

# 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

- **Starověk**

- Soupis občanů, půdy a všeho, co tvořilo základ státu
- **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, potom nalezené rozdíly pravděpodobně nezpůsobila jen **náhoda**, ale i nějaký **jiný faktor**.

# Oblasti využití statistiky v medicíně

- **Zvládání variability**
  - Variabilita: biologická, podmínek, měřících přístrojů - hodnocení variability, variabilita náhodná x nenáhodná
- **Diagnostika nemocí a identifikace zdravotních problémů společnosti**
  - Pravděpodobnostní závěry na základě mnoha údajů z předchozích obdobných případů (popis příznaků nemoci x počátek thalidomidové aféry)
- **Prognóza léčby a odhad přínosu zdravotnických programů**
  - Pravděpodobnostní odhad dalšího průběhu léčby (vychází z minulých zkušeností podobnými případy)
  - Také o aplikacích populačních zdr. opatřeních se vedou záznamy, které umožňují odhadovat úspěšnost příštích opatření



# Oblasti využití statistiky v medicíně

- **Výběr vhodného medicínského postupu**
  - Dřívější zkušenosti + klinické zkoušky + další důležité aspekty dané metody (ekon. náklady, riziko pro společnost)
- **Řízení systému péče o zdraví**
  - Využívání soustavy rutinních statistik doplňovaných o výběrová šetření
    - velikost a struktura populace,
    - informace o populačních procesech – rození, umírání, migrace,
    - zdravotní stav populace,
    - životní prostředí,
    - životní styl,
    - zdravotnický systém

# 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 (konečný n. nekonečný rozsah)

**Vyčerpávající** (úplné) šetření

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

**Výběrové** šetření

*Výběr a ZS spojuje **statistická indukce**  
(zobecnění výsledků z výběru na ZS)*

# 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.

Mačování není metodou náhodného výběru.

# 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 (deskriptivní) statistika**
  - **východisko k usuzování z výběru na základní soubor (tj. indukci)**
- **Induktivní statistika**
  - odhady parametrů ZS z výběrových charakteristik
  - testování statistických hypotéz
  - hodnocení závislostí kvantit. i kvalit. veličin

# Deskriptivní statistika

*Popis a technické zpracování dat:*

- ❑ **Statistické třídění**

**cíl:** uspořádat a zpřehlednit velký soubor dat

- ❑ **Prezentace dat** (konstrukce **tabulek** a **grafů**)

**cíl:** znázornit rozložení četností sledovaných znaků

- ❑ **Statistické charakteristiky (ukazatele)**

**cíl:** charakterizovat sledované znaky pomocí výstižných ukazatelů



# Třídění

*Rozdělení souboru dat do skupin ( tříd, intervalů)  
podle předem určených třídících znaků.*

- zpřehlednění souboru dat
- popis struktury souboru
- rozložení četností
- produktem třídění je **tabulka  
rozdělení(rozložení) četností**

**Způsob třídění závisí na typu veličiny**

# Třídění: typy veličin (znaků)

**KVALITATIVNÍ** (kategoriální) – **slovní** určení, *nelze měřit číselně*, lze pouze klasifikovat do různých kategorií (pohlaví, věk, ...)

1. **Nominální** – lze vyjádřit pouze slovně, nelze seřadit
  - a) **alternativní** – existují pouze 2 varianty (kuřák x nekuřák, muž x žena)
  - b) **množné** – existují  $> 2$  varianty (diagnózy, barva vlasů, ...)
2. **Ordinální** – lze je seřadit dle nějaké míry (ZŠ – SŠ – VŠ, silný – slabý kuřák – nekuřák)

**KVANTITATIVNÍ** – lze vyjádřit *pouze číselně* (jejich obměny charakterizovány polohou na číselné ose)

1. **Diskrétní** (nespojité) – nabývají oddělených hodnot, vyjádřeny celými čísly (počet cigaret, počet onemocnění)
2. **Spojité** – jejich hodnoty na sebe plynule navazují, desetinná čísla (výška, hmotnost, ...)  
v praxi lze spojité znaky převést na diskrétní

# 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 (shrnutí do tříd)** ve prospěch přehlednosti (! ztráta informací!)
- **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)**

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

# Třídění kvantitativních veličin

- Stejně dlouhé intervaly
- Nestejně dlouhé intervaly  
(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 (**kombinační**).

	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 (**tabulka**)
- Tvar rozložení četností (**graf**)
  - Symetrické x asymetrické, jednovrcholové x dvouvrcholové
  - Výběr vhodného ukazatele pro popis souboru (ukazatel polohy, variability)



# 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
<b>celkem</b>		<b>200</b>	<b>100,0</b>		

# Prezentace dat v grafech

- **Kvalitativní veličiny**

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

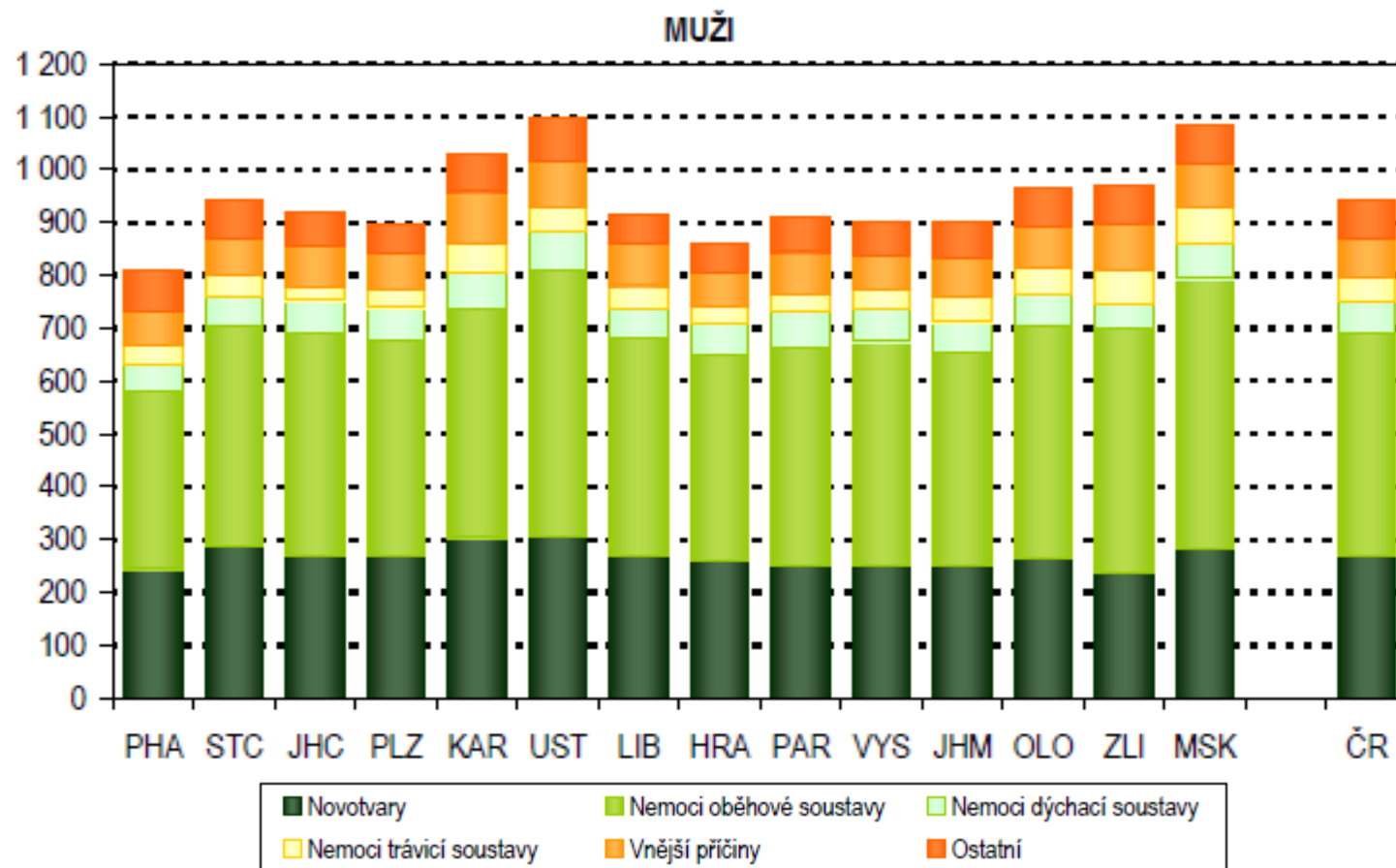
- **Kvantitativní veličiny**

- **Bodový graf**

- Sloupcový graf (plošný graf)
- Histogram (spojnicový graf)
- Polygon četností (spojnicový graf)

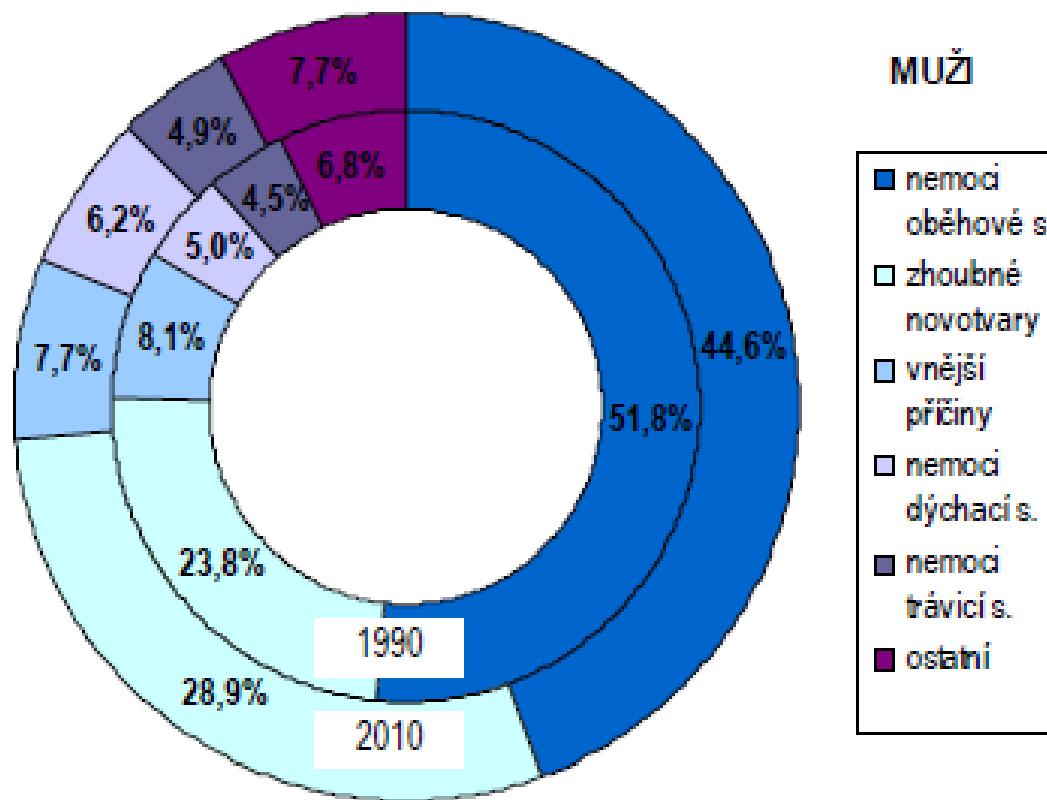
# Sloupcový graf

## 2. Standardizovaná úmrtnost podle příčin smrti a kraje bydliště (na 100 000 osob)



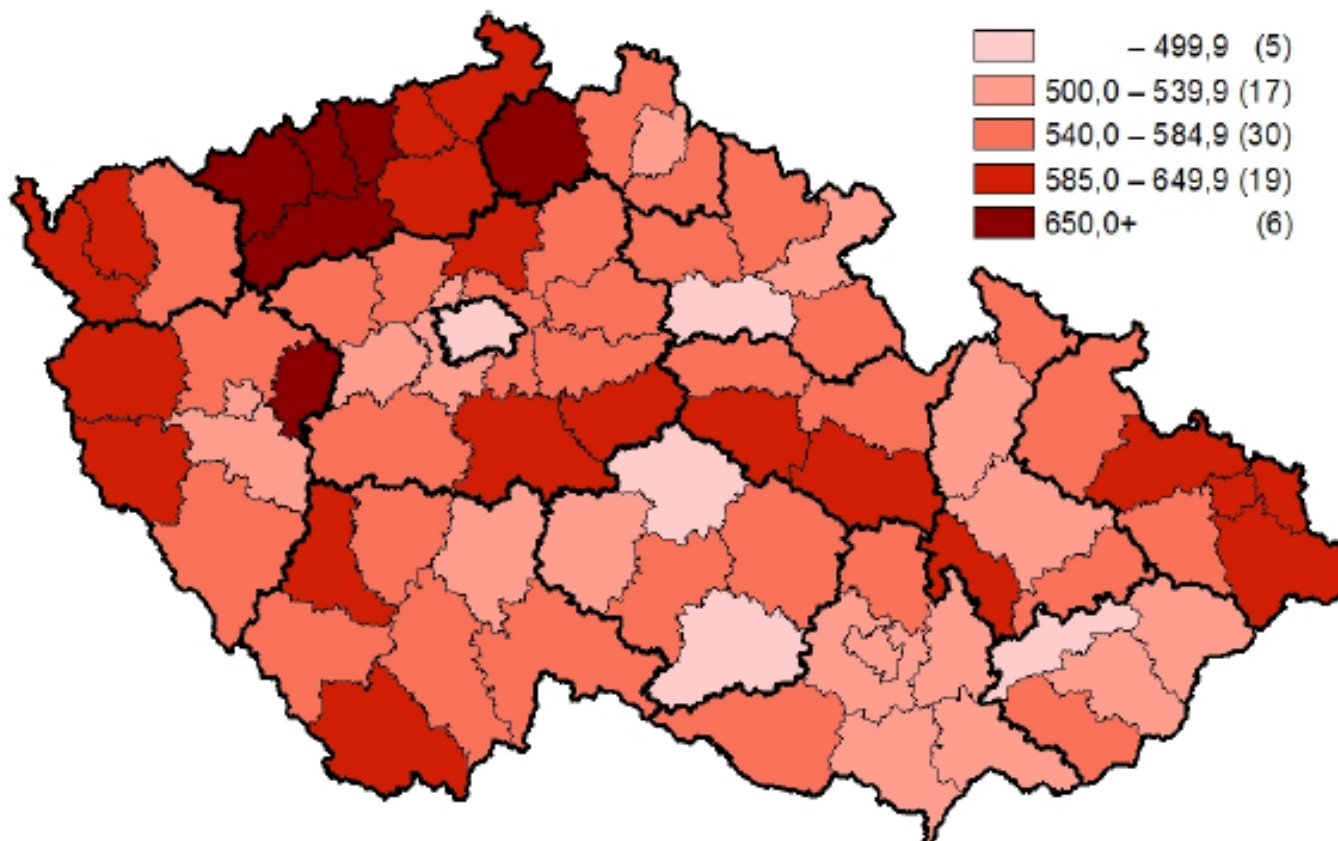
# Výsečový (kruhový) graf

Struktura zemřelých podle příčin v letech 1990 a 2010

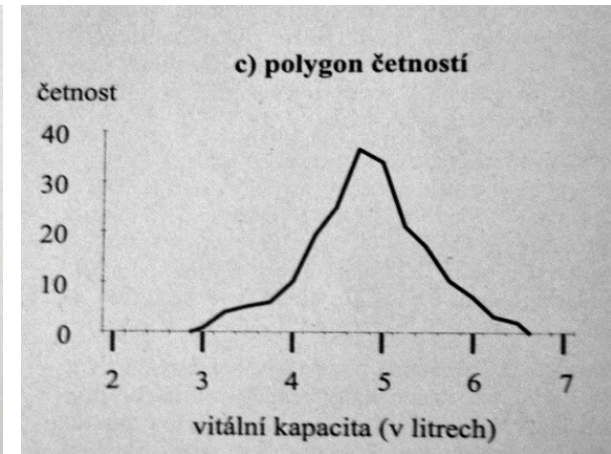
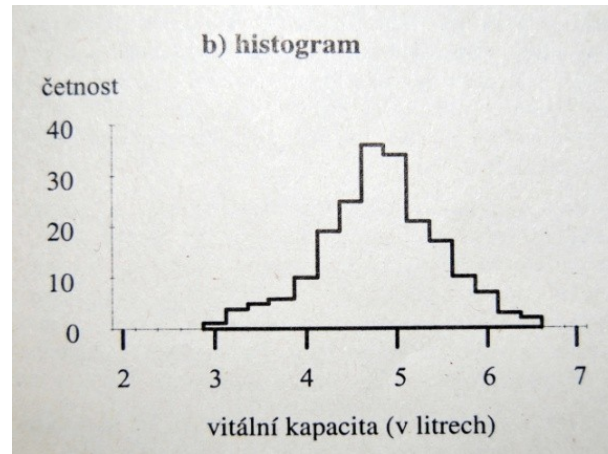
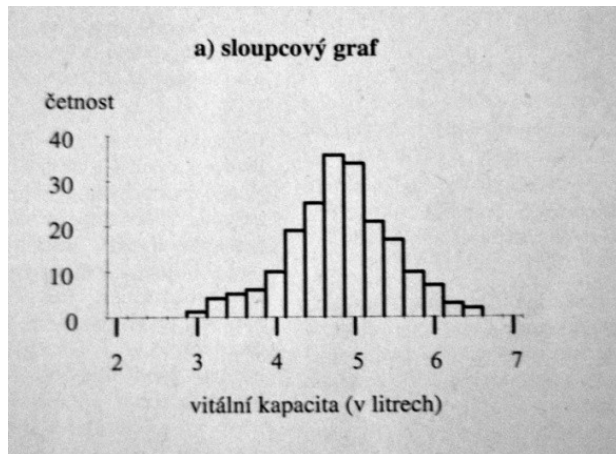


# Kartogram

7. Standardizovaná úmrtnost žen (na 100 000 osob)



# Prezentace dat v grafech

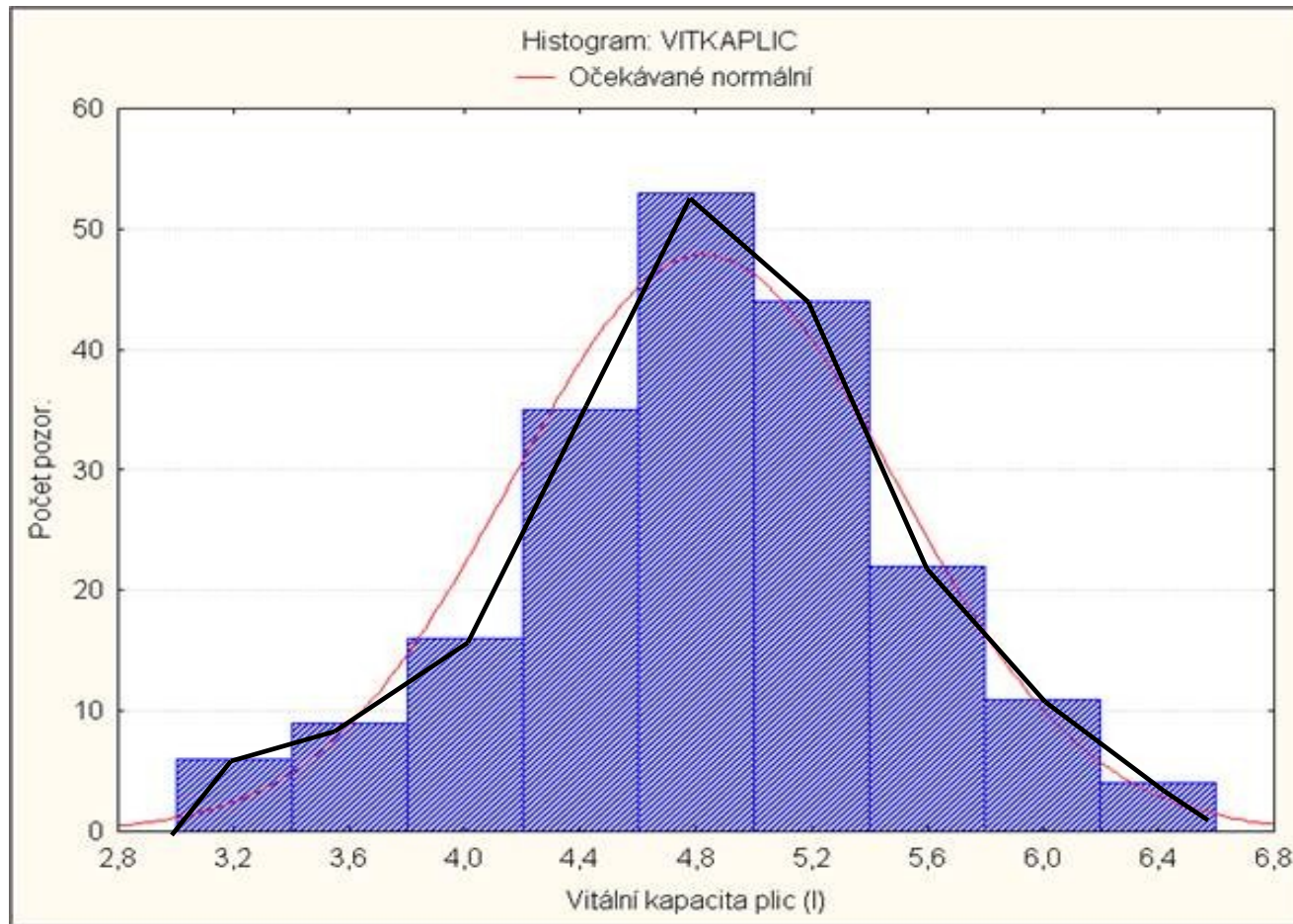


osa **X** : naměřené hodnoty sledování veličiny  
osa **Y** : četnost intervalů (abs. nebo v %)

## Tvar rozložení četností:

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

# Prezentace kvantitativních dat





# Statistické ukazatele

**Kvalit.veličiny - relativní ukazatele** - (viz rutinní statistiky – ukazatele frekvence, ukazatele struktury, indexy)

**Kvantit. veličiny**

## 1) střední hodnoty (ukazatele polohy)

- aritmetický průměr
- medián
- modus
- kvantil, percentil

## 2) ukazatele variability

- rozpětí
- rozptyl – směrodatná odchylka – variační koeficient
- kvantily, percentily (nejméně dva)

**Volba ukazatele:**

1. **Tvar (typ)rozložení (symetrické X asymetrické)**
2. **Typ sledovaného znaku**

# Statistické ukazatele

Ukazatelé polohy i variability charakterizují rozdělení NV jak ve **výběru** (výběrové charakteristiky), tak v celém **základním souboru** (parametry).

- **Výběrové charakteristiky** – náhodné veličiny, jejichž hodnotu počítáme z dat výběrového souboru. Jejich hodnota se mění náhodně výběr od výběru.
- **Parametry** základního souboru- pro daný ZS pevná čísla(neměnné konstanty), jejichž hodnotu neznáme

# Ukazatele polohy

- Většina hodnot, kterých může NV nabývat, se **kupí kolem** nějakého **pevného bodu**, zpravidla kolem středu rozdělené četností.
- Tento bod charakterizuje polohu souboru na číselné ose a ukazatele vystihující tuto vlastnost se nazývají **ukazatele polohy**.

# Ukazatele polohy

- **Aritmetický průměr ( $m$ ):**
  - sečteme pozorované hodnoty a vydělíme je počtem sledovaných jednotek
- **Medián ( $m_e$ ):** pořadová charakteristika
  - hodnota, která je právě uprostřed všech pozorování, která jsme seřadili podle velikosti ( u sudého  $n$  = průměr ze 2 prostř.hodnot)
- **Modus ( $m_o$ ):** nejčastější hodnota, nejvíce typická, leží v modálním intervalu (tj. třída (kategorie) s nejvyšší četností)
- **Kvantil (percentil, decil, kvartil)**
  - pořadový ukazatel, obměna mediánu (=5.decil, 50. percentil)

# 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 ASYMETRICKÝCH ROZLOŽENÍ je MEDIÁN a MODUS.**

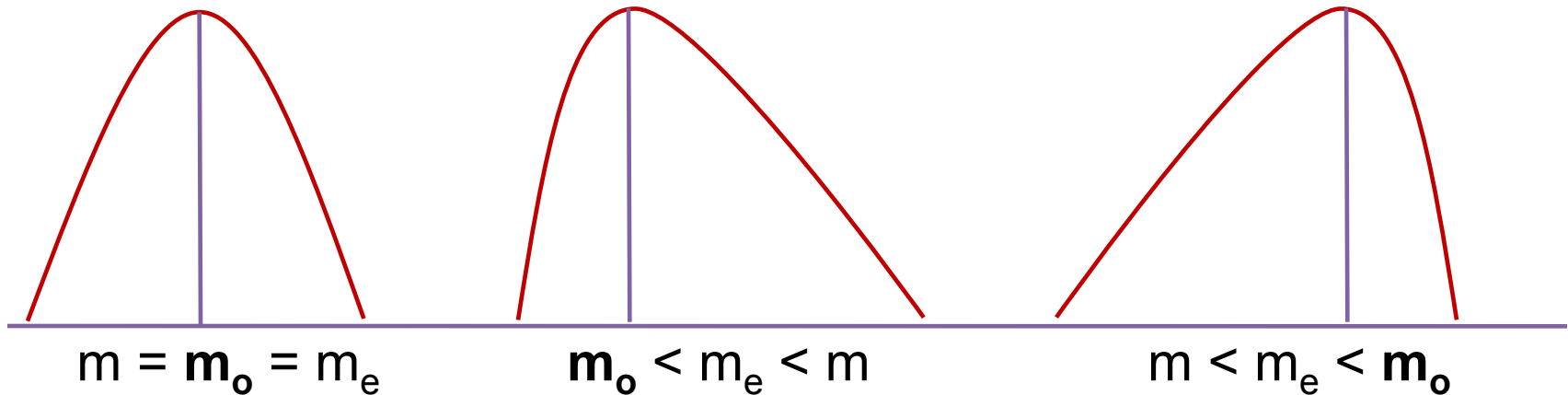
# Ukazatele polohy

- Ukazatele polohy u symetrického a asymetrického rozložení

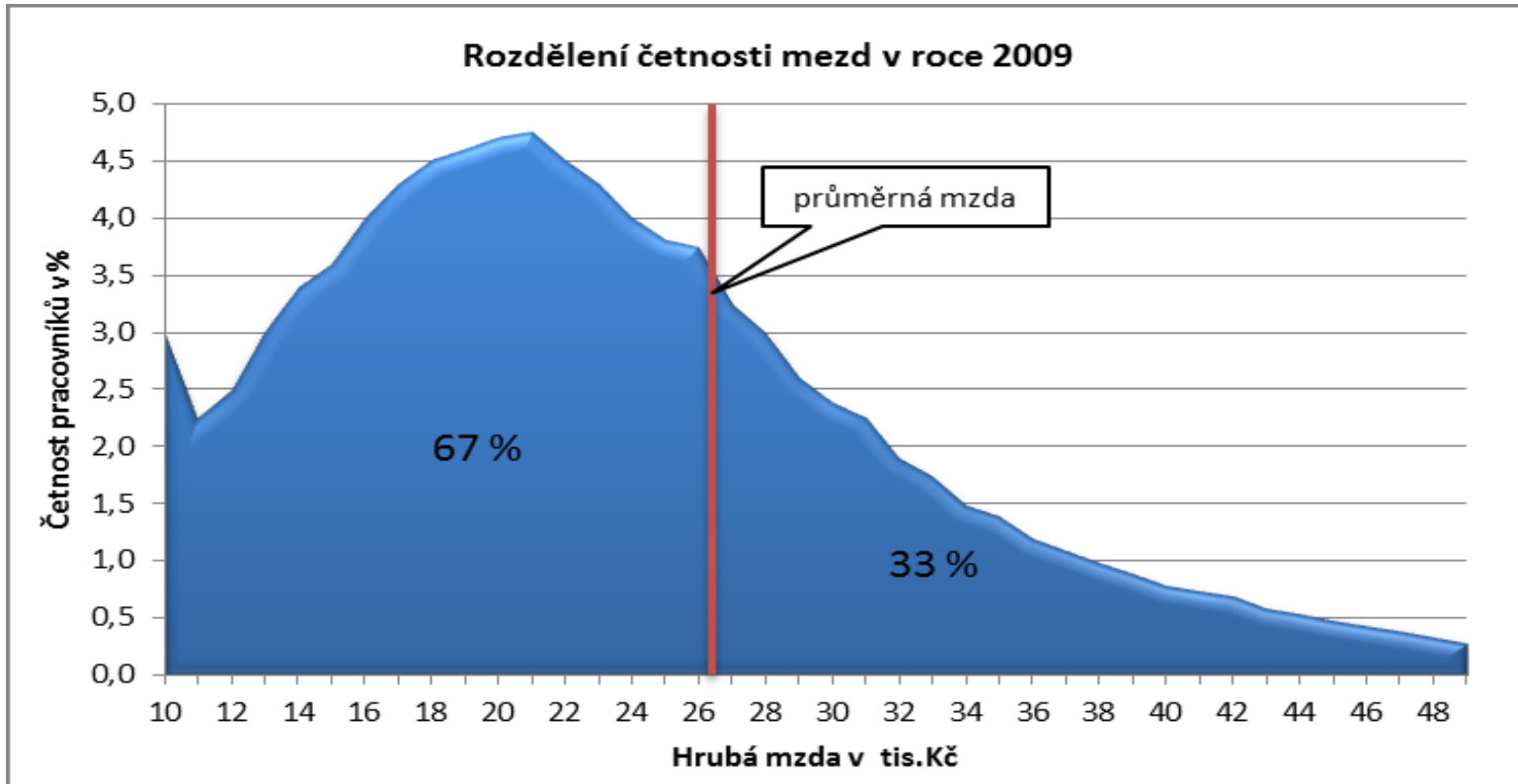
symetrické

pravostr. asym.

levostr. asym.



# Ukazatele polohy



$$m = 26\,700 \quad m_0 = 20\,000 \quad m_e = 22\,000$$

# ÚKOL:

Z výběru 200 mužů (z roztríděných dat), ve kterém jsme měřili VKP určete hodnotu.

- Medián ( $m_e$ )
- $P_{25}$
- $P_{75}$

Výsledek interpretujte.



# 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

- Hodnoty, kterých nabývá NV **kolísají** v **určitém rozmezí kolem středních hodnot.**
- Ukazatelé kvantifikující míru tohoto kolísání (rozptýlení) se nazývají **ukazatelé variability.**

# Ukazatele variability

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

- **Rozpětí** (u malých souborů, kde  $n \leq 10$ )
- **Rozptyl - směrodatná odchylka (nejč.) - variační koeficient**
  - uvádějí se s aritmetickým průměrem ( u symetrický 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  $n \leq 10$

## Rozptyl ( $s^2$ ):

- Průměr čtverců odchylek aritmetického průměru od jednotlivých měření:

- $$s^2 = \frac{\sum(x_i - m)^2}{n-1}$$

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

$$\begin{array}{r} 4,96 - 3,08 = 1,88 \\ - 4,42 = 0,54 \\ - 5,05 = -0,09 \\ - 5,67 = -0,71 \\ - 6,59 = -1,63 \end{array} \quad \begin{array}{r} 3,53 \\ 0,29 \\ 0,01 \\ 0,50 \\ 2,66 \end{array}$$

$$s^2 = 6,99/4 = 1,75$$

- Udává se ve čtvercích jednotek sledovaného znaku, tj. zde v litrech<sup>2</sup>

# Ukazatele variability

## Směrodatná odchylka (s):

- Odmocněný rozptyl,  $s = \sqrt{s^2}$
  - Ukazatel variability udávaný ve stejných jednotkách jako sledovaný znak
  - 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 :  $(s/m) \times 100$
- Je-li větší 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 měrných jednotkách

Př.: VKP (I), 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	$m = 253 : 5 = 50,60$
50	- 0,6	0,36	$s^2 = 9,20 : 4 = 2,30$
50	- 0,6	0,36	
51	0,4	0,16	$s = \sqrt{2,3} = 1,52$
53	2,4	5,76	
<hr/>			
253	0,0	9,20	$v.k. = (1,52 : 50,60) \cdot 100$ $= 3,00\%$

# Ukazatele variability pro asymetrická rozložení četností

## Kvantily – percentily, decily, kvartily

- Kvantily dělí soubor dat uspořádaných podle velikosti na části obsahující stejný podíl z celkového počtu jednotek
- Variabilitu vyjadřujeme pomocí dvou kvantilů (percentilů, decilů)
- 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**

# Jaké charakteristiky použít? (dle typu rozložení)

- **Symetrická rozdělení**
  - průměr
  - směrodatná odchylka
- **Asymetrická rozdělení**
  - **modus, medián, 2 percentily**  
(např. P10 P25 me P75 P90)

*Transformace (např. logaritmická) – převede nesymetr. rozdělení na symetr., pak lze použít  $m$  a  $s$*

# Úkol:

Máme soubor 200 hodnot VCP, které jsme naměřili ve výběru 200 mužů (40-50 let)...

$$\underline{n = 200, m = 4,824, s = 0,668}$$

Stanovte (pomocí směr.odchylky a průměru) hranice - tj. intervaly, ve kterých se nachází 68 %, 95% a 99,7% naměřených hodnot VCP. (na 2 desetinná místa)

## Úkol:

Které percentily odpovídají  
jednonásobku a dvojnásobku  
směrodatné odchylky ?  
(tj. intervaly  $m \pm s$  a  $m \pm 2s$   
vyjádřete pomocí percentilů.)

# Děkuji za pozornost



# Percentilové růstové grafy

**Auxologie** – obor, který se komplexně zabývá růstem a vývojem člověka.

- umožňují pediatrům a rodičům, aby podle návodu připojeného ke grafům průběžně hodnotili všechna základní růstová data dítěte od narození až do jeho osmnácti let (tělesná výška, tělesná hmotnost, obvod hlavy, obvod paže, ...)
- Zároveň je grafy seznamují s **variabilitou** těchto základních antropometrických rozměrů pro každou věkovou skupinu chlapců a dívek současné české populace
- Zcela snadno tak lze zjistit, kolik např. měří nejmenší děti (3. -10. percentil), jak vysoké jsou největší děti (90. – 97. percentil) a kolik měří dítě zcela průměrné (50. percentil).