Explorační analýza dat

Popisné statistiky

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Petr Seďa

Pravděpodobnost a statistika II

2025

Čím se zabývá teorie pravděpodobnosti?

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

- **Teorie pravděpodobnosti** je matematická disciplína, jejíž logická struktura je budována **axiomaticky.** To znamená, že její základ tvoří několik tvrzení (takzvaných axiomů), která vyjadřují základní vlastnosti pravděpodobnosti a všechna další tvrzení jsou z nich **odvozena** deduktivně. Popisuje zákonitosti týkající se **náhodných jevů**, tj. jevů, které (přinejmenším z hlediska pozorovatele) mohou a nemusí nastat. Hledá **pravděpodobnosti určitých výsledků** (náhodných jevů), známe-li základní soubor (populaci).

- **Matematická statistika** je věda zahrnující **studium dat** vykazujících **náhodná kolísání,** ať už jde o data získaná pečlivě připraveným pokusem provedeným pod stálou kontrolou experimentálních podmínek v laboratoři, či o data provozní, případně o data získaná počítačovými simulacemi (např. metodou Monte-Carlo).

Co je to statistika?

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

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 webové stránky, statistika střel na branku, statistika nehodovosti, ekonomické statistiky, …).

Co vypovídá statistika o jednotlivci?

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Donald Trump

podnikatel

politik (prezident)

Američan

Ve statistice nezkoumáme jednotlivce jako individualitu, ale jako anonymního nositele některého **znaku** (činnosti, vlastnosti).

Základní pojmy

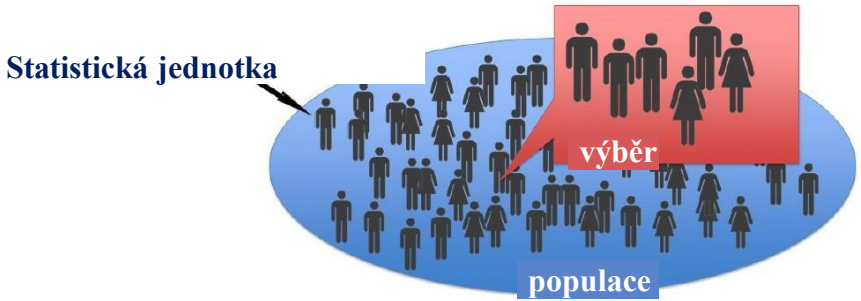
SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

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, …).



Typy statistických znaků

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Podle způsobu zpracování:

znaky kvalitativní – vyjadřují kvalitu (úroveň, míru):

a) znaky nominální - všechny hodnoty rovnocenné, mají stejnou váhu

např. národnost, profese, rodinný stav

b) znaky ordinální - hodnoty lze uspořádat podle významu

např. dosažené vzdělání, platový stupeň, spokojenost

znaky kvantitativní – vyjadřují kvantitu (počet, číslo):

a) znaky metrické - s hodnotami lze plnohodnotně počítat

např. počet dětí, hrubá měsíční mzda, hospodářský výsledek

Jak určit typ statistického znaku (proměnných)

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Při porovnávání dvou hodnot kvantitativního metrického znaku se lze ptát **„kolikrát?“** nebo **„o kolik?“** ve smyslu hodnoty.
Při porovnání dvou hodnot kvalitativního ordinálního znaku se lze ptát pouze **„o kolik pozic?“** ve smyslu pořadí.
Hodnoty kvalitativního nominálního znaku **nelze porovnávat** (uvedené otázky nemají smysl).

Odpovězte si:

Může existovat nominální znak s číselnými hodnotami?

Může mít znak číselné i nečíselné hodnoty?

Jaký znak je „známka ze zkoušky“ – nominální, ordinální, metrický?

Třídění kvalitativních znaků

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Podle počtu různých hodnot (obměn):

znaky alternativní (dichotomické) – pouze dvě různé hodnoty

např. členství v komoře, pohlaví, plátce DPH

znaky množné – více než dvě různé hodnoty

např. dosažené vzdělání, rodinný stav, platový stupeň

Podle formy vyjádření hodnot (obměn):

znaky slovní (alfanumerické) – obměny vyjádřené slovně

např. barva vlasů, spokojenost, známka u zkoušky

znaky číselné (numerické) – obměny vyjádřené číselně

např. IČO, platový stupeň, známka u zkoušky

Třídění kvantitativních znaků

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Podle způsobu srovnávání hodnot (obměn):

znaky intervalové – o kolik je jedna hodnota větší než druhá

např. hospodářský výsledek, teplota ve stupních Celsia

znaky poměrové – kolikrát je jedna hodnota větší než druhá

např. počet dětí, hrubá měsíční mzda, teplota v Kelvinech

Podle počtu hodnot (obměn):

znaky diskrétní – konečný nebo spočetný počet hodnot

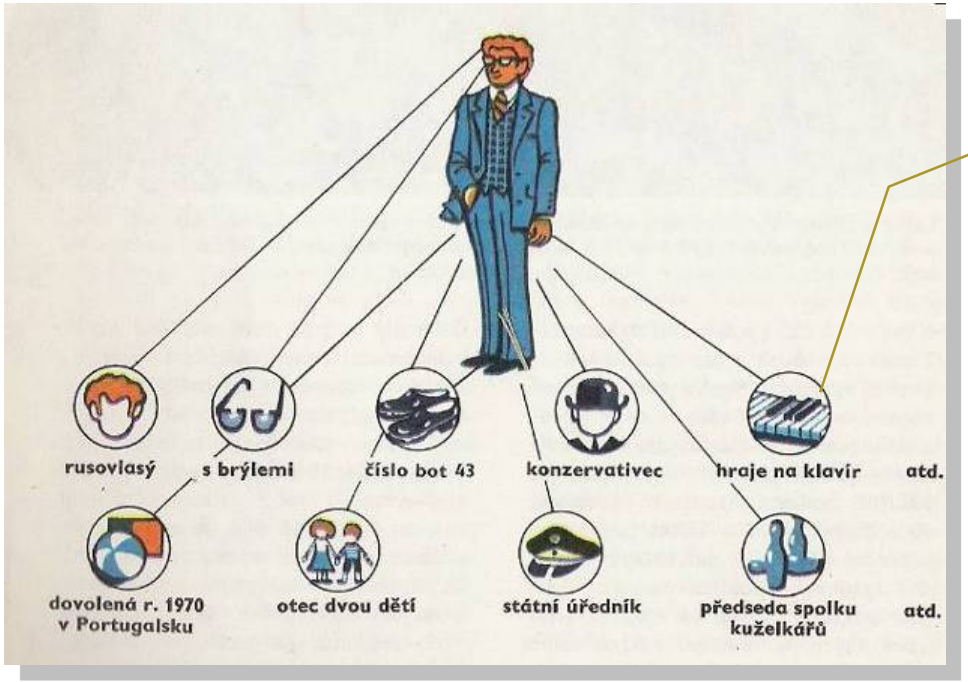
např. počet dětí, věk, počet zakázek

znaky spojité – libovolné hodnoty z daného intervalu

např. hospodářský výsledek, doba obsluhy, hrubá měsíční mzda

Typy statistických znaků (proměnných)

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ



hraje na klavír

znak:
záliby (koníčky)

určení:
znak kvalitativní
nominální
množný
slovní

Typy statistických znaků (proměnných)

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Časová značka	Pohlaví	Výška	Váha	Přivyděláváte si během studia?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně 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	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	Hlídání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v tydnu	praxe v oboru během studia	24	26

Respondent (proband) – označení statistické jednotky v dotazníkovém šetření.

Popište strukturu datového souboru v závislosti na pohlaví respondentů.

Typy statistických znaků (proměnných)

Časová značka	Pohlaví	Výška	Váha	Přivyděláváte si během studia?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně 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	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	Hlídání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v tydnu	praxe v oboru během studia	24	26

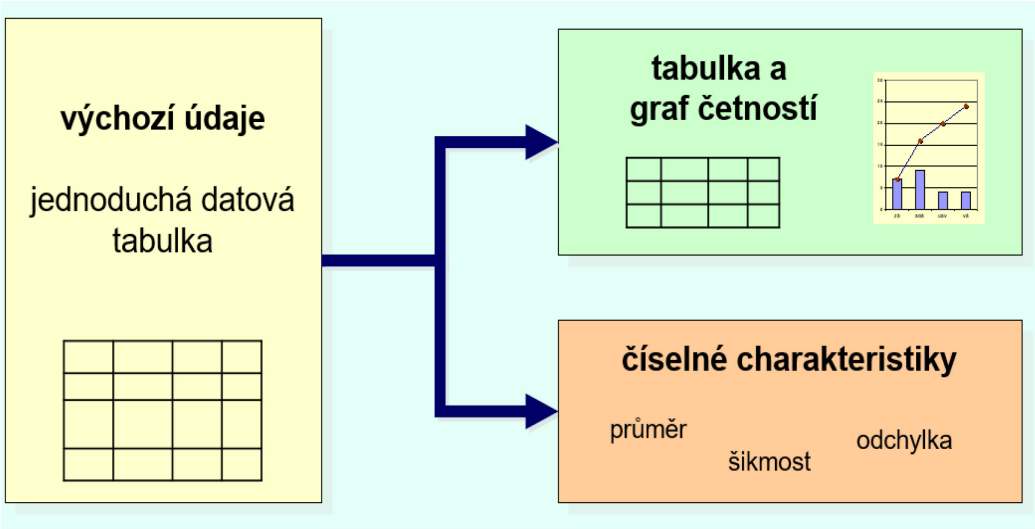
Statistický znak je nějaká měřitelná (zjistitelná) charakteristika prvků základního souboru.

a) **Kvantitativní znak** – znak, jehož varianty mají číselné hodnoty (má smysl posuzovat rozdíly a poměry)

b) **Kvalitativní znak** – znak, jehož varianty se liší kvalitou (může jít i o číselné hodnoty)

Jak správně roztřídit data?

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ



a) **řádky** tabulky představují jednotlivé **třídy** (kategorie)

b) **sloupce** tabulky vyjadřují **četnosti** (počty jednotek)

Třídy mohou být:

a) jednotlivé obměny třídícího znaku,

b) množiny hodnot třídícího znaku (například intervaly).

Každá hodnota třídícího znaku musí náležet **právě jedné** třídě.

Popisná statistika - tabulka četností

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

NÁZEV (HODNOTA) TŘÍDY	ABSOLUTNÍ ČETNOSTI	RELATIVNÍ ČETNOSTI	KUMULATIVNÍ ABSOLUTNÍ ČETNOSTI	KUMULATIVNÍ RELATIVNÍ ČETNOSTI
třída 1	n_1	p_1	$kn_1 = n_1$	$kp_1 = p_1$
třída 2	n_2	p_2	$kn_2 = n_1 + n_2$	$kp_2 = p_1 + p_2$
…	…	…	…	…
třída m	n_m	p_m	$kn_m = n$	$kp_m = 1$
CELKEM	n	1	x	x

Četnost (absolutní četnost) n_i představuje počet prvků třídy i.

Relativní četnost p_i vyjadřuje část z celku (obvykle v %).

Platí: $p_i = \frac{n_i}{n}$, kde n je velikost souboru.

Popisná statistika - tabulka četností

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

NÁZEV (HODNOTA) TŘÍDY	ABSOLUTNÍ ČETNOSTI	RELATIVNÍ ČETNOSTI	KUMULATIVNÍ ABSOLUTNÍ ČETNOSTI	KUMULATIVNÍ RELATIVNÍ ČETNOSTI
třída 1	n_1	p_1	$kn_1 = n_1$	$kp_1 = p_1$
třída 2	n_2	p_2	$kn_2 = n_1 + n_2$	$kp_2 = p_1 + p_2$
…	…	…	…	…
třída m	n_m	p_m	$kn_m = n$	$kp_m = 1$
CELKEM	n	1	x	x

Absolutní kumulativní četnost kn_i je počet hodnot menších nebo rovných dané třídě:
$kn_i = n_1 + n_2 + \cdots + n_i = \sum_{k=1}^{i} n_k.$

Relativní kumulativní četnost $kp_i = \frac{kn_i}{n}$.

Popisná statistika - tabulka četností

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

NÁZEV (HODNOTA) TŘÍDY	ABSOLUTNÍ ČETNOSTI	RELATIVNÍ ČETNOSTI	KUMULATIVNÍ ABSOLUTNÍ ČETNOSTI	KUMULATIVNÍ RELATIVNÍ ČETNOSTI
třída 1	n_1	p_1	$kn_1 = n_1$	$kp_1 = p_1$
třída 2	n_2	p_2	$kn_2 = n_1 + n_2$	$kp_2 = p_1 + p_2$
…	…	…	…	…
třída m	n_m	p_m	$kn_m = n$	$kp_m = 1$
CELKEM	n	1	x	x

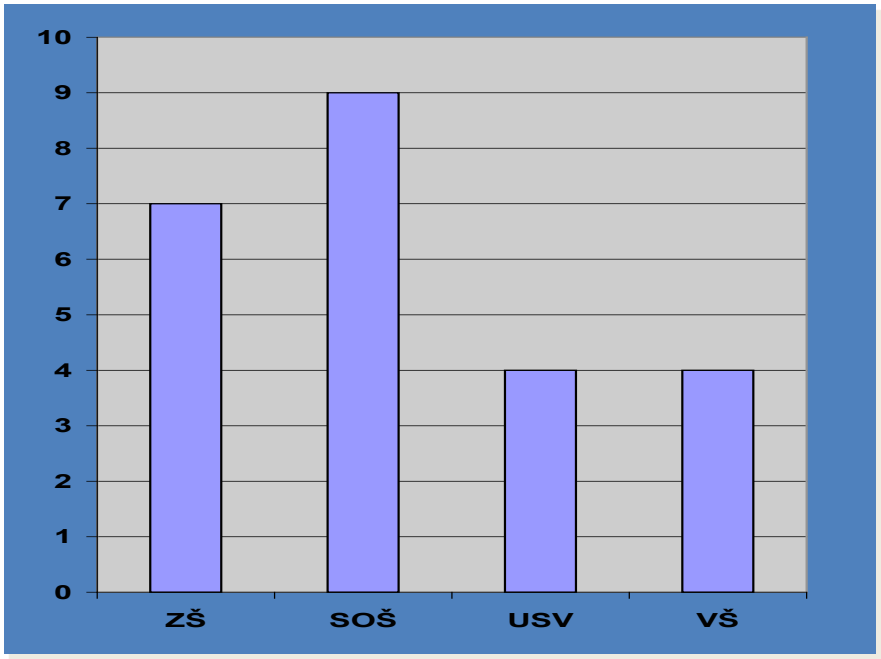
Platí: $kn_i = kn_{i-1} + n_i$ a $kp_i = kp_{i-1} + p_i$ pro $i = 2, 3, \dots, m$.

Popisná statistika – sloupcový graf

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Sloupcový graf četností

Vhodný pro **nominální a ordinální proměnné**.



osa *x* – jednotlivé **kategorie** seřazené abecedně (nominální znaky) nebo podle uspořádání

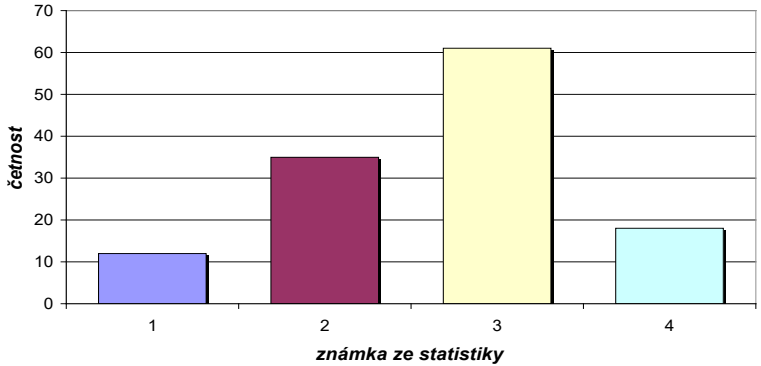
osa *y* – vynášené **četnosti**

- absolutní četnosti
- relativní četnosti
- kumulativní četnosti

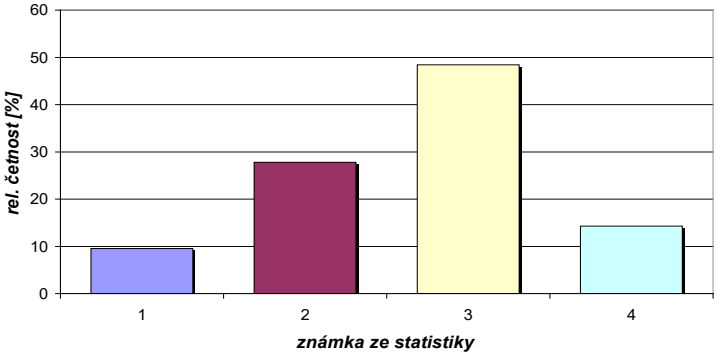
Popisná statistika - sloupcový graf

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

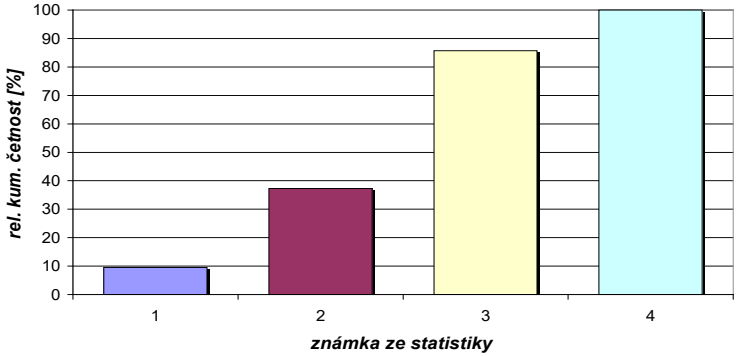
Sloupcový graf absolutních četností



Sloupcový graf relativních četností



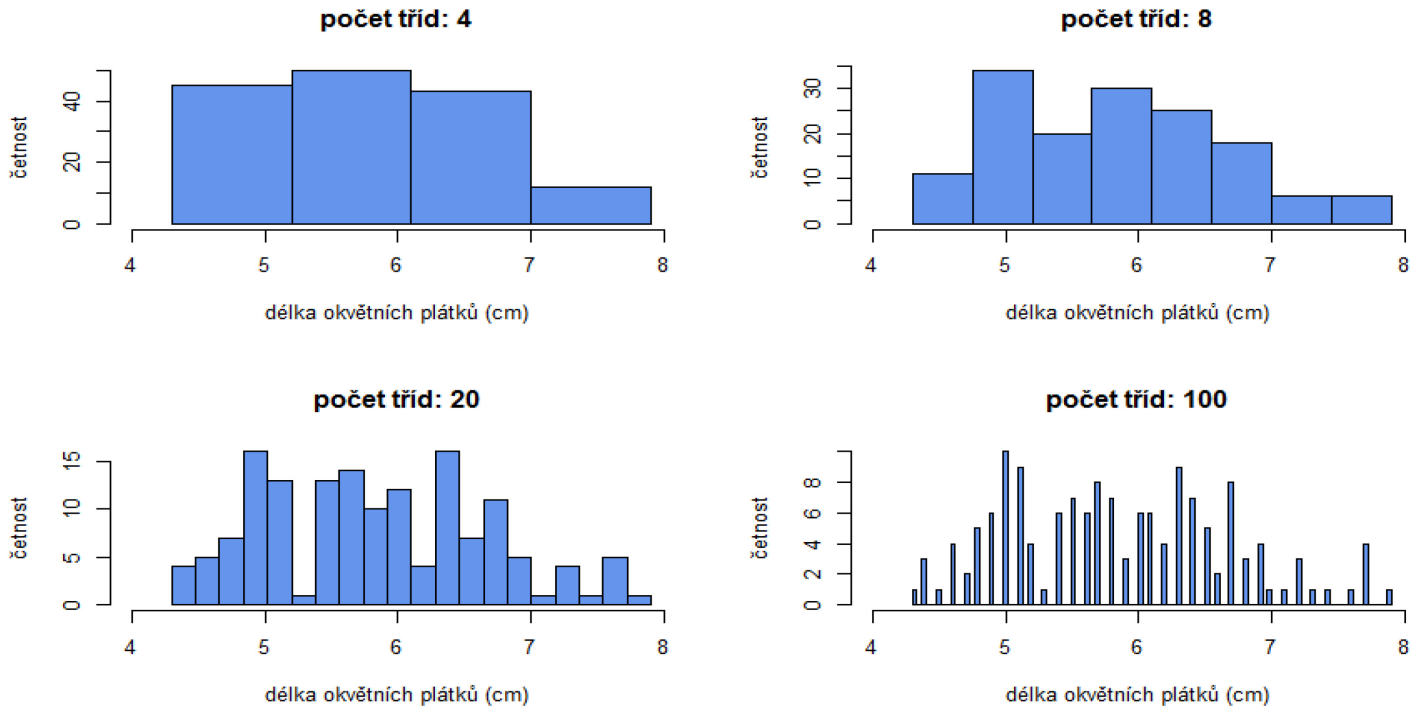
Sloupcový graf relativních kumulativních četností



- absolutní a relativní četnosti mají stejný průběh, liší se pouze v ose y
- kumulativní četnosti mají vždy neklesající průběh

Popisná statistika - histogram

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ



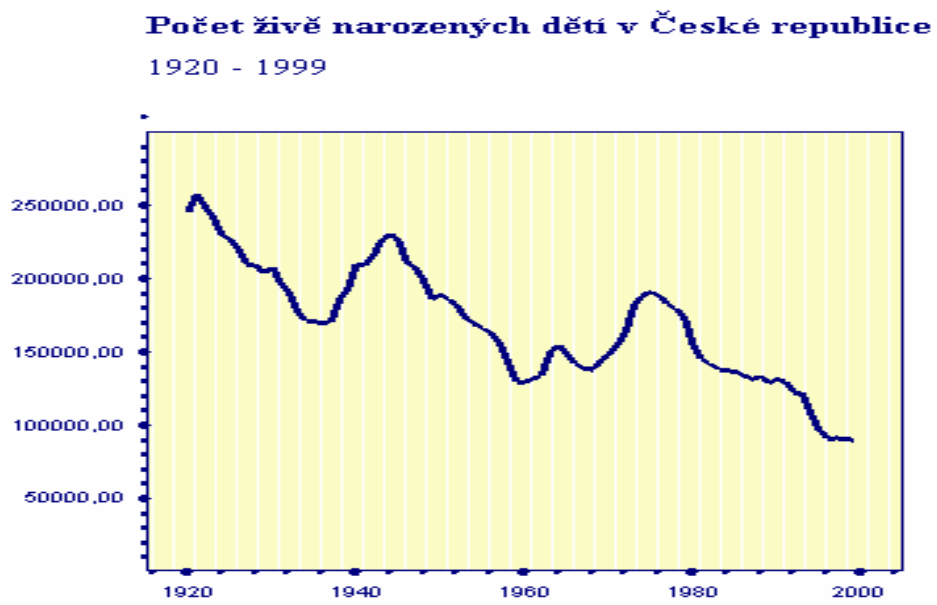
Tvar histogramů závisí na počtu tříd (sloupečků)!

Popisná statistika - polygon

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Spojnicový graf četností (polygon)

Vhodné pro **ordinální a metrické proměnné**.

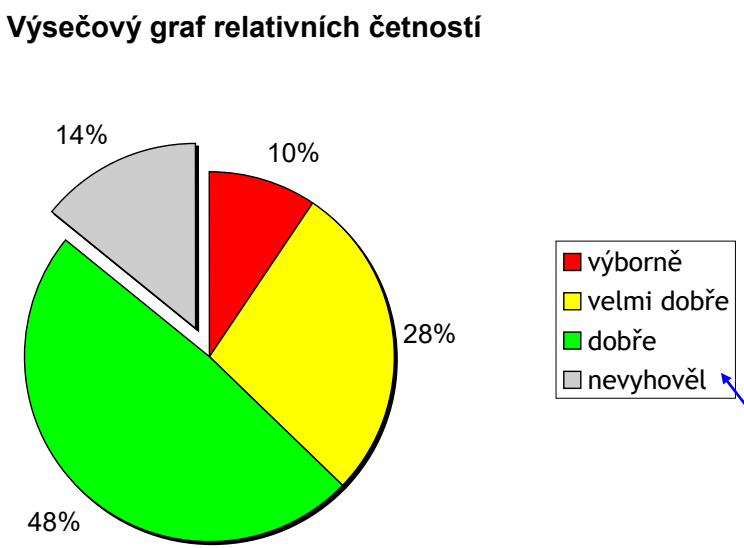


Zvláštní případ:
graf časové řady kdy kategoriální proměnná je čas.

Popisná statistika – výsečový graf

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Výsečový graf četností.
Vhodné zejména pro **nominální proměnné** s malým počtem obměn.



Plochy výsečí a středové úhly úměrné četnostem.

Popisky obvykle v %.

Důležité obměny lze zvýraznit vysunutím.

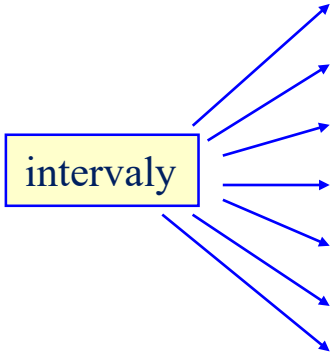
legenda grafu

Popisná statistika – intervalové rozdělení četností

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Vhodné pro **spojité metrické znaky** (s velkým počtem obměn).

Rozdělení oboru hodnot třídícího znaku na jednotlivé intervaly.



VĚK	četnosti		kumul. četnosti	
	abs.	rel.	abs.	rel.
26 - 30	6	25,0%	6	25,0%
31 - 35	5	20,8%	11	45,8%
36 - 40	3	12,5%	14	58,3%
41 - 45	5	20,8%	19	79,2%
46 - 50	2	8,3%	21	87,5%
51 - 55	2	8,3%	23	95,8%
56 - 60	1	4,2%	24	100,0%
Celkem	24	100,0%	x	x

Popisná statistika – jak vytvořit intervaly

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

a) počet tříd (intervalů) v rozmezí 5 až 20

- malý počet tříd – malá informační hodnota
- velký počet tříd – nepřehledná tabulka

b) hranice intervalů dobře zapamatovatelná čísla

- dělitelná 5, 10, 20, …

c) intervaly jednoznačně pokrývají celý obor hodnot

- hraniční body intervalů patří pouze jednomu z nich

d) intervaly stejně široké

- srovnatelnost intervalů mezi sebou

e) oba krajní intervaly mají nenulové četnosti

Popisná statistika – odhad počtu tříd

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Sturgesovo pravidlo pro odhad počtu tříd k:

$$k \approx 1 + 3{,}3 \cdot \log n$$

n	k
10	4
20	5
50	7
100	8
500	10
1000	11

Odhad šířky intervalu h:

$$h \approx \frac{x_{max} - x_{min}}{k}$$

a) vypočtené hodnoty jsou pouze doporučením

b) skutečné hodnoty → přehlednost tabulky

(zaokrouhlené hranice a šířky tříd)

Číselné charakteristiky

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Agregují informaci o statistickém znaku do několika málo hodnot.

Jsou stručnější a přehlednější než výchozí data.

Snaží se charakterizovat rozdělení hodnot znaku.

Základní typy charakteristik

a) míry polohy – umístění hodnot znaku (na číselné ose),

b) míry variability - rozptýlení hodnot kolem typické polohy,

c) míry tvaru rozdělení – symetrie, koncentrace hodnot znaku.

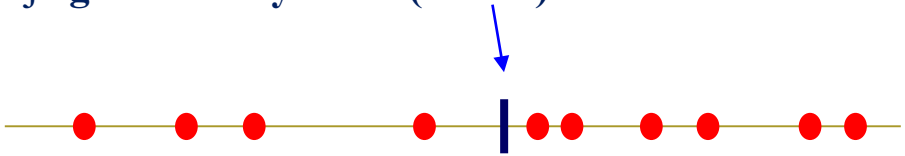
Míry polohy

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Určují polohu (pomyslný střed) statistického znaku.

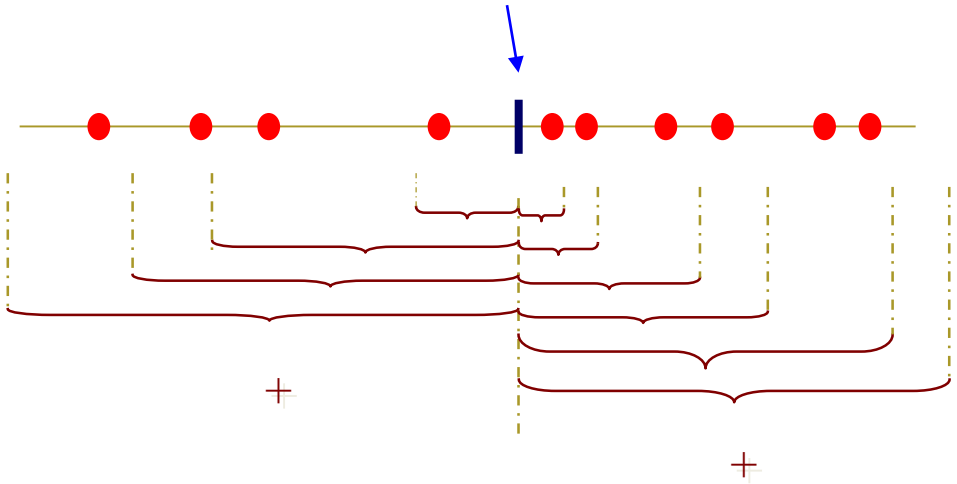
Střední hodnota – aritmetický průměr: $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Vyjadřuje **geometrický střed (těžiště)** statistického znaku na číselné ose:



Co je to geometrický střed?

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ



součet vzdáleností od průměru hodnot **menších** než průměr = součet vzdáleností od průměru hodnot **větších** než průměr

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Nejednoznačná terminologie

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Místo pojmu **průměr** se v praxi často používají následující pojmy:

a) průměrná hodnota,

b) střední hodnota,

c) prostřední hodnota,

d) charakteristická hodnota,

e) typická hodnota,

f) očekávaná hodnota.

Vlastnosti aritmetického průměru

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Součet odchylek všech hodnot znaku od aritmetického průměru je roven nule:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Přičteme-li ke všem hodnotám znaku stejné číslo, zvětší se o toto číslo také aritmetický průměr:

$$\overline{x + a} = \bar{x} + a$$

Vynásobíme-li všechny hodnoty znaku stejným číslem, zvětší se stejným způsobem i aritmetický průměr:

$$\overline{ax} = a \cdot \bar{x}$$

Vážený aritmetický průměr

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Střední hodnota pro tabulku rozdělení četností.

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \cdots + x_k \cdot n_k}{n} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}$$

počet tříd (kategorií)

četnosti jednotlivých tříd

jednotlivé hodnoty znaku

velikost souboru

$$\bar{x} = x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_k \cdot p_k = \sum_{i=1}^{k} x_i \cdot p_i$$

Jiné typy průměrů

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Kromě aritmetického průměru existují také další typy průměrů:

a) **Harmonický průměr**: $\bar{x}_H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$

b) **Geometrický průměr**: $\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Kdy se který z těchto průměrů používá?

a) Geometrický průměr: např. výpočet průměrného tempa růstu.

b) Harmonický průměr: například při výpočtu průměrné rychlosti na úsecích stejné délky. Dále jsou-li hodnoty znaku nerovnoměrně rozloženy kolem aritmetického průměru, nebo když jsou hodnoty extrémně nízké či vysoké.

Další míry polohy – modus a medián

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Modus $\hat{x}$ – **nejčetnější hodnota** znaku v souboru (vyskytuje se v souboru nejčastěji)

- vhodné zejména pro nominální znaky,
- nemusí být určen jednoznačně.

Medián $\tilde{x}$ – **prostřední hodnota** znaku v souboru uspořádaného podle velikosti znaku

- vhodné pro ordinální a nesymetrické znaky,
- není důležitá hodnota, ale pořadí,
- u sudého počtu prvků souboru se medián počítá jako průměr ze dvou hodnot nejbližších středu.

Jakou střední hodnotu použít?

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

(Aritmetický) průměr - u číselných znaků, které nevykazují extrémní hodnoty.

Medián - u číselných znaků s extrémy, u ordinálních nečíselných znaků.

Modus - u nominálních nečíselných znaků.

Otázka k zamyšlení:

Proč aritmetický průměr není vhodnou střední hodnotou pro znak „měsíční příjem zaměstnance“?

Příklad – portfolio akcií

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Příklad 1.1: U portfolia akcií vypočítejte střední hodnotu ceny akcie, modus a medián.

cena akcie	počet
200 Kč	3
300 Kč	5
500 Kč	2
1 000 Kč	1
1 500 Kč	1

modus (nejčetnější hodnota)
300 Kč

medián: $n = 12$

$$\tilde{x} = \frac{x_6 + x_7}{2} = \frac{300 + 300}{2} = 300.$$

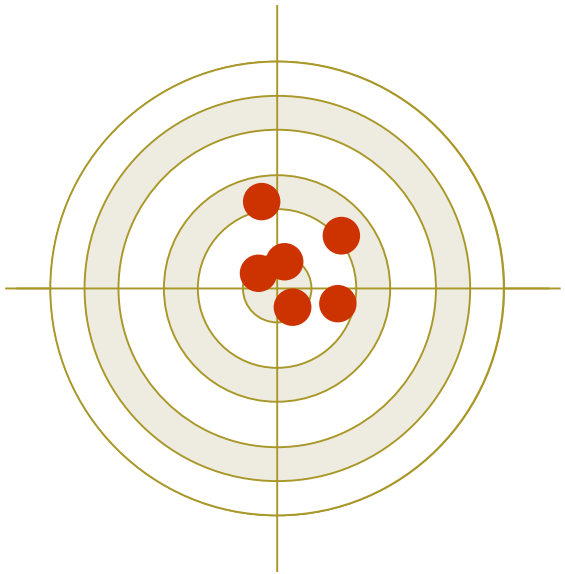
průměrná cena akcie:

$$\bar{x} = \frac{200 \cdot 3 + 300 \cdot 5 + 500 \cdot 2 + 1000 \cdot 1 + 1500 \cdot 1}{3 + 5 + 2 + 1 + 1} = \frac{5600}{12} = 466{,}67$$

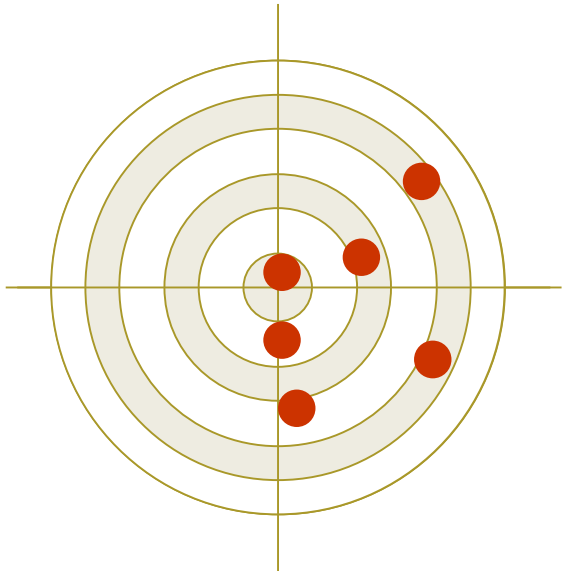
Variabilita znaku

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Variabilita určuje, jak se hodnoty znaku **liší od průměru**.



malý rozptyl



velký rozptyl

Rozptyl – ukazatel variability

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Rozptyl - variabilita znaku v souboru:

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

průměrný čtverec odchylek od průměru

nezáleží na znaménku odchylky

Vzorec vhodnější pro ruční výpočet:

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

Vlastnosti rozptylu

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Rozptyl konstanty (znaku, který nemění svou hodnotu) je roven nule:

$$s^2(a) = 0.$$

Přičteme-li ke všem hodnotám statistického znaku stejné číslo, rozptyl se nezmění:

$$s^2(x + a) = s^2(x).$$

Vynásobíme-li všechny hodnoty statistického znaku stejným číslem (např. k-krát), zvětší se rozptyl znaku dvojmocí této hodnoty (tj. k^2-krát):

$$s^2(ax) = a^2 \cdot s^2\,(x).$$

Další ukazatele variability

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Směrodatná odchylka:

$$s = \sqrt{s^2}$$

průměrná odchylka od průměru
(tzv. kvadratický průměr)

ve stejných jednotkách jako původní znak

Variační koeficient:

$$V_x = \frac{s}{\bar{x}}$$

a) použití pro znaky s nezápornými hodnotami
b) srovnání znaků s různou velikostí hodnot
c) obvykle se vyjadřuje v % (x 100)

Příklad – portfolio akcií

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Příklad 1.2: U portfolia akcií vypočítejte rozptyl, směrodatnou odchylku a variační koeficient cen akcií.

cena akcie	počet
200 Kč	3
300 Kč	5
500 Kč	2
1 000 Kč	1
1 500 Kč	1

Rozptyl ceny akcie:

$$s^2 = \frac{200^2 \cdot 3 + 300^2 \cdot 5 + 500^2 \cdot 2 + 1000^2 \cdot 1 + 1500^2 \cdot 1}{12} - 466{,}67^2 = 142219$$

Směrodatná odchylka a variační koeficient:

$$s = \sqrt{s^2} = \sqrt{142219} = 377$$

$$V_x = \frac{s}{\bar{x}} = \frac{377}{466{,}67} = 0{,}808 = 80{,}8\%$$

Závěr: vysoká variabilita znamená, že střední hodnota (průměr) není dobrým reprezentantem znaku.

Jak chápat směrodatnou odchylku?

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Čebyševova nerovnost:

V intervalu $(\bar{x} - k \cdot s; \bar{x} + k \cdot s)$ se nachází nejméně $1 - \frac{1}{k^2}$ hodnot znaku pro $(k > 1)$.

Pravidlo 6 sigma:

Všechny hodnoty znaku, které se nacházejí ve vzdálenosti větší než 3 směrodatné odchylky od průměru, se považují za extrémní.

Normovaná hodnota z

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Určuje vzdálenost hodnoty znaku od střední hodnoty (v násobcích směrodatné odchylky):

$$z_i = \frac{x_i - \bar{x}}{s}$$

$z_i > 0$ hodnota je větší než průměr

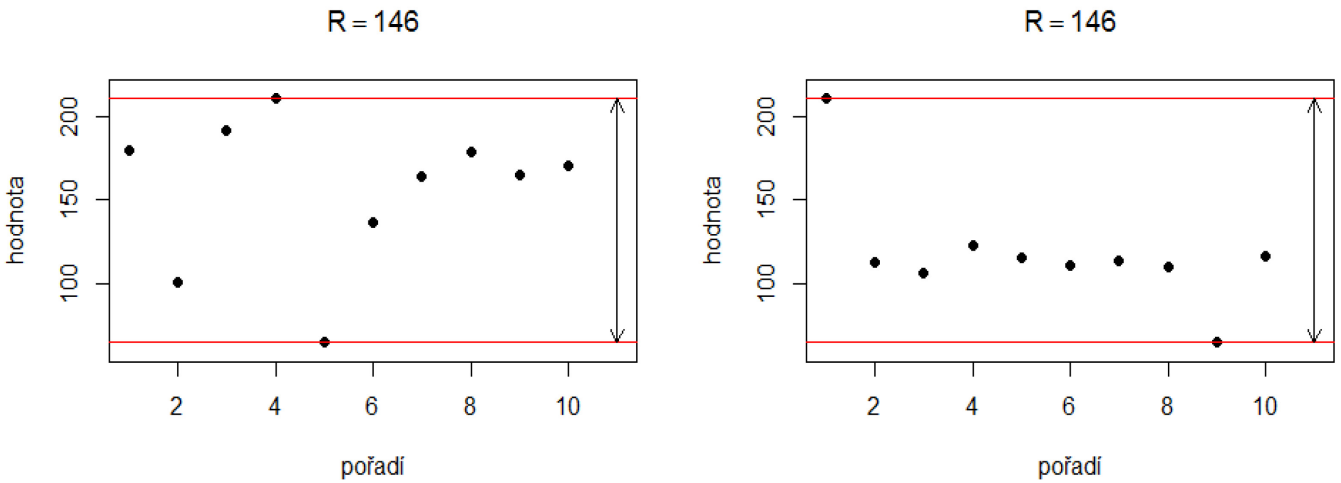
$z_i < 0$ hodnota je menší než průměr

Pravidlo 6 sigma:
Hodnoty z větší než 3 (menší než -3) značí extrémní hodnoty.

Pozn. Někdy se normovaná hodnota označuje též jako u.

Variační rozpětí

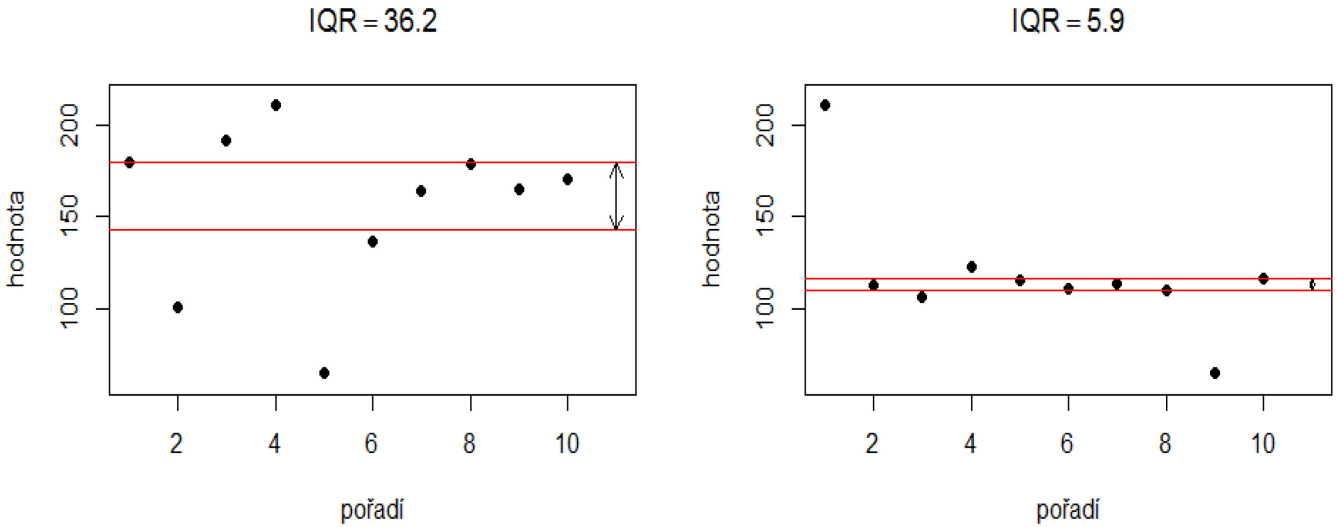
SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ



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

Mezikvartilové rozpětí

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ



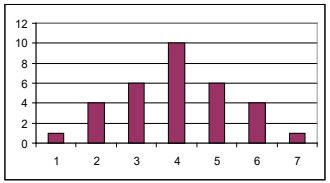
Mezikvartilové (interkvartilové) rozpětí: $IQR = x_{0,75} - x_{0,25}$

Míry tvaru rozdělení

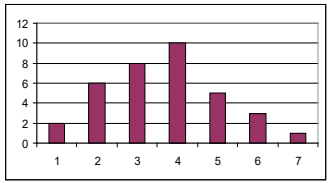
SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

šikmost - vyjadřuje asymetrii rozložení hodnot znaku

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} z_i^3$$

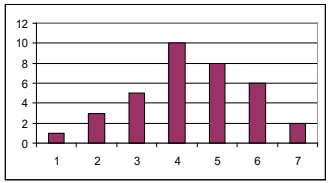


$\alpha = 0$



$\alpha > 0$

kladné sešikmení

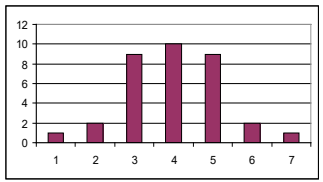


$\alpha < 0$

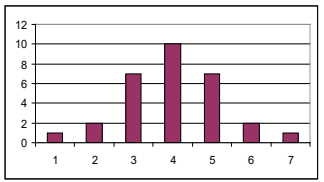
záporné sešikmení

špičatost - vyjadřuje koncentraci hodnot znaku

$$\beta = \frac{1}{n}\sum_{i=1}^{n} z_i^4 - 3$$

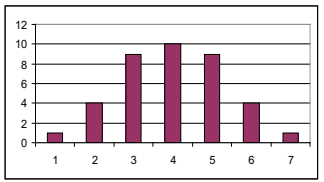


$\beta = 0$



$\beta > 0$

špičaté rozdělení



$\beta < 0$

ploché rozdělení

Příklad – portfolio akcií

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Příklad 1.3: U portfolia akcií vypočítejte šikmost a špičatost cen akcií.

Šikmost: $\alpha = \frac{(-0{,}68)^3 \cdot 3 + (-0{,}42)^3 \cdot 5 + 0{,}08^3 \cdot 2 + 1{,}35^3 \cdot 1 + 2{,}62^3 \cdot 1}{12} = 1{,}59$

kladné sešikmení – vyšší koncentrace menších hodnot

Špičatost: $\beta = \frac{(-0{,}68)^4 \cdot 3 + (-0{,}42)^4 \cdot 5 + 0{,}08^4 \cdot 2 + 1{,}35^4 \cdot 1 + 2{,}62^4 \cdot 1}{12} - 3 = 1{,}27$

kladná špičatost – vyšší koncentrace hodnot kolem průměru

Kvantily – specifické míry polohy

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Kvantil x_p

Odděluje p% nejnižších hodnot od zbytku souboru:

- medián $x_{50\%} = x_{0,5}$
- kvartily $x_{25\%}$ $x_{50\%}$ $x_{75\%}$
- decily $x_{10\%}$ $x_{20\%}$ ... $x_{90\%}$
- percentily $x_{1\%}$ $x_{2\%}$ ... $x_{99\%}$

z_p je pořadí kvantilu v rámci uspořádaného znaku: $z_p = \frac{n \cdot p}{100} + 0{,}5$.

Kvartily a box-plot

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

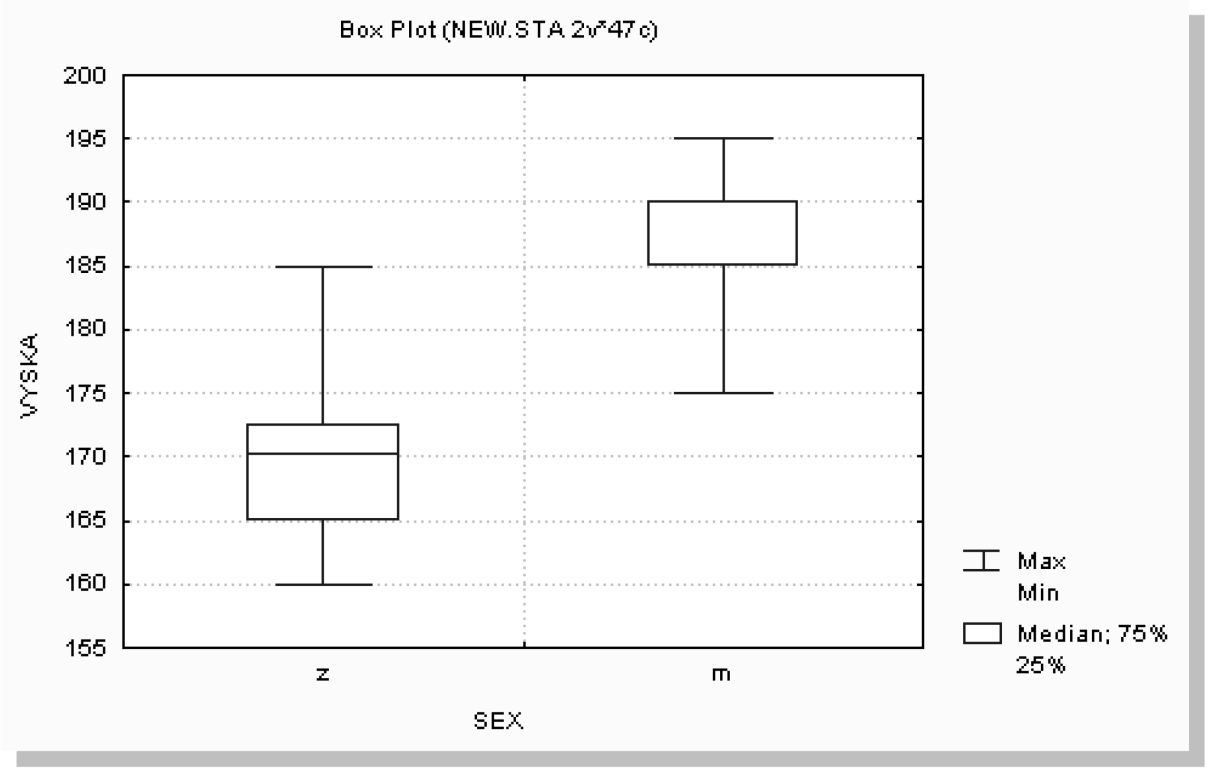
Kvartily rozdělují uspořádaný soubor na 4 stejně početné části.

box plot – graf kvartilů

Box-plot slouží k porovnávání rozdělení různých znaků

Box-plot v praxi

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ



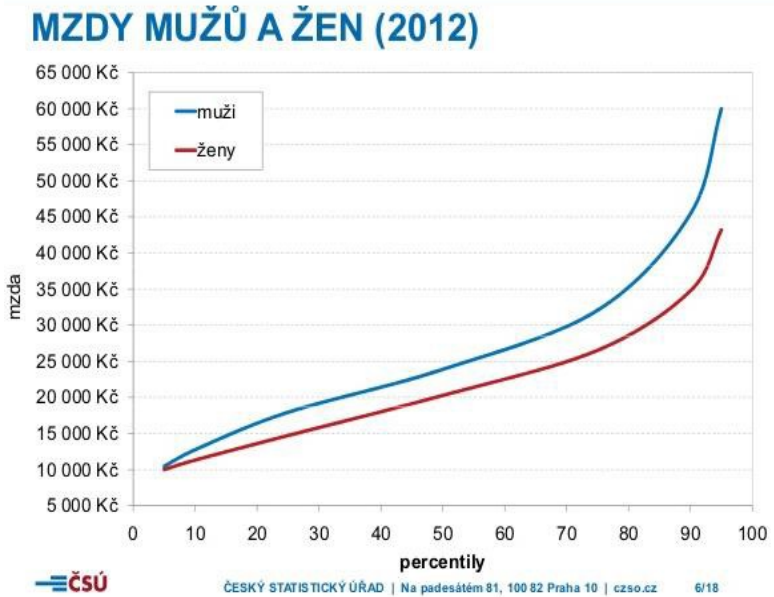
Kvantily v praxi

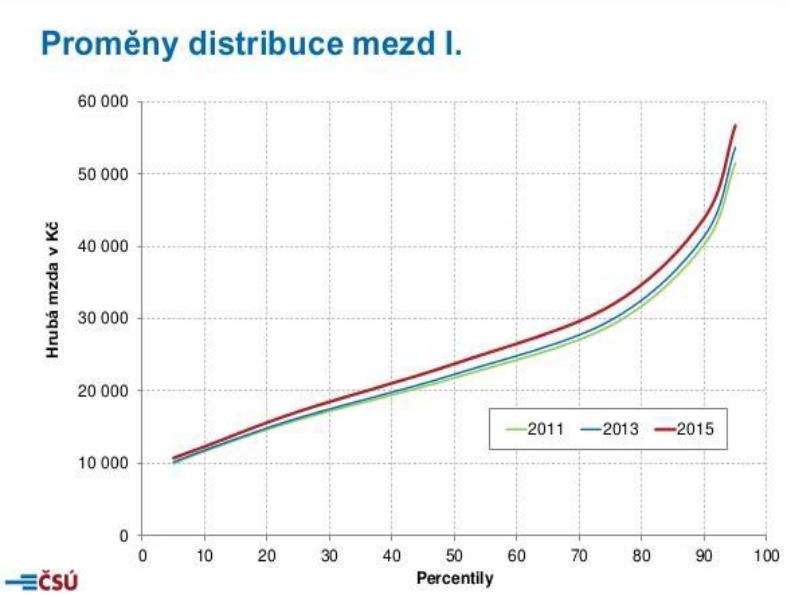
SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Zdroj: http://www.statistikaamy.cz/2018/09/kvantily-kvartily-decily-percentily/

Kvantily v praxi

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ





Charakteristiky intervalového rozdělení

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Střední hodnota – vážený aritmetický průměr:

$$\bar{x} = \frac{\bar{x}_1 \cdot n_1 + \bar{x}_2 \cdot n_2 + \cdots + \bar{x}_k \cdot n_k}{n} = \frac{\sum_{i=1}^{k} \bar{x}_i \cdot n_i}{n}.$$

Neznáme-li průměry tříd, nahradíme je středy intervalů.

Kdy je takto vypočtená střední hodnota přesná a kdy je pouze odhadem?

Charakteristiky intervalového rozdělení

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Rozptyl:

Známe-li rozptyly jednotlivých tříd:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot (\bar{x}_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{k} n_i \cdot s_i^2.$$

meziskupinový rozptyl

vnitroskupinový rozptyl

Nahradíme-li průměr třídy jejím středem – Sheppardova korekce:

$$s_{cor}^2 = s^2 - \frac{1}{12}h^2,$$ kde h je šířka třídy.

kompenzuje nadhodnocení rozptylu při náhradě průměru středem

Charakteristiky intervalového rozdělení

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Modus a kvantily – lineární interpolace

Jak upřesnit odhad modu?

$$\hat{x} = \bar{x}_i - \frac{n_{i+1} - n_{i-1}}{n_{i+1} - 2n_i + n_{i-1}} \cdot \frac{h}{2}$$

četnost intervalu předcházejícího nejčetnějšímu

Jak interpolovat kvantil?

$$\tilde{x}_p = d_i + \left(z_p - k n_{i-1}\right) \cdot \frac{h}{n_i}$$

kumulativní četnost

četnost intervalu s hledaným kvantilem

dolní mez intervalu

Odlehlá pozorování

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Takové 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:

a) hrubými chybami, překlepy, prokazatelným selháním lidí či techniky ...

b) 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.

Odlehlá pozorování

SLEZSKÁ UNIVERZITA
MATEMATICKÝ ÚSTAV V OPAVĚ

Identifikace odlehlých pozorování:

Metoda vnitřních hradeb:

$$\left[\left(x_i < x_{0,25} - 1{,}5 \cdot IQR\right) \vee \left(x_i > x_{0,75} + 1{,}5 \cdot IQR\right)\right] \rightarrow x_i \text{ je } \textbf{odlehlým pozorováním}$$

Dolní mez vnitřních hradeb

Horní mez vnitřních hradeb

Metoda vnějších hradeb:

$$\left[\left(x_i < x_{0,25} - 3 \cdot IQR\right) \vee \left(x_i > x_{0,75} + 3 \cdot IQR\right)\right] \rightarrow x_i \text{ je } \textbf{extrémním pozorováním}$$

Dolní mez vnějších hradeb

Horní mez vnějších hradeb

Literatura

1. HENDL, Jan. *Přehled statistických metod zpracování dat.* Praha: Portál, 2004. ISBN 80-7178-820-1. **(kapitola 3).**

2. RIEČANOVÁ Z. a kol. *Numerické metody a matematická štatistika*. Bratislava: Alfa, 1987. ISBN 063-559-87. **(podkapitola 7. 3).**

Děkuji za pozornost.