



Statisztika feladatok

Informatikai Tudományok Doktori Iskola



feladat

Legyen az X valószínűségi változó adott. Legyen továbbá $X_1, X_2, \dots, X_n, \dots$ statisztikai minta, amelynek eloszlásfüggvénye X -ével azonos. Rögzítsük most az $x \in \mathbb{R}$ valós számot. Ekkor X eloszlásfüggvénye az x pontban a paraméter: $\vartheta = \vartheta(\mathbf{P}) = F_{\mathbf{P}}(x)$. Akkor az $F_n(x)$ empirikus eloszlásfüggvény értéke a ϑ eloszlásfüggvény-érték torzítatlan, négyzetes középben konzisztens becslése.

Bizonyítandó, hogy: $\mathbf{E}_{\mathbf{P}}(F_n(x)) = \vartheta \quad \sigma_{\mathbf{P}}^2(F_n(x)) \rightarrow 0$



Megoldás

$$0 \leq nF_n(x) \leq n$$

$$\mathbf{P}(nF_n(x) = k) =$$

$\mathbf{P}(k \text{ db } i \text{ indexre } X_i < x, (n-k) \text{ db } j \text{ indexre } j \neq i \text{ } X_j \geq x)$

$$= \binom{n}{k} [F_{\mathbf{P}}(x)]^k [1 - F_{\mathbf{P}}(x)]^{n-k}$$

$$\text{azaz } nF_n(x) \in B(n, \vartheta) \implies \mathbf{E}_{\mathbf{P}}(nF_n(x)) = n\vartheta$$

$$\sigma_{\mathbf{P}}^2(nF_n(x)) = n\vartheta(1-\vartheta) \implies \sigma_{\mathbf{P}}^2(F_n(x)) = \frac{\vartheta \cdot (1-\vartheta)}{n} \rightarrow 0$$



feladat

Legyen az X valószínűségi változó adott. Tegyük fel, hogy X egyenletes eloszlású valószínűségi változó a $[0, \vartheta]$ intervallumon, ahol $\vartheta > 0$ ismeretlen paraméter.

Most $\forall \mathbf{P} \in \mathcal{P}$ -re $F_{X,\vartheta}(x) = \frac{x}{\vartheta}$, $\frac{dF_{X,\vartheta}(x)}{dx} = f_{X,\vartheta}(x) = \frac{1}{\vartheta}$, $x \in (0, \vartheta)$,

$$\mathbf{E}_{\vartheta} X = \frac{\vartheta}{2}, \quad \sigma_{\vartheta}^2 X = \frac{\vartheta^2}{12}.$$

Tekintsük az alábbi statisztikákat:

$$T_1 = \frac{n+1}{n} X_n^*, \quad T_2 = X_n^* + X_1^*, \quad T_3 = \frac{n+1}{n-1} (X_n^* - X_1^*), \quad T_4 = 2\bar{X}_n$$

Igazoljuk, hogy torzítatlan statisztikák! Melyik a leghatásosabb közöttük?



Megoldás

$$\mathbf{E}_{\vartheta} T_4 = \mathbf{E}_{\vartheta} 2\bar{X}_n = 2\mathbf{E}_{\vartheta} \bar{X}_n = 2\mathbf{E}_{\vartheta} X = 2 \frac{\vartheta}{2} = \vartheta \implies T_4 \text{ torzítatlan.}$$

$$\sigma_{\vartheta}^2 T_4 = 4\sigma_{\vartheta}^2 \bar{X}_n = 4 \frac{\sigma_{\vartheta}^2 X}{n} = 4 \frac{\vartheta^2}{12n} = \frac{\vartheta^2}{3n} \rightarrow 0 \implies$$

T_4 négyzetes középben konzisztens.

Az X_n^* eloszlásfüggvénye:

$$\mathbf{P}(X_n^* < x) = [F_{X,\vartheta}(x)]^n = \left(\frac{x}{\vartheta}\right)^n, \quad x \in [0, \vartheta]$$

$\implies$ sűrűségfüggvénye

$$f_{n,\vartheta}(x) = n \frac{x^{n-1}}{\vartheta^n}, \quad x \in (0, \vartheta).$$



Megoldás

$$\mathbf{E}_{\vartheta} X_n^* = \int_0^{\vartheta} x f_{n,\vartheta}(x) dx = \int_0^{\vartheta} n \frac{x^n}{\vartheta^n} dx = n \frac{1}{\vartheta^n} \left[\frac{x^{n+1}}{n+1} \right]_0^{\vartheta} = \frac{n}{n+1} \vartheta$$

$$\implies \mathbf{E}_{\vartheta} T_1 = \vartheta, \text{ torzítatlan.}$$

$$\begin{aligned} \sigma_{\vartheta}^2 T_1 &= \mathbf{E}_{\vartheta} T_1^2 - (\mathbf{E}_{\vartheta} T_1)^2 = \left(\frac{n+1}{n}\right)^2 \int_0^{\vartheta} x^2 n \frac{x^{n-1}}{\vartheta^n} dx - \vartheta^2 = \\ &= \frac{(n+1)^2}{n} \frac{1}{\vartheta^n} \left[\frac{x^{n+2}}{n+2} \right]_0^{\vartheta} - \vartheta^2 = \frac{(n^2 + 2n + 1 - n^2 - 2n) \vartheta^2}{n(n+2)} = \frac{\vartheta^2}{n(n+2)} \rightarrow 0. \end{aligned}$$

$\implies T_1$ is négyzetes középben konzisztens.



Megoldás

Az X_1^* eloszlásfüggvénye:

$$\mathbf{P}(X_1^* < x) = 1 - [1 - F_{X,\vartheta}(x)]^n = 1 - \left(\frac{\vartheta - x}{\vartheta}\right)^n, \quad x \in [0, \vartheta]$$

$$\implies f_{1,\vartheta}(x) = n \frac{(\vartheta - x)^{n-1}}{\vartheta^n}, \quad x \in (0, \vartheta).$$

$$\mathbf{E}_\vartheta T_2 = \mathbf{E}_\vartheta X_n^* + \mathbf{E}_\vartheta X_1^* = \frac{n}{n+1}\vartheta + \int_0^\vartheta x f_{1,\vartheta}(x) dx = \frac{n}{n+1}\vartheta + \frac{n}{\vartheta^n} \int_0^\vartheta x(\vartheta - x)^{n-1} dx \stackrel{*}{=}$$

$$\stackrel{*}{=} \frac{n}{n+1}\vartheta - \frac{n}{\vartheta^n} \int_\vartheta^0 (\vartheta - y) y^{n-1} dy = \frac{n}{n+1}\vartheta + \frac{n}{\vartheta^{n-1}} \left[\frac{y^n}{n} \right]_0^\vartheta - \frac{n}{\vartheta^n} \left[\frac{y^{n+1}}{n+1} \right]_0^\vartheta = \vartheta,$$

azaz T_2 is torzítatlan.

$$\mathbf{E}_\vartheta T_3 = \frac{n+1}{n-1} (\mathbf{E}_\vartheta X_n^* - \mathbf{E}_\vartheta X_1^*) = \frac{n+1}{n-1} \left(\frac{n}{n+1}\vartheta - \frac{1}{n+1}\vartheta \right) = \vartheta \implies \text{torzítatlan.}$$



Megoldás

$$\sigma_\vartheta^2 T_2 = \sigma_\vartheta^2 X_n^* + \sigma_\vartheta^2 X_1^* + 2 \operatorname{cov}_\vartheta(X_n^*, X_1^*).$$

X_1^* és X_n^* nem függetlenek, így ki kell számolnunk a kovarianciájukat:

$$\begin{aligned} \mathbf{P}(X_1^* < x, X_n^* < y) &= \mathbf{P}(X_n^* < y) - \mathbf{P}(X_1^* \geq x, X_n^* < y) = \\ &= [F_{X,\vartheta}(y)]^n - \mathbf{P}(x \leq X_1 < y, x \leq X_2 < y, \dots, x \leq X_n < y) = \\ &= [F_{X,\vartheta}(y)]^n - \prod_{i=1}^n \mathbf{P}(x \leq X_i < y) = \\ &= [F_{X,\vartheta}(y)]^n - [F_{X,\vartheta}(y) - F_{X,\vartheta}(x)]^n, \quad x, y \in [0, \vartheta], \quad x < y. \end{aligned}$$

(Ez az együttes eloszlásfüggvényük.)



Megoldás

X_1^* és X_n^* együttes sűrűségfüggvénye így:

$$\begin{aligned} \frac{\partial^2 \mathbf{P}(X_1^* < x, X_n^* < y)}{\partial x \partial y} &= \\ &= n(n-1) [F_{X,\vartheta}(y) - F_{X,\vartheta}(x)]^{n-2} f_{X,\vartheta}(y) f_{X,\vartheta}(x) = n(n-1) \frac{(y-x)^{n-2}}{\vartheta^n}. \\ \operatorname{cov}_\vartheta(X_n^*, X_1^*) &= \int_0^\vartheta \int_0^y xy n(n-1) \left(\frac{y-x}{\vartheta}\right)^{n-2} \frac{1}{\vartheta^2} dx dy - \mathbf{E}_\vartheta X_n^* \mathbf{E}_\vartheta X_1^* = \\ &= \int_0^\vartheta \left(\int_0^y (y^2 - yu) n(n-1) \left(\frac{u}{\vartheta}\right)^{n-2} \frac{1}{\vartheta^2} du \right) dy - \frac{n}{(n+1)^2} \vartheta^2 = \\ &= \int_0^\vartheta \frac{y^{n+1}}{\vartheta^n} dy - \frac{n}{(n+1)^2} \vartheta^2 = \\ &= \frac{1}{n+2} \vartheta^2 - \frac{n}{(n+1)^2} \vartheta^2 = \frac{1}{(n+2)(n+1)^2} \vartheta^2. \end{aligned}$$



Megoldás

$$\begin{aligned}\mathbf{E}(X_1^*)^2 &= \vartheta^2 - \frac{2n}{n+1}\vartheta^2 + \frac{n}{n+2}\vartheta^2 = \\ &= \frac{(n^2 + 3n + 2 - 2n^2 - 4n + n^2 + n)\vartheta^2}{(n+1)(n+2)} = \frac{2}{(n+1)(n+2)}\vartheta^2, \\ \sigma_\vartheta^2 X_1^* &= \frac{2}{(n+1)(n+2)}\vartheta^2 - \frac{1}{(n+1)^2}\vartheta^2 = \frac{n}{(n+1)^2(n+2)}\vartheta^2. \\ \sigma^2 X_n^* &= \int_0^\vartheta x^2 \cdot n \cdot \frac{x^{n-1}}{\vartheta^n} dx - \left(\frac{n \cdot \vartheta}{n+1} \right)^2 = \frac{n}{n+2}\vartheta^2 - \frac{n^2}{(n+1)^2}\vartheta^2 = \\ &= \frac{n\vartheta^2}{(n+1)^2(n+2)}.\end{aligned}$$



Megoldás

$$\begin{aligned}\sigma_\vartheta^2 T_2 &= \sigma_\vartheta^2 X_n^* + \sigma_\vartheta^2 X_1^* + 2\mathbf{cov}_\vartheta(X_n^*, X_1^*) = \\ &= \frac{n}{(n+1)^2(n+2)}\vartheta^2 + \frac{n}{(n+1)^2(n+2)}\vartheta^2 + \frac{2}{(n+1)^2(n+2)}\vartheta^2 = \\ &= \frac{2}{(n+1)(n+2)}\vartheta^2 \rightarrow 0, \quad \Rightarrow T_2 \text{ is négyzetes középben konzisztens.} \\ \sigma_\vartheta^2 T_3 &= \left(\frac{n+1}{n-1} \right)^2 (\sigma_\vartheta^2 X_n^* + \sigma_\vartheta^2 X_1^* - 2\mathbf{cov}_\vartheta(X_n^*, X_1^*)) = \\ &= \frac{n}{(n-1)(n+2)}\vartheta^2 \rightarrow 0, \quad \Rightarrow T_3 \text{ is négyzetes középben konzisztens.}\end{aligned}$$

Az eredményt összegezve: $\sigma_\vartheta^2 T_1 < \sigma_\vartheta^2 T_2 < \sigma_\vartheta^2 T_3 < \sigma_\vartheta^2 T_4$.



feladat

Igazoljuk az alábbi állítást!

A lineáris statisztikák között a X_n statisztika a leghatásosabb, azaz, ha tetszőleges $c_1, c_2, \dots, c_n$, $\sum_{i=1}^n c_i = 1$ valós súlyokkal tekintjük a $t_n = \sum_{i=1}^n c_i \cdot X_i$ lineáris becslést, akkor t_n torzítatlan, és $\sigma_{\mathbf{P}}^2 \bar{X}_n \leq \sigma_{\mathbf{P}}^2 t_n$.

Először is megjegyezzük, hogy a $c_i = \frac{1}{n}$ ($i = 1, 2, \dots, n$) súlyválasztással az átlagstatisztikát kapjuk, tehát az átlagstatisztika is lineáris becslés.



Megoldás

Legyen $\varepsilon_i = c_i - \frac{1}{n}$ ($i = 1, 2, \dots, n$). Ekkor $\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n c_i - 1 = 0$.

$$\begin{aligned}\sigma_{\mathbf{P}}^2 \left(\sum_{i=1}^n c_i X_i \right) &= \sum_{i=1}^n c_i^2 \sigma_{\mathbf{P}}^2 X_i = \sigma_{\mathbf{P}}^2 X \cdot \sum_{i=1}^n c_i^2 = \sigma_{\mathbf{P}}^2 X \cdot \sum_{i=1}^n \left(\varepsilon_i + \frac{1}{n} \right)^2 = \\ &= \sigma_{\mathbf{P}}^2 X \left(\sum_{i=1}^n \varepsilon_i^2 + \frac{2}{n} \sum_{i=1}^n \varepsilon_i + \frac{1}{n} \right) \geq \frac{\sigma_{\mathbf{P}}^2 X}{n} = \sigma_{\mathbf{P}}^2 \bar{X}_n\end{aligned}$$

$0 \leq$

$= 0$



feladat

Legyen $X_1, X_2, \dots, X_n$ egy statisztikai minta. Adjuk meg az eloszlás ismeretlen paraméterének (paramétereinek) maximum likelihood becslését és momentumos becslését, ha a minta eloszlása

- $B(N, \vartheta)$, $\vartheta \in (0, 1)$, N adott természetes szám
- $Po(\vartheta)$, $\vartheta > 0$,
- $G(\vartheta)$, $\vartheta \in (0, 1)$,
- $E\left(\frac{1}{\vartheta}\right)$, $\vartheta > 0$,
- $N(\vartheta, \sigma_0)$, $\sigma_0 > 0$ adott, ϑ ismeretlen,
- $N(\vartheta_1, \vartheta_2)$, $\vartheta_2 > 0$, ϑ_1 ismeretlenek.



a.) A maximum likelihood statisztika számolása:

$$L(x_1, x_2, \dots, x_n, \vartheta) = \prod_{i=1}^n \binom{N}{x_i} \cdot \vartheta^{x_i} \cdot (1 - \vartheta)^{N - x_i}$$

$$l(x_1, x_2, \dots, x_n, \vartheta) = \sum_{i=1}^n \ln \binom{N}{x_i} + \ln \vartheta \cdot \sum_{i=1}^n x_i + \ln(1 - \vartheta) \cdot \sum_{i=1}^n (N - x_i)$$

$$\frac{\partial l}{\partial \vartheta} = \frac{\sum_{i=1}^n x_i}{\vartheta} - \frac{N \cdot n - \sum_{i=1}^n x_i}{1 - \vartheta} = 0$$

$$\sum_{i=1}^n x_i - \vartheta \cdot \sum_{i=1}^n x_i - \vartheta \cdot N \cdot n + \vartheta \cdot \sum_{i=1}^n x_i = 0$$

$$\vartheta = \frac{\bar{x}_n}{N}.$$

Tehát, a maximum likelihood statisztika: $\frac{\bar{x}_n}{N}$.

A becslés számítása a momentumok módszerével:

$$EX_1 = N \cdot \vartheta \approx \bar{x}_n \implies \vartheta \approx \frac{\bar{x}_n}{N}.$$

Tehát a momentumok módszerével ugyanazt a becslést kapjuk.



Megoldás

b.) A maximum likelihood statisztika számolása:

$$L(x_1, x_2, \dots, x_n, \vartheta) = \prod_{i=1}^n \frac{\vartheta^{x_i}}{x_i!} \cdot e^{-\vartheta},$$

$$l(x_1, x_2, \dots, x_n, \vartheta) = \ln \vartheta \cdot \sum_{i=1}^n x_i - \ln \prod_{i=1}^n x_i! - n \cdot \vartheta$$

$$\frac{\partial l}{\partial \vartheta} = \frac{\sum_{i=1}^n x_i}{\vartheta} - n = 0 \implies \vartheta = \bar{x}_n$$

Tehát, a maximum likelihood statisztika az átlag: $\bar{X}_n$.

A becslés számítása a momentumok módszerével:

$$\mathbf{E}X_1 = \vartheta \approx \bar{X}_n \implies \vartheta \approx \bar{X}_n.$$

A momentumok módszerével ugyanazt a statisztikát kaptuk.



Megoldás

c.) A maximum likelihood statisztika számolása:

$$L(x_1, x_2, \dots, x_n, \vartheta) = \prod_{i=1}^n (1 - \vartheta)^{x_i - 1} \cdot \vartheta,$$

$$l(x_1, x_2, \dots, x_n, \vartheta) = \ln(1 - \vartheta) \cdot \left(\sum_{i=1}^n x_i - n \right) + n \cdot \ln \vartheta,$$

$$\frac{\partial l}{\partial \vartheta} = \frac{n - \sum_{i=1}^n x_i}{1 - \vartheta} + \frac{n}{\vartheta} = 0,$$

$$\vartheta \cdot n - \vartheta \cdot \sum_{i=1}^n x_i + n - \vartheta \cdot n = 0 \implies \vartheta = \frac{1}{\bar{x}_n}.$$

Tehát, a maximum likelihood statisztika az átlag reciproka: $\frac{1}{\bar{X}_n}$.

A becslés számítása a momentumok módszerével:

$$\mathbf{E}X_1 = \frac{1}{\vartheta} \approx \bar{X}_n \implies \vartheta \approx \frac{1}{\bar{X}_n}.$$

A momentumok módszerével ugyanazt a statisztikát kaptuk.



Megoldás

d.) A maximum likelihood statisztika számolása:

$$L(x_1, x_2, \dots, x_n, \vartheta) = \prod_{i=1}^n \frac{1}{\vartheta} e^{-\frac{1}{\vartheta} \cdot x_i},$$

$$l(x_1, x_2, \dots, x_n, \vartheta) = -n \cdot \ln \vartheta - \frac{1}{\vartheta} \cdot \sum_{i=1}^n x_i,$$

$$\frac{\partial l}{\partial \vartheta} = -\frac{1}{\vartheta^2} \sum_{i=1}^n x_i - \frac{n}{\vartheta} = 0,$$

$$\sum_{i=1}^n x_i - \vartheta \cdot n = 0 \implies \vartheta = \bar{x}_n.$$

Tehát, a maximum likelihood statisztika az átlag: $\bar{X}_n$.

A becslés számítása a momentumok módszerével:

$$\mathbf{E}X_1 = \vartheta \approx \bar{X}_n \implies \vartheta \approx \bar{X}_n.$$

A momentumok módszerével ugyanazt a statisztikát kaptuk.



Megoldás

e.) A maximum likelihood statisztika számolása:

$$L(x_1, x_2, \dots, x_n, \vartheta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi} \cdot \sigma_0} \cdot e^{-\frac{(x_i - \vartheta)^2}{2 \cdot \sigma_0^2}},$$

$$l(x_1, x_2, \dots, x_n, \vartheta) = n \cdot \ln \left(\frac{1}{\sqrt{2\pi} \cdot \sigma_0} \right) - \frac{\sum_{i=1}^n (x_i - \vartheta)^2}{2 \cdot \sigma_0^2},$$

$$\frac{\partial l}{\partial \vartheta} = \frac{1}{\sigma_0^2} \cdot \sum_{i=1}^n (x_i - \vartheta) = 0 \implies \vartheta = \bar{x}_n.$$

Tehát, a maximum likelihood statisztika az átlag: $\bar{X}_n$.

A becslés számítása a momentumok módszerével:

$$\mathbf{E}X_1 = \vartheta \approx \bar{X}_n \implies \vartheta \approx \bar{X}_n.$$

A momentumok módszerével ugyanazt a statisztikát kaptuk.



Megoldás

f.) A maximum likelihood statisztikák számolása:

$$L(x_1, x_2, \dots, x_n, \vartheta_1, \vartheta_2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi} \cdot \vartheta_2} \cdot e^{-\frac{(x_i - \vartheta_1)^2}{2 \cdot \vartheta_2^2}},$$

$$l(x_1, x_2, \dots, x_n, \vartheta_1, \vartheta_2) = n \cdot \ln \left(\frac{1}{\sqrt{2\pi} \cdot \vartheta_2} \right) - \frac{\sum_{i=1}^n (x_i - \vartheta_1)^2}{2 \cdot \vartheta_2^2},$$

$$\frac{\partial l}{\partial \vartheta_1} = \frac{1}{\vartheta_2^2} \cdot \sum_{i=1}^n (x_i - \vartheta_1) = 0 \implies \vartheta_1 = \bar{x}_n,$$

$$\frac{\partial l}{\partial \vartheta_2} = -\frac{n}{\vartheta_2} + \frac{1}{\vartheta_2^3} \cdot \sum_{i=1}^n (x_i - \vartheta_1)^2 = 0 \implies \vartheta_2^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \vartheta_1)^2 = s_n^2.$$

Tehát, a maximum likelihood statisztikák az átlag és az empirikus szórás:

$\bar{X}_n, S_n$.

A becslések számítása a momentumok módszerével:

$$\mathbf{E}X_1 = \vartheta_1 \approx \bar{X}_n \implies \vartheta_1 \approx \bar{X}_n,$$

$$\mathbf{E}X_1^2 = \vartheta_2^2 + \vartheta_1^2 \approx \frac{1}{n} \sum_{i=1}^n X_i^2,$$

$$\vartheta_2^2 = \frac{\sum_{i=1}^n (X_i - \vartheta_1)^2}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n} = S_n^2.$$

A momentumok módszerével ugyanazokat a statisztikákat kaptuk.



feladat

Mutassuk meg, hogy az átlagstatisztika normális esetben nem csak torzítatlan, erősen konzisztens becslés, hanem hatásos is!

Feltételek:

Legyen az X valószínűségi változó adott. Legyen továbbá $X_1, X_2, \dots, X_n, \dots$ statisztikai minta, amelynek eloszlásfüggvénye X -ével azonos valamilyen m, D_0 paraméterű normális eloszláshoz tartozik, ahol $D_0 > 0$ ismert, m ismeretlen. Ennél a feladatnál az ismeretlen paraméter tehát a normális eloszlás várható értéke: $\vartheta = m = \mathbf{E}_P X$.

Bizonyítandó, hogy a várható értékre nincs kisebb szórású torzítatlan becslés a mintaátlagnál!



Megoldás

Ha $\sigma_{\vartheta}^2 t = \frac{1}{I_{\vartheta}(\vartheta)}$ akkor t biztosan hatásos statisztika!

Ez teljesül, ha $\exists v(\vartheta) : \frac{1}{L_{\vartheta}(\mathbf{x})} \frac{\partial L_{\vartheta}(\mathbf{x})}{\partial \vartheta} (= \frac{\partial \ln L_{\vartheta}(\mathbf{x})}{\partial \vartheta}) = v(\vartheta) \cdot (t(\mathbf{x}) - g(\vartheta))$

$I_n(\vartheta) = \sigma_{\vartheta}^2 \left(\frac{\partial \ln L_{\vartheta}(\mathbf{X})}{\partial \vartheta} \right)$ a Fisher-féle információ mennyiség

A minta együttes sűrűségfüggvénye, a likelihood függvény most:

$$L_m(\mathbf{x}) = \prod_{i=1}^n \varphi_{m, D_0}(x_i) = \left(\frac{1}{\sqrt{2\pi D_0}} \right)^n e^{-\frac{1}{2D_0} \sum_{i=1}^n (x_i - m)^2}$$



Megoldás

$$\ln L_m(\mathbf{x}) = \sum_{i=1}^n \ln \varphi_{m, D_0}(x_i) = -n \ln(\sqrt{2\pi D_0}) - \frac{1}{2D_0^2} \sum_{i=1}^n (x_i - m)^2,$$

$$\frac{\partial \ln L_m(\mathbf{x})}{\partial m} = \sum_{i=1}^n \frac{\partial}{\partial m} \ln \varphi_{m, D_0}(x_i) = \frac{1}{D_0^2} \sum_{i=1}^n (x_i - m) = \frac{n}{D_0^2} (\bar{x}_n - m)$$

Mivel teljesült a feltétel, az átlagstatisztika tényleg hatásos!



feladat

Mutassuk meg, hogy az átlagstatisztika exponenciális esetben is nem csak torzítatlan, erősen konzisztens becslés, hanem hatásos is!

Feltételek:

Legyen X egy valószínűségi változó. Legyen továbbá $X_1, X_2, \dots, X_n$ statisztikai minta, amelynek eloszlásfüggvénye X -ével azonos valamilyen ismeretlen $\lambda > 0$ paraméterű exponenciális eloszláshoz tartozik.

Bizonyítandó, hogy a várható értékre nincs kisebb szórású torzítatlan becslés a mintaátlagnál! A bizonyítást az előző példánál megmutatott módon végezzük.



Megoldás

$\vartheta = \frac{1}{\lambda} = \mathbf{E} \mathbf{p} X$. A minta együttes sűrűségfüggvénye most:

$$L_{\vartheta}(\mathbf{x}) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} = \frac{1}{\vartheta^n} e^{-\frac{1}{\vartheta} \sum_{i=1}^n x_i}$$

$$\begin{aligned} \frac{\partial \ln L_{\vartheta}(\mathbf{x})}{\partial \vartheta} &= \frac{\partial}{\partial \vartheta} \left(\ln \frac{1}{\vartheta^n} e^{-\frac{1}{\vartheta} \sum_{i=1}^n x_i} \right) = \frac{\partial}{\partial \vartheta} \left(-n \ln \vartheta - \frac{1}{\vartheta} \sum_{i=1}^n x_i \right) = -\frac{n}{\vartheta} + \frac{1}{\vartheta^2} \sum_{i=1}^n x_i = \\ &= \frac{n}{\vartheta^2} (\bar{x}_n - \vartheta) \implies \bar{x}_n \text{ hatásos becslés.} \end{aligned}$$



feladat

Mutassuk meg, hogy az átlagstatisztika Poisson esetben is nem csak torzítatlan, erősen konzisztens becslés, hanem hatásos is! Ebben a példában az alapsokaság eloszlása diszkrét!

Feltételek:

Legyen X diszkrét valószínűségi változó. Legyen továbbá $X_1, X_2, \dots, X_n$ statisztikai minta, amelynek eloszlásfüggvénye X -ével azonos valamilyen ismeretlen $\lambda > 0$ paraméterű Poisson-eloszláshoz tartozik. Ezzel a példánál legyen az ismeretlen paraméter a Poisson-eloszlás elméleti várható értéke: $\vartheta = \lambda = \mathbf{E} \mathbf{p} X$.

$L_{\vartheta}(\mathbf{x})$ most a minta együttes eloszlása lesz:

$$\begin{aligned} L_{\vartheta}(\mathbf{x}) &= \mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n \mathbf{P}(X_i = x_i) = \prod_{i=1}^n \frac{\vartheta^{x_i}}{(x_i)!} e^{-\vartheta} = \\ &= \frac{\vartheta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i)!} e^{-n\vartheta} \end{aligned}$$



Megoldás

A log-likelihood függvény most:

$$\ln L_{\vartheta}(\mathbf{X}) = (\ln \vartheta) \sum_{i=1}^n x_i - \ln \left(\prod_{i=1}^n (x_i)! \right) - n \cdot \vartheta$$

ϑ szerinti deriválás után:

$$\frac{\partial \ln L_{\vartheta}(\mathbf{x})}{\partial \vartheta} = \frac{1}{\vartheta} \sum_{i=1}^n x_i - n = \frac{n}{\vartheta} (\bar{x}_n - \vartheta) \implies \bar{x}_n \text{ hatásos becslése } \vartheta\text{-nak.}$$



feladat

Legyen most az $X_1, X_2, \dots, X_n$ minta eloszlása $U(0, \vartheta)$, ahol $\vartheta > 0$ ismeretlen paraméter. Láttuk a 3. példában, hogy a $T_1 = \frac{n+1}{n} X_n^*$ statisztika torzítatlan becslés volt $g(\vartheta) = \vartheta$ -ra, ahol $\sigma^2 T_1 = \frac{\vartheta^2}{n(n+2)}$. Számoljuk ki ebben az esetben az $\frac{1}{I_n(\vartheta)}$ információs alsó határt!

$$I_1(\vartheta) = \int_0^{\vartheta} \left(\frac{\partial}{\partial \vartheta} \frac{1}{\vartheta} \right)^2 dx = \frac{1}{\vartheta^2},$$

azaz

$$\frac{1}{I_n(\vartheta)} = \frac{1}{n I_1(\vartheta)} = \frac{\vartheta^2}{n}.$$

Az a meglepő eredményt kaptuk, hogy a T_1 torzítatlan becslés szórásnégyzete kisebb, mint a Cramer-Rao-tételben az információs alsó határ!

Mi lehet ennek az oka???



Megoldás

Az ellentmondás abból adódik, hogy az egyenletes eloszlás esetén nem teljesülnek a Cramer-Rao-tétel e) regularitási feltételei. Most

$$L_\vartheta(\mathbf{x}) = \left(\frac{1}{\vartheta} \right)^n, \quad \forall x_i \in (0, \vartheta),$$

és

$$\frac{\partial}{\partial \vartheta} \int_{\mathbb{R}^n} L_\vartheta(\mathbf{x}) d\mathbf{x} = 0,$$

míg

$$\int_{\mathbb{R}^n} \frac{\partial}{\partial \vartheta} L_\vartheta(\mathbf{x}) d\mathbf{x} = \int_0^{\vartheta} \int_0^{\vartheta} \dots \int_0^{\vartheta} -n \cdot \frac{1}{\vartheta^{n+1}} d\mathbf{x} = -\frac{n}{\vartheta}.$$

Tehát a Cramer-Rao-egyenlőtlenséggel nem igazolható most, hogy T_1 hatásos lenne! (Nem biztos, hogy hatásos!)



feladat

Mutassuk meg, hogy az átlagstatisztika elégséges normális esetben!

Legyen X valószínűségi változó. Legyen továbbá $X_1, X_2, \dots, X_n$ statisztikai minta, amelynek eloszlásfüggvénye X -ével azonos valamilyen m, D_0 paraméterű normális eloszláshoz tartozik $\forall \mathbf{P} \in \mathcal{P}$ -re, ahol $D_0 > 0$ ismert, m ismeretlen. Az ismeretlen paraméter a normális eloszlás elméleti várható értéke: $\vartheta = m = \mathbf{E}_{\mathbf{P}} X$.



Megoldás

Az átlagstatisztika teljesen független, $N(\frac{\theta}{n}, \frac{D_0}{n})$ eloszlású valószínűségi változók konvolúciója, tehát maga is normális eloszlású, θ és $\frac{D_0}{n}$ paraméterekkel. Így az $X_1, X_2, \dots, X_n$ minta együttes $X_n = y$ -ra vett feltételes sűrűségfüggvénye:

$$f_{X_1, X_2, \dots, X_n | \bar{X}_n}(x_1, x_2, \dots, x_n | y) = \begin{cases} \frac{f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)}{f_{\bar{X}_n}(y)}, & \text{ha } ny = \sum_{i=1}^n x_i \\ 0 & \text{egyébként} \end{cases}.$$

Mivel

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = (\sqrt{2\pi})^{-n} D_0^{-n} e^{-\frac{1}{2D_0^2} \sum_{i=1}^n (x_i - \theta)^2}$$

és

$$f_{\bar{X}_n}(y) = \frac{\sqrt{n}}{\sqrt{2\pi} D_0} e^{-\frac{n}{2D_0^2} (y - \theta)^2},$$



Megoldás

ezért

$$f_{X_1, X_2, \dots, X_n | \bar{X}_n}(x_1, x_2, \dots, x_n | y) = \begin{cases} \frac{1}{\sqrt{n}(\sqrt{2\pi} D_0)^n} e^{-\frac{1}{2D_0^2} \left[\sum_{i=1}^n (x_i - \theta)^2 - n(y - \theta)^2 \right]}, & \text{ha } ny = \sum_{i=1}^n x_i \\ 0 & \text{egyébként} \end{cases}.$$

Mivel $\sum_{i=1}^n (x_i - \theta)^2 - n(y - \theta)^2 = \sum_{i=1}^n x_i^2 - n y^2$, ha $\sum_{i=1}^n x_i = n y \implies$ a feltételes sűrűségfüggvény nem függ a paramétértől, amiből már következik az állítás.



feladat

Mutassuk meg, hogy az átlagstatisztika elégséges exponenciális esetben!

Legyen az X valószínűségi változó adott. Legyen továbbá $X_1, X_2, \dots, X_n$ statisztikai minta, amely eloszlásfüggvénye X -ével azonos valamilyen $\theta = \frac{1}{\lambda}$ paraméterű exponenciális eloszláshoz tartozik. Az ismeretlen paraméter tehát az exponenciális eloszlás várható értéke: $\mathbf{E}_\theta X = \theta$.



Megoldás

Az átlagstatisztika teljesen független, $E\left(\frac{1}{\vartheta}\right)$ eloszlású valószínűségi változók konvolúciója, eloszlása $n, \frac{n}{\vartheta}$ paraméterű gamma eloszlás, melynek sűrűségfüggvénye:

$$f_{\tilde{X}_n}(x) = \left(\frac{n}{\vartheta}\right)^n \frac{x^{n-1} e^{-\frac{nx}{\vartheta}}}{(n-1)!} \quad x > 0.$$

A minta együttes sűrűségfüggvénye most

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) = \frac{1}{\vartheta^n} e^{-\frac{\sum_{i=1}^n x_i}{\vartheta}} \quad \forall x_i > 0.$$

Az

$$f_{X_1, X_2, \dots, X_n | \tilde{X}_n}(x_1, x_2, \dots, x_n | y) = \begin{cases} \frac{f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)}{f_{\tilde{X}_n}(y)}, & \text{ha } ny = \sum_{i=1}^n x_i \\ 0 & \text{egyébként} \end{cases}$$

képletbe behelyettesítve:

$$f_{X_1, X_2, \dots, X_n | \tilde{X}_n}(x_1, x_2, \dots, x_n | y) = \begin{cases} \frac{\frac{1}{\vartheta^n} e^{-\frac{\sum_{i=1}^n x_i}{\vartheta}}}{\left(\frac{n}{\vartheta}\right)^n \frac{y^{n-1} e^{-\frac{ny}{\vartheta}}}{(n-1)!}}, & \text{ha } ny = \sum_{i=1}^n x_i \\ 0 & \text{egyébként} \end{cases}.$$



Megoldás

Egyszerűsítések után:

$$f_{X_1, X_2, \dots, X_n | \tilde{X}_n}(x_1, x_2, \dots, x_n | y) = \begin{cases} \frac{(n-1)! e^{-\frac{\sum_{i=1}^n x_i}{\vartheta}}}{n^n y^{n-1}}, & \text{ha } ny = \sum_{i=1}^n x_i \\ 0 & \text{egyébként} \end{cases} = \begin{cases} \frac{(n-1)!}{n^n y^{n-1}}, & \text{ha } ny = \sum_{i=1}^n x_i \\ 0 & \text{egyébként} \end{cases}.$$

Látható, hogy a függvény nem függ a paramétertől, azaz az átlagstatisztika ebben az esetben is elégséges becslést ad.



feladat

Mutassuk meg, hogy az átlagstatisztika elégséges Poisson esetben!

Legyen az X diszkrét valószínűségi változó adott. Legyen továbbá $X_1, X_2, \dots, X_n$ statisztikai minta, amelynek eloszlásfüggvénye X -ével azonos valamilyen $\vartheta > 0$ paraméterű Poisson-eloszláshoz tartozik. Az ismeretlen paraméter tehát a Poisson-eloszlás várható értéke: $\vartheta = \mathbf{E}_P X$.



Megoldás

Az átlagstatisztika eloszlása most:

$$\mathbf{P}(n \cdot \bar{X}_n = y) = \mathbf{P}\left(\sum_{i=1}^n X_i = y\right) = \frac{(n\theta)^y}{y!} e^{-n\theta} \quad y = 0, 1, 2, \dots$$

A minta együttes eloszlása:

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n \mathbf{P}(X_i = x_i) = \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\theta}.$$

Így a mintának az átlagra vonatkozó feltételes eloszlása:

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid n\bar{X}_n = y) = \frac{\prod_{i=1}^n \mathbf{P}(X_i = x_i)}{\mathbf{P}(n\bar{X}_n = y)} = \frac{y!}{n^y \prod_{i=1}^n x_i!},$$

ha $y = \sum_{i=1}^n x_i$, ami nem függ a paramétertől, azaz az átlagstatisztika a Poisson-eloszlás esetén is elégséges.



feladat

Mutassuk meg, hogy az első mintaelem önmagában nem elégséges!

Vizsgáljuk meg a $t = X_1$ „statisztikát”!



Megoldás

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid X_1 = y) = \begin{cases} \frac{\prod_{i=2}^n \mathbf{P}_\theta(X_i = x_i)}{\mathbf{P}_\theta(X_1 = y)}, & \text{ha } x_1 = y \\ 0, & \text{ha } x_1 \neq y \end{cases}$$

$$= \begin{cases} \prod_{i=2}^n \mathbf{P}_\theta(X_i = x_i), & \text{ha } x_1 = y \\ 0, & \text{ha } x_1 \neq y \end{cases}$$

ami látható, hogy tartalmazza a paramétert.



feladat

Konfidencia intervallum szerkesztése az ismeretlen szórásra normális eloszlás esetében.

Legyen $X_1, X_2, \dots, X_n$ $N(m, D)$ eloszlásból származó statisztikai minta, ahol $D > 0$ és $m \in \mathbb{R}$ is ismeretlen. Szerkesszünk D -re adott $0 < \varepsilon < 1$ mellett $(1 - \varepsilon)$ -szintű konfidenciaintervallumot!



Megoldás

A Lukács-tételre hivatkozva megint: $\frac{(n-1)s_n^{*2}}{D^2} \in \chi_{n-1}^2$.

Az $n-1$ szabadságfokú χ^2 -eloszlás táblázatból megadhatók olyan $0 < c_1 < c_2$ számok, hogy

$$1 - \varepsilon = \mathbf{P} \left(c_1 < \frac{(n-1)s_n^{*2}}{D^2} < c_2 \right)$$

teljesüljön. (A c_1, c_2 értékek nyilván kielégítik a $\mathbf{P}(\chi_{n-1}^2 > c_1) = 1 - \frac{\varepsilon}{2}$ és $\mathbf{P}(\chi_{n-1}^2 > c_2) = \frac{\varepsilon}{2}$ feltételeket.) Egyszerű átrendezéssel kapjuk, hogy

$$1 - \varepsilon = \mathbf{P} \left(\sqrt{\frac{(n-1)}{c_2}} s_n^* < D < \sqrt{\frac{(n-1)}{c_1}} s_n^* \right),$$

azaz a $T_1 = \sqrt{\frac{(n-1)}{c_2}} s_n^*$, $T_2 = \sqrt{\frac{(n-1)}{c_1}} s_n^*$ statisztikáknál $(1 - \varepsilon)$ -szintű konfidenciaintervallum lesz D -re.



feladat

Szerkesszünk $1 - \varepsilon$ megbízhatósági szintű konfidencia-intervallumot az exponenciális eloszlás λ paraméterére!

Használjuk fel az alábbi segédtelet:

Segédtelet: Legyen $X_1, X_2, \dots, X_n$ $E(\lambda)$ eloszlásból származó statisztikai minta. Ekkor

a) $\lambda X_i \in E(1)$,

b) $\sum_{j=1}^n \lambda X_j = \lambda n \bar{X}_n \in (n, 1)$, azaz $n, 1$ paraméterű gamma eloszlású,

$$f_{\Gamma}(x) = \frac{x^{n-1}}{(n-1)!} e^{-x} \quad (x > 0)$$

sűrűségfüggvényvel.

Az eloszlás neve: $n, 1$ paraméterű Gamma-eloszlás (Jelölés: $\Gamma(n, 1)$)



Megoldás

A segédítélet bizonyítása:

a) $\mathbf{P}(\lambda X_j < x) = \mathbf{P}(X_j < \frac{x}{\lambda}) = 1 - e^{-\lambda \frac{x}{\lambda}} = 1 - e^{-x} \Rightarrow \lambda X_j \in E(1).$

b) $\varphi_{\lambda X_j}(t) = \mathbf{E} e^{i\lambda X_j t} = \int_0^\infty e^{itx} e^{-x} dx = \left[\frac{1}{it-1} e^{x(it-1)} \right]_0^\infty = \frac{1}{1-it}.$

$\varphi_{\lambda n X_n}(t) = \prod_{j=1}^n \varphi_{\lambda X_n}(t) = \left(\frac{1}{1-it} \right)^n \Rightarrow f_{\lambda n X_n}(x) = \frac{x^{n-1} e^{-x}}{(n-1)!},$ mert a karakterisztikus függvénye:

$$\begin{aligned} \int_0^\infty e^{itx} \frac{x^{n-1}}{(n-1)!} e^{-x} dx &= \left[\frac{x^{n-1}}{(n-1)!} \frac{1}{it-1} e^{x(it-1)} \right]_0^\infty - \frac{1}{(n-2)!} \frac{1}{it-1} \int_0^\infty x^{n-2} e^{x(it-1)} dx = \\ &= 0 - \frac{1}{(n-2)!} \frac{1}{it-1} \left[x^{n-2} \frac{1}{it-1} e^{x(it-1)} \right]_0^\infty + \frac{1}{(n-3)!} \left(\frac{1}{it-1} \right)^2 \int_0^\infty x^{n-3} e^{x(it-1)} dx = \\ &= \frac{1}{(n-3)!} \left(\frac{1}{it-1} \right)^2 \int_0^\infty x^{n-3} e^{x(it-1)} dx = \dots = (-1)^n \left(\frac{1}{it-1} \right)^n = \left(\frac{1}{1-it} \right)^n. \end{aligned}$$



Megoldás

Ezek alapján a konfidencia-intervallum szerkesztése:

Az $n, 1$ paraméterű gamma-eloszláshoz tartozó táblázatból kiolvashatók olyan $0 < \gamma_1 < \gamma_2$ számok, amelyekkel

$$1 - \varepsilon = \mathbf{P}(\gamma_1 < \lambda n \bar{X}_n < \gamma_2) = \mathbf{P}\left(\frac{\gamma_1}{n \bar{X}_n} < \lambda < \frac{\gamma_2}{n \bar{X}_n}\right),$$

ahol az $T_1 = \frac{\gamma_1}{n \bar{X}_n}, T_2 = \frac{\gamma_2}{n \bar{X}_n}$ statisztika pár lesz $(1 - \varepsilon)$ -szintű konfidenciaintervallum λ -ra.

A γ_1, γ_2 számokat úgy kell meghatározni, hogy $\mathbf{P}(0 < \Gamma(n, 1) < \gamma_1) = \mathbf{P}(\Gamma(n, 1) > \gamma_2) = \frac{\varepsilon}{2}$ legyen.



feladat

Egy szövőgyárban a fonalak szakítószilárdságát mérik kp-ban. Egy 12 elemű mérés-sorozatban az átlagra $\bar{x}_{12} = 3,25$ kp adódott. A mérés szórása ismert, $\sigma_0 = 0,3$ kp.

a.) Állapítsuk meg a fonál szakítószilárdság konfidencia intervallumát 95%-os szinten.

b.) Legalább hány mérést kellene elvégezni ahhoz, hogy 95%-os megbízhatósági szint esetén a konfidencia intervallum hossza ne legyen 0,2 kp-nál nagyobb?

c.) Hogyan módosul a konfidencia intervallum, ha a mérés szórását nem ismerjük, a standard empirikus szórásra $s_{12} = 0,3$ kp adódik?

d.) Ebben az esetben legalább hány mérés kell a legfeljebb 0,2 kp szélességű, 95%-os megbízhatóságú konfidencia intervallumhoz?



Megoldás

a.) Jelölje ϑ a fonalak szakítószilárdságának várható értékét! A feltételek miatt $T = \frac{\bar{X}_{12} - \vartheta}{\sigma_0} \cdot \sqrt{12}$ standard normális eloszlást követ. Így $\mathbf{P}(|T| < u_{0,05}) = 2\Phi(u_{0,05}) - 1 = 0,95 \implies \Phi(u_{0,05}) = 0,975 \implies u_{0,05} = 1,96$. Tehát $\mathbf{P}\left(\left|\frac{\bar{X}_{12} - \vartheta}{\sigma_0} \cdot \sqrt{12}\right| < 1,96\right) =$

$$= \mathbf{P}\left(3,25 - \frac{1,96 \cdot 0,3}{\sqrt{12}} < \vartheta < 3,25 + \frac{1,96 \cdot 0,3}{\sqrt{12}}\right) =$$

$$= \mathbf{P}(3,08 < \vartheta < 3,42) = 0,95.$$

Tehát a 12 elemű mintából számolható 85%-os konfidencia intervallum $[3,08, 3,42]$, az intervallum hossza: 0,34 (kp).



Megoldás

b.) Ha a mintaelemszám n lenne, amiből az átlag $\bar{x}_n = 3,25$ kp lenne, akkor a 95%-os konfidencia intervallum hossza:

$$\frac{2 \cdot u_{0,05} \cdot \sigma_0}{\sqrt{n}} = \frac{1,176}{\sqrt{n}} < 0,2$$

$$\sqrt{n} > 5,88 \implies n \geq 35.$$

Azaz legalább 35 elemű minta kellene.



Megoldás

c.) Ha nem ismerjük a szórást, akkor a $T = \frac{\bar{X}_{12} - \vartheta}{s_{12}} \cdot \sqrt{12}$ Student eloszlású, 11 szabadságfokkal.

$$\mathbf{P}(|T| < t_{0,05}) = 0,95$$

A Student (t-) eloszlás táblázatból kiolvasható, hogy $t_{0,05} = 2,201$. Ezzel ϑ -ra a $[3,06, 3,44]$ konfidencia intervallum adódik, aminek hossza 0,38 kp.



Megoldás

d.) Ha az statisztikák n elemű mintából származtak volna, akkor a konfidencia intervallum hossza:

$$\frac{2 \cdot t_{0,05} \cdot s_{12}}{\sqrt{n}} = \frac{1,3206}{\sqrt{n}} < 0,2$$

$$\sqrt{n} > 6,653 \Rightarrow n \geq 45.$$

Azaz legalább 45 elemű minta kellene.



feladat

Egy gépsoron gyártott termék átmérője az /1/-es beállítás mellett $N(10, 1)$ a /2/-es beállítás mellett pedig $N(10.25, 1)$ eloszlást követ. Adva van a termékek egy halmaza, amelyről nem tudjuk, hogy az /1/ vagy a /2/ beállítás mellett gyártották. A technológus az alábbi döntési metódust alkalmazza: ha egy 25 elemű minta átlaga 10.1 alatt marad, akkor az /1/-es beállítás, különben a /2/-es beállítás mellett szavaz. Kérdések:

- Mennyi ekkor az /1/ és a /2/ melletti helytelen döntés valószínűsége?
- Hány elemű mintát kellene venni ahhoz, hogy a „helytelenül /2/-est tippelünk” esemény valószínűsége legfeljebb 0,01 legyen?
- Az előző pontban kiszámolt mintaelemszám mellett mennyi annak az eseménynek a valószínűsége, hogy „tévesen /1/-es mellett tippelünk”?
- 100 elemű minta esetén mi legyen az elválasztási érték (10,1 helyett), hogy annak a valószínűsége, hogy „helytelenül /1/-es beállításra szavazunk” 0,02 legyen?



Megoldás

a.) Az /1/ beállítás mellett helytelen döntés esetén a /2/-es beállítás a helytálló, tehát $X_i \in N(10.25, 1)$ és $\bar{X}_{25} \in N(10.25, \frac{1}{5})$. Mégis az történt, hogy teljesült az $\bar{X}_{25} < 10.1$ reláció, hiszen /1/ mellett döntöttünk. Ennek valószínűsége:

$$P(\bar{X}_{25} < 10.1) = \Phi\left(\frac{10.1 - 10.25}{0.2}\right) = \Phi(-0.75) = 1 - \Phi(0.75)$$

= 0,2266. Tehát az /1/ melletti hibás döntés valószínűsége: 0,2266.

Az /1/ igazsága esetén $X_i \in N(10, 1)$ és $\bar{X}_{25} \in N(10, \frac{1}{5})$.

Így

$$P(\bar{X}_{25} \geq 10.1) = 1 - P(\bar{X}_{25} < 10.1) = 1 - \Phi\left(\frac{10.1 - 10}{0.2}\right) = 1 - \Phi(0.5)$$

= 0,3085. Ez adja a /2/ beállítás mellett hibás döntés valószínűségét.



Megoldás

b.) Egy általános n elemű mintával számolva: $\bar{X}_n \in N\left(10, \frac{1}{\sqrt{n}}\right)$.

$$\mathbf{P}(\bar{X}_n \geq 10,1) = 1 - \mathbf{P}(\bar{X}_n < 10,1) = 1 - \Phi(0,1 \cdot \sqrt{n}) < 0,01,$$

$$0,99 < \Phi(0,1 \cdot \sqrt{n}).$$

Mivel $\Phi(2,34) = 0,9904$, így $0,1 \cdot \sqrt{n} \geq 2,34 \implies n \geq 548$.

Tehát legalább 548 megfigyelés kellene.



Megoldás

c.) Ha $n = 548$, akkor $\bar{X}_{548} \in N\left(10,25, \frac{1}{23,41}\right)$,

$$\begin{aligned}\mathbf{P}(\bar{X}_{548} < 10,1) &= \Phi\left(\frac{10,1 - 10,25}{\frac{1}{23,41}}\right) = \Phi(-3,5115) = \\ &= 1 - \Phi(3,5115) \approx 0,0003.\end{aligned}$$



Megoldás

d.) Milyen nagyságú K legyen a lüszób? $\bar{X}_{100} \in N\left(10,25, \frac{1}{10}\right)$,

$$K = ? : \mathbf{P}(\bar{X}_{100} < K) = \Phi\left(\frac{K - 10,25}{0,1}\right) \leq 0,02.$$

$$0,98 \approx \Phi(2,06) \leq \Phi(10 \cdot (10,25 - K))$$

Tehát $0,206 = 10,25 - K \implies K = 10,044$. Ha a vágási határ $K = 10,044$ lenne, teljesülne a hibás döntésre tett 0,02-es korlátozási feltétel.



feladat

Egy automata darabolónak 1200 mm hosszúságú acélszalagokat kell levágnia. Előzetes adatfelvételtől ellenőriztük, hogy a gép által készített darabok hossza normális eloszlású valószínűségi változónak tekinthető $\sigma_0 = 3$ mm szórással. Ellenőrizni akarjuk a gép beállításának helyes voltát. Ezért a gyártmányokból 16 db szalagot véletlenszerűen kiválasztunk, és lemérünk. Az adatok az alábbiak voltak mm-ben:

1193, 1196, 1198, 1195, 1198, 1199, 1204, 1193,
1203, 1201, 1196, 1200, 1191, 1196, 1198, 1191.

Vizsgálja meg, hogy van-e szignifikáns eltérés az előírt mérettől!



Megoldás

Mivel ismerjük a minta szórását, egymintás u-próbával dönthetünk a nullhipotézisről. Az átlagra: $\bar{x}_{16} = 1197,875$ adódik. Ha igaz a nullhipotézis, az átlagstatisztika eloszlása:

$$\bar{X}_{16} \in N\left(1200, \frac{3}{4}\right) \Rightarrow T = \frac{\bar{X}_{16} - 1200}{0,75} \in N(0, 1).$$

Döntünk $\varepsilon = 0,05$ szignifikancia szinten a nullhipotézisről. A kritikus értéket a

$$2\Phi(u_{0,05}) - 1 = 0,95 \iff \Phi(u_{0,05}) = 0,975$$

Összefüggésből számoljuk: $u_{0,05} = 1,96$. A T próbastatisztika számított értéke:

$$T = \frac{1197,875 - 1200}{0,75} = -2,83.$$

Mivel abszolút értékben ez nagyobb a kritikus értéknél, ezen a szignifikancia szinten elvetjük a nullhipotézist, vagyis a szalagok hossza jelentősen eltér a beállítási értéktől, újra kell kalibrálni a vágógépet.



feladat

Bizonyos gyártmány hosszúságát 1000 mm az előírt érték. Tudjuk, hogy a hosszúság ingadozása az előírt érték körül normális eloszlású. Tekintjük az alábbi 10 elemű mintát:

983, 1002, 998, 996, 1002, 983, 994, 991, 1005, 986.

Elfogadható-e az a nullhipotézis, hogy a hosszúság várhatóértéke 1000?





Megoldás

Nem ismert a minta szórása, ezért egymintás t-próbával dönthetünk a nullhipotézisről.

$$\bar{x}_{10} = 994, s_{10}^2 = 64,8, s_{10}^* = 8,055$$

A próbastatisztika számított értéke:

$$T = \frac{994 - 1000}{8,055} \approx -2,356$$

Az $\varepsilon = 0,05$ kritikus értékhez és $f = 9$ szabadsági fokhoz tartozó kritikus érték a Student-eloszlás táblázatából:

$$t_{0,05} = 2,262.$$

Mivel a számított érték abszolút értéke nagyobb a kritikus értéknél, ezért nem fogadjuk el a nullhipotézist, azaz a gyártott termékek hossza szignifikánsan eltérnek a hipotetikus értéktől. (Állítani kell a gyártósoron, mert nem az előírt méretű termékeket termelnek.)



feladat

A textiliparban minőségellenőrzésnél fontos annak a vizsgálata, hogy két, minta alapján minősített tétel azonos tulajdonságúnak tekinthető-e vagy sem. Az első $n_1 = 10$ elemű mintatételben az egységnyi fonalhossz tömegének átlagára $\bar{x}_{n_1} = 12,3$ gr, a második $n_2 = 12$ elemű mintatételben ugyanakkor az átlagra $\bar{x}_{n_2} = 11,6$ gr adódott. Feltételezve, hogy a tömegmérés pontossága (a szórás) $\sigma_0 = 18$ gr, döntünk 99%-os szignifikancia szinten arról, hogy a két tétel fonaltömegei azonosaknak tekinthetők-e!





Megoldás

Kétmintás u-próbával dönthetünk arról a nullhipotézisről, hogy a minták várható értékei nem különböznek jelentősen. A próbastatisztika számított értéke most:

$$\frac{\bar{x}_{n_1} - \bar{x}_{n_2}}{\sqrt{\frac{\sigma_0^2}{n_1} + \frac{\sigma_0^2}{n_2}}} = \frac{12,3 - 11,6}{\sqrt{\frac{9}{80}}} \approx 0,696.$$

Ha az $\varepsilon = 0,05$ szignifikancia szinten döntünk, akkor a kritikus érték $u_{0,05} = 1,65$ lesz. Látható, hogy a számított érték abszolút értéke ennél lényegesen kisebb, azaz a nullhipotézis elfogadható, vagyis a két tétel fonalsúlyainak különbsége nem szignifikánsan tér el egymástól.



feladat

Ki akarjuk mutatni, hogy egy kezelés amelyet állatokon végeztek, hatást gyakorol az állat testsúlyának növekedésére. Annál a $n_1 = 10$ állatnál, ahol nem volt speciális kezelés, az időegység alatt mért súlynövekedés

61, 52, 47, 51, 58, 64, 60, 55, 49, 53

(kg) volt. (Ez volt, az ún. kontrollcsoport.) Az állatok azon $n_2 = 12$ létszámú csoportjánál, ahol a kezelést elvégezték ugyanezen idő alatt a súlygyarapodások értékei

53, 59, 63, 67, 60, 57, 73, 65, 58, 68, 62, 71

(kg) voltak. Igaz-e az a nullhipotézis, hogy a kezelés nem növeli az állatok testsúlyát?





Megoldás

Ezúttal kétmintás t-próbát kell végrehajtanunk, hiszen a testsúlymérés pontosságát, azaz a minták szórásait nem ismerjük. Azonban, hogy a próbát alkalmazhassuk, előbb el kell végeznünk az F-próbát annak ellenőrzésére, hogy a minták azonos szórásúak-e. A mintákból az alábbi statisztikákat kapjuk:

$$n_1 = 10, n_2 = 12, \bar{x}_{n_1} = 55, \bar{x}_{n_2} = 63, s_{n_1}^2 = 31,1111, s_{n_2}^2 = 36.$$

Az F-statisztika számított értéke:

$$F = \frac{36}{31,1111} \approx 1,15714.$$

Az $\varepsilon = 0,05$ szignifikancia-szinthez, az $f_1 = 11, f_2 = 9$ szabadságfokokhoz tartozó kritikus érték az F táblázatból:

$$F_{krit} = 3,102485.$$

Mivel ennél a számított érték jóval kisebb, ezen a szinten a szórások között nem tapasztaltunk lényeges különbséget, azaz a kétmintás t-próba elvégezhető.



Megoldás

A próbastatisztika számított értéke most:

$$T = \frac{\bar{x}_{n_1} - \bar{x}_{n_2}}{\sqrt{\frac{f_1 \cdot s_{n_1}^2 + f_2 \cdot s_{n_2}^2}{f_1 + f_2}}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} = \frac{55 - 63}{\sqrt{\frac{280 + 396}{20}}} \cdot \sqrt{\frac{120}{22}} \approx -3,21.$$

Az $\varepsilon = 0,05$ szignifikancia-szinthez és az $f = f_1 + f_2 = 20$ szabadságfokokhoz tartozó kritikus érték a Student-táblázatból:

$$t_{krit} = 2,086.$$

Mivel a számított érték abszolút értéke nagyobb ennél, ezért ezen a szinten nem fogadhatjuk el azt a nullhipotézist, hogy a kezelés nem ér el súlynövekedést az állatoknál. Magyarul, a kezelés bevezethető, mert az állatok súlya növekedik tőle.



feladat

Egy konzervgyárban adagolóautomata tölti a dobozokat. Az egy dobozba töltendő anyag tömegének várható értékére az előírás $\mu = 500$ gr. Mintavétel során az alábbi értékeket kapták grammokban:

483, 502, 498, 496, 502, 483, 494, 491, 505, 486.

Döntsen 95%-os szignifikancia szinten, hogy teljesül-e a várható értékre az $\mu = 500$ gr előírás!





Megoldás

A kérdést egymintás t-próbával tudjuk eldönteni. Az adatokból számolt statisztikák:

$$n = 10, \bar{x}_{10} = 494, s_{10}^2 = 64,89, s_{10}^* = 8,05536.$$

A próbastatisztika számított értéke:

$$T = \frac{494 - 500}{8,05536} \cdot \sqrt{10} = -2,355$$

Az $\varepsilon = 0,05$ szignifikancia-szinthez és az $f = 9$ szabadságfokhoz tartozó kritikus érték a Student-táblázatból:

$$t_{krit} = 2,262.$$

Ezen a szignifikancia szinten nem fogadjuk el a nullhipotézist, a minta átlaga szignifikánsan eltér a beállítási értéktől.



feladat

Egy rovarirtószer hatását vizsgálják két különböző rovar (légy és darázs) esetében. A mérgez hatásfokát az ún. reakcióidővel mérik, azaz azt az időt jegyzik fel, ami alatt a méreggel reakcióba került rovar lebénul, és a hátára esik. Korábbi vizsgálatokban igazolták, hogy a reakcióidők logaritmus normális eloszlást követ. A légyek esetében egy $n_1 = 14$ elemű, a darázsok esetében egy $n_2 = 15$ elemű mintát kaptak. Az adatokat nagyság szerint rendezve olvashatjuk az alábbi táblázatban:

legyek adata	0,477	0,699	0,845	0,954	0,954	1	1,079	1,301	1,38	1,532	1,634	1,663	1,763	2,146	
darázsok adata	0,301	0,699	0,699	0,845	0,903	0,954	1,146	1,255	1,38	1,415	1,415	1,532	1,568	1,623	1,954

Van-e szignifikáns különbség a két rovar esetében a hatás szempontjából?





Megoldás

A megoldást független két-mintás t-próbával végezzük el. Először meg kell győződni a minták szórásnégyzeteinek egyezéséről F-próbával. A statisztikák most:

$$\bar{x}_{n_1} = 1,245, \bar{x}_{n_2} = 1,18, s_{n_1}^2 = 0,2, s_{n_2}^2 = 0,18$$

A próbastatisztika számított értéke:

$$F = \frac{s_{n_1}^2}{s_{n_2}^2} = \frac{0,2}{0,18} \approx 1,11$$

Az $\varepsilon = 0,05$ szignifikancia szinthez, az $f_1 = 13, f_2 = 14$ szabadságfokokhoz tartozó kritikus érték az F-eloszlás táblázatából: $F_{krit} = 2,51$. Mivel $F < F_{krit}$, ezért a két minta varianciájára tett nullhipotézist elfogadjuk.

A kétmintás t-póba statisztikájának számított értéke:

$$T = \frac{\bar{x}_{n_1} - \bar{x}_{n_2}}{\sqrt{(n_1 - 1) \cdot s_{n_1}^2 + (n_2 - 1) \cdot s_{n_2}^2}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2}} \approx -0,4$$

Az $\varepsilon = 0,05$ szignifikancia szinthez és az $n_1 + n_2 - 2 = 27$ szabadsági fokhoz tartozó kritikus értéket a Student-táblázatból olvashatjuk ki: $t_{krit} = 2,052$. Mivel $|T| < t_{krit}$, a nullhipotézist elfogadhatjuk, azaz a két rovarra gyakorolt hatása a rovarirtószernek azonos.



feladat

Egy klinikán két altatószer hatását hasonlítják össze. Összesen $n = 15$ kísérleti személynek (k.sz.) beadják először az A , majd néhány hét múlva a B altatót, és mérik hány perc alatt hat a gyógyszer, mennyi idő alatt kerül a k.sz. a mélyalvás stádiumába a beszédés után. Az adatokat az alábbi táblázatban láthatjuk:

k.sz. sorszáma	1	2	3	4	5	6	7
A hatása (p)	19,81	15,48	16,53	20,13	13,45	16,37	16,07
B hatása (p)	12,52	15,26	12,77	11,02	14,99	12,61	17,62

k.sz.	8	9	10	11	12	13	14	15
A	13,70	12,65	17,28	13,09	18,38	16,29	15,42	13,15
B	17,78	14,53	17,86	18,74	17,30	16,98	16,79	13,06

Van-e szignifikáns különbség a két gyógyszer hatása között? A döntést $\varepsilon = 0,05$ szignifikancia-szinten hozzuk meg!





Megoldás

A nullhipotézisről párosított t-próbával döntünk. Ha X_i jelöli az A gyógyszerhez tartozó adatokat, Y_i pedig a B gyógyszerre vonatkozó adatokat ($i = 1, 2, \dots, 15$), akkor a $Z_i = X_i - Y_i$ mintára kell az $m_0 = 0$ hipotetikus értékhez tartozó egymintás t-próbát elvégeznünk. Az adatokból számolható statisztikák:

$$\begin{aligned} n &= 15 \\ \bar{x}_{15} &= 15,44, \bar{y}_{15} = 16,36, \bar{z}_{15} = -0,92 \\ s_{Z,15}^2 &= 3,76 \\ f &= 14, \text{ a szabadságfok} \end{aligned}$$

A próbastatisztika számított értéke:

$$T = \frac{\bar{z}_{15}}{s_{Z,15}} \cdot \sqrt{15} \approx -0,947.$$

Az $\varepsilon = 0,05$ szignifikancia szinthez és az $f = 14$ szabadsági fokhoz tartozó kritikus értéket a Student-táblázatból olvashatjuk ki: $t_{krit} = 2,145$. Mivel $|T| < t_{krit}$, a nullhipotézist elfogadhatjuk, azaz a két altatószer azonos hatásfokú. (Ezek alapján, ha felmerül, az olcsóbbat fogja támogatni a TB...)



feladat

Valaki sokszor játszik ne neved korán társasjátékot és szorgalmasan feljegyezte egy táblázatba, hogy $n = 1000$ dobás alatt milyen gyakorisággal kapott különböző értékeket. A gyakoriság táblázat alapján állíthatjuk-e $\varepsilon = 0,05$ szignifikancia-szinten, hogy a dobókocka szabályos?

dobott érték	1	2	3	4	5	6
dobott gyakoriság	182	154	162	175	151	176





Megoldás

Tiszta illeszkedésvizsgálatot hajtunk végre χ^2 -próbával. A csoportok száma: $r = 6$. A hipotetikus eloszlás diszkrét egyenletes, ha a kocka tényleg szabályos: $p_i = \frac{1}{6}$, $i = 1, 2, 3, 4, 5, 6$. A próbat statisztika számítása:

$$T = \frac{(182 - 166,7)^2}{166,7} + \frac{(154 - 166,7)^2}{166,7} + \frac{(162 - 166,7)^2}{166,7} + \frac{(175 - 166,7)^2}{166,7} + \frac{(151 - 166,7)^2}{166,7} + \frac{(176 - 166,7)^2}{166,7} \approx 4,915.$$

Az $\varepsilon = 0,05$ szignifikancia-szinthez és az $f = 5$ szabadságfokokhoz tartozó kritikus érték a χ^2 -táblázatból:

$$c_{krit} = 11,05.$$

Mivel $T < c_{krit}$ teljesül, elfogadjuk, hogy a kocka szabályos.



feladat

Egy textilüzemben korábbi tapasztalatok azt mutatták, hogy a fonalszakadások száma egy bizonyos géptípus és fonal esetén Poisson eloszlású $\lambda = 8$ paraméterrel. Vizsgáljuk meg újabb adatokkal, hogy fennáll-e a Poisson eloszlás!

A fonalszakadások száma	Gyakoriság	Az elméleti Poisson eo. érték
0	-	0.00033
1	-	0.00268
2	-	0.01073
3	-	0.02863
4	1	0.05725
5	4	0.09160
6	9	0.12214
7	7	0.13959
8	8	0.13959
9	10	0.12408
10	8	0.09926
11	10	0.07219
12	5	0.04513
13	7	0.02962
14	2	0.01692
15	3	0.00903
16	1	0.00451
17	-	0.00212
18	-	0.00160
Összesen	75	1



Megoldás...

Tiszta illeszkedésvizsgálatot hajtottunk végre a $Po(8)$ -eloszláshoz χ^2 -próbával. A táblázat alapján $r = 5$ kategóriát állapítottunk meg összevonással. Az összevonásokat azért tesszük, hogy az egyes csoportokba eső mintaelem-számok meghaladhassák a 10-et.

összevont értékek	csoportgyakoriság	csoportvalószínűség	$n \cdot p_i$
0 – 6	14	0,31336	23,502
7 – 8	15	0,27918	20,9385
9 – 10	18	0,22334	16,7505
11 – 12	15	0,12032	9,024
13 – ∞	13	0,0638	4,785



Megoldás

Ebből a táblázatból számolható a próbastatisztika:

$$T = \sum_{i=1}^5 \frac{(\nu_i - n \cdot p_i)^2}{n \cdot p_i} =$$

$$= \frac{(14 - 23,502)^2}{23,502} + \frac{(15 - 20,9385)^2}{20,9385} + \frac{(18 - 16,7505)^2}{16,7505} +$$

$$+ \frac{(15 - 9,024)^2}{9,024} + \frac{(13 - 4,785)^2}{4,785} \approx 23,68.$$

Az $\varepsilon = 0,05$ szignifikancia-szinthez és az $f = 4$ szabadságfokhoz tartozó kritikus érték a χ^2 -táblázatból:

$$c_{krit} = 9,488.$$

Mivel a számított érték ennél jóval nagyobb, az a döntés hozható, hogy a minta nem illeszkedik a $Po(8)$ -eloszláshoz!



feladat

A megadott százéves minta alapján döntünk arról a nullhipotézisről, hogy Budapest levegőjének januári középhőmérséklete normális eloszlást követ vagy sem! Az ellenőrzést a Kolmogorov-Szmirnov próbával végezzé el! A levegő január havi közepes hőmérséklete, Budapest, 1861-1960 [°C]

1861	-0,2	1890	-1,5	1920	-0,9
1862	-2,7	1891	-0,4	1921	-0,2
1863	1,9	1892	0,1	1922	0,9
1864	-7,9	1893	0,2	1923	0,4
1865	0,3	1894	7,7	1924	-1,4
1866	0,1	1895	0,9	1925	-2,1
1867	6,7	1896	-0,9	1926	-1,0
1868	-0,9	1897	1,0	1927	-0,9
1869	-2,7	1898	-1,1	1928	4,9
1870	-0,0	1899	-2,1	1929	-2,9
1871	-2,0	1900	-0,9	1930	0,3
1872	-0,4	1901	-0,9	1931	1,7
1873	1,7	1902	-0,0	1932	-0,0
1874	-1,1	1903	-2,4	1933	-2,1
1875	-1,1	1904	-0,0	1934	-0,0
1876	-0,7	1905	0,9	1935	-0,9
1877	1,7	1906	1,0	1936	-0,0
1878	-2,4	1907	-0,0	1937	-0,0
1879	-2,4	1908	-0,0	1938	-0,0
1880	-2,4	1909	-0,0	1939	-0,0
1881	-0,1	1910	-0,0	1940	-0,0
1882	0,0	1911	0,0	1941	0,0
1883	0,0	1912	0,0	1942	0,0
1884	0,0	1913	0,0	1943	0,0
1885	0,0	1914	0,0	1944	0,0
1886	0,0	1915	0,0	1945	0,0
1887	0,0	1916	0,0	1946	0,0
1888	0,0	1917	0,0	1947	0,0
1889	0,0	1918	0,0	1948	0,0
1890	0,0	1919	0,0	1949	0,0
1891	0,0	1920	0,0	1950	0,0
1892	0,0	1921	0,0	1951	0,0
1893	0,0	1922	0,0	1952	0,0
1894	0,0	1923	0,0	1953	0,0
1895	0,0	1924	0,0	1954	0,0
1896	0,0	1925	0,0	1955	0,0
1897	0,0	1926	0,0	1956	0,0
1898	0,0	1927	0,0	1957	0,0
1899	0,0	1928	0,0	1958	0,0
1900	0,0	1929	0,0	1959	0,0
1901	0,0	1930	0,0	1960	0,0



Megoldás...

Az illeszkedés tesztjét Kolmogorov–Szmirnov-próbával fogjuk végrehajtani. Az adatokból az átlagra és szórásra kapott értékek:

$$\bar{x}_{100} = -1,028, s_{100} = 2,789232.$$

Ebből megadható a normális eloszlás hipotetikus eloszlásfüggvénye:

$$F(x) = \Phi\left(\frac{x + 1,028}{2,789232}\right).$$

A minta rendezése, az empirikus eloszlásfüggvény számolása után kapjuk a próbat statisztika számított értékét:

$$T = \sqrt{n} \cdot \sup |F_n(x) - F(x)| = 10 \cdot 0,053772 = 0,53772.$$

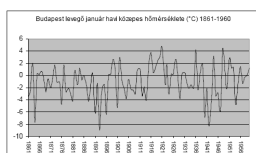


Megoldás

A Kolmogorov–Szmirnov-próba K_{krit} kritikus értékeit az alábbi táblázat tartalmazza:

$1 - \varepsilon$	K_{krit}
0,9	1,23
0,95	1,36
0,99	1,63
0,999	1,96

Az $\varepsilon = 0,05$ szignifikancia-színthez tartozó kritikus érték: $c_{krit} = 1,36$. Ez alapján igazolni tudtuk az hőmérsékleti adatsornak a normális eloszláshoz való illeszkedését.





feladat

Az előző adatsoron (budapesti januári középhőmérsékletek) ellenőrizze a normalitás teljesülését χ^2 -próbával is.



Megoldás

A várható értéket és a szórást a mintából becsüljük, tehát becsléses illeszkedésvizsgálatról lesz szó, a becsült paraméterek száma 2, amivel majd csökkenteni kell a szabadságfokot. $m \approx \bar{x}_{100} = -1,02, \sigma \approx s_{100}^* = 1,67$. A hipotetikus eloszlásfüggvény:

$$F_0(x) = \Phi\left(\frac{x + 1,02}{1,67}\right).$$

A csoportok számát $r = 8$ -ban megszabva, a csoportbaesés valószínűségeit az alábbi képlettel számoljuk:

$$p_i = F_0(C_i) - F_0(C_{i-1}).$$



Megoldás

i	$[C_{i-1}, C_i)$ az intervallum	v_i a gyakoriság	p_i a valószínűség
1	$[-10, -4)$	13	0,1466
2	$[-4, -3)$	9	0,0966
3	$[-3, -2)$	15	0,1243
4	$[-2, -1)$	14	0,1458
5	$[-1, 0)$	9	0,1453
6	$[0, 1)$	16	0,1237
7	$[1, 2)$	12	0,0948
8	$[2, 5)$	12	0,1229

A próbatatiztikát ezekből az adatokból számoljuk:

$$T = \sum_{i=1}^r \frac{(\nu_i - n \cdot p_i)^2}{n \cdot p_i} \approx 7,243.$$

Az $\varepsilon = 0,05$ szignifikancia-szinthez és a $f = r - 1 - 2 = 5$ szabadsági fokokhoz tartozó kritikus érték: $c_{krit} = 11,07$. Ez alapján igazolni tudtuk az hőmérsékleti adatsornak a normális eloszláshoz való illeszkedését, hiszen $T < c_{krit}$.



feladat

Vizsgálja meg a homogenitásvizsgálatra vonatkozó χ^2 -próbával, hogy a budapesti januári középhőmérsékletek 1861-1910 és 1911-1960 között azonos eloszlásúnak tekinthetők-e?



Megoldás

A mintanagyságok most egyformák: $n_1 = n_2 = 50$. Az értékészletet $r = 4$ csoportba osztva az adatok:

i	$[C_{i-1}, C_i)$	ν_i	μ_i
1	$[-10, -3)$	12	10
2	$[-3, -1)$	17	12
3	$[-1, +1)$	15	10
4	$[+1, 5)$	6	18

Ebből a táblázatból számolható ki a próbastatisztika:

$$T = n_1 \cdot n_2 \sum_{i=1}^r \frac{\left(\frac{\nu_i}{n_1} - \frac{\mu_i}{n_2}\right)^2}{\nu_i + \mu_i} \approx 8,045.$$

Az $\varepsilon = 0,05$ szignifikancia-szinthez és a $f = r - 1 = 3$ szabadsági fokhoz tartozó kritikus érték: $c_{krit} = 7,81$. Ez alapján elvetjük, hogy az adatsor első 50 évének eloszlása megegyezne a második 50 év eloszlásával, hiszen $T > c_{krit}$. (Úgy tűnik, 50 év alatt megváltozott Budapest mikroklimája...)



feladat

Egy újonnan kifejlesztett müzli ötféle magot (A, B, C, D és E) tartalmaz, melyek százalékos megoszlása a terméken lévő tájékoztató szerint 35%, 25%, 20%, 10%, illetve 10%. Egy véletlenül kiválasztott zacskóban az alábbi mennyiségi megoszlást találtuk:

Összetevő	A	B	C	D	E
Szem (db)	184	145	100	68	63



A: 35%
B: 25%
C: 20%
D: 10%
E: 10%



Döntsünk $\varepsilon = 0,1$ szignifikancia-szinten arról, hogy a minta összetétele megfelel-e a csomagoláson feltüntetett eloszlásnak!



Megoldás

A diszkrét eloszlások illeszkedésére vonatkozó χ^2 -próbával dönthetünk a kérdésről. Összesen $n = 560$ szem volt a zacskóban, az elméleti eloszlás, amihez az illeszkedést ellenőrizzük:

$$p_A = 0,35, p_B = 0,25, p_C = 0,2, p_D = 0,1, p_E = 0,1$$

A csoportok száma most $r = 5$. A próbastatisztika számított értéke:

$$\begin{aligned} T &= \frac{(184 - 560 \cdot 0,35)^2}{560 \cdot 0,35} + \frac{(145 - 560 \cdot 0,25)^2}{560 \cdot 0,25} + \frac{(100 - 560 \cdot 0,2)^2}{560 \cdot 0,2} + \\ &\quad + \frac{(68 - 560 \cdot 0,1)^2}{560 \cdot 0,1} + \frac{(63 - 560 \cdot 0,1)^2}{560 \cdot 0,1} = \\ &= \frac{144}{196} + \frac{25}{140} + \frac{144}{112} + \frac{144}{56} + \frac{169}{56} \approx 5,6454. \end{aligned}$$

Az $\varepsilon = 0,1$ szignifikancia-szinthez és az $f = r - 1 = 4$ szabadsági fokhoz tartozó kritikus érték: $c_{krit} = 7,779$. Mivel $T < c_{krit}$, elfogadjuk, hogy a megvizsgált zacskó tartalma megfelel a meghirdetett eloszlásnak.



feladat

Egy kutatócsoport azt vizsgálta, van-e összefüggés egy bizonyos betegség lefolyásának súlyossága és a betegek életkora között. A vizsgálat során $n = 200$ beteg adatait gyűjtötték össze, majd azokat csoportosították a betegség súlyossági foka és a páciens életkora szerint. Eredményül az alábbi táblázatot kapták:

		Életkor		
		40 alatti	40-60	60 fölötti
Lefolyás	enyhe	41	34	9
	közepes	25	25	12
	súlyos	6	33	15

A függetlenségre vonatkozó χ^2 -próbával döntünk $\varepsilon = 0,01$ szignifikancia szinten, hogy van-e összefüggés a betegek életkora és a betegség lefolyásának súlyossága között!



Megoldás

A kialakított csoportok száma: $r = 3, s = 3$. A próbatasztikához ki kell számolni a peremösszegeket.

Megfigyelt gyakoriságok ($k_{ij}, i = 1, \dots, r, j = 1, \dots, s$):

41	34	9	84
25	25	12	62
6	33	15	54
72	92	36	200



Megoldás

Várt gyakoriságok ($k_{ij}^* = \frac{k_{i \cdot} \cdot k_{\cdot j}}{n}, i = 1, \dots, r, j = 1, \dots, s$):

30.24	38.64	15.12	84
22.32	28.52	11.16	62
19.44	24.84	9.72	54
72	92	36	200

A próbatasztika ezekből az adatokból számolva:

$$T = \sum_{i=1}^r \sum_{j=1}^s \frac{(k_{ij} - k_{ij}^*)^2}{k_{ij}^*} \approx 22.5230.$$



Megoldás

Ha független a betegség foka a kortól, azaz igaz a függetlenségre tett null-hipotézis, akkor T eloszlása $f = (r - 1)(s - 1) = 4$ szabadságfokú χ^2 -eloszást kell követnie. Az $\varepsilon = 0,01$ szignifikancia-szinthez és az $f = 4$ szabadsági fokhoz tartozó kritikus értéket a táblázatból olvashatjuk ki: $c_{krit} = 13,277$. Látható, hogy még ezen az „enyhe” szignifikancia-szinten is $T > c_{krit}$ teljesül, azaz a betegség súlyossága összefügg a páciens korával.



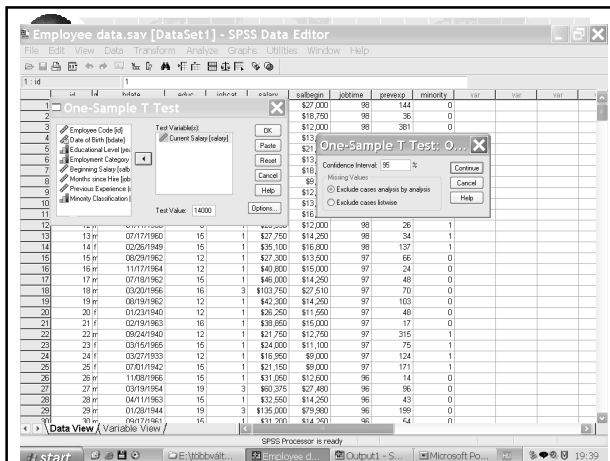
feladat

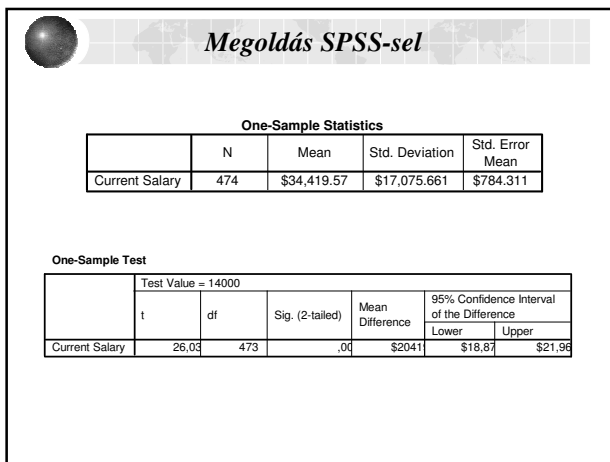
Banki alkalmazottak személyi adatait tartalmazó állomány 474 esetből álló *employee data* adatmátrixa. Ellenőrizzük azt a feltevést, hogy az átlagfizetés 14 000 \$

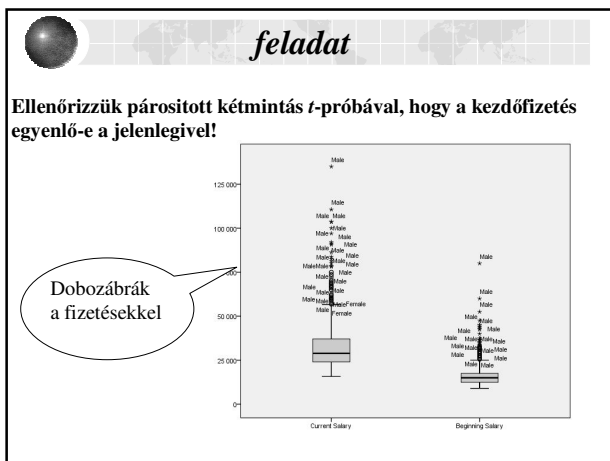
A változók jelentése

id	a dolgozó kódja
gender	a dolgozó neme (m-férfi, f-nő)
bdate	születési dátum
educ	képzési szint (években)
jobcat	beosztás (1-tisztviselő, 2-biztonsági, 3-menedzser)
salary	jelenlegi fizetés
salbegin	kezdőfizetés
jobtime	hány hónapja alkalmazták
prevexp	betanítási idő (hónapokban)
minority	hátrányos helyzet (0-nincs, 1-van)

id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority
1	m	02/03/1962	12	1	\$40,300	\$14,250	97	103	0
2	m	06/23/1968	12	1	\$36,250	\$11,250	97	48	0
3	f	07/26/1929	16	1	\$39,650	\$15,000	97	17	0
4	f	04/15/1947	15	1	\$21,750	\$12,750	97	315	1
5	m	02/09/1965	15	1	\$24,000	\$11,100	97	75	1
6	m	06/22/1968	12	1	\$16,250	\$9,000	97	124	1
7	m	04/26/1966	15	1	\$21,150	\$9,000	97	171	1
8	f	05/05/1965	15	1	\$31,050	\$12,800	96	14	0
9	f	01/23/1946	19	3	\$60,375	\$27,600	96	96	0
10	f	02/13/1946	15	1	\$32,550	\$14,250	96	43	0
11	f	02/07/1960	12	1	\$19,000	\$79,900	96	199	0
12	m	01/11/1966	15	1	\$31,000	\$14,250	96	14	0
13	m	07/17/1960	12	1	\$20,000	\$11,100	97	75	1
14	f	02/26/1949	19	3	\$60,375	\$27,600	96	96	0
15	m	06/29/1962	12	1	\$20,000	\$11,100	97	75	1
16	m	11/17/1964	15	1	\$20,000	\$11,100	97	75	1
17	m	01/09/1962	12	1	\$20,000	\$11,100	97	75	1
18	m	03/20/1966	12	1	\$20,000	\$11,100	97	75	1
19	m	06/19/1962	12	1	\$20,000	\$11,100	97	75	1
20	m	01/23/1940	12	1	\$36,250	\$11,250	97	48	0
21	f	02/19/1963	16	1	\$39,650	\$15,000	97	17	0
22	m	09/24/1940	12	1	\$21,750	\$12,750	97	315	1
23	m	03/15/1965	15	1	\$24,000	\$11,100	97	75	1
24	m	03/27/1933	12	1	\$16,250	\$9,000	97	124	1
25	m	07/01/1942	15	1	\$21,150	\$9,000	97	171	1
26	m	11/09/1965	15	1	\$31,050	\$12,800	96	14	0
27	m	03/19/1964	19	3	\$60,375	\$27,600	96	96	0
28	m	04/11/1963	15	1	\$32,550	\$14,250	96	43	0
29	m	01/29/1944	19	3	\$195,000	\$79,900	96	199	0
30	m	06/17/1963	15	1	\$31,000	\$14,250	96	14	0





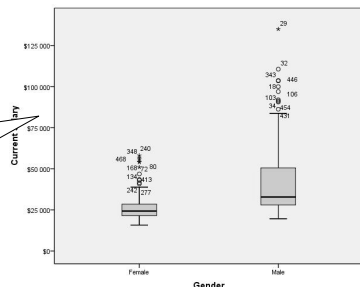




feladat

Ellenőrizzük független kétmintás t -próbával, hogy a nők és férfiak fizetése egyenlő-e!

Dobozábrák a fizetésekkel



Employee data.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analysis Graphs Utilities Window Help

1: id

id	id	idate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var	var	var	var
1	1	02/03/1962	15	3	\$57,000	\$27,000	98	144	0				
2	2	05/23/1966	16	3	\$103,750	\$27,510	97	70	0				
3	3	07/05/1929	15	3	\$12,000	\$9,381	0						
4	4	04/15/1947	15	3	\$13,200	\$9,190	0						
5	5	02/09/1965	16	3	\$13,500	\$9,136	0						
6	6	06/22/1958	15	3	\$18,750	\$9,114	0						
7	7	04/06/1956	15	3	\$9,750	\$12,750	97	315	1				
8	8	05/06/1966	15	3	\$31,050	\$12,800	96	14	0				
9	9	01/03/1946	15	3	\$60,375	\$27,480	96	96	0				
10	10	02/13/1946	12	1	\$26,250	\$11,550	97	48	0				
11	11	02/07/1965	16	1	\$38,850	\$15,000	97	17	0				
12	12	01/11/1966	12	1	\$26,250	\$11,550	97	48	0				
13	13	07/17/1960	15	3	\$12,000	\$9,190	0						
14	14	02/05/1949	15	3	\$135,000	\$79,980	96	199	0				
15	15	09/29/1962	15	3	\$131,300	\$14,251	96	44	0				
16	16	11/17/1964	15	3	\$135,000	\$79,980	96	199	0				
17	17	07/19/1962	15	3	\$135,000	\$79,980	96	199	0				
18	18	03/20/1966	15	3	\$135,000	\$79,980	96	199	0				
19	19	06/19/1962	12	1	\$42,300	\$14,250	97	103	0				
20	20	01/23/1940	12	1	\$26,250	\$11,550	97	48	0				
21	21	02/19/1963	16	1	\$38,850	\$15,000	97	17	0				
22	22	06/04/1940	12	1	\$21,750	\$12,750	97	315	1				
23	23	03/15/1966	15	3	\$31,050	\$12,800	96	14	0				
24	24	03/27/1933	12	1	\$16,950	\$9,000	97	124	1				
25	25	07/01/1942	15	3	\$21,150	\$9,000	97	171	1				
26	26	11/09/1966	15	3	\$31,050	\$12,800	96	14	0				
27	27	03/19/1964	19	3	\$60,375	\$27,480	96	96	0				
28	28	04/11/1963	15	3	\$32,550	\$14,250	96	43	0				
29	29	01/29/1944	19	3	\$135,000	\$79,980	96	199	0				
30	30	06/17/1961	15	3	\$131,300	\$14,251	96	44	0				

Reports

Descriptive Statistics

Tables

Compare Means

General Linear Model

Mixed Models

Correlate

Regression

Loglinear

Classify

Data Reduction

Scale

Nonparametric Tests

Time Series

Survival

Multiple Response

Missing Value Analysis...

Complex Samples

SPSS Processor is ready

Independent-Samples T Test

start

E:\Jobbált... Employee d... Output1 - S... Microsoft Po... 19:56

Employee data.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analysis Graphs Utilities Window Help

1: id

id	id	idate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var	var	var	var
1	1	02/03/1962	15	3	\$57,000	\$27,000	98	144	0				
2	2	05/23/1966	16	3	\$103,750	\$27,510	97	70	0				
3	3	07/05/1929	15	3	\$12,000	\$9,381	0						
4	4	04/15/1947	15	3	\$13,200	\$9,190	0						
5	5	02/09/1965	16	3	\$13,500	\$9,136	0						
6	6	06/22/1958	15	3	\$18,750	\$9,114	0						
7	7	04/06/1956	15	3	\$9,750	\$12,750	97	315	1				
8	8	05/06/1966	15	3	\$31,050	\$12,800	96	14	0				
9	9	01/03/1946	15	3	\$60,375	\$27,480	96	96	0				
10	10	02/13/1946	12	1	\$26,250	\$11,550	97	48	0				
11	11	02/07/1965	16	1	\$38,850	\$15,000	97	17	0				
12	12	01/11/1966	12	1	\$26,250	\$11,550	97	48	0				
13	13	07/17/1960	15	3	\$12,000	\$9,190	0						
14	14	02/05/1949	15	3	\$135,000	\$79,980	96	199	0				
15	15	09/29/1962	15	3	\$131,300	\$14,251	96	44	0				
16	16	11/17/1964	15	3	\$135,000	\$79,980	96	199	0				
17	17	07/19/1962	15	3	\$135,000	\$79,980	96	199	0				
18	18	03/20/1966	15	3	\$135,000	\$79,980	96	199	0				
19	19	06/19/1962	12	1	\$42,300	\$14,250	97	103	0				
20	20	01/23/1940	12	1	\$26,250	\$11,550	97	48	0				
21	21	02/19/1963	16	1	\$38,850	\$15,000	97	17	0				
22	22	06/04/1940	12	1	\$21,750	\$12,750	97	315	1				
23	23	03/15/1966	15	3	\$31,050	\$12,800	96	14	0				
24	24	03/27/1933	12	1	\$16,950	\$9,000	97	124	1				
25	25	07/01/1942	15	3	\$21,150	\$9,000	97	171	1				
26	26	11/09/1966	15	3	\$31,050	\$12,800	96	14	0				
27	27	03/19/1964	19	3	\$60,375	\$27,480	96	96	0				
28	28	04/11/1963	15	3	\$32,550	\$14,250	96	43	0				
29	29	01/29/1944	19	3	\$135,000	\$79,980	96	199	0				
30	30	06/17/1961	15	3	\$131,300	\$14,251	96	44	0				

Reports

Descriptive Statistics

Tables

Compare Means

General Linear Model

Mixed Models

Correlate

Regression

Loglinear

Classify

Data Reduction

Scale

Nonparametric Tests

Time Series

Survival

Multiple Response

Missing Value Analysis...

Complex Samples

SPSS Processor is ready

Independent-Samples T Test

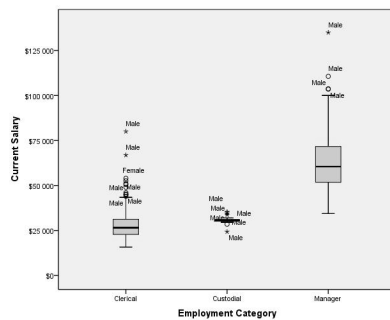
start

E:\Jobbált... Employee d... Output1 - S... Microsoft Po... 19:57

Group Statistics					
	Gender	N	Mean	Std. Deviation	Std. Error Mean
Current Salary	Male	258	\$41,441.78	\$19,499.214	\$1,213.968
	Female	216	\$26,031.92	\$7,558.021	\$514.258

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower Upper	
Current Salary	Equal variances assumed	119,669	,000	10,945	47	,000	15459,98	1,467,908	12,643,322	18,176,401
	Equal variances not assumed			11,688	344,362	,000	15459,98	1,318,400	12,816,718	18,002,996

Ellenőrizzük egyszerű csoportosítással, hogy a munkakörökben azonos-e a fizetés!



The screenshot shows the SPSS Data Editor window for the file 'Employee Data.sav'. The 'Data' tab is selected, displaying a list of variables on the left and a data grid on the right. The variables list includes 'id', 'age', 'sex', 'education', 'marital', 'income', 'job', 'tenure', 'promotion', 'minority', 'y1978', 'y1979', 'y1980', 'y1981', 'y1982', 'y1983', 'y1984', 'y1985', 'y1986', 'y1987', 'y1988', 'y1989', 'y1990', 'y1991', 'y1992', 'y1993', 'y1994', 'y1995', 'y1996', 'y1997', 'y1998', 'y1999', 'y2000', 'y2001', 'y2002', 'y2003', 'y2004', 'y2005', 'y2006', 'y2007', 'y2008', 'y2009', 'y2010', 'y2011', 'y2012', 'y2013', 'y2014', 'y2015', 'y2016', 'y2017', 'y2018', 'y2019', 'y2020'. The data grid shows the first 20 rows of data, with columns for 'id', 'age', 'sex', 'education', 'marital', 'income', 'job', 'tenure', 'promotion', 'minority', 'y1978', 'y1979', 'y1980', 'y1981', 'y1982', 'y1983', 'y1984', 'y1985', 'y1986', 'y1987', 'y1988', 'y1989', 'y1990', 'y1991', 'y1992', 'y1993', 'y1994', 'y1995', 'y1996', 'y1997', 'y1998', 'y1999', 'y2000', 'y2001', 'y2002', 'y2003', 'y2004', 'y2005', 'y2006', 'y2007', 'y2008', 'y2009', 'y2010', 'y2011', 'y2012', 'y2013', 'y2014', 'y2015', 'y2016', 'y2017', 'y2018', 'y2019', 'y2020'.

SPSS Data Editor window showing 'Employee data.sav' with a One-Way ANOVA dialog box open. The dialog box shows 'Current Salary (salary)' as the dependent variable and 'Employment Category (jobcat)' as the factor. The data table shows columns for id, jobcat, salary, and other variables.

Megoldás SPSS-sel

Descriptives

Current Salary

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Clerical	363	\$27,838.54	\$7,567.895	\$397.217	\$27,057.40	\$28,619.68	\$15,750	\$80,000
Custodial	27	\$30,936.89	\$2,114.616	\$406.958	\$30,102.37	\$31,775.40	\$24,300	\$35,250
Manager	84	\$63,977.80	\$18,244.776	\$1,990.668	\$60,018.44	\$67,937.16	\$34,410	\$135,000
Total	474	\$34,419.57	\$17,075.661	\$784.311	\$32,876.40	\$35,960.73	\$15,750	\$135,000

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8.9E+010	2	4.472E+010	434.481	.000
Within Groups	4.8E+010	471	102929714.5		
Total	1.4E+011	473			

Robust Tests of Equality of Means

Current Salary

	Statistic ^a	df1	df2	Sig.
Welch	162.200	2	117.312	.000

a. Asymptotically F distributed.

Megoldás SPSS-sel

Multiple Comparisons

Dependent Variable: Current Salary

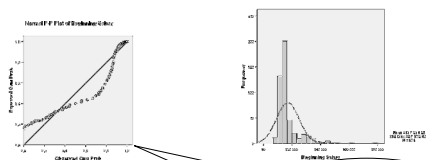
	(I) Employment Category	(J) Employment Category	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	Clerical	Custodial	-\$3,100.349*	\$2,023.760	.126	-\$7,077.06	\$876.37
	Manager	Clerical	-\$36,139.258*	\$1,228.352	.000	-\$38,552.99	-\$33,725.53
	Manager	Manager	\$3,100.349	\$2,023.760	.126	-\$876.37	\$7,077.06
	Custodial	Manager	-\$33,038.909*	\$2,244.409	.000	-\$37,449.20	-\$28,628.62
Tamhane	Clerical	Custodial	-\$36,139.258*	\$1,228.352	.000	-\$38,552.99	-\$33,725.53
	Manager	Clerical	-\$3,100.349*	\$2,023.760	.126	-\$7,077.06	\$876.37
	Manager	Manager	\$3,100.349	\$2,023.760	.126	-\$876.37	\$7,077.06
	Custodial	Manager	-\$33,038.909*	\$2,244.409	.000	-\$37,449.20	-\$28,628.62

*. The mean difference is significant at the .05 level.



feladat

Ellenőrizzük a fizetés illeszkedését a normálishoz!
Grafikusan, majd egy mintás *Kolmogorov-Szmirnov* próbával!



Grafikus vizsgálat
alapján nem tűnik
józnak az illeszkedés!

Employee data.sav [DataSet1] - SPSS Data Editor

	id	idate	educ	jobcat	salary	subbegin	jobtime	prevexp	minority	sex	mar	rel	rel	rel	rel
1	1	02/03/1962	12	1	\$42,300	\$11,000	97	144	0						
2	2	06/23/1968	16	1	\$36,250	\$11,000	97	36	0						
3	3	07/26/1929	12	1	\$38,850	\$15,000	97	381	0						
4	4	04/15/1947	12	1	\$21,750	\$12,750	97	190	0						
5	5	02/09/1965	16	1	\$38,850	\$15,000	97	138	0						
6	6	06/23/1968	16	1	\$21,150	\$9,000	97	67	0						
7	7	04/26/1966	15	1	\$31,050	\$12,000	96	114	0						
8	8	05/06/1966	19	3	\$63,375	\$27,493	96	0	0						
9	9	01/23/1946	16	1	\$32,550	\$14,250	96	116	0						
10	10	02/13/1946	19	3	\$135,000	\$79,980	96	244	0						
11	11	02/07/1960	16	1	\$31,200	\$14,250	96	26	0						
12	12	01/17/1966	12	1	\$42,300	\$15,000	97	24	0						
13	13	07/17/1960	15	1	\$46,000	\$14,250	97	48	0						
14	14	02/06/1940	12	1	\$21,750	\$12,750	97	315	1						
15	15	06/29/1962	16	1	\$36,250	\$11,000	97	75	1						
16	16	11/17/1964	16	1	\$38,850	\$15,000	97	124	1						
17	17	07/19/1962	15	1	\$21,150	\$9,000	97	171	1						
18	18	03/20/1956	16	1	\$38,850	\$15,000	97	17	0						
19	19	08/19/1962	12	1	\$42,300	\$15,000	97	103	0						
20	20	01/23/1940	12	1	\$36,250	\$11,000	97	48	0						
21	21	02/19/1963	16	1	\$38,850	\$15,000	97	17	0						
22	22	08/24/1940	12	1	\$21,750	\$12,750	97	315	1						
23	23	03/15/1966	15	1	\$31,050	\$12,000	96	14	0						
24	24	03/27/1933	12	1	\$16,950	\$9,000	97	124	1						
25	25	07/01/1942	15	1	\$21,150	\$9,000	97	171	1						
26	26	11/09/1966	15	1	\$31,050	\$12,000	96	14	0						
27	27	03/19/1954	19	3	\$63,375	\$27,493	96	96	0						
28	28	04/11/1963	15	1	\$32,550	\$14,250	96	43	0						
29	29	01/29/1944	19	3	\$135,000	\$79,980	96	199	0						
30	30	06/17/1963	16	1	\$31,200	\$14,250	96	44	0						

Employee data.sav [DataSet1] - SPSS Data Editor

One-Sample Kolmogorov-Smirnov

Test Variable List: Employee Code [id]

Test Distribution: Normal

Exact: [X]

Options: [X]

	id	idate	educ	jobcat	salary	subbegin	jobtime	prevexp	minority	sex	mar	rel	rel	rel	rel
1	1	02/03/1962	12	1	\$42,300	\$15,000	97	144	0						
2	2	06/23/1968	16	1	\$36,250	\$11,000	97	36	0						
3	3	07/26/1929	12	1	\$38,850	\$15,000	97	381	0						
4	4	04/15/1947	12	1	\$21,750	\$12,750	97	190	0						
5	5	02/09/1965	16	1	\$38,850	\$15,000	97	138	0						
6	6	06/23/1968	16	1	\$21,150	\$9,000	97	67	0						
7	7	04/26/1966	15	1	\$31,050	\$12,000	96	114	0						
8	8	05/06/1966	19	3	\$63,375	\$27,493	96	0	0						
9	9	01/23/1946	16	1	\$32,550	\$14,250	96	116	0						
10	10	02/13/1946	19	3	\$135,000	\$79,980	96	244	0						
11	11	02/07/1960	16	1	\$31,200	\$14,250	96	26	0						
12	12	01/17/1966	12	1	\$42,300	\$15,000	97	24	0						
13	13	07/17/1960	15	1	\$46,000	\$14,250	97	48	0						
14	14	02/06/1940	12	1	\$21,750	\$12,750	97	315	1						
15	15	06/29/1962	16	1	\$36,250	\$11,000	97	75	1						
16	16	11/17/1964	16	1	\$38,850	\$15,000	97	124	1						
17	17	07/19/1962	15	1	\$21,150	\$9,000	97	171	1						
18	18	03/20/1956	16	1	\$38,850	\$15,000	97	17	0						
19	19	08/19/1962	12	1	\$42,300	\$15,000	97	103	0						
20	20	01/23/1940	12	1	\$36,250	\$11,000	97	48	0						
21	21	02/19/1963	16	1	\$38,850	\$15,000	97	17	0						
22	22	08/24/1940	12	1	\$21,750	\$12,750	97	315	1						
23	23	03/15/1966	15	1	\$31,050	\$12,000	96	14	0						
24	24	03/27/1933	12	1	\$16,950	\$9,000	97	124	1						
25	25	07/01/1942	15	1	\$21,150	\$9,000	97	171	1						
26	26	11/09/1966	15	1	\$31,050	\$12,000	96	14	0						
27	27	03/19/1954	19	3	\$63,375	\$27,493	96	96	0						
28	28	04/11/1963	15	1	\$32,550	\$14,250	96	43	0						
29	29	01/29/1944	19	3	\$135,000	\$79,980	96	199	0						
30	30	06/17/1963	16	1	\$31,200	\$14,250	96	44	0						



One-Sample Kolmogorov-Smirnov Test

		Beginning Salary
N		474
Normal Parameters ^{a,b}	Mean	\$17,016.09
	Std. Deviation	\$7,870.638
Most Extreme Differences	Absolute	,252
	Positive	,252
	Negative	-,170
Kolmogorov-Smirnov Z		5,484
Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.

b. Calculated from data.



Regressziós feladat

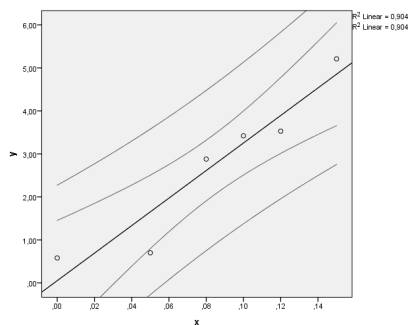
Vizsgáljuk meg az x független változó és az y függő változó közötti összefüggést! Az x független változó értéke pontosan beállítható, az y függő változó értéke azonban a Y valódi érték körül ingadozik. A mérési adatok az alábbi táblázatban láthatók, az y értéke szerint növekvő sorrendbe rendezve. A tényleges mérési sorrendet a táblázat második oszlopa tartalmazza. Feltételezve, hogy y normális eloszlású, valamint azt hogy az y és x közötti függvénykapcsolat lineáris, adjunk becslést az egyenes paramétereire!

No	mérési sorrend	x	y
1	3	0	0.58
2	5	0.05	0.7
3	4	0.08	2.88
4	2	0.1	3.42
5	1	0.12	3.53
6	6	0.15	5.21



Scatter ábra az adatokkal

a 95%-os konfidencia intervallummal és a 95%-os jóslási határral





SPSS táblázatok

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	x ^b		Enter

a. Dependent Variable: y
b. All requested variables entered.

Model Summary^a

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.951 ^a	.904	.880	.62136

a. Predictors: (Constant), x
b. Dependent Variable: y

R^2

s_r

R^2_{adj}



Egyváltozós lineáris regresszió ismétlés nélküli mérések esetén, σ^2 konstans

A) Ha $\sum x_i \neq 0$

$(y_i - b_0 - bx_i)^2 = \min.$

a b_0 és b becslések egymástól nem függetlenek

$b_0 = a - b\bar{x}$

A normálegyenletek:

$$\frac{\partial \phi}{\partial b_0} = -2 \sum [y_i - b_0 - bx_i] = 0$$

$$\frac{\partial \phi}{\partial b} = -2 \sum [y_i - b_0 - bx_i]x_i = 0$$

Átrendezve:

$$\sum y_i = nb_0 + b \sum x_i$$

$$\sum y_i x_i = b_0 \sum x_i + b \sum x_i^2$$



A normálegyenletek az $\hat{Y}_i = a + b(x_i - \bar{x})$ modell illesztésekor

$$\frac{\partial \phi}{\partial a} = -2 \sum [y_i - a - b(x_i - \bar{x})] = 0$$

$$\frac{\partial \phi}{\partial b} = -2 \sum [y_i - a - b(x_i - \bar{x})](x_i - \bar{x}) = 0$$

Átrendezve:

$$\sum y_i = na + b \sum (x_i - \bar{x})$$

$$\sum y_i(x_i - \bar{x}) = a \sum (x_i - \bar{x}) + b \sum (x_i - \bar{x})^2$$

Az a és b becslések egymástól függetlenek, mert

$$\sum (x_i - \bar{x}) = 0$$

$$\bar{x} = \frac{\sum x_i}{n}$$



$$\sum y_i = na \quad \text{és} \quad \sum y_i(x_i - \bar{x}) = b \sum (x_i - \bar{x})^2$$

tehát az a és b becsült paraméterek egymástól függetlenül kaphatók meg a két normálegyenletből:

$$a = \frac{\sum_i y_i}{n} \quad b = \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{Y} = a + b(x_i - \bar{x}) ; \quad E(\hat{Y}_i) = Y_i = \alpha + \beta(x_i - \bar{x})$$



A becslések tulajdonságai

$$E(a) \equiv E\left(\frac{\sum y_i}{n}\right) = \alpha \quad \text{Var}(a) = \frac{\sum \sigma^2}{(n)^2} = \frac{\sigma^2}{n}$$

$$E(b) = \beta \quad \text{Var}(b) = \frac{\sum (x_i - \bar{x})^2 \sigma^2}{\left(\sum (x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$E(\hat{Y}) = E[a + b(x - \bar{x})] = E(a) + E(b)(x - \bar{x})$$

$$E(\hat{Y}) = \alpha + \beta(x - \bar{x}) = Y$$

$$\text{Var}(\hat{Y}) = \text{Var}(a) + (x - \bar{x})^2 \text{Var}(b) = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right]$$



Konfidencia határok

$$s_a = \frac{s_r}{\sqrt{n}} \quad s_b = \frac{s_r}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$s_{\hat{Y}} = s_r \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} = \sqrt{s_a^2 + s_b^2 (x - \bar{x})^2}$$

$$s_{b_0} = s_{\hat{Y}(x=0)} = \sqrt{s_a^2 + s_b^2 \bar{x}^2}$$

A konfidenciatartományok a t -eloszlás alapján számíthatók.



Konfidencia intervallum az átlaghoz

$$P\left(\hat{y}_i - t_{\frac{\epsilon}{2}, n-2} \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{n \cdot s_x^2}} \leq M_i \leq \hat{y}_i + t_{\frac{\epsilon}{2}, n-2} \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{n \cdot s_x^2}}\right) = 1 - \epsilon$$

$$M_i = E(Y | X = x_i)$$

Az átlag konfidencia-intervalluma a mintapontok kb. $1-\epsilon$ %-át tartalmazza. A sáv az x átlagának a közelében a legvékonyabb.



Jóslási intervallum

$$s_{y-\hat{Y}} = s_r \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} = \sqrt{s_r^2 + s_a^2 + s_b^2 (x - \bar{x})^2}$$

$$\text{intervallum: } \hat{Y}(x) \pm t_{\alpha/2} s_{y-\hat{Y}}$$

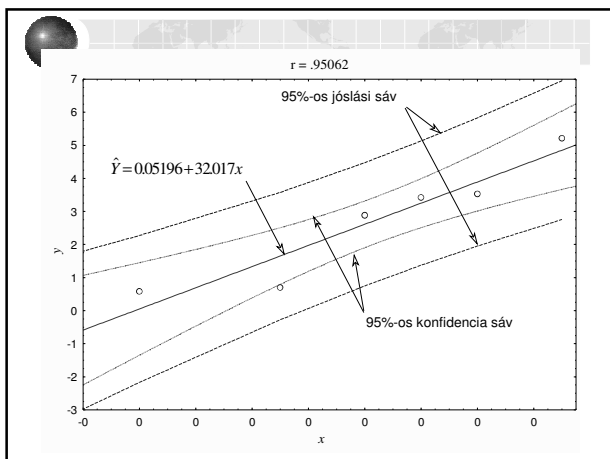
$(1 - \alpha)$ a valószínűsége annak, hogy x adott értékénél egy későbbi mérés eredménye a számított intervallumba esik.



Jóslási intervallum

Az adott x_i -hez tartozó Y_i egyedi értéket tartalmazza az alábbi $1-\epsilon$ szintű konfidencia intervallum:

$$P\left(\hat{y}_i - t_{\frac{\epsilon}{2}, n-2} \cdot s_r \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{n \cdot s_x^2}} \leq Y_i \leq \hat{y}_i + t_{\frac{\epsilon}{2}, n-2} \cdot s_r \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{n \cdot s_x^2}}\right) = 1 - \epsilon$$



Determinációs együttható

“Regression”

“Residual”

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

“Total”

$$R^2_{adj} = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$$

A képletek magyarázata

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = SSE + SSR$$

d.f.: n-1 = n-2 + 1

$$s_r = \sqrt{\frac{SSE}{n-2}}$$

s_r reziduális szórás

$$s_a = \frac{s_r}{\sqrt{n}} \quad s_b = \frac{s_r}{\sqrt{\sum (x_i - \bar{x})^2}}$$

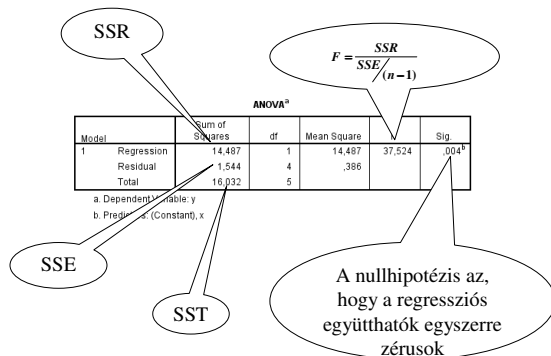
Az együtthatók szórásai

A konfidenciartományok a t-eloszlás alapján számíthatók

$$s_{\hat{y}} = s_r \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} = \sqrt{s_a^2 + s_b^2 (x - \bar{x})^2}$$

$$s_{b_0} = s_{\hat{y}(x=0)} = \sqrt{s_a^2 + s_b^2 \bar{x}^2}$$

ANOVA-táblázat



Regressziós együtthatók

Coefficients ^a										
Model	Unstandardized Coefficients			Standardized Coefficients			95.0% Confidence Interval for B			Collinearity Statistics
	B	Std. Error		Beta	t	Sig.	Lower Bound	Upper Bound		
1										
(Constant)	,952	,504			,193	,923	-1,347	1,451		
x	32,817	5,227		,951	6,126	,004	17,505	48,129	,951	,951

a. 0,051958

A tapasztalati regressziós egyenes képlete:

$$\hat{y} = 0,05196 + 32,017x$$
