7. SEMINÁŘ

DESKRIPTIVNÍ STATISTIKA

Statistika

- Nedostatečná znalost cílů, metod a možností statistiky
  - Nezájem a nedůvěra

    X
  - Přílišné přeceňování statistiky

- „S pomocí statistiky je jednoduché lhát, bez ní je ale těžké říci pravdu“.

  A. Dunkels

Počátky - popisná statistika

- **Statistika jako popis státu:**
  - **Popis a soupis** zemědělského, hospodářského a politického stavu země a obyvatelstva
  - **Vyčerpávající šetření** – zachycení veškerého obyvatelstva pomocí sčítání lidu a vedení podrobných záznamů o demografických, geografických a hospodářských jevech
  - Heslo: čísla, stále více a stále úplnější

Moderní (induktivní) statistika

- 30. léta 20. století – rozvoj **teorie pravděpodobnosti** a revoluce ve statistice
- **Výběrová šetření** – nové možnosti:
  - hlubší analýza výběrového souboru,
  - zkoumání mnoha dosud nezkoumaných jevů,
  - zobecnění výsledků pomocí postupů induktivní statistiky.
- Heslo dnešní statistiky: výběr

Statistika – základní pojmy

Statistika jako vědní obor

- Jejím předmětem jsou **hromadné jevy**
  - Vlastnosti, znaky a události, které se vyskytují ve velkém množství.
- Zabývá se **sběrem, popisem a analýzou dat.**
- **Data**
  - zjištěné (naměřené) hodnoty určitých vlastností
  - hodnoty jednotlivých vlastností se vyznačují **variabilitou**
- **Variabilita dat**
  - Důsledek působení velkého množství drobných **NÁHODNÝCH** vlivů, z nichž každý výslednou hodnotu sledované vlastnosti ovlivňuje jen nepatrně.

Náhoda ve statistice

- **Přirozený jev**, který lze zkoumat exaktními metodami **teorie pravděpodobnosti**.
- Má svoje **zákonitosti**, jsou-li sledované vlastnosti určovány pouze náhodnými vlivy, podléhají zákonitostem náhody.
- Pokud zjištěné údaje neodpovídají těmto zákonitostem, nezpůsobuje rozdíly v hodnotách vlastnosti jen **náhoda**, ale **působení** nějakého jiného faktoru.

Induktivní a deduktivní úvaha

Aplikace statistických metod se váže ke dvěma typům uvažování:

- **Deduktivní úvaha:** využívání obecných znalostí k rozhodování v jednotlivých případech.

- **Induktivní úvaha:** zobecnění poznatků z jednotlivých případů na všechny možné případy.

Základní a výběrový soubor

- **Základní soubor** – soubor jednotek, jejichž vlastnosti chceme poznat.

- **Výběrový soubor** – ta část souboru, u které skutečně probíhá statistické šetření.

Výběrový soubor

- Vypovídá jen o tom základním souboru, ze kterého byl odvozen.

- **Reprezentativnost** výběrového souboru (dobře reprezentuje všechny známé i neznámé charakteristiky základního souboru).

- **Náhodný výběr** – je získán postupem, kdy každý prvek základního souboru má na začátku výběru stejnou naději být vybrán.

Metody náhodného výběru

1. **Prostý náhodný výběr** – losováním, pomocí tabulek (generátoru) náhodných čísel.
2. **Náhodný výběr mechanický** (systematický) – vytvoříme seznam jednotek, ze kterého vybereme např. každou stou osobu , přičemž první osobu vybereme metodou prostého náhodného výběru.
3. **Náhodný výběr oblastní** (stratifikovaný) – rozdělení do oblastí (strat) – např. rozdělíme soubor na muže a ženy a vybíráme prostým NV takový počet mužů a žen, aby byl zachován poměr mužů a žen v základním souboru.

Etapy statistického šetření

1) Plán šetření (cíl, studium literatury, statistická jednotka, základní soubor, sledované znaky, způsob a přesnost měření, forma záznamu, způsob a rozsah výběru, statistické zpracování, pracovní a testované hypotézy, přínos a náklady výzkumu, pilotní studie).
2) Sběr dat (dodržování pravidel těmi, kdo sběr dat provádějí).
3) Popis a technické zpracování (deskriptivní statistika)
4) Rozbory a závěry (induktivní statistika)

Dvě základní oblasti statistiky

- **Popisná statistika**
- **Induktivní statistika**

Deskriptivní statistika - popis dat

Deskriptivní statistika

- statistické třídění
- prezentace dat
- statistické charakteristiky

Statistické třídění

- zpřehlednění souboru dat
- popis struktury souboru
- rozložení četností

Způsob třídění závisí na typu veličiny.

Třídění: typy veličin (znaků)

- Věk, pohlaví, výška hmotnost, VKP, nemoc, vzdělání. kuřáctví
- 50ti-letý muž, měří 170 cm, váží 90 kg, vitální kapacitu plic má 4,62 l, prodělal zánětlivé plicní onemocnění, má středoškolské vzdělání a je nekuřák.

KVALITATIVNÍ

- nominální

  – alternativní
  – množné

- ordinální

KVANTITATIVNÍ

- diskrétní

- spojité

Třídění kvalitativních veličin

- Kategorie třídění jsou předem dány.

- Jde o výčet všech hodnot, kterých může sledovaný znak nabývat (např. znak vzdělání – hodnoty znaku: ZŠ, SŠ, VŠ).

Třídění kvantitativních veličin

- Vytváříme třídy teprve na základě získaných dat
- Dochází k **redukci dat** ve prospěch přehlednosti

- **Vytváření intervalů:**
  - počet intervalů
  - délka intervalů
  - hranice intervalů

- **Musíme brát v úvahu:**
  - počet dat (velikost souboru)
  - přesnost měření
  - cíl třídění

Prezentace dat

Tab. 1.: Rozložení vitální kapacity plic u 200 mužů ve věku 40-50 let (v litrech)

interval	střed	absolut. četnost
3,00 – 3, 39	3,20	**6**
3,40 – 3,79	3,60	**9**
3,80 – 4,19	4,00	**16**
4,20 – 4,59	4,40	**36**
4,60 – 4,99	4,80	**52**
5,00 – 5,39	5,20	**44**
5,40 – 5,79	5,60	**22**
5,80 – 6,19	6,00	**11**
6,20 – 6,59	6,40	**4**
celkem		**200**

Třídění kvantitativních veličin

Výskyt dětské nemoci podle věku:

Věk	Abs. četnost
1	18
2	43
3	50
4	60
5	36
6	25
7	22
8	21
9	6
10	5
11-15	14
16-20	3

Třídění jednostupňové a vícestupňové

- Třídění podle jednoho znaku.
- Třídění podle dvou a **více** znaků současně.

	CELKEM
Nekuřák	120
Slabý kuřák	60
Silný kuřák	20
CELKEM	200

	ZŠ	SŠ	VŠ	CELKEM
Nekuřák	20	40	60	120
Slabý kuřák	35	10	15	60
Silný kuřák	12	7	1	20
CELKEM	67	57	76	200

Prezentace dat

Prezentace dat v tabulkách a grafech

- Četnost jednotlivých kategorií
- Tvar rozložení četností
  - Symetrické x asymetrické
  - Jednovrcholové x dvouvrcholové
  - Výpočet ukazatelů polohy (a variability)
  - Výběr vhodného teoretického rozložení četností při odhadu parametrů a testování hypotéz

Prezentace dat v tabulkách

- Výsledky třídění uvádíme v tabulkách – tzv. **tabulky rozdělení četností.**
- **Četnosti:**
  - absolutní
  - relativní
  - kumulativní absolutní
  - kumulativní relativní

Prezentace dat

Tab. 1.: Rozložení vitální kapacity plic u 200 mužů ve věku 40-50 let (v litrech)

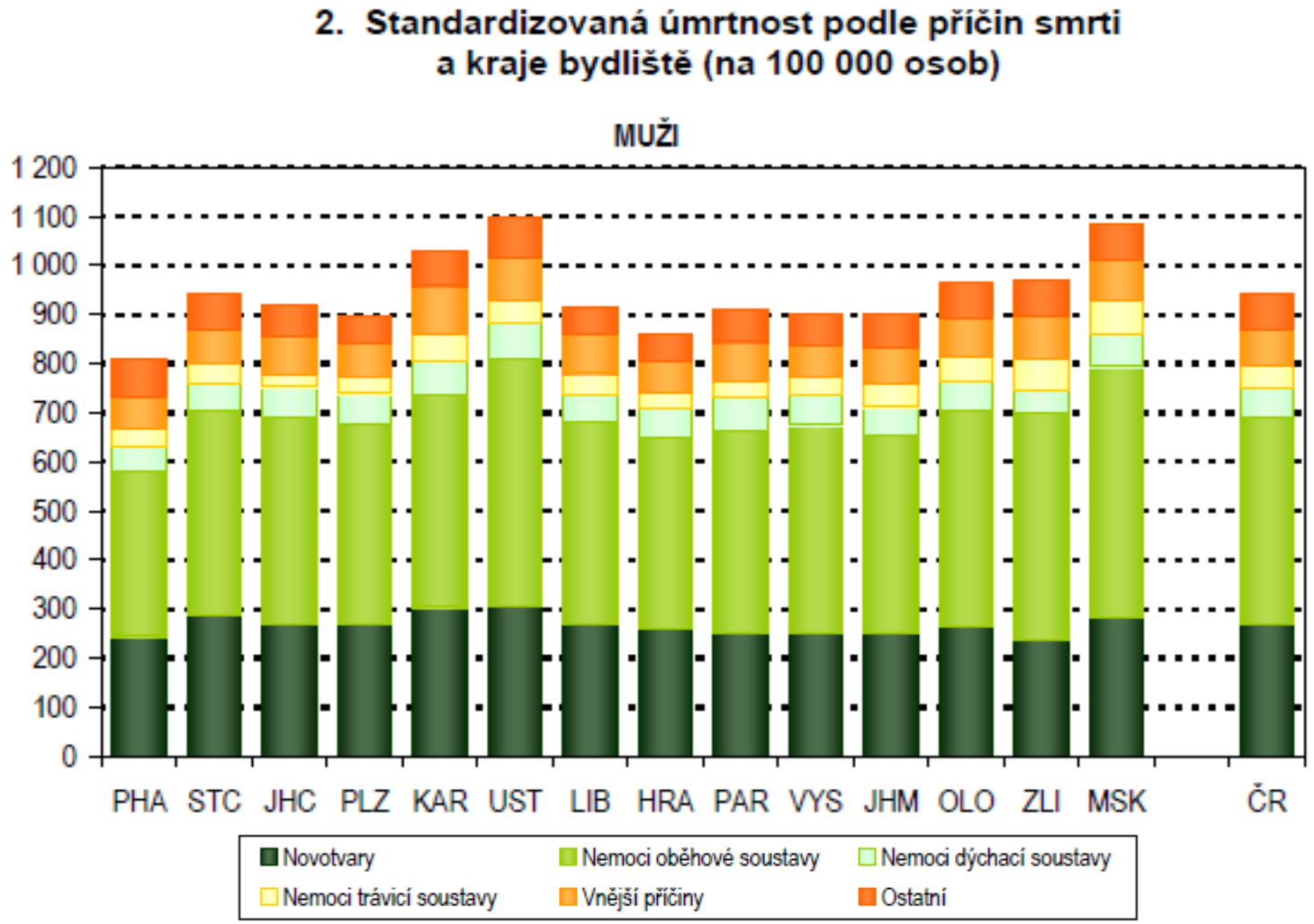
interval	střed	četnost		kumulativní četnost	
		absolut.	relat. %	absolut.	relat. %
3,00 – 3, 39	3,20	6	3,0	6	3,0
3,40 – 3,79	3,60	9	4,5	15	7,5
3,80 – 4,19	4,00	16	8,0	31	15,5
4,20 – 4,59	4,40	36	18,0	67	35,5
4,60 – 4,99	4,80	52	26,0	119	59,5
5,00 – 5,39	5,20	44	22,0	163	81,5
5,40 – 5,79	5,60	22	11,0	185	92,5
5,80 – 6,19	6,00	11	5,5	196	98,0
6,20 – 6,59	6,40	4	2,0	200	100,0
celkem		**200**	**100,0**		

Prezentace dat v grafech

- **Kvalitativní veličiny**
  - Sloupcový graf (sloupce oddělené mezerou)
  - Výsečový graf (struktura)
  - Kartogram (regionální srovnání)

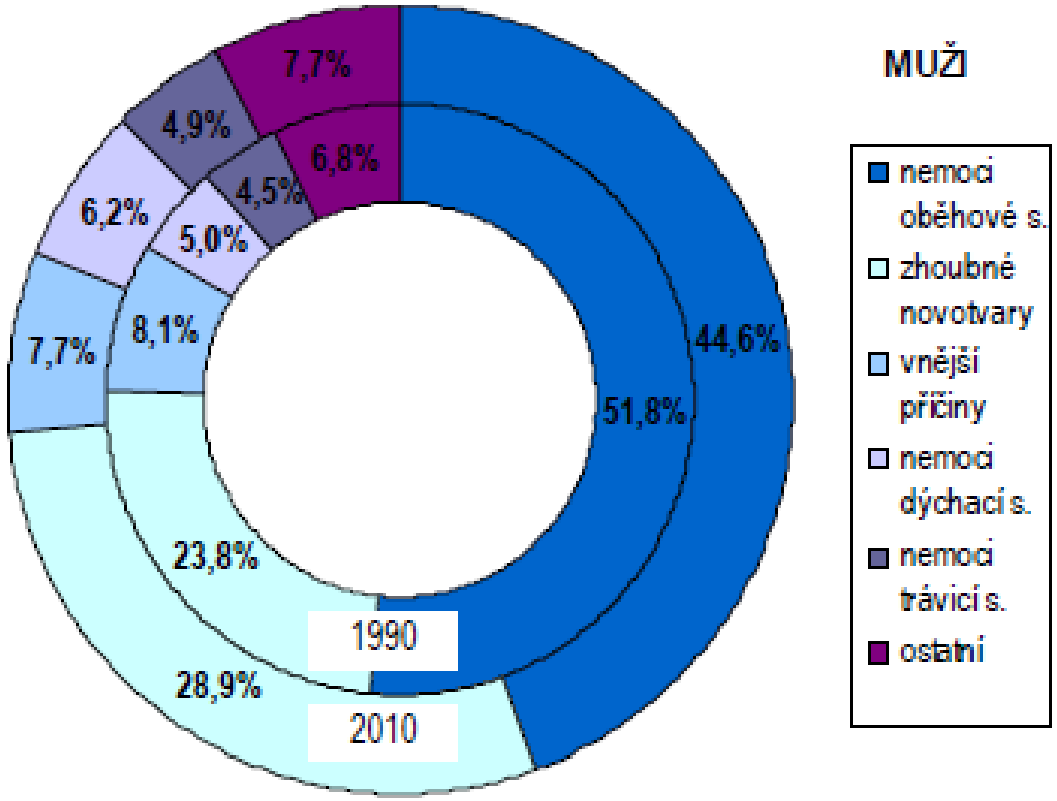
- **Kvantitativní veličiny**
  - Sloupcový graf
  - Histogram
  - Polygon četností

Sloupcový graf

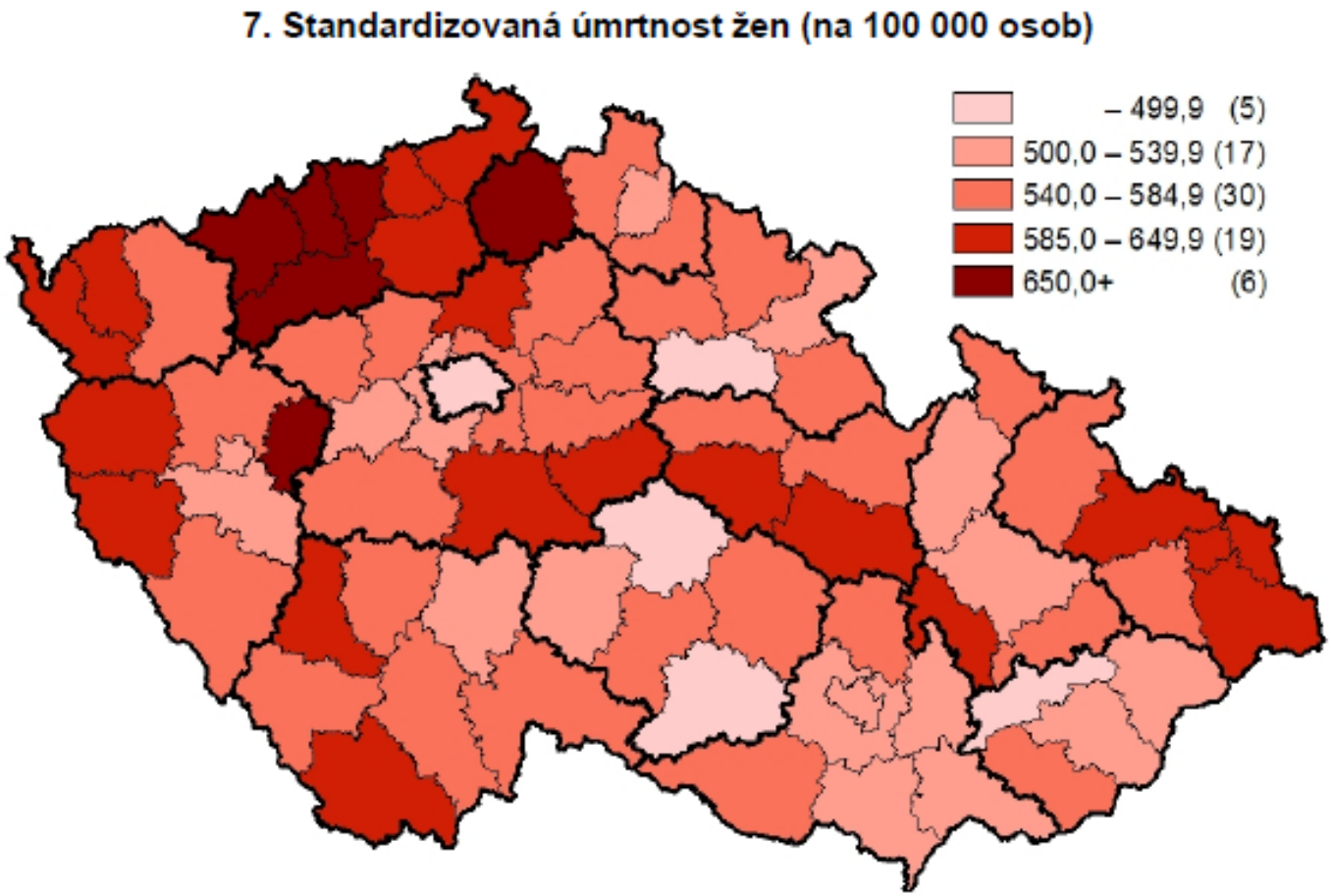


Výsečový graf

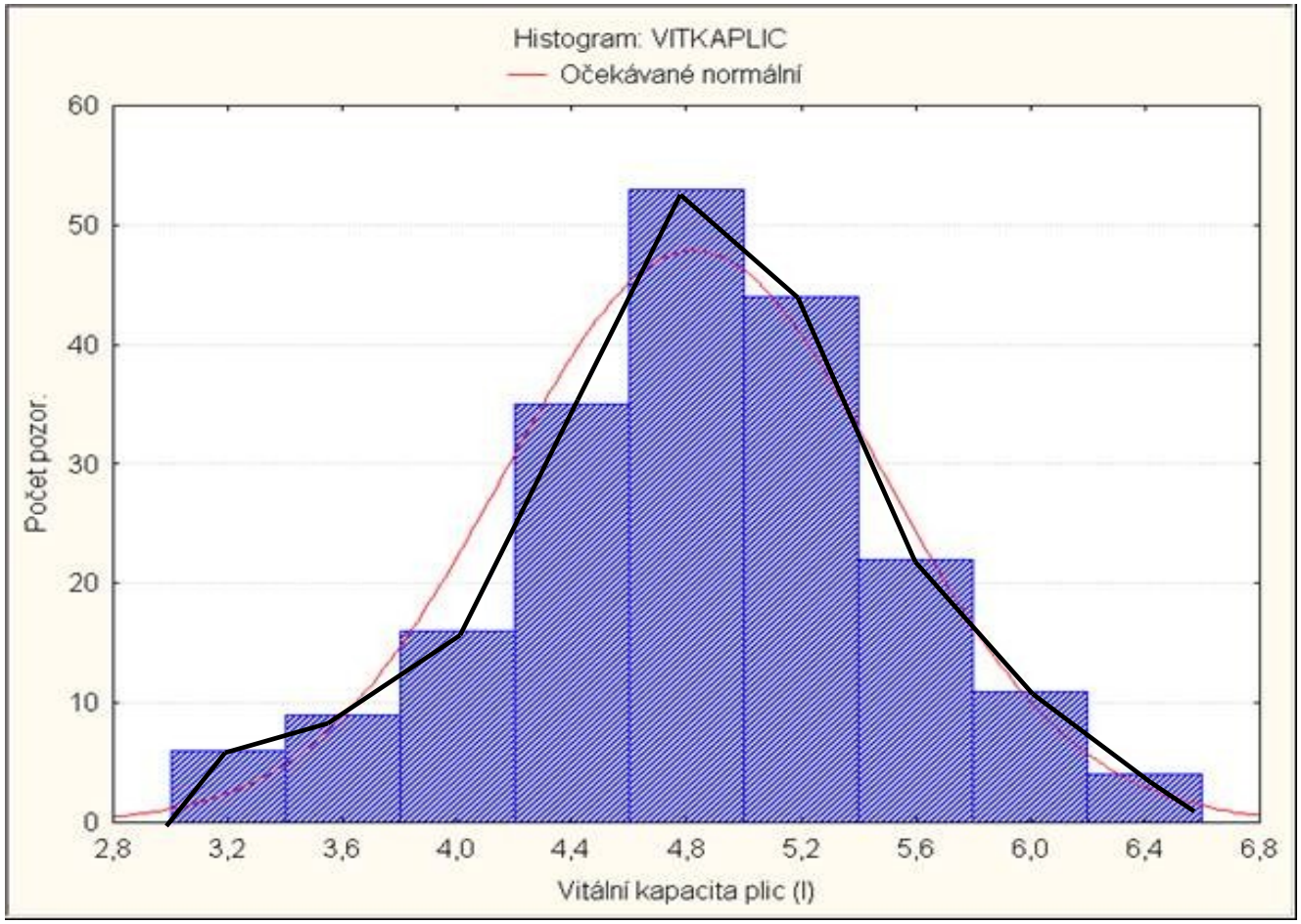
Struktura zemřelých podle příčin v letech 1990 a 2010



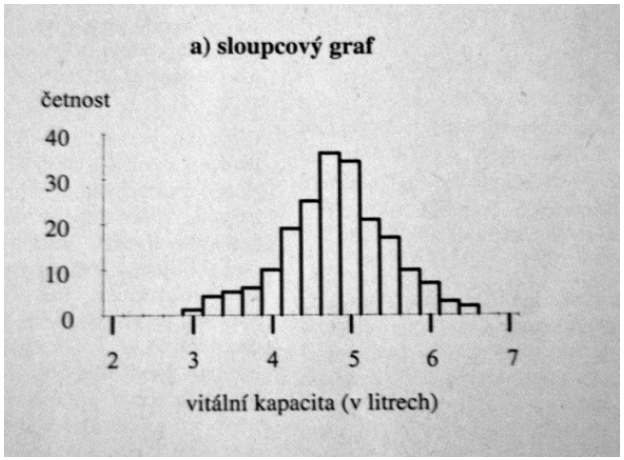
Kartogram

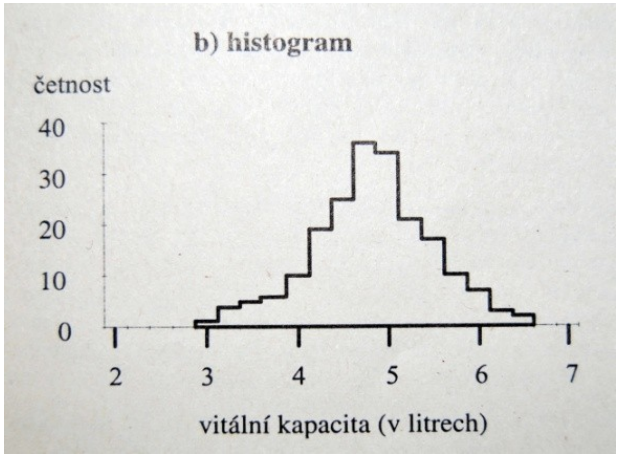

7. Standardizovaná úmrtnost žen (na 100 000 osob)
– 499,9 (5)
500,0 – 539,9 (17)
540,0 – 584,9 (30)
585,0 – 649,9 (19)
650,0+ (6)

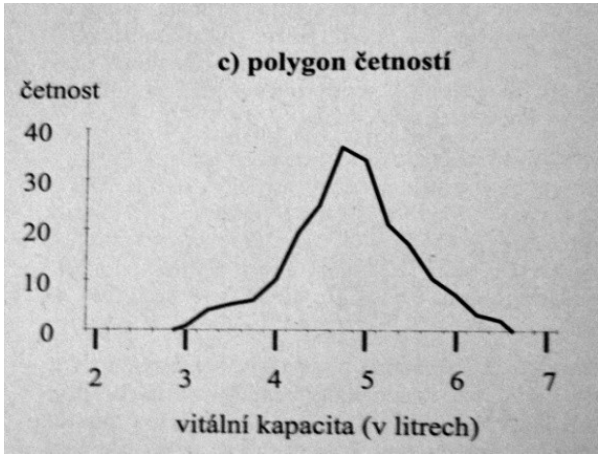
Prezentace kvantitativních dat



Prezentace dat v grafech







osa **X** : naměřené hodnoty sledování veličiny
osa **Y** : četnost intervalů (abs. nebo v %)

Tvar rozložení četností:

- Symetrické x asymetrické
- Jednovrcholové x vícevrcholové
- Podoba s teoretickými modely rozložení četností

Statistické ukazatele

a) **relativní ukazatele**

b) **střední hodnoty (ukazatele polohy)**

c) **ukazatele variability**

VOLBA VHODNÝCH UKAZATELŮ POLOHY A VARIABILITY ZÁVISÍ NA TYPU SLEDOVANÉHO ZNAKU (nominální x ordinální x intervalový) **A NA TVARU ROZLOŽENÍ ČETNOSTÍ** (symetrické x asymetrické)**.**

Ukazatele polohy

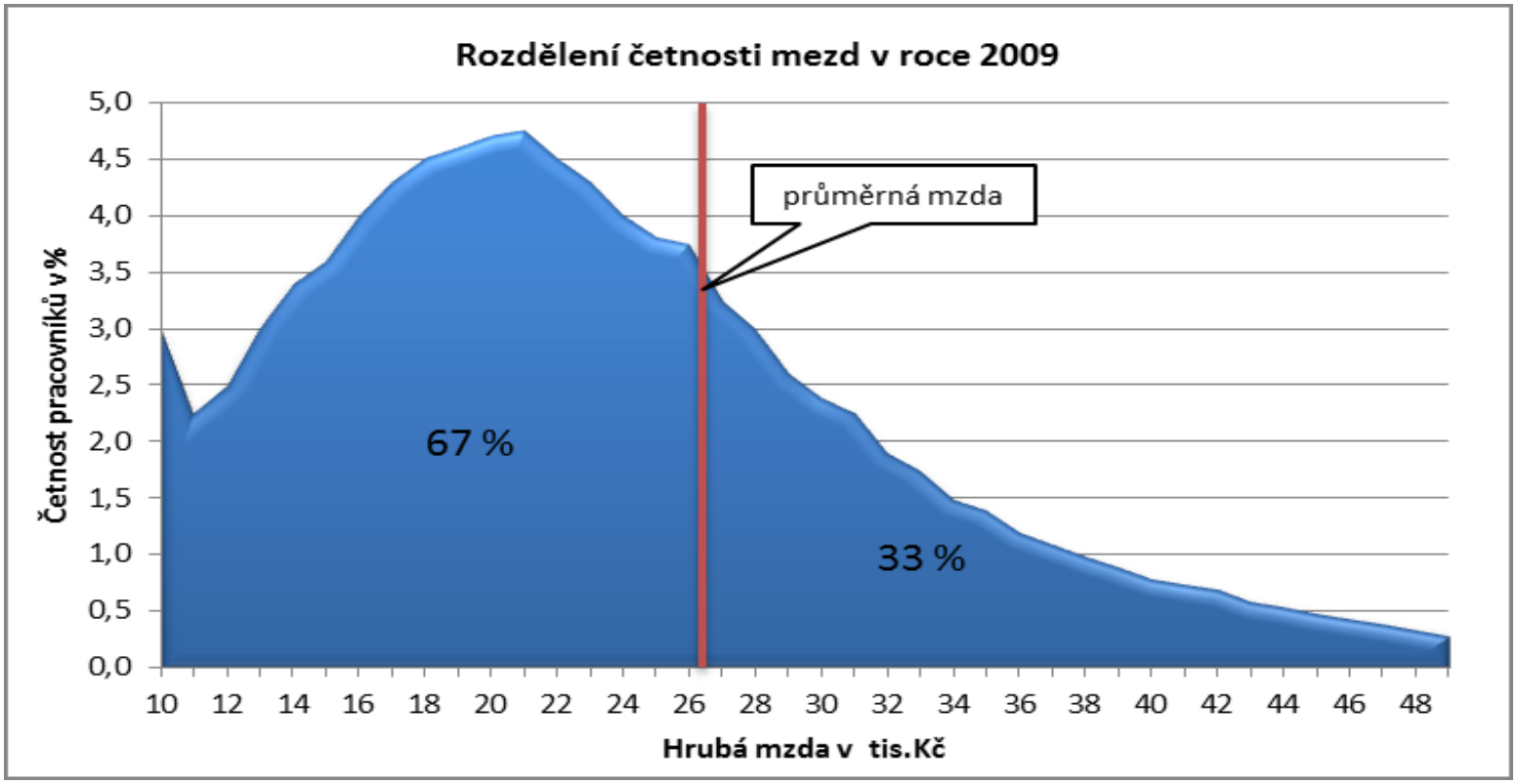
- **Aritmetický průměr (m):**
  - sečteme pozorované hodnoty a vydělíme je počtem sledovaných jednotek
- **Medián (m_e):**
  - hodnota, která je právě uprostřed všech pozorování, která jsme seřadili podle velikosti
- **Modus (m_o):**
  - třída (kategorie) s nejvyšší četností
- **Kvantil (percentil, decil, kvartil)**
  - pořadový ukazatel, obměna mediánu

Ukazatele polohy

- **Typ veličiny:**
  - nominální: modus
  - ordinální: modus, medián, percentil (kvantil)
  - intervalové: modus, medián, percentil (kvantil), průměr

- **POZOR NA INTERPRETACI ARITMETICKÉHO PRŮMĚRU U ASYMETRICKÝCH ROZLOŽENÍ.**
- **ARITMETICKÝ PRŮMĚR JE CITLIVÝ NA VYCHÝLENÉ HODNOTY.**
- **VHODNĚJŠÍM UKAZATELEM POLOHY U ASYM. ROZLOŽENÍ MŮŽE BÝT V URČ. PŘÍPADECH MEDIÁN.**

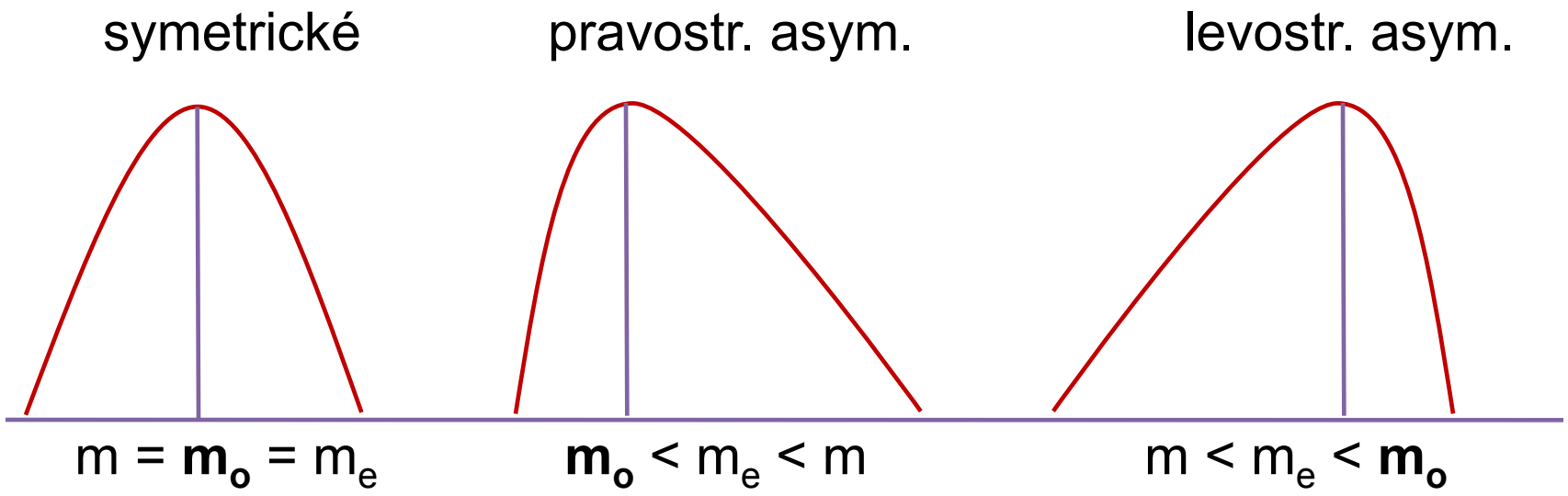
Ukazatele polohy



$m = 26\,700 \quad m_o = 20\,000 \quad m_e = 22\,000$

Ukazatele polohy

- Ukazatele polohy u symetrického a asymetrického rozložení



Ukazatele variability

Proč nestačí ukazatele polohy k výstižnému popisu dat?

Př.

1. sk.: 3,08 4,42 5,05 5,67 6,59 m = 4,96

2. sk.: 4,86 4,90 4,91 5,03 5,11 m = 4,96

Obě skupiny mají stejný průměr, liší se ale kolísáním hodnot, tj. **VARIABILITOU**

Ukazatele variability

Spolu se střední hodnotou by se měl vždy udávat příslušný ukazatel variability!

- **Rozpětí** (u malých souborů, kde **n ≤ 10)**
- **Rozptyl - směrodatná odchylka (nejč.) – variační koeficient**
    - uvádějí se s aritmetickým průměrem (u symetrických rozdělení)
- **Kvantily** (percentily, decily, kvartily)
    - uvádějí se s modem či medián (asymetrický rozdělení)
    - lze je ale samozřejmě použít i s aritmetickým průměrem

Ukazatele variability

Rozpětí:

- max. - min. Pro n ≤ 10

Rozptyl (s^2):

- Průměr čtverců odchylek aritmetického průměru od jednotlivých měření:
- $s^2 = \frac{\Sigma(xi - m)2}{n}$

1.sk.: 3,08 4,42 5,05 5,67 6,59 m = 4,96

4,96 – 3,08 = 1, 88	3,53	
– 4,42 = 0,54	0,29	
– 5,05 = - 0,09	0,01	$s^2 = 6,99/5 = 1,40$
– 5,67 = -0,71	0,50	
– 6,59 = -1, 63	2,66	

- Udává se ve čtvercích jednotek sledovaného znaku, tj. zde v litrech^2

Ukazatele variability

Směrodatná odchylka (s):

- Odmocněný rozptyl, $s=\sqrt{s^2}$
- Ukazatel variability udávaný ve stejných jednotkách jako sledovaný znak.
- Za předpokladu normálního rozdělení četností vypovídá o tom, **o kolik se většina hodnot sledovaného znaku odchyluje od průměru.**

$m \pm 1s$	interval, ve kterém leží 68% naměřených hodnot
$m \pm 2s$	interval, ve kterém leží 95% naměřených hodnot
$m \pm 3s$	interval, ve kterém leží 99% naměřených hodnot

– **Příklad:** vypočítejte, v jakém intervalu leží 68% hodnot VKP
V našem souboru 200 mužů.

Ukazatele variability

Variační koeficient (v.k.)

- Relativní ukazatel variability
- Udává, jaký podíl tvoří směrodatná odchylka z průměru.
- Je-li vetší než 50%, pak je soubor natolik nesourodý, že nemá smysl ho charakterizovat aritmetickým průměrem.

Ukazatele variability

Variační koeficient (v.k.)

- **Slouží ke srovnání variability 2 souborů, jejichž průměry se značně liší**

  Př.: VKP u mužů a u žen

M:	m = 4, 80	s = 0,66	v.k. = 13,8%
Ž:	m = 3, 90	s = 0,42	v.k. = 10,8%

- **Slouží ke srovnání variability znaků uváděných v různých jednotkách**

  Př.: VKP (l), výška (cm) a hmotnost mužů (kg)

VKP:	m = 4,80	s = 0,66	v.k. = 13,8%
Výška:	m = 178	s = 4	v.k. = 2,2%
Hmotnost:	m = 82	s = 6	v.k. = 7,3%

Příklad

Porodní délka 5 novorozenců v cm:

49, 50, 50, 51, 53

Vypočítejte:

- Aritmetický průměr
- Rozptyl
- Směrodatnou odchylku
- Variační koeficient

Příklad - řešení

x_i	$x_i - m$	$(x_i - m)^2$
49	- 1,6	2,56
50	- 0,6	0,36
50	- 0,6	0,36
51	0,4	0,16
53	2,4	5,76
253	0,0	9,20

$m = 253 : 5 = 50,60$

$s^2 = 9,20 : 5 = 1,84$

$s = \sqrt{1,84} = 1,35$

$v.k. = (1,35 : 50,60) \cdot 100$
$= 1,98\ \%$

Ukazatele variability

Kvantily – percentily, decily, kvartily

- Kvantily dělí soubor dat uspořádaných podle velikosti na části obsahující stejný podíl z celkového počtu jednotek
- Variabilita se určuje pomocí intervalu, ve kterém se pohybuje nejčastěji 80% ($P_{10} – P_{90}$) nebo 50% ($P_{25} – P_{75}$) pozorování.
- Postup výpočtu:
  1. Určíme hodnotu pozorování, které představuje 10. percentil = dolní hranice intervalu
  2. Určíme hodnotu pozorování, které představuje 90. percentil = horní hranice intervalu

- **Vhodné ukazatele variability pro asymetrická rozložení.**

Ukazatele variability

Kvantily – percentily, decily, kvartily

- Kvantily dělí soubor dat uspořádaných podle velikosti na části obsahující stejný podíl z celkového počtu jednotek
- Variabilita se určuje pomocí intervalu, ve kterém se pohybuje nejčastěji 80% ($P_{10} – P_{90}$) nebo 50% ($P_{25} – P_{75}$) pozorování.
- Postup výpočtu:
  1. Určíme hodnotu pozorování, které představuje 10. percentil = dolní hranice intervalu
  2. Určíme hodnotu pozorování, které představuje 90. percentil = horní hranice intervalu

- **Vhodné ukazatele variability pro asymetrická rozložení.**