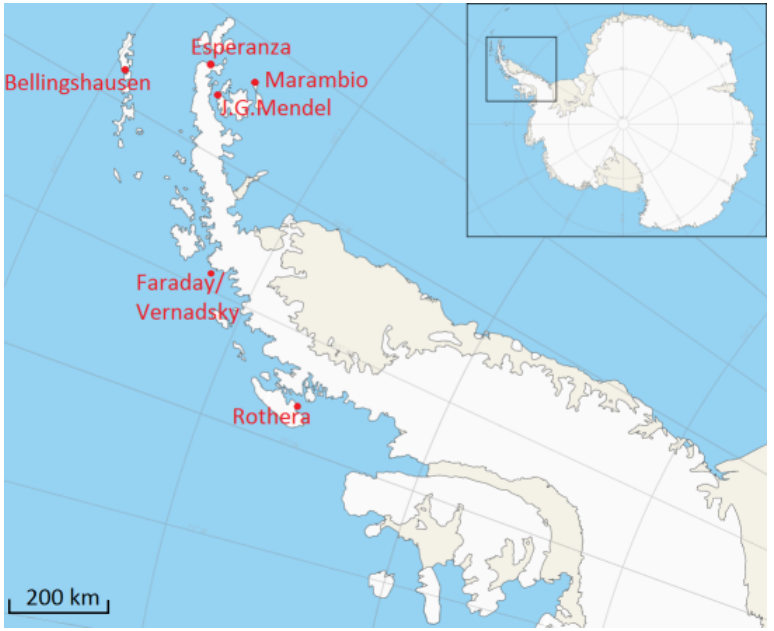
Odhady parametrů základního souboru

Cvičení 6
Statistické metody a zpracování dat 1 (podzim 2016)
Brno, říjen–listopad 2016
Ambrožová Klára

Motivační příklad

- Mám průměrné roční teploty vzduchu z 8 stanic v oblasti Antarktického poloostrova z let 2005–2015.



READER	1 2005	2 2006	3 2007	4 2008	5 2009	6 2010	7 2011	8 2012	9 2013	10 2014	
Ro	-4,50833333	-4,10833333	-3,68333333	-2,7	-3,88333333	-2,71666667	-3,78333333	-3,75	-4,35	-3,74166667	-5,(
Faraday	-3,29166667	-1,96666667	-2,36666667	-1,80833333	-2,24166667	-1,5	-2,8	-2,44166667	-3,09166667	-2,60833333	-4,3
Bel	-1,94166667	-1,3	-3,24166667	-1,14166667	-2,71666667	-1,55833333	-2,91666667	-2,75833333	-2,59166667	-2,28333333	-2,7
Esp	-4,4	-3,33333333	-7,025	-3,55833333	-5,71666667	-3,775	-5,475	-5,475	-4,71666667	-4,68333333	
Ma	-7,525	-6,41666667	-10,7833333	-6,84166667	-9,58333333	-7,425	-9,45833333	-9,375	-8,46666667	-8,35833333	-8,3
JRI	-6,00363845	-4,56857068	-9,1553138	-5,47369866	-8,25154814	-5,80089297	-8,02302831	-8,08738599	-7,27670253	-6,98573058	-6,9
Bibby	-6,95770314	-5,8223198	-10,2228753	-6,14441924	-9,30447946	-7,24184803	-9,06120345	-9,09187158	-8,41396823	-8,23276116	-8,1

Motivační příklad

- Výběrové průměry
  a směrodatné odchylky:

Proměnná	Popisné statistiky (READ	
	Průměr	Smodch
2005	-4,94686	1,87352737
2006	-3,93084	1,76376026
2007	-6,63974	3,30787608
2008	-3,95259	2,08075198
2009	-5,95681	2,91538752
2010	-4,28825	2,37906579
2011	-5,93108	2,70911377
2012	-5,85418	2,79757432
2013	-5,55819	2,31093017
2014	-5,27050	2,41354165
2015	-5,83517	1,94999144

- Výše uvedené statistiky platí pro náš výběr (tj. 8 míst, na nichž se měří), ale platí i pro základní soubor (celou oblast Antarktického poloostrova)?

Trocha teorie

- Můžeme počítat statistické charakteristiky pro náš výběr, např.

	Výběrový soubor	Základní soubor	Odhad
– Výběrový průměr	$\bar{x}$	μ	$\hat{\mu}$
– Výběrový rozptyl	s^2	σ^2	$\hat{\sigma}^2$
– Výběrová směrodatná odchylka	s	σ	$\hat{\sigma}$
– a spousta dalších ☺			

- V praxi ale potřebujeme znát statistické charakteristiky základního souboru, resp. jak moc se naše výběrové stat. charakteristiky liší od stat. charakteristik základního souboru?

Odhady

- Odhad bodový:
  - vyjádření jedním číslem
  - nevýhoda: neznáme riziko, že dané číslo není skutečnou charakteristikou základního souboru!

  - Odhad aritmetického průměru: $\bar{x} = \hat{\mu}$
  - Odhad rozptylu: $s^2 = \hat{\sigma}^2$
  - Odhad směrodatné odchylky: $s = \hat{\sigma}$

Pro tyto tři charakteristiky platí, že výběrové charakteristiky jsou nestranným odhadem charakteristik základního souboru.

Záhada směrodatné odchylky

aneb proč bylo na přednášce, že směrodatná odchylka výběrového souboru není nezkresleným odhadem směrodatné odchylky základního souboru?

- **Popisná statistika:**
  - Zabývá se pouze popisem vlastností našeho výběru
  - Definice směrodatné odchylky (výběrového souboru):

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \qquad (1)$$

- **Matematická statistika a statistika pravděpodobnosti:**
  - Uvědomuje si, že náš soubor je pouze výběr, a že bude třeba zobecňovat
  - Definice směrodatné odchylky (výběrového souboru):

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} \qquad (2)$$

Sm. odchylka ze vzorce *(1)* tedy není nezkresleným odhadem, hodnota ze vzorce *(2)* je nezkresleným odhadem

Tohle počítá program STATISTICA. Hodnotu jako ze vzorce (1) lze získat takto:

(1) = *(2)* * √(n-1) / n

Poznámka k bodovému odhadu průměru

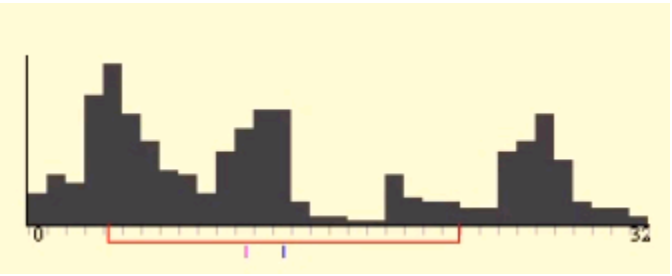
- Směrodatná chyba průměru (Standard error of the mean):
  - směrodatná odchylka výběrových průměrů, která závisí na směrodatné odchylce základního souboru (σ) a velikosti výběrů (n)

$$\sigma_{x\text{-}bar} = (\sigma^2/n)^{1/2}$$

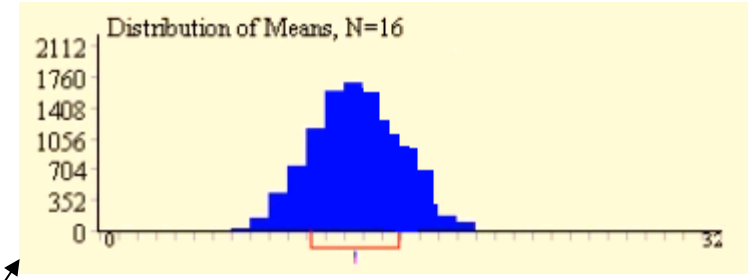
  - v praxi se používá výpočet na základě 1 náhodného výběru o velikosti n a směrodatné odchylce s

$$s_{x\text{-}bar} = (s^2/n)^{1/2}$$

  - Je zjevné, že čím větší je náš výběr n,tím je menší chyba (a menší riziko, že jsme se v odhadu zmýlili)



Provedeme-li např. 1000 výběrů z tohoto souboru o velikosti n=16, jejich průměry lze vykreslit ve vedlejším grafu



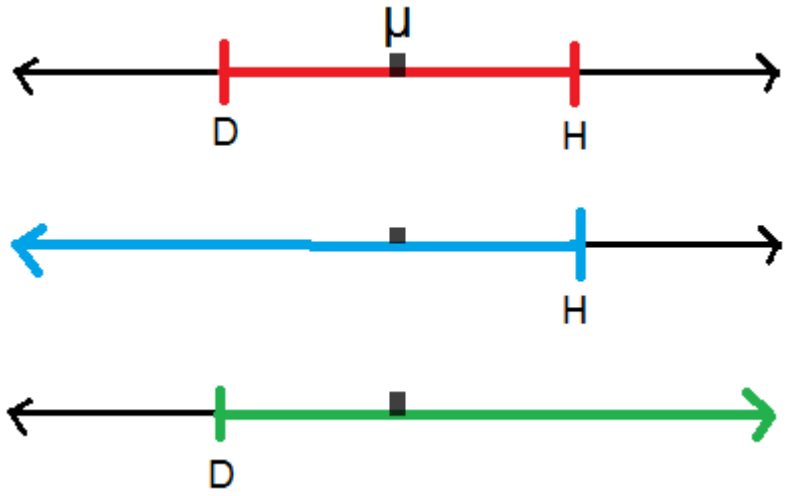
Intuitivně by všechny výběrové průměry měly být rovny průměru základního souboru, prakticky tomu tak není. Soubor složený z těchto výběrových průměrů má svou variabilitu, která bude tím menší, čím větší je n.

Odhady

- Odhad intervalový
  - Odhad intervalem hodnot
  - známe riziko, s nímž se reálná hodnota v tomto intervalu nenachází

  - Existují tři typy intervalů:

1. Oboustranný interval spolehlivosti
   - Odhadovaná charakteristika se v intervalu (D,H) nachází s pravděpodobností 1-α
2. Pravostranný interval spolehlivosti
   - Odhadovaná charakteristika je menší než H s pravděpodobností 1-α
3. Levostranný interval spolehlivosti
   - Odhadovaná charakteristika je větší než D s pravděpodobností 1-α



α – hladina významnosti (riziko). Zvolíme-li *α=0.05*, pak bude odhad. charakteristika v daném intervalu s pravděpodobností *95 %* (0.95).

Intervalové odhady

PŘEDPOKLAD NORMÁLNÍHO ROZDĚLENÍ!

– Intervalový odhad aritmetického průměru (n>30):

D H

$$\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}}$$

– Intervalový odhad rozptylu:

$$\frac{(n-1)\cdot s^2}{\chi^2_{\frac{\alpha}{2},(n-1)}} \leq \sigma^2 \leq \frac{(n-1)\cdot s^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}}$$

– Intervalový odhad směrodatné odchylky: meze získáme odmocněním hodnot D a H

$\bar{x}$ ... výběrový aritm. průměr

s ... výběrová směrodatná odchylka

n ... počet hodnot ve výběru

z, χ^2 kvantily normovaného normálního (resp. Chí2) rozdělení pro dané n a α

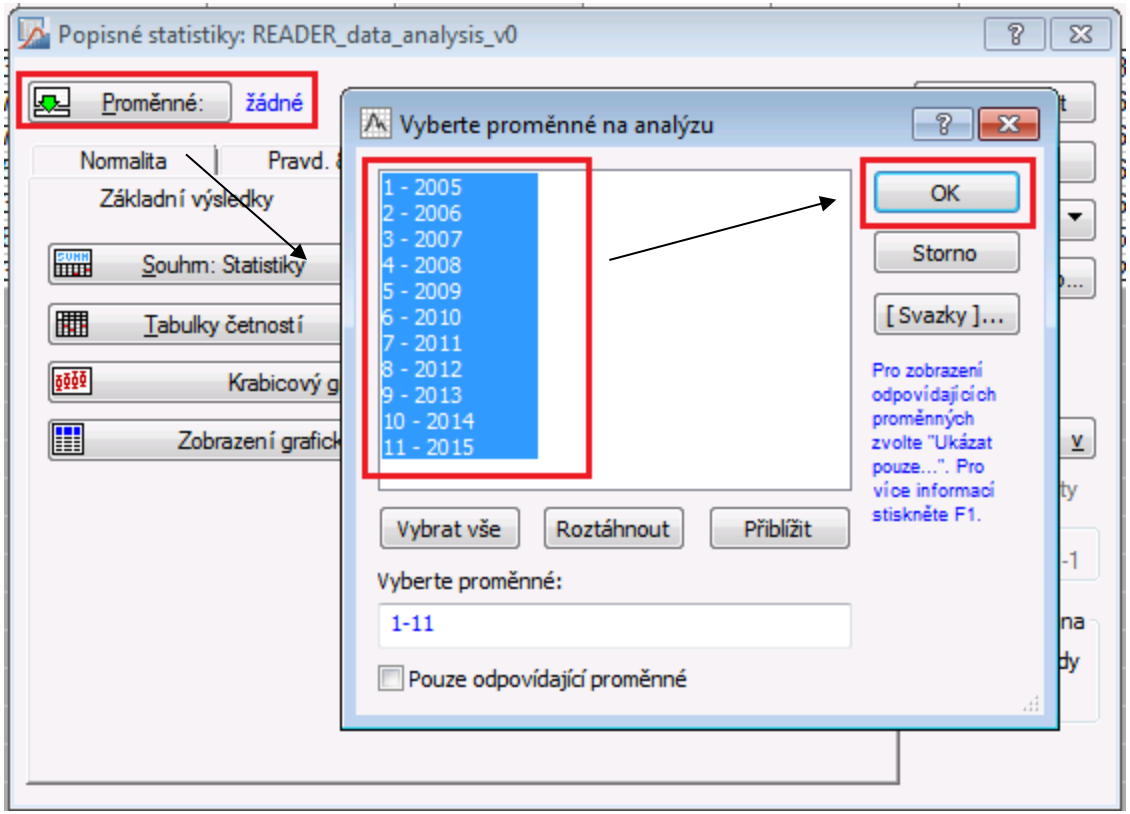
α = 0.05 -> z=1.96

α = 0.01 -> z=2.576

χ2 potřeba nalézt v tabulkách, protože závisí na n

Zpět k motivačnímu případu…

- Bodový a intervalový odhad:
  - Statistiky – Základní statistiky – Popisné statistiky
  - Proměnné: Vše*



* ve vašem případě může jít i o 1 sloupec

Zpět k motivačnímu případu…

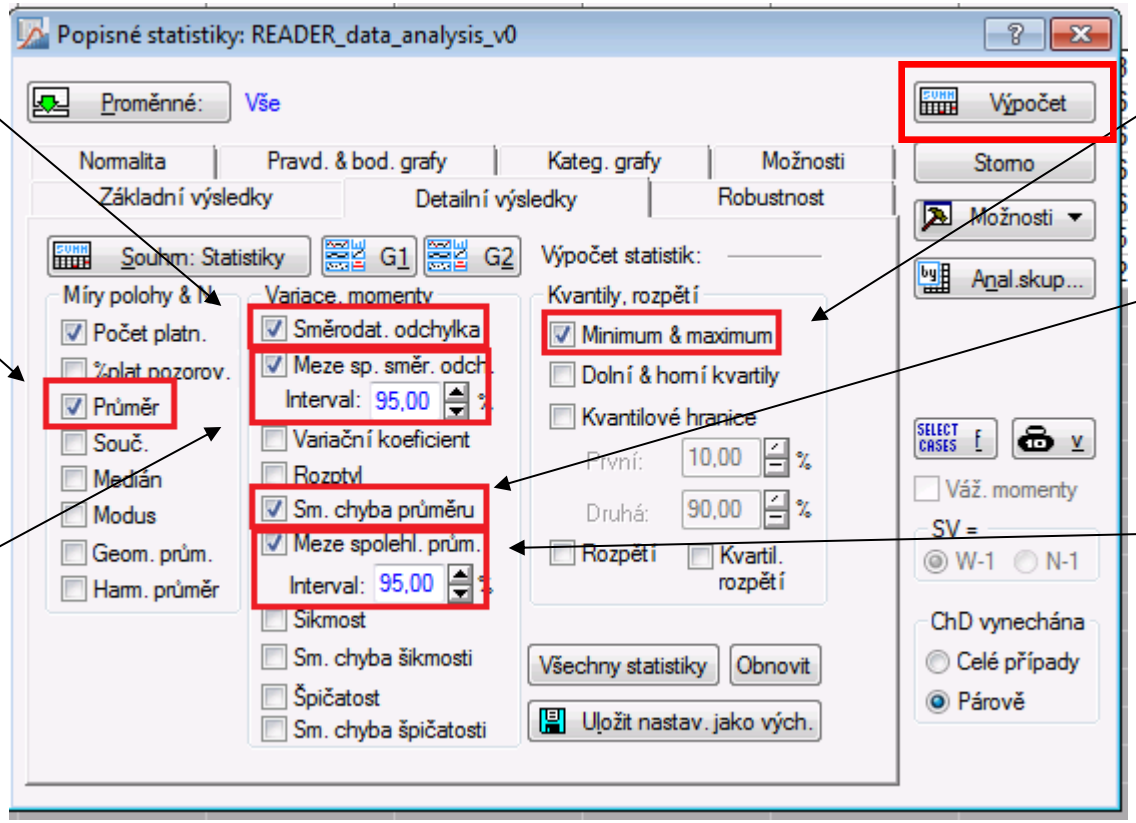
- Bodový a intervalový odhad:
  - Nastavení modulu:* -> klikneme na Výpočet

Bodový odhad sm. odchylky

Bodový odhad průměru

Minimum a maximum

Směrodatná chyba průměru



Výpočet intervalu spolehlivosti sm. odchylky (zde na hladině spolehlivosti 95 %!!!!!!!!)

Výpočet intervalu spolehlivosti průměru (zde na hladině spolehlivosti 95 %!!!!!!!!)

* ve cvičení máte 2 zadání a ne vždy musíte počítat všechno!

Zpět k motivačnímu příkladu

- Bodový a intervalový odhad

Bodový odhad průměru

Dolní hranice intervalu spolehlivosti průměru

Horní hranice intervalu spolehlivosti průměru

Bodový odhad sm. odchylky

Dolní hranice intervalu spolehlivosti sm. odchylky

Horní hranice intervalu spolehlivosti sm. odchylky

Popisné statistiky (READER_data_analysis_v0)

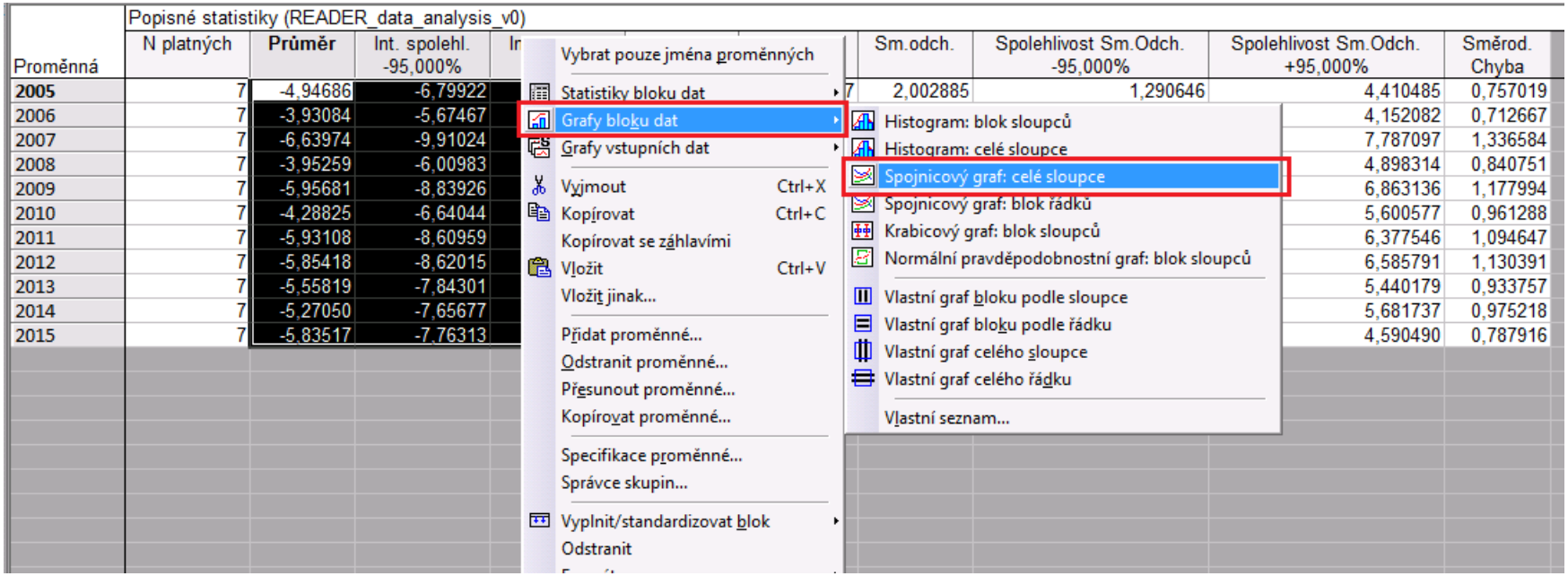
Proměnná	N platných	Průměr	Int. spolehl. -95,000%	Int. spolehl. 95,000%	Minimum	Maximum	Sm.odch.	Spolehlivost Sm.Odch. -95,000%	Spolehlivost Sm.Odch. +95,000%	**Směrod. Chyba**
2005	7	-4,94686	-6,79922	-3,09450	-7,5250	-1,94167	2,002885	1,290646	4,410485	0,75701
2006	7	-3,93084	-5,67467	-2,18701	-6,4167	-1,30000	1,885539	1,215029	4,152082	0,71266
2007	7	-6,63974	-9,91024	-3,36924	-10,7833	-2,36667	3,536268	2,278748	7,787097	1,33658
2008	7	-3,95259	-6,00983	-1,89535	-6,8417	-1,14167	2,224417	1,433400	4,898314	0,84075
2009	7	-5,95681	-8,83926	-3,07436	-9,5833	-2,24167	3,116680	2,008368	6,863136	1,17799
2010	7	-4,28825	-6,64044	-1,93606	-7,4250	-1,50000	2,543328	1,638904	5,600577	0,96128
2011	7	-5,93108	-8,60959	-3,25258	-9,4583	-2,80000	2,896164	1,866269	6,377546	1,09464
2012	7	-5,85418	-8,62015	-3,08821	-9,3750	-2,44167	2,990733	1,927208	6,585791	1,13039
2013	7	-5,55819	-7,84301	-3,27337	-8,4667	-2,59167	2,470488	1,591966	5,440179	0,93375
2014	7	-5,27050	-7,65677	-2,88423	-8,3583	-2,28333	2,580185	1,662654	5,681737	0,97521
2015	7	-5,83517	-7,76313	-3,90721	-8,3583	-2,79167	2,084629	1,343321	4,590490	0,78791

Některé věci jsou logické, např. bodový odhad hodnoty je vždy mezi dolní a horní hranicí intervalu… Pokud to tak nemáte, něco bylo uděláno špatně.

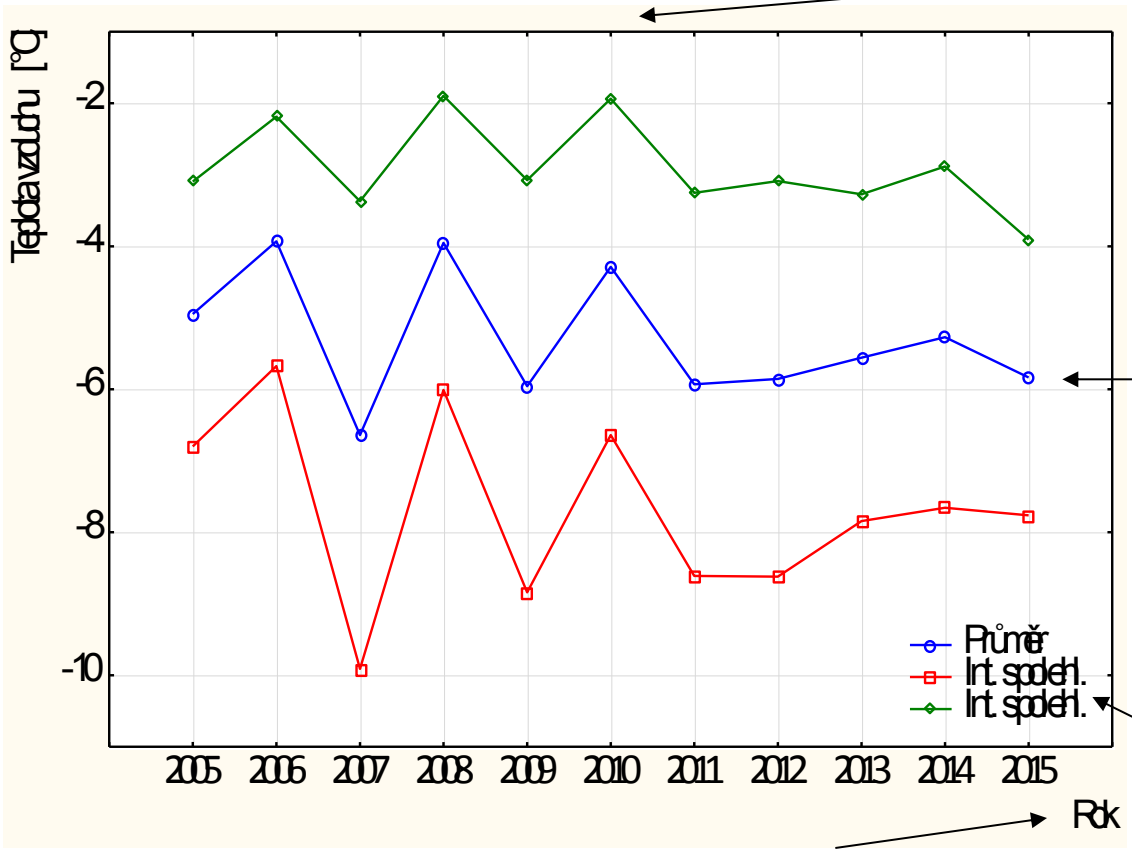
Zpět k motivačnímu příkladu

- Tvorba spojnicového grafu pro bodový a intervalové odhady průměru

Nejprve levým tlačítkem označit sloupce průměru
a obou hranic intervalů spolehlivosti a kliknout pravým tlačítkem



Zpět k motivačnímu příkladu



Název lze odstranit označením a kliknutím na Delete

Křivka dole: dolní hranice intervalu, křivka uprostřed: bodový odhad průměru, křivka nahoře: horní hranice intervalu

Legendu lze posouvat po kliknutí pravým tlačítkem: „Změnit na plovoucí text“

Název osy lze změnit kliknutím pravým tlačítkem: Možnosti grafu – Osa – Název

Čeho si třeba všimnout do závěru: Nejnižší teplota byla v r. 2007, kdy byl také zjevně nejširší interval spolehlivosti. Tzn. v tomto roce je bodový odhad méně spolehlivý!

Cvičení č. 6

6.1. Zadání: Proveďte bodový a intervalový odhad průměru a směrodatné odchylky základního souboru pro **95% a 99% interval spolehlivosti**. Jako výběrový soubor použijte řadu průměrných ročních teplot vzduchu na stanici Praha, Klementinum za období 120 let od ..... do ..... (viz. cvičení 2).

6.2. Zadání: Z průměrných měsíčních hodnot teploty vzduchu Vámi zpracovávané stanice (viz. cvičení 3) určete pro každý měsíc intervalový odhad průměru **na hladině spolehlivosti 95 %** a dále **směrodatnou chybu průměru, [aritmetický průměr, minimum a maximum]** . Hodnoty aritmetického průměru a intervalového odhadu vyneste do vhodného typu grafu, tak abyste mohli názorně prezentovat rozdíly mezi jednotlivými měsíci. V závěru porovnejte intervalový odhad pro jednotlivé měsíce a interpretujte - o čem vypovídá? Jak souvisí např. s variabilitou studované veličiny v daném měsíci?

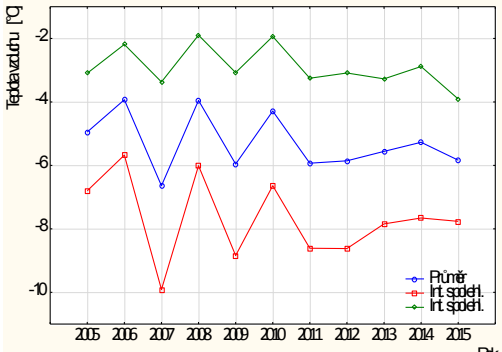
Cvičení 6

Požadovaný výstup cvičení:

- V zadání uvést Vaše období ze cv. 2 a Vaši stanici ze cv.3
- 6.1: Tabulka s odhady
- 6.2: Tabulka s aritm. průměry, dolní a horní hranicí intervalů a sm. chybou průměru; spojnicový NEBO krabicový graf
- **Závěr:**
  - 6.1: Všimněte si např. šířky intervalů pro 99% a 95% interval spolehlivosti – který je širší?
  - 6.2: Všimněte si, jak se šířka intervalu měří v průběhu roku? Čím to může být způsobeno?

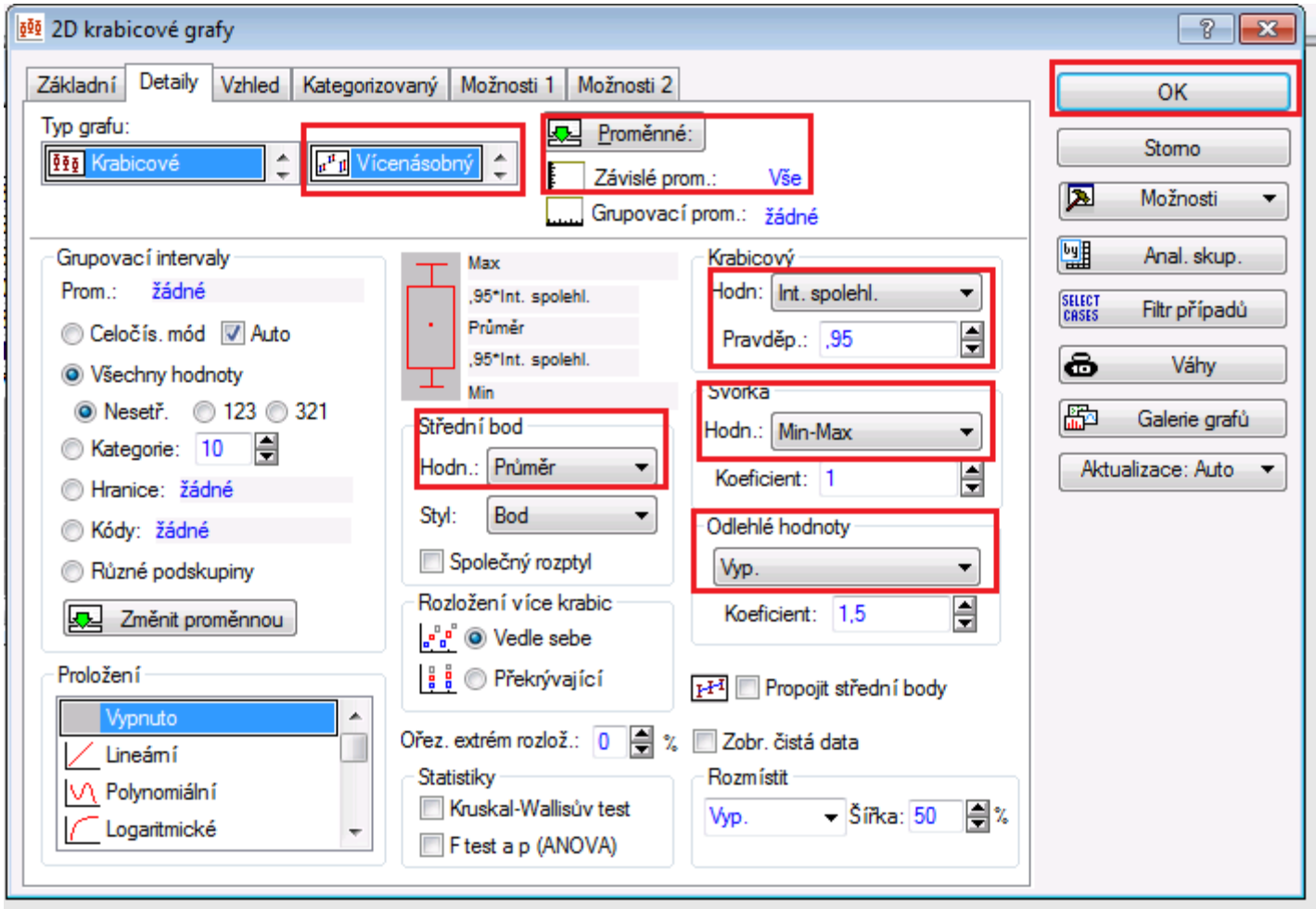
Bodový odhad aritmetického průměru	-4.95
Dolní hranice 95% intervalu spolehlivosti průměru	-6.80
Horní hranice 95% intervalu spolehlivosti průměru	-3.09
Dolní hranice 99% intervalu spolehlivosti průměru	-7.75
Horní hranice 99% intervalu spolehlivosti průměru	-2.14
Bodový odhad směrodatné odchylky	2.00
Dolní hranice 95% intervalu spolehlivosti sm.odch.	1.29
Horní hranice 95% intervalu spolehlivosti sm.odch.	4.41
Dolní hranice 99% intervalu spolehlivosti sm.odch.	1.14
Horní hranice 99% intervalu spolehlivosti sm.odch.	5.97

	Průměr	Dolní hranice 95% int. Spolehlivosti průměru	Horní hranice 95% int. Spolehlivosti průměru	Minimum	Maximum	Směrodatná chyba
2005	-4.95	-6.80	-3.09	-7.53	-1.94	0.76
2006	-3.93	-5.67	-2.19	-6.42	-1.30	0.71
2007	-6.64	-9.91	-3.37	-10.78	-2.37	1.34
2008	-3.95	-6.01	-1.90	-6.84	-1.14	0.84
2009	-5.96	-8.84	-3.07	-9.58	-2.24	1.18
2010	-4.29	-6.64	-1.94	-7.43	-1.50	0.96
2011	-5.93	-8.61	-3.25	-9.46	-2.80	1.09
2012	-5.85	-8.62	-3.09	-9.38	-2.44	1.13
2013	-5.56	-7.84	-3.27	-8.47	-2.59	0.93
2014	-5.27	-7.66	-2.88	-8.36	-2.28	0.98
2015	-5.84	-7.76	-3.91	-8.36	-2.79	0.79



Poznámka ke cvičení: Nastavení pro tvorbu krabicového grafu

- Grafy – 2D grafy – Krabicové grafy
- Proměnné: Závislé (označit všechny měsíce)
- Karta Detaily



I tento graf je třeba upravit (osy, názvy, legenda…)

Zdroje

- BRÁZDIL, Rudolf. Statistické metody v geografii :cvičení. 3. vyd. Brno: Vydavatelství Masarykovy univerzity, 1995. 177 s. ISBN 80-210-1260-9.
- BUDÍKOVÁ, Marie. Základní pojmy matematické statistiky (přednáška). Brno: Masarykova univerzita, 27.9. 2016.
- DOBROVOLNÝ, Petr. Z1069 Statistické metody a zpracování dat:IV. Odhady parametrů. Brno: Masarykova univerzita, 27.9.2016.
- KHAN ACADEMY. KhanAcademy. <https://www.khanacademy.org/> 27.9.2016.
- STATSOFT. Nápověda k programu STATISTICA. 27.9.2016.
- Část dat použitých v příkladu pochází z databáze READER <https://legacy.bas.ac.uk/met/READER/>.