

VADÍ - NEVADÍ ANEB STATISTIKA KOLEM NÁS

Martina Litschmannová

Vysoká škola báňská – Technická univerzita Ostrava

Fakulta elektrotechniky a informatiky, Katedra aplikované matematiky



Obsah

Část 1

- Analýza dat – Základní pojmy
- Popisná statistika kvalitativního znaku – tabulky četnosti, vizualizace
- Jak to vypadá v praxi

Část 2

- Popisná statistika kvantitativního znaku – míry polohy, vizualizace
– míry variability, zaokrouhlování (spíše nestihneme)

Co je to statistika?

Google – $83 \cdot 10^6$ odkazů (čeština), $1,3 \cdot 10^9$ odkazů (angličtina)

- Teoretická disciplína, která se zabývá metodami sběru a analýzy dat (matematická statistika vs. aplikovaná statistika)
- Číselný údaj „syntetizující“ vlastnosti datových souborů (četnost, průměr, rozptyl, ...)
- Uspořádaný datový soubor (statistika přístupů na web. stránky, statistika střel na branku, statistika nehodovosti, ekonomické statistiky, ...)

Co vypovídá statistika o jednotlivci?



Donald Trump



podnikatel



politik (prezident)

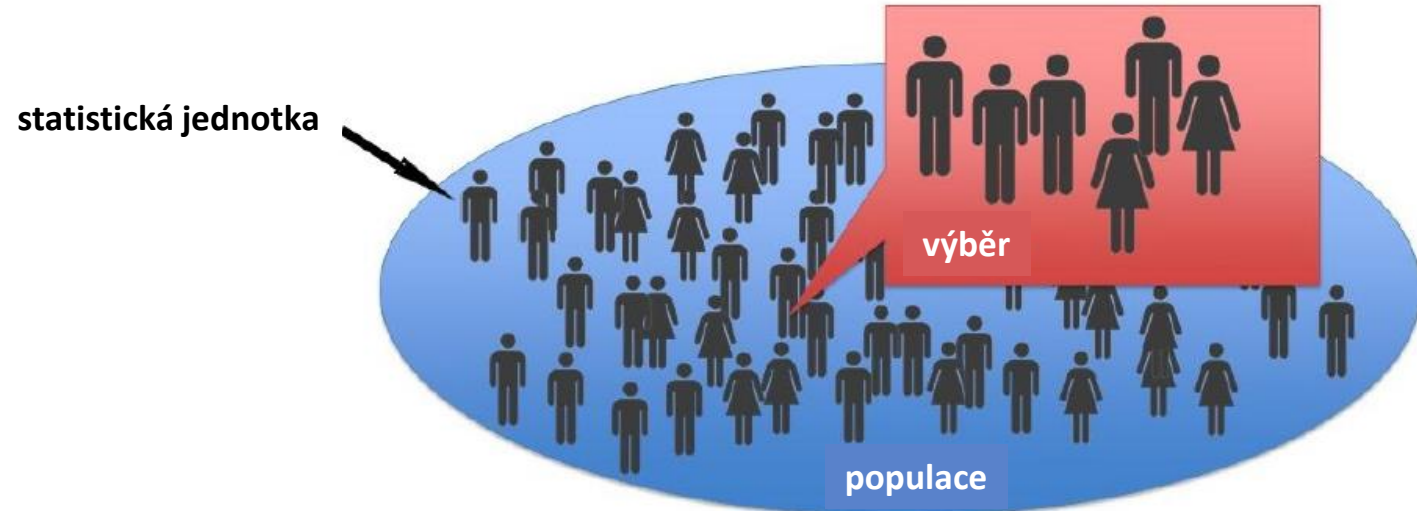


američan

- Statistika nezkoumá jednotlivce jako individualitu, ale jako anonymního nositele některého znaku (činnosti, vlastnosti).

Základní pojmy

- **Populace** (základní soubor) je soubor nějakých prvků, o kterém chceme statistickými metodami něco vypovídat. Definuje se výčtem nebo pomocí zvolené vlastnosti. O každém prvku umíme rozhodnout, zda do populace patří či nikoliv.
- **Výběr** je část dané populace, která má sloužit k odvození závěrů platných pro celou populaci. (Pozor na reprezentativnost výběru!)
- **Statistická jednotka** je prvek populace.
- **Statistický znak (proměnná)** je nějaká měřitelná (zjistitelná) charakteristika statistické jednotky (hmotnost, pohlaví, ...).



Jak zapsat výsledky statistického šetření?

„nějaká“ tabulka

kapacita akumulátoru (mAh)							
po 5 cyklech				po 100 cyklech			
Výrobce A	Výrobce B	Výrobce C	Výrobce D	Výrobce A	Výrobce B	Výrobce C	Výrobce D
1946,5	2006,5	1881,8	1806,9	1780,4	1654,2	1663,3	1668,4
1963,5	1991,5	1890,4	1788,1	1751,4	1663,1	1641,1	1641,9
1934,3	1988,8	1865,7	1775	1743,5	1633,3	1621,5	1620
1934,8	1975,4		1805,4	1727,4	1642,2		1685,8
1939,9	1998,4		1775,7	1728,8	1656,7		1610,5

Nevýhody:

- Obtížnější analýza pomocí statistického software.
- Chybí jednoznačná identifikace příslušných statistických jednotek.

Jak zapsat výsledky statistického šetření?

Standardní datový formát (datová matice)

id	kap5	kap100	vyrobce
1	1946.5	1780.4	A
2	1963.5	1751.4	A
3	1934.3	1743.5	B
4	1934.8	1727.4	B
5	1939.9	1728.8	C
6	1925.9	1767.5	C
7	2023	1838.7	D
8	1952.5	1734.1	A
9	1894.7	1688.8	D



- Každý **řádek** matice obsahuje **údaje o jedné statistické jednotce**.
- **V prvním sloupci** (nebo jako popisky řádků) se obvykle uvádí **identifikační číslo** statistické jednotky (důležité pro jednoznačné spárování s konkrétní statistickou jednotkou, zejména při poskytování anonymizovaných dat zpracovateli).

Vlastní pokus o výběrové šetření

- Dotazník pro studenty (např. pomocí Google Apps) - <http://goo.gl/forms/Z289s0ALPY>

Dotazník pro výukové účely - Brigáda při studiu

Vážené studentky, vážení studenti,

prosím Vás o vyplnění krátkého anonymního dotazníku na téma brigáda při studiu. Výsledky dotazníkového šetření budou použity pro prezentaci analýzy dat v rámci výuky.

Předem moc děkuji za váš čas a ochotu,
Martina Litschmannová

*Povinné pole

Identifikace respondenta

1. Pohlaví *

Označte jen jednu elipsu.

- ☐ žena
☐ muž

2. Výška (cm) *

.....

3. Váha (kg) *

.....

4. Přivyděláváte si v rámci prezenčního studia na brigádách? *

Označte jen jednu elipsu.

- ☐ ano *Přeskočte na otázku 5.*
☐ ne *Přeskočte na otázku 8.*

Brigáda při studiu

5. Jak často brigádu máte? *

Označte jen jednu elipsu.

- ☐ každý pracovní den
☐ pouze o víkendu
☐ nepravidelně
☐ Jiné:

6. Jak byste svou brigádu charakterizoval(a)? *

Označte jen jednu elipsu.

- ☐ praxe v oboru během studia
☐ "kancelářská" práce mimo obor mého studia
☐ manuální práce
☐ Jiné:

7. Kolik času týdně obvykle věnujete brigádě? *

Uvedte čas v hodinách.

.....

Studium

8. Kolik času týdně obvykle věnujete studiu? *

Uvedte čas v hodinách.

.....

Základní pojmy

- **Populace** (základní soubor) je soubor nějakých prvků, o kterém chceme statistickými metodami něco vypovídat.
- **Výběr** je část dané populace, která má sloužit k odvození závěrů platných pro celou populaci.
- **Statistická jednotka** je prvek populace.
- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika statistické jednotky.

Datová matice

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídaní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

Základní pojmy

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.

Základní pojmy

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.

Základní pojmy

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.
 - **Kvantitativní znak** – znak, jehož varianty mají číselné hodnoty (má smysl posuzovat rozdíly a poměry)

Základní pojmy

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.
 - **Kvantitativní znak** – znak, jehož varianty mají číselné hodnoty (má smysl posuzovat rozdíly a poměry)

Základní pojmy

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.
 - **Kvantitativní znak** – znak, jehož varianty mají číselné hodnoty (má smysl posuzovat rozdíly a poměry)
 - **Kvalitativní znak** – znak, jehož varianty se liší kvalitou (může jít i o číselné hodnoty – např. známka z matematiky)

Základní pojmy

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.
 - **Kvantitativní znak** – znak, jehož varianty mají číselné hodnoty (má smysl posuzovat rozdíly a poměry)
 - **Kvalitativní znak** – znak, jehož varianty se liší kvalitou (může jít i o číselné hodnoty – např. známka z matematiky)

Základní pojmy

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídnání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.
 - **Kvantitativní znak** – znak, jehož varianty mají číselné hodnoty (má smysl posuzovat rozdíly a poměry)
 - **Kvalitativní znak** – znak, jehož varianty se liší kvalitou (může jít i o číselné hodnoty – např. známka z matematiky)
 - **Alternativní (dichotomický) znak** – znak, který nabývá pouze dvou variant

Základní pojmy

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídnání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

- **Statistický znak** je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.
 - **Kvantitativní znak** – znak, jehož varianty mají číselné hodnoty (má smysl posuzovat rozdíly a poměry)
 - **Kvalitativní znak** – znak, jehož varianty se liší kvalitou (může jít i o číselné hodnoty – např. známka z matematiky)
 - **Alternativní (dichotomický) znak** – znak, který nabývá pouze dvou variant

**Popisná statistika
aneb
Jak ~~efektivně~~ efektivně popsat a vizualizovat data**

Část 1
Kvalitativní znak

Popisná statistika – Kvalitativní znak

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

Respondent (proband) – označení statistické jednotky v dotazníkovém šetření

Popište strukturu datového souboru v závislosti na pohlaví respondentů.

Popisná statistika – Kvalitativní znak

- Tabulka četností

Tabulka četností		
Varianta znaku	Absolutní četnost	Relativní četnost
x_i	n_i	p_i
x_1	n_1	$p_1 = n_1/n$
x_2	n_2	$p_2 = n_2/n$
x_k	n_k	$p_k = n_k/n$
Celkem:	$n_1 + n_2 + \dots + n_k = n$	1

+ **Modus** (varianta, které znak nabývá s nejvyšší četností)

Popisná statistika – Kvalitativní znak

- Tabulka četností

Tabulka četností		
Pohlaví	Absolutní četnost	Relativní četnost
muž	66	0,776470588
žena	19	0,223529412
Celkem:	85	1,000000000

Popisná statistika – Kvalitativní znak

- Tabulka četností

Tabulka četností		
Pohlaví	Absolutní četnost	Relativní četnost (%)
muž	66	77,6470588
žena	19	22,3529412
Celkem:	85	100,0000000

Jak zaokrouhlovat relativní četnosti?

1,0 % ... 0,85 osob
0,1 % ... 0,085 osob



Popisná statistika – Kvalitativní znak

- Tabulka četností

Tabulka četností		
Pohlaví	Absolutní četnost	Relativní četnost (%)
muž	66	77,6470588
žena	19	22,3529412
Celkem:	85	100,0000000

} Pozor na zaokrouhlovací chybu!

Jak zaokrouhlovat relativní četnosti?

1,0 % ... 0,85 osob
0,1 % ... 0,085 osob



Popisná statistika – Kvalitativní znak

- Tabulka četností

Tabulka četností		
Pohlaví	Absolutní četnost	Relativní četnost (%)
muž	66	78
žena	19	22
Celkem:	85	100

Pozor na zaokrouhlovací chybu!

Součet musí být 100 %!

Jak zaokrouhlovat relativní četnosti?

1,0 % ... 0,85 osob
0,1 % ... 0,085 osob



Příklad demonstrující problém zaokrouhlovací chyby

TABULKA ROZDĚLENÍ ČETNOSTI		
Typ pasažéra	Absolutní četnosti	Relativní četnosti (%)
Muž	77	37,37864
Žena	85	41,26214
Dítě	44	21,35922
Celkem:	206	100,00000

1 % ... 2,06 osob

0,1 % ... 0,206 osob

Jak zaokrouhlit relativní četnost?



Příklad demonstrující problém zaokrouhlovací chyby

TABULKA ROZDĚLENÍ ČETNOSTI		
Typ pasažéra	Absolutní četnosti	Relativní četnosti (%)
Muž	77	37,4
Žena	85	41,3
Dítě	44	21,4
Celkem:	206	100,1



POZOR na
zaokrouhlovací
chybu!



Příklad demonstrující problém zaokrouhlovací chyby

TABULKA ROZDĚLENÍ ČETNOSTI		
Typ pasažéra	Absolutní četnosti	Relativní četnosti (%)
Muž	77	37,4
Žena	85	41,3
Dítě	44	21,3
Celkem:	206	100,0

Dopočet do 100 %!



Popisná statistika – Kvalitativní znak

- Tabulka četností

Tabulka četností		
Pohlaví	Absolutní četnost	Relativní četnost (%)
muž	?	78
žena	?	22
Celkem:	85	100

Relativní četnosti používejte pouze jako doplněk absolutních četností!



Popisná statistika – Kvalitativní znak

- Tabulka četností

Tabulka četností		
Pohlaví	Absolutní četnost	Relativní četnost (%)
muž	66	78
žena	19	22
Celkem:	85	100

Určete modus proměnné pohlaví.

Popisná statistika – Kvalitativní znak

- Tabulka četností

Tabulka četností		
Pohlaví	Absolutní četnost	Relativní četnost (%)
muž	66	78
žena	19	22
Celkem:	85	100

Modus = muž

Mezi respondenty převažovali muži.

Popisná statistika – Kvalitativní znak

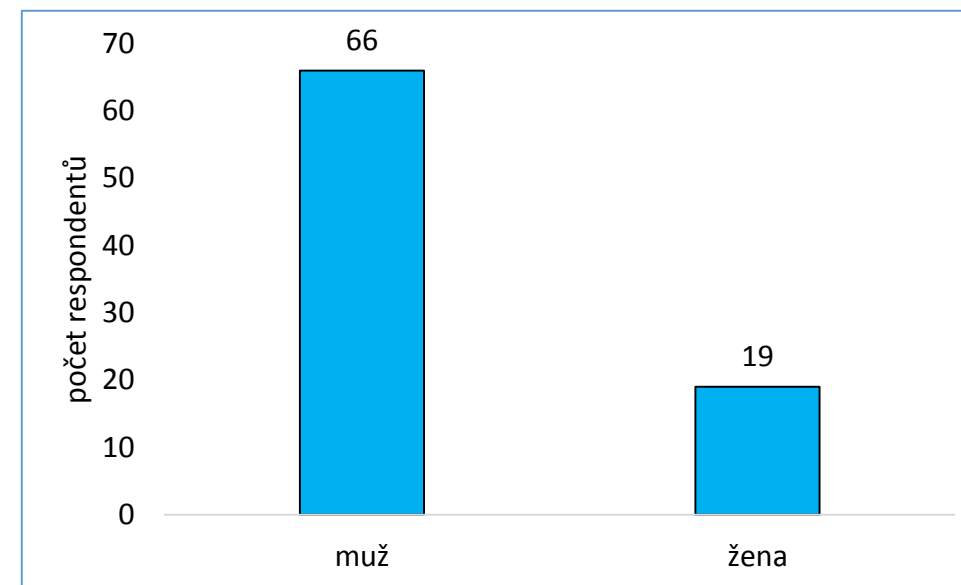
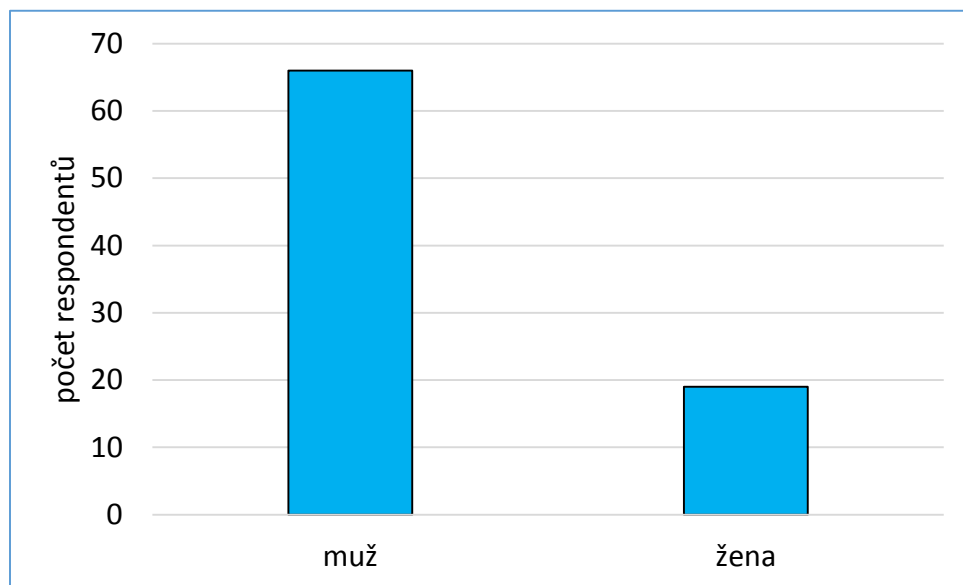
- Tabulka četností

Tabulka četností		
Pohlaví	Absolutní četnost	Relativní četnost (%)
muž	66	78
žena	19	22
Celkem:	85	100

Jak data vizualizovat?

Popisná statistika – kvalitativní znak

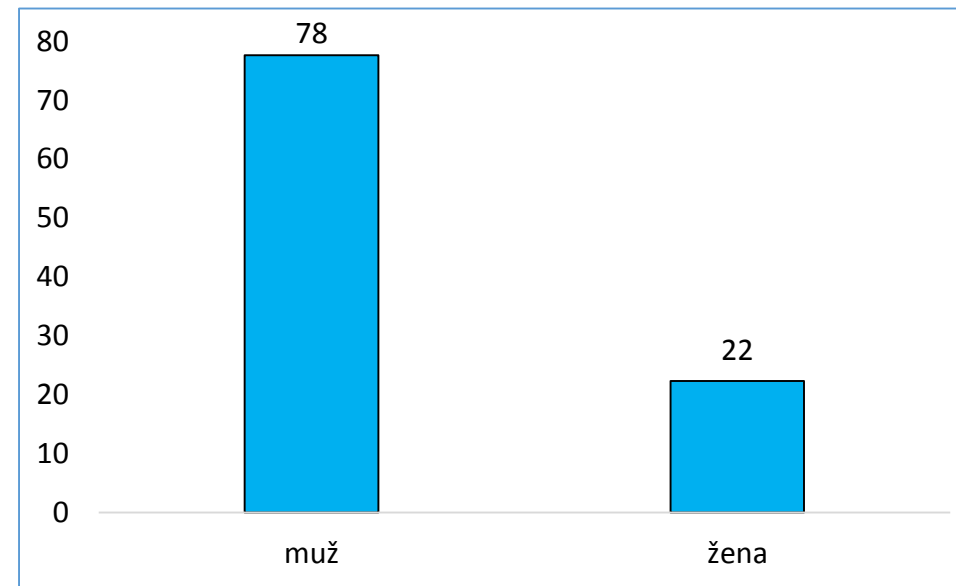
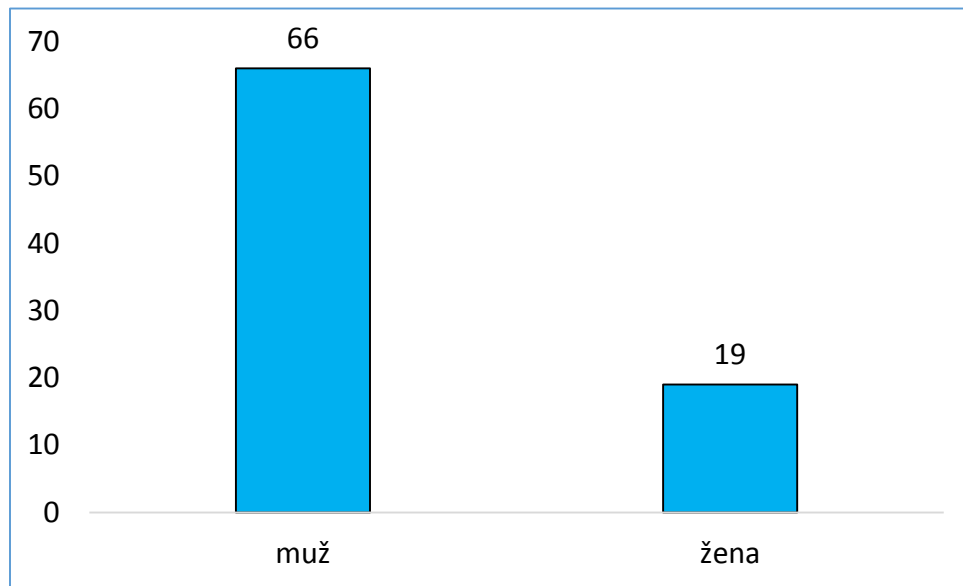
Sloupcový graf (Bar Chart)



- Nejsou-li v grafu uvedeny absolutní (relativní) četnosti, obvykle je nedokážeme „od oka“ přesně odečíst.

Popisná statistika – kvalitativní znak

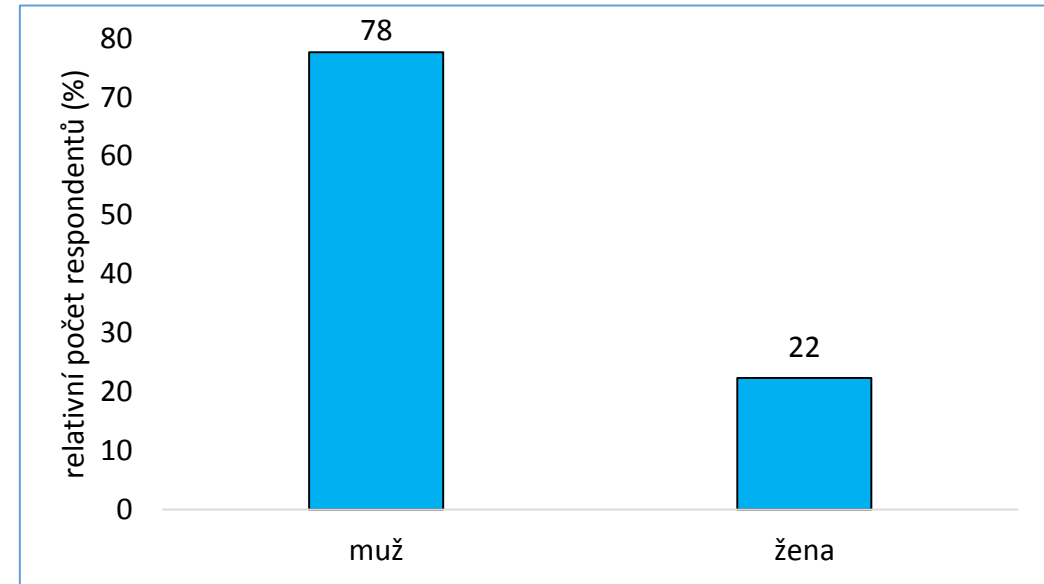
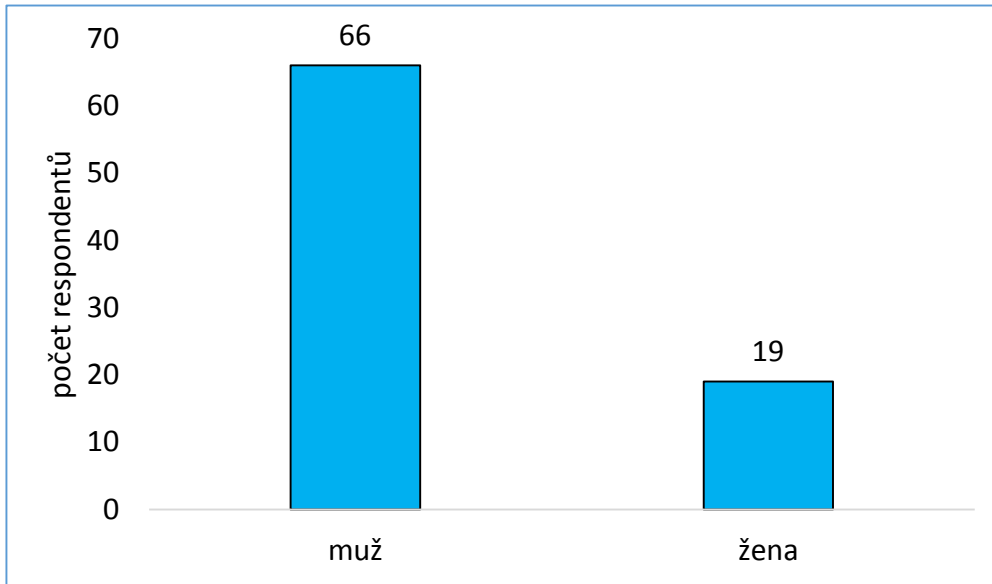
Sloupcový graf (Bar Chart)



- Nejsou-li v grafu uvedeny absolutní (relativní) četnosti, obvykle je nedokážeme „od oka“ přesně odečíst.

Popisná statistika – kvalitativní znak

Sloupcový graf (Bar Chart)



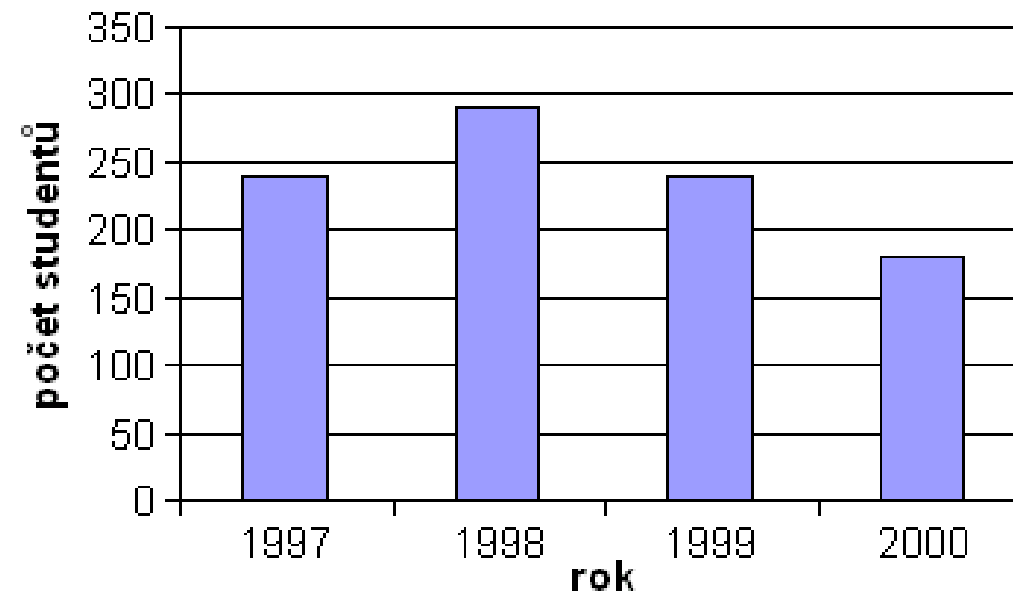
- Nejsou-li v grafu uvedeny absolutní (relativní) četnosti, obvykle je nedokážeme „od oka“ přesně odečíst.
- Pozor na uvádění popisu os!

Popisná statistika – Kvalitativní znak

Zdroj: Srovnávací testy pro žáky 9. tříd

Určete pravdivost tvrzení:

V žádných dvou letech nebyl počet studentů stejný.

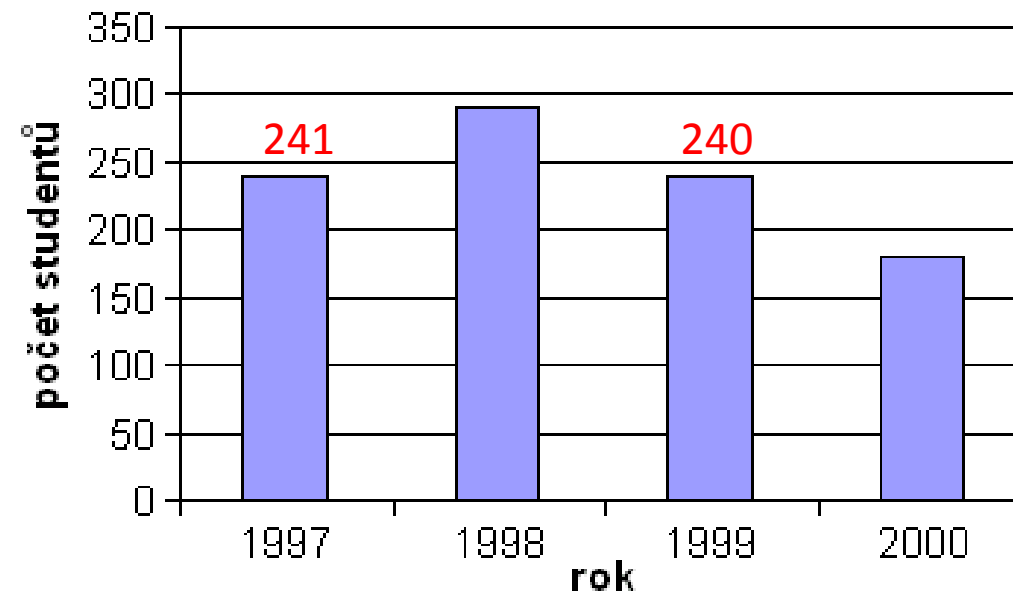


Popisná statistika – Kvalitativní znak

Zdroj: Srovnávací testy pro žáky 9. tříd

Určete pravdivost tvrzení:

V žádných dvou letech nebyl počet studentů stejný.

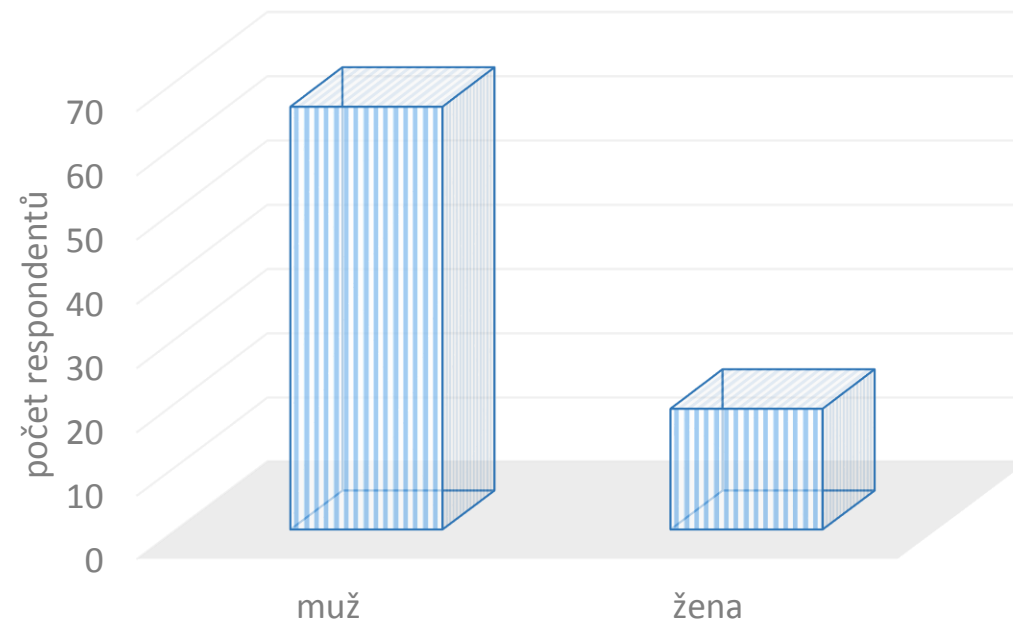


Pozor na omezenou
vypovídací schopnost grafů!



Popisná statistika – Kvalitativní znak

- Sloupcový graf

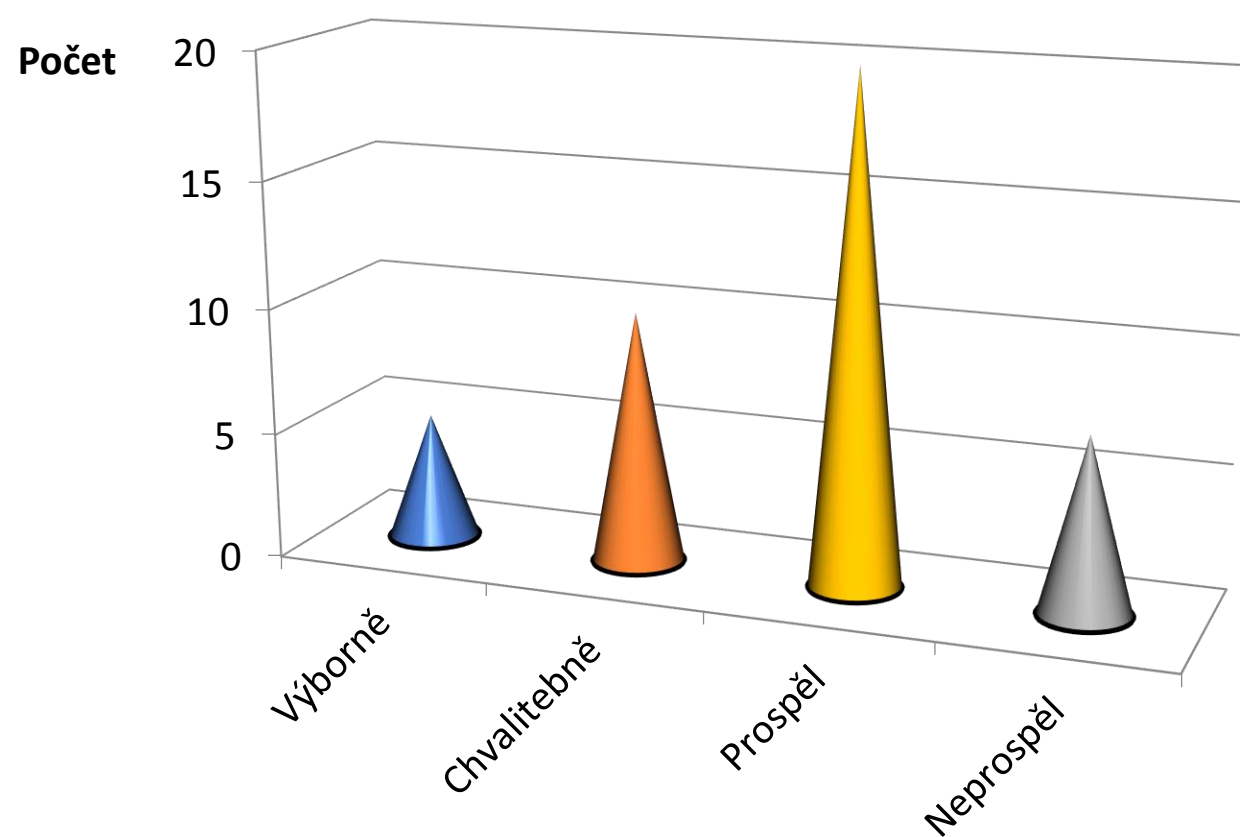


Nezapomínejte, že méně mnohdy znamená více ...



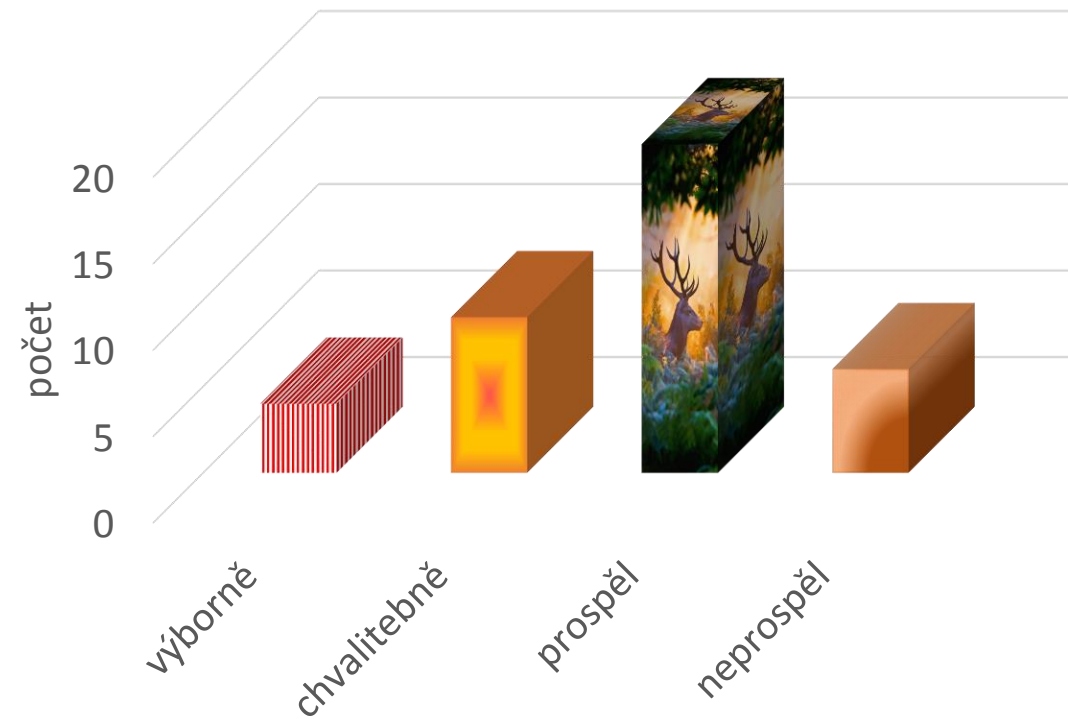
Popisná statistika – Kvalitativní znak

- I takto může vypadat sloupcový graf ...



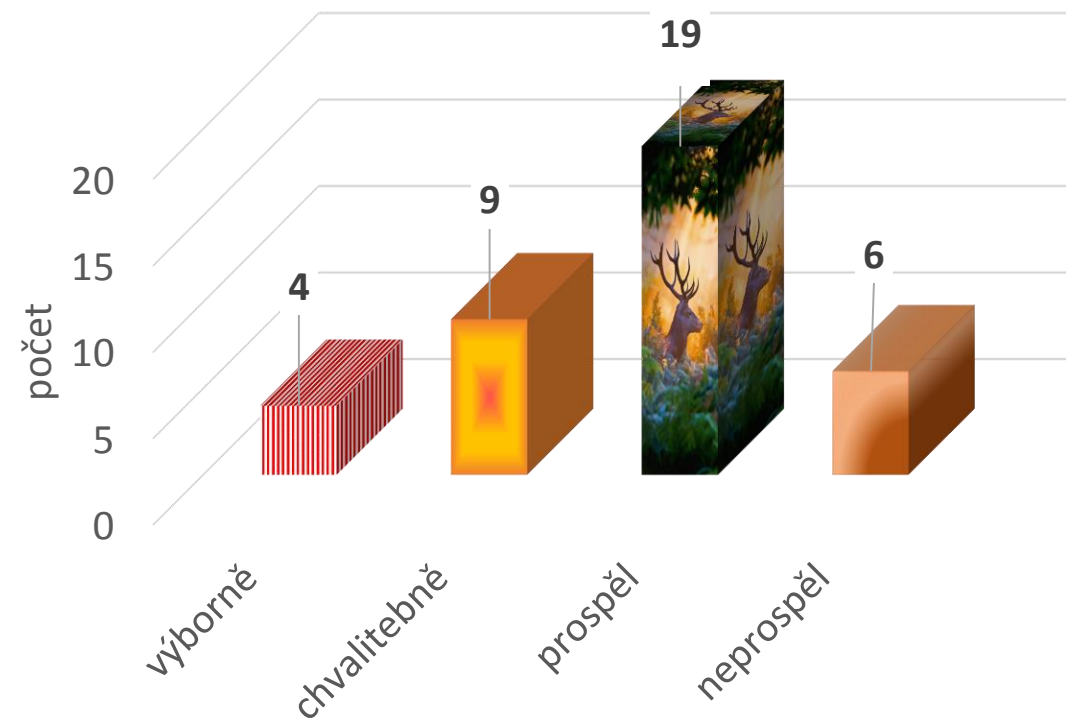
Popisná statistika – Kvalitativní znak

- I takto může vypadat sloupcový graf ...



Popisná statistika – Kvalitativní znak

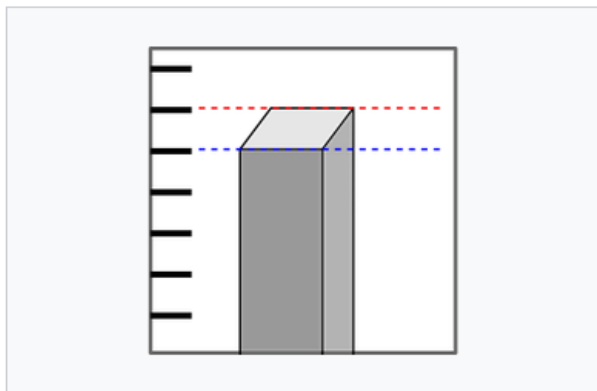
- I takto může vypadat sloupcový graf ...



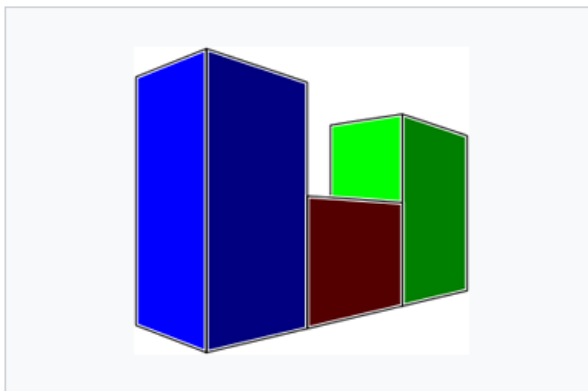
Popisná statistika – Kvalitativní znak

- 3D sloupcový graf

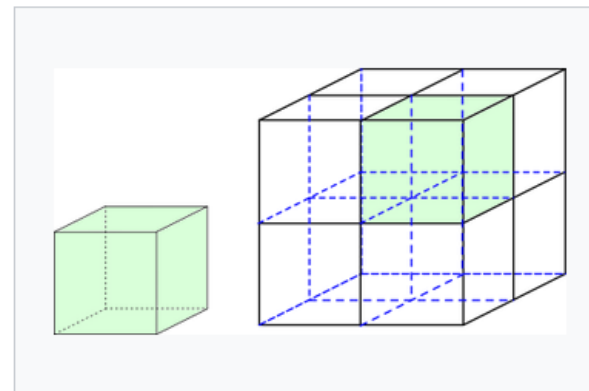
Opravdu musí být v každé efektní prezentaci / publikaci?



3D graf může být pro čtenáře matoucí...



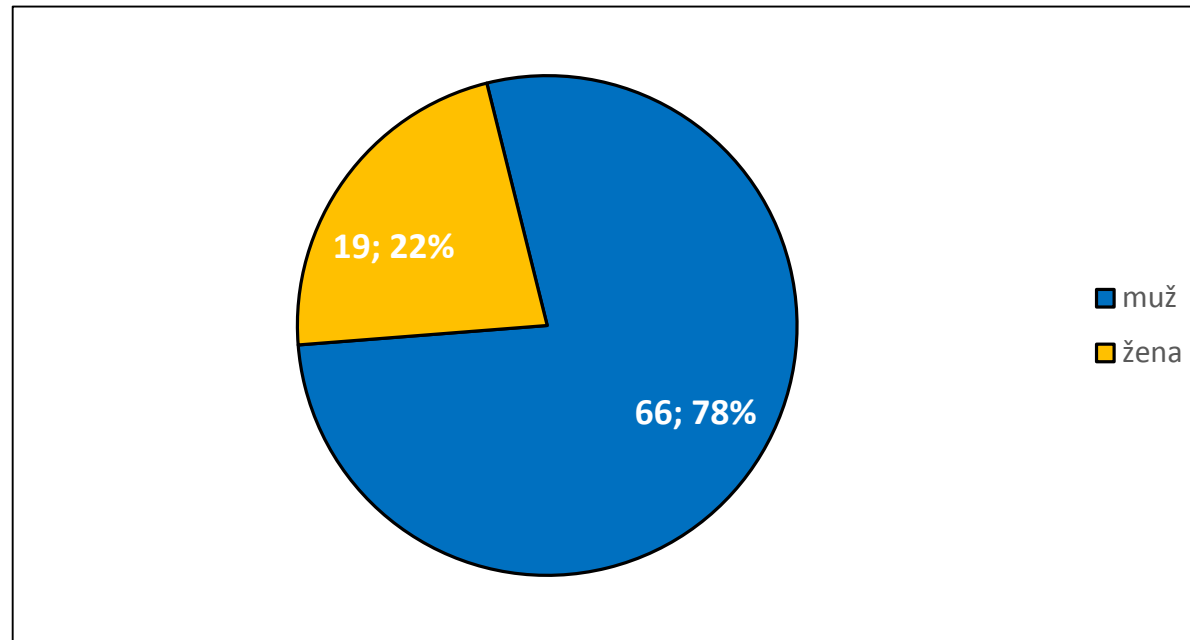
Modrý a zelený sloupec mají stejnou výšku. Odpovídá to tomu, jak obrázek vnímáte?



Zvětšíte-li k krát stranu krychle, její objem (to, co vnímáte) se zvětší k^3 krát.

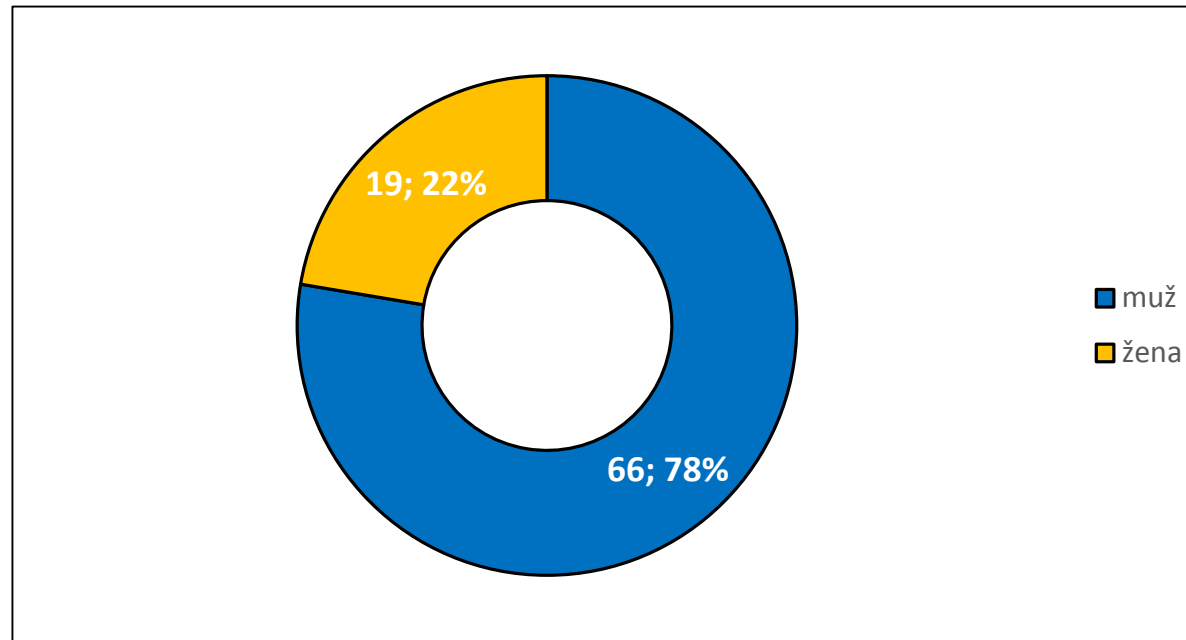
Popisná statistika – Kvalitativní znak

- Výsečový graf



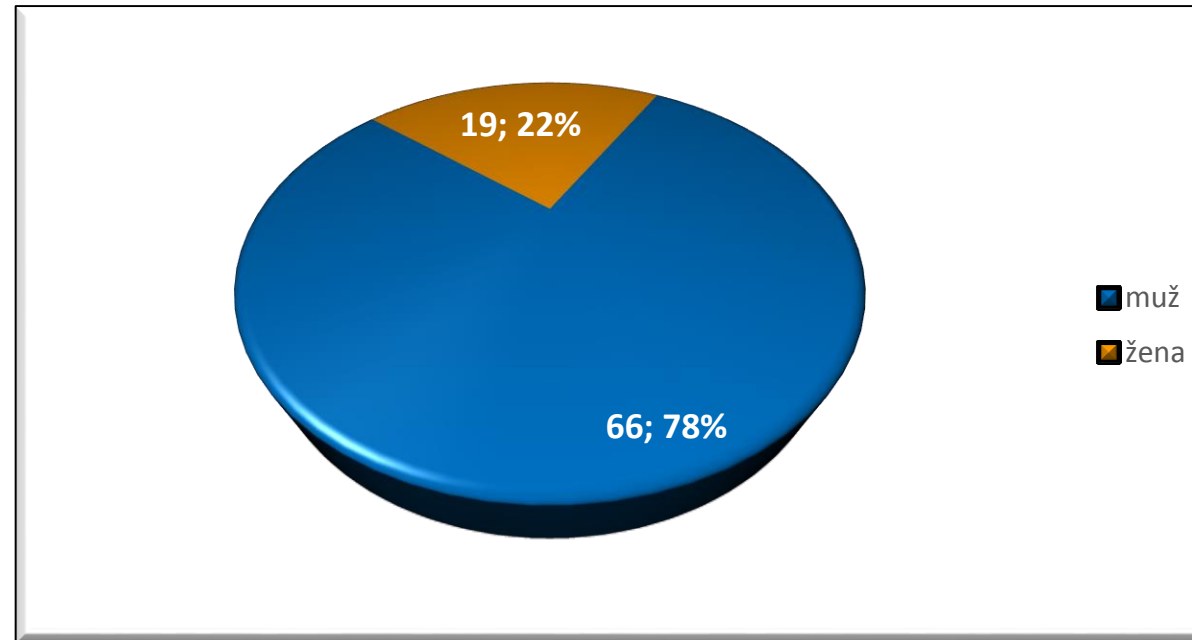
Popisná statistika – Kvalitativní znak

- Prstencový graf



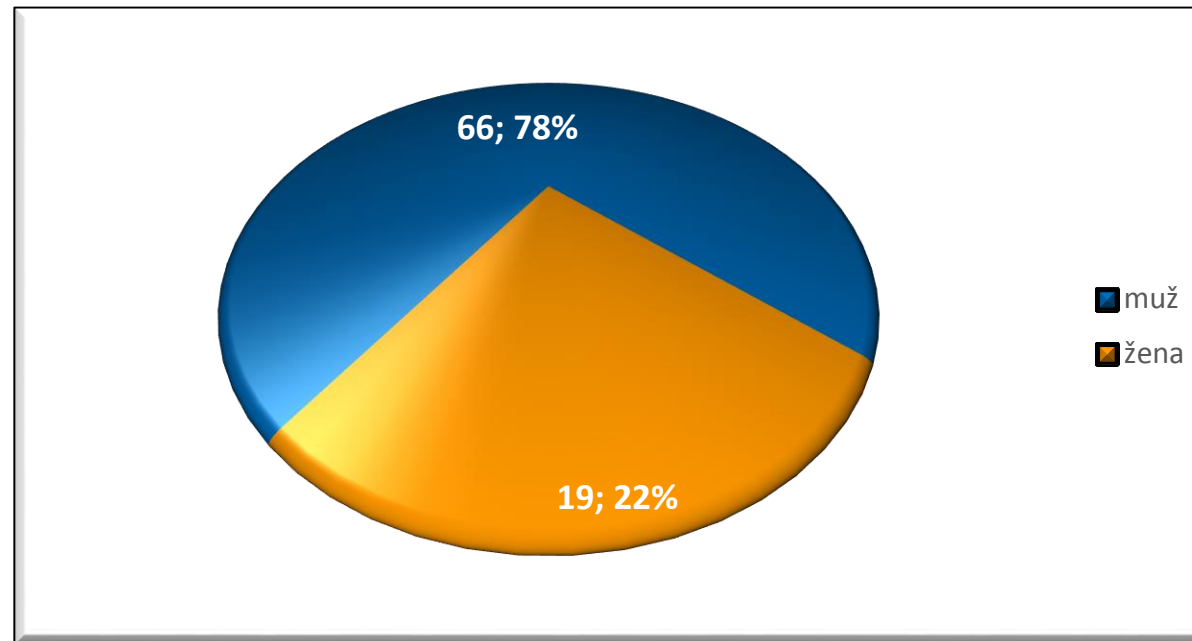
Popisná statistika – Kvalitativní znak

- 3D výsečový graf



Popisná statistika – Kvalitativní znak

- 3D výsečový graf



Pozor na vypovídací
schopnost 3D grafů!

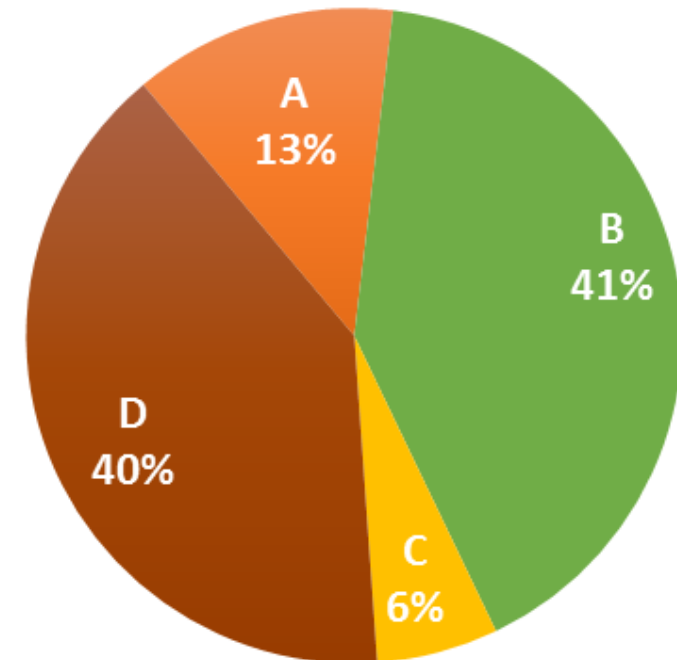
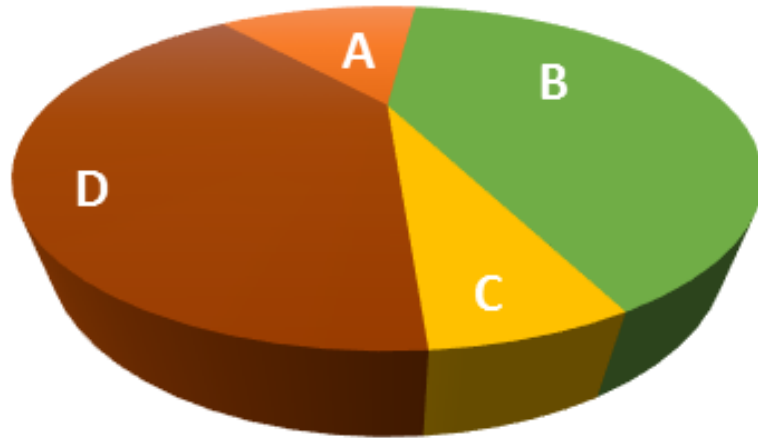


Popisná statistika – Kvalitativní znak

- 3D vs 2D výsečový graf

Jaký je poměr výsečí A a C?

Jaký je poměr výsečí B a D?



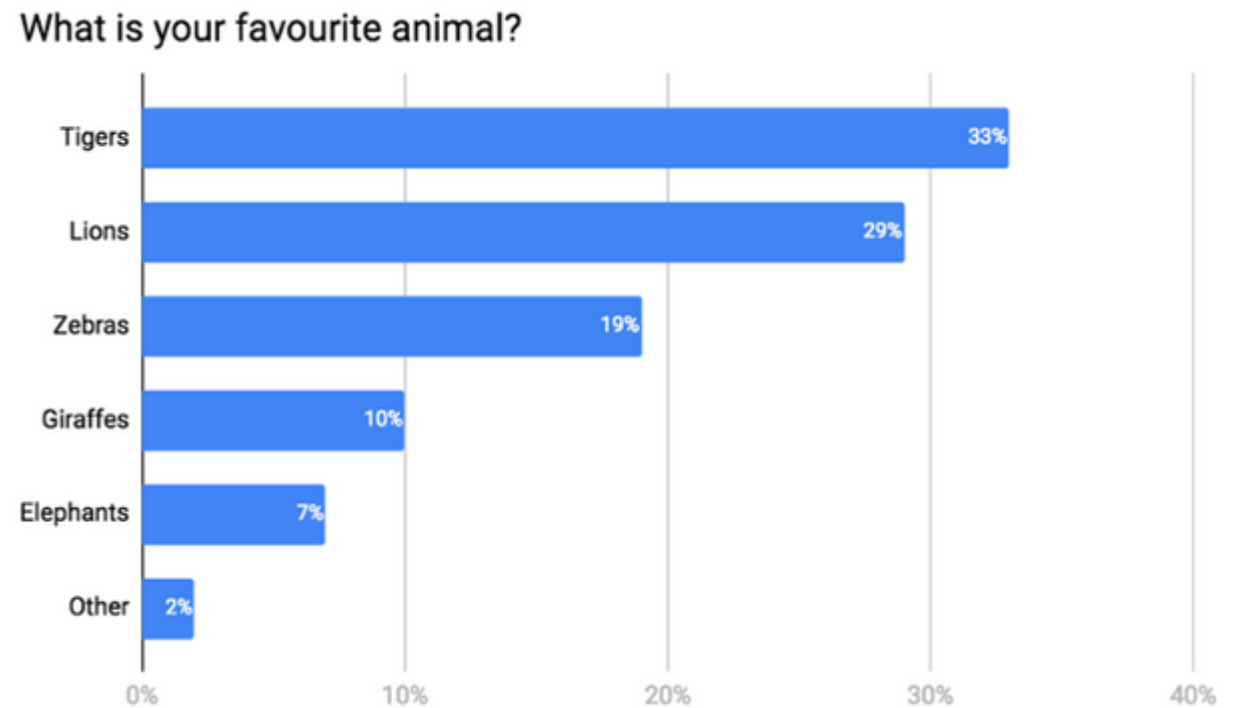
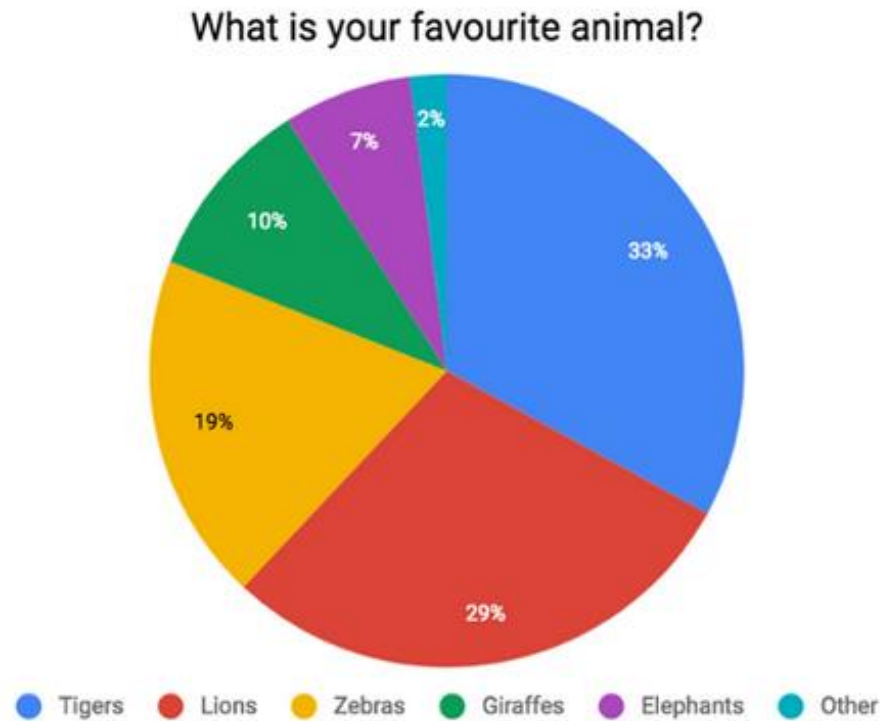
Anketa

Jste pro navýšení hodinové dotace matematiky?



Nezapomínejte, že relativní četnosti byste měli uvádět pouze jako doplněk četností absolutních!

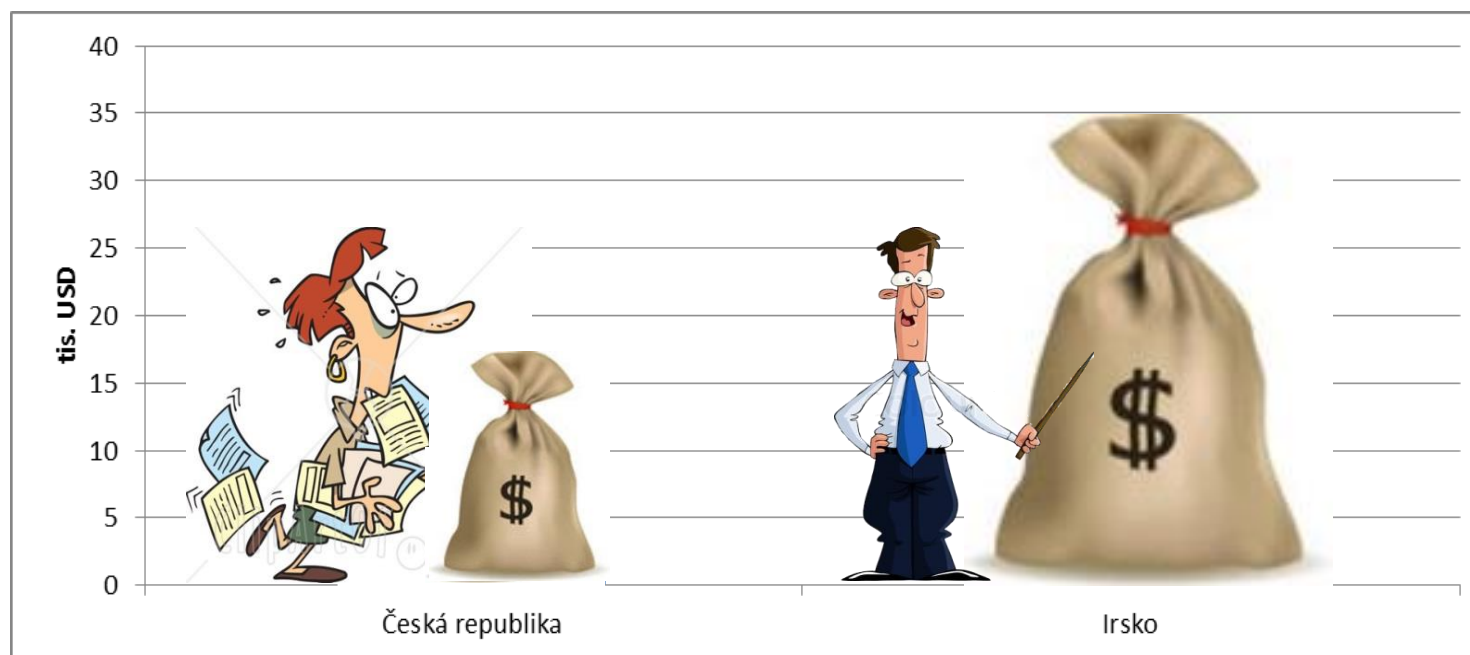
Je výsečový graf tou správnou volbou?



*Are Tigers more popular than Lions? How much more popular are Zebras than Giraffes?
Tigers are more popular than Lions! Zebras are twice as popular as Giraffes!*

Obrázkové grafy – užiteční pomocníci?

Obrázkové grafy



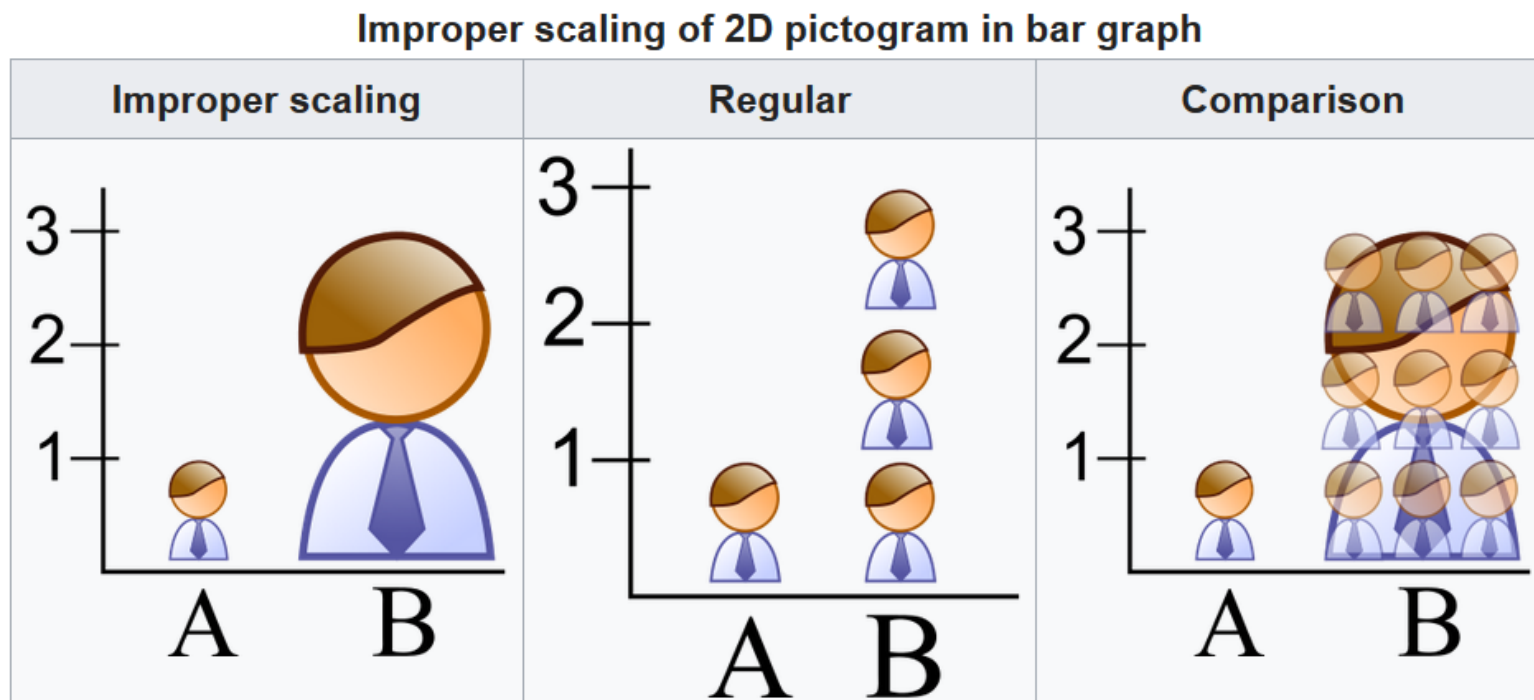
*Srovnání průměrných ročních nástupních platů učitelů středních škol
v ČR (17 244 \$) a Irsku (34 604 \$)*

Obrázkové grafy



*Srovnání průměrných ročních nástupních platů učitelů středních škol
v ČR (17 244 \$) a Irsku (34 604 \$)*

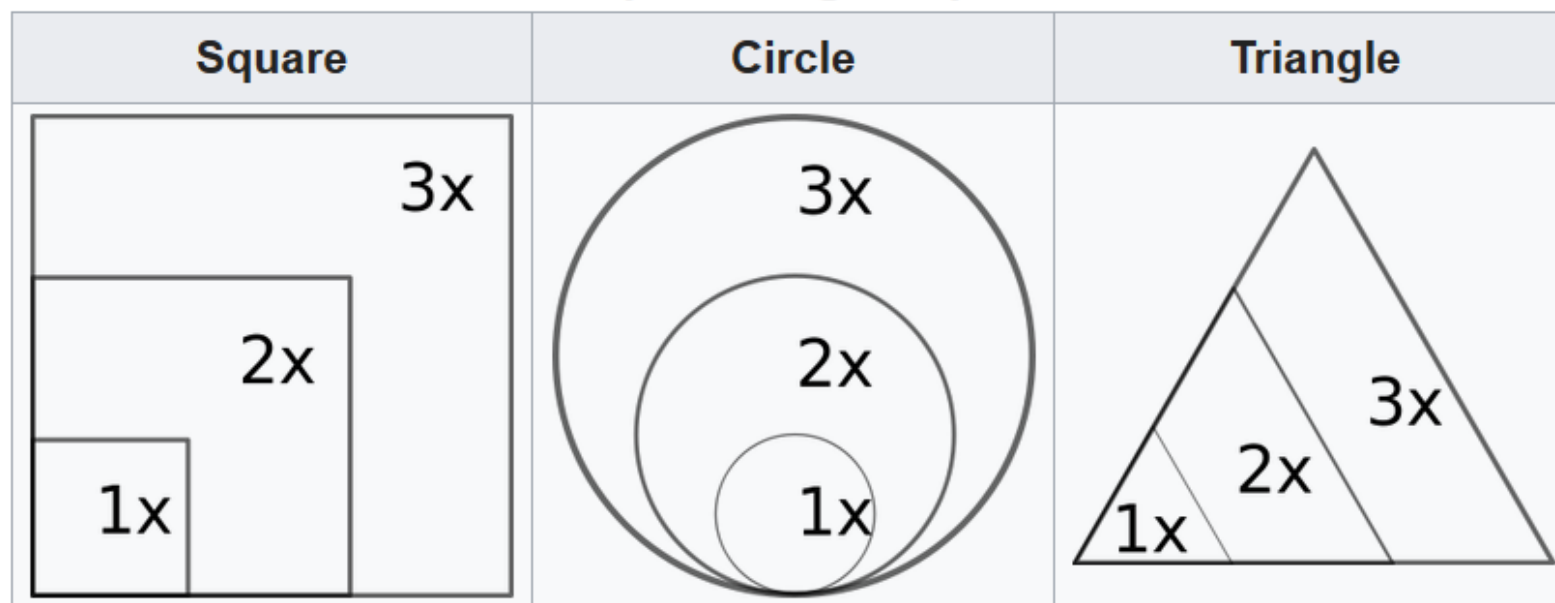
Obrázkové grafy



Zdroj: UTTS, Jessica M. Seeing through statistics. 3rd ed. Belmont, CA: Thomson, Brooks/Cole, c2005. ISBN 0-534-39402-7.

Obrázkové grafy

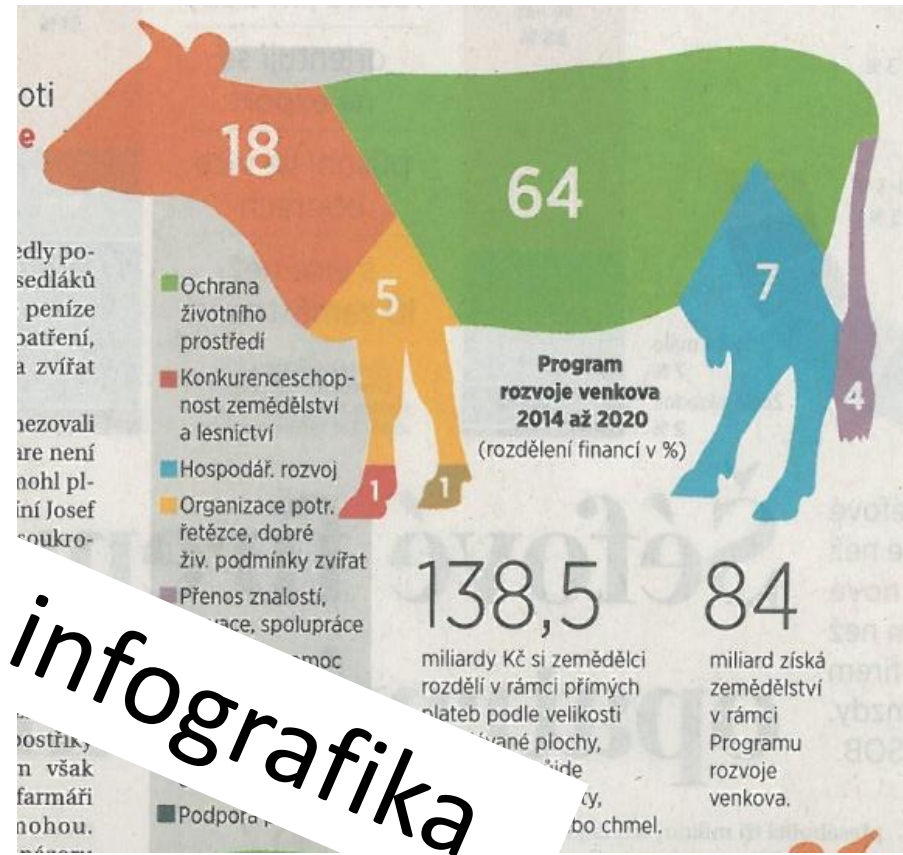
2D shape scaling comparison



Zdroj: UTTS, Jessica M. Seeing through statistics. 3rd ed. Belmont, CA: Thomson, Brooks/Cole, c2005. ISBN 0-534-39402-7.

Pár příkladů z praxe

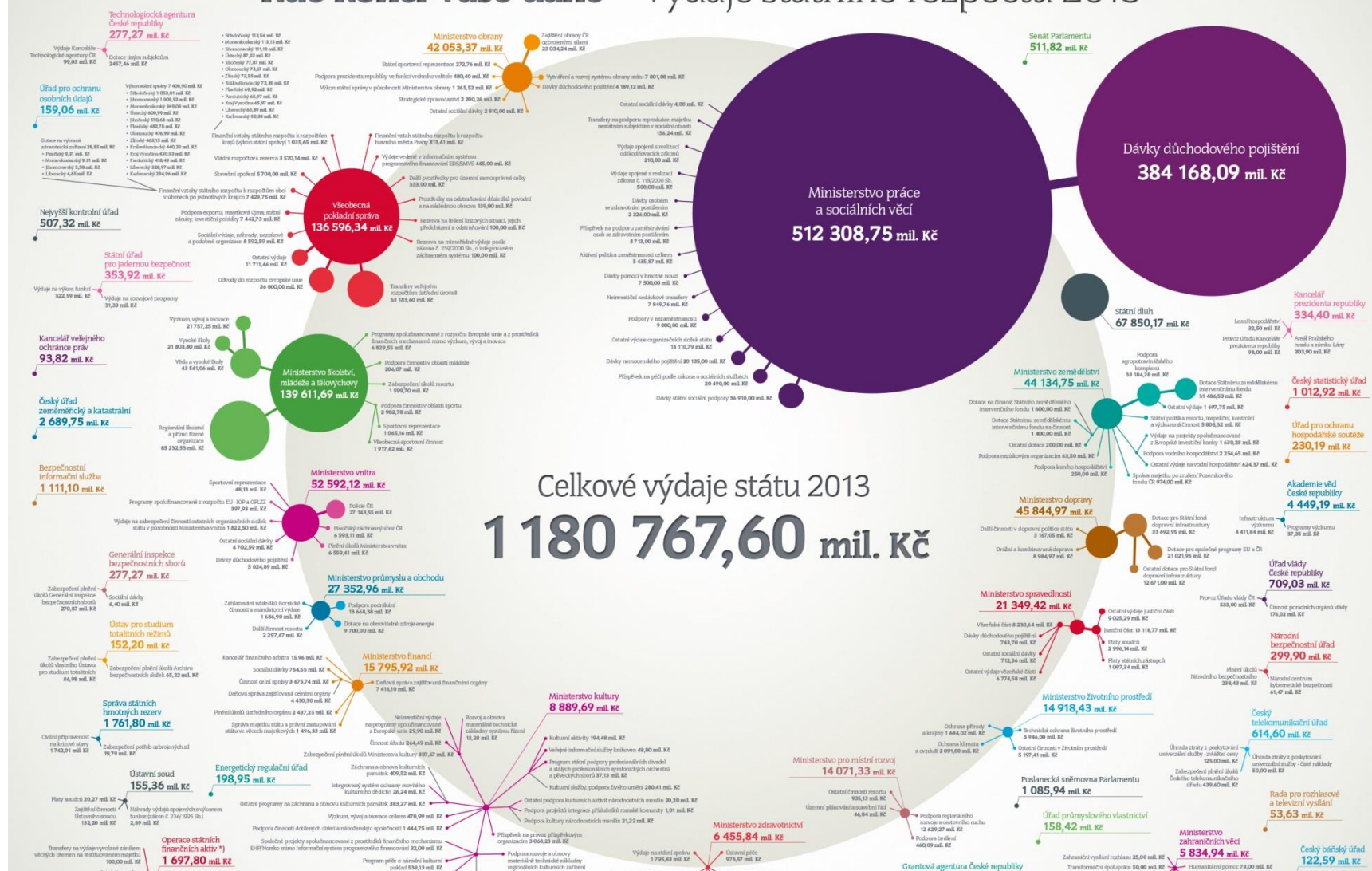
Obrázkové grafy – užiteční pomocníci?



(Zdroj: Mf Dnes, 10. 7. 2014:

Zemědělci si rozdělí miliardy. Krávy a vepři se budou mít lépe.

Kde končí Vaše daně – Výdaje státního rozpočtu 2013



„Úžasná infografika o výdajích státního rozpočtu České republiky v roce 2013“

Zdroj: <http://www.estat.cz/zpravy/informace-k-projektum/kde-konci-vase-dane/>

[illegible]

Zdroj: <http://www.estat.cz/zpravy/informace-k-projektum/kde-konci-vase-dane/>

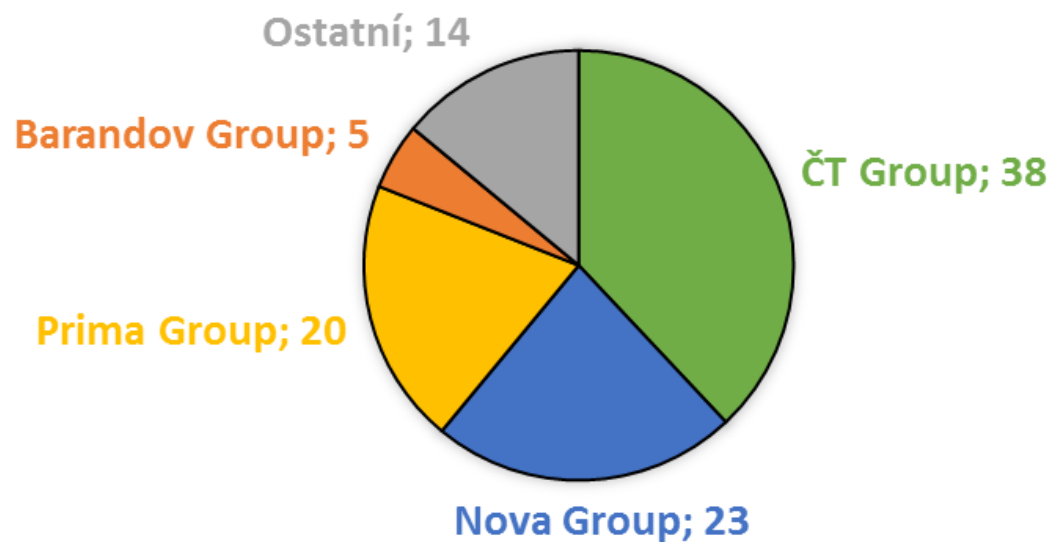
Příklad s klobásou



Příklad s klobásou

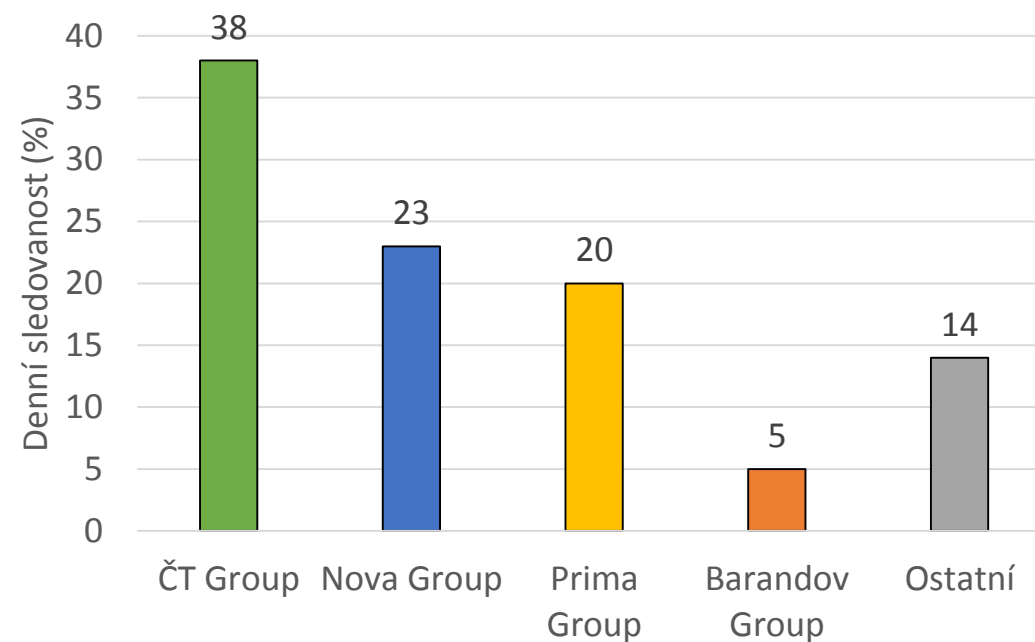
2. místo: klobása z trhů na náměstí Míru (45 Kč)	
Obsah masa:	79,73 %
Obsah tuku:	24,94 %
Obsah bílkovin:	12,4 %
Obsah vody:	55,56 %
Obsah kolagenu:	2,6 %
Sójová bílkovina:	méně než 0,7 %

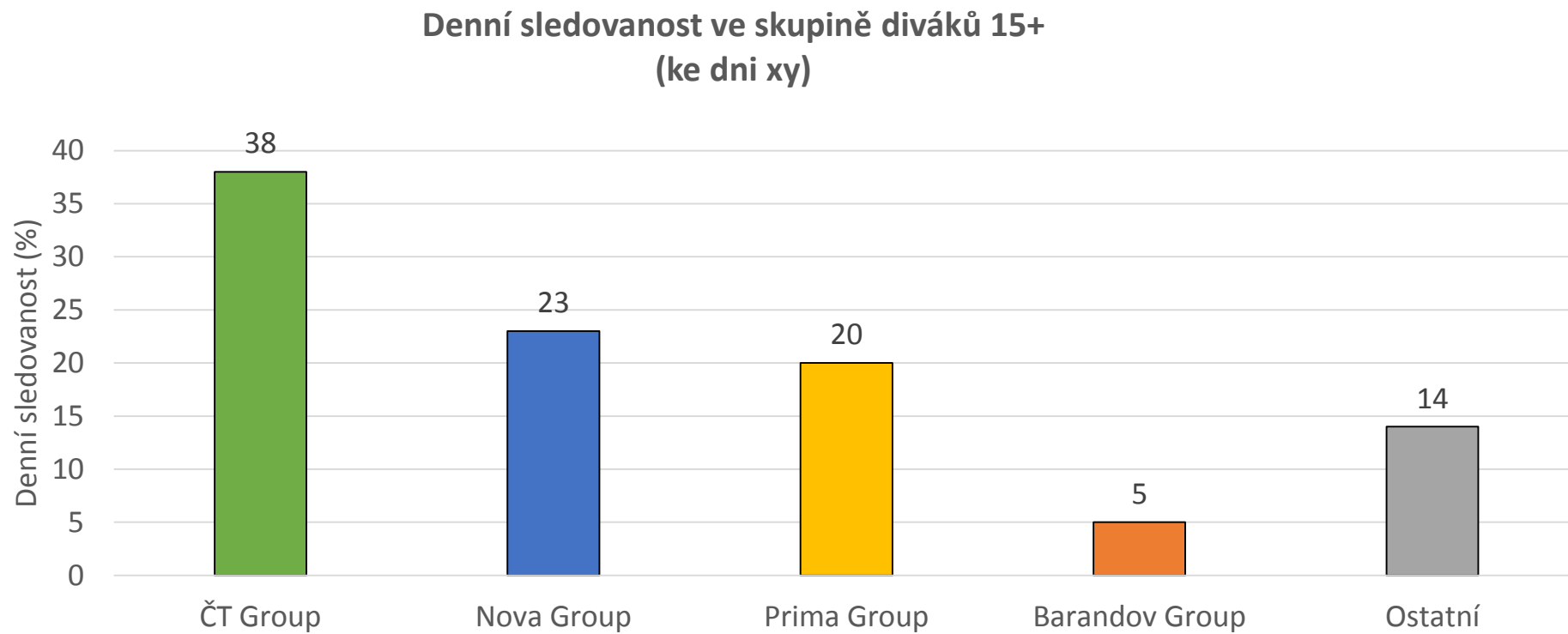
Denní sledovanost (%) ve skupině diváků 15+
(ke dni xy)



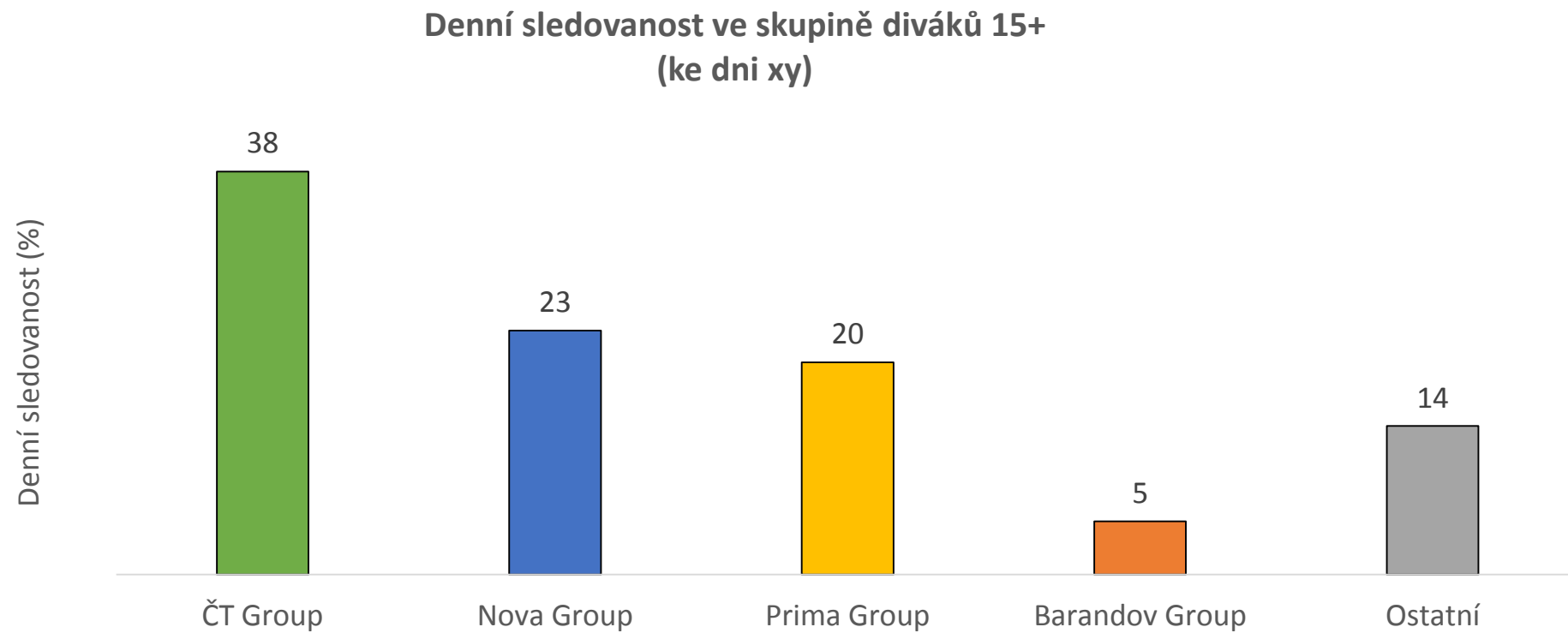
Ilustrativní (smyšlená) data

Denní sledovanost ve skupině diváků 15+
(ke dni xy)

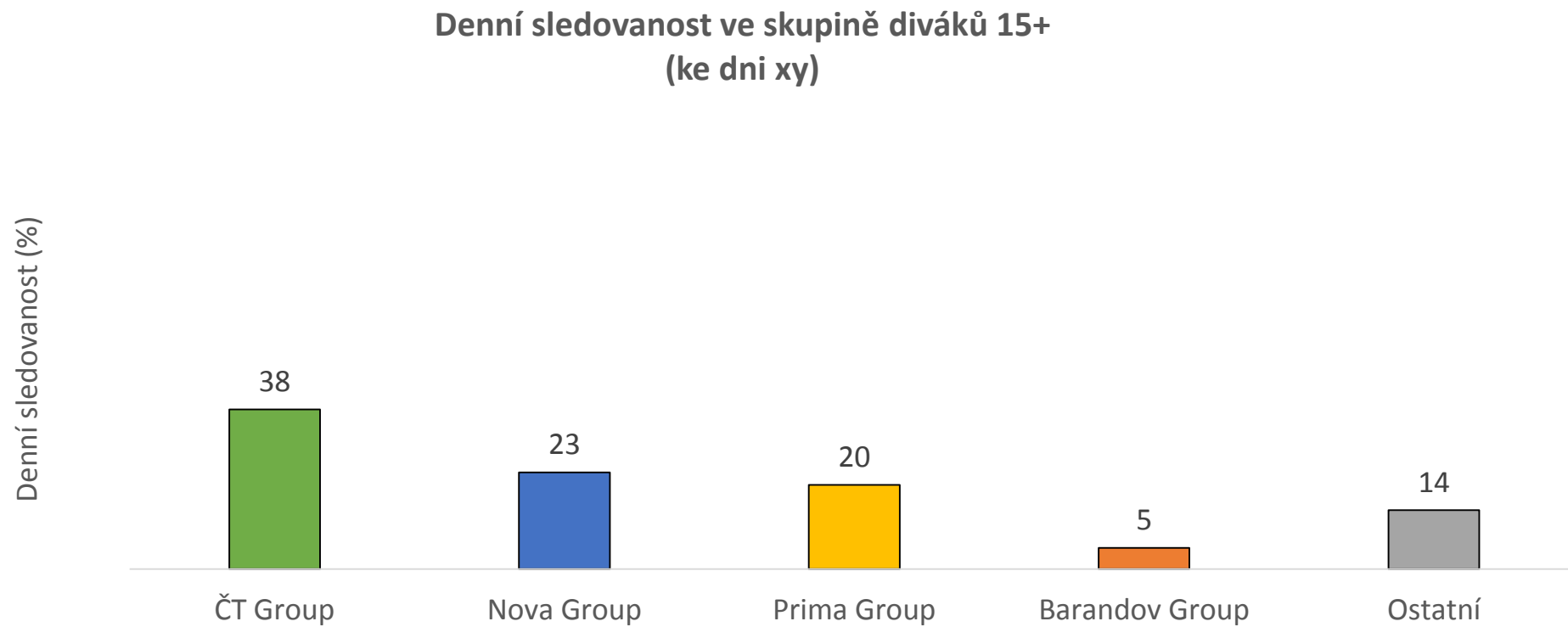




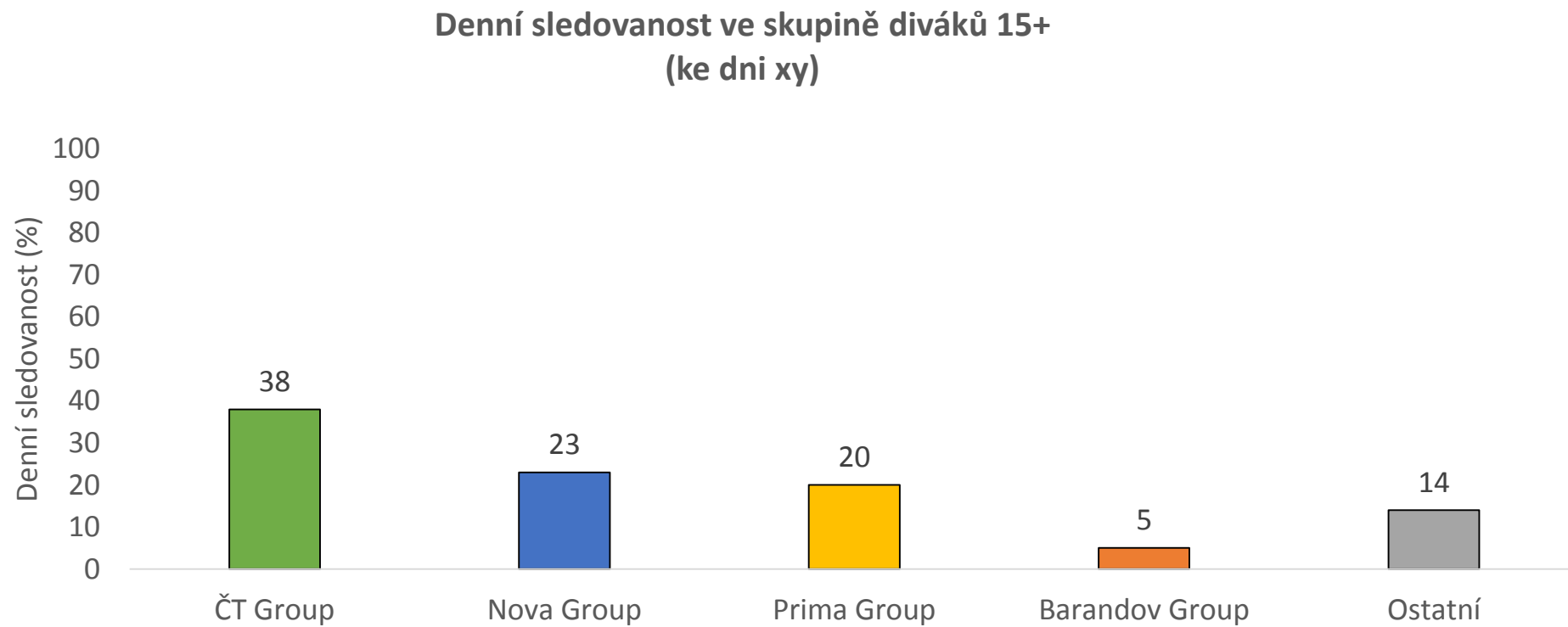
Ilustrativní (smyšlená) data



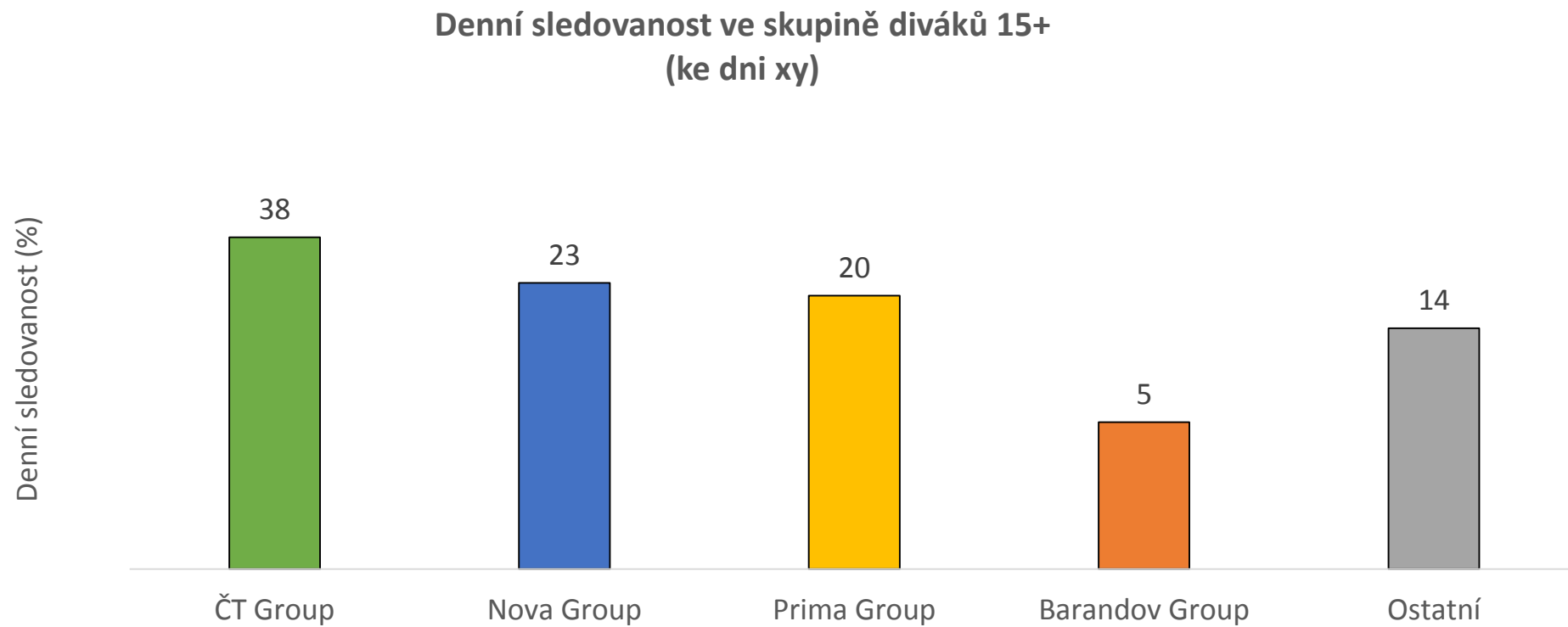
Ilustrativní (smyšlená) data



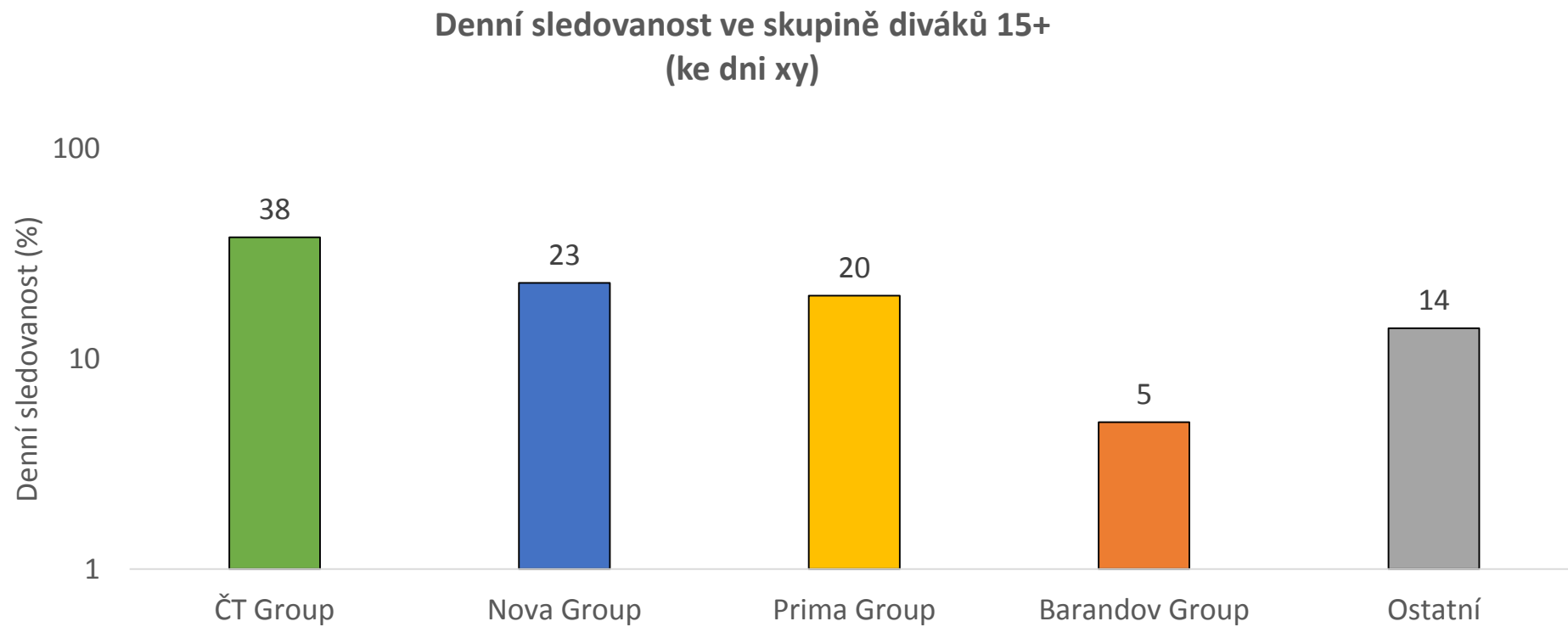
Ilustrativní (smyšlená) data



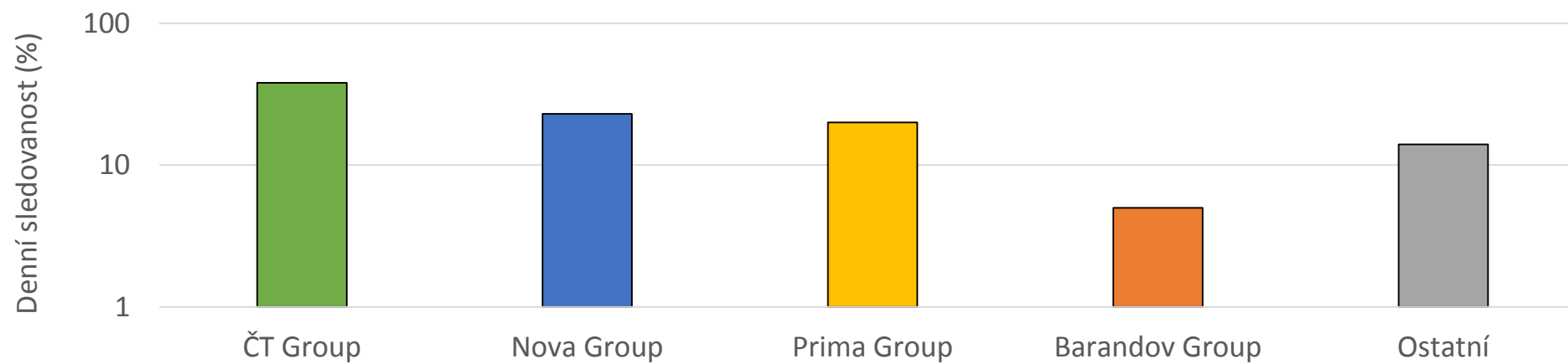
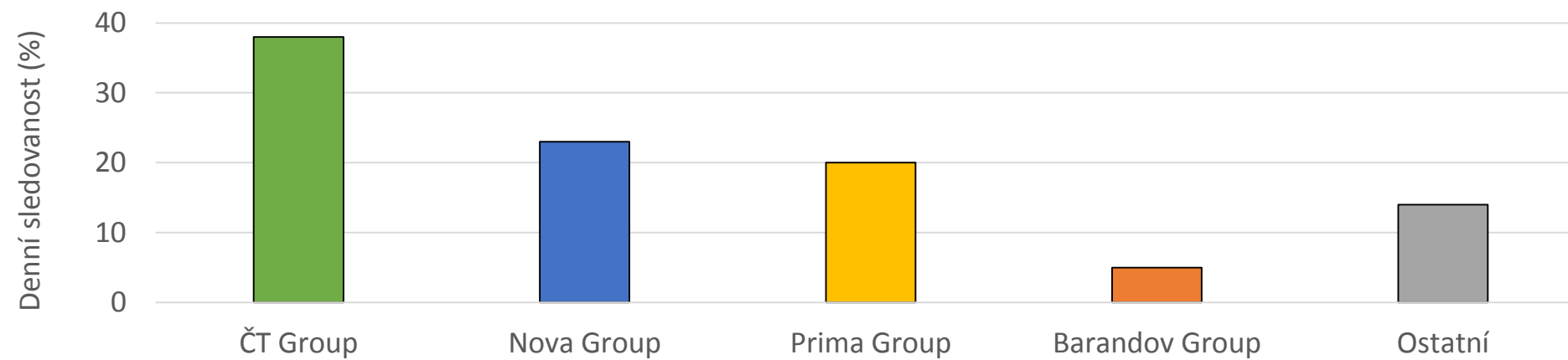
Ilustrativní (smyšlená) data

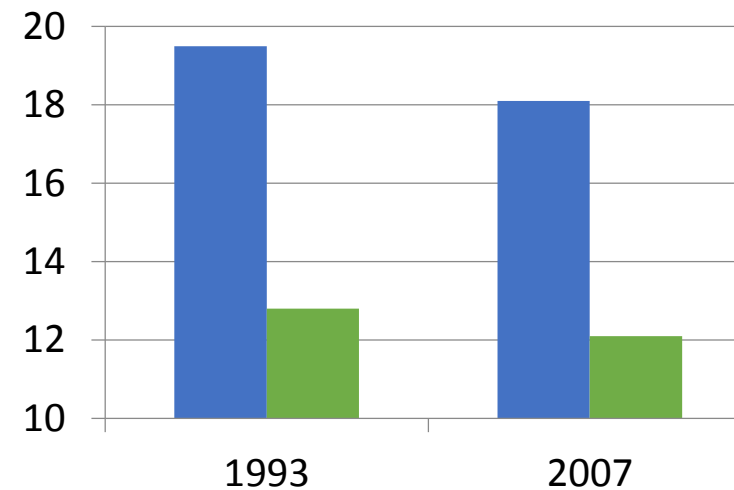
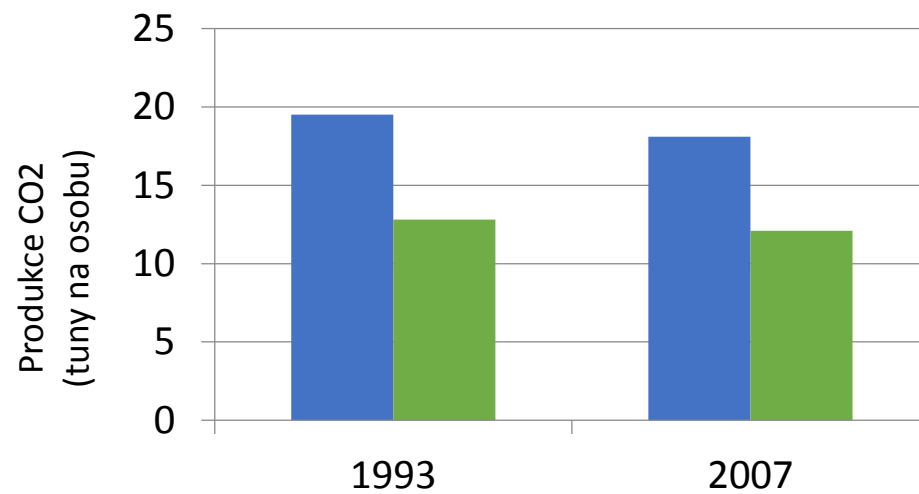


Ilustrativní (smyšlená) data

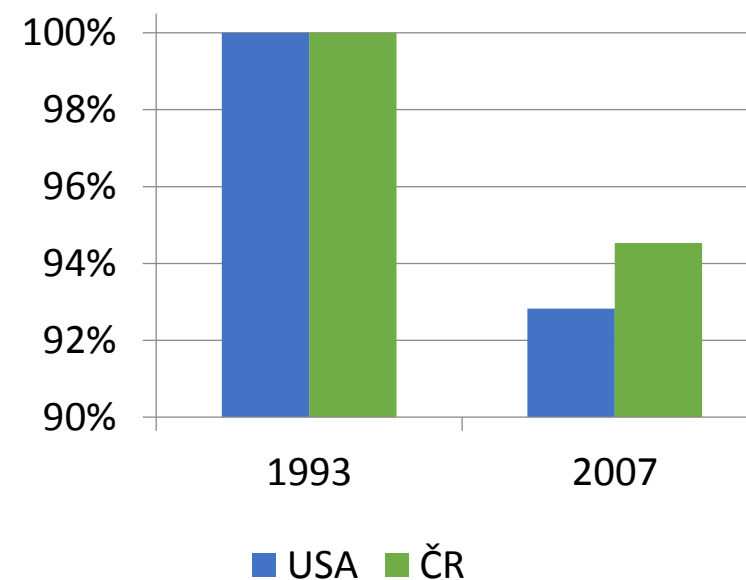
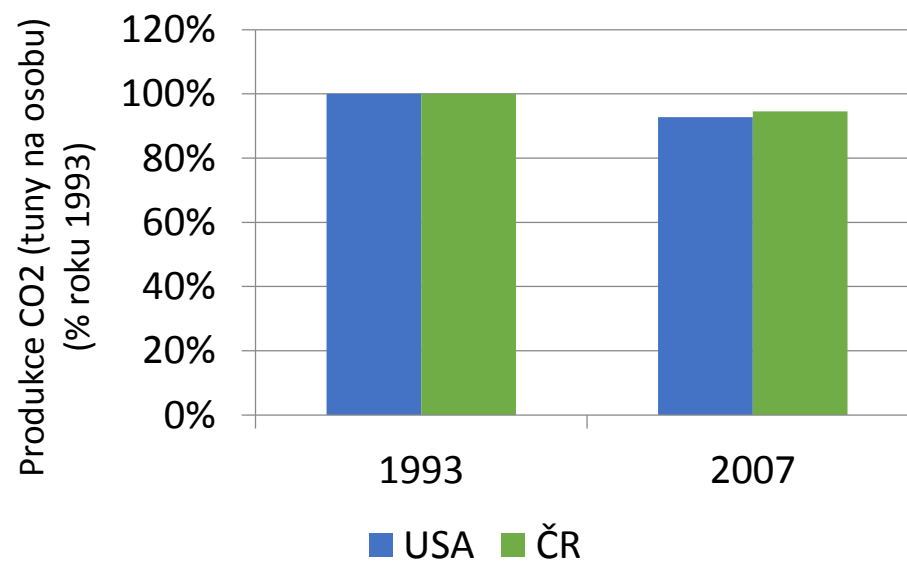


Ilustrativní (smyšlená) data





Který z grafů je „správný“?



**Popisná statistika
aneb
Jak ~~efektivně~~ efektivně popsat a vizualizovat data**

Část 2 / 1

Kvantitativní znak – míry polohy, vizualizace

Popisná statistika - Kvantitativní znak

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídnání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

Popište a vizualizujte hmotnost respondentů.

Popisná statistika - Kvantitativní znak

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídní dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v týdnu	praxe v oboru během studia	24	26

Míry polohy: průměr, kvantily

Míry variability: rozptyl, směrodatná odchylka, variační koeficient

Popisná statistika - Kvantitativní znak

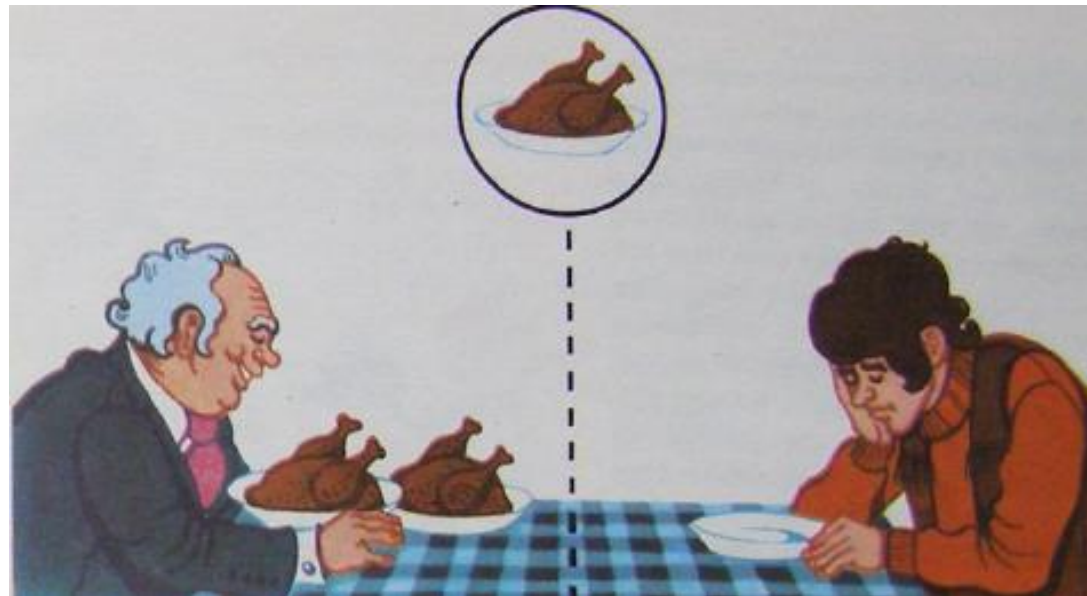
- Míry polohy

- **(Aritmetický) průměr:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Popisná statistika – kvantitativní znak

Míry polohy

- **Aritmetický průměr:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$



Zdroj: Swoboda Helmut, Moderní statistika, 1977

Popisná statistika – kvantitativní znak

Míry polohy

- Aritmetický průměr: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$



Průměrná produkce kuřat (na osobu):
1,0 (denně)

Popisná statistika – kvantitativní znak

Míry polohy

- Aritmetický průměr: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Průměrná mzda přivádí Čechy k zuřivosti: Kdo z vás má 27 tisíc?

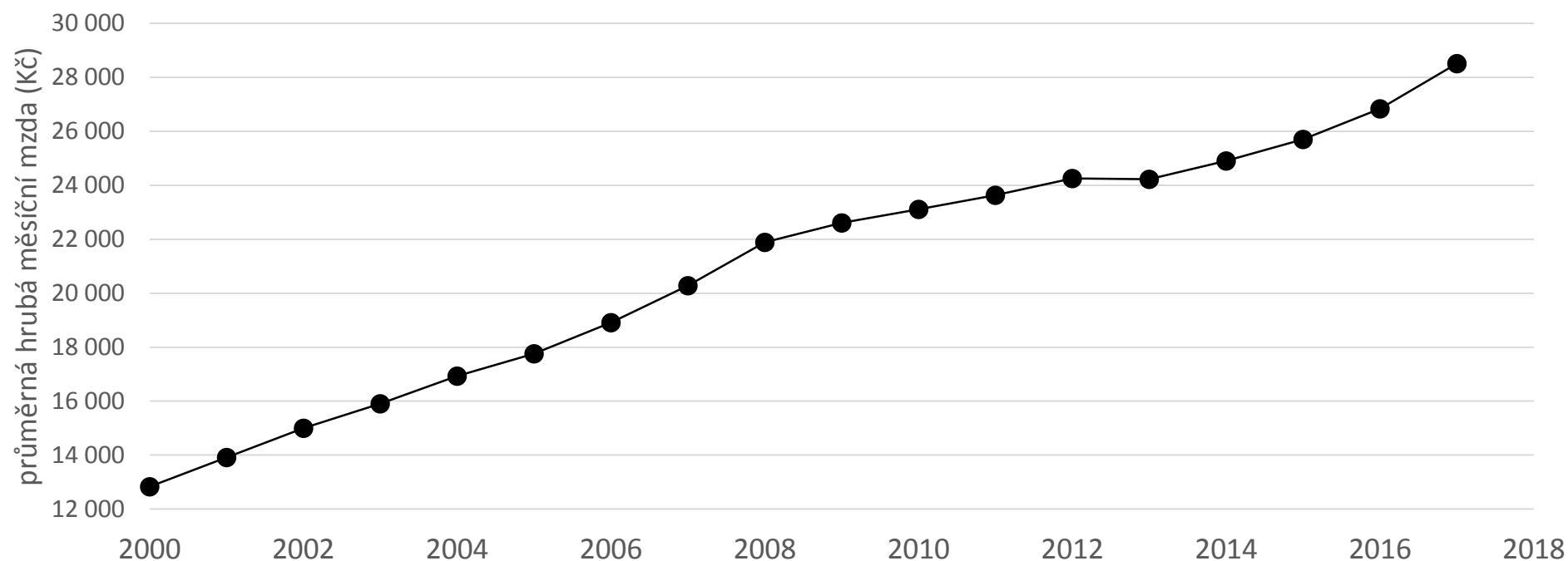


Zdroj: Blesk, 9.4.2013

Popisná statistika – kvantitativní znak

Míry polohy

- Aritmetický průměr: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

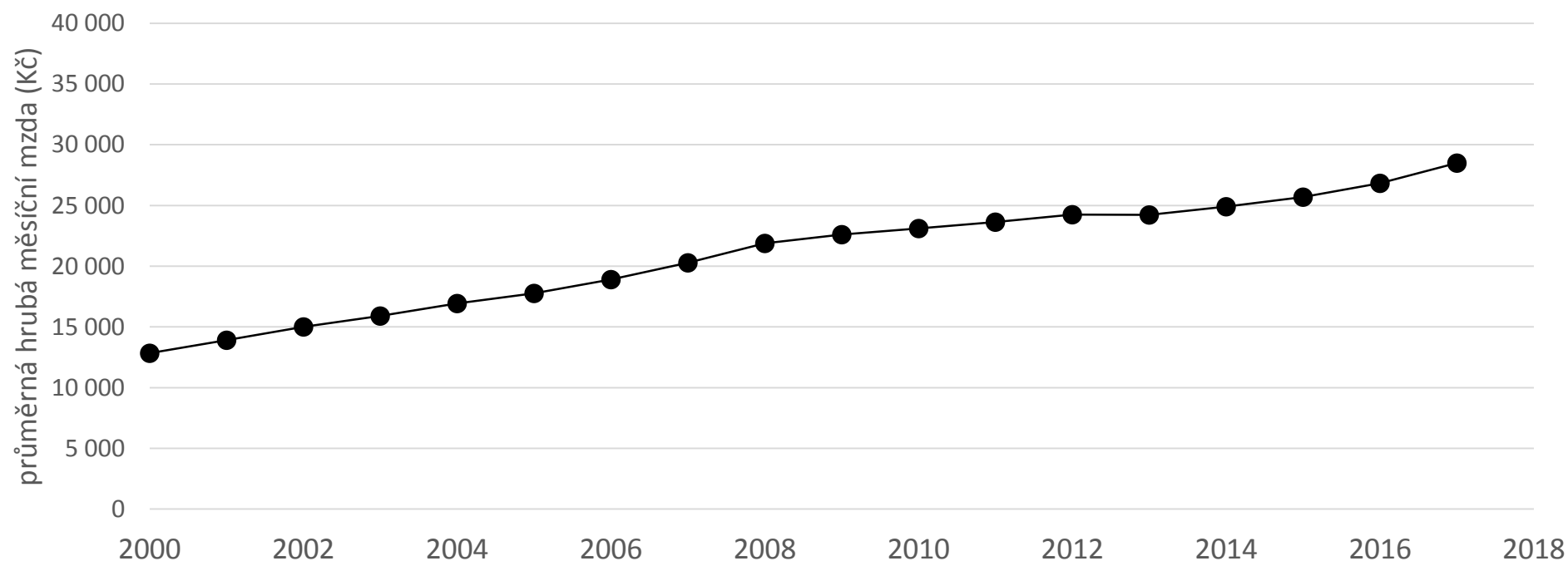


Zdroj dat: [Veřejná databáze ČSÚ](#)

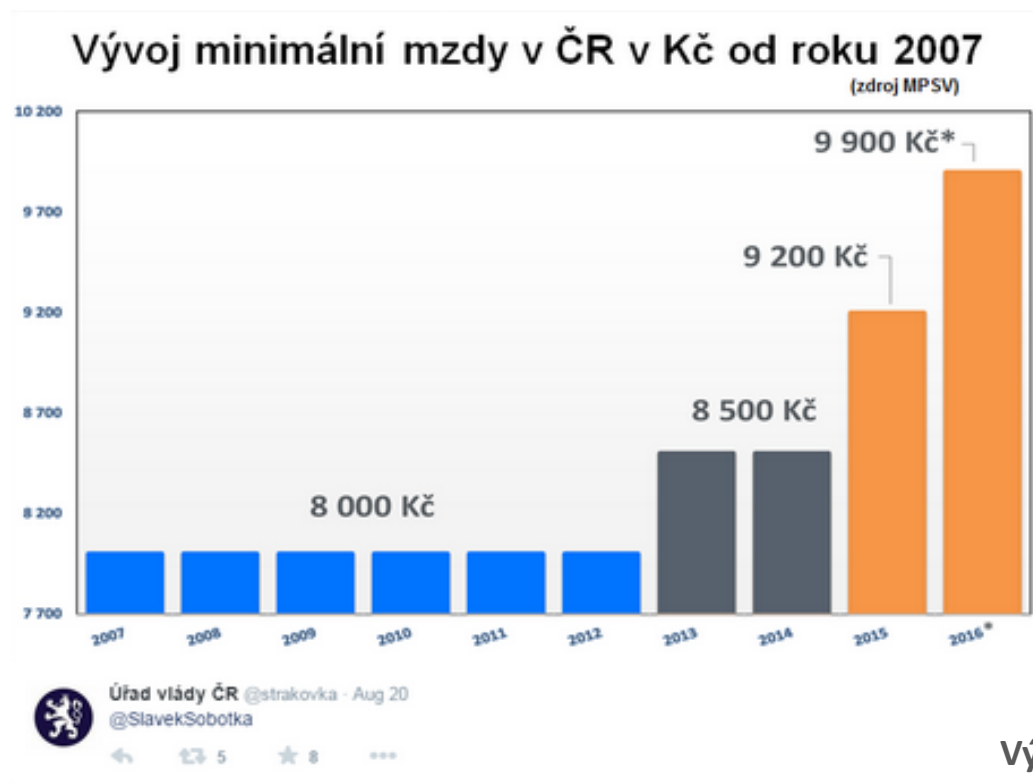
Popisná statistika – kvantitativní znak

Míry polohy

- **Aritmetický průměr:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

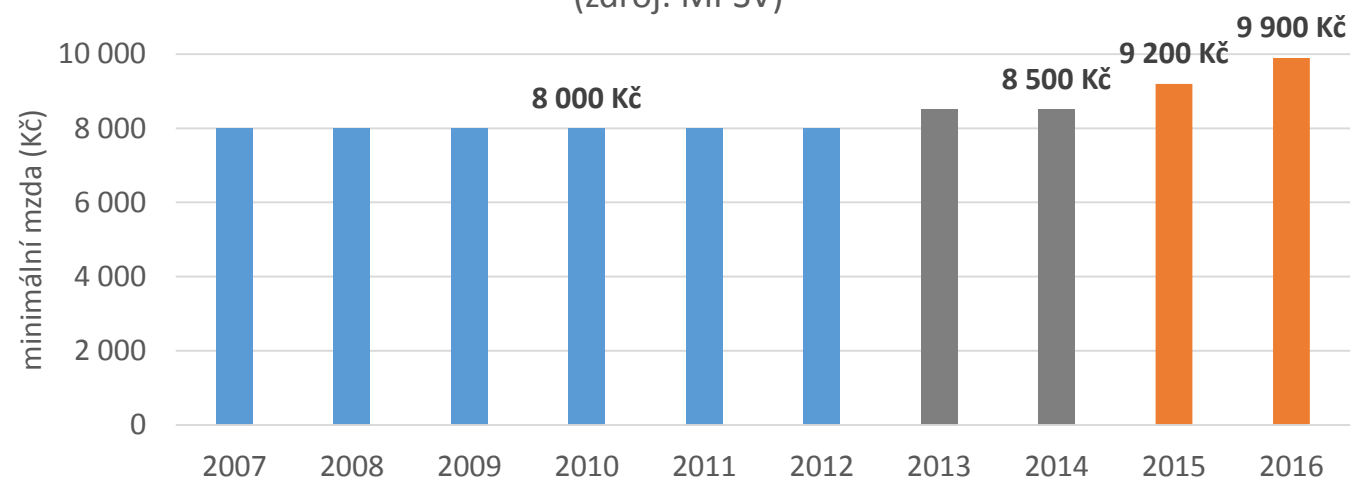


Zdroj dat: [Veřejná databáze ČSÚ](#)



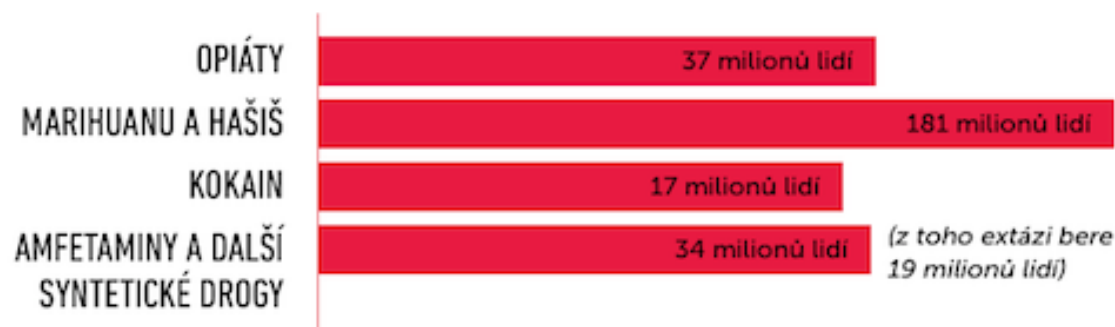
Zdroj: Twitter @strakovka
(20. srpna 2015)

Vývoj minimální mzdy v ČR od roku 2007
(zdroj: MPSV)



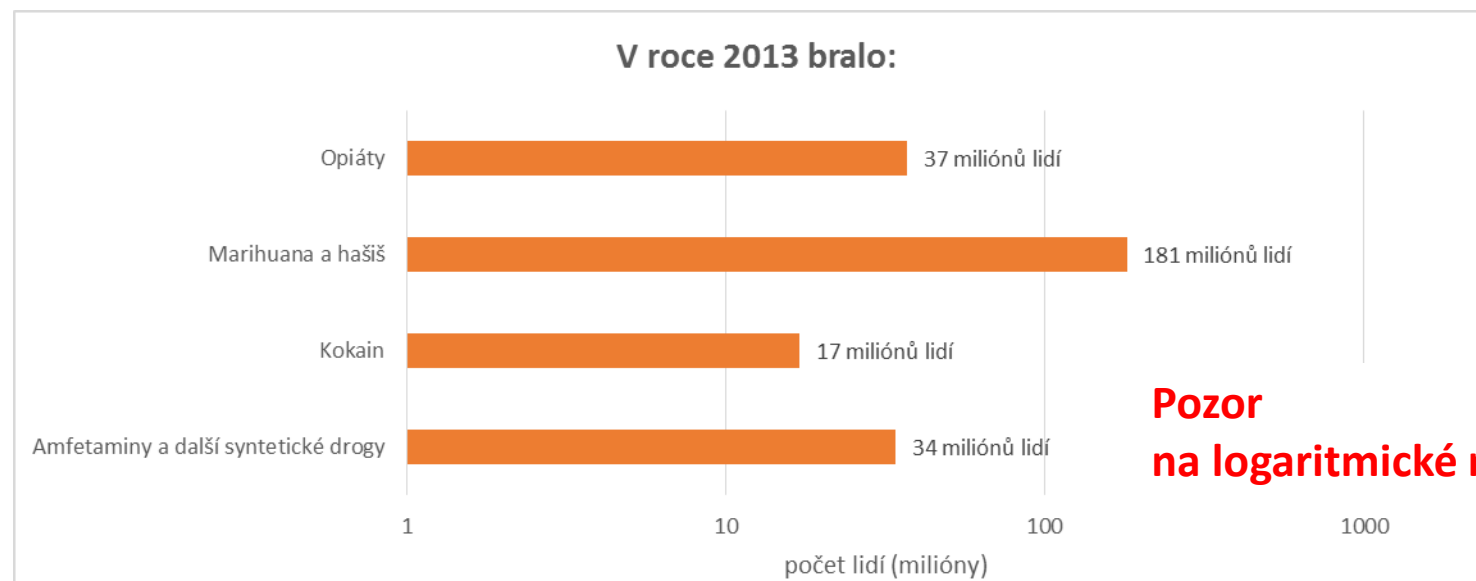
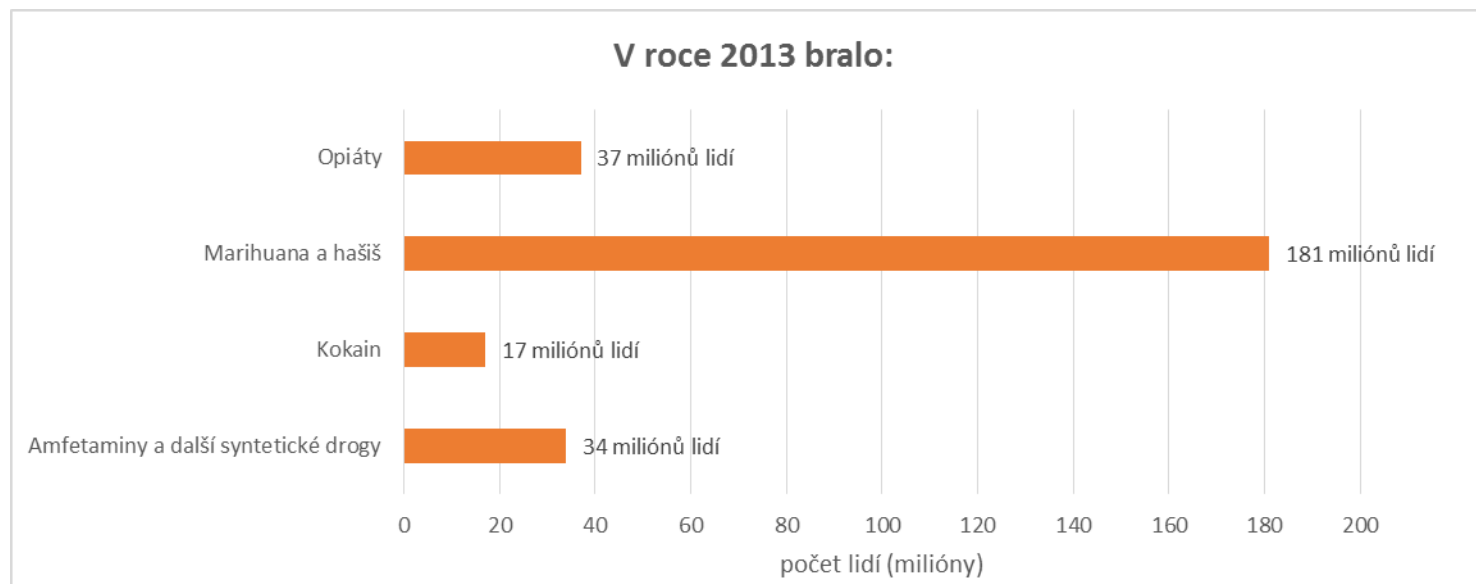
BILANCE ZÁVISLOSTI

V ROCE 2013 BRALO:



- Celkový počet lidí užívajících ilegální drogy se odhaduje za rok 2013 na 246 milionů. To je o 3 miliony více než v roce 2012, ale vzhledem k růstu celkového počtu obyvatel Země se drží podíl konzumentů drog na stabilní úrovni.
- Odhadem asi 27 milionů z těchto uživatelů drog má vážné zdravotní problémy nebo je na drogách závislá. Asi 12 milionů z těchto „problémových“ uživatelů si píchá drogy do žíly. Celkem 1,65 milionu narkomanů, kteří berou drogy injekčně, je HIV-pozitivních.
- Jen každý šestý problémový narkoman má přístup k lékařské péči.
- Ročně umírá na užívání drog 183 000 lidí.

Zdroj: Dotyk, týdeník, 34. číslo, 21. 8. 2015, ISSN: 1805-9465





Bohuslav Sobotka

@SlavekSobotka



Follow

Naše práce na lepších vztazích ČR-Čína má výsledky. Český export do Číny se zvýšil mezi 2013 a 2015 o 8 mld z 37,5 miliardy na 45,5 miliardy

[View translation](#)

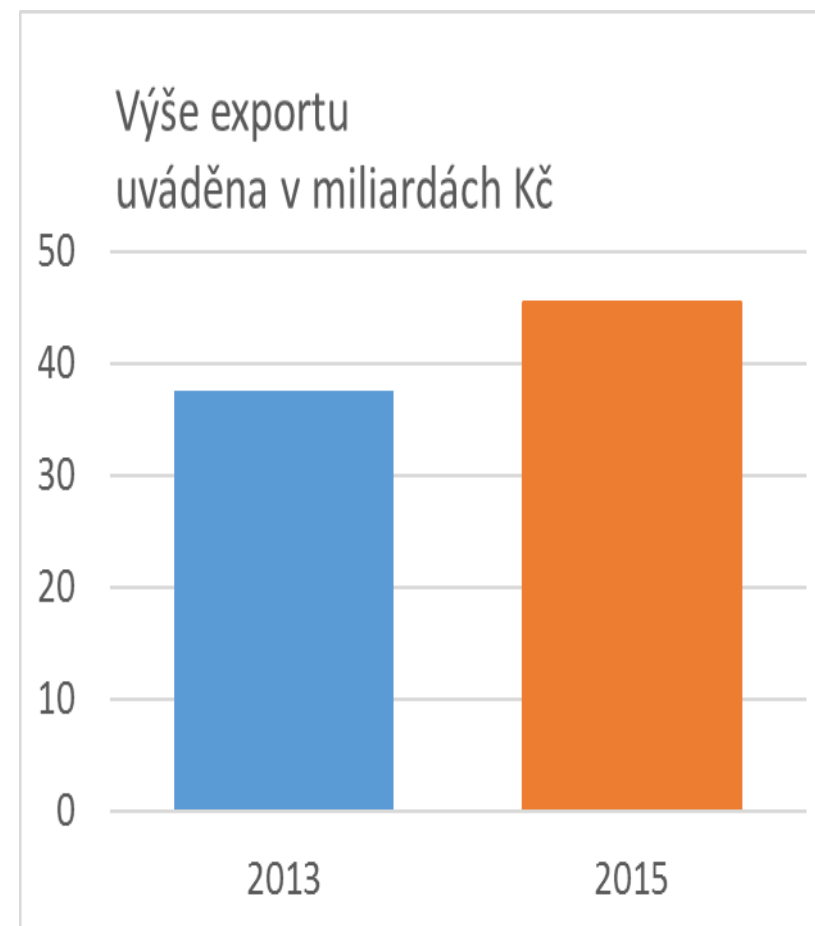


Bohuslav Sobotka
@SlavekSobotka

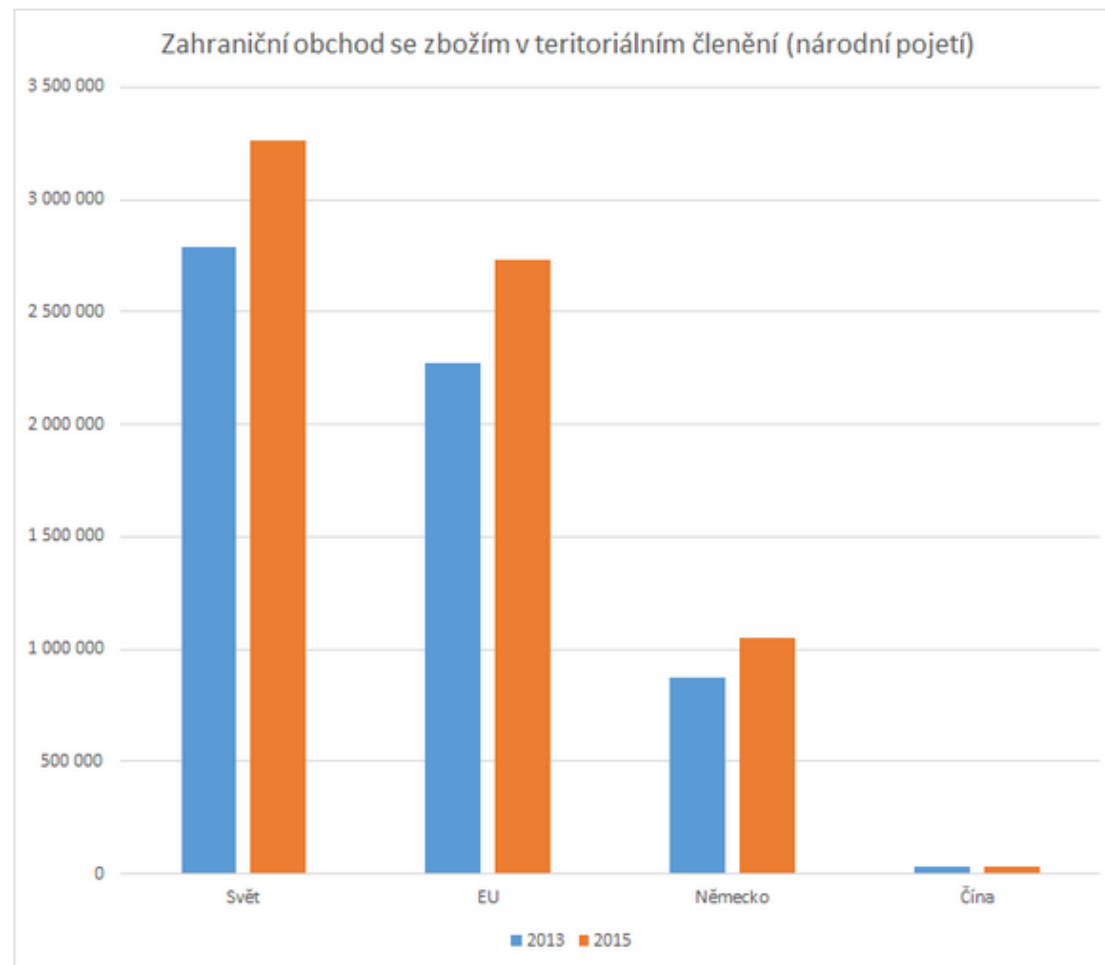


Follow

Naše práce na lepších vztazích ČR-Čína má výsledky. Český export do Číny se zvýšil mezi 2013 a 2015 o 8 mld z 37,5 miliardy na 45,5 miliardy



Český export



Zdroj: <https://www.souki.cz/kouzelne-grafy>

Popisná statistika – kvantitativní znak

Míry polohy

- **Aritmetický průměr:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

POZOR!

- Průměr je číslo, které nemusí patřit do definičního oboru analyzovaného znaku. (např. průměrný počet dětí jedné ženy)

Popisná statistika – kvantitativní znak

Míry polohy

- **Aritmetický průměr:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

V malé vesnici někde v Americe žije 6 lidí, jejichž roční plat je uveden níže.

\$25 000 \$27 000 \$29 000

\$35 000 \$37 000 \$38 000

Určete průměrný plat obyvatel této vesnice.

(\$31 830)

Do vesnice se přistěhoval Bill Gates, jehož roční příjem je \$40 000 000.

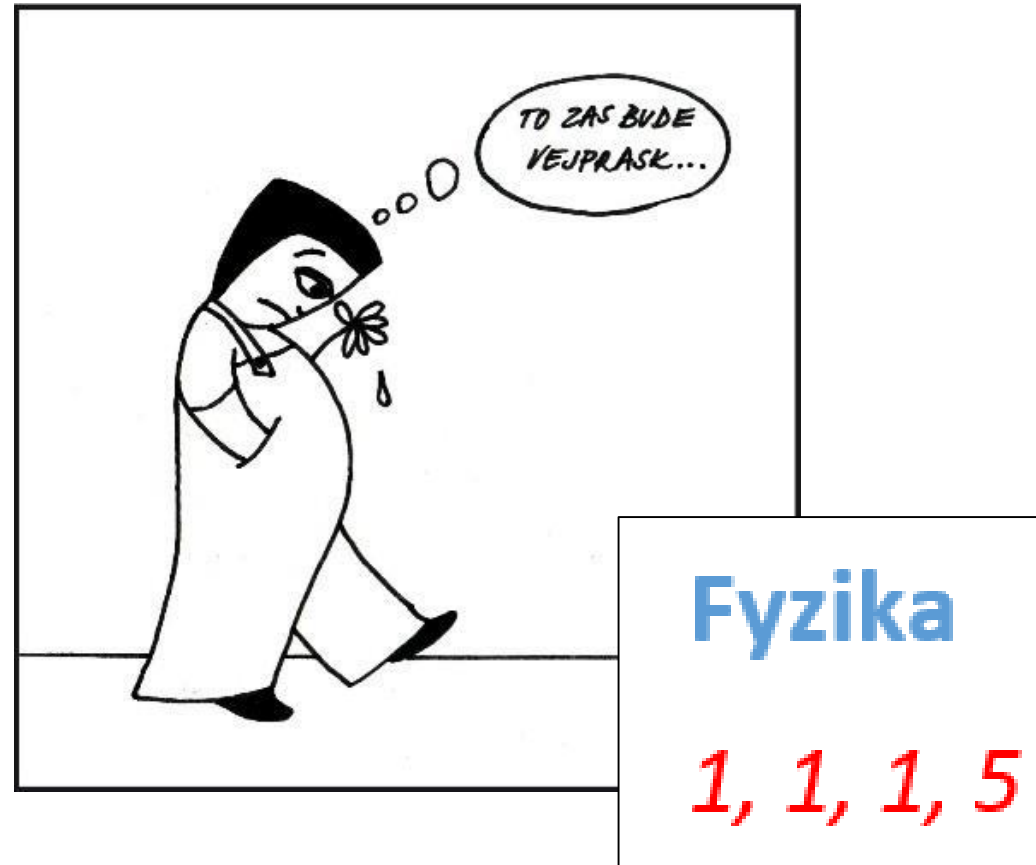
\$25 000 \$27 000 \$29 000

\$35 000 \$37 000 \$38 000 \$40 000 000

Určete průměrný plat obyvatel této vesnice.

(\$5 741 571)

Popisná statistika – kvantitativní znak



Popisná statistika - Kvantitativní znak

- **Míry polohy**

- **(Aritmetický) průměr:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

POZOR!

- Pozor na interpretaci průměru! Nepřisuzujme mu vlastnosti, které nemá!
- Průměr je číslo, které nemusí patřit do definičního oboru analyzovaného znaku. (např. průměrný počet dětí jedné ženy)
- Průměr není rezistentní vůči odlehlým pozorováním!



Popisná statistika – kvantitativní znak

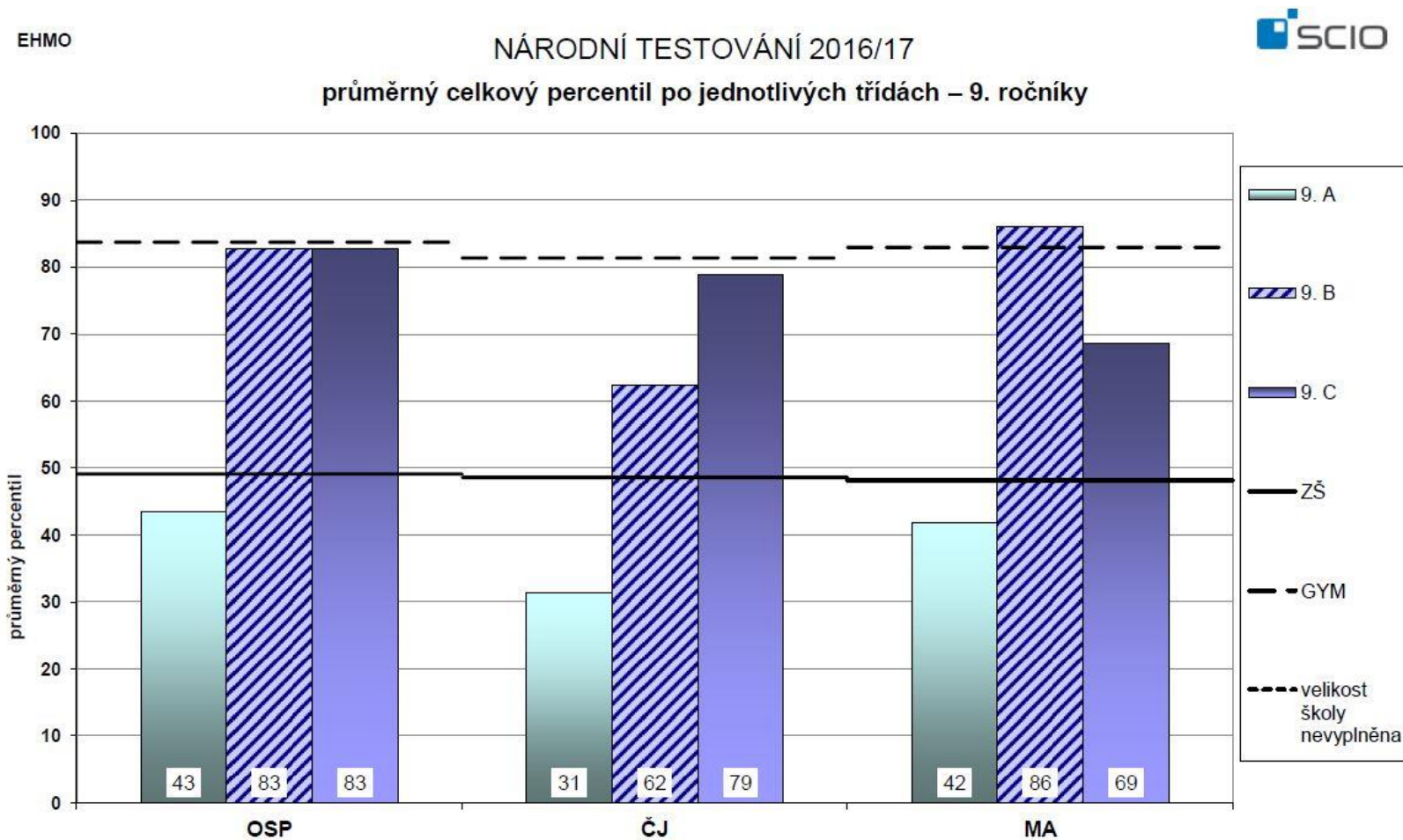
Míry polohy

- **Aritmetický průměr:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
- **Medián *Med*** (50% kvantil – 50 % hodnot je menších nebo rovných mediánů)
- **Dolní kvartil *Q1*** (25% kvantil – 25 % hodnot je menších nebo rovných dolnímu kvartilu)
- **Horní kvartil *Q3*** (75% kvantil – 75 % hodnot je menších nebo rovných hornímu kvartilu)
- **100p% kvantil** – 100p % hodnot je menších nebo rovných 100p% kvantilu

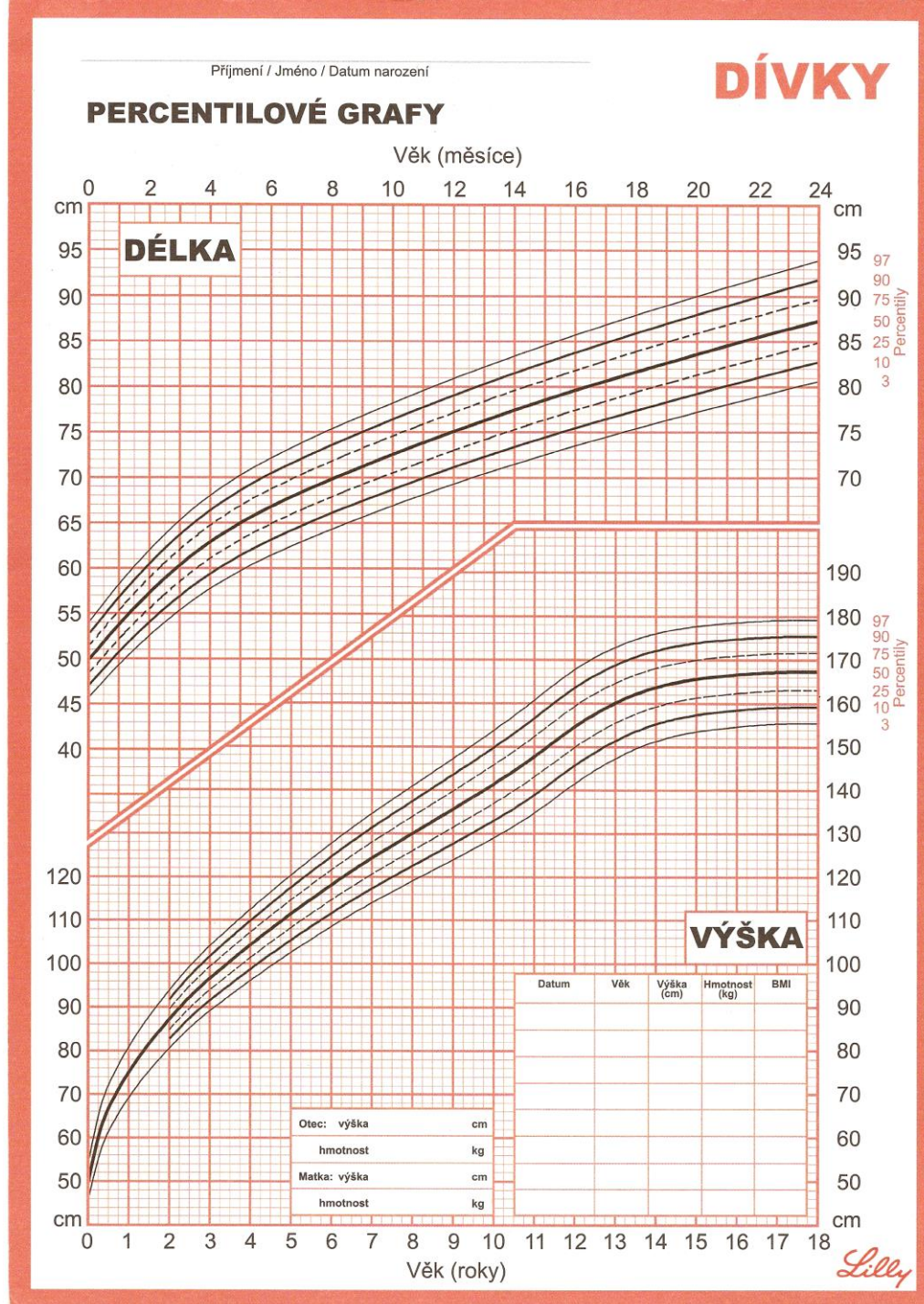
Speciální typy kvantilů:

- Kvartily
- Decily
- Percentily

Kvantily v praxi



Graf znázorňuje průměrné celkové percentily všech tříd 9. ročníku vaší školy. Zároveň je zde pro porovnání vyznačen průměrný percentil všech žáků 9. ročníku ZŠ a odpovídajícího ročníku víceletých gymnázií.



Odlehlá pozorování

- ty hodnoty proměnné, které se mimořádně liší od ostatních hodnot a tím ovlivňují např. vypovídací hodnotu průměru.

Jak postupovat v případě, že v datech identifikujeme odlehlá pozorování?

- V případě, že odlehlost pozorování je způsobena:
 - hrubými chybami, překlepy, prokazatelným selháním lidí či techniky ...
 - důsledky poruch, chybného měření, technologických chyb ...

tzn., známe-li příčinu odlehlosti a předpokládáme-li, že již nenastane, jsme oprávněni tato pozorování vyloučit z dalšího zpracování.

- V ostatních případech je nutno zvážit, zda se vyloučením odlehlých pozorování nepřipravíme o důležité informace o jevech vyskytujících se s nízkou četností.

Jak identifikovat odlehlá pozorování?

Metoda vnitřních hradeb

$$[x_i < x_{0,25} - 1,5 \cdot IQR] \vee [x_i > x_{0,75} + 1,5 \cdot IQR] \Rightarrow x_i \text{ je odlehlé pozorování}$$

dolní mez vnitřních hradeb

horní mez vnitřních hradeb

Poznámka: $IQR = x_{0,75} - x_{0,25}$ (interkvartilové rozpětí)

Jak identifikovat extrémní pozorování?

Metoda vnějších hradeb

$$[x_i < x_{0,25} - 3 \cdot IQR] \vee [x_i > x_{0,75} + 3 \cdot IQR] \Rightarrow x_i \text{ je extrémní pozorování}$$

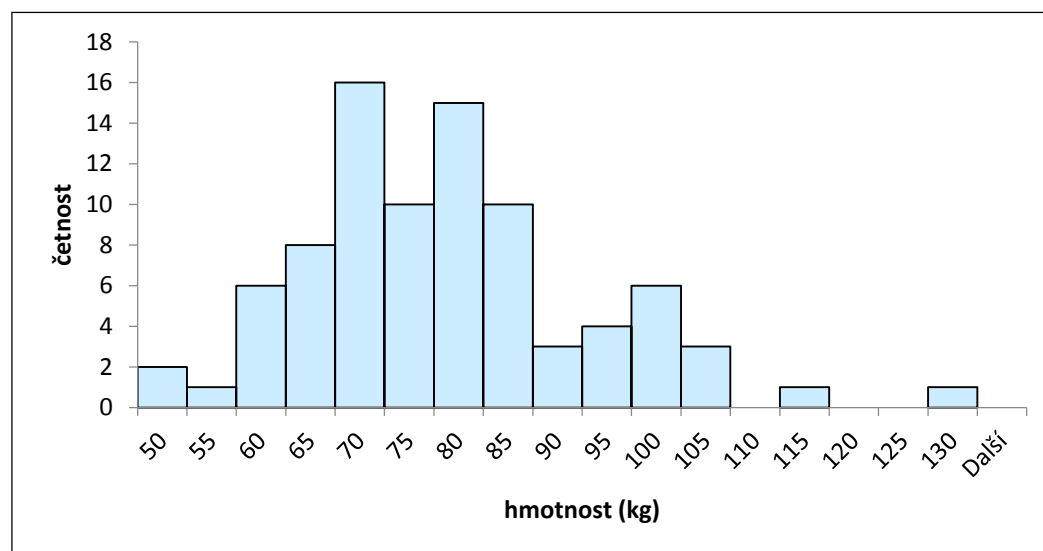
dolní mez vnějších hradeb

horní mez vnějších hradeb

Poznámka: $IQR = x_{0,75} - x_{0,25}$ (interkvartilové rozpětí)

Popisná statistika - Kvantitativní znak

- Vizualizace

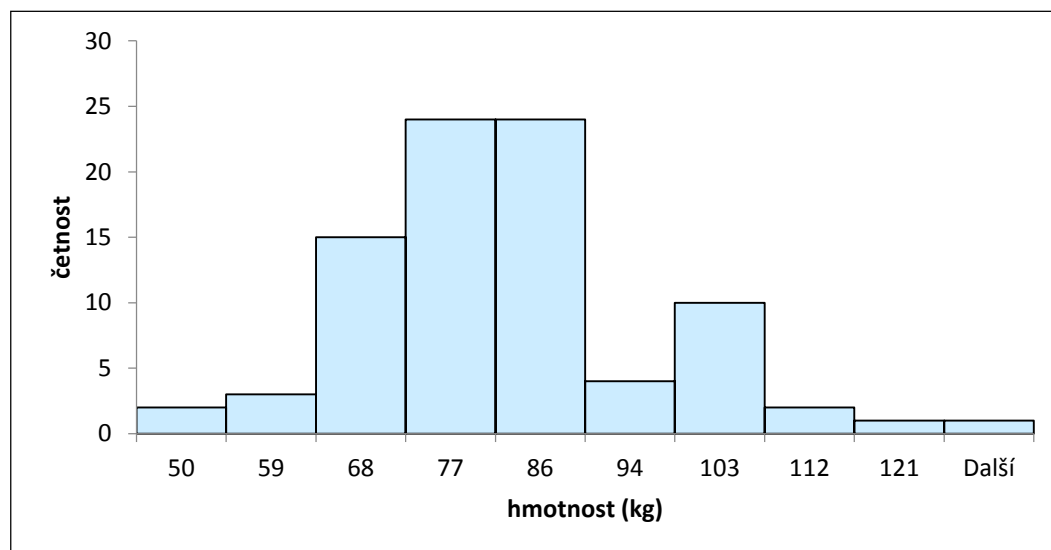


Histogram

Míry polohy	Hmotnost (kg)
Minimum	50
Dolní kvartil	68
Průměr	78
Medián	76
Horní kvartil	85
Maximum	130

Popisná statistika - Kvantitativní znak

- Vizualizace



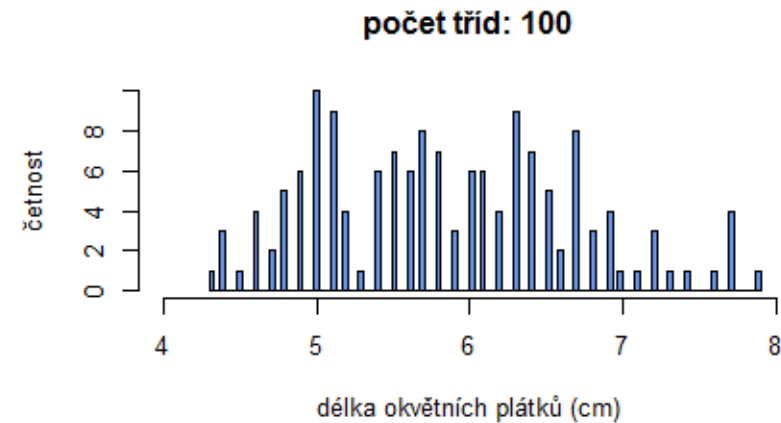
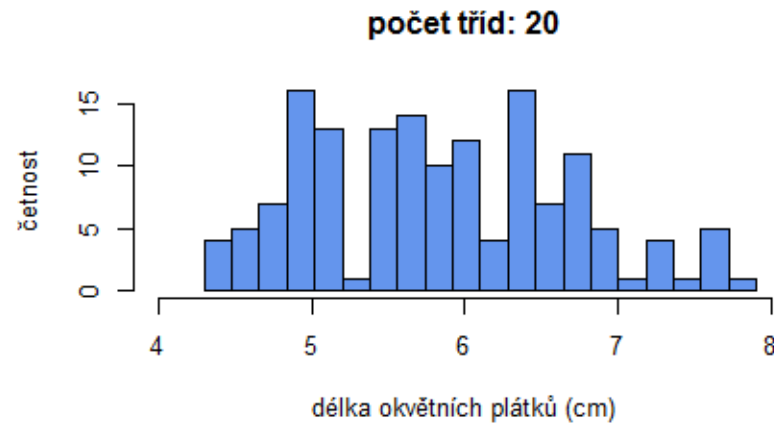
Histogram

Míry polohy	Hmotnost (kg)
Minimum	50
Dolní kvartil	68
Průměr	78
Medián	76
Horní kvartil	85
Maximum	130

Tvar histogramů závisí na počtu tříd (sloupečků)!



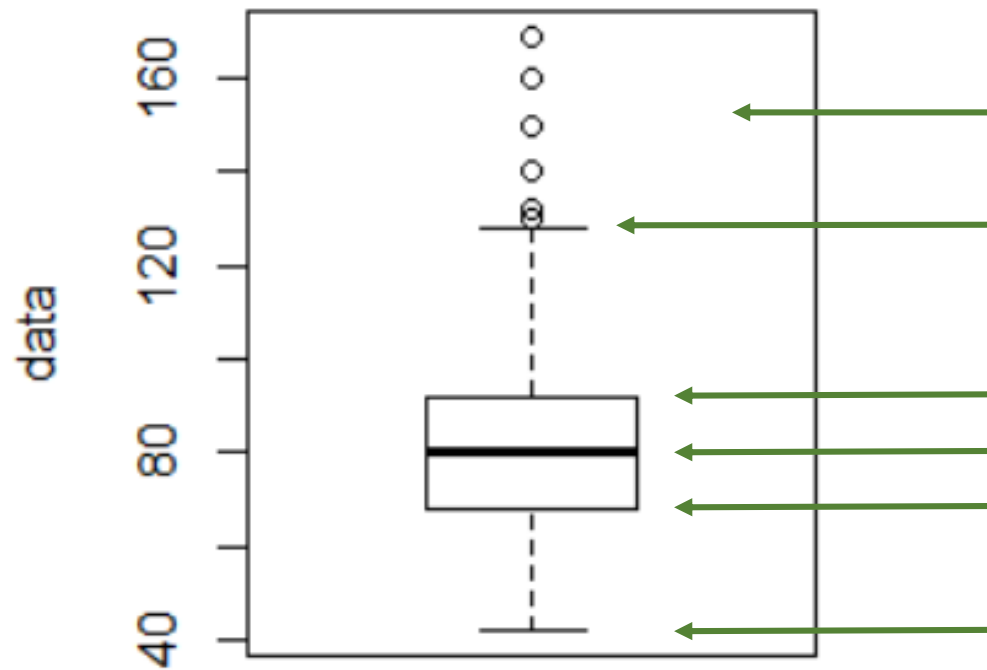
Popisná statistika - Kvantitativní znak



Tvar histogramů závisí na počtu tříd (sloupečků)!



Popisná statistika - Kvantitativní znak



odlehlá pozorování, tj. defaultně: data ležící vně vnitřních hradeb

není definováno jednoznačně, v R je to defaultně nastaveno jako:

horní mez vnitřních hradeb

`max(data)[data < horní mez vnitřních hradeb]`

horní kvartil

medián

dolní kvartil

není definováno jednoznačně, v R je to defaultně nastaveno jako:

dolní mez vnitřních hradeb

`min(data)[data > dolní mez vnitřních hradeb]`

Krabicový graf (angl. Box plot)

Popisná statistika aneb Jak ~~efektivně~~ efektivně popsat a vizualizovat data

Část 2 / 2

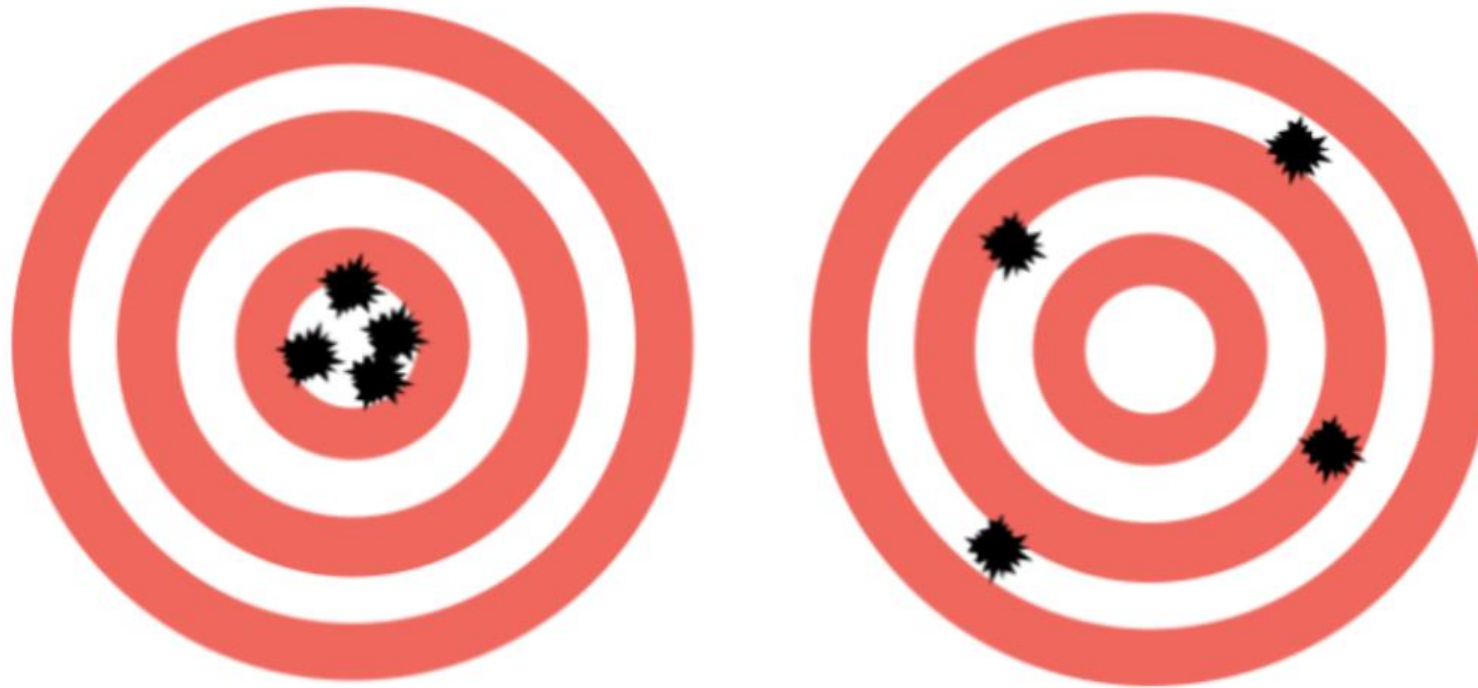
Kvantitativní znak – míry polohy, vizualizace

(Předpokládám, že to nestihneme ☹, ale nadšenci mají podklady pro samostudium ☺)

Popisná statistika – kvantitativní znak

Míry variability

K čemu nám jsou dobré?



Zdroj: <http://blackswanfarming.com/>

Popisná statistika – kvantitativní znak

Míry variability

- **Výběrový rozptyl:** $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
 - POZOR! – Jednotka rozptylu je kvadrátem jednotky analyzovaného znaku.
- **Výběrová směrodatná odchylka:** $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$
 - Neumožňuje srovnání variability znaků s různými jednotkami.
- **Variační koeficient:** $V = \frac{s}{|\bar{x}|} \cdot 100 (\%)$
 - Čím nižší var. koeficient, tím homogennější soubor.
 - $V > 50 \%$ značí silně rozptýlený soubor. (empirické doporučení)
- **Rozpětí:** $R = \max - \min$
- **Interkvartilové rozpětí:** $IQR = \tilde{x}_{0,75} - \tilde{x}_{0,25}$

Proč se pro směrodatnou odchylku někdy používá symbol s a jindy symbol σ ?

Míry variability

- **Výběrová směrodatná odchylka:** $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$
- **Populační směrodatná odchylka:** $\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$
 - N ... rozsah populace, tj. pro výpočet musíme mít k dispozici všechna data ze základního souboru (populace), tj. musíme provést úplné šetření.
 - Lze ukázat, že nejlepším odhadem populační směrodatné odchylky je výběrová směrodatná odchylka:

$$\hat{\sigma} \cong s$$

Jakou představu o variabilitě dat nám dává směrodatná odchylka?

Obecně platí tzv. Chebysheva nerovnost

Mají-li data libovolné rozdělení s konečným průměrem (μ) a konečnou sm. odchylkou (σ), pak nejméně $\left(1 - \frac{1}{k^2}\right) \cdot 100$ % variant leží v intervalu $(\mu - k\sigma; \mu + k\sigma)$.

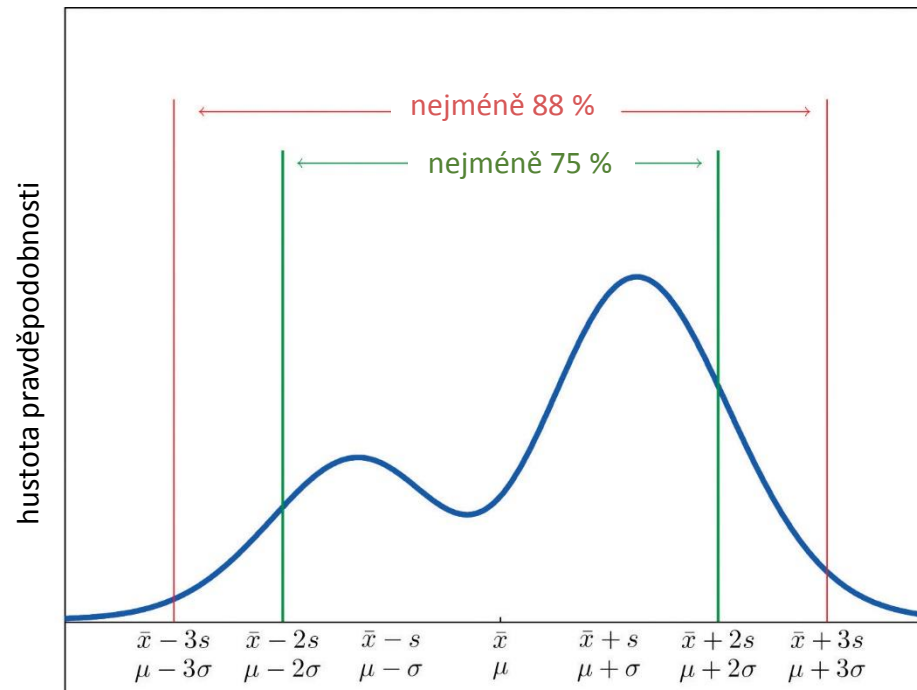
$$\forall k > 0: P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$$

k	$P(\mu - k\sigma < X < \mu + k\sigma)$ (tj. pravděpodobnost, že data leží v intervalu $(\mu - k\sigma; \mu + k\sigma)$)
1	> 0 %
2	> 75 %
3	> 88 %

Jakou představu o variabilitě dat nám dává směrodatná odchylka?

Obecně platí tzv. Chebyschevova nerovnost

Mají-li data libovolné rozdělení s konečným průměrem (μ) a konečnou sm. odchylkou (σ), pak nejméně $\left(1 - \frac{1}{k^2}\right) \cdot 100\%$ variant leží v intervalu $(\mu - k\sigma; \mu + k\sigma)$.



Jakou představu o variabilitě dat nám dává směrodatná odchylka?

Pro data, která mají normální rozdělení platí:

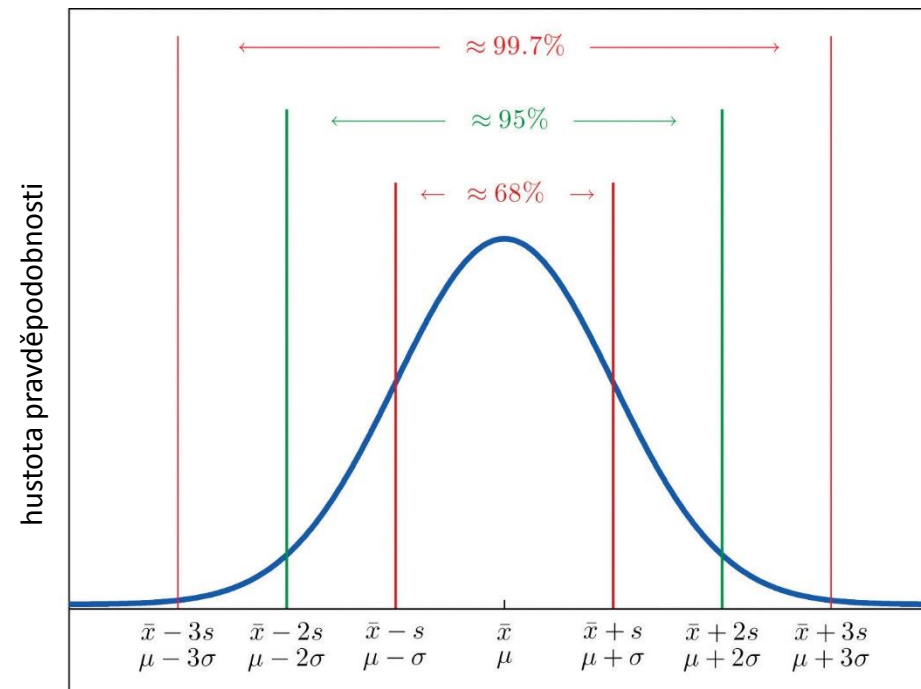
Mají-li data normální rozdělení (obálka histogramu odpovídá Gaussově křivce (zvonovitý tvar)) s konečným průměrem (μ) a konečnou sm. odchylkou (σ), pak:

k	$P(\mu - k\sigma < X < \mu + k\sigma)$ (tj. pravděpodobnost, že data leží v intervalu $(\mu - k\sigma; \mu + k\sigma)$)
1	$\cong 68 \%$
2	$\cong 95 \%$
3	$\cong 99,7 \%$

Jakou představu o variabilitě dat nám dává směrodatná odchylka?

Pro data, která mají normální rozdělení platí:

Mají-li data normální rozdělení (obálka histogramu odpovídá Gaussově křivce (zvonovitý tvar)) s konečným průměrem (μ) a konečnou sm. odchylkou (σ), pak:



Popisná statistika - kvantitativní znak

Míry variability	Hmotnost (kg)
Rozptyl	215,3893
Směrodatná odchylka	14,67615
Variační koeficient (%)	18,90576

Jak zaokrouhlovat výběrové charakteristiky?

Směrodatnou odchylku zaokrouhlujeme nahoru na 1-2 platné cifry.

Míry polohy zaokrouhlujeme následně na stejný řád.

Popisná statistika - kvantitativní znak

Jak zaokrouhlovat výběrové charakteristiky?

Směrodatnou odchylku zaokrouhlujeme nahoru na 1-2 platné cifry.

Míry polohy zaokrouhlujeme následně na stejný řád.

Míry polohy	Váha (kg)	zaokrouhleno
minimum	50	
dolní kvartil	68	
průměr	77,62791	
medián	75,5	
horní kvartil	84,5	
maximum	130	
Míry variability		
směrodatná odchylka	14,67615	
variační koeficient (%)	18,90576	

Popisná statistika - kvantitativní znak

Jak zaokrouhlovat výběrové charakteristiky?

Směrodatnou odchylku zaokrouhlujeme nahoru na 1-2 platné cifry.

Míry polohy zaokrouhlujeme následně na stejný řád.

Míry polohy	Váha (kg)	zaokrouhleno
minimum	50	
dolní kvartil	68	
průměr	77,62791	
medián	75,5	
horní kvartil	84,5	
maximum	130	
Míry variability		
směrodatná odchylka	14,67615	15
variační koeficient (%)	18,90576	

zaokrouhlujeme **nahoru** na 1-2 platné cifry

Popisná statistika - kvantitativní znak

Jak zaokrouhlovat výběrové charakteristiky?

Směrodatnou odchylku zaokrouhlujeme nahoru na 1-2 platné cifry.

Míry polohy zaokrouhlujeme následně na stejný řád.

Míry polohy	Váha (kg)	zaokrouhleno
minimum	50	
dolní kvartil	68	68
průměr	77,62791	78
medián	75,5	76
horní kvartil	84,5	85
maximum	130	
Míry variability		
směrodatná odchylka	14,67615	15
variační koeficient (%)	18,90576	

zaokrouhlujeme na stejný řád
jako směrodatnou odchylku

Popisná statistika - kvantitativní znak

Jak zaokrouhlovat výběrové charakteristiky?

Směrodatnou odchylku zaokrouhlujeme nahoru na 1-2 platné cifry.

Míry polohy zaokrouhlujeme následně na stejný řád.

Míry polohy	Váha (kg)	zaokrouhleno
minimum	50	50
dolní kvartil	68	68
průměr	77,62791	78
medián	75,5	76
horní kvartil	84,5	85
maximum	130	130
Míry variability		
směrodatná odchylka	14,67615	15
variační koeficient (%)	18,90576	

← nezaokrouhlujeme
(údaj vybrán z datového souboru)

← nezaokrouhlujeme
(údaj vybrán z datového souboru)

Popisná statistika - kvantitativní znak

Jak zaokrouhlovat výběrové charakteristiky?

Směrodatnou odchylku zaokrouhlujeme nahoru na 1-2 platné cifry.

Míry polohy zaokrouhlujeme následně na stejný řád.

Míry polohy	Váha (kg)	zaokrouhleno
minimum	50	50
dolní kvartil	68	68
průměr	77,62791	78
medián	75,5	76
horní kvartil	84,5	85
maximum	130	130
Míry variability		
směrodatná odchylka	14,67615	15
variační koeficient (%)	18,90576	18,9

zaokrouhlujeme na desetiny %
(potřebujeme srovnávat s 50 %)

DĚKUJI ZA POZORNOST!

martina.litschmannova@vsb.cz

