

Úvod do kvantitativního výzkumu, měřítka centrální tendence a rozptylu

METODOLOGICKÝ PROSEMINÁŘ II

TÝDEN 5 | 21. BŘEZNA 2018

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 politologie
- je důležité být si vědom nedostatků a vyvažovat je využitím kvalitativních dat



juffrouj.wordpress.com

Typy proměnných

- nominální (nominal)
 - o dvou hodnotách lze říci pouze to, zda jsou stejné či různé
 - např. pohlaví, politická strana, typ politického režimu apod.

kvalitativní data

- ordinální (ordinal)
 - u hodnot je navíc možné určit jejich pořadí („co je více a co méně“)
 - např. úroveň dosaženého vzdělání, míra spokojenosti s politickým režimem apod.

lze užít jako oba typy dat

- intervalová (interval)
 - lze vypočítat, o kolik je jedna hodnota větší (menší) než druhá
 - např. míra růstu HDP, výše inflace apod.

kvantitativní data
- poměrová (ratio)
 - lze vypočítat, kolikrát je jedna hodnota větší (menší) než druhá – existuje absolutní nula, pouze kladné hodnoty
 - např. počet hlasů získaných ve volbách, počet poslanců hlasujících pro zákon apod.

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

The screenshot shows the website of the Czech Statistical Office (ČSÚ) for the 2011 census results. The page is titled "Sčítání lidu, domů a bytů" (Census of population, households and dwellings). The navigation bar includes "Statistiky", "Vydáváme", "Databáze, registry", "Klasifikace, číselníky", "Výkazy, sběr dat", and "O ČSÚ". The main content area is divided into three columns: "Data", "Analýzy, komentáře", and "Související informace". The "Data" column lists various publications and databases. The "Analýzy, komentáře" column lists analyses and reports. The "Související informace" column lists related information. A sidebar on the right titled "Průřezové statistiky" (Cross-sectional statistics) lists various statistical categories.

ČESKÝ STATISTICKÝ ÚŘAD

Kontakty Odkazy Časté dotazy Nápověda

Statistiky Vydáváme Databáze, registry Klasifikace, číselníky Výkazy, sběr dat O ČSÚ

Úvod > Statistiky > Sčítání lidu, domů a bytů Vytisknout

Sčítání lidu, domů a bytů

f t e +

Data

- > [Výsledky SLDB 2011 \(VDB\)](#)
- > [Publikace SLDB 2011](#)
- > [Databáze SLDB 2011 na CD/DVD](#)
- > [Applikace Census Hub](#)
- > [Publikace SLDB 2001](#)
- > [Historie sčítání \(včetně vybraných dat\)](#)
- > [Historický lexikon obcí České republiky - 1869 - 2005](#)
- > [Historický lexikon obcí České republiky - 1869 - 2011](#)
- > [Výsledky minulých sčítání](#)

Analýzy, komentáře

- > [Analýzy SLDB 2011](#)
- > [Analýzy SLDB 2001](#)
- > [Tiskové zprávy](#)

Související informace

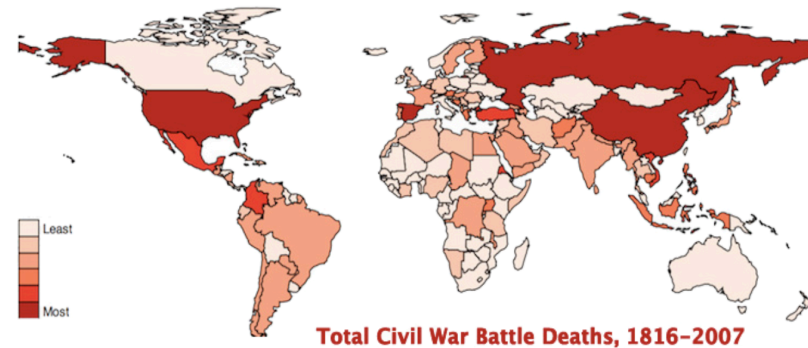
- > [Webové stránky SLDB 2011](#)
- > [Otevřená data SLDB 2011](#)
- > [Sčítání v zahraničí](#)

Průřezové statistiky

- > [Cizinci](#)
- > [Genderové statistiky](#)
- > [Senioři](#)
- > [Souhrnná data o ČR](#)
- > [Regionální statistiky](#)
- > [Makroekonomické údaje](#)
- > [Mezinárodní data](#)

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

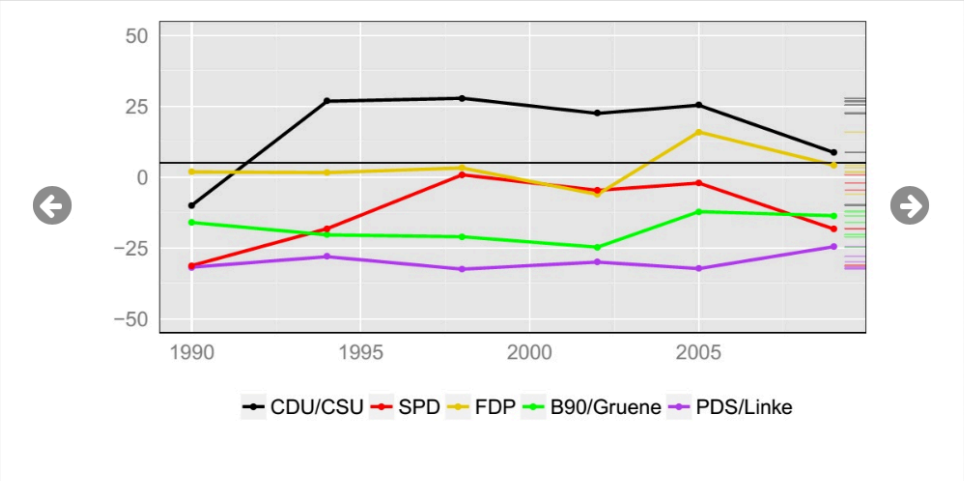
Manifesto Project

INFO DATA CORPUS & DOCUMENTS EXPLORE manifestoR
manifestata
API PUBLICATIONS   LOGIN
SIGNUP

Project description

The Manifesto Project provides the scientific community with parties' policy positions derived from a content analysis of parties' electoral manifestos. It covers over 1000 parties from 1945 until today in over 50 countries on five continents. The DFG-funded MARPOR project continues the work of the Manifesto Research Group (MRG) and the Comparative Manifestos Project (CMP). On this website you find the Manifesto Project Dataset containing the parties' policy preferences generated by the project. You also find coded and uncoded election manifestos of the parties in the dataset as well as information and links to many applications for the dataset, related projects and publications etc.

[\[+ read more\]](#)



Year	CDU/CSU	SPD	FDP	B90/Gruene	PDS/Linke
1990	-10	-35	5	-15	-30
1995	25	-15	5	-20	-25
2000	25	5	5	-20	-30
2005	25	-5	15	-15	-30
2017	10	-15	5	-15	-25

Left Right of German Parties

manifesto-project.wzb.eu

Chapel Hill Expert Survey

1999-2014 Chapel Hill Expert Survey (CHES) trend file

DESCRIPTION

The Chapel Hill expert surveys estimate party positioning on European integration, ideology and policy issues for national parties in a variety of European countries. The first survey was conducted in 1999, with subsequent waves in 2002, 2006, 2010, and 2014. The number of countries increased from 14 Western European countries in 1999 to 24 current or prospective EU members in 2006 to 31 countries in 2014. In this time, the number of national parties grew from 143 to 268. The 2014 survey includes all EU member states, plus parties in Norway, Switzerland, and Turkey. Separate surveys were conducted in the Balkan candidate countries. Questions on parties' general position on European integration, several EU policies, general left/right, economic left/right, and social left/right are common to all surveys. More recent surveys also contain questions on non-EU policy issues, such as immigration, redistribution, decentralization, and environmental policy.

DATASETS

- [Dataset \(STATA\)](#) (in [csv](#))
- [Codebook](#)

HOW TO CITE

When using the 1999-2014 trend file, cite the following:

chesdata.eu

European Election Studies

The screenshot shows the homepage of the European Election Studies (EES) website. At the top, there is a dark grey navigation bar with the EES logo on the left, the text "European Election Studies" in the center, and icons for CONTACT, SITEMAP, TWITTER, GESIS, and SEARCH on the right. Below this is a white horizontal menu with links for Home, European Election Studies, EES Study Components, Bibliography, Blog, and News. The main content area has a teal background. It features a title "The European Election Studies (EES)" followed by a paragraph describing the studies' focus on electoral participation, EU political community, and public sphere. Below this are four white boxes, each representing a study component: Voter Study, Euromanifesto Study, Elite Study, and Media Study. Each box contains a brief description and a button with the study name and a right-pointing arrow.

ees european election studies

European Election Studies

CONTACT SITEMAP TWITTER GESIS SEARCH

Home European Election Studies EES Study Components Bibliography Blog News

The European Election Studies (EES)

are about electoral participation and voting behaviour in European Parliament elections, but more than that. They are also concerned with the evolution of an EU political community and a European public sphere, with citizens' perceptions of and preferences about the EU political regime, and with their evaluations of EU political performance. European Election Studies include more components than the postelection surveys among representative samples of voters. These additional components include content analyses of party manifestos ("Euromanifestos"), elite surveys and content analyses of media news.

Voter Study
The Voter Study consists of nationally representative post-election surveys.
[Voter Study](#) >

Euromanifesto Study
The Euromanifestos Study analyses the manifestos issued ahead of EP elections.
[Euromanifesto Study](#) >

Elite Study
The Elite Study Every involved a comprehensive study of political elites.
[Elite Study](#) >

Media Study
The media study content-analyses main broadcaster in the EU member states.
[Media Study](#) >

<http://europeanelectionstudies.net>

ParlGov



Explore

Data ▾

Documentation ▾

Parliaments and governments database

Project description

ParlGov is a data infrastructure for political science and contains information for all EU and most OECD democracies (*37 countries*). The database combines approximately *1500 parties*, *920 elections* (8400 results), and *1400 cabinets* (3500 parties).

Examples

Party: [Deu SPD](#) —

Election: [Aut 1990](#) — · (EP [Gbr 1989](#))

Cabinet: [Jpn 2009](#) —

Information

[Credits](#)

[FAQ](#)

[Data](#) · 2015

Latest news

09 Sep — [Snippet · cabinets party ...](#)

30 Aug — [ParlGov snippets on Github](#)

06 Aug — [Romania cabinet added](#) ·

Activity

[All news](#)

[Twitter](#)

<http://www.parlgov.org>

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

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
 - $\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \dots + x_{n-1} + x_n$
 - např. $X = \{7, 2, 1, \dots, 46, 35\}$
 - $\sum_{i=1}^n x_i = 7 + 2 + 1 + \dots + 46 + 35$

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 = $(4 + 5) \div 2$

= **4.5**

wikipedia.org

Modus (*mode*)

- nejčastěji se vyskytující hodnota v číselné řadě
- využíváme v případě, že nás zajímá nejvyšší četnost (to není často, ale příkladem je třeba výsledek volebního systému first-past-the-post)
- pokud se všechny hodnot vyskytují v řadě jen jednou, modus zde není
 - 1,3,5,6,7
 - modus = \emptyset
- 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

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

Příklad

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

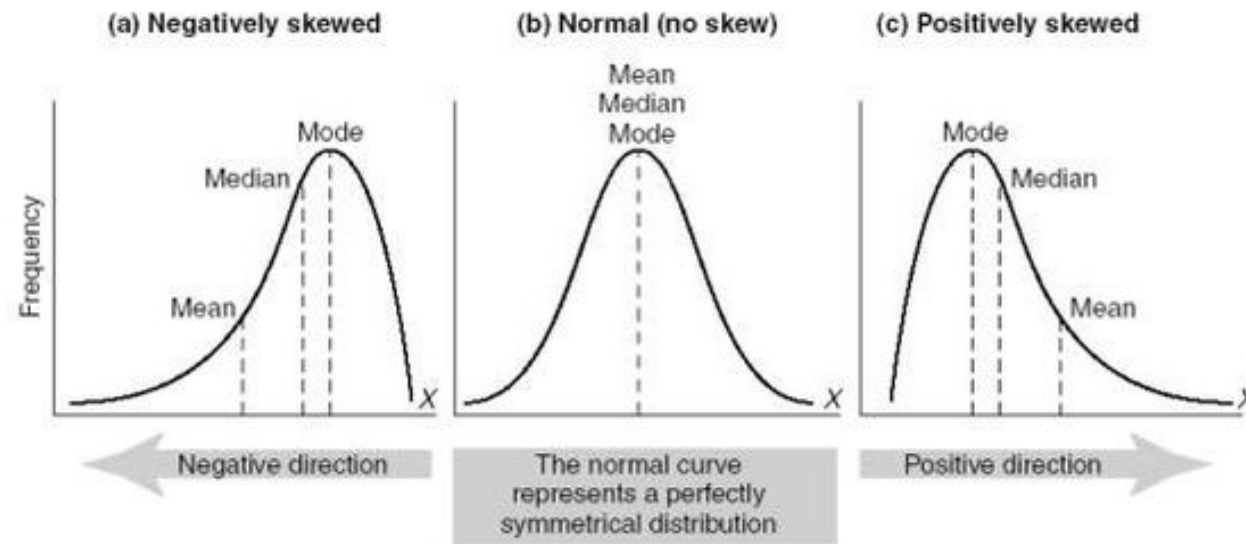
- průměr

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

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

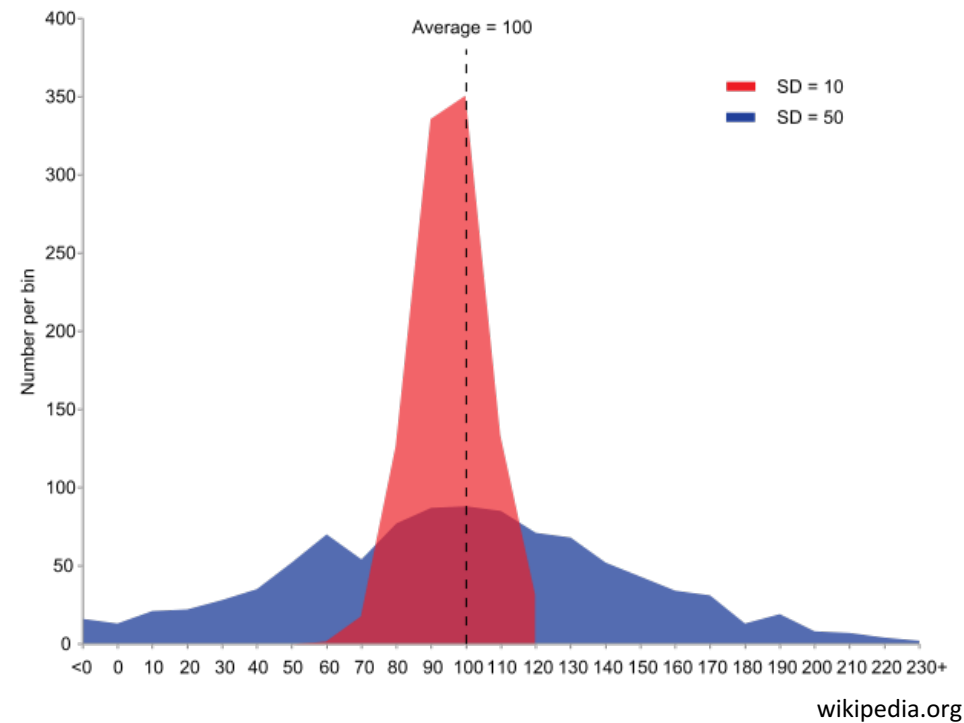
Měřítko centrální tendence

- 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*)



- 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

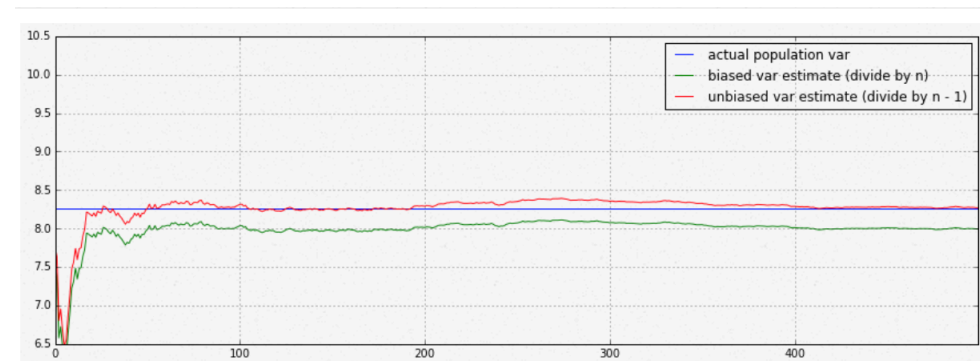
- výpočet rozptylu pro celou populaci

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

- výpočet rozptylu pro vzorek populace

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

- Proč v případě vzorku populace používáme „n-1“? Hodnoty x_n jsou blíže průměru vzorku \bar{x} než průměru populace μ . Z toho důvodu je využita korekce $n - 1$, aby se s^2 stal nezaujatým odhadem σ^2 .



mortada.net

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

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

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

Příklad

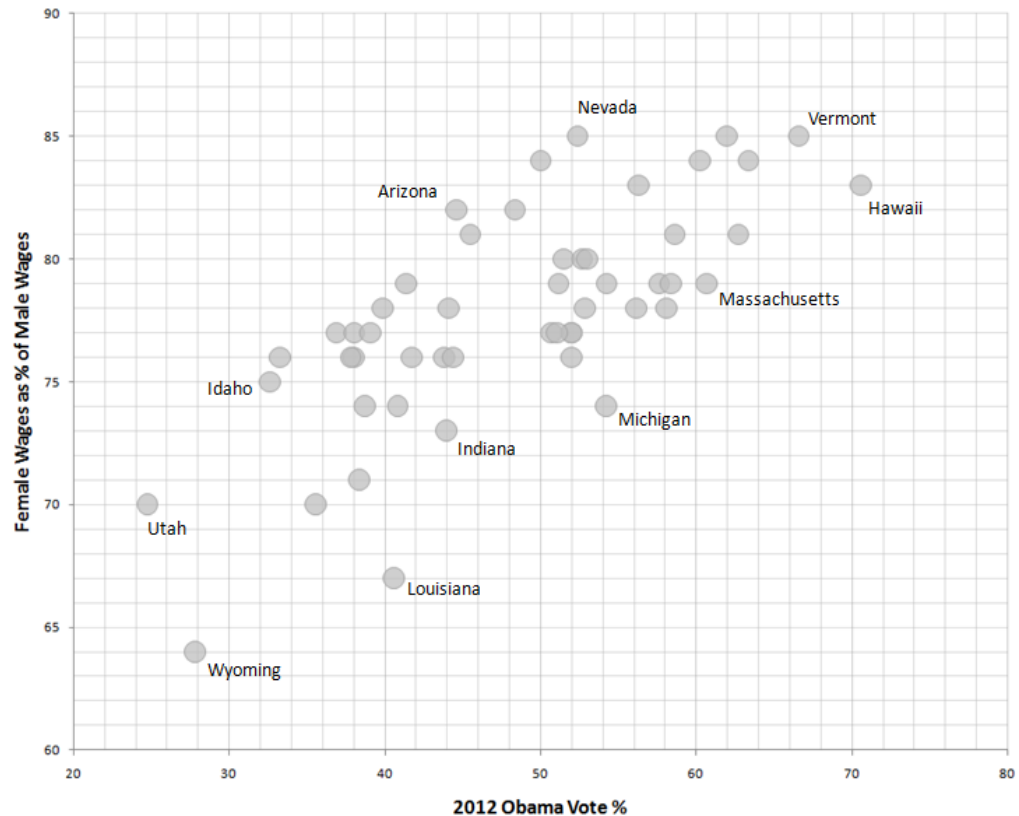
- vypočtěte rozptyl a směrodatnou odchylku pro následující číselnou řadu vzorku populace
 - $X = \{1, 2, 5, 2, 5, 9\}$

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{(1 - 4)^2 + (2 - 4)^2 + (5 - 4)^2 + (2 - 4)^2 + (5 - 4)^2 + (9 - 4)^2}{6 - 1} \\ &= \frac{9 + 4 + 1 + 4 + 1 + 25}{5} = \frac{44}{5} = 8.8 \end{aligned}$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} = \sqrt{8.8} \cong 2.97$$

Zobrazení kvantitativních dat - bodový graf (*scatter plot*)

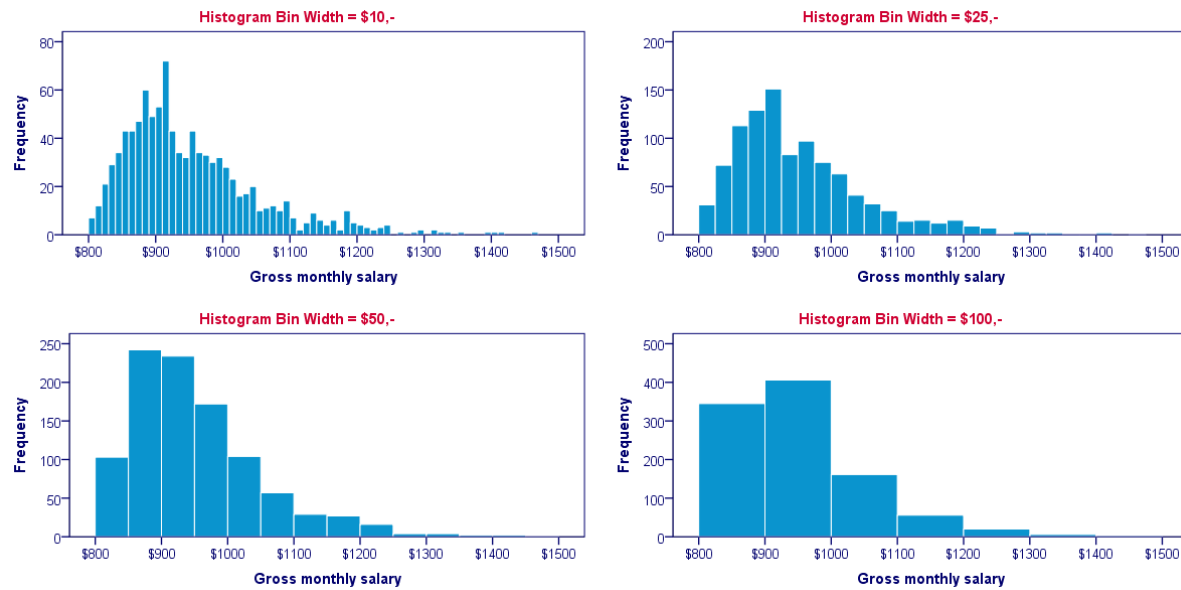
- přímé zobrazení pozorování ve 2D prostoru
- nejužitečnější ze všech statistických zobrazení
- na horizontální ose x je zobrazena nezávisle proměnná X
- na vertikální ose y je zobrazena závisle proměnná Y
- je možné pozorovat vztahy mezi proměnnými



justinholman.com

Zobrazení kvantitativních dat - histogram

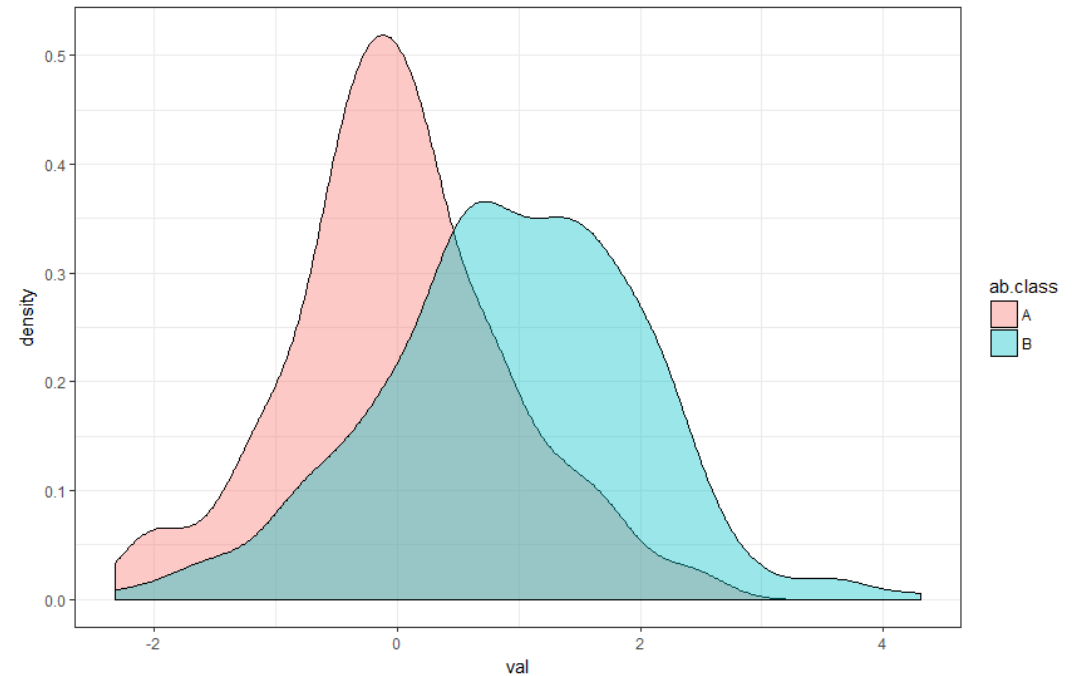
- ukazuje distribuci hodnot proměnné
- je to sloupcový graf, jehož výška ukazuje, kolik pozorování spadá do dané kategorie
- jednotlivé kategorie mohou být různě detailní (zobrazení vybíráme podle situace)



spss-tutorials.com

Zobrazení kvantitativních dat - hustota rozložení (*density plot*)

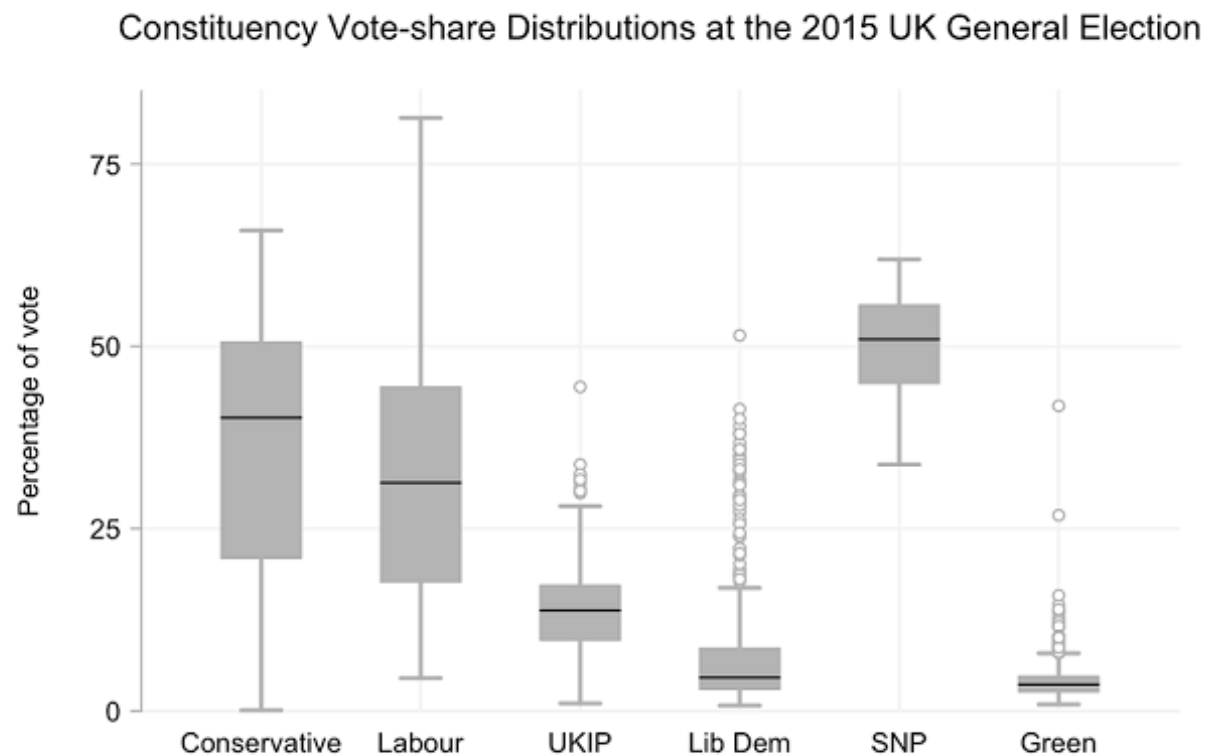
- rozložení hodnot také můžeme zakreslit pomocí zprůměrování a zahlazení jednotlivých frekvencí z histogramu
- grafy hustoty rozložení jsou zvláště cenné, když se zajímáme o rozložení pravděpodobností výskytu
- na ose x jsou jednotlivé nabývané hodnoty
- osa y je obtížně interpretovatelná (záleží na konkrétní situaci) – zásadní je všimnout si tvaru, vrcholů apod.



stackoverflow.com

Zobrazení kvantitativních dat - krabicový diagram (*boxplot*)

- komplexní zobrazení distribuce hodnot proměnné
- krabice vyznačuje první, druhý (medián; černá čára) a třetí kvartil
- „vousy“ (šedé vertikální čáry s horizontálním ukončením) vyznačují krajní hodnoty mezi horním, respektive dolním kvartilem, a 1,5násobkem vzdálenosti mezi prvním a třetím kvartilem
- odlehlé hodnoty jsou od horního či dolního kvartálu vzdálený více než 1,5násobku vzdálenosti mezi prvním a třetím kvartilem (vyznačeny prázdnými body)



infragistics.com

Shrnutí

- 4 typy proměnných – nominální, ordinální, intervalové a poměrové
- existuje mnoho zdrojů kvantitativních dat
- pro zjednodušený popis využíváme měřítka centrální tendence a rozptylu
- pozor u měřítek rozptylu na to, kdy pracujeme s celou populací a kdy jenom se vzorkem
- různé možnosti grafického zobrazení kvantitativních dat