

Popisná statistika

jsou to jednoduché metody, ke kterým není potřeba žádná teorie.

Patří sem např.: tabulka četností, histogram, výběrový medián, výběrový průměr, rozptyl, výběrová odchylka, výběrová směrodatná odchylka, katicový diagram, rozptylový diagram atd.

Tabulka četností: máme naměřené hodnoty $x_1, x_2, \dots, x_n$

upřesňujeme je: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

někdy $a_1 < a_2 < \dots < a_m$ jsou všechny možným různě hodnoty souboru

$m_1, m_2, \dots, m_m$ jsou jejich četnosti; $n = \sum_{j=1}^m m_j$

vytvříme tabulku četností

a_1	a_2	...	a_m
m_1	m_2	...	m_m

nebo relativních četností

a_1	a_2	...	a_m
m_1/n	m_2/n	...	m_m/n

(relativní četnost $\frac{m_i}{n}$ je axiomatickým předpokladem, že náhodná veličina X má jako hodnotu a_i)

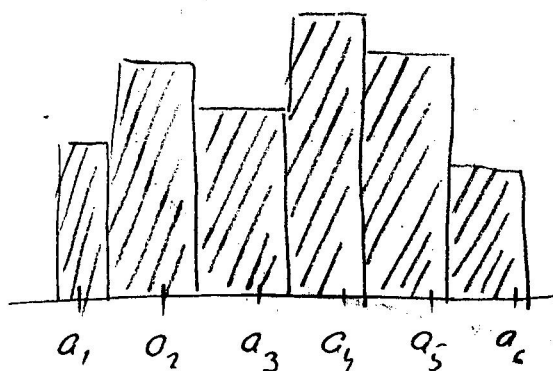
nebo kumulativních četností

a_1	a_2	...	a_m
N_1	N_2	...	N_m

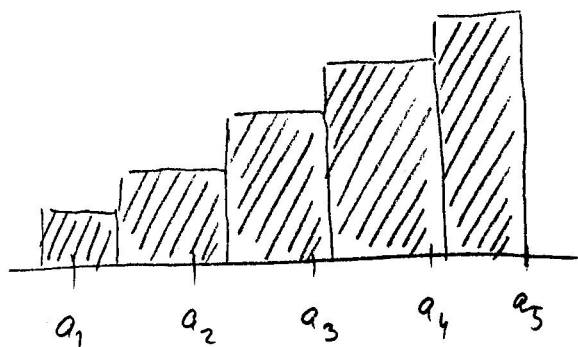
kde $N_j = \sum_{i=1}^j m_i$

(relativní kumulativní četnost $\frac{N_j}{n}$ je axiomatickým předpokladem)
 $P(X \leq j) = P(X < j+1) = F(j+1)$... distribuční funkce

Histogram je grafická podoba tabulek četosti



histogram relativních četostí
je odhadem hustoty



histogram kumulativních
relativních četostí je
odhadem distribuční
funkce

Výběrový průměr $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ je odhadem střední hodnoty
 $EX = \sum_i X_i P(X=X_i)$

Výběrový medián je "prostřední" hodnota v uspořádaném
souboru, tj.
$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{po } n \text{ liché} \\ \frac{1}{2} (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{po } n \text{ sudé} \end{cases}$$

je odhadem medianu $\tilde{\mu}$, který je definován vztahem

$$F(\tilde{\mu}) \leq \frac{1}{2}, \quad F(\tilde{\mu} + 0) \geq \frac{1}{2}$$

limita zprava

Medan je nultýňí hodnota

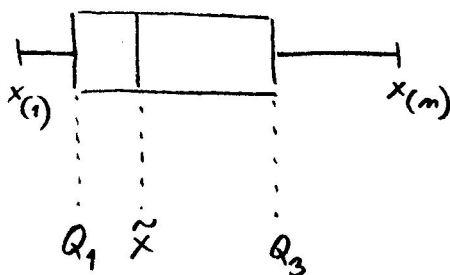
Výběrová rozpětí $R = X_{(n)} - X_{(1)}$

Průměrná odchylka $d = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$

Výběrový rozptyl $M^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ je odhadem rozptylu
 $E(X - EX)^2$

odmocnina z něj je výběrová směrodatná odchylka

Krácový diagram vypadá nějak takhle (nemá to jedinou možnost):



hde $\tilde{x}$ je median

Q_1 je dolní kvartil $x_{0,25}$

Q_3 je horní kvartil $x_{0,75}$

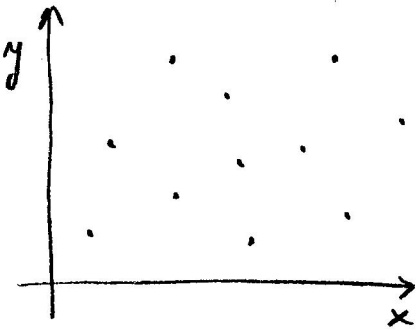
(obecně p -tý kvartil $x_p = \begin{cases} x_{(\lceil np \rceil + 1)} & \text{pro } np \neq \lceil np \rceil \\ \frac{1}{2} (x_{(np)} + x_{(np+1)}) & \text{pro } np = \lceil np \rceil \end{cases}$
 ale

$R_Q = Q_3 - Q_1$ je kvartilová rozpětí

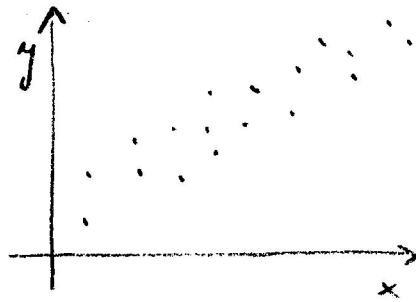
Rozptyľný diagram sa používa ke mäsavíme závislosti medzi
veličinami X a Y

máme poradení $x_1, x_2, \dots, x_n$ a $y_1, y_2, \dots, y_n$

vytvoríme dvojice $(x_1, y_1), \dots, (x_n, y_n)$ a naneseme je do grafu



ľahšie to vypadá,
keďže jsou X a Y nezávislé



ľahšie je závislosť
(přímá závislosť)

Jste k tabulkám čísel: Pokud je n souboru přilis mnoho různých
hodnot, rozdělíme jejich obor na intervaly (třidy):

$$(-\infty, t_0), (t_0, t_1), (t_1, t_2), \dots, (t_m, \infty)$$

a každé třídě zvolíme střed $a_j = \frac{t_{j-1} + t_j}{2}$

a_0 se volí tak, aby to byl střed intervalu (a_0, a_1)

$$a_{m+1} \quad \text{---} \quad \text{---} \quad t_m \quad \text{---} \quad \text{---} \quad a_{m+1}$$

vezme třídni číseli n_j jako počet hodnot x_i , které patří do
jednotlivých intervalů

vedeme tabulku třídni číseli

data file: csu2002.sta [14 cases with 3 variables]

0 CASE NAME	1 VEK	2 PLAT	3 DETI
Praha	41.500	19879.00	8.400
Stredoces	39.500	15534.00	9.300
Jihocesky	39.000	14029.00	9.200
Plzensky	39.700	14840.00	8.900
Karlovarsk	38.300	13535.00	9.700
Ustecky	38.300	14269.00	10.000
Liberecky	38.600	14081.00	9.700
Kralovehra	39.600	13753.00	9.000
Pardubicky	39.000	13539.00	9.200
Vysocina	38.700	13441.00	9.200
Jihomoravs	39.500	14059.00	9.000
Olomoucky	39.000	13373.00	8.900
Zlinsky	39.000	13843.00	8.800
Moravskosl	38.500	14923.00	9.100

Údaje Českého statistického úřadu na rok 2002:

příměří věk

příměří mada v Kč

počet živě narozených dětí na 1000 obyvatel

STAT. BASIC STATS	Descriptive Statistics (csu2002.sta)						
Variable	Valid N	Mean	Confid. -95.000%	Confid. +95.000%	Median	Minimum	Maximum
VEK	14	39.16	38.68	39.63	39.00	38.30	41.50
PLAT	14	14507.00	13543.81	15470.19	14044.00	13373.00	19879.00
DETI	14	9.17	8.93	9.41	9.15	8.40	10.00

STAT. BASIC STATS	Descriptive Statistics (csu2002.sta)	
Variable	Std.Dev.	Standard Error
VEK	.820	.2191
PLAT	1668.207	445.8471
DETI	.412	.1102

Kalendāri charakteristikas "apmērojamie" pēc vecuma laika.

Pozor - neapmierinātība ar šiem datiem, jo tie ir vidējais
 piemērs, nevis reālā situācija!

vidējais vecums ir 39,3

— " — plat 15 854

— " — proc. bērni 9,1

Ja redz, ka vidējais un medians ir liels

Krabicový diagram pro věk:

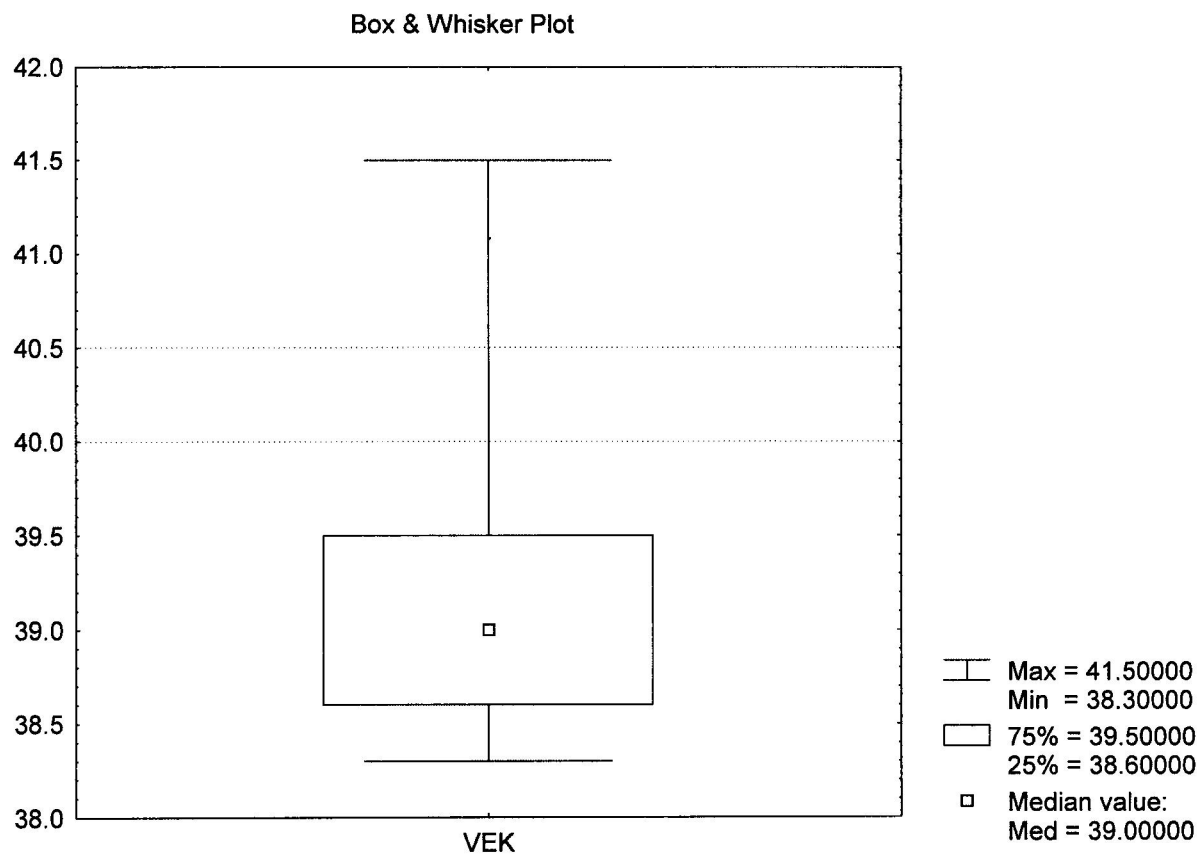


Diagram založený na mediánu a kvartilovém rozpětí
není symetrický

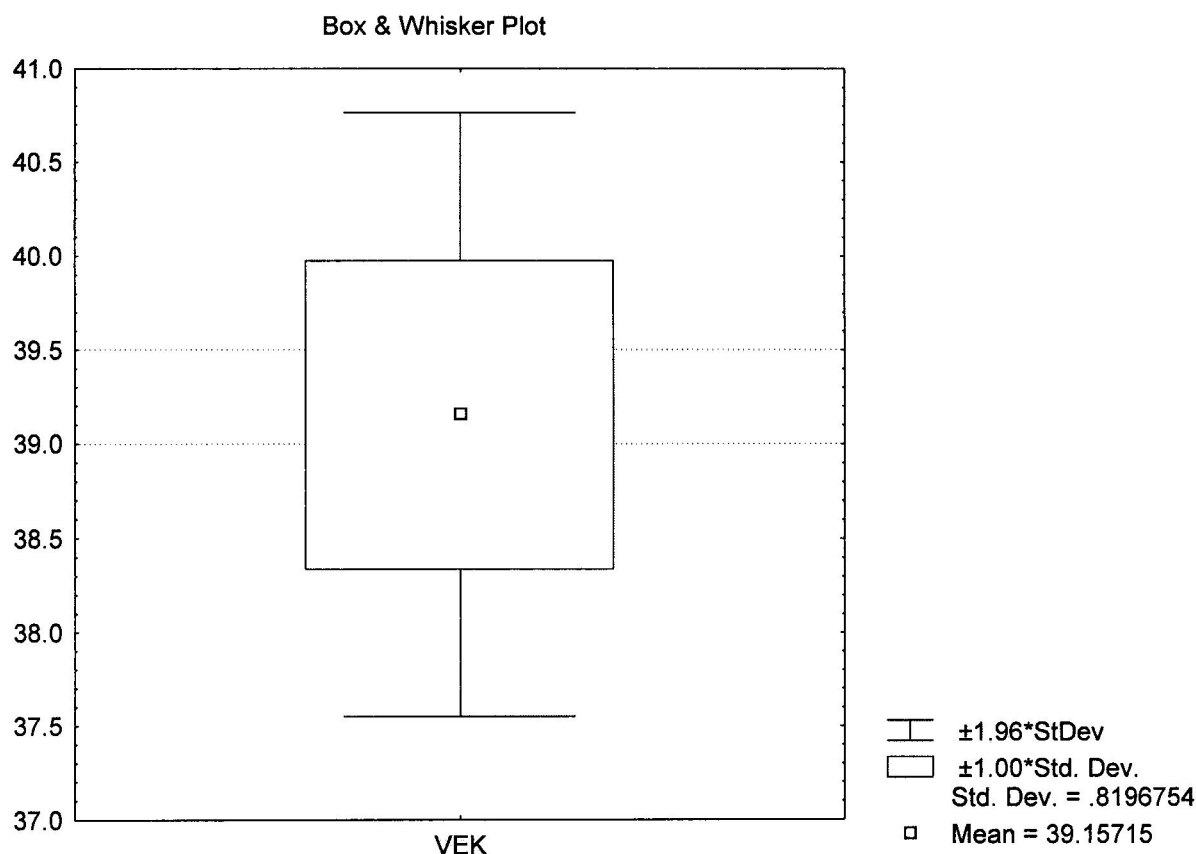


Diagram založený na průměru a směrodatné odchylce
je symetrický

Předechozí diagram lépe vystihuje rozložení počtu obyvatel
podle věku : polovina obyvatel je ve věku do 39 let
pouze čtvrtina je starší než 39,5 roku

Něco takového z diagramu založeném na průměru není
patrné

krabicový diagram pro plat:

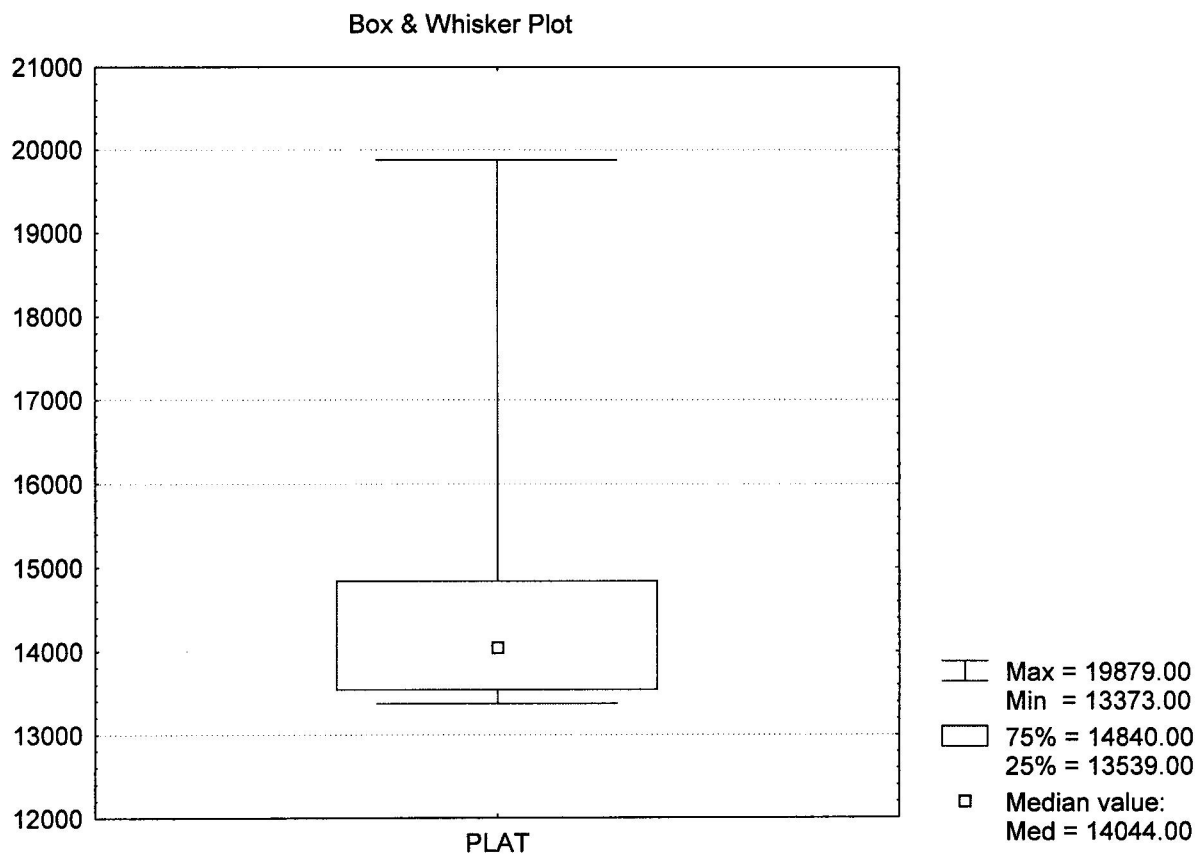


Diagram založený na mediánu a kvartilovém rozpětí

je ještě více nesymetrický než v případě věku

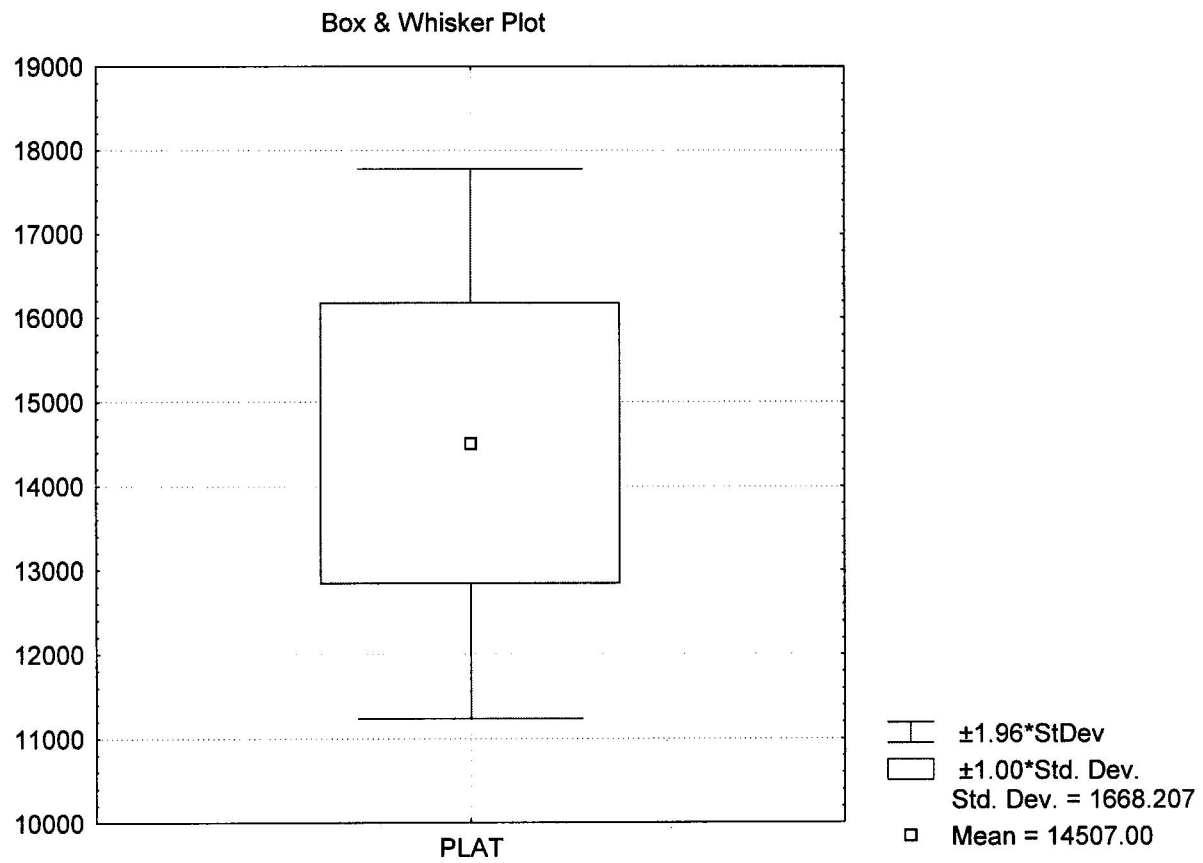


Diagram založený na průměru a směrodatné odchylce

krabicový diagram pro počet narozených dětí:

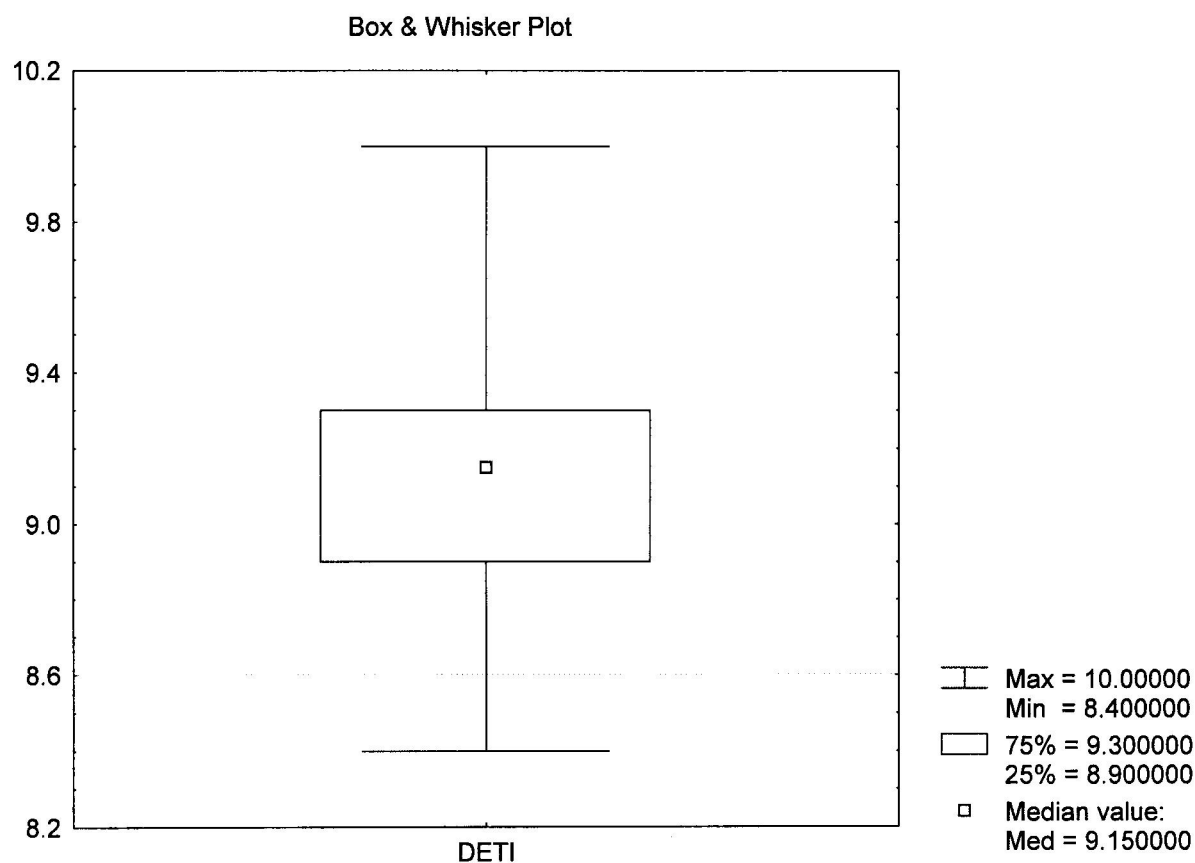


Diagram založený na mediánu a mezikvartilovém rozpětí

tady rozdíl v symetrii není velký

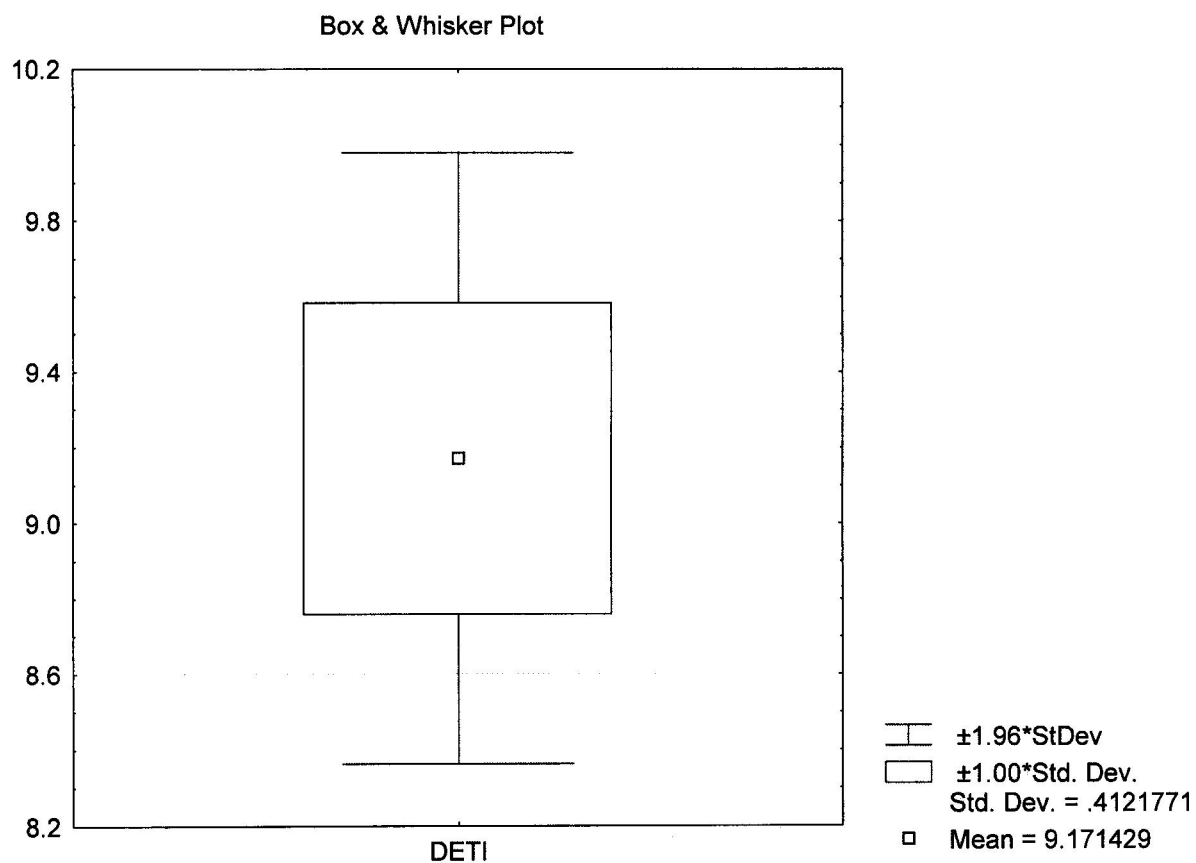


Diagram založený na průměru a směrodatné odchylce