag **Range** eljárásával is kiszámíthatunk.

> *A := min(csapolas); B := max(csapolas);*
terjedelem := trunc(Range(csapolas));

A := 116

B := 187

terjedelem := 71

(1.5.2)

Tehát az adott hónapban a legkevesebb csapolás 116 tonna volt és a legtöbb 187 tonna. Így a csapolás terjedelme $71 = (187 - 116)$ tonna, amely az ingadozás mértékét jellemzi.

Számítsuk ki a minta $\bar{X} = \frac{\sum_{k=1}^N x_k}{N}$ **átlagát** a **Statistics** csomag **Mean** eljárásával!

> *átlag := Mean(csapolas)*

átlag := 152.944444444444

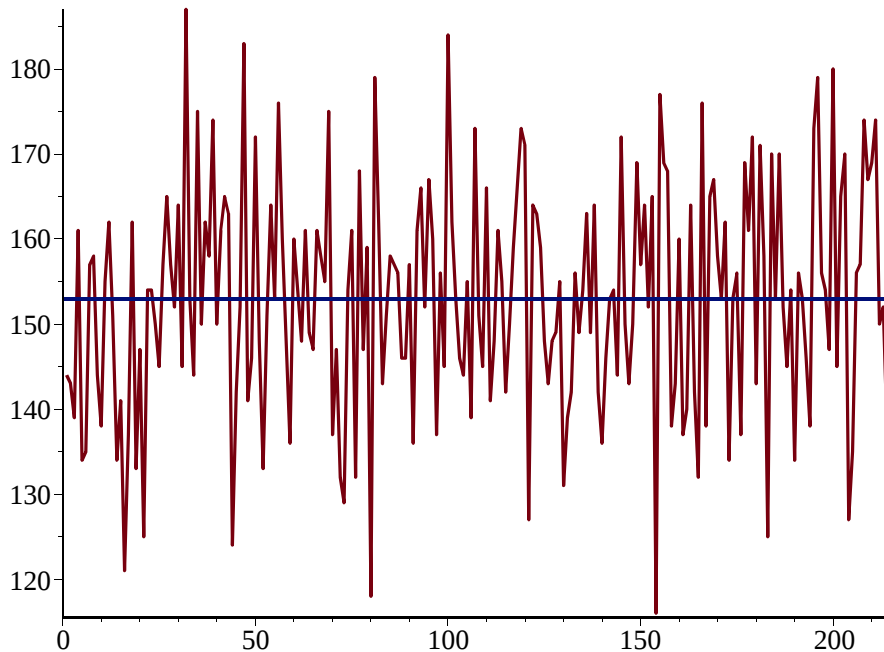
(1.5.3)

Tehát a csapolások átlag értéke az adott hónapban 152.9 tonna. Rajzoljuk fel a csapolás időbeli

váltakozását és az átlag értéket azonos koordináta-rendszerben! Az ilyen módon rajzolt értékeket idősortnak is nevezzük.

> `plot([seq([k, csapolas[k]], k = 1..N)], átlag, 0..N, title
= A csapolások mennyiségeinek egymásutáni váltakozása)`

A csapolások mennyiségeinek egymásutáni váltakozása



▼ 14.6. Gyakoriság értékek és a hisztogramok. Módusz.

14.6.1. DEFINÍCIÓ. GYAKORISÁG

Legyenek adva az "a" és "b" értékek a $\min(x_1, x_2, \dots, x_n) \leq a \leq b \leq \max(x_1, x_2, \dots, x_n)$ határok között! Ekkor a

$$\{k : a \leq x_k < b\}$$

halmaz elemeinek száma megadja, hogy a minta értéke milyen gyakorisággal esik az **[a,b)** balról zárt és jobbról nyitott intervallumba! Az $a = b$ esetben egyetlen érték gyakoriságát adja meg a felírt halmaz elemeinek száma.

14.6.2. PÉLDA. A kohó csapolási példa folytatása

(a) Rajzoljuk fel az adatsort és vizsgáljuk az $[a, b) = [135, 140)$ intervallumba eső csapolási értékek számát!

(b) Rajzoljuk fel a gyakoriság hisztogramot!

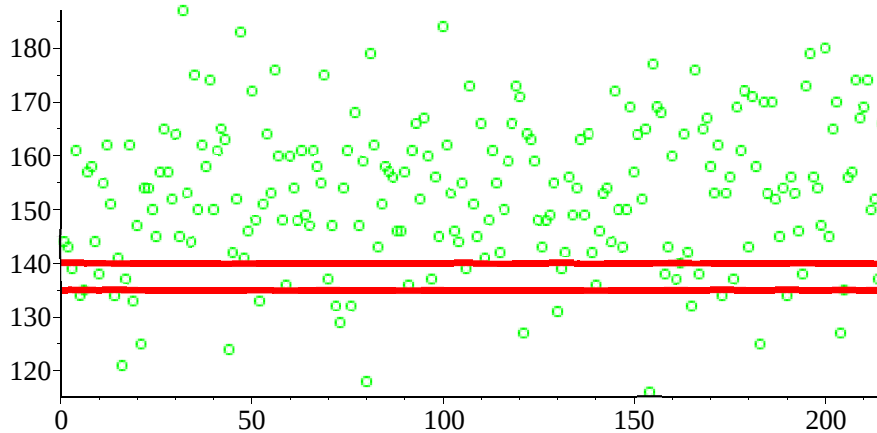
Megoldás

> `pontok := plot([seq([k, csapolas[k]], k = 1..N)], style = point, symbol = circle, color`

= green) :

> *hatarok := plot([135, 140], 0 ..N, thickness = 3, color = red) :*
plots[display]([pontok, hatarok], title
= A csapolások mennyiségek gyakorisága egy adott intervallumban)

A csapolások mennyiségek gyakorisága egy adott intervallumban



A **Statistics** csomag **FrequencyTable** utasítása kiszámolja a megadott intervallumokba eső minta adatok gyakoriságát, relatív gyakoriságát és ezek halmozott összegeit! Az intervallumok hosszát a **bins** opcióval szabályozhatjuk, amely megadja a részintervallumok számát, ahány egyenlő részre kell osztani a minta terjedelmét.

> *gyak := FrequencyTable(Array(csapolas), bins = terjedelem)*

gyak := $\left[\begin{array}{l} 1..71 \times 1..5 \text{ Array} \\ \text{Data Type: anything} \\ \text{Storage: rectangular} \\ \text{Order: Fortran_order} \end{array} \right]$ **(1.6.1)**

Mivel a **bins** = a terjedelem, ezért az intervallumok hossza 1 lesz és ezekbe egyetlen egész szám esik. (Ugyanis a csapolt mennyiségek egész tonnában vannak adva!)

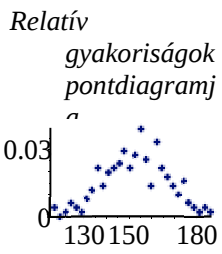
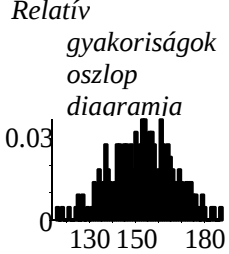
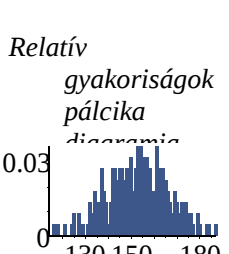
Irassuk ki az **intervallumokat**, a **gyakoriságokat** és a **halmozott relatív gyakoriságokat**!

> *fejléc := [Intervallumok, Gyakoriságok, Halmozott relatív gyakoriságok] :*

> *Gyakoriság értékek := matrix(terjedelem + 1, 3, [fejléc, seq([gyak_{k,1}, gyak_{k,2}, gyak_{k,5}], k*
= 1 ..terjedelem)])

<i>Intervallumok</i>	<i>Gyakoriságok</i>	<i>Halmozott relatív gyakoriságok</i>
116. ...117.	1.	0.4629629630
117. ...118.	0.	0.4629629630
118. ...119.	1.	0.9259259259
119. ...120.	0.	0.9259259259
120. ...121.	0.	0.9259259259
121. ...122.	1.	1.3888888889
122. ...123.	0.	1.3888888889
123. ...124.	0.	1.3888888889
124. ...125.	1.	1.851851852
125. ...126.	2.	2.7777777778
126. ...127.	0.	2.7777777778
127. ...128.	2.	3.703703704
128. ...129.	0.	3.703703704
129. ...130.	1.	4.166666667
130. ...131.	0.	4.166666667
131. ...132.	1.	4.629629630
132. ...133.	3.	6.018518519
133. ...134.	2.	6.944444444
134. ...135.	4.	8.796296296
135. ...136.	2.	9.722222222
136. ...137.	3.	11.11111111
137. ...138.	6.	13.888888889
138. ...139.	4.	15.74074074
139. ...140.	3.	17.12962963
140. ...141.	1.	17.59259259
141. ...142.	3.	18.98148148
142. ...143.	6.	21.75925926
143. ...144.	6.	24.53703704
144. ...145.	5.	26.85185185
145. ...146.	6.	29.62962963
146. ...147.	6.	32.40740741
147. ...148.	5.	34.72222222
148. ...149.	6.	37.50000000

A **Statistics** csomag **FrequencyPlot** és **Histogram** utasításával a gyakoriság diagramot tudjuk felrajzoltatni, **pontszerűen**, vagy **oszlop** diagrammal vagy **pálcika** diagrammal.

> <i>FrequencyPlot</i>(<i>csapolas</i>, <i>style = point</i>, <i>title</i> = <i>Relatív gyakoriságok</i> <i>pontdiagramja</i>) 	> <i>Histogram</i>(<i>csapolas</i>, <i>binwidth = 1</i>, <i>title</i> = <i>Relatív gyakoriságok</i> <i>oszlop diagramja</i>); 	> <i>H := Histogram</i>(<i>csapolas</i>, <i>discrete = true</i>, <i>title</i> = <i>Relatív gyakoriságok</i> <i>pálcika diagramja</i>) : <i>H</i>; 
---	--	--

14.6.3. DEFINÍCIÓ. MÓDUSZ

A $minta = (x_1, x_2, \dots, x_n)$ azon x_M elemét, amely a legtöbbször vagy **leggyakrabban fordul elő** az adatsorban **móduszának** nevezzük.

14.6.4. PÉLDA. A kohó csapolási példa folytatása

Számoljuk ki a kohó csapolási adatok leggyakrabban előforduló értékét!

Rajzoljuk be az adatok gyakoriság hisztogram ábrájába az átlag és a módusz értékeket!

Megoldás

A **Statistics** csomag **Mode** eljárása megadja egy mintában leggyakrabban előforduló értéket!

> *Leggyakoribb* := *Mode*(*csapolas*);

Átlag = átlag

Leggyakoribb := 161.

Átlag = 152.944444444444

(1.6.3)

Erre az adatsorra a 161 érték gyakorisága 8, de nemcsak ennek a gyakorisága 8, hanem a 153 és a 154 értékeknek is. Ez azt jelenti, hogy a módusz nem egyértelműen definiált, a legnagyobb gyakoriság több különböző értékre is előfordulhat. Több **maximum helye** is lehet a gyakoriság hisztogramnak.

> *érték* = *Gyakoriság értékek*[47, 1], *gyakorisága* = *Gyakoriság értékek*[47, 2];

érték = *Gyakoriság értékek*[39, 1], *gyakorisága* = *Gyakoriság értékek*[39, 2];

érték = *Gyakoriság értékek*[40, 1], *gyakorisága* = *Gyakoriság értékek*[40, 2]

érték = 161..162., *gyakorisága* = 8.

érték = 153..154., *gyakorisága* = 8.

érték = 154..155., *gyakorisága* = 8.

(1.6.4)

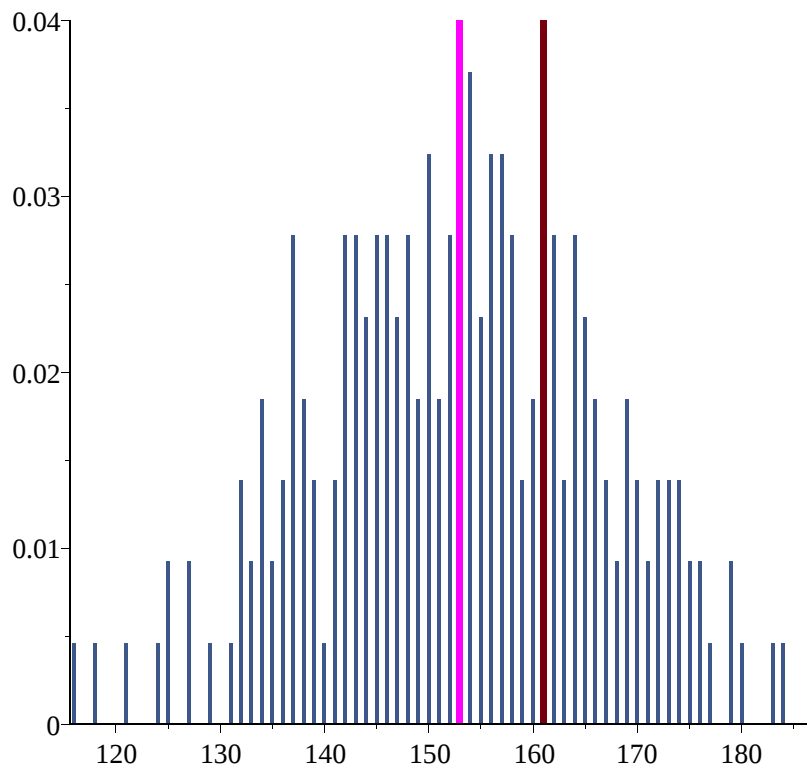
Rajzoljuk be az **átlagot** és a kapott **móduszt** a gyakoriság hisztogramba!

> *rajzM* := *plot*([[*Leggyakoribb*, 0], [*Leggyakoribb*, 0.04]], *thickness* = 3) :

rajzA := *plot*([[*átlag*, 0], [*átlag*, 0.04]], *color* = *magenta*, *thickness* = 3) :

plots[*display*]([*H*, *rajzM*, *rajzA*], *title* = *Az átlag és a módusz*)

Az átlag és a módusz



▼ 14.7. Halmazott gyakoriságok és az empirikus eloszlásfüggvény. Medián és a box plot.

A **FrequencyTable** utasítással kapott mátrix 5-ik oszlopában a relatív gyakoriságok halmazott összege található!

A halmazott relatív gyakoriságok leírásához az $X_1, X_2, X_3, \dots, X_n$ minta elemeket célszerű nagyság szerint növekvőleg rendezni: $Y_1 < Y_2 < Y_3 < \dots < Y_n$.

14.7.1. DEFINÍCIÓ. EMPIRIKUS ELOSZLÁSFÜGGVÉNY

Az

$$F_n(x) = \begin{cases} 0 & x \leq Y_1 \\ \frac{k}{n} & (Y_k < x \leq Y_{k+1}) \quad (k = 1, 2, \dots, n-1) \\ 1 & Y_n < x \end{cases} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(x_i \leq x)$$

függvényt a minta elemek empirikus vagy tapasztalati eloszlás függvényének nevezzük.

Tehát a függvény az $[Y_k, Y_{k+1}]$ intervallumban a $\frac{k}{n}$ konstans értéket veszi fel, amely a halmazott

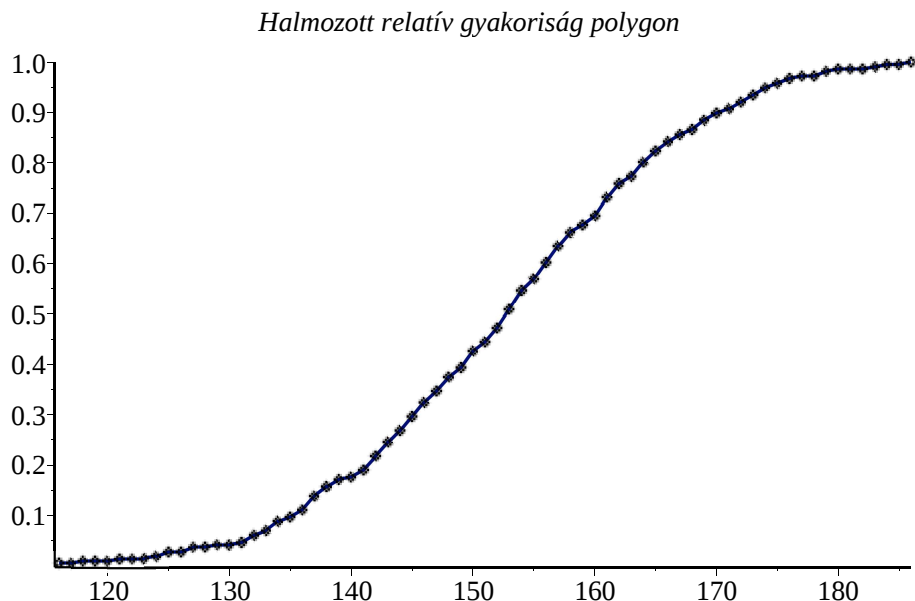
relatív gyakoriság.

14.7.2. PÉLDA. A kohó csapolási példa folytatása

Rajzoljuk fel a kohó csapolási adatok empirikus eloszlás függvényének vonalas diagramját!

Megoldás

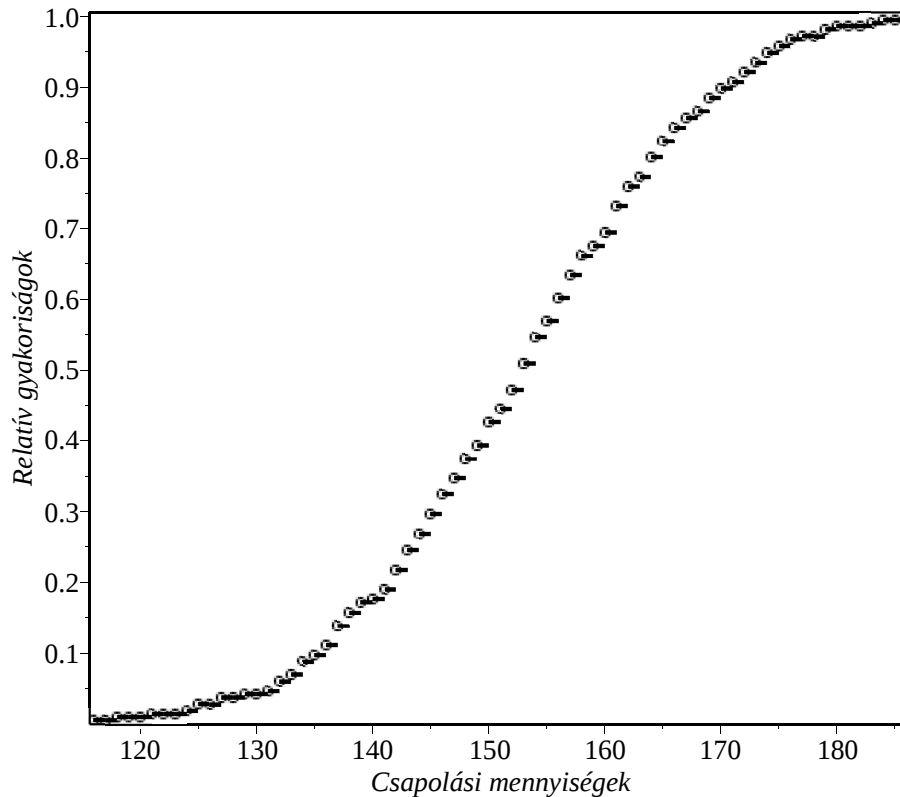
```
> S := LineChart(gyak[1..terjedelem, 5], xcoords = [seq(k, k = A..B)], scale = relative, title  
= Halmazott relatív gyakoriság polygon) : S;
```



A diszkrét eloszlásoknál is használt eloszlásfüggvényt rajzoló eljárással valóban lépcsős függvényt kaphatunk.

```
> plotDEF := proc(x, Fx, xlabel, ylabel)  
  local n, m, j, pList, ptsB, ptsT :  
  > n := nops(x) : m := nops(Fx) :  
  > ptsT := [seq([xj, Fxj], j = 1..n)] : ptsB := [seq([xj+1, Fxj], j = 1..n-1)] :  
  pList := plots[pointplot](ptsT, axes = boxed, style = point, symbol = circle, symbolsize = 8,  
    labels = [xlabel, ylabel], labeldirections = [horizontal, vertical], title  
    = "Empirikus eloszlásfüggvény lépcsős függvénye") :  
  for j to n-1 do  
    pList := pList, plots[pointplot]([ptsTj, ptsBj], connect = true, thickness = 2)  
  end do :  
  > plots[display]([pList])  
end proc:  
plotDEF([seq(k, k = A..B-1)], convert( $\frac{\text{gyak}_{1..terjedelem, 5}}{100}$ , list),  
  Csapolási mennyiségek, Relatív gyakoriságok)
```

Empirikus eloszlásfüggvény lépcsős függvénye



A **Medián** a minta középtértékét jelöli, tehát a nagyság szerint növekvően rendezett sorban a középen álló elem értéke. Ha van képen álló elem, akkor ennek értéke és egyébként a két középső elem számtani közepe.

14.7.3. DEFINÍCIÓ. MEDIÁN

Az $Y_1 < Y_2 < Y_3 < \dots < Y_n$ növekvőleg rendezett minta **medián** értéke a középen álló elem értéke, ha ez létezik. Külön kell választanunk a definíciót páros n és páratlan n esetekre!

Ha a minta $n = 2 \cdot m + 1$ elemszáma páratlan, akkor $\text{Medián}(Y) = Y_{m+1}$ a középen álló elem.

Ha a minta $n = 2 \cdot m$ elemszáma páros, akkor $\text{Medián}(Y) = \frac{Y_m + Y_{m+1}}{2}$ a két középső elem számtani átlaga.

A következő példa azt mutatja, hogy bizonyos esetekben a medián jobban jellemzi a statisztikai értékeket, mint az átlag.

14.7.4. PÉLDA. Amikor a medián jobban jellemzi a sokaságot, mint az átlag

Tekintsünk egy 5 fős csoportot, akik egy munkahelyen dolgoznak. A 4 alkalmazott fizetése legyen 100, 100, 100, 100 és a főnök fizetése legyen 1000 ezer forint. (Tehát mindegyik érték ezer forintban értendő!)

Hasonlítsuk össze az **átlag fizetést** és a **fizetések medián** értékét! Melyik jellemzi jobban a csoport tagjainak fizetését?

Megoldás

Számoljuk ki a *Statistics* csomag *Mean* eljárásával az átlag fizetést és a *Median* eljárással a fizetések mediánját!

> *fizetések* := [100, 100, 100, 100, 1000]

> *átlag fizetés* = *Mean*(*fizetések*), *a fizetések mediánja* = *Median*(*fizetések*)

fizetések := [100, 100, 100, 100, 1000]

átlag fizetés = 280., *a fizetések mediánja* = 100.

(1.7.1)

Az átlag $\frac{100 + 100 + 100 + 100 + 1000}{5} = \frac{1400}{5} = 280$ (ezer) forint. A medián érték a nagyság középső

szerint növekvőleg rendezett 100, 100, 100, 100, 1000 számok közepe. Amely a megjelölt 100 érték az, mert összesen 5 szám van és harmadik elemtől balra is 2 és jobbra is 2 elem van. Ezért ez a 100 áll középen.

Tehát az átlag 280 (e) forint és a medián 100 (e) forint. Látható, hogy az átlag szépíti azt a tényt, hogy valójában mindenki fizetése 100 (e) forint, kivéve a főnökét. ♣

14.7.5. PÉLDA. A kohó csapolási példa folytatása

Számoljuk ki a csapolási adatok **medián** értékét!

(i) **Szemléltessük** a medián értéket az empirikus eloszlás függvény segítségével!

(ii) Rajzoljuk fel az adatok **box plot** diagramját és magyarázzuk a vonalak jelentését!

(iii) Számoljuk ki a **kvartiliseket**!

Megoldás

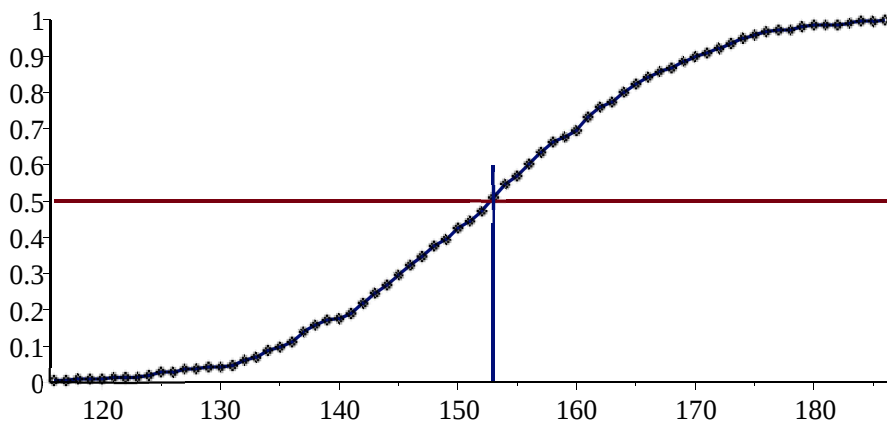
A *Statistics* csomag **Median** eljárása megadja a csapolt mennyiségek közép értékét vagy mediánját!

> *kozep* := *Median*(*csapolas*);

> *vonalak* := *plot*([[[A, 0.5], [B, 0.5]], [[*kozep*, 0], [*kozep*, 0.6]]) :

> *plots*[*display*]([*S*, *vonalak*], *ytickmarks* = 10, *title* = *A medián és a halmazott relatív gyakoriság polygon*)
kozep := 153.

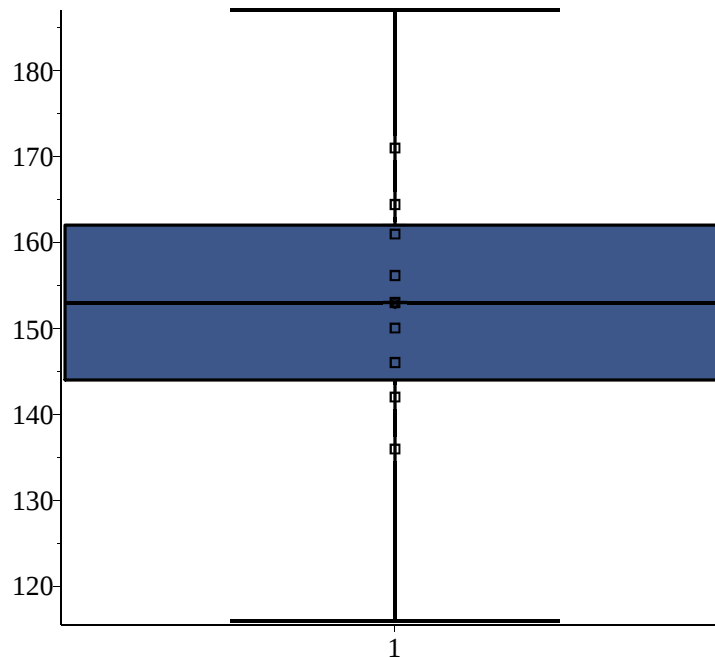
A medián és a halmazott relatív gyakoriság polygon



Tehát a medián az az m érték a vízszintes tengelyen, amelyre $P(X \leq m) = \frac{1}{2}$ és $P(m \leq X) = \frac{1}{2}$ teljesül. Vagyis az adatok 50% -a kisebb a mediánnál és 50%-a nagyobb a mediánnál. Tehát az $y = \frac{1}{2}$ valószínűségnél húzott konstans függvény az empirikus eloszlás függvényt a medián értéknél metszi. Jelen esetben a medián és az átlag majdnem egyenlő, mert $m = 153$ és $\mu = 152.94$.

Az adatok grafikus megjelenítése közül a box rajz az, amelyik **mutatja** a medián értéket is. Nézzük a csapolási adatok box megjelenítését.

> **BoxPlot(csapolas)**



A rajz egy téglalapból és a hozzá kapcsolt két függőleges, talpas ún. bajuszból áll. Az alsó bajusz az adatok **minimum**ától indul és a téglalap alsó oldaláig ér. A felső bajusz az adatok **maximum**át és a doboz felső oldalát köti össze. A függőleges tengely skálájáról leolvasható, hogy most az adatok a $[116, 187]$ értékek között helyezkednek el.

A dobozban levő vízszintes vonal az adatok medián értékét jelöli, amely most $m=153$ értéknél van. A doboz alsó oldala az értékek első kvartilisének vagy negyedének van, míg a doboz felső oldala az adatok harmadik kvartilisének mutatja. A **kvartilisek** az adatsort úgy osztják 4 részre, hogy mindegyik részbe az adatok $\frac{1}{4}$ része kerül. Így a második kvartilis megegyezik a mediánnal. A

box plotról leolvasható, hogy az adatok nagyrésze a medián közelében van, mert a doboz magassága a terjedelmen belül keskeny.

A *Statistics* csomag *Quartile* eljárásával kapjuk a kvartiliseket.

> $Q_1 := \text{Quartile}(\text{csapolas}, 1)$; $Q_2 := \text{Quartile}(\text{csapolas}, 2)$; $Q_3 := \text{Quartile}(\text{csapolas}, 3)$

$Q_1 := 144.$

$Q_2 := 153.$

$$Q_3 := 162.$$

(1.7.2)

Tehát az $n = 216$ csapolási adat negyede, tehát 54 adat van az alábbi intervallumok mindegyikében
 $[116, 144)$, $[144, 153)$, $[153, 162)$, $[162, 187]$

Így a Q_2 kvartilis megegyezik a medián értékkel. ♣

A medián az alábbi nevezetes minimum tulajdonságát igazoljuk folytonos eloszlásra!

14.7.6. TÉTEL. A MEDIÁN MINIMUM TULAJDONSÁGÁRÓL

Az X valószínűségi változó m mediánja a következő c értékre vonatkozó $\min_{c \in \mathbb{R}} E(|X - c|)$

minimum feladat megoldása

$$E(|X - m|) = \min_{-\infty < c < \infty} E(|X - c|)$$

Bizonyítás

Folytonos eloszlás esetén igazoljuk az állítást, melynek sűrűség függvénye $f(x)$.

$$E(|X - c|) = \int_{-\infty}^{\infty} |x - c| f(x) dx = \quad / \text{ az integrált válasszuk } c\text{-vel két részintervallumra!}$$

$$= \int_{-\infty}^c |x - c| f(x) dx + \int_c^{\infty} |x - c| f(x) dx = \quad / \text{ hagyjuk el az abszolút értékeket és vegyük figyelembe az}$$

előjeleket! /

$$= \int_{-\infty}^c (c - x) f(x) dx + \int_c^{\infty} (x - c) f(x) dx = \quad / \text{ használjuk az integrál linearitását és az}$$

$$F(x) = \int_{-\infty}^x f(u) du \quad \text{az eloszlás függvénye!} /$$

$$= c F(c) - \left(\int_{-\infty}^c x f(x) dx \right) + \int_c^{\infty} x f(x) dx - c (1 - F(c)) = \quad / \text{ vonjunk össze!}$$

$$= 2 c F(c) - c \int_{-\infty}^c x f(x) dx + \int_c^{\infty} x f(x) dx.$$

Deriváljuk c -szerint az egyenletet, hogy megkeressük c -re a minimum helyet!

$$\frac{\partial}{\partial c} E(|X - c|) = 2 F(c) + 2 c f(c) - 1 - 2 c f(c) = 2 F(c) - 1$$

Tehát $\frac{\partial}{\partial c} E(|X - c|) = 0$ akkor és csakis akkor, ha $F(c) = \frac{1}{2}$ teljesül. Ez a hely valóban a medián és ezen a helyen az $E(|X - c|)$ függvénynek c -ben valóban minimuma van, mert a második derivált nem negatív

$$\frac{\partial^2}{\partial c^2} E(|X - c|) = 2 f(c) > 0.$$

Ezzel igazoltuk, hogy az $Y = |X - c|$ valószínűségi változó várható értékének a $c = m$ medián helyen minimuma van. Q.e.d.

▼ 14.8. A minta korrigált szórása.

Legyen az $X_1, X_2, X_3, \dots, X_n$ véletlen minta közös várható értéke μ és a minta átlaga $\bar{X}_n = \frac{\sum_{k=1}^n X_k}{n}$.

14.8.1. DEFINÍCIÓ. KORRIGÁLT SZÓRÁSNÉGYZET VAGY VARIANCIA

A

$$Var_X = S_n^2 = \frac{\sum_{k=1}^n (X_k - \mu)^2}{n - 1}$$

menyiséget a minta korrigált szórásnégyszetének vagy varianciájának nevezzük. Ennek négyzetgyöke a korrigált szórá

$$S_n = \sqrt{\frac{\sum_{k=1}^n (X_k - \mu)^2}{n - 1}}$$

A **Statistics** csomag **StandardDeviation** eljárása kiszámolja a korrigált szórást és a **Variance** eljárás pedig a varianciát!

> *szórás := StandardDeviation(csapolas);*
variancia = Variance(csapolas); Szórásnégyszet = szórás^2;
szórás := 13.4671578696201

variancia = 181.364341085272

Szórásnégyszet = 181.364341085272

(1.8.1)

Vajon miért kell (n-1)-gyel osztani és miért nem n-nel? A kérdésre a következő tétel adja meg a választ!

TÉTEL. A korrigált szórásnégyszet a variancia torzítatlan becslése

Ha minden $k=1,2,\dots,n$ esetén $Var(X_k) = \sigma^2$, akkor $E(S_n^2) = \sigma^2$.

BIZONYÍTÁS

Mivel $Var(X_k) = E(X_k^2) - E(X_k)^2$, ezért átrendezés után

$$E(X_k^2) = \sigma^2 + \mu^2.$$

Láttuk, hogy $E\left(\frac{\sum_{k=1}^n X_k}{n}\right) = \mu$ és $Var\left(\frac{\sum_{k=1}^n X_k}{n}\right) = \frac{\sigma^2}{n}$, ezért a fenti azonosság szerint

$$E\left(\left(\frac{\sum_{k=1}^n X_k}{n}\right)^2\right) = \frac{\sigma^2}{n} + \mu^2$$

A korrigált szórásnégyszet képletét átrendezve

$$(n-1) S_n^2 = \sum_{k=1}^n X_k^2 - n \left(\frac{\sum_{k=1}^n X_k}{n} \right)^2$$

Ahonnán a várható értékre áttérve

$$\begin{aligned} E((n-1) S^2) &= E\left(\sum_{k=1}^n X_k^2\right) - n E\left(\left(\frac{\sum_{k=1}^n X_k}{n}\right)^2\right) = \text{ / az } E(\cdot) \text{ várható érték linearitása miatt /} \\ &= \sum_{k=1}^n E(X_k^2) - n \cdot E\left(\left(\frac{\sum_{k=1}^n X_k}{n}\right)^2\right) = n E(X_k^2) - n \cdot E\left(\left(\frac{\sum_{k=1}^n X_k}{n}\right)^2\right) \quad \text{ / a fenti képletek alapján /} \\ &= n (\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) = (n-1) \sigma^2 \end{aligned}$$

Tehát $E(S_n^2) = \sigma^2$. Vagyis a szórásnégyzet torzítatlan becslése a korrigált szórásnégyzet! Q.e.d.

Megjegyzés

Ha n nagy, akkor természetesen az n és $(n-1)$ osztó között kicsi az eltérés, így nem sokkal különbözne az az érték, ha $(n-1)$ helyett n -nel osztanánk!

▼ 14.9. Elméleti sűrűségfüggvény megbízhatósági becslése. Khi-négyzet modell teszt.

A csapolások gyakoriság hisztogramja hasonlít egy normál eloszlásra. Ezt gyakorlati szempontból az alábbi kettős indokkal magyarázhatjuk.

A csapolásnál egyrészt "túl kevés" vasat egyszerre nem célszerű készíteni, mert gazdaságtalan. Lehetőleg tele kell tenni az üstöt, amiben készítik a folyékony vasat. Másrészt "túl sok" vasat egyszerre nem tudnak készíteni, mert nem fér bele az üstbe. Tehát nagyon sokszor a középtér körül fognak csapolni, de a véletlenek miatt ritkán előfordulnak az ettől eltérő csapolások is. Képezzük azt a **normál eloszlást**, melynek várható értéke és szórása megegyezik az eddig számolt átlag és korrigált szórással! Majd rajzoljuk fel a sűrűség függvényét és a hisztogramot közös koordináta-rendszerben!

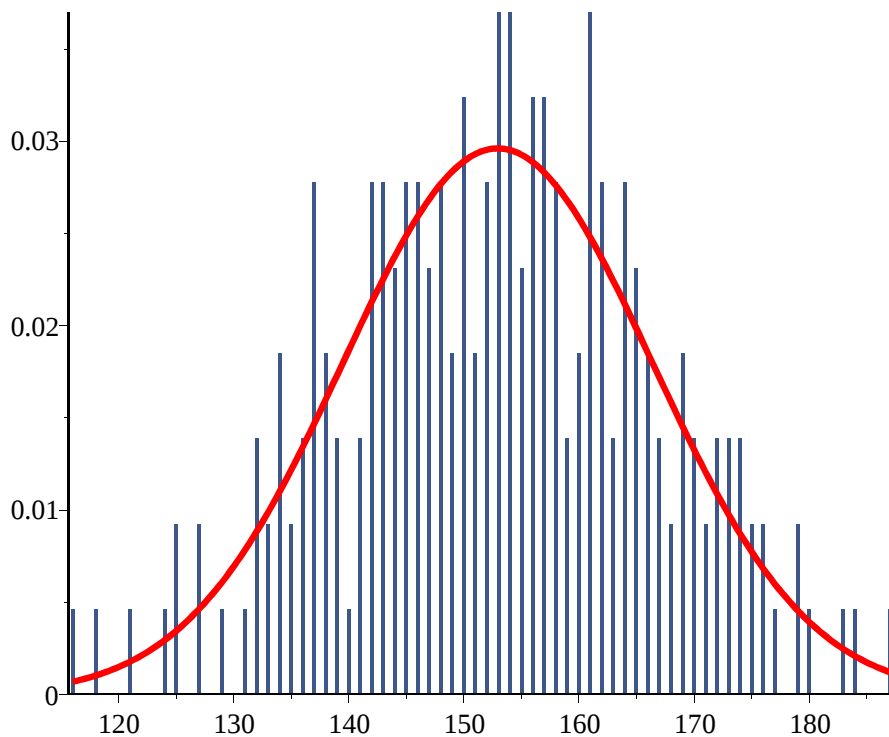
> $X := \text{RandomVariable}(\text{Normal}(\text{átlag}, \text{szórás})) :$

> $\text{suruseg} := \text{DensityPlot}(X, \text{range} = A..B, \text{thickness} = 3, \text{color} = \text{red}) :$

$\text{plots[display]}([H, \text{suruseg}], \text{title}$

$= \text{A relatív gyakoriságok és a normál eloszlás sűrűség függvénye jó illeszkedést mutatnak});$

A relatív gyakoriságok és a normál eloszlás sűrűség függvénye jó illeszkedést mutatnak



A normál eloszlás jól illeszkedik a mért értékekre! Hogyan lehetne mérni az illesztés jóságát?
A Maple **Statistics** csomagjába beépítették az adatok **eloszlására vonatkozó teszt vizsgálatot** khi-négyzet módszer felhasználásával, melynek neve **ChiSquareSuitableModelTest**!
Az eljárás válaszainak teljes megjelenítése érdekében a **Statistics** csomag **infolevel** nevű változóját 1 értékre állítjuk, így az eljárás által kiszámított eredmények és a magyarázatok mindegyike megjelenik!

```
> infolevel[Statistics] := 1 :
    statisztika := ChiSquareSuitableModelTest(csapolas,
        Normal(átlag, szórás), bins = terjedelem, level = 0.05);
```

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: 71
Distribution: ChiSquare(70)
Computed statistic: 56.1595
```

Computed pvalue: 0.88489
 Critical value: 90.5312253977732

Result: [Accepted]

There is no statistical evidence against the null hypothesis

statisztika := hypothesis = true, criticalvalue = 90.5312253977732, distribution **(1.9.1)**

= ChiSquare(70), pvalue = 0.884889738382919, statistic = 56.1595179234177

A **ChiSquareSuitableModelTest** eljárás

első paramétere a csapolási minta értékeiből képezett lista;

második paramétere annak az eloszlásnak a neve adott paraméterekkel együtt, amellyel közelíteni szeretnénk;

harmadik paramétere a részekre osztások számát adja meg;

az utolsó paraméter az úgy nevezett **szignifikancia vagy megbízhatósági szint**, melynek alapfeltételezett értéke 5%, azaz level=0.05. Ezt az értéket nem feltétlenül kell megadni.

Értelmezzük a Maple-től kapott válaszokat!

A Maple elsőként kiírja az úgynevezett **null hipotézist**, mely szerint

- a kivett **minta a megadott eloszlásból való**.

Ezt a **H0 hipotézist** kell eldönti khi-négyzet módszer segítségével az adatok ismeretében! A válasz a hipotézisre vagy **igaz** vagy **hamis** lesz!

A H0 hipotézis **tagadása** szerepel a következő sorban, amit **alternatív hipotézisnek** nevezünk.

Jelöljük ezt **H1**-gyel, amely így néz ki:

- **H1- alternatív hipotézis:** a kivett minta **nem a megadott eloszlásból való!**

A számításnál figyelembe vett érték közök száma 71, ezt jelzi a **bins =71**. A számoláshoz használt eloszlás a **ChiSquare(70)** eloszlás.

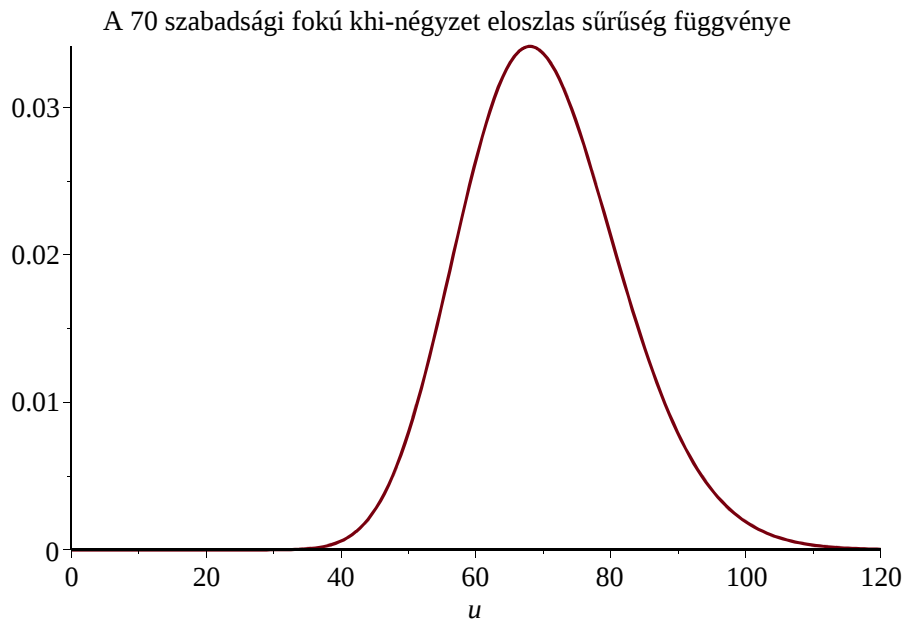
Tegyük egy kis **kitérőt** a szóban forgó **70 szabadsági fokú khi-négyzet eloszlás** világába! Adjuk meg **sűrűség függvényét** és annak grafikonját!

> *Khi70 := RandomVariable(ChiSquare(70)) : suruseg := PDF(Khi70, u)*

> *rajz70 := plot(suruseg, u = 0..120, title
 = "A 70 szabadsági fokú khi-négyzet eloszlás sűrűség függvénye") :
 rajz70*

suruseg :=

$$\left\{ \begin{array}{ll} 0 & u < 0 \\ \frac{1}{10144121732653119949813597183199268946575360000000} u^{34} e^{-\frac{1}{2}u} & \text{otherwise} \end{array} \right.$$



Felvettünk egy 70 szabadsági fokkal rendelkező khi70 nevű és **ChiSquare(70)** típusú valószínűségi változót a **RandomVariable** eljárás segítségével, majd ennek a *sűrűség függvényét* a **PDF** eljárás (**P**robability **D**ensity **F**unction) segítségével megkaptuk az **u** független változó exponenciális függvényeként! A sűrűségfüggvény ábráját láthatjuk a [0, 120] intervallumon! Mint minden rendes sűrűség függvény így ez is rendelkezik azzal a tulajdonsággal, hogy nem-negatív és a teljes görbe alatti területe 1. Ellenőrizzük ezt!

> *Int(suruseg, u = 0 .. infinity) = int(suruseg, u = 0 .. infinity);*

$$\int_0^{\infty} \begin{cases} 0 & u < 0 \\ \frac{1}{10144121732653119949813597183199268946575360000000} u^{34} e^{-\frac{1}{2}u} & \text{otherwise} \end{cases} du = 1 \quad (1.9.2)$$

Mutassuk most meg, hogy a **ChiSquareSuitableModelTest** által visszaadott
criticalvalue = 90.53122540

ún. kritikus érték olyan, amelytől számított görbe alatti területe végtelenig éppen 0.05 lesz! Ez az $\alpha=0.05$ a megadott **szignifikancia szint**! Ehhez nem kell mást tenni, mint integrálni a sűrűség függvényt a [90.53122540, ∞) tartományon!

> *statisztika[2];*
Int(suruseg , u = rhs(statisztika[2]) .. infinity) = int(suruseg, u = rhs(statisztika[2])
.. infinity);

$$criticalvalue = 90.5312253977732$$

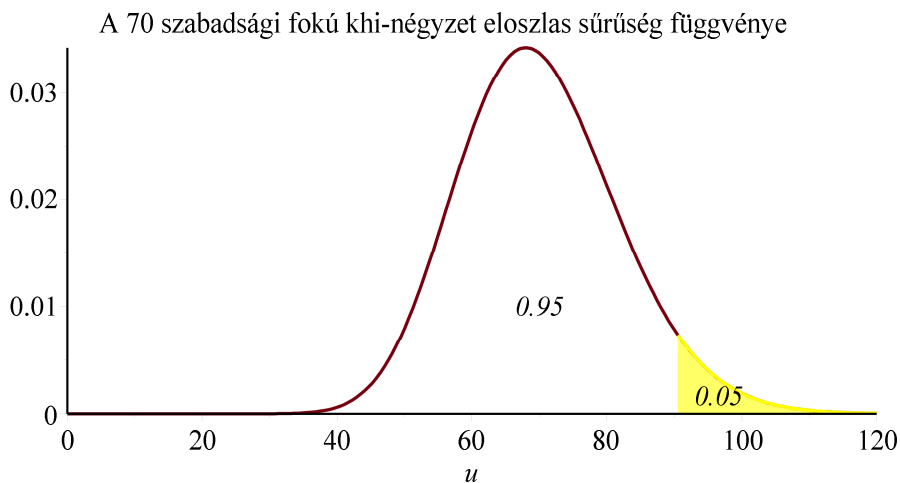
$$(1.9.3) \quad \int_{90.5312253977732}^{\infty} \frac{1}{10144121732653119949813597183199268946575360000000} u^{34} e^{-\frac{1}{2}u} du \quad u < 0 \quad otherwise$$

$$= 0.05000000027$$

Honnan vettük a kritikus értéket? Ez az eljárás által visszaadott érték sorozat második eleme. A választ ugyanis megőriztük a "statisztika" névben!

Fessük be sárga színűre a khi-négyszet sűrűség függvény görbéje alatti azon tartományt, amelynek területe 0.05 és írjuk bele a nagy területbe a maradék 0.95 értéket! Így látható, hogy melyik az a terület, amelynek mértéke 0.05 és melyik az, amelynek 0.95 .

```
> fest70 := plot(suruseg, u = rhs(statisztika[2]) ..120, filled = true, color = yellow) :
s70 := plots[textplot]([ [96.5, 0.0015, `0.05`], [70, 0.01, `0.95`]]) :
abra70 := plots[display]([rajz70, fest70, s70]) : abra70;
```



A H_0 hipotézis **elfogadása** vagy **visszautasítása** illetve a H_0 hipotézis **igaz** vagy **hamis** volta, két különböző tény! Az igaz hipotézist szeretnénk elfogadni illetve a hamis hipotézist szeretnénk visszautasítani. Előfordulnak azonban a statisztikai döntések során hibák. Az [elfogadás, visszautasítás] és az [igaz, hamis] párosítása 4-féleképpen lehetséges, melyet az alábbi 2x2-es táblázattal szemléltethetünk!

LEHETSÉGES PÁROSÍTÁSOK	A H_0 hipotézis igaz	A H_0 hipotézis hamis
A H_0 hipotézist elfogadjuk	A módszer jól működik	II. fajú hiba
A H_0 hipotézist visszautasítjuk	I. fajú hiba	A módszer jól működik

Abban az esetben amikor

- az igaz H_0 hipotézist elfogadjuk, vagy
- a hamis H_0 hipotézist visszautasítjuk,

akkor nem követünk el hibát. Ilyen esetekben a módszer jól működik! Ezek az esetek a 2x2-es táblázat főátlójában láthatók!

Van azonban két rossz eset, amelyeket a véletlenek miatti **hibáknak** kell minősítenünk.

1. eset: Amikor az igaz H_0 hipotézist visszautasítjuk vagy

2. eset: Amikor a hamis H_0 hipotézist elfogadjuk.

Az 1.esetet **I. fajú** a 2. esetet **II. fajú hibának** nevezzük.

Amennyiben a χ^2 teszt segítségével egy H_0 hipotézist 5%-os szignifikancia szinten elfogadjuk, akkor az annyit jelent, hogy ennek a döntésnek az I. fajú hibája kisebb lesz, mint 0.05. Tehát egy valóban igaz H_0 hipotézist az esetek kevesebb, mint 5%-ában utasítunk vissza. Még gyakorlatiasabban szólva, ha ugyanarra az **igaz hipotézisre** vonatkozóan ismételt 20 kísérletet végzünk el (egymástól függetlenül 20-szor mérünk csapolási adatokat), akkor a χ^2 teszt segítségével visszautasított esetek száma várhatóan nem lesz több, mint 1.

Sajnos a χ^2 teszt segítségével nem tudjuk a másodfajú hibát (β) kordában tartani! Csak az I. fajú hiba (α) mértékét lehet visszaszorítani.

A gyakorlatban a szokásos számok a **szignifikancia szintre** vonatkozóan a 0.001 , 0.01 és 0.05 értékek.

Azt mondjuk, hogy

$\alpha = 5\%$ és 1% közötti érték **majdnem szignifikáns**

$\alpha = 1\%$ és 0.1% közötti érték **szignifikáns**

$\alpha = 0.1\%$ alatti érték **erősen szignifikáns**.

Ezek után a statisztikai hipotézis eldöntése úgy működik, hogy a lista adataiból az eljárás kiszámít egy ún. **statisztikai értéket**, amelyet számolt statisztikának vagy "computed statistics"-nak nevez! Ezt két helyen is megadja! Egyrészt a szöveges részben, másrészt a visszaadott értékek utolsó, ötödik elemeként!

> *statisztika*[5];

statistic = 56.1595179234177

(1.9.4)

Mint látható a **számolt statisztika** 56.15951820 lett! Össze kell hasonlítani ezt a számolt értéket a korábbiakban χ^2 -ből kapott 90.53122540 **kritikus értékkel**!

A H_0 hipotézis elfogadásának feltétele: A χ^2 sűrűségéből számolt kritikus érték legyen

nagyobb, mint a számolt statisztikai érték, azaz
statisztikai érték < kritikus érték

A H_0 hipotézis visszautasításának feltétele: A χ^2 sűrűségből számolt kritikus érték legyen **kisebb**, mint a számolt statisztikai érték.

> *statisztika[5] < statisztika[2];*
statisztika[1];

(*statistic = 56.1595179234177*) < (*criticalvalue = 90.5312253977732*)

hypothesis = true

(1.9.5)

Statisztikai következtetés:

Mivel a csapolási adatokból számolt

statisztikai érték = 56.15951820 < a χ^2 sűrűségből számolt kritikus érték = 90.53122540 ,
ezért jelen esetben a H_0 hipotézist el **kell fogadni!**

Ezt a döntést olvashatjuk az eljárás által visszaadott

Result: [Accepted]

sorban is, amely annyit jelent, hogy a hipotézist elfogadjuk 95%-os megbízhatósági szinten!
Magyarázatként szolgál a következő mondat:

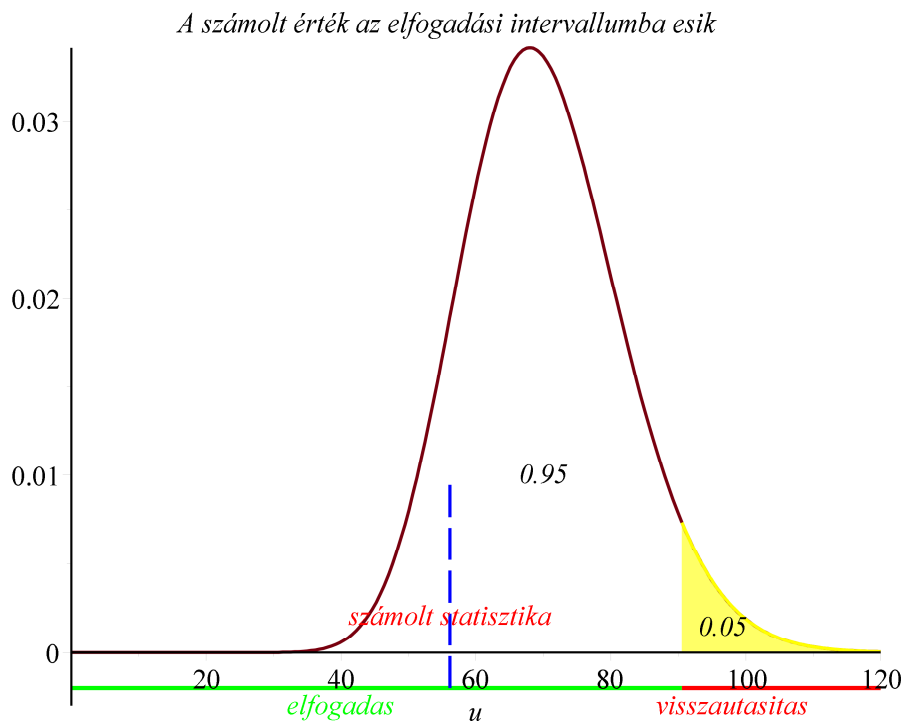
There is no statistical evidence against the null hypothesis.

Azaz a H_0 null hipotézis elfogadása ellen nem szól semmilyen statisztikai bizonyosság vagy
evidencia! Ez tehát esetünkben annyit jelent, hogy a

csapolási mennyiségek **normál eloszlást követnek**, melynek *átlaga* = 153 és *szórása* = 13.5

A megoldás végén felrajzolunk egy ábrát, amely pirossal szemlélteti az **elutasítási intervallumot**
és zölddel az **elfogadási intervallumot**. Továbbá berajzoltuk az ábrába a **számított statisztikai értéket** is, amely a zöld tartományba esik!

> *zold := plot([[0, -0.002], [rhs(statisztika_2), -0.002]], color = green, thickness = 2) :*
piros := plot([[rhs(statisztika_2), -0.002], [120, -0.002]], color = red, thickness = 2) :
igaz := plots[textplot]([[40, -0.003, elfogadas]], color = green) :
hamis := plots[textplot]([[98, -0.003, visszautasitas]], color = red) :
stat70 := plot([[rhs(statisztika_5), -0.002], [rhs(statisztika_5), 0.01]], color = blue,
linestyle = 3) :
mutat := plots[textplot]([[rhs(statisztika_5), 0.002, számolt statisztika]], color = red) :
plots[display]([abra70, zold, piros, igaz, hamis, stat70, mutat], title
= A számolt érték az elfogadási intervallumba esik)



▼ 14.10. Elméleti ellenőrző kérdések

- 14.10.1. Mit nevezünk sokaságnak?
- 14.10.2. Mit nevezünk mintának?
- 14.10.3. Mit értünk statisztika alatt?
- 14.10.4. Mit jelent az, hogy az átlag a várható érték torzítatlan becslése?
- 14.10.5. Mit jelent az, hogy az átlag a várható érték konzisztens becslése?
- 14.10.6. Mit jelent a várható érték minimum tulajdonsága?
- 14.10.7. Mit értünk egy minta móduszán?
- 14.10.8. Hogyan értelmezzük az empirikus eloszlás függvényt?
- 14.10.9. Milyen érték a minta mediánja?
- 14.10.10. Milyen minimum tulajdonsága van a mediánnak?
- 14.10.11. A adatok milyen jellemzőit mutatja a boxplot diagram?
- 14.10.12. Hogyan számoljuk a korrigált szórást?
- 14.10.13. Mit értünk az alatt, hogy a korrigált szórás négyzet torzítatlan becslése a varianciának?
- 14.10.14. Milyen H_0 hipotézist tesztlünk, a khi-négyzet modell teszt segítségével?
- 14.10.15. Mi a jelentése a szignifikancia szintnek? Milyen szignifikancia szintek megfelelőek?
- 14.10.16. Milyen feltétel esetén fogadjuk el a H_0 hipotézist a modell tesztnél?
- 14.10.17. Milyen hibát jelöl a teszt első fajú hibája?
- 14.10.18. Milyen hibát jelöl a teszt második fajú hibája?

▼ 14.11. Gyakorló feladatok

14.11.1. feladat.

Az építő mérnököknek felügyelni kell a folyó átlagos napi **vízhozamát** a vizierőmű megbízható működése érdekében. Az alábbi adatok a Namekagon folyón májusban mért napi átlagos vízhozamot tartalmazza különböző években.

vízhozam: = [602.0, 517.5, 572.5, 392.4, 505.8, 547.5, 389.1, 497.2, 794.8, 657.6, 904.7, 595.5, 611.9, 482.9, 698.6, 606.7, 986.4, 567.7, 400.1, 634.9, 448.4, 479.1, 1156.0, 718.5, 575.6, 743.3, 1146.0, 461.6, 644.0, 480.8, 429.1, 626.9, 833.9, 889.0, 752.6, 516.5, 817.2, 895.8, 572.2, 563.7, 679.3, 738.0, 618.9, 390.8, 550.9, 425.9, 760.6]

(i) Számoljuk ki az adatok terjedelmét, átlagát, korrigált szórását, móduszát, mediánját!

(ii) Rajzoljuk fel az adatok váltakozását az átlaggal, mediánnal és a kvartilisekkel együtt!

(iii) Rajzoljuk fel az adatok box diagramját!

(iv) Rajzoljuk fel a gyakoriság hisztogramot azon normál eloszlás sűrűség függvényével együtt, melynek várható értéke és szórása megegyezik az adatok átlagával és korrigált szórásával!

(v) Teszteljük $\alpha = 0.05$ szignifikancia szinten, hogy az adatok normál eloszlást követnek-e!

14.11.2. feladat.

Jelölje X az USA Kalifornia államának Oroville városa mellett folyó Feather River folyó éves áradási szintjének maximumát. Az alábbi adatokat regisztrálták 1902- és 1960 között bizonyos alapszinthez viszonyítva.

áradások: = [41, 102, 118, 81, 128, 230, 16, 140, 31, 75, 16, 17, 122, 81, 42, 80, 28, 66, 23, 62, 36, 22, 42, 64, 56, 94, 185, 14, 80, 12, 23, 9, 20, 59, 85, 19, 185, 8, 152, 84, 110, 108, 25, 60, 54, 46, 37, 17, 46, 92, 13, 59, 113, 55, 203, 83, 102, 35, 135]

Készítsük el az adatok statisztikai jellemzőit és hisztogramját!

A fenti áradási maximumokra illesszünk exponenciális eloszlást!

Teszteljük a

H_0 : az adatok exponenciális eloszlású sokaságból valók

hipotézist 5%-os szignifikancia szinten!

Változtassuk az osztásközök számát!

14.11.3. feladat.

Egy ipari vállalat napi kén-oxid kibocsátását mutatja az alábbi táblázat.

kén_oxid: = [15.8, 26.4, 17.3, 11.2, 23.9, 24.8, 18.7, 13.9, 9.0, 13.2, 22.7, 9.8, 6.2, 14.7, 17.5, 26.1, 12.8, 28.6, 17.6, 23.7, 26.8, 22.7, 18.0, 20.5, 11.0, 20.9, 15.5, 19.4, 16.7, 10.7, 19.1, 15.2, 22.9, 26.6, 20.4, 21.4, 19.2, 21.6, 16.9, 19.0, 18.5, 23.0, 24.6, 20.1, 16.2, 18.0, 7.7, 13.5, 23.5, 14.5, 14.4, 29.6, 19.4, 17.0, 20.8, 24.3, 22.5, 24.6, 18.4, 18.1, 8.3, 21.9, 12.3, 22.3, 13.3, 11.8, 19.3, 20.0, 25.7, 31.8, 25.9, 10.5, 15.9, 27.5, 18.1, 17.9, 9.4, 24.1, 20.1, 28.5]

(i) Számoljuk ki az adatok terjedelmét, átlagát, korrigált szórását, móduszát, mediánját!

(ii) Rajzoljuk fel az adatok váltakozását az átlaggal, mediánnal és a kvartilisekkel együtt!

(iii) Rajzoljuk fel az adatok box diagramját!

(iv) Rajzoljuk fel a gyakoriság hisztogramot azon normál eloszlás sűrűség függvényével együtt, melynek várható értéke és szórása megegyezik az adatok átlagával és korrigált szórásával!

(v) Teszteljük $\alpha = 0.05$ szignifikancia szinten, hogy az adatok normál eloszlást követnek-e!

14.11.4. feladat.

A következő adatok az USA Yellowstone Nemzeti Parkja egyik legnevezetesebb gejzírének, az "Old Faithful" (Öreg Hűséges) -nek nevezett gejzír **kitörési időtartamait** mutatja percekben mérve. A 222 kitörést 1978 augusztusa és 1979 augusztusa között mérték. (Lásd az adatokat Applied Linear Regression, 2nd Edition, by Sanford Weisberg, pp. 231 and 234. és a http://www.math.hope.edu/swanson/statlabs/lab_20.html internet címen)

kitörések := [4.4, 3.9, 4, 4, 3.5, 4.1, 2.3, 4.7, 1.7, 4.9, 1.7, 4.6, 3.4, 4.3, 1.7, 3.9, 3.7, 3.1, 4, 1.8,

4.1, 1.8, 3.2, 1.9, 4.6, 2, 4.5, 3.9, 4.3, 2.3, 3.8, 1.9, 4.6, 1.8, 4.7, 1.8, 4.6, 1.9, 3.5, 4, 3.7, 3.7, 4.3, 3.6, 3.8, 3.8, 3.8, 2.5, 4.5, 4.1, 3.7, 3.8, 3.4, 4, 2.3, 4.4, 4.1, 4.3, 3.3, 2, 4.3, 2.9, 4.6, 1.9, 3.6, 3.7, 3.7, 1.8, 4.6, 3.5, 4, 3.7, 1.7, 4.6, 1.7, 4, 1.8, 4.4, 1.9, 4.6, 2.9, 3.5, 2, 4.3, 1.8, 4.1, 1.8, 4.7, 4.2, 3.9, 4.3, 1.8, 4.5, 2, 4.2, 4.4, 4.1, 4.1, 4, 4.1, 2.7, 4.6, 1.9, 4.5, 2, 4.8, 4.1, 4.1, 4.2, 4.5, 1.9, 4.7, 2, 4.7, 2.5, 4.3, 4.4, 4.4, 4.3, 4.6, 2.1, 4.8, 4.1, 4, 4, 4.4, 4.1, 4.3, 4, 3.9, 3.2, 4.5, 2.2, 4.7, 4.6, 2.2, 4.8, 4.3, 3.8, 4, 4.1, 1.8, 4.4, 4, 2.2, 5.1, 1.9, 5, 4.4, 4.5, 3.8, 4.3, 4.4, 2.2, 4.8, 1.9, 4.7, 1.8, 4.8, 2, 4.4, 2.5, 4.3, 4.4, 1.9, 4.7, 4.3, 2.2, 4.7, 2.3, 4.6, 3.3, 4.2, 2.9, 4.6, 3.3, 4.2, 2.6, 4.6, 3.7, 1.8, 4.7, 4.5, 4.5, 4.8, 2, 4.8, 1.9, 4.7, 2, 5.1, 4.3, 4.8, 3, 2.1, 4.6, 4, 2.2, 5.1, 2.9, 4.3, 2.1, 4.7, 4.5, 1.7, 4.2, 4.3, 1.7, 4.4, 4.2, 2.2, 4.7, 4, 1.8, 4.7, 1.8, 4.5, 2.1, 4.2, 2.1, 5.2, 2] :

- (i) Számítsuk a **kitörési időtartamok** legkisebb, legnagyobb értékét, terjedelmét, átlagát, szórását!
- (ii) Rajzoljuk fel a **boxplot** diagramot! Magyarázzuk a rajzon található rajzi elemek jelentését!
- (iii) Számoljuk ki a **kvartiliseket** és a **mediánt**! Rajzoljuk fel az adatok váltakozását és a kvartiliseket! Magyarázzuk a kvartilisek jelentését!
- (iv) Mutassuk meg, hogy a kitörési időtartamok nem követnek **normál eloszlást**! Vessük össze a gyakoriság hisztogramot és a normál eloszlás sűrűség függvényét!
- (v) Rajzoljuk fel az **empirikus eloszlás függvényt** az *EmpiricalDistribution* eljárás segítségével! Hasonlítsuk össze az empirikus eloszlás függvényt és a normál eloszlás eloszlás függvényével!

14.11.5. feladat.

A következő adatok az USA Yellowstone Nemzeti Parkja egyik legnevezetesebb gejzirének, az "Old Faithful" (Öreg Hűséges) -nek nevezett gejzír kitörések között eltelt **várakozási** idők mutatja percekben mérve. A 222 kitörést 1978 augusztusa és 1979 augusztusa között mérték. (Lásd az adatokat Applied Linear Regression, 2nd Edition, by Sanford Weisberg, pp. 231 and 234. és a http://www.math.hope.edu/swanson/statlabs/lab_20.html internet címen)

várakozások := [78, 74, 68, 76, 80, 84, 50, 93, 55, 76, 58, 74, 75, 80, 56, 80, 69, 57, 90, 42, 91, 51, 79, 53, 82, 51, 76, 82, 84, 53, 86, 51, 85, 45, 88, 51, 80, 49, 82, 75, 73, 67, 68, 86, 72, 75, 75, 66, 84, 70, 79, 60, 86, 71, 67, 81, 76, 83, 76, 55, 73, 56, 83, 57, 71, 72, 77, 55, 75, 73, 70, 83, 50, 95, 51, 82, 54, 83, 51, 80, 78, 81, 53, 89, 44, 78, 61, 73, 75, 73, 76, 55, 86, 48, 77, 73, 70, 88, 75, 83, 61, 78, 61, 81, 51, 80, 79, 82, 80, 76, 56, 82, 47, 76, 61, 75, 72, 74, 69, 78, 52, 91, 66, 71, 75, 81, 77, 74, 70, 83, 53, 82, 62, 73, 84, 58, 82, 77, 75, 77, 77, 53, 75, 78, 51, 81, 52, 76, 73, 84, 72, 89, 75, 57, 81, 49, 87, 43, 94, 45, 81, 59, 82, 80, 54, 75, 73, 57, 80, 51, 77, 66, 77, 60, 86, 62, 75, 67, 69, 84, 58, 90, 82, 71, 80, 51, 80, 62, 84, 51, 81, 83, 84, 72, 54, 75, 74, 51, 91, 60, 80, 54, 80, 70, 60, 86, 78, 51, 83, 76, 51, 90, 71, 49, 88, 52, 79, 61, 81, 48, 84, 63] :

- (i) Számítsuk a **várakozási időtartamok** legkisebb, legnagyobb értékét, terjedelmét, átlagát, szórását!
- (ii) Rajzoljuk fel a **boxplot** diagramot! Magyarázzuk a rajzon található rajzi elemek jelentését!
- (iii) Számoljuk ki a **kvartiliseket** és a **mediánt**! Rajzoljuk fel az adatok váltakozását és a kvartiliseket! Magyarázzuk a kvartilisek jelentését!
- (iv) Mutassuk meg, hogy a várakozási időtartamok nem követnek **normál eloszlást**! Vessük össze a gyakoriság hisztogramot és a normál eloszlás sűrűség függvényét!
- (v) Rajzoljuk fel az **empirikus eloszlás függvényt** az *EmpiricalDistribution* eljárás segítségével! Hasonlítsuk össze az empirikus eloszlás függvényt és a normál eloszlás eloszlás függvényével!

14.11.6. feladat.

Azonos típusú beton rudak szakítási szilárdságára vonatkozó kísérletek eredményeit (lb/inch²) mutatja az alábbi táblázat

szakítás:= [320,380,340,410,380,340,360,350,320,370, 350,340,350,360,370,350,380,370,300,420, 370,390,390,440,330,390,330,360,400,370, 320,350,360,340,340,350,350,390,380,340, 400,360,350,390,400,350,360,340,370,420,

420,400,350,370,330,320,390,380,400,370,
390,330,360,380,350,330,360,300,360,360,
360,390,350,370,370,350,390,370,370,340,
370,400,360,350,380,380,360,340,330,370,
340,360,390,400,370,410,360,400,340,360];

- (i) Számítsuk ki az adatok statisztikai jellemzőit!
- (ii) Rajzoljuk fel a gyakoriság hisztogramot!
- (iii) Illeszzünk normál eloszlást a gyakoriság adatokra!
- (iv) Teszteljük a
 H_0 : az adatok normál eloszlású sokaságból valók

hipotézist χ^2 módszerrel 5%-os szignifikancia szinten!

14.11.7. feladat.

Egy forgalmas autópályán az autók közötti követési időközöket mérték másodpercekben. Az autópálya egy adott helyén az alábbi követési időközök adódtak egymás után.

Követési idők=

[4.1,3.5,2.2,2.7,2.7,4.1,3.4,1.8,3.1,2.1,
2.1,1.7,2.3,3.0,4.1,3.2,2.2,2.3,1.5,1.1,
2.5,4.7,1.8,4.8,1.8,4.8,4.9,3.1,5.7,5.7,
3.1,2.0,2.9,5.9,2.1,3.0,4.4,2.1,2.6,2.7,
3.2,2.5,1.7,2.0,2.7,1.2,9.0,1.8,2.1,5.4,
2.1,3.8,4.5,3.3,2.1,2.1,7.1,4.7,3.1,1.7,
2.2,3.1,1.7,3.1,2.3,8.1,5.7,2.2,4.0,2.7,
1.5,1.7,4.0,6.4,1.5,2.2,1.2,5.1,2.7,2.4,
1.7,1.2,2.7,7.0,3.9,5.2,2.7,3.5,2.9,1.2,
1.5,2.7,2.9,4.1,3.1,1.9,4.8,4.0,3.0,2.7]

- (i) Számítsuk ki az adatszámot, terjedelmet, az átlagot és a korrigált szórást!
- (ii) Rajzoljuk fel az adatsor váltakozását az átlaggal együtt!
- (iii) Osszuk fel a terjedelmet 4 egyenlő részekre és rajzoljuk fel a mérési pontokat, a sávokkal együtt!
- (iv) Számítsuk ki a gyakoriságokat és rajzoljuk fel a gyakoriság hisztogramot! Változtassuk a közök számát! Milyen eloszlást kaptunk?
- (v) A követési időkre illesszünk exponenciális eloszlást!
Teszteljük a
 H_0 : az adatok exponenciális eloszlású sokaságból valók
hipotézist 5%-os szignifikancia szinten!
- (vi) Teszteljük a

H_0 : az adatok Gamma- eloszlású sokaságból valók

hipotézist 5%-os szignifikancia szinten!

Határozzuk meg a Gamma - eloszlás két paraméterét!

(vii) Adjuk össze a követési időket kettesével! Illesszünk a kapott mintára Gamma - eloszlást!

Keressük meg a Gamma - eloszlás két paraméterét!

Hasonlítsuk össze a kapott paraméter értékét a (vi) pontbeli Gamma-eloszlás paraméterével!

Teszteljük a

H_0 : az összeg Gamma- eloszlású sokaságból való

hipotézist 5%-os szignifikancia szinten!

14.11.8. feladat.

Az alábbi 4 független minta 4 populációból való!

minta1:=[5.098, 2.739,2.146,5.006,4.016,9.026,4.965,5.016,6.195,4.523]

minta2:=[4.627, 5.061,2.787, 4.181,3.617,3.605, 6.03,4.745,2.340,6.934]

minta3:=[3.021,6.173,7.602,6.250,1.875,6.996,4.850,6.661,6.360,7.052]

minta4:=[7.390,5.666,6.616,7.868,2.428,6.740,7.605,10.868,1.739, 1.996]

(i) Használjuk a boxplot diagramot annak eldöntésére, hogy a 4 minta ugyanabból a populációból való-e!

Miért igen vagy miért nem?

(ii) Rajzoljuk fel az adatok gyakoriság hisztogramját olyan normál eloszlás sűrűség függvényével együtt, melynek várható értéke az adatok átlaga és szórása az adatok korrigált szórása.

14.11.9. feladat.

Az alábbi adatsor egy hallgatói csoport Zh pontszámait mutatja nagyság szerint csökkenő sorrendben.

pontszámok:=[90,90,89,88,85,85,84,82,82,82,

81,81,81,80,79,79,78,76,75,74,

72,71,70,66,65,63,62,62,61,59,

58,58,57,56,56,53,48,44,40,35,33]

(i) Számoljuk ki pontszámok átlagát, szórását, a mediánt és módozt!

(ii) Rajzoljuk fel a pontszámok box diagramját és a gyakoriság hisztogramját!

(iii) Normál eloszlásúak-e a pontszámok 5%-os szignifikancia szinten?

14.11.10. feladat.

Mérnökök tesztelték a fejlesztés alatt álló alumínium ötvözet nyomószilárdságát, amelyet új generációs repülőgépekhez készítettek. A mért nyomószilárdság értékek az alábbiak.

nyomó_szilárdságok:=[66.4,67.7,68.0,68.0,68.3,68.4,68.6,68.8,68.9,69.0,69.1,

69.2,69.3,69.3,69.5,69.5,69.6,69.7,69.8,69.8,69.9,70.0,

70.0,70.1,70.2,70.3,70.3,70.4,70.5,70.6,70.6,70.8,70.9,

71.0,71.1,71.2,71.3,71.3,71.5,71.6,71.6,71.7,71.8,71.8,

71.9,72.1,72.2,72.3,72.4,72.6,72.7,72.9,73.1,73.3,73.5,

74.2,74.5,75.3]

(i) Számoljuk ki az adatok terjedelmét, átlagát, korrigált szórását, módoztát, mediánját!

(ii) Rajzoljuk fel az adatok váltakozását az átlaggal, mediánnal és a kvartilisekkel együtt!

(iii) Rajzoljuk fel az adatok box diagramját!

(iv) Rajzoljuk fel a gyakoriság hisztogramot azon normál eloszlás sűrűség függvényével együtt, melynek várható értéke és szórása megegyezik az adatok átlagával és korrigált szórásával!

(v) Teszteljük $\alpha = 0.05$ szignifikancia szinten, hogy az adatok normál eloszlást követnek-e!

14.11.11. feladat.

Az alábbi adatok fonál szakítási erejét mutatja uncia (kb. 28 gr) egységben mérve.

fonal:=[32.5,15.2,35.4,21.3,28.4,26.9,34.6,29.3,24.5,31.0,

21.2,28.3,27.1,25.0,32.7,29.5,30.2,23.9,23.0,26.4,

27.3,33.7,29.4,21.9,29.3,17.3,29.0,36.8,29.2,23.5,

20.6,29.5,21.8,37.5,33.5,29.6,26.8,28.7,34.8,18.6,

25.4,34.1,27.5,29.6,22.2,22.7,31.3,33.2,37.0,28.3,

36.9,24.6,28.9,24.8,28.1,25.4,34.5,23.6,38.4,24.0]

(i) Számoljuk ki az adatok terjedelmét, átlagát, korrigált szórását, módoztát, mediánját!

(ii) Rajzoljuk fel az adatok váltakozását az átlaggal, mediánnal és a kvartilisekkel együtt!

(iii) Rajzoljuk fel az adatok box diagramját!

(iv) Rajzoljuk fel a gyakoriság hisztogramot azon normál eloszlás sűrűség függvényével együtt, melynek várható értéke és szórása megegyezik az adatok átlagával és korrigált szórásával!

(v) Teszteljük $\alpha = 0.05$ szignifikancia szinten, hogy az adatok normál eloszlást követnek-e!

14.11.12. feladat.

Az alábbi adatok nyílt lángnak kitett kárpitos bútor szövetek gyulladási idejét mutatja század másodpercekben mérve.

gyulladás:=[2.58,2.51,4.04,6.43,1.58,4.32,2.20,4.19,

4.79,6.20,1.52,1.38,3.87,4.54,5.12,5.15,

5.50,5.92,4.56,2.46,6.90,1.47,2.11,2.32,
6.75,5.84,8.80,7.40,4.72,3.62,2.46,8.75,
2.65,7.86,4.71,6.25,9.45,12.80,1.42,1.92,
7.60,8.79,5.92,9.65,5.09,4.11,6.37,5.40,
11.25,3.90,5.33,8.64,7.41,7.95,10.60,3.81,
3.78,3.75,3.10,6.43,1.70,6.40,3.24,1.79,
4.90,3.49,6.77,5.62,9.70,5.11,4.50,2.50,
5.21,1.76,9.20,1.20,6.85,2.80,7.35,11.75]

(i) Számoljuk ki az adatok terjedelmét, átlagát, korrigált szórását, móduszát, mediánját!

(ii) Rajzoljuk fel az adatok váltakozását az átlaggal, mediánnal és a kvartilisekkel együtt!

(iii) Rajzoljuk fel az adatok box diagramját!

(iv) Rajzoljuk fel a gyakoriság hisztogramot azon normál eloszlás sűrűség függvényével együtt, melynek várható értéke és szórása megegyezik az adatok átlagával és korrigált szórásával!

(v) Teszteljük $\alpha = 0.05$ szignifikancia szinten, hogy az adatok normál eloszlást követnek-e!

14.11.13. feladat.

Mérték a munkások termelékenységét egy üzemben 2 héten keresztül úgy, hogy a munkások által készített elfogadható termékek számát írták fel 100 munkás esetén.

termelékenység:=

[65,36,49,84,79,56,28,43,67,36,
43,78,37,40,68,72,55,62,22,82,
88,50,60,56,57,46,39,57,73,65,
59,48,76,74,70,51,40,75,56,45,
35,62,52,63,32,80,64,53,74,34,
76,60,48,55,51,54,45,44,35,51,
21,35,61,45,33,61,77,60,85,68,
45,53,34,67,42,69,52,68,52,47,
63,65,55,61,73,50,53,59,41,54,
41,74,82,58,26,35,47,50,38,70]

(i) Számoljuk ki az adatok terjedelmét, átlagát, korrigált szórását, móduszát, mediánját!

(ii) Rajzoljuk fel az adatok váltakozását az átlaggal, mediánnal és a kvartilisekkel együtt!

(iii) Rajzoljuk fel az adatok box diagramját!

(iv) Rajzoljuk fel a gyakoriság hisztogramot azon normál eloszlás sűrűség függvényével együtt, melynek várható értéke és szórása megegyezik az adatok átlagával és korrigált szórásával!

(v) Teszteljük $\alpha = 0.05$ szignifikancia szinten, hogy az adatok normál eloszlást követnek-e!

14.11.14. feladat.

Mérték 20 közepes méretű teszt autó városi úton az üzemanyag fogyasztását mérföld/gallon (mpg) mérték egységben.

fogyasztás:=

[19.7,21.5,22.5,22.2,22.6,
21.9,20.5,19.3,19.9,21.7,
22.8,23.2,21.4,20.8,19.4,
22.0,23.0,21.1,20.9,21.3
19.7,21.5,22.5,22.2,22.6,
21.9,20.5,19.3,19.9,21.7,
22.8,23.2,21.4,20.8,19.4,
22.0,23.0,21.1,20.9,21.3]

(i) Számoljuk ki az adatok terjedelmét, átlagát, korrigált szórását, móduszát, mediánját!

(ii) Rajzoljuk fel az adatok váltakozását az átlaggal, mediánnal és a kvartilisekkel együtt!

(iii) Rajzoljuk fel az adatok box diagramját!

(iv) Rajzoljuk fel a gyakoriság hisztogramot azon normál eloszlás sűrűség függvényével együtt, melynek várható értéke és szórása megegyezik az adatok átlagával és korrigált szórásával!

(v) Teszteljük $\alpha = 0.05$ szignifikancia szinten, hogy az adatok normál eloszlást követnek-e!

14.11.15. feladat.

Az alábbi adatok a napsugárzás intenzitására, különböző napokon mért eredmények $\left[\frac{\text{watt}}{\text{m}^2} \right]$

mértékegységben, valahol dél Spanyolországban.

sugárzás:=[562, 869, 708, 775, 775, 704, 809, 856, 655, 806,
878, 909, 918, 558, 768, 870, 918, 940, 946, 661,
820, 898, 935, 952, 957, 693, 835, 905, 939, 955,
960, 498, 653, 730, 753]

- (a) Számoljuk ki az intenzitási értékek átlagát és szórását!
- (b) Rajzoljuk fel az adatsor váltakozását az átlaggal együtt!
- (c) Rajzoljuk fel az intenzitási adatok gyakoriság hisztogramját a terjedelem $n=15$ részre osztásával!
- (d) Rajzoljuk be a gyakoriság hisztogramba annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása egyezik az adatok átlagával és szórásával!
- (e) Vajon a napsugárzás intenzitása normál eloszlást követ-e 95%-os megbízhatósággal, ha $n=15$ részre osztjuk a terjedelmet?
- (f) Van-e olyan n felosztása a terjedelemnek, amely esetén az eloszlás nem lesz normál?

14.11.16. feladat.

A felhők "beoltását" régóta tanulmányozzák, mint az időjárás módosításának egyik eljárását (lásd pl. a Simpson, Alsen, és Eden által a "Technometrics" ben közölt "A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification", Vol. 17, pp. 161–166. cikket)

A cikkben a csapadék mennyiségét mérték (kat. hold x láb) térfogat egységben, amelyet 20 olyan véletlenszerűen kiválasztott felhőre jegyeztek fel, amelyet ezüst nitráttal oltottak be. A csapadék mennyiségét az alábbi adatsor mutatja:

csapadékok:=[18.0, 30.7, 19.8, 27.1, 22.3, 18.8, 31.8, 23.4, 21.2, 27.9, 31.9, 27.1, 25.0, 24.7, 26.9, 21.8, 29.2, 34.8, 26.7, 31.6]

- (a) Rajzoljuk fel a csapadék gyakoriság hisztogramját!
- (b) Rajzoljuk be a gyakoriság hisztogramba annak a normál eloszlásnak a sűrűség függvényét, melynek várható értékét és szórását az adatokból számítjuk!
- (c) A csapadék mennyisége normál eloszlást követ-e 95%-os megbízhatósággal?
- (d) Számítsuk ki az átlag konfidencia intervallumát $\alpha=0.05$ szignifikancia szinten!
- (e) Teszteljük a H_0 : az eső átlag = 27 hipotézist $\alpha=0.05$ szignifikancia szinten, melynek alternatív hipotézise legyen a H_1 : az eső *átlaga* $\neq 27$ hipotézis!

(f) Mennyi mérést kell végezni ahhoz, hogy az eső várható értékére 95%-os megbízhatósági szinten $L = 2$ [kat. hold x láb] hosszú konfidencia intervallum adódjon?

14.11.17. feladat.

Egy félvezető termékeket gyártó cég olyan terméket állít elő, amelyet a személyi számítógépek központi vezérlő egységében használnak fel. A termék sebessége megahertzben mérve egy fontos jellemzője a terméknek, mert alapvetően meghatározza az árát. Az alábbi táblázat tartalmazza 120 termék sebességi adatát [megaHertz] mértékegységben.

sebességek:=

[680, 669, 719, 699, 670, 710, 722, 663, 658, 634, 720, 690,
677, 669, 700, 718, 690, 681, 702, 696, 692, 690, 694, 660,
649, 675, 701, 721, 683, 735, 688, 763, 672, 698, 659, 704,
681, 679, 691, 683, 705, 746, 706, 649, 668, 672, 690, 724,
652, 720, 660, 695, 701, 724, 668, 698, 668, 660, 680, 739,
717, 727, 653, 637, 660, 693, 679, 682, 724, 642, 704, 695,
704, 652, 664, 702, 661, 720, 695, 670, 656, 718, 660, 648,
683, 723, 710, 680, 684, 705, 681, 748, 697, 703, 660, 722,

662, 644, 683, 695, 678, 674, 656, 667, 683, 691, 680, 685,
681, 715, 665, 676, 665, 675, 655, 659, 720, 675, 697, 663]

(a) Számoljuk ki a sebesség értékek **átlagát, szórását**, legkisebb- és legnagyobb értéket és a **terjedelmet**!

(b) Rajzoljuk fel az adatsor váltakozását az átlaggal együtt!

(c) Rajzoljuk fel a sebesség adatok gyakoriság **hisztogramját** a terjedelem $n=10$ egyenlő részre osztásával!

(d) A leggyakrabban előforduló sebesség érték nagyobb vagy kisebb, mint az átlag?

(e) Rajzoljuk be a gyakoriság hisztogramba annak a **normál eloszlásnak** a sűrűség függvényét, melynek várható értéke és szórása egyezik az adatok átlagával és szórásával!

(f) Vajon a félvezetők sebessége **normál eloszlást** követ-e 95%-os megbízhatósággal, ha $n=20$ részre osztjuk a terjedelmet?

(g) A félvezetők hány százalékának nagyobb a sebessége, mint 700 MegaHertz? Használjuk a FrequencyTable eljárást a számításhoz.

14.11.18. feladat.

A következő táblázat egy űrhajó hajtó motor O-gyűrűjének hőmérsékletére vonatkozó mérési adatokat tartalmazza Fahrenheit ($^{\circ}\text{F}$) mértékegységben.

(Az adatok "Presidential Commission on the Space Shuttle Challenger Accident" jelentésből valók)

homerseklet:=[84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31]

(a) Számítsuk ki a motor átlag hőmérsékletét és a hőmérséklet szórását!

(b) Melyik hőmérsékleti értéket mérték legtöbbször? Mennyi ennek a gyakorisága?

(c) Rajzoljuk fel azonos koordinátarendszerben a gyakoriság hisztogramot és az olyan normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása megegyezik az adatok várható értékével és szórásával!

(d) Teszteljük 95%-os megbízhatósági szinten, hogy a mérési adatok normál eloszlásúak-e!

14.11.19. feladat

Akkumulátorok élettartamát mutatja az alábbi táblázat években mérve.

akkumulátor:=[4.1, 5.2, 2.8, 4.9, 5.6, 4.0, 4.1, 4.3, 5.4,
4.5, 6.1, 3.7, 2.3, 4.5, 4.9, 5.6, 4.3, 3.9,
3.2, 5.0, 4.8, 3.7, 4.6, 5.5, 1.8, 5.1, 4.2,
6.3, 3.3, 5.8, 4.4, 4.8, 3.0, 4.3, 4.7, 5.1]

(a) Számoljuk ki az akkumulátorok élettartamának átlagát, szórását, legkisebb- és legnagyobb értéket és a terjedelmet!

(b) Rajzoljuk fel az adatsor váltakozását az átlaggal együtt!

(c) Rajzoljuk fel az élettartamok gyakoriság hisztogramját a terjedelem $n=10$ egyenlő részre osztásával!

(d) A leggyakrabban előforduló élettartam nagyobb vagy kisebb, mint az átlag?

(e) Rajzoljuk be a gyakoriság hisztogramba annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása egyezik az adatok átlagával és szórásával!

(f) Vajon az akkumulátorok élettartama normál eloszlást követ-e 95%-os megbízhatósággal, ha $n=10$ részre osztjuk a terjedelmet?

(g) Az akkumulátorok hány százaléka működik 5 év eltelte után is? Használjuk a FrequencyTable eljárást a számításhoz

14.11.20. feladat

Férfi mérnök hallgatók magasság értékeit jegyezték fel az USA egyik egyetemén. A magasság értékek inchben ($=2.54\text{ cm}$) mérve a következők:

magassag:=[69, 67, 69, 70, 65, 68, 69, 70, 71, 69, 66, 67, 69, 75, 68, 67, 68,
69, 70, 71, 72, 68, 69, 69, 70, 71, 68, 72, 69, 69, 68, 69, 73, 70,

73, 68, 69, 71, 67, 68, 65, 68, 68, 69, 70, 74, 71, 69, 70, 69]

- (a) Számoljuk ki a magasságok átlagát, szórását, legkisebb- és legnagyobb értéket és a terjedelmet!
- (b) Rajzoljuk fel az adatsor váltakozását az átlaggal együtt!
- (c) Rajzoljuk fel a magasságok gyakoriság hisztogramját a terjedelem 10 egyenlő részre osztásával!
- (d) A leggyakrabban előforduló magasság nagyobb vagy kisebb, mint az átlag?
- (e) Rajzoljuk be a gyakoriság hisztogramba annak a normál eloszlásnak a sűrűség függvényét, melynek várható értéke és szórása egyezik az adatok átlagával és szórásával!
- (f) Vajon a magasság értékek normál eloszlást követnek-e 95%-os megbízhatósággal, ha 10 részre osztjuk a terjedelmet?
- (g) A mérnök hallgatók hány százaléka magasabb 72 inch (=182.9 cm)-nél? Használjuk a FrequencyTable eljárást a számításhoz