Statistické metody a zpracování dat

II. Popisné statistické metody

Petr Dobrovolný

Popisné (deskriptivní) metody

Činíme závěry pouze z určitého zpracovávaného souboru – **výběrového**, popisujeme jen to, co bylo zjištěno, bez zobecňování

Deskriptivní metody:

1. přehledné vyjádření výsledků pomocí **četnostních tabulek** a **grafů**
2. Výpočty a grafické znázornění základních **popisných statistických** charakteristik

Popisná statistika společně s tzv. **explorační (průzkumovou) analýzou dat** obvykle tvoří počátek vlastní statistické analýzy.

Rozdělení četností

• Statistické údaje jednotlivých statistických souborů pro další zpracování uspořádáváme

• U jednotek statistického souboru můžeme na základě kvantitativních hodnot zjišťovat jejich **četnost – frekvenci**.

• Četnost - počet prvků se stejnou hodnotou statistického znaku

• Používáme ho pro **nespojité znaky** a při malém počtu variant (počet členů domácnosti).

Příklad:

U 20 náhodně vybraných domácností byl sledován počet členů domácnosti: 1,3,4,3,4,3,3,2,2,1,1,2,2,1,4,5,4,3,2,2

Počet členů	1	2	3	4	5
četnost	4	6	5	4	1

Skupinové rozdělení četností

• Pro spojité znaky udáváme počet prvků s hodnotami znaku patřícími do určitého intervalu (třídy).

• Jednotky statistického souboru roztřídíme podle velikosti do několika intervalů.

• Dolní a horní hranice (mez) intervalu udává, jakou nejmenší a největší hodnotu znaku do daného intervalu zařadíme.

• Délka či **šířka intervalu** je kladný rozdíl dvou po sobě následujících dolních (horních) mezí.

• Krajní interval může být otevřený (neuzavřený).

• U skupinového rozdělení četností zastupuje hodnoty znaku **střed intervalu** (x_s).

Skupinové rozdělení četností

Interval hodnot znaku	Střed intervalu x_s	Četnosti			
		absolutní n_i	relativní F_i	kumulované abs. N_i	kumulované rel. F_i
1 - 5	3	1	0,02	1	0,02
6 - 10	8	3	0,06	4	0,08
11 - 15	13	11	0,22	15	0,30
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
41 - 45	43	5	0,10	50	1,00
Σ		50	1,00		

Zásady pro stanovení hranic intervalů:

• každý interval je určen horní a dolní hranicí

• každý interval musí být vymezen tak, abychom mohli každý prvek jednoznačně zařadit

• intervaly se nesmí překrývat

• má-li být rozdělení četností použito k výpočtu dalších statistik, musí mít intervaly stejnou šířku

• šířka intervalu nesmí být velká – aby nesetřela zvláštnosti rozdělení hodnot, ale ani malá – aby nevzniklo více intervalů s nulovou četností (optimum 5 – 20).

• počty intervalů (m) lze určovat subjektivně i pomocí vzorců:

$$m \approx \sqrt{n} \qquad m \leq 5 \cdot \log n$$

Sturgesovo pravidlo $m = 1 + 3{,}3 \log_{10}(n)$

	A	B
1	Rok	td
2	1771	8,4
3	1772	10,9
4	1773	10,0
5	1774	10,2
6	1775	10,7
7	1776	8,8
8	1777	8,9
9	1778	10,2
10	1779	10,4
11	1780	8,9
12	1781	10,3
13	1782	9,0
14	1783	10,1
15	1784	8,4
16	1785	7,9
17	1786	7,4
18	1787	10,2
19	1788	9,8
20	1789	10,2
21	1790	10,2
22	1791	11,1
23	1792	10,3
24	1793	10,6

Četnosti

- absolutní
- relativní
- kumulované

Interval hodnot			Četnost		Kumulovaná	
dolní mez	horní mez	střed	absolutní	relativní	absolutní	relativní
7,01	7,50	7,25	6	0,027	6	0,027
7,51	8,00	7,75	7	0,032	13	0,059
8,01	8,50	8,25	22	0,100	35	0,158
8,51	9,00	8,75	33	0,149	68	0,308
9,01	9,50	9,25	41	0,186	109	0,493
9,51	10,00	9,75	49	0,222	158	0,715
10,01	10,50	10,25	40	0,181	198	0,896
10,51	11,00	10,75	15	0,068	213	0,964
11,01	11,50	11,25	8	0,036	221	1,000
Suma			221	1		

Vícerozměrné rozdělení četností

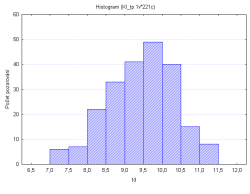
- třídění se realizuje podle dvou či více znaků
- tzv. **kombinační tabulka**
- slouží ke zkoumání závislostí studovaných znaků (**korelační tabulka**)
- pokud znaky nabývají pouze dvou hodnot - **asociační tabulka**

	Inteligenční kvocient – IQ							Součet Σ
	95	98	100	105	110	112	120	
Obratnost 15		1						1
16		1						1
18	1	1						2
20	1	2	3	4				10
22			3	3	1			7
25				1	3	1		5
26				1	2	1		4
32						1		1
37							1	1
Σ	2	5	6	9	6	3	1	32

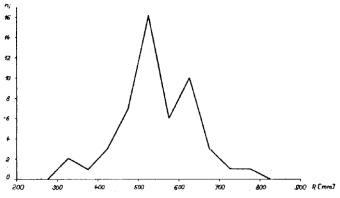
Grafické znázornění rozdělení četností

- Pravoúhlá soustava souřadnic, osa x – intervaly hodnot znaku, osa y – četnosti hodnot
- **Histogram** – typ sloupkového diagramu
- **Polygon** – spojnicový diagram
- **Čára kumulovaných četností** – součtová čára, četnosti vynášíme k horní hranici intervalu
- Graf relativních kumulovaných četností umožňuje odvození kvantilů

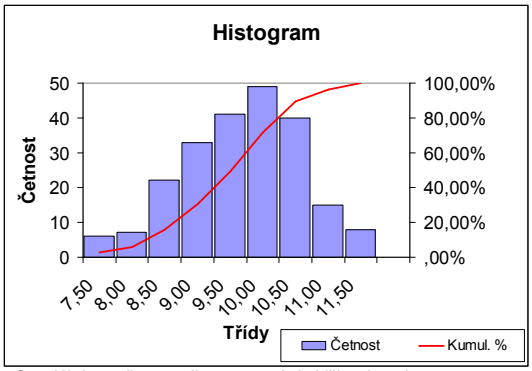
Histogram



Polygon



Součtová čára



Speciální typy četnostního zpracování - Věková struktura obyvatel (strom života)

Popisná statistika

K čemu je to dobré?

- jednoduše popsat chování statistického souboru dat (kondenzace dat)
- porovnat více souborů mezi sebou

Jednoduchý příklad: Vystihnout průměrnou teplotu vzduchu lokality za určité období

Složitý příklad: Vystihnout průměrné chování lidí nakupujících v určitém supermarketu

Základní statistické charakteristiky

- Charakteristiky úrovně
- Charakteristiky variability
- Charakteristiky asymetrie
- Charakteristiky špičatosti

Výchozí data – způsob výpočtu

- z reálných hodnot
- ze skupinového rozdělení četností (reálné hodnoty seskupené do intervalů)

Charakteristiky úrovně

(střední hodnoty, míry polohy, míry centrální tendence)

Jedná se o čísla, která reprezentují jednotlivé hodnoty statistického znaku, udávají polohu, charakterizují obecnou velikost jevu.

Aritmetický průměr – úhrn hodnot kvantitativního statistického znaku dělený rozsahem souboru. Statistický znak X nabývá hodnot $x_1, x_2, \ldots x_n$. Aritmetický průměr bude:

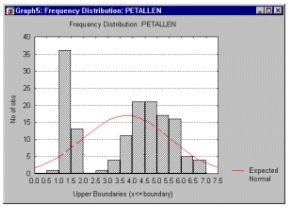
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Vlastnosti aritmetického průměru

- součet kladných odchylek se rovná součtu odchylek záporných
- suma čtverců odchylek od průměru je vždy menší než suma čtverců odchylek od jakékoliv jiné hodnoty
- přičteme-li ke všem hodnotám znaku konstantu, průměr se zvětší o tuto konstantu
- znásobí-li se všechny hodnoty znaku konstantou k , průměr se k-krát zvětší
- průměr součtu dvou proměnných se rovná součtu obou průměrů

Vlastnosti aritmetického průměru

- Geometricky si lze aritmetický průměr představit jako těžiště.
- Průměr musí být **typický** (většina hodnot je blízká průměru).
- Typický je tehdy, blíží-li se nejčetnější hodnotě.



- Aby aritmetický průměr vhodně vystihoval úroveň studovaného souboru rozdělení hodnot znaku musí být jednovrcholové.
- Aritmetický průměr má smysl jen tehdy, jestliže má nějaký smysl součet hodnot.
- Průměr, pokud je uvedený samotný, může být silně zavádějící.

Aritmetický průměr

Skládá-li se soubor z k skupin o rozsazích n_i s průměry $\bar{x}_i$ platí pro celkový průměr souboru:

$$\bar{x} = \frac{\sum_{i=1}^{k} \bar{x}_i n_i}{\sum_{i=1}^{k} n_i}$$

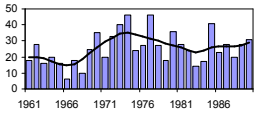
Vážený aritmetický průměr

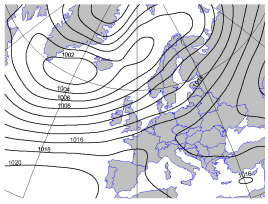
$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots x_k n_k}{n_1 + n_2 + n_3 + \ldots n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$$

Vážený aritmetický průměr

Příklady použití:

- k výpočtu aritmetického průměru z rozdělení četností
- shlazování časových řad
- výpočet množství studovaného prvku v ploše (váha – plocha území v rozmezí intervalu izolinií
- výpočet průměrné denní teploty vzduchu





Geometrický průměr

n-tá odmocnina součinu z řady hodnot znaku. Používá se u souborů, jejichž hodnoty tvoří geometrickou posloupnost.

Prostý geometrický průměr $\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots x_n}$

Vážený geometrický průměr $\bar{x}_{gv} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \cdot \ldots x_n^{n_n}}$

Použití:
- počítá se pouze z hodnot, které jsou kladné
- v případě, kdy má smysl součin hodnot studovaného jevu
- k určení tzv. tempa růstu v časových řadách.
- obvykle se používá pro veličiny měřené na logaritmické stupnici.

Geometrický průměr - příklad

Růst cen určitého zboží byl postupně 20 %, 10 %, poté 15 % pokles a 10 % růst.

Potom průměrný růst je roven $(1{,}20 \cdot 1{,}10 \cdot 0{,}85 \cdot 1{,}10)^{1/4} \cong 1{,}054$, tzn. průměrný růst je přibližně 5,4 %.

Koeficienty růstu produkce závodu pro jednotlivá období:

období	roční koef. růstu	počet roků (n_i)
1996-2001	1,04	5
2002/2001	1,07	1
2002-2005	1,05	3
2006/2005	1,04	1
Σ	x	10

$$\bar{x}_{gv} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \cdot \ldots x_n^{n_n}} = \sqrt{1{,}04^5 \cdot 1{,}07^1 \cdot 1{,}05^3 \cdot 1{,}04^1} = 1{,}046$$

Průměrný koeficient růstu produkce závodu za posledních 10 roků je 4,6%

Geometrický průměr - příklad použití:

Nalezení průměrného přírůstku obyvatel, kdy populace na určité ploše roste geometricky

časový okamžik	počet jedinců
t1	3 000
t2	9 000
t3	27 000

Geometrický průměr je vhodný pro použití v situacích, když je rozdělení hodnot asymetrické a logaritmická transformace jej opět vrací k symetrii.

Harmonický průměr

Počet jednotek souboru dělený součtem reciprokých hodnot. Používá se pro charakterizování průměrné rychlosti změny – k popisu intenzitních ukazatelů.

Prostý harmonický průměr $\bar{x}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

Vážený harmonický průměr $\bar{x}_{hv} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

Používá se tam, kde má smysl sčítat převrácené hodnoty.

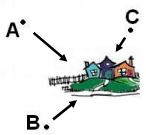
Harmonický průměr – příklady použití

Výpočet celkové průměrné rychlosti dojíždějících do centra.

Vzhledem k rozdílné dopravní propustnosti, průměrná rychlost se výrazně mění na jednotlivých úsecích cesty.

K výpočtu celkové průměrné rychlosti je pak vhodnější využít harmonického průměru

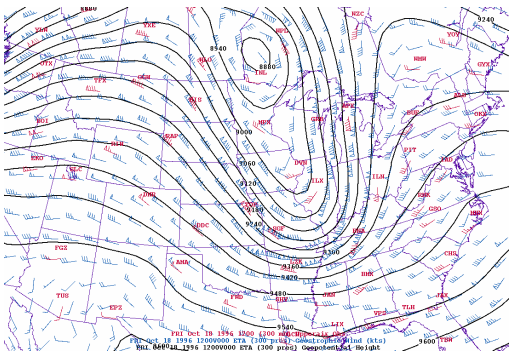
Dostupnost místa:
- z bodu A…….. 30 min.
- z bodu B…….. 20 min.
- z bodu C…….. 6 min.



$$\bar{x}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots \frac{1}{x_n}} = \frac{3}{\frac{1}{3} + \frac{1}{2} + \frac{1}{10}} = \frac{3}{3+2+10} = \frac{3}{15} = \frac{1}{5} = 12 \text{ min}$$

Harmonický průměr – příklady použití

Příklad 2: Určení průměrné rychlosti tzv. geostrofického větru ze vzdáleností dvou izobar



Kvadratický průměr

Prostý kvadratický průměr $\bar{x}_k = \sqrt{\frac{x_1^2 + x_2^2 + x_3^2 + \ldots x_n^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$

Vážený kvadratický průměr $\bar{x}_{kv} = \sqrt{\frac{x_1^2 n_1 + x_2^2 n_2 + x_3^2 n_3 + \ldots x_n^2 n_k}{n_1 + n_2 + \ldots n_k}} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{\sum_{i=1}^{k} n_i}}$

Nahrazuje individuální hodnoty řady tak, že se nemění součet jejich čtverců

Pokud hodnoty znaku x nejsou stejné, potom platí:

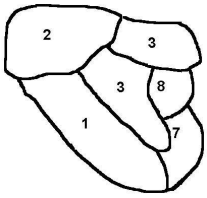
$$\bar{x}_h < \bar{x}_g < \bar{x} < \bar{x}_k$$

Modus $\hat{x}$

- Nejčetnější (typická) hodnota kvantitativního znaku studovaného souboru
- U rozdělení četností – modální interval závisí na šířce intervalů (subjektivní vliv – modus je nestabilní hodnota).
- V grafu frekvenční funkce je modus hodnota, ve které tato dosahuje vrcholu.
- Má velký význam u nespojitých veličin a u kvalitativních znaků. Umožňuje popisovat nominální data (Auto je nejčastěji využívaným dopravním prostředkem).

Modus - příklad použití:

Určení dominantní třídy v rámci studované plochy



Aritmetický průměr: 4

Modus: 3

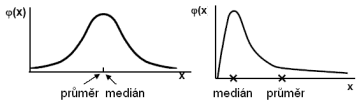
Modus - vlastnosti:

- Některá rozdělení mohu mít více modů – např. bimodální. Takovéto soubory mají dva mody. A nebo žádná hodnota nemusí dominovat.
- Výhodné je použití modu při porovnání souborů, pokud jde o typické hodnoty znaku.
- Výpočet modu z rozdělení četností:

$$\hat{x} = L + h\frac{n_2}{n_1 + n_2}$$

kde L je dolní hranice modálního intervalu, h je šířka modálního intervalu n_1 je četnost intervalu předcházejícího před modálním intervalem a n_2 četnost intervalu následujícího za modálním

Medián $\tilde{x}$



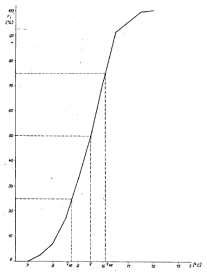
- Medián je prvek řady, uspořádané v neklesajícím pořadí, který ji dělí tak, že polovina prvků má hodnotu větší, druhá polovina větší, než je hodnota mediánu.
- Medián není ovlivněn extrémními hodnotami, ale jejich počtem.
- Porovnáním mediánu dvou souborů lze získat informaci o tendenci k vyššímu (nižšímu) výskytu extrémních hodnot.
- Někdy lépe charakterizuje úroveň souboru než průměr.
- Lze ho stanovit z řady uspořádaných hodnot a nebo ho určit z rozdělení četností.

Kvantily

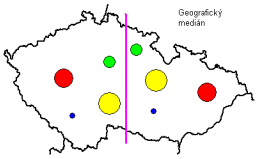
- Medián dělí statistický soubor na poloviny.
- Analogickým dělením souboru na více částí získáme **kvantily** (kvartily, decily percentily)

Dolní kvartil $\tilde{x}_{25}$

Horní kvartil $\tilde{x}_{75}$



Medián i kvantity lze snadno určit z čáry kumulovaných četností



Geografický medián – linie rozdělující plochu, na níž se vyskytuje studovaný jev na dvě části, tak aby hodnota jevu byla v obou částech stejná.

Aritmetický střed

• Aritmetický průměr min. a max. hodnoty znaku.

• Extrémy se často značně liší od ostatních hodnot – jsou netypické, často nahodilé, mají však význam samy o sobě.

$$x_{st} = \frac{x_{\max} + x_{\min}}{2}$$

Useknutý (trimmed) průměr

$$\tilde{u}_T = \frac{\tilde{u}_{0,25} + 2 \cdot \tilde{u}_{0,5} + \tilde{u}_{0,75}}{4}$$

Použití měr centrální tendence

Aritmetický průměr použijeme:

- pro data intervalová a poměrová, ne pro data kategoriální
- je-li rozdělení symetrické
- hodláme-li použít statistických testů

Medián použijeme v případech, kdy:

- data jsou získána minimálně v ordinálním měřítku
- chceme znát střed rozdělení dat
- data mohou obsahovat odlehlé hodnoty
- je-li rozdělení silně zešikmené

Modus použijeme v případech, kdy:

- data jsou získána minimálně v ordinálním měřítku
- má-li rozdělení více vrcholů
- chceme-li o rozdělení získat jen základní přehled
- míníme-li slovem „průměrný“ nejčastější hodnotu

Kritéria pro výběr nejvhodnější míry úrovně

Závisí na těchto faktorech

- vlastnostech použité míry úrovně
- typu řešené úlohy
- typu rozložení dat

ČLOVĚK S PRŮMĚRNÝM PLATEM V PRŮMĚRNÉM BYTĚ V PRŮMĚRNÉM TÝDNU ŠIROKO DALEKO NEEXISTUJE.

Omezení měr úrovně

Omezení spočívají v porovnávání průměrů dvou výběrových souborů bez ohledu na tvar rozložení.

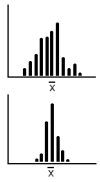
Dva soubory se shodnou hodnotou aritmetického průměru mohu mít zcela odlišné rozložení hodnot.

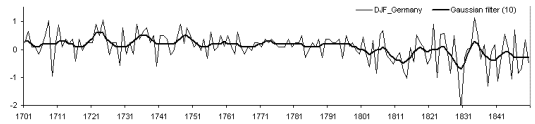
Je nutné uvažovat také charakteristiky popisující míry proměnlivosti a koncentrace kolem střední hodnoty

Charakteristiky variability

• Popisují **stupeň proměnlivosti** statistického znaku v daném statistickém souboru.

• Vypovídají také o tom, jak dobře vystihuje použitá míra úrovně jednotlivé hodnoty souboru.





Míry variability

- založené na vybraných hodnotách znaku v souboru
- založené na všech hodnotách znaku v souboru

Charakteristiky variability

Variační rozpětí $R = x_{\max} - x_{\min}$

Kvantilové odchylky – kladné odchylky jednotlivých kvantilů (kvartilová, decilová, percentilová odchylka).

Kvartilová odchylka $Q = \frac{(\tilde{x}_{75} - \tilde{x}) + (\tilde{x} - \tilde{x}_{25})}{2} = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{2}$

Variační rozpětí a kvantilové odchylky nejsou založeny na všech hodnotách studovaného souboru – neberou tedy ohled na rozdělení hodnot

Průměrné odchylky

• Jsou definovány jako aritmetický průměr absolutních odchylek jednotlivých hodnot znaku od střední hodnoty.

• Absolutní hodnota odstraňuje kompenzaci kladných a záporných odchylek.

• Ukazují na odlišnost prvků od střední hodnoty.

Průměrná odchylka od průměru $\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$

Z rozdělení četností se průměrná odchylka od průměru počítá formou váženého aritmetického průměru absolutních odchylek – jako váhy se používají četnosti n_i:

$$\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| \cdot n_i}{\sum_{i=1}^{k} n_i}$$

Střední diference

• Aritmetický průměr absolutních hodnot všech možných vzájemných rozdílů *n* jednotlivých hodnot studovaného znaku *x*.

• Je vhodnou mírou variability znaku u souborů s malým rozsahem.

$$\Delta = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{n(n-1)}$$

Nejpoužívanější míry variability jsou založeny na všech hodnotách souboru

Rozptyl s^2

Je definován jako průměr ze čtverců odchylek jednotlivých hodnot znaku od jejich aritmetického průměru:

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

Rozptyl měří velikost proměnlivosti, avšak v jednotkách čtverců odchylek.

Výpočet rozptylu ze skupinového rozdělení četností:

$$s^2 = \frac{\sum_{i=1}^{k}(x_s - \bar{x})^2 \cdot n_i}{\sum_{i=1}^{k} n_i}$$

kde x_s jsou středy intervalů a k je počet intervalů.

Směrodatná odchylka

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

• Druhá odmocnina z rozptylu.

• Je vyjádřením proměnlivosti v jednotkách původních dat. Je absolutní mírou variability.

• Má největší použití pro porovnání proměnlivosti více souborů.

• Má velký význam pro vymezení třídních intervalů za předpokladu normálního rozdělení.

Výpočet směrodatné odchylky ze skupinového rozdělení četností:

$$s = \sqrt{\frac{\sum_{i=1}^{k}(x_s - \bar{x})^2 \cdot n_i}{\sum_{i=1}^{k} n_i}}$$

Vlastnosti rozptylu a směrodatné odchylky

• Rozptyl hodnot znaku v celém souboru se rovná součtu aritmetického průměru skupinových rozptylů a rozptylu skupinových průměrů.

• Přidáním konstanty k jednotlivým znakům se jejich rozptyl ani směrodatná odchylka nemění.

• Násobíme-li jednotlivé znaky konstantou, jejich rozptyl je násoben čtvercem této konstanty a směrodatná odchylka je násobena touto konstantou.

• Násobíme-li váhy konstantou, rozptyl ani směrodatná odchylka se nemění.

(*Modifikace výpočtu rozptylu a směrodatné odchylky pro základní soubor – viz. odhady parametrů)*

Variační koeficient

• Nejpoužívanější relativní míra proměnlivosti.

• Poměr směrodatné odchylky k průměru (směrodatná odchylka vyjádřená v procentech průměru):

$$v = \frac{s}{\bar{x}} \cdot 100$$

Slouží k porovnání proměnlivosti více souborů o nestejné úrovni (průměru).

Příklad:

Charakteristiky naměřené na dvou objektech mají stejnou směrodatnou odchylku avšak výrazně jiný aritmetický průměr hodnot.

Charakteristika	Stanice č. 1	Stanice č. 2
X1	6	56
X2	8	58
X3	10	60
X4	12	62
X5	16	66
X6	18	68
Aritmetický průměr	11,67	61,67
Směrodatná odchylka	4,23	4,23
Variační koeficient	39,5	7,5

Charakteristiky asymetrie - šikmosti (SKEWNESS)

Charakterizují nesouměrnost rozdělení četností. Dávají představu o tvaru rozdělení.

Míry šikmosti založené na variačním rozpětí

$$s = \frac{x_{max} + x_{min} - 2\tilde{x}}{x_{max} - x_{min}}$$

Míry šikmosti založené na rozpětí kvantilů

$$s_i = \frac{\tilde{x}_{100-i} + \tilde{x}_i - 2\tilde{x}}{\tilde{x}_{100-i} - \tilde{x}_i} \qquad s_{25} = \frac{\tilde{x}_{75} + \tilde{x}_{25} - 2\tilde{x}}{\tilde{x}_{75} - \tilde{x}_{25}}$$

Koeficient asymetrie α

Aritmetický průměr z třetích mocnin odchylek jednotlivých hodnot znaku od aritmetického průměru vyjádřených v jednotkách směrodatné odchylky.

Pro ideálně symetrické rozdělení nabývá hodnoty 0.

Ze skupinového rozdělení četností se koeficient asymetrie vypočte:

$$\alpha = \frac{\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^3}{s^3 \sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^3}{n \cdot s^3}$$

Umožňuje objektivní porovnání dvou histogramů.

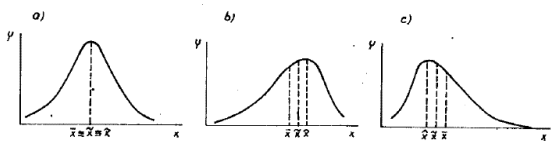
Koeficient asymetrie α

Podle hodnoty koeficientu asymetrie rozlišujeme rozdělení

a) souměrné $\alpha = 0$

b) zešikmené zleva (záporná asymetrie) $\alpha < 0$

c) zešikmené zprava (kladná asymetrie) $\alpha > 0$



Charakteristiky špičatosti (KURTOSIS)

• Popisují koncentraci prvků souboru v blízkosti určité hodnoty znaku.

• Dávají představu o rozdělení s ohledem na jeho „špičatost“ či „plochost“.

• Vyšší hodnoty charakteristik špičatosti mají soubory, u kterých jsou prvky souboru více koncentrovány kolem uvažované hodnoty znaku.

Míra koncentrace kolem mediánu

$$K = \frac{x_{max} - x_{min}}{\tilde{x}_{75} - \tilde{x}_{25}}$$

Koeficient špičatosti (exces) ε

Průměrná hodnota součtu čtvrtých odmocnin odchylek hodnot znaku od průměru měřených v jednotkách směrodatné odchylky.

Jedná se o bezrozměrné číslo. Ze skupinového rozdělení četností se koeficient špičatosti vypočte:

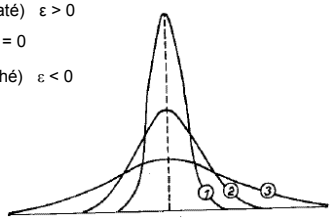
$$\varepsilon = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^4 \cdot n_i}{n \cdot s^4} - 3$$

Špičatost (resp. plochost) rozdělení je tím větší, čím více se hodnota ε odlišuje od nuly.

Koeficient špičatosti (exces) ε

Podle hodnoty koeficientu špičatosti rozlišujeme rozdělení

1. kladně zašpičatělé (špičaté) $\varepsilon > 0$
2. normálně zašpičatělé $\varepsilon = 0$
3. záporně zašpičatělé (ploché) $\varepsilon < 0$



Obě uvedené míry dávají informaci o tom, do jaké míry se rozdělení studovaného souboru liší od normálního. Mají využití v aplikacích tzv. parametrických testů.

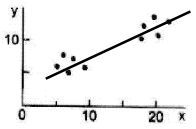
Průzkumová analýza dat
(EDA - Exploratory Data Analysis)

• Souhrn metod popisné statistiky, které předchází vlastnímu statistickému zpracování.

• Cílem je **ověřit** některé vlastnosti vstupního datového souboru, které jsou nezbytnými předpoklady pro vlastní statistické metody zpracování.

• EDA se zaměřuje na grafické a tabelační znázorňování dat

• Každá analýza by měla začínat pečlivým zkoumáním struktury dat



Průzkumová analýza dat
(EDA - Exploratory Data Analysis)

EDA zahrnuje především:

• výpočet charakteristik úrovně a variability
• analýzu odlehlých hodnot
• studium histogramu s cílem ověření normality rozdělení
• konstrukci grafů
• ověření homogenity vstupních dat
• ověření stacionarity vstupních dat

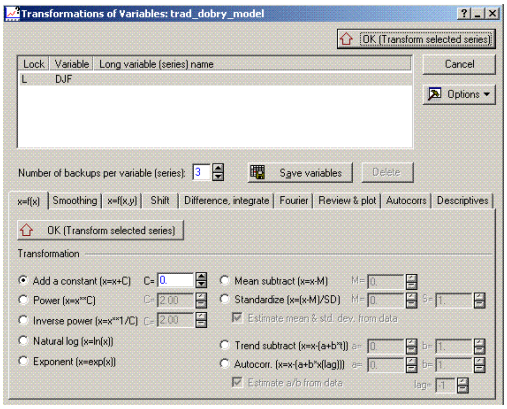
Výsledkem EDA je závěr o event. potřebě transformace vstupních dat

Transformace dat

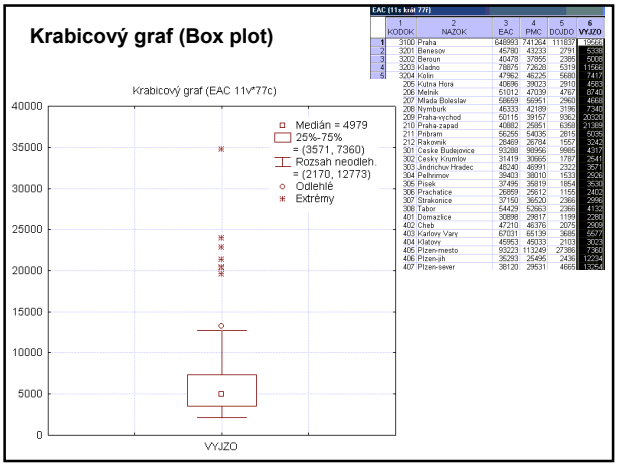
Cíle: úprava dat pro následnou analýzu, splnění požadavků některých statistických metod, zjednodušení výpočtu, …

• funkční transformace
• standardizace
• transformace do pořadí
• transformace na percentily, …

Příklad transformace dat ve programu Statistica



Krabicový graf (Box plot)



Krabicový graf – porovnání více souborů