Úvod, data v médiích

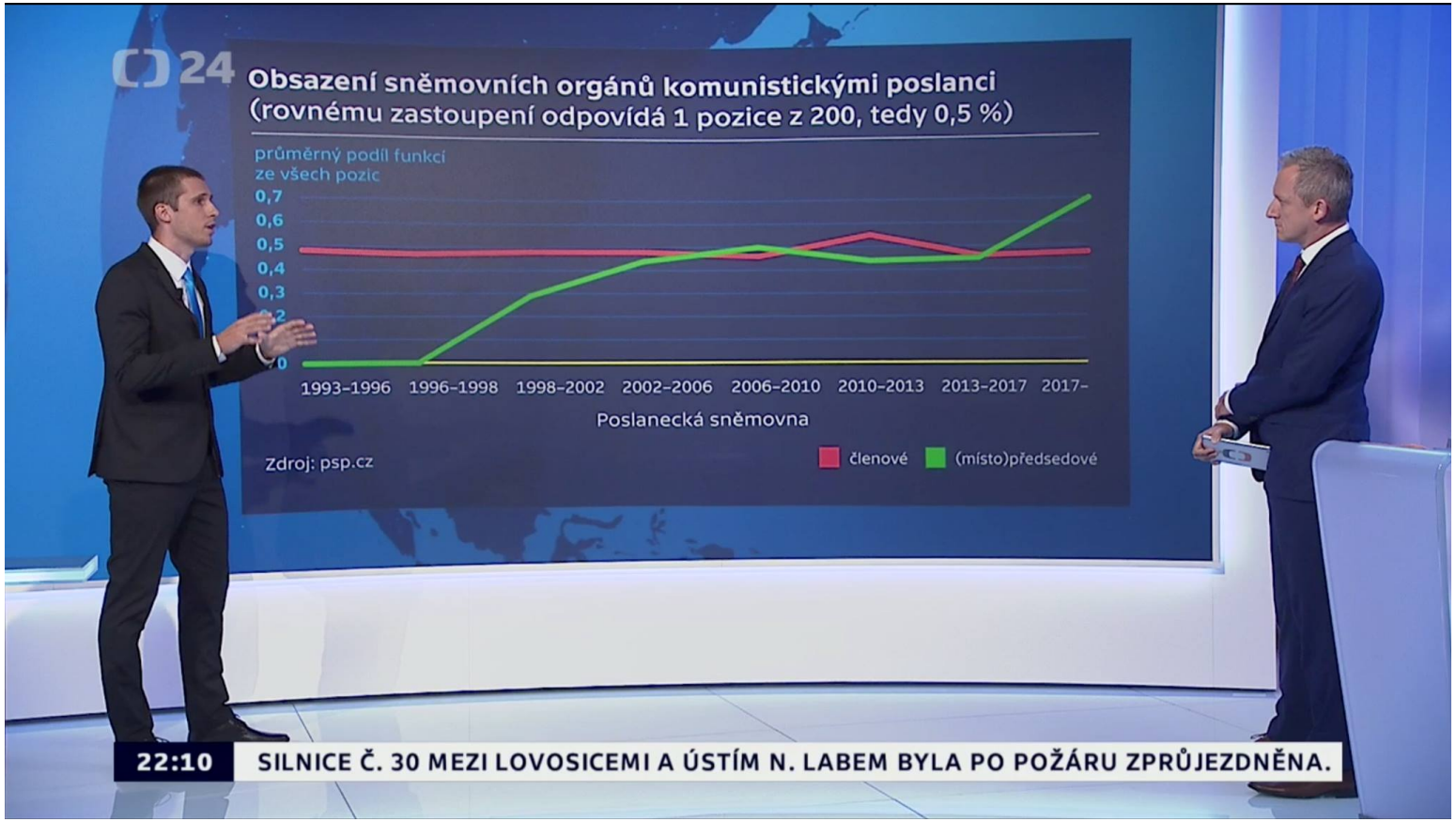
ANALÝZA DAT

18. ÚNORA 2021

Proč jsme tady?

theschoolrun.com

Proč já?


ČT24
Obsazení sněmovních orgánů komunistickými poslanci
(rovnému zastoupení odpovídá 1 pozice z 200, tedy 0,5 %)
průměrný podíl funkcí
ze všech pozic
0,7
0,6
0,5
0,4
0,3
1993–1996
1996–1998
1998–2002
2002–2006
2006–2010
2010–2013
2013–2017
2017–
Poslanecká sněmovna
Zdroj: psp.cz
členové
(místo)předsedové
22:10
SILNICE Č. 30 MEZI LOVOSICEMI A ÚSTÍM N. LABEM BYLA PO POŽÁRU ZPRŮJEZDNĚNA.

Proč já?

Analýza dat

1. Úvod, data v médiích
   - obsah předmětu a nároky na splnění, role dat v médiích
2. Práce s daty
   - typy dat, replikace výstupů datové žurnalistiky
3. Volební průzkumy
   - logika a ideální podoby prezentace výsledků volebních průzkumů
4. Vizualizace dat
   - nejlepší způsoby grafické prezentace dat

Analýza dat

5. Statistické usuzování
   - logika aplikace výsledků dat ze vzorku na celou populaci
6. Základní statistické nástroje
   - korelace a t-test
7. Regresní analýza
   - nejvíce využívaná metoda v kauzálním sociálněvědním výzkumu

Analýza dat

8. Základní popisná statistika v prostředí MS Excel
   - základní práce s daty v programu
9. Postupy zobecňování v prostředí MS Excel
   - použití základních statistických nástrojů v programu
10. Sociálněvědní archivy a velké mezinárodní výzkumy
    - představení významných zdrojů dat
11. Etické aspekty výzkumu a publikace jeho výsledků
    - základní etické aspekty výzkumné práce

Organizace výuky

- povinný předmět pro specializaci Mediální studia
- čtvrtek 12:30–13:50
- 11 týdnů
- výuka bude probíhat on-line: https://cesnet.zoom.us/j/96177417282

Moodle

- sdílení prezentací z hodin a podkladů pro semináře
- plnění testů
- MOODLE pro výuku 1 - https://dl1.cuni.cz
- do systému se přihlašujte pomocí údajů CAS UK (https://dl.cuni.cz/?page_id=17):
  - uživatelské jméno: 8místné číslo na ISIC kartě pod fotografií (stejné pro přihlášení do SISu)
  - heslo: stejné jako pro SIS
- přihlásit se do kurzu:
  - Analýza dat (JKB045)
    (https://dl1.cuni.cz/course/view.php?id=9109)

Požadavky na absolvování

- studium literatury (věcné pochopení postupů a metod), účast na přednáškách, samostatná práce, konzultace
- výsledná známka bude odvozena od složení dvou testů v Moodle
  - jejich smyslem není testovat znalost složitější matematiky, ale spíše logiky a pochopení věci
- test č. 1 (40 % výsledné známky)
  - obsahem literatura a přednášky z prvních sedmi setkání
  - absolvování po 7. přednášce semestru
- test č. 2 (60 % výsledné známky)
  - obsahem literatura a přednášky z celého semestru
  - k dispozici budou 3 termíny (po jednom v červnu, červenci a září)

Kontakt

Lukáš Hájek

lukas.hajek@fsv.cuni.cz
čtvrtek 14:00–15:20 v on-line místnosti
https://meet.google.com/tpv-dcbw-ddj (nebo kdykoliv po dohodě e-mailem v on-line formě)

Petr Soukup

petr.soukup@fsv.cuni.cz
úterý 8:30–9:00 v on-line místnosti
https://cesnet.zoom.us/j/4677639176?pwd=dVAyWkJISGRrOGp2ZUJRc0I2Y3p2UT09

fórum na Moodle

Data v médiích

Data v médiích

THE UPSHOT

The Complete List of Trump's Twitter Insults (2015-2021)

This list documents all the verbal attacks Mr. Trump posted on Twitter, from when he declared his candidacy in June 2015 to Jan. 8, when Twitter permanently barred him.

By Kevin Quealy

nytimes.com

Data

- kvalitativní data
  - data nečíselné povahy
  - příklady jsou texty (vstupy moderátorů), obrazy (vizuály volebních kampaní) nebo zvukové záznamy (záznam pořadu)
  - pracuje se s nimi jako s jedinečnými, až samostatnými prvky
  - je možné sledovat podobnost vlastností a obsahu
- kvantitativní data
  - data číselné povahy
  - příkladem jsou počty (pořadů), procenta (sledovanosti) nebo průměry (počtu diváků)
  - pracuje se s nimi jako s reprezentanty určité kategorie
  - zpracování probíhá pomocí kvantitativních metod

Výhody a nevýhody různých dat

KVALITATIVNÍ DATA

- ✓ široká paleta druhů
- ✓ existují i tam, kde kvantitativní data ne
- ✓ odhalují nové souvislosti

- ✗ náročné zpracování
- ✗ riziko subjektivního zpracování
- ✗ hrozba snížené přesvědčivosti

KVANTITATIVNÍ DATA

- ✓ snadná práce s daty
- ✓ dostupnost a množství
- ✓ přesvědčivost

- ✗ častá povrchnost
- ✗ mohou být špatně interpretována
- ✗ neumožňují odhalovat nové proměnné

Práce s daty

- už v rámci sběru dat dbejte na jejich třídění a popis – ušetříte si spoustu následné práce
- zaznamenávejte si informace související se sběrem dat (datum sběru, odmítnutí respondentů, …) i jejich následným tříděním (definice nových proměnných, význam jednotlivých označení, ...)
- k práci s daty je vhodné využít speciální software
  - kvantitativní data: Excel, R, SPSS, Stata atd.
  - kvalitativní data: Nvivo, ATLAS.ti, QDA Miner, MAXQDA atd.
- pozor na chyby při práci s daty – mohou se postupně násobit
- pozor na subjektivní dezinterpretace dat ve snaze „odhalit zajímavější výsledky“

Kvantitativní data

- informace, které na sebe berou numerickou podobu
- je možné je statisticky zpracovávat
- jejich hlavní výhodou je schopnost efektivně popsat velké množství případů
- jejich hlavní nevýhodou je neschopnost hlubšího porozumění jednotlivým případům
- význam kvantitativních dat v dnešním světě roste, a to včetně oblasti médií
- je důležité být si vědom nedostatků a vyvažovat je využitím kvalitativních dat

juffrouj.wordpress.com

Zdroje kvantitativních dat

- zdarma dostupné je velké množství velmi kvalitních kvantitativních dat
- univerzitní i další vědecké instituce pravidelně poskytují datasety na nejrůznější témata přímo pro vědecké účely a pravidelně je aktualizují
- zpravidla je pro použití nutná pouze registrace a souhlas s podmínkami užití
- datasety soukromých společností jsou pak často dostupné za poplatek
- alternativou jsou primární data samotných zkoumaných institucí (Poslanecká sněmovna, OSN, OECD atd.)
- další možností je stáhnutí datasetů již publikovaných článků (pokud dataset není dostupný, napište autorovi a požádejte o něj)
- v případě využití datasetů platí stejná logika citování jako v případě jiných zdrojů!

Sčítání lidu, domů a bytů; ČSÚ

CZSO.CZ

The Correlates of War Project

1 2 3

You are here: Home / Data Sets

Data Sets

COW Country Codes

The list of states with COW abbreviations and ID numbers

Read More…

State System Membership (v2016)

This data set records the fluctuating composition of the state system since 1816. It also identifies countries corresponding to the standard Correlates of War country codes.

Read More…

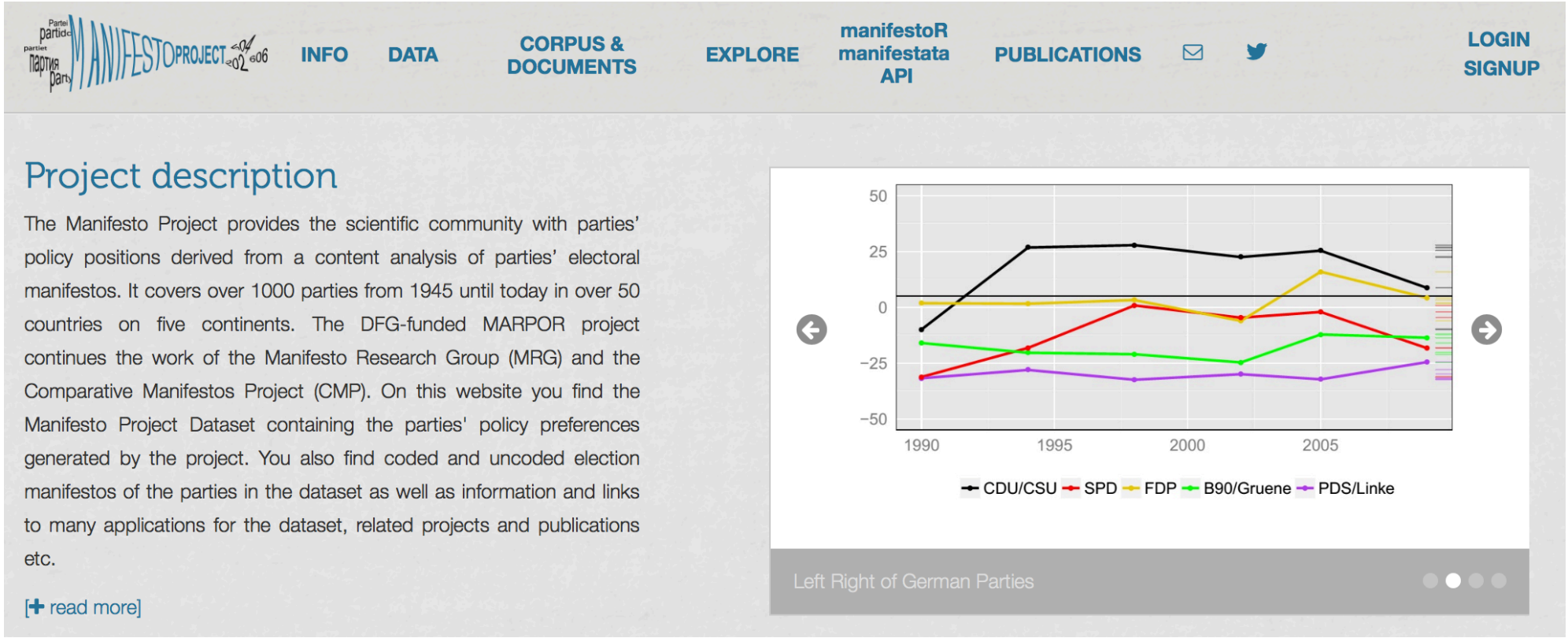
COW War Data, 1816 - 2007 (v4.0)

The new list of wars that will be included in the COW war databases is available. Non-state War data set (v4.0), Intra-state War data set (v4.0), Inter-state War data set (v4.0), and Extra-state War data set (v4.0) are now available.

Read More…

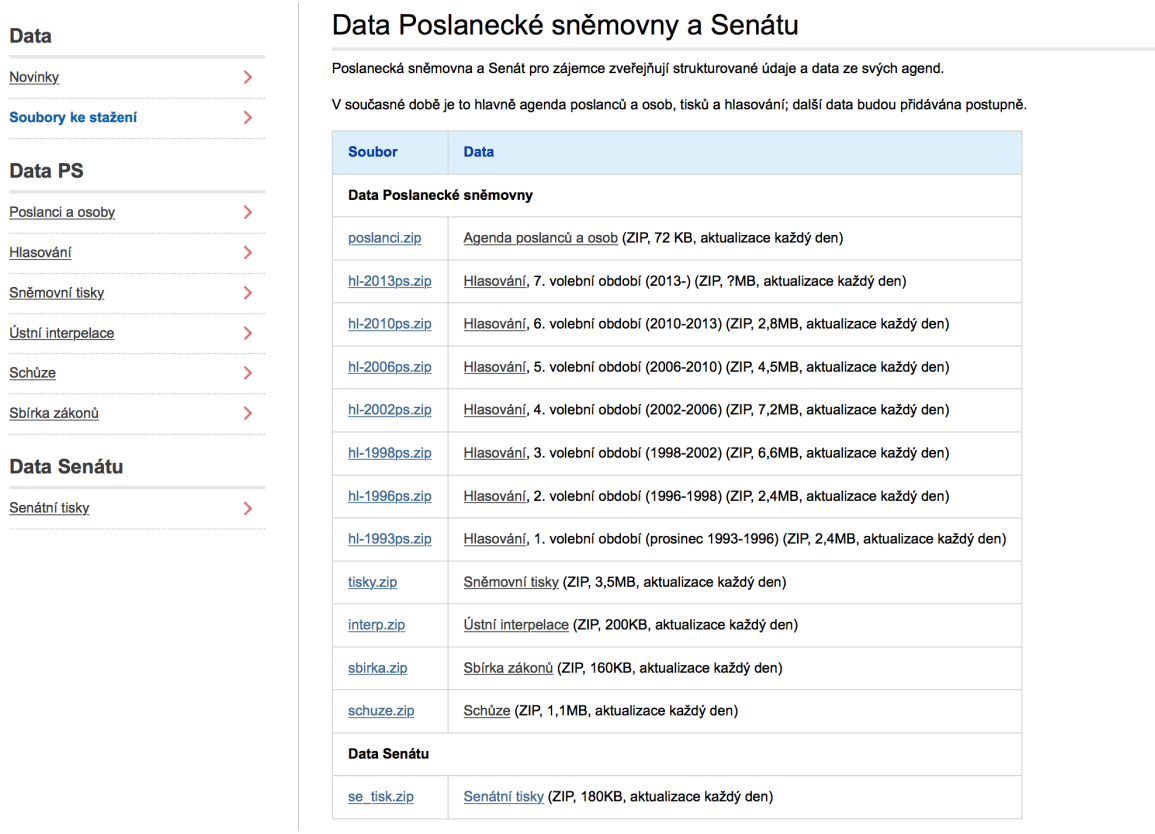
correlatesofwar.org

Manifesto Project



manifesto-project.wzb.eu

Data Poslanecké sněmovny a Senátu



Data

Novinky

Soubory ke stažení

Data PS

Poslanci a osoby

Hlasování

Sněmovní tisky

Ústní interpelace

Schůze

Sbírka zákonů

Data Senátu

Senátní tisky

Data Poslanecké sněmovny a Senátu

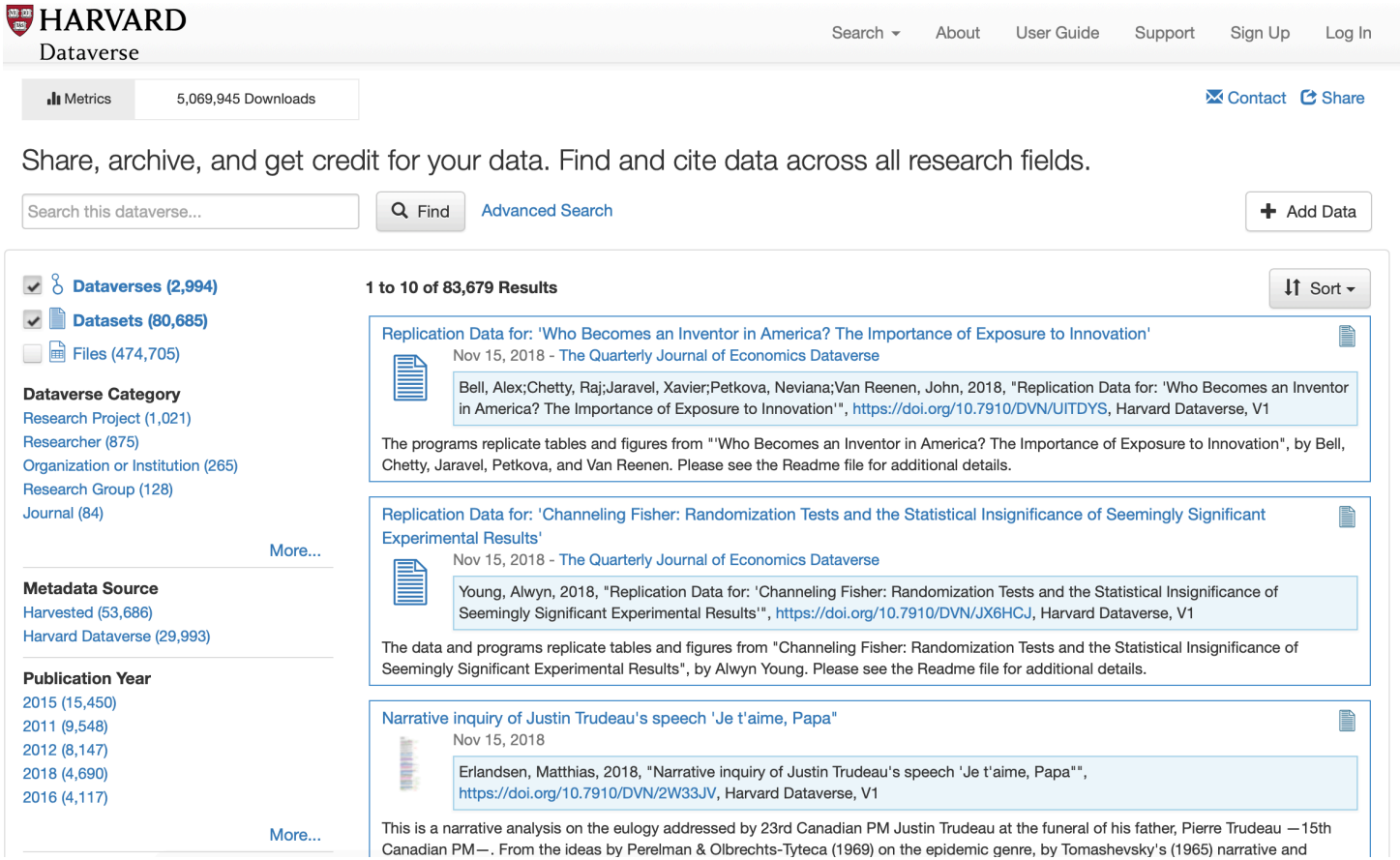
Poslanecká sněmovna a Senát pro zájemce zveřejňují strukturované údaje a data ze svých agend.

V současné době je to hlavně agenda poslanců a osob, tisků a hlasování; další data budou přidávána postupně.

Soubor	Data
Data Poslanecké sněmovny	
poslanci.zip	Agenda poslanců a osob (ZIP, 72 KB, aktualizace každý den)
hl-2013ps.zip	Hlasování, 7. volební období (2013-) (ZIP, ?MB, aktualizace každý den)
hl-2010ps.zip	Hlasování, 6. volební období (2010-2013) (ZIP, 2,8MB, aktualizace každý den)
hl-2006ps.zip	Hlasování, 5. volební období (2006-2010) (ZIP, 4,5MB, aktualizace každý den)
hl-2002ps.zip	Hlasování, 4. volební období (2002-2006) (ZIP, 7,2MB, aktualizace každý den)
hl-1998ps.zip	Hlasování, 3. volební období (1998-2002) (ZIP, 6,6MB, aktualizace každý den)
hl-1996ps.zip	Hlasování, 2. volební období (1996-1998) (ZIP, 2,4MB, aktualizace každý den)
hl-1993ps.zip	Hlasování, 1. volební období (prosinec 1993-1996) (ZIP, 2,4MB, aktualizace každý den)
tisky.zip	Sněmovní tisky (ZIP, 3,5MB, aktualizace každý den)
interp.zip	Ústní interpelace (ZIP, 200KB, aktualizace každý den)
sbirka.zip	Sbírka zákonů (ZIP, 160KB, aktualizace každý den)
schuze.zip	Schůze (ZIP, 1,1MB, aktualizace každý den)
Data Senátu	
se_tisk.zip	Senátní tisky (ZIP, 180KB, aktualizace každý den)

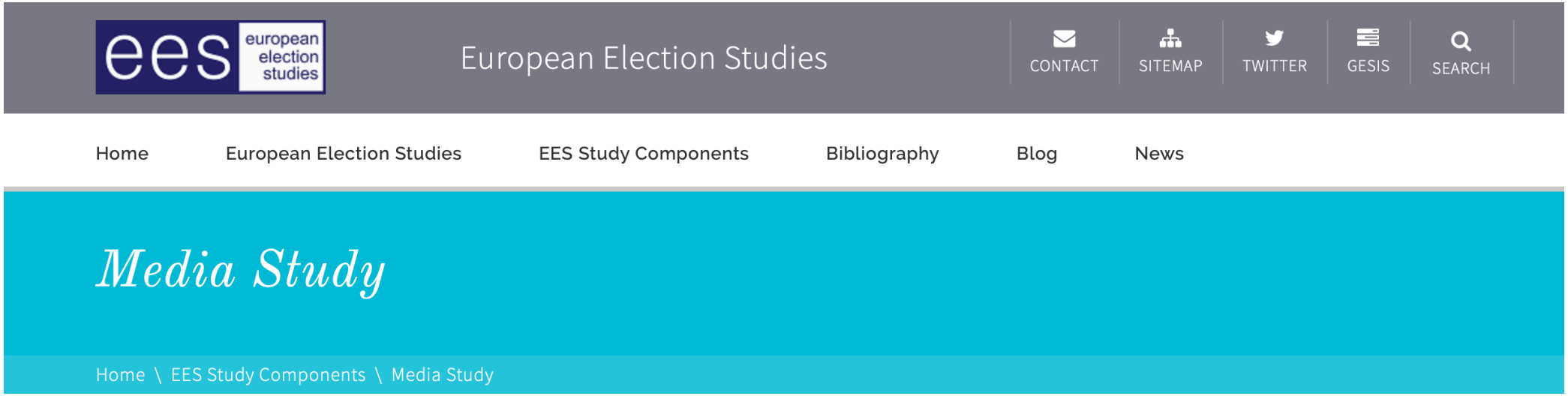
psp.cz

Harvard Dataverse



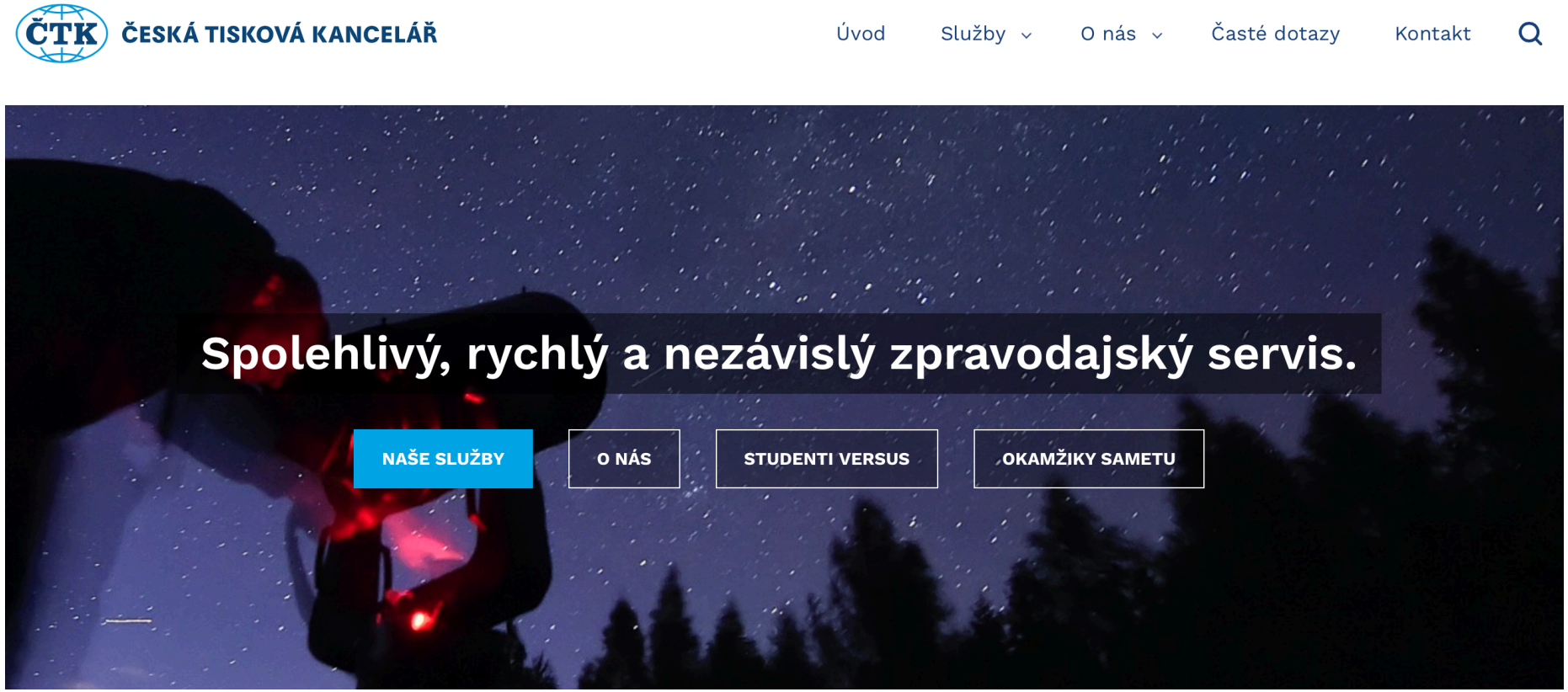
dataverse.harvard.eu

European Election Studies



europeanelectionstudies.net

Infobanka ČTK

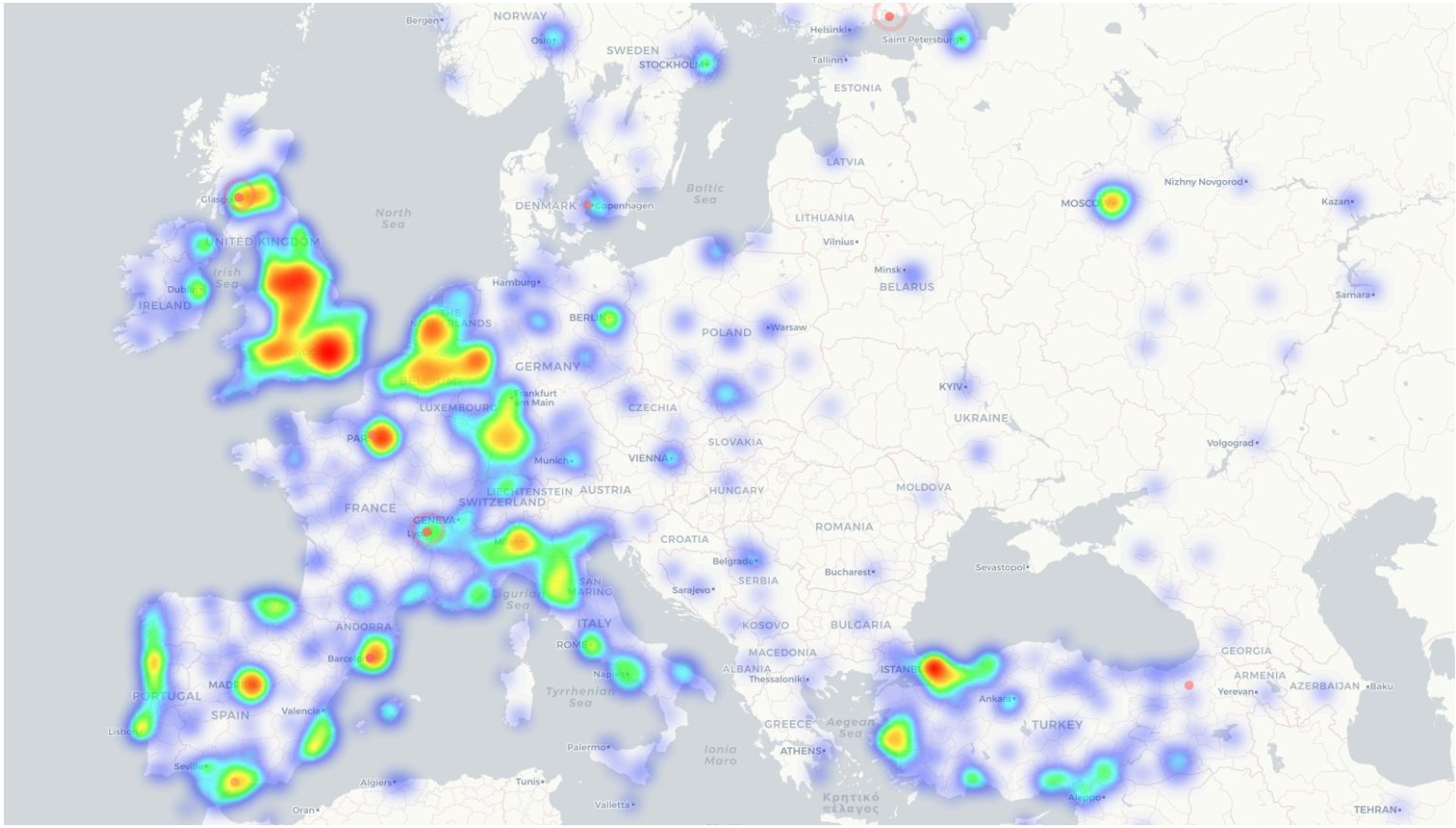


ctk.cz

Mediální archiv NEWTON

mediasearch.newtonmedia.eu

#onemilliontweetmap



onemilliontweetmap.com

Čísla mohou být matoucí

3	100	1619	1078	66,58	1078	1059	98,24	81	160	27	41	15	257	67	124	287
2	100	975	571	58,56	571	568	99,47	33	91	16	24	19	156	54	100	75
1	100	1305	877	67,2	876	873	99,66	51	136	30	31	19	269	55	194	88
1	100	1614	1043	64,62	1043	1041	99,81	85	192	23	49	23	288	56	168	157
2	100	2258	1579	69,93	1579	1567	99,24	124	222	43	38	21	417	72	274	356
89	100	81929	49213	60,07	49183	48852	99,33	2421	6938	902	1920	1156	11692	3047	7568	13208
1	100	1700	1096	64,47	1095	1094	99,91	53	183	33	42	19	298	80	192	194
7	100	936	599	64	599	599	100	36	117	19	23	12	200	43	104	45
1	100	482	336	69,71	336	333	99,11	20	67	10	9	2	81	32	45	67
1	100	2092	1401	66,97	1400	1387	99,07	69	235	36	53	30	357	94	212	301
9	100	1264	830	65,66	830	826	99,52	37	174	21	34	12	248	67	108	125
1	100	594	386	64,98	386	383	99,22	9	87	10	8	15	94	30	65	65
3	100	1588	927	58,38	926	921	99,46	35	170	32	34	22	256	74	208	90
3	100	2036	1361	66,85	1358	1352	99,56	74	238	22	49	12	328	82	190	357
4	100	1403	932	66,43	932	916	98,28	46	161	26	23	10	284	98	170	98
1	100	1268	806	63,56	806	800	99,26	58	152	29	26	12	232	49	134	108
1	100	1198	771	64,36	771	767	99,48	27	135	26	33	8	203	56	150	129
4	100	3579	2000	55,88	1996	1990	99,7	71	430	60	77	49	481	140	415	267
1	100	1243	795	63,96	795	790	99,37	44	132	21	20	20	224	52	113	164
1	100	490	317	64,69	317	313	98,74	25	68	20	12	5	70	22	40	51
2	100	1914	1226	64,05	1226	1222	99,67	78	237	32	36	20	302	89	203	225
1	100	606	409	67,49	409	406	99,27	39	70	18	7	6	110	38	86	32
1	100	474	300	63,29	300	296	98,67	7	83	11	8	7	71	36	45	28
1	100	678	466	68,73	466	465	99,79	32	116	7	12	10	114	21	89	64
1	100	152	95	62,5	95	95	100	1	18	4	1	2	26	6	27	10
17	100	8139	5003	61,47	4996	4967	99,42	284	833	143	163	100	1570	323	875	676
4	100	675	466	69,04	466	461	98,93	26	87	24	11	11	150	29	74	49
4	100	2631	1544	58,68	1544	1536	99,48	80	230	41	43	23	488	90	300	241
1	100	942	626	66,45	626	619	98,88	36	112	15	22	10	196	53	112	63
1	100	416	217	52,16	217	217	100	9	62	4	7	2	34	14	60	25
2	100	641	331	51,64	331	329	99,4	13	63	4	13	7	110	17	59	43

Měření centrální tendence a rozptylu

- měřítka centrální tendence
  - průměr
  - medián
  - modus
- měřítka rozptylu
  - rozpětí
  - rozptyl
  - směrodatná odchylka

Zápis dat

- číselná řada
  - $X = \{x_1, x_2, x_3, \dots, x_n\}$
  - např. $X = \{7,2,1, \dots, 46,35\}$
  - x_i je i-tý prvek množiny, např. $x_3 = 1$ (tj. zde i=3)
- suma
  - $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n$
  - např. $X = \{7,2,1, \dots, 46,35\}$
  - $\sum_{i=1}^{n} x_i = 7 + 2 + 1 + \cdots + 46 + 35$

Průměr (*mean*)

- „nejcennější“ střední hodnota shrnující velké množství dat do jednoho čísla
- průměr je využíván v dalších analýzách
- slabinou průměru je vliv extrémních hodnot na něj
- součet všech hodnot dělený jejich počtem (aritmetický průměr)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- vážený průměr
  - jednotlivé hodnoty x jsou vyváženy další proměnnou a proto mají ve výsledném průměru různou váhu

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i * x_i}{\sum_{i=1}^{n} w_i}$$

Medián (*median*)

- střední hodnota číselné řady
- nedokáže zachytit postavení extrémních pozorování
- v případě, že číselná řada má sudý počet prvků, mediánem je průměr dvou hodnot uprostřed

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

wikipedia.org

Modus (*mode*)

- nejčastěji se vyskytující hodnota v číselné řadě
- využíváme v případě, že nás zajímá nejvyšší četnost (například vyřčených slov v pořadu)
- pokud se všechny hodnot vyskytují v řadě jen jednou, modus zde není
  - 1,3,5,6,7
  - modus = Ø
- pokud se více hodnot objevuje stejně frekventovaně a přitom více než jednou, je zde více modů
  - 1,2,3,3,4,5,6,6,7,8,8
  - modus = 3,6,8

Příklad

- vypočtěte průměr, medián a modus pro následující číselnou řadu
  - $X = \{1,2,5,2,5,9\}$

Příklad

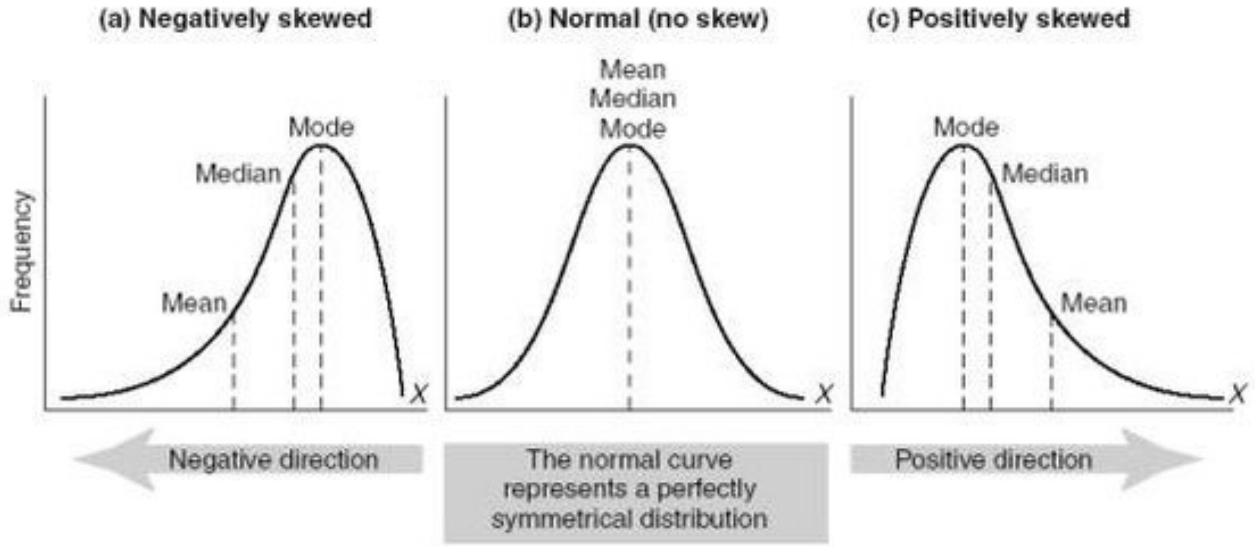
- vypočtěte průměr, medián a modus pro následující číselnou řadu
  - $X = \{1,2,5,2,5,9\}$
- průměr

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+5+2+5+9}{6} = \frac{24}{6} = 4$$

- medián
  - $X = \{1,2,\underline{2},\underline{5},5,9\}$ -> medián = 3,5
- modus
  - 2 a 5

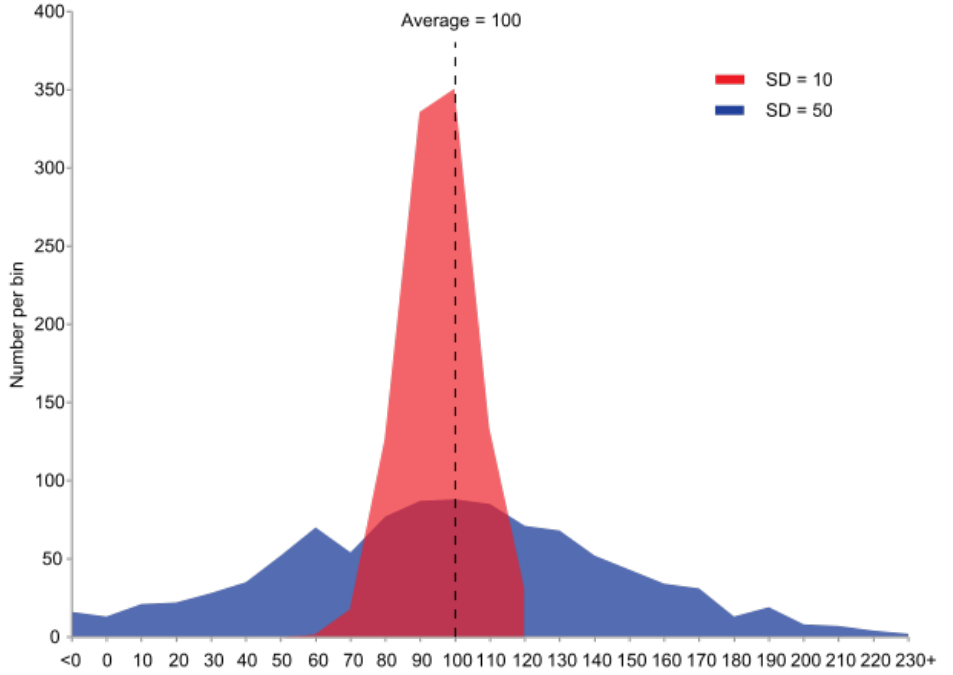
Měřítka centrální tendence

- v případě symetrického rozložení dat mají průměr, medián a modus identické hodnoty
- v případě asymetrického rozložení dat se liší a v tu chvíli nabývají o to více na významu při popisu rozložení dat
- pomocí tří čísel jsme schopni popsat dlouhou (a samu o sobě nepřehlednou) číselnou řadu



http://luigigallo.info

Rozpětí (*range*)



wikipedia.org

- určuje rozpětí hodnot (vzdálenost mezi nejmenší a největší hodnotou v číselné řadě)
- $R = x_{MAX} - x_{MIN}$

Rozptyl (*variance*)

- distribuce dat mohou mít stejný průměr, medián i modus, ale jejich rozložení přitom vypadá odlišně
- rozptyl určuje čtverec průměrné vzdálenosti hodnot od průměru, zatímco kontrolujeme počet pozorování
- výpočet rozptylu populace:
  - $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- POZOR, pokud byste někdy počítali rozptyl pro vzorek, dělitel je n-1 (týká se to pokročilejší úrovně statistiky)

Směrodatná odchylka (*standard deviation*)

- rozptyl je třeba standardizovat tak, aby jeho jednotka odpovídala jednotce původní proměnné v našich datech
  - ze čtverce se odmocninou dostaneme zpět na původní jednotky
- proto spočítáme odmocninu rozptylu a dostaneme směrodatnou odchylku:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Příklad

- výpočet rozptylu a směrodatné odchylky pro číselnou řadu populace:
  - $X = \{1,2,5,2,5,9\}$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N} = \frac{(1-4)^2+(2-4)^2+(5-4)^2+(2-4)^2+(5-4)^2+(9-4)^2}{6}$$
$$= \frac{9+4+1+4+1+25}{6} = \frac{44}{6} = 7{,}3$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}} = \sqrt{7{,}3} \cong 2{,}7$$

Shrnutí

- základem dobré analýzy je správný sběr dat a následná bezchybná práce s nimi
- existuje mnoho zdrojů dat a stále přibývají
- pro porozumění využíváme měřítka centrální tendence a další deskriptivní nástroje
- v rámci kvantitativního přístupu k analýze dat je zásadní uvědomovat si jeho výhody i nevýhody
- zároveň je více než žádoucí propojení s kvalitativním přístupem

Mnoho úspěchů v novém semestru ☺