Deskriptivní statistika

POLB1139

19.10. 2020

Navazujeme na minule…

Máme data

Máme datovou matici

Kombinace proměnných a případů

Co s tím?

Chci data sumarizovat, vizualizovat, podívat se na strukturu dat

Deskriptivní analýza

Explorace dat v rámci jedné proměnné

Jednorozměrná analýza

Nehledáme rozdíly ani souvislosti mezi proměnnými (vícerozměrná analýza)

Dobrý první krok vždy

Vizualizace

Deskriptivní analýza

Záleží na úrovni proměnných podle měření

Různé typy proměnných = různé možnosti

Vždy ale začínáme popisem

Prostor pro odhalování chyb (měření)

Kategorická data

Nominální a ordinální proměnná

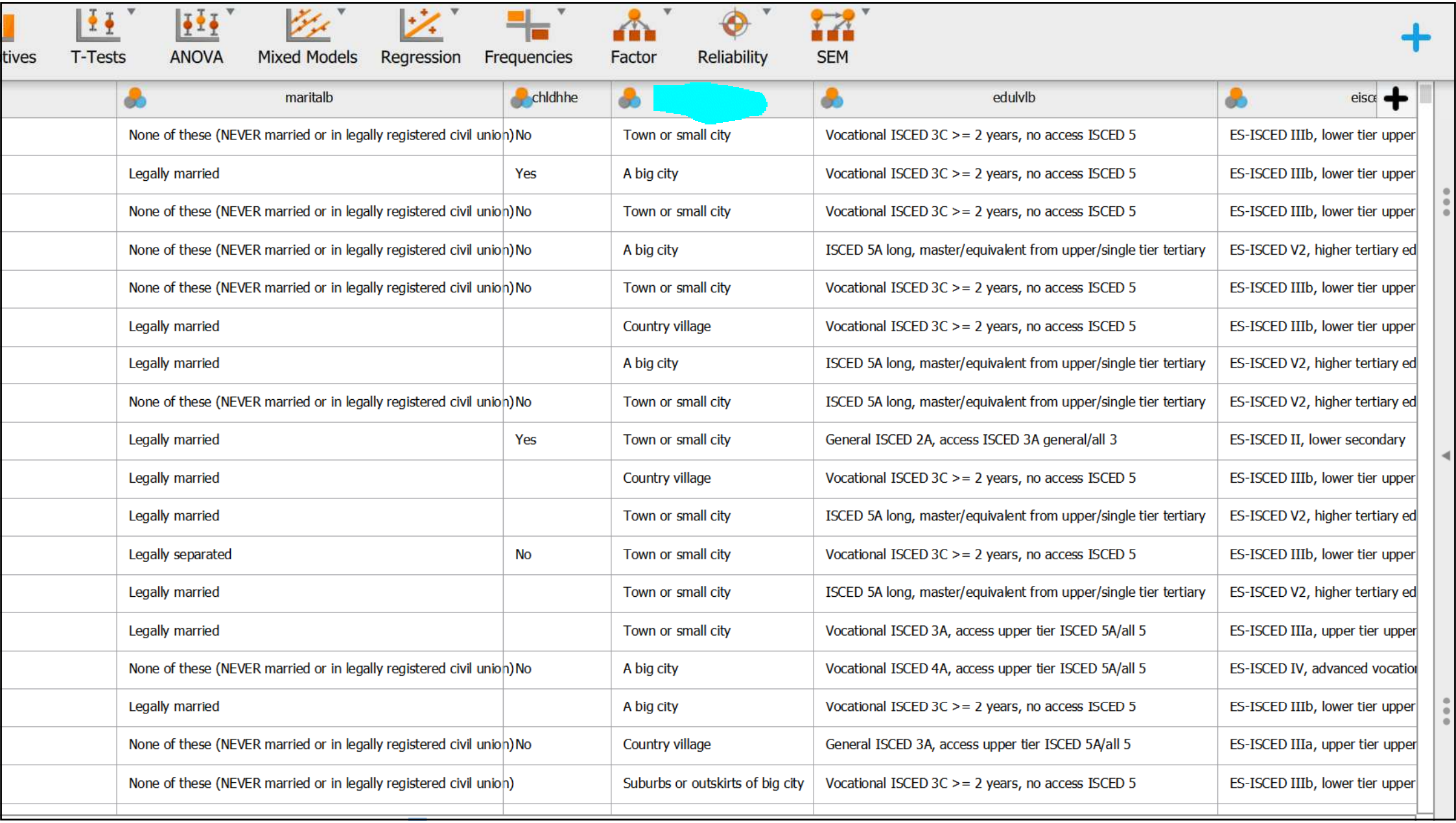
Čím se vyznačují?

Co s nimi tedy můžeme dělat?

Můžeme se podívat, kolik máme případů pro jednotlivé kategorie

Jak je distribuovaná proměnná napříč kategoriemi

Neprovádíme žádné výpočty



	maritalb	chldhhe	[redacted]	edulvlb	eisc
	None of these (NEVER married or in legally registered civil union)	No	Town or small city	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper
	Legally married	Yes	A big city	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper
	None of these (NEVER married or in legally registered civil union)	No	Town or small city	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper
	None of these (NEVER married or in legally registered civil union)	No	A big city	ISCED 5A long, master/equivalent from upper/single tier tertiary	ES-ISCED V2, higher tertiary ed
	None of these (NEVER married or in legally registered civil union)	No	Town or small city	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper
	Legally married		Country village	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper
	Legally married		A big city	ISCED 5A long, master/equivalent from upper/single tier tertiary	ES-ISCED V2, higher tertiary ed
	None of these (NEVER married or in legally registered civil union)	No	Town or small city	ISCED 5A long, master/equivalent from upper/single tier tertiary	ES-ISCED V2, higher tertiary ed
	Legally married	Yes	Town or small city	General ISCED 2A, access ISCED 3A general/all 3	ES-ISCED II, lower secondary
	Legally married		Country village	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper
	Legally married		Town or small city	ISCED 5A long, master/equivalent from upper/single tier tertiary	ES-ISCED V2, higher tertiary ed
	Legally separated	No	Town or small city	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper
	Legally married		Town or small city	ISCED 5A long, master/equivalent from upper/single tier tertiary	ES-ISCED V2, higher tertiary ed
	Legally married		Town or small city	Vocational ISCED 3A, access upper tier ISCED 5A/all 5	ES-ISCED IIIa, upper tier upper
	None of these (NEVER married or in legally registered civil union)	No	A big city	Vocational ISCED 4A, access upper tier ISCED 5A/all 5	ES-ISCED IV, advanced vocatio
	Legally married		A big city	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper
	None of these (NEVER married or in legally registered civil union)	No	Country village	General ISCED 3A, access upper tier ISCED 5A/all 5	ES-ISCED IIIa, upper tier upper
	None of these (NEVER married or in legally registered civil union)		Suburbs or outskirts of big city	Vocational ISCED 3C >= 2 years, no access ISCED 5	ES-ISCED IIIb, lower tier upper

Tabulka četností (frequency table)

Frequency Tables

Frequencies for domicil

domicil	Frequency	Percent	Valid Percent	Cumulative Percent
A big city	873	36.405	36.405	36.405
Suburbs or outskirts of big city	88	3.670	3.670	40.075
Town or small city	734	30.609	30.609	70.684
Country village	697	29.066	29.066	99.750
Farm or home in countryside	6	0.250	0.250	100.000
Missing	0	0.000		
Total	2398	100.000		

Absolutní četnost

Relativní četnost

Kumulativní procenta

Frequency Tables

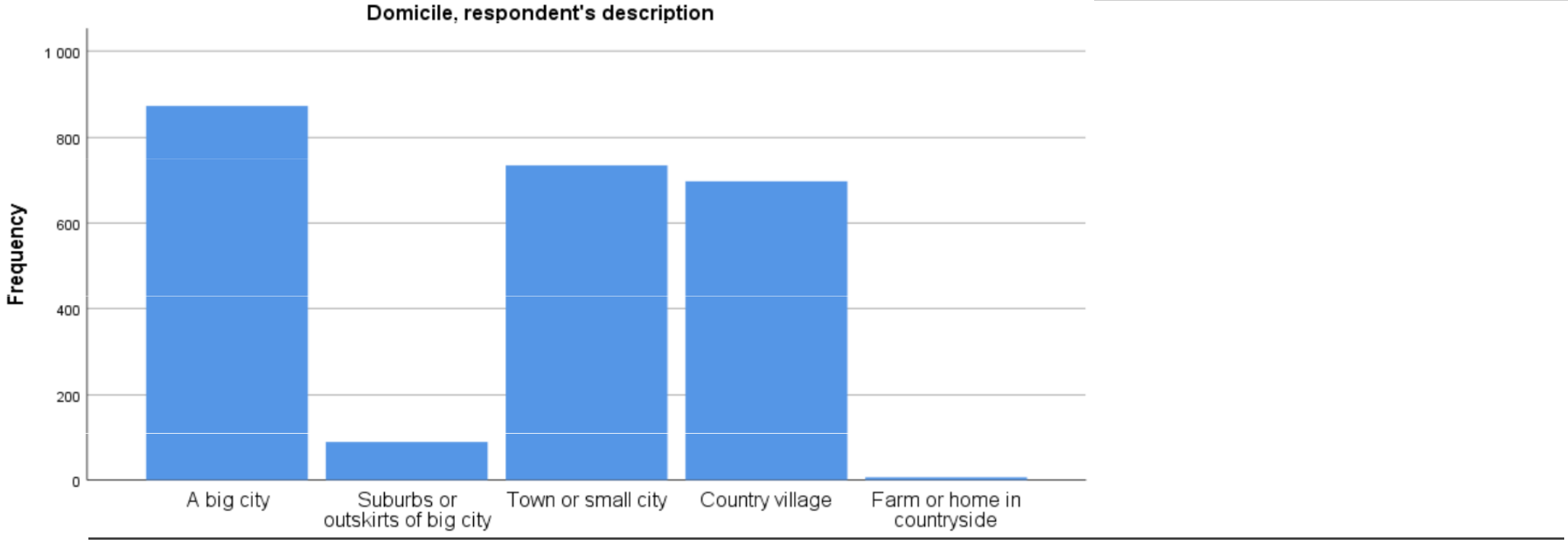
Frequencies for domicil

domicil	Frequency	Percent	Valid Percent	Cumulative Percent
A big city	873	36.405	36.405	36.405
Suburbs or outskirts of big city	88	3.670	3.670	40.075
Town or small city	734	30.609	30.609	70.684
Country village	697	29.066	29.066	99.750
Farm or home in countryside	6	0.250	0.250	100.000
Missing	0	0.000		
Total	2398	100.000		

Vizualizace

Sloupcový graf (bar chart)

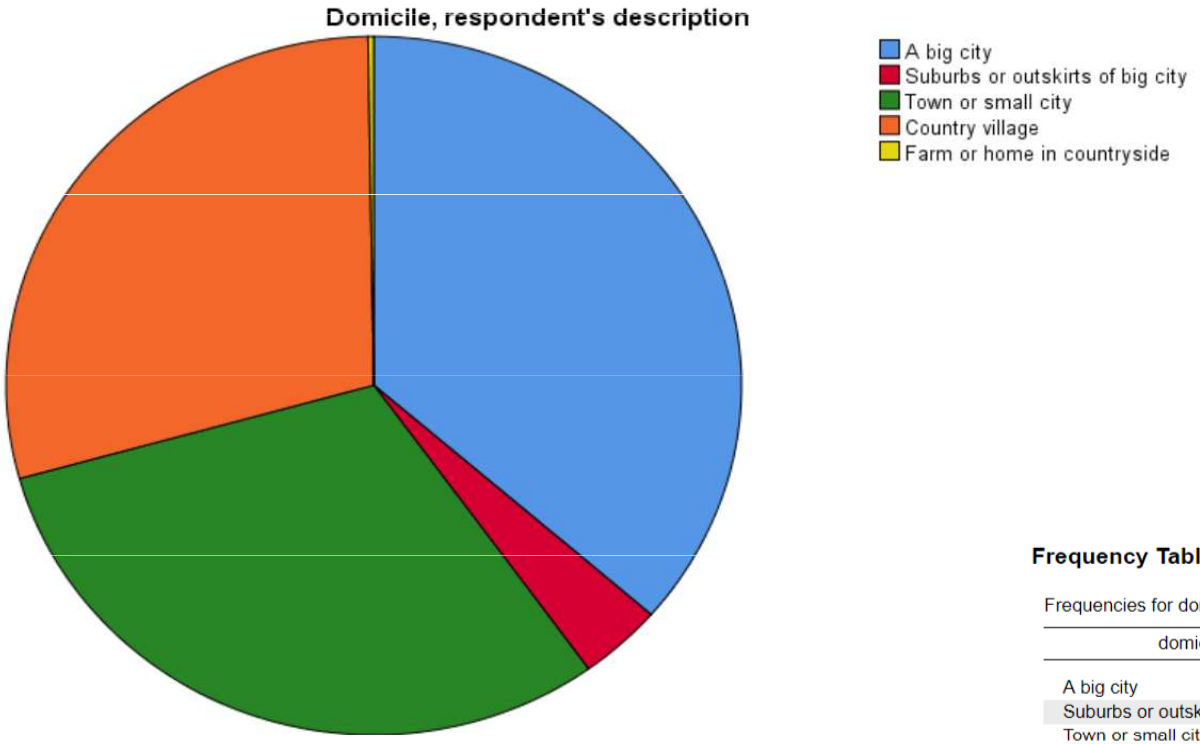
Koláčový graf (pie chart) – nedoporučuje se



Frequency Tables

Frequencies for domicil

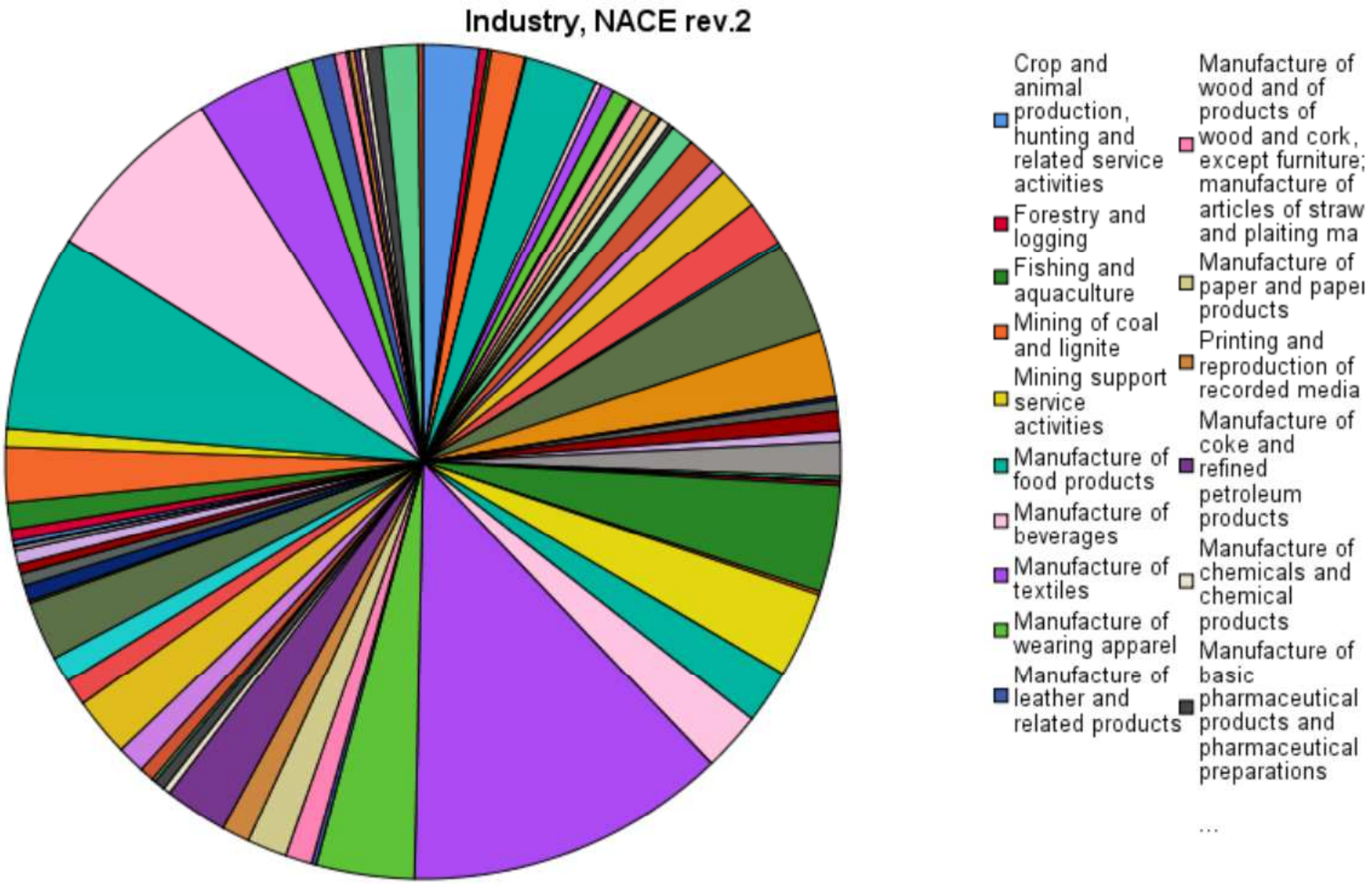
domicil	Frequency	Percent	Valid Percent	Cumulative Percent
A big city	873	36.405	36.405	36.405
Suburbs or outskirts of big city	88	3.670	3.670	40.075
Town or small city	734	30.609	30.609	70.684
Country village	697	29.066	29.066	99.750
Farm or home in countryside	6	0.250	0.250	100.000
Missing	0	0.000		
Total	2398	100.000		



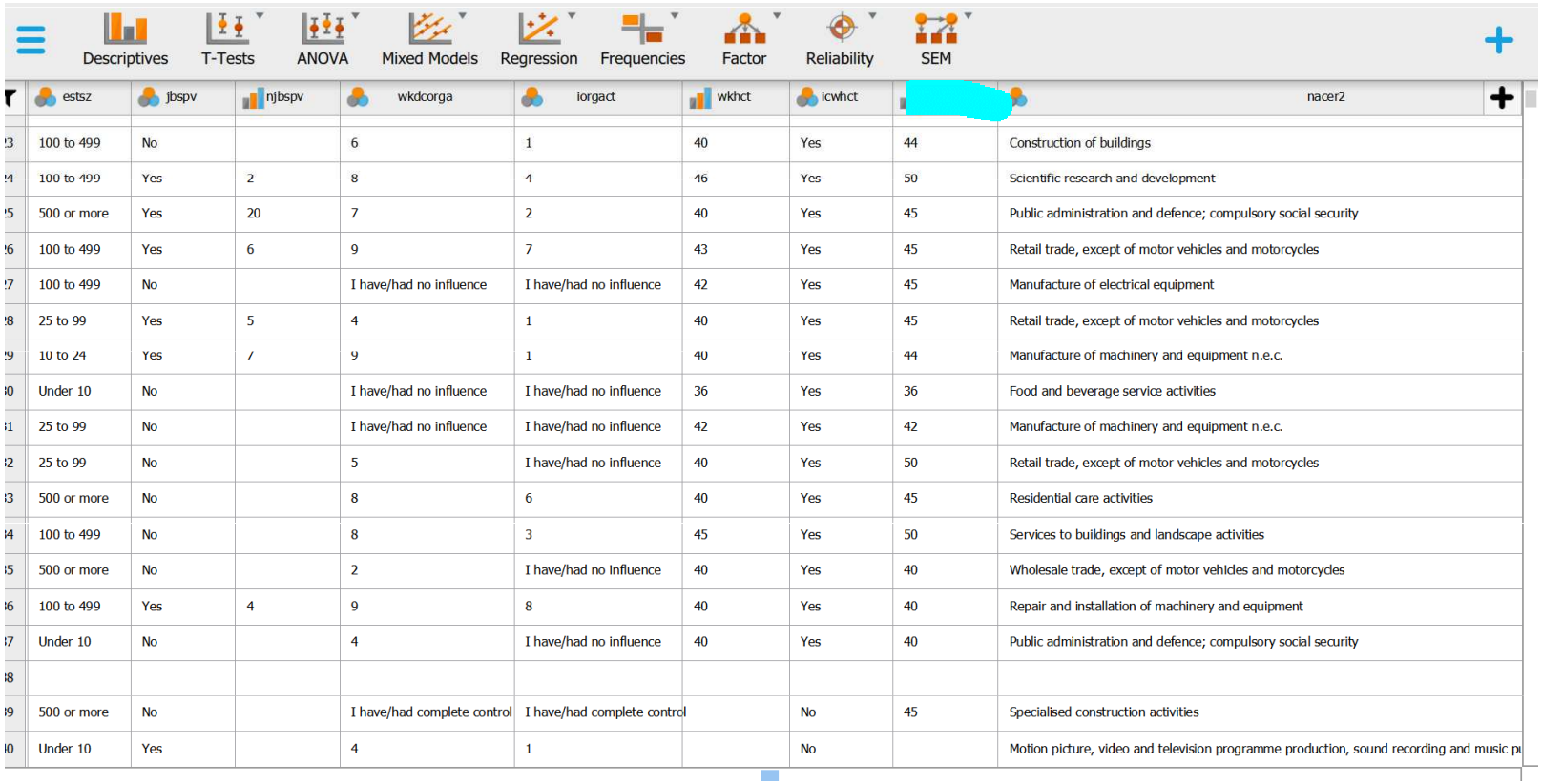
Frequency Tables

Frequencies for domicil

domicil	Frequency	Percent	Valid Percent	Cumulative Percent
A big city	873	36.405	36.405	36.405
Suburbs or outskirts of big city	88	3.670	3.670	40.075
Town or small city	734	30.609	30.609	70.684
Country village	697	29.066	29.066	99.750
Farm or home in countryside	6	0.250	0.250	100.000
Missing	0	0.000		
Total	2398	100.000		

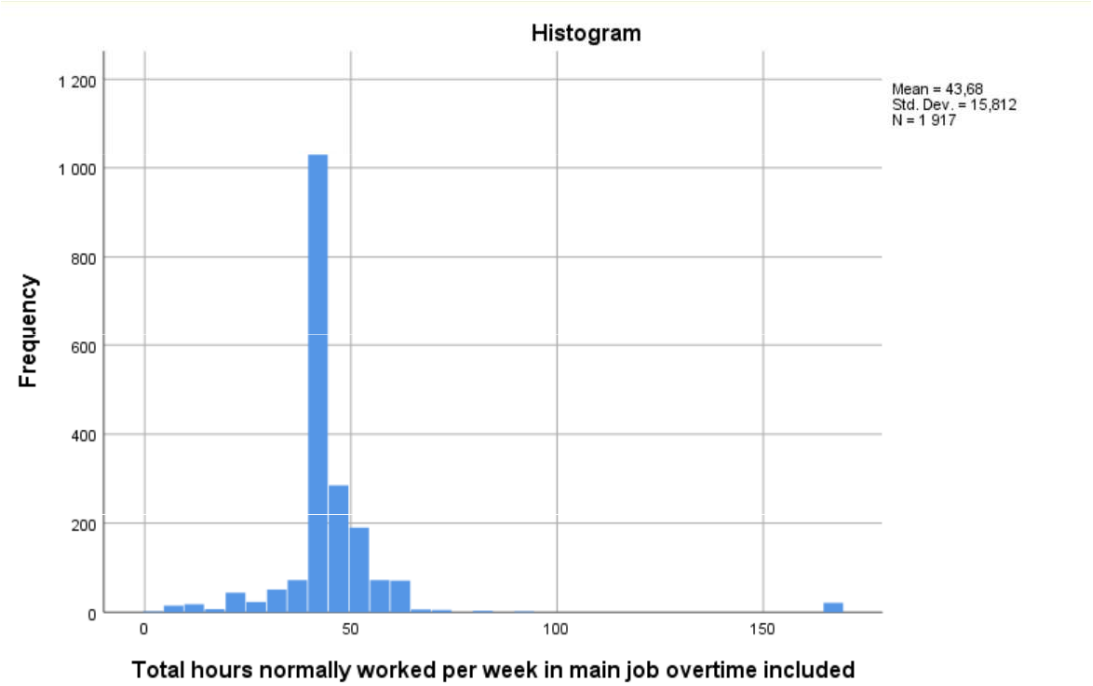

Industry, NACE rev.2
Crop and animal production, hunting and related service activities
Forestry and logging
Fishing and aquaculture
Mining of coal and lignite
Mining support service activities
Manufacture of food products
Manufacture of beverages
Manufacture of textiles
Manufacture of wearing apparel
Manufacture of leather and related products
Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting ma
Manufacture of paper and paper products
Printing and reproduction of recorded media
Manufacture of coke and refined petroleum products
Manufacture of chemicals and chemical products
Manufacture of basic pharmaceutical products and pharmaceutical preparations
...

Co kvantitativní (kardinální proměnné?)



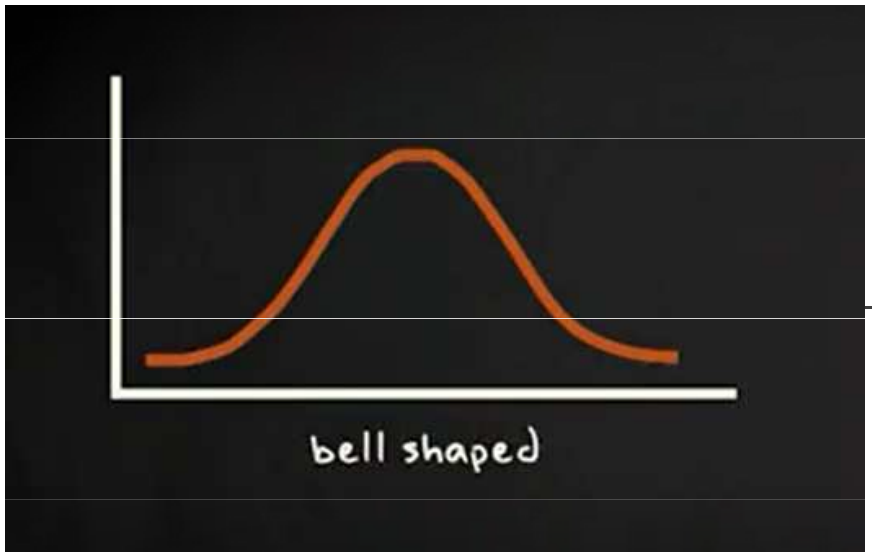
	estsz	jbspv	njbspv	wkdcorga	iorgact	wkhct	icwhct	[redacted]	nacer2
3	100 to 499	No		6	1	40	Yes	44	Construction of buildings
4	100 to 499	Yes	2	8	1	46	Yes	50	Scientific research and development
5	500 or more	Yes	20	7	2	40	Yes	45	Public administration and defence; compulsory social security
6	100 to 499	Yes	6	9	7	43	Yes	45	Retail trade, except of motor vehicles and motorcycles
7	100 to 499	No		I have/had no influence	I have/had no influence	42	Yes	45	Manufacture of electrical equipment
8	25 to 99	Yes	5	4	1	40	Yes	45	Retail trade, except of motor vehicles and motorcycles
9	10 to 24	Yes	7	9	1	40	Yes	44	Manufacture of machinery and equipment n.e.c.
0	Under 10	No		I have/had no influence	I have/had no influence	36	Yes	36	Food and beverage service activities
1	25 to 99	No		I have/had no influence	I have/had no influence	42	Yes	42	Manufacture of machinery and equipment n.e.c.
2	25 to 99	No		5	I have/had no influence	40	Yes	50	Retail trade, except of motor vehicles and motorcycles
3	500 or more	No		8	6	40	Yes	45	Residential care activities
4	100 to 499	No		8	3	45	Yes	50	Services to buildings and landscape activities
5	500 or more	No		2	I have/had no influence	40	Yes	40	Wholesale trade, except of motor vehicles and motorcycles
6	100 to 499	Yes	4	9	8	40	Yes	40	Repair and installation of machinery and equipment
7	Under 10	No		4	I have/had no influence	40	Yes	40	Public administration and defence; compulsory social security
8									
9	500 or more	No		I have/had complete control	I have/had complete control		No	45	Specialised construction activities
0	Under 10	Yes		4	1		No		Motion picture, video and television programme production, sound recording and music pu

Histogram

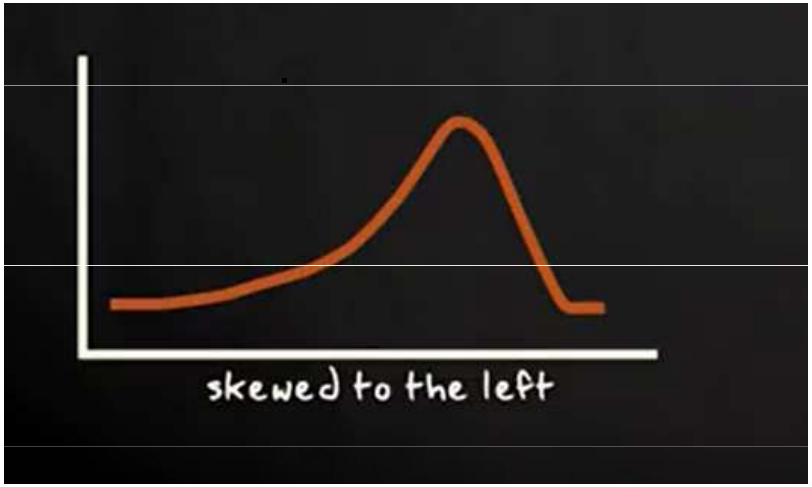

Histogram
Mean = 43,68
Std. Dev. = 15,812
N = 1 917
Frequency
1 200
1 000
800
600
400
200
0
0
50
100
150
Total hours normally worked per week in main job overtime included

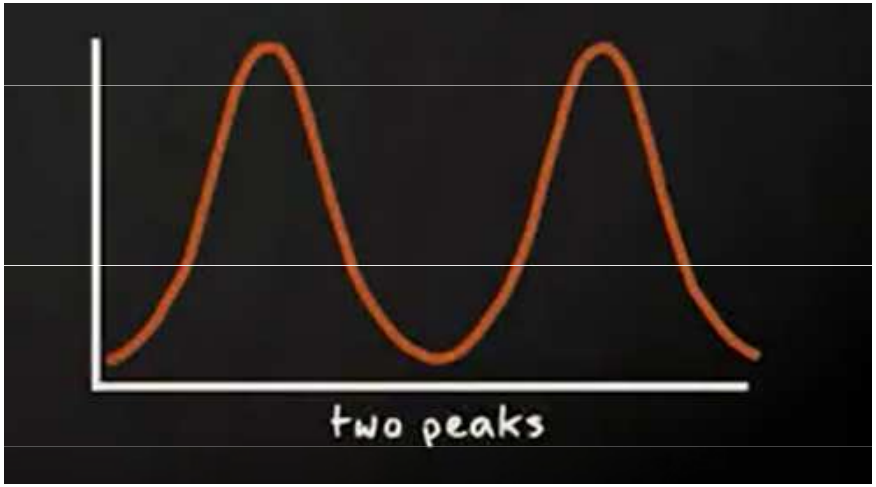
Histogram


43
Histogram
Mean = 43,68
Std. Dev. = 15,812
N = 1 917
Frequency
1 200
1 000
800
600
400
200
0
0
50
100
150
Total hours normally worked per week in main job overtime included


bell shaped


skewed to the right


skewed to the left


two peaks

Distribuce

Kolik má vrcholů?

Je symetrická (zvonový tvar, symetrie kolem střední hodnoty)

Symetrické rozložení = šikmost (skewness) blízk nule

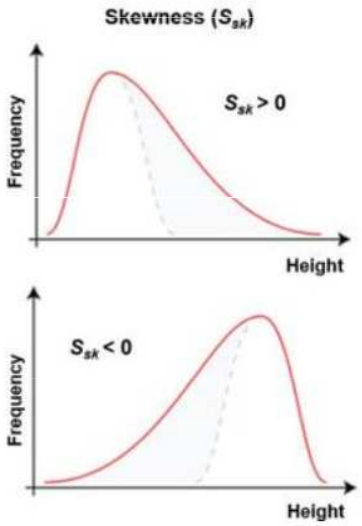
Pozitivní zešikmění (zešikmení zprava)

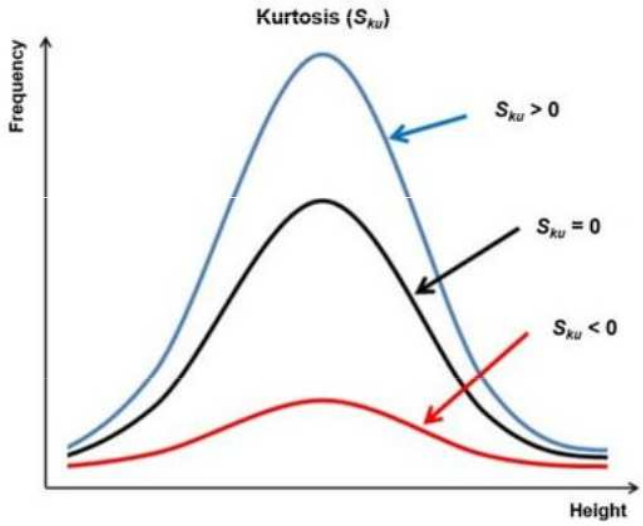
Negativní zešikmení (zleva)

Špičatost (kurtosis) = rozdělení blízko středu

Pozitivní špičatost = špičaté rozložení

Negativní špičatost = ploché/placaté





Centrální tendence distribuce

Užitečné k sumarizaci

Modus

Medián

Průměr

Centrální tendence distribuce

Užitečné k sumarizaci

Nominální data – modus

Ordinální data – modus, medián

Kategorická data – modus, medián, průměr

MODUS

Nejčastější hodnota

Který sloupec sloupcového grafu nebo histogramu je nejvyšší???

Který kus koláčového grafu je největší?

Může jich být více – pak máme multimodální distribuci

Více modů (třeba bimodální)

Můžu použít pro kardinální i kategorická data

MEDIÁN

Středová hodnota, rozděluje data set na dvě poloviny hodnot

Seřadíme hodnoty vzestupně

Najdeme tu, která leží uprostřed data setu (jednodušší pro matice s lichým počtem hodnot)

Hodnota, pod kterou leží 50 % hodnot a nad kterou leží 50 % hodnot

V kategorických datech = mediánová kategorie (kumulativní četnost zahrnuje 50% případů pod mediánem)

50. percentil

Výhoda: je stabilní, není citlivý na extrémní hodnoty

Medián: příklad

Počet odpracovaných hodin týdně pro 11 lidí: 45, 20, 56, 33, 18, 70, 40, 8, 40, 48, 59

Seřadíme vzestupně: 8, 18, 20, 33, 40, 40, 45, 48, 56, 59, 70

Najdu hodnotu, co leží uprostřed (na 6. místě): 8, 18, 20, 33, 35, 40, 45, 48, 56, 59, 70

SUDÝ POČET ČÍSEL: 8, 18, 20, 33, 35, 40, 45, 48, 56, 59

- Je to hodnota uprostřed dvou prostředních naměřených hodnot = (35+40)/2 = 37,5

PRŮMĚR

Jednoduchý aritmetický průměr

$$\bar{x} = \frac{\Sigma x}{n}$$

Pozor na extrémní hodnoty

Průměrný počet nohou člověka?

Průměr: příklad

Měsíční plat tří random lidí v baru:

25 000, 32 0000, 40 000 Kč

$\bar{x}$ = (25 000 + 32 000 + 40 000)/3 = 97 000/3 = 32 333,3 Kč

Průměr: příklad

Jaký je průměr?

$\bar{x}$ = 86 750 Kč

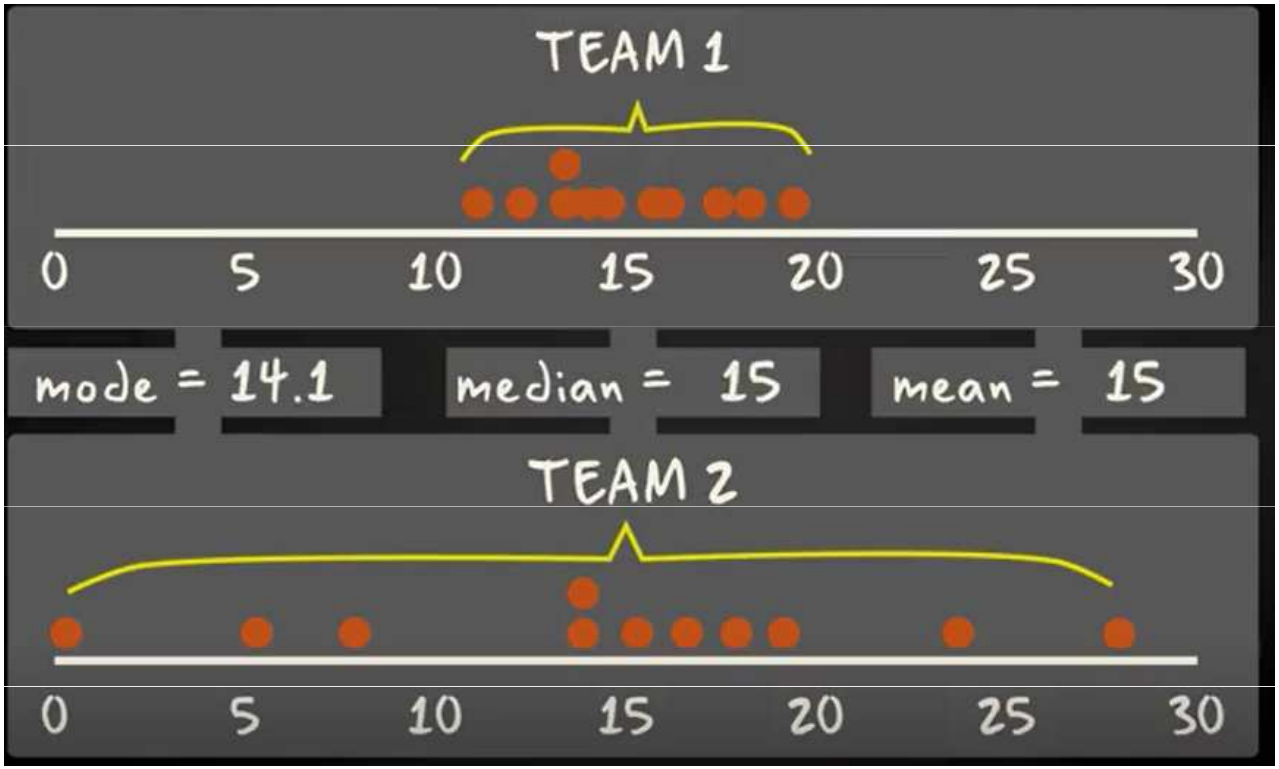
Míry centrální tendence

Jsou jednou z cenných informací o distribuci dat

ALE!!!

Nestačí. Potřebujeme znát i míru rozptýlenosti těch dat (dispersion)

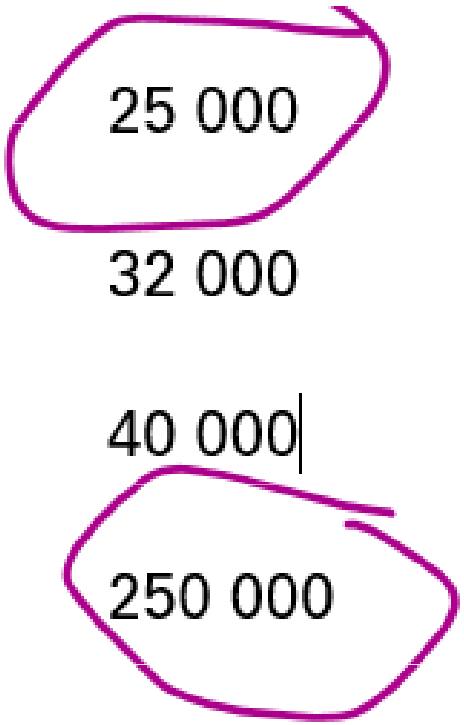
Příklad: potetovaná plocha z celkové plochy těl fotbalových hráčů??



ROZPĚTÍ (range)

250 000 – 25 000 = 225 000

Variační rozpětí je taky citlivé na extrémní hodnoty



Tetování: dva týmy

8,5

27,7

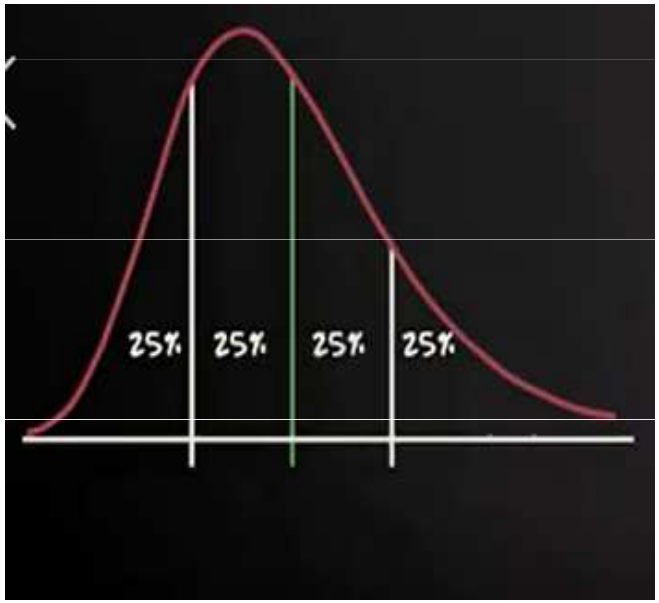
MEZIKVARTILOVÉ ROZPĚTÍ

Interquartile range (IQR)

Nebere v úvahu odlehlé hodnoty (výhoda)

Kvartil – hodnoty, které dělá soubor na čtyři stejně velké části

Je to první, druhý a třetí kvartil



MEZIKVARTILOVÉ ROZPĚTÍ

Interquartile range (IQR)

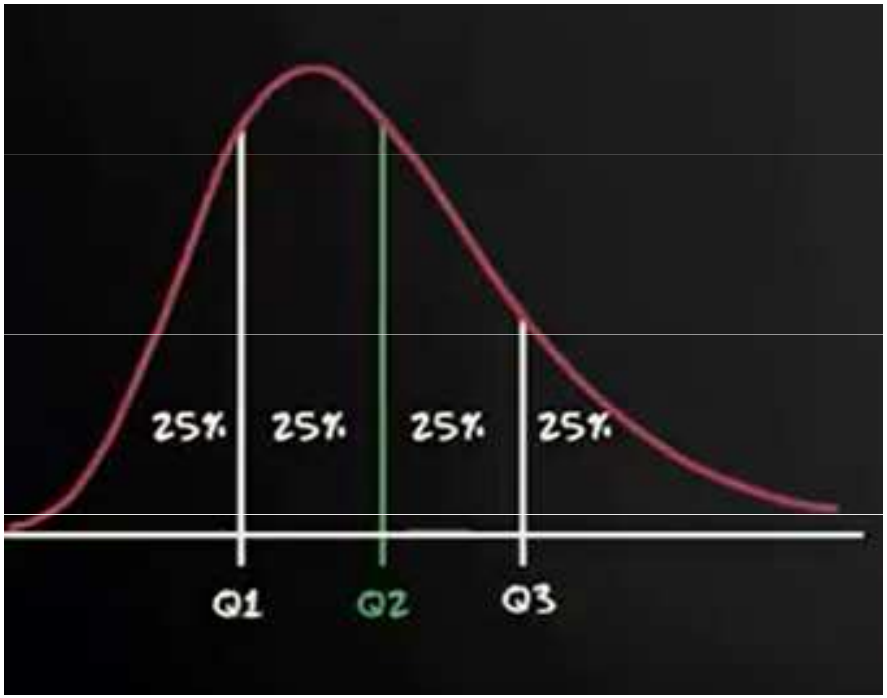
Nebere v úvahu odlehlé hodnoty (výhoda)

Kvartil – hodnoty, které dělá soubor na čtyři stejně velké části

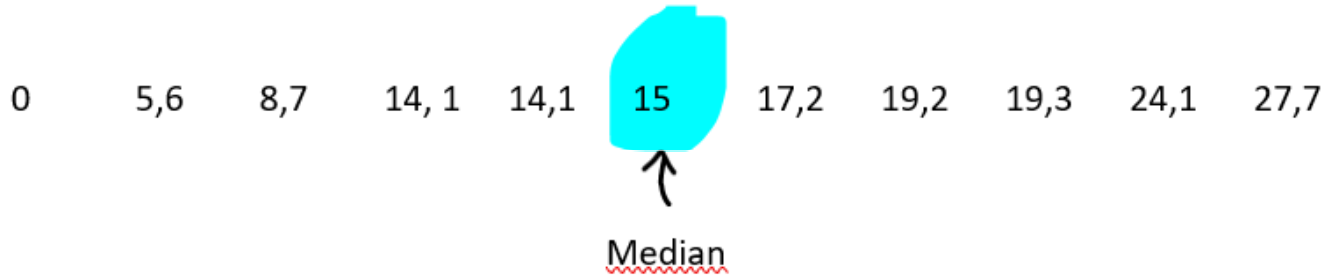
Je to první, druhý a třetí kvartil

IQR je rozdíl mezi Q3 a Q1

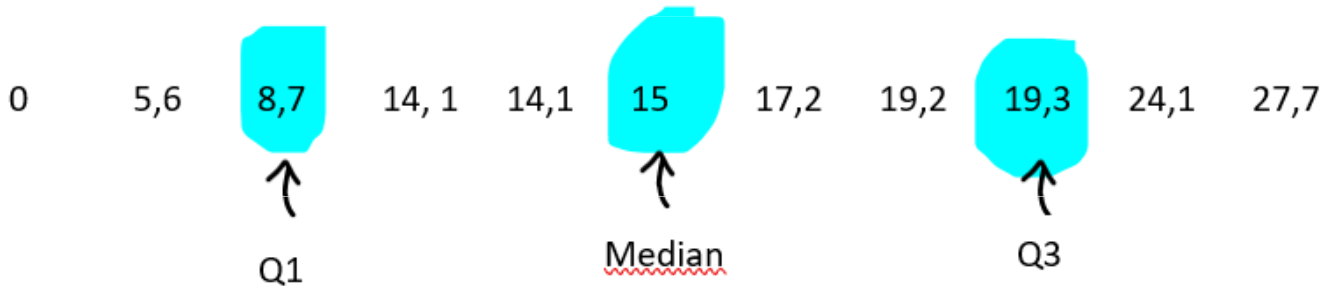
IQR = Q3 – Q1



Najít Q2

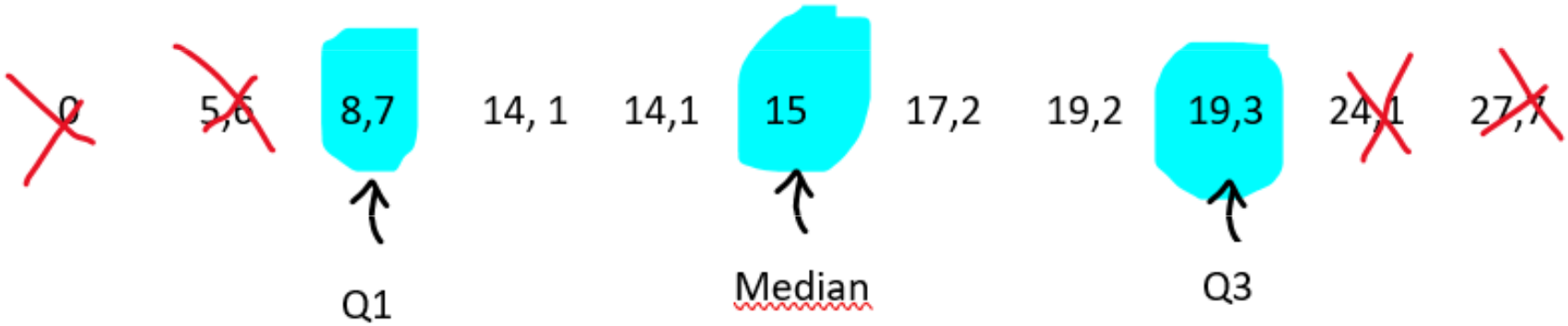


Najdu Q1 a Q3



IQR = 19,3 – 8,7 = 10,6

Je to vhodné proto, že IQR nebere v úvahu velmi nízké a velmi vysoké hodnoty



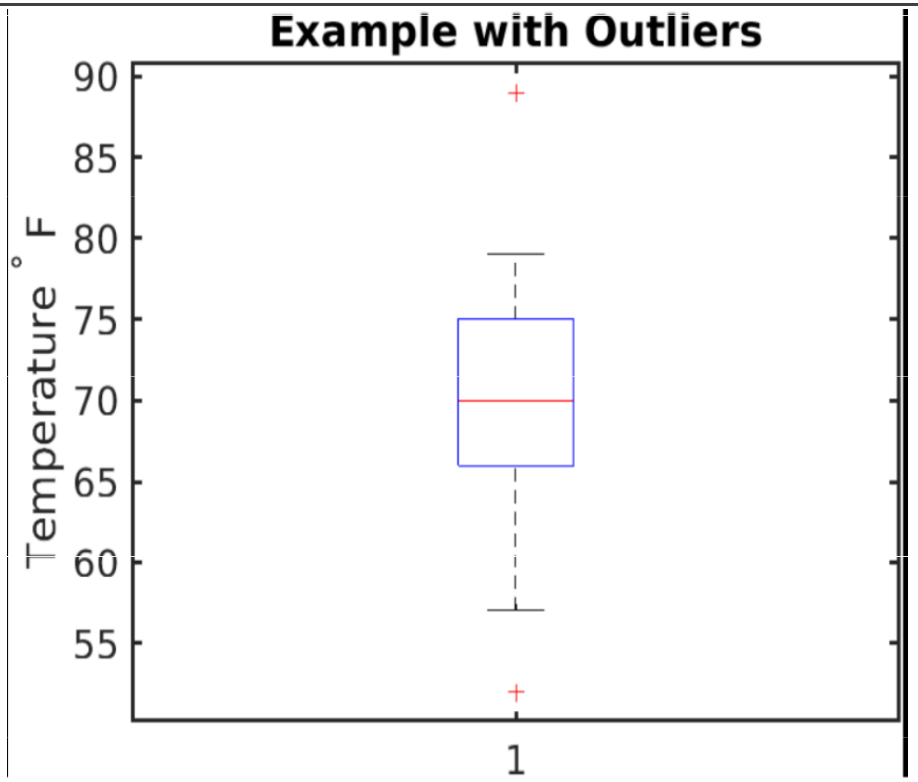
Co když mě zajímají odlehlé hodnoty???

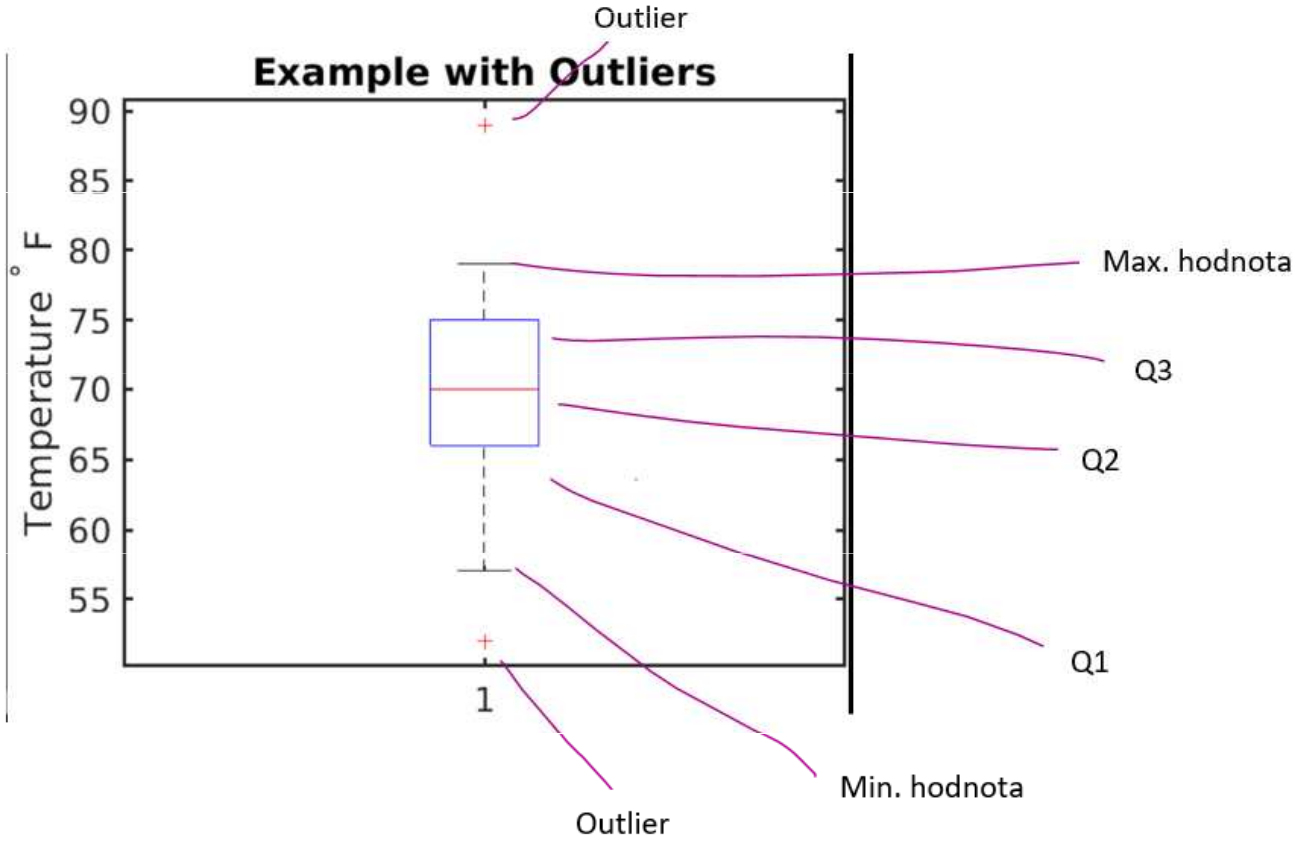
<Q1 – 1,5(IQR)

>Q3 + 1,5(IQR)

Za těmito hodnotami leží tzv. outliers

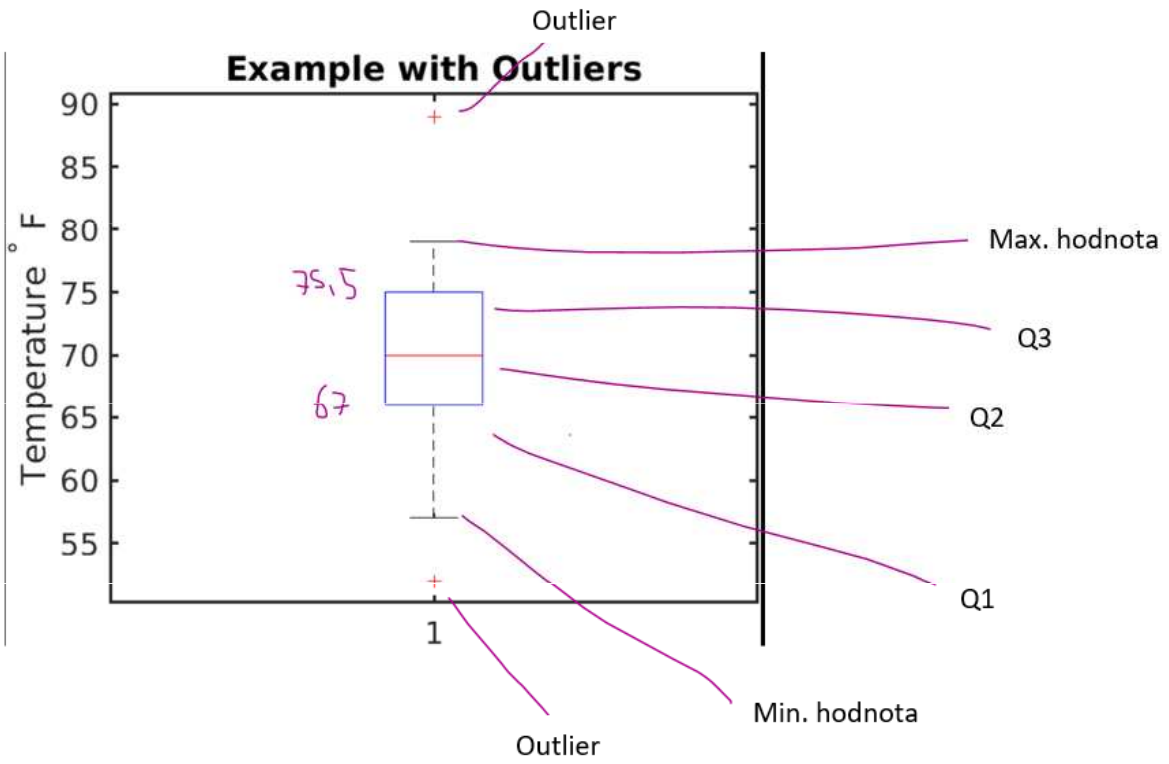
BOXPLOT: krabicový graf




Example with Outliers
Temperature °F
90
85
80
75
70
65
60
55
1
Outlier
Max. hodnota
Q3
Q2
Q1
Min. hodnota
Outlier

Za jakými hodnotami leží odlehlé případy????

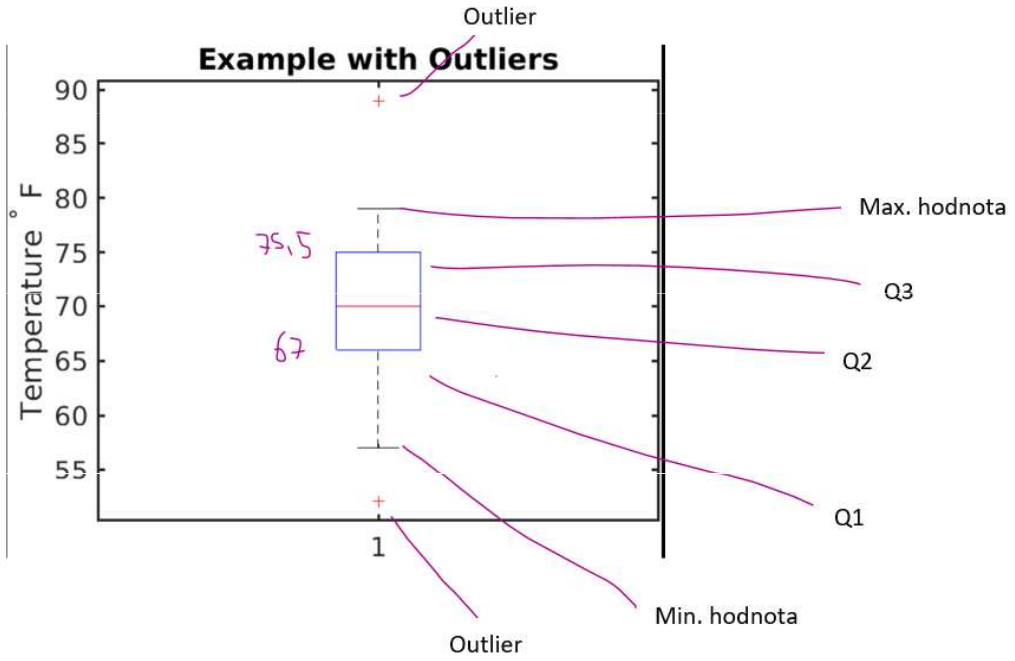
< Q1 - 1,5(IQR) a > Q3 + 1,5(IQR)



Za jakými hodnotami leží odlehlé případy????

< Q1 - 1,5(IQR) a > Q3 + 1,5(IQR)

IQR = 75,5 – 67 = 8,5
67 – 1,5(8,5) = 54,2
75,5 + 1,5(8,5) = 88,25



ROZPTYL (variance)

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

ROZPTYL

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Hráči	x	x - x̄	$(x - \bar{x})^2$
Hráč 1	0	-15	$(-15)^2 = 225$
Hráč 2	24,1	9,1	82,81
Hráč 3	5,6	-9,4	88,36
Hráč 4	14,1	-0,9	0,81
Hráč 5	17,2	2,2	4,84
Hráč 6	8,7	-6,3	39,69
Hráč 7	19,2	4,2	17,64
Hráč 8	14,1	-0,9	0,81
Hráč 9	27,7	12,7	161,29
Hráč 10	15	0	0
Hráč 11	19,3	4,3	18,49
		0	639,74

$\bar{x} = 15$

$\Sigma(x - \bar{x})^2$ suma čtverců

n-1 = 10

639,74/10 = 63,97

ROZPTYL

Čím větší rozptyl, tím větší variabilita dat.

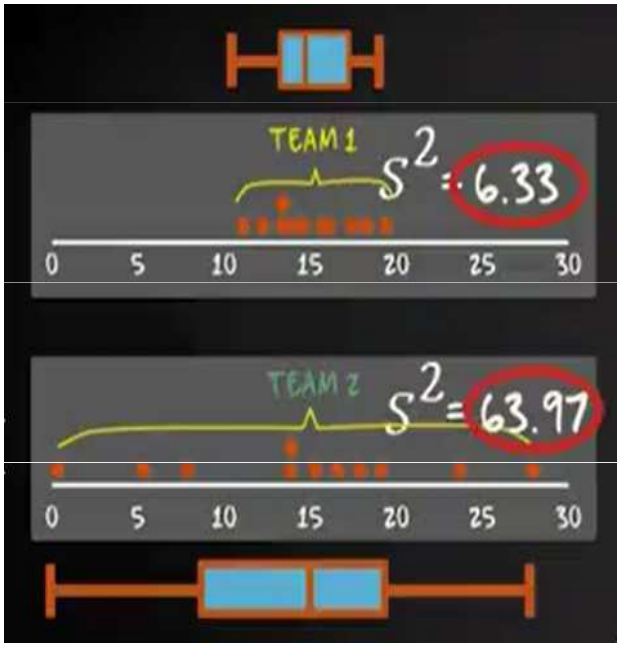
Tím více jsou rozptýlena.

Indikuje to rozptyl a vizuálně i krabicový graf

(příklad s fotbalisty)

Co je nevýhoda?

Je udán ve stupnici měřené proměnné ale na druhou.



SMĚRODATNÁ ODCHYLKA (standard deviation)

Odmocníme hodnotu rozptylu!

Čím větší SD, tím větší variabilita v datech.

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

V našem fotbalovém týmu byl rozptyl 63,97, tím pádem směrodatná odchylka je 8.

Jako míra disperze dat se používá nejčastěji.