Deskriptivní statistika

7. seminář

STATISTIKA – úvod (1)

Nezbytné charakteristiky vědeckého výzkumu:

- **přesnost**
- **správnost**
- **spolehlivost**

- Uplatňování matem. modelů, testování hypotéz, srovnání s kontrolní skupinou.

→ umožnila až aplikace metod (po II. svět. válce) **MATEMATICÉ STATISTIKY**

STATISTIKA – úvod (2)

Těžiště statistických metod spočívá v

- racionálním přístupu k řešení problému
- plánování výzkumu
- správné interpretaci a objektivizaci závěrů

poč. 20.stol. → náplň STATISTIKY: hodnocení **ø veličin** + obohaceno o studium jejich **variability**

- vypracována řada MATEMATICKÝCH MODELŮ
- základem STAT: **TEORIE PRAVDĚPODOBNOSTI**
- **Přechod od pouhé registrace k analytickému myšlení**

STATISTIKA – úvod (3)

- Věda, která se zabývá studiem hromadných jevů.
- věda, jejímž předmětem studia jsou **výsledky HROMADNÉHO POZOROVÁNÍ** → jejich sběr, analýza + využití pro rozhodování a předpovědi

  Hromadné jevy – výsledky neomezeně opakovaných pokusů nebo výsledky pozorování na na velkých souborech - např. narození, úmrtí, onemocnění,…
- Hromadné jevy, které nelze před provedením pokusu nebo pozorování zcela přesně předvídat → **HROMADNÉ NÁHODNÉ JEVY**

Pozn.: NÁHODY nelze vyloučit, ale lze je studovat exaktními metodami

Role STATISTIKY v medicíně

využití na úrovni:

- populační – úroveň a vývoj zdrav.situace; zvažuje zdrav.stav lidí, determinanty zdraví + možnost jejich ovlivnění
- individuální – stanovení správné diagnózy (mnoho kvalitat.+kvantitat.údajů) + odhad prognózy léčby

Statistika

- je důležitým nástrojem při šetřeních zdravotního stavu populace a jeho determinant
- vychází z ní moderní epidemiologické metody

Statistika – věda, jež se zabývá výsledky hromadných pozorování, jejich sběrem, analýzou a využitím pro rozhodování a předpovědi.

Data - zjištěné (naměřené) **vlastnosti** statistických jednotek (jednotky stat. šetření)

vlastnosti stat. j., jež ji vymezují – **znaky určující**

vlastnosti stat. j, jež jsou sledovány – **znaky zkoumané** (variabilita)

- hodnoty jednotlivých sledovaných vlastností se vyznačují **variabilitou**

Variabilita dat – důsledkem působení velkého počtu drobných **náhodných** vlivů

Náhoda – přirozený jev, který lze zkoumat exaktními metodami teorie pravděpodobnosti

- neodpovídá-li variabilita dat variabilitě, kterou způsobuje náhoda, pak to lze statistickými metodami určit

Statistická úvaha

Aplikace statistických metod se úzce váže na záměry a úvahy vědeckého pracovníka: **deduktivní úvaha**

induktivní úvaha

Deduktivní úvaha: od obecných zkušeností k jednotlivým (konkrétní)

Induktivní úvaha: od jednotlivých zkušeností k obecným.

Statistické šetření

Vyčerpávající vs. výběrové šetření

- od 30. let 20. století rozvoj teorie pravděpodobnosti a induktivní statistiky

Základní a výběrový soubor

- **ZS**: souhrn prvků (osob, případů nemoci, pokusů), jejichž vlastnosti chceme poznat
- **VS**: vybraná část ZS, kterou budeme skutečně zkoumat (měření, dotazníky, testy)

Náhodný výběr

- každý prvek ZS má na začátku výběru stejnou pravděpodobnost, že bude vybrán do VS

Usuzujeme-li induktivně z vlastností stat. jednotek VS na vlastnosti všech stat. jednotek, hovoříme o **stat. indukci**

Objektivizujeme induktivní závěry pomocí **teorie pravděpodobnosti.**

Náhodný výběr - typy výběrů

podmínka reprezentativnosti

1. **Prostý náhodný výběr** – losováním nebo pomocí tabulek náhod. č.
2. **Náhodný výběr mechanický** (systematický) – např. počáteční písmeno příjmení
3. **Náhodný výběr oblastní** (stratifikovaný) – rozdělení do oblastí, strat a dále výběr z každého vzorku prost. náhod. výběrem nebo systemat. v.
4. **Párový výběr** (mačování) – např. k osobám s jistou vlastností a nemocí osoby se stejnou vlastností a bez nemoci

Pozn.: reprezentativnost může být porušena i při sběru dat, např.

- *neúplné chybějící údaje*
- *nevhodně zvolené otázky*
- *nejednoznačnost odpovědi atd.*

Statistické šetření

Etapy statistického šetření

1. Plán šetření
2. Sběr dat
3. Popis a technické zpracování
4. Rozbory a závěry

J. W. Goethe: „Kdo splete první knoflík, ten se už pořádně nezapne“

Výběrový soubor

Deskriptivní statistika (popis souboru)

1. **Třídění**

   cíl: uspořádat +zpřehlednit velký soubor dat

2. **Prezentace dat**

   tabulky +grafy

   cíl: znázornit rozložení četností sledovaných znaků

3. **Statistické charakteristiky**

   cíl: charakterizovat sledované znaky pomocí výstižných ukazatelů

Způsob třídění závisí na typu veličiny

KVALITATIVNÍ – ***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 kritérií (ZŠ – SŠ – VŠ, silný – slabý kuřák – nekuřák)

KVANTITATIVNÍ – lze vyjádřit ***pouze číselně polohou na číselné ose***

1. **Diskrétní** – vyjádřeny celými čísly (počet cigaret, počet onemocnění)
2. **Spojité –** desetinná čísla (výška, hmotnost, …)

   v praxi lze spojité znaky převést na diskrétní

Třídění veličin

Statistickým tříděním rozumíme rozdělení statistického souboru do skupin podle předem určených třídících znaků.

Třídění kvalitativních veličin

- kategorie jsou ***předem dány***
- jde o výčet všech hodnot, kterých může veličina nabývat (barva očí – modrá, hnědá, zelená, …)

Třídění kvantitativních veličin

- kategorie (třídy) ***vytváříme*** teprve na základě předem získaných dat
- Dochází k ***redukci*** dat ve prospěch přehlednosti

Pozn. Znaky sloužící za podklad třídění musí vyjadřovat podstatu zkoumaného jevu a musí být voleny podle cíle prováděného výzkumu.

Třídění

a) Třídění jednostupňové – rozdělení souboru podle kuřáckých návyků

b) Třídění dvoustupňové – rozdělení souboru podle kuřáckých návyků a vzdělání

c) Třídění třístupňové – rozdělení souboru podle kuřáckých návyků, vzdělání a pohlaví

a

	CELKEM
Nekuřák	389
Slabý kuřák	274
Silný kuřák	261
CELKEM	924

b

	ZŠ	SŠ	VŠ	CELKEM
Nekuřák	269	74	46	389
Slabý kuřák	213	44	17	274
Silný kuřák	197	50	14	261
CELKEM	679	168	77	924

c

MUŽI	ZŠ	SŠ	VŠ	CELKEM
Nekuřák	130	39	27	196
Slabý kuřák	110	24	8	142
Silný kuřák	140	32	8	180
CELKEM	380	95	43	518

ŽENY	ZŠ	SŠ	VŠ	CELKEM
Nekuřák	139	35	19	193
Slabý kuřák	103	20	9	132
Silný kuřák	57	18	6	81
CELKEM	299	73	34	406

Vytváření intervalů

1. **Rozpětí** –od největší naměřené hodnoty odečteme nejmenší
   6,59 – 3,08 = **3,51**
2. **Stanovení počtu intervalů** – závisí na mnoha faktorech (velikost souboru, podrobnost,…) (5-20)
3. **Délka intervalu** – rozpětí/ počet intervalů (např. 10)
   3,51/10 = **0,351** délka 1 intervalu
   pravidlo: a) okrouhlé číslo
   b) ne víc deset. míst než měřená veličina
   zaokrouhlit na **0,40**
4. **Hranice intervalu** – počátek – od nejmenšího čísla 3,08 tj. 3,00

   1. interval 3,00 – 3,39 nebo <3,00 – 3,4)
   2. interval 3,40 – 3,79 <3,4 – 3,8)

Tabulka: Vitální kapacita plic –prezentace dat

OD	DO	Tabulka četností:VITKAPLIC Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
3,000000<=	x<3,400000	6	6	3,00000	3,0000
3,400000<=	x<3,800000	9	15	4,50000	7,5000
3,800000<=	x<4,200000	16	31	8,00000	15,5000
4,200000<=	x<4,600000	35	66	17,50000	33,0000
4,600000<=	x<5,000000	53	119	26,50000	59,5000
5,000000<=	x<5,400000	44	163	22,00000	81,5000
5,400000<=	x<5,800000	22	185	11,00000	92,5000
5,800000<=	x<6,200000	11	196	5,50000	98,0000
6,200000<=	x<6,600000	4	200	2,00000	100,0000
6,600000<=	x<7,000000	0	200	0,00000	100,0000
ChD		0	200	0,00000	100,0000

Proměnná	Popisné statistiky (Tabulka_vitální kapacita.sta) Průměr	Sm.odch.	Minimum	Maximum	N	Počet ChD
VITKAPLIC	4,825550	0,667008	3,060000	6,590000	200	0

Prezentace dat tabulky + grafy

Četnost – kolik z naměřených hodnot spadá do jednotlivých intervalů

Kumulativní četnost – součet všech předchozích intervalů

15 mužů (6 +9) má VC < 3,8

Relativní četnost - % z celkového počtu měření

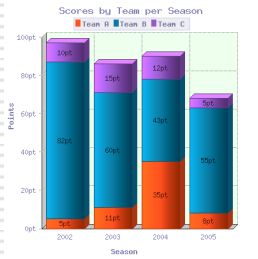
četnost 6 ……. 3% z 200

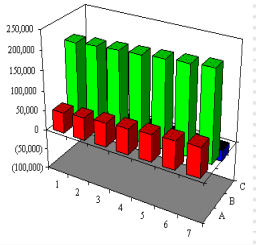
Kumulativní četnost – obdoba kumulativní četnosti v %

Grafy – tvar rozložení

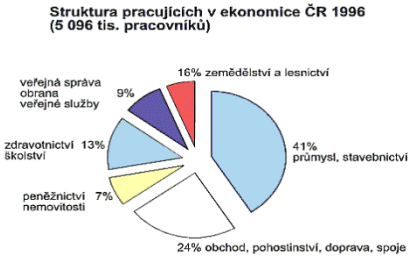
Kvalitativní veličiny

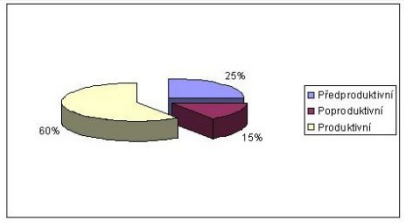
- **Sloupcový graf** (sloupce oddělené mezerou)





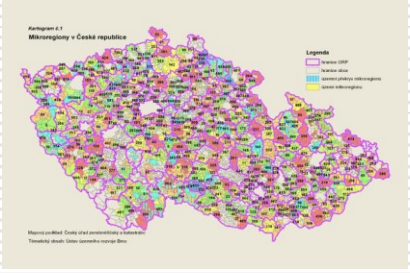
- **Výsečový graf** (struktura)

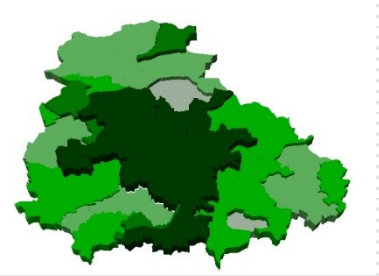




Zdroj: Český statistický úřad

- **Kartogram** (regionální srovnání)





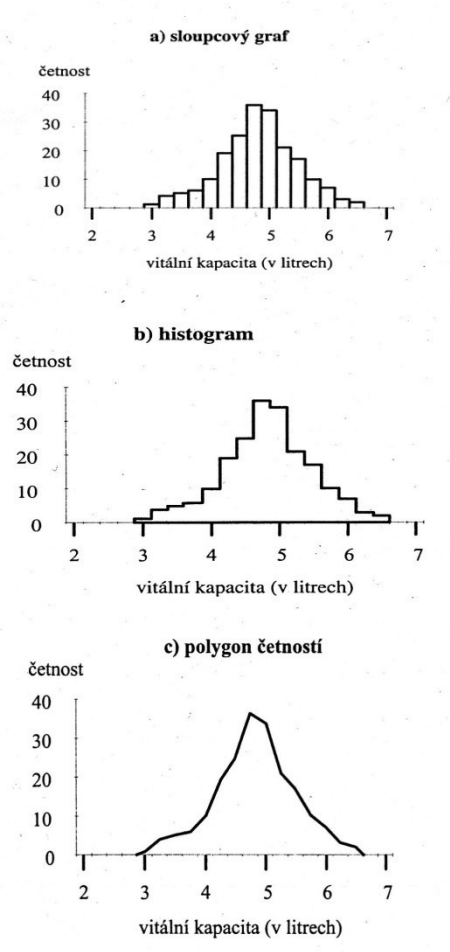
Kvantitativní veličiny

- **Sloupcový graf** (plošné grafy)

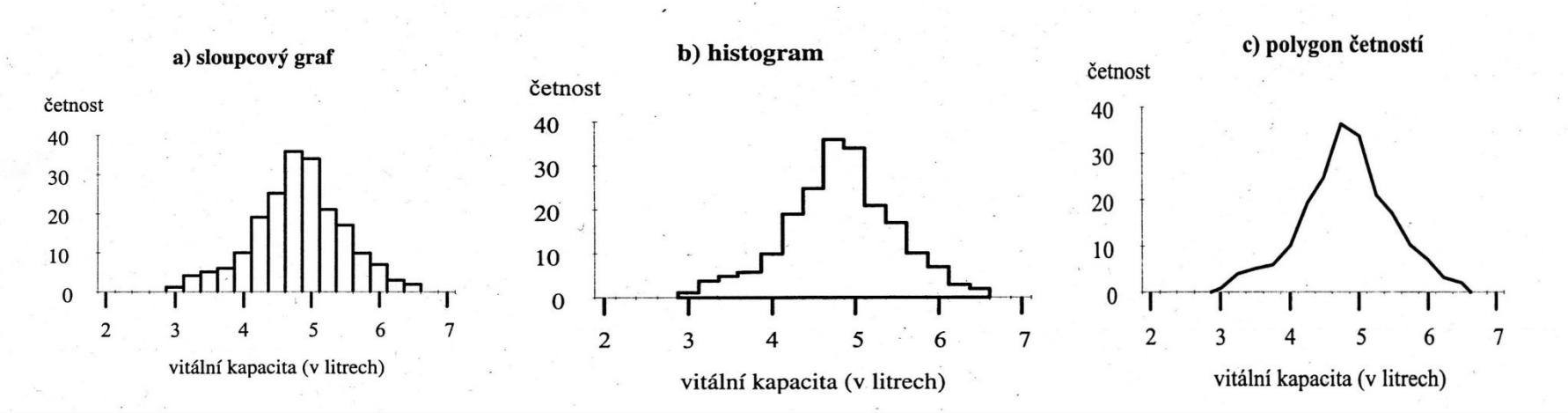
Pro rozložení znaku v několika souborech:

- **Histogram** (pouze obrysy sloupců) (spojnicové grafy)

- **Polygon** (středy sloupců se spojí křivkou) (spojnicové grafy)



Grafy znázorňující frekvenci rozložení veličiny



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

Tvar rozložení četností:

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

Statistické charakteristiky

Výběrové charakteristiky – charakteristiky náhodných veličin ve V.S. (mění se výběr od výběru).

Parametry – charakteristiky náhodných veličin v Z.S. (neměnné konstanty).

Statistické charakteristiky

a) **relativní ukazatele** – viz RS

b) **ukazatele polohy** (střední hodnoty) – **aritmetický průměr**
- medián
- modus
- kvantil, percentil

c) **ukazatele variability – variační šíře (rozpětí)**
- rozptyl
– směrodatná odchylka
– variační koeficient

Volba ukazatele:

1. **Tvar rozložení** (sym. **X** asym.)
2. **Typ sledovaného znaku**

nominální: modus
ordinální: modus, medián, percentil
intervalové: medián, modus, percentil, aritmetický průměr

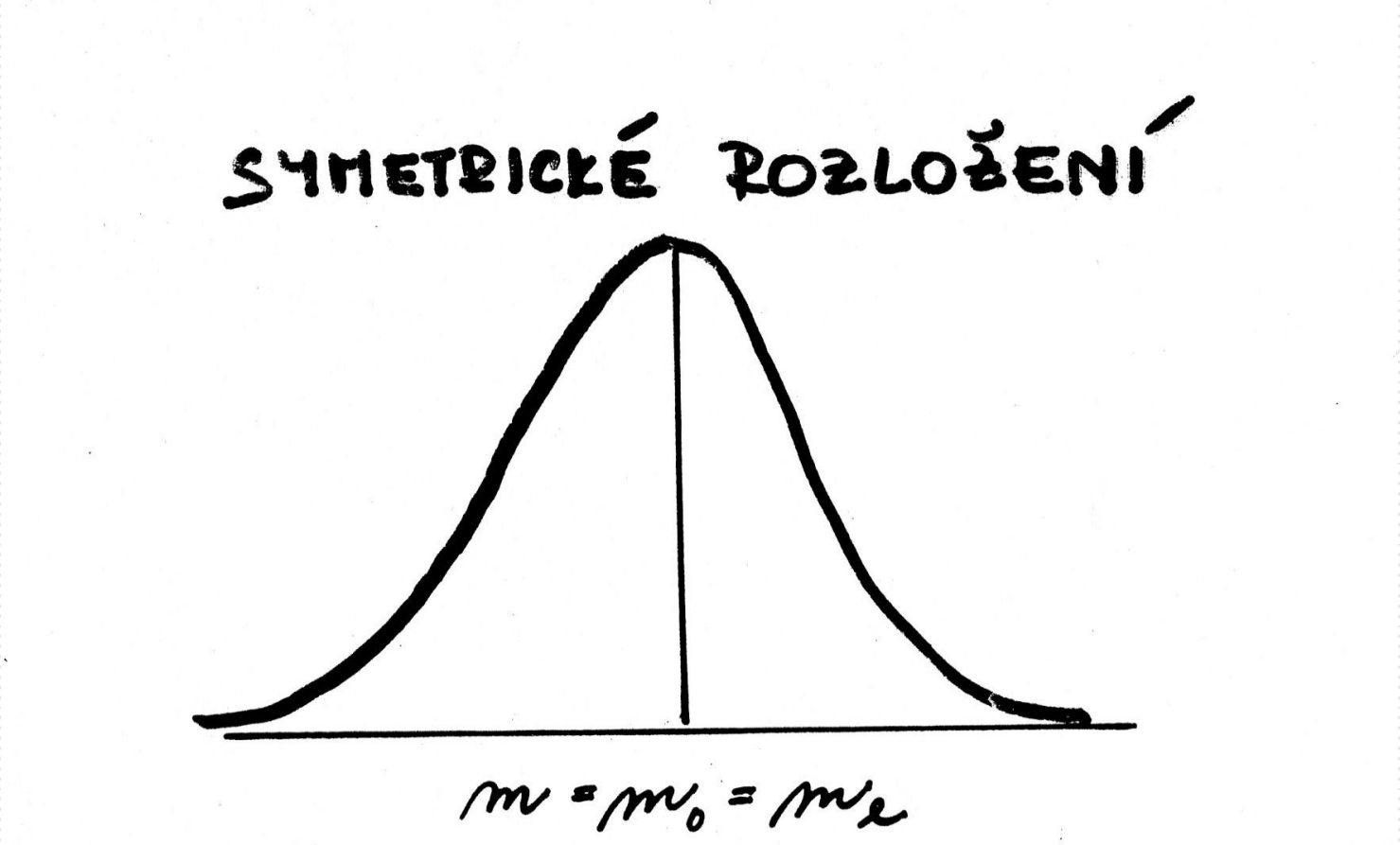
Ukazatele polohy (střední hodnoty)

- kde se hodnota nachází nad osou X

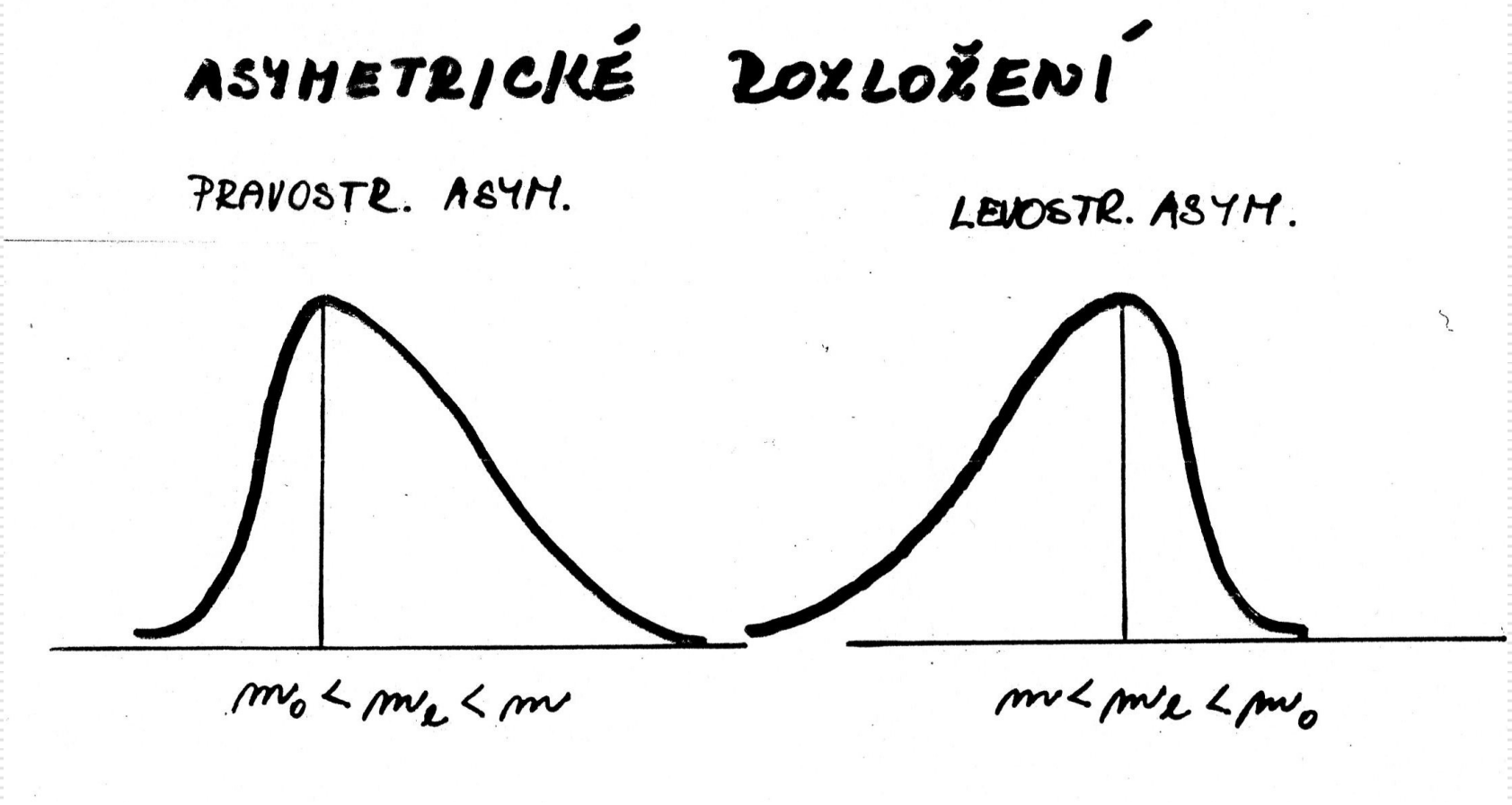
- **Aritmetický průměr (m)** – součet pozorovaných hodnot dělený počtem sledovaných jednotek
- **Medián (m_e)** – hodnota, která je právě uprostřed všech pozorování, která jsou seřazena podle velikosti (u sudého počtu – průměr ze dvou prostředních hodnot)
- **Modus (m_o)** – hodnota s nejvyšší četností (nejčastější)
- **Kvantil** (obecný název), **percentil** – pořadový ukazatel, medián je 50 percentil

aritmetický průměr nemá smysl počítat u asymetrických rozložení (náchylný k extrémním hodnotám)

Symetrické rozložení hodnot

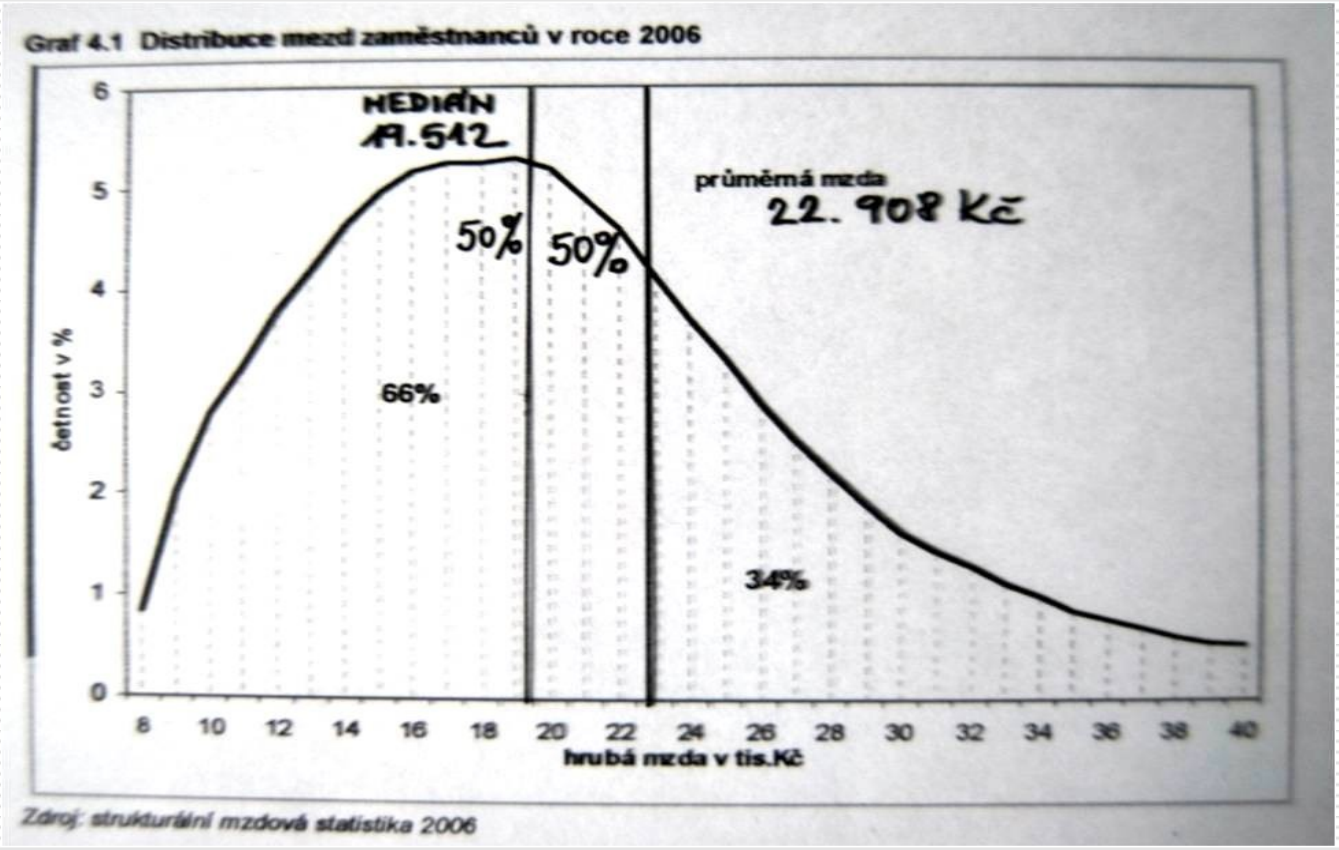


Asymetrické rozložení hodnot



Př.: distribuce mezd zaměstnanců 2006

Obr.



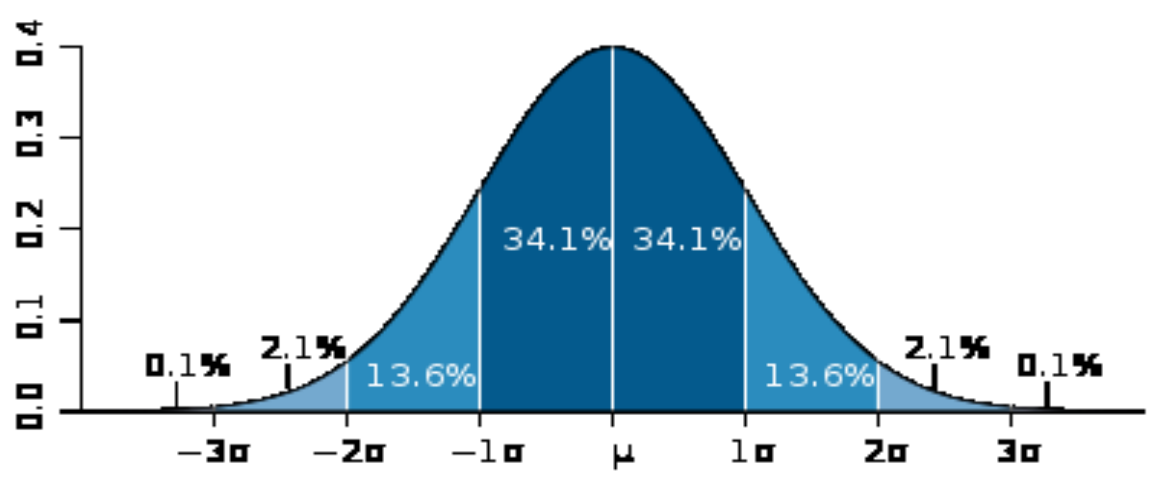
Zrádnost aritmetického průměru

Stará historka popisuje loutkové divadlo, které si dělalo marketing profilu svých diváků. Použili aritmetického průměru a dospěli k závěru, že jejich představení jsou nejatraktivnější pro generaci třicátníků.

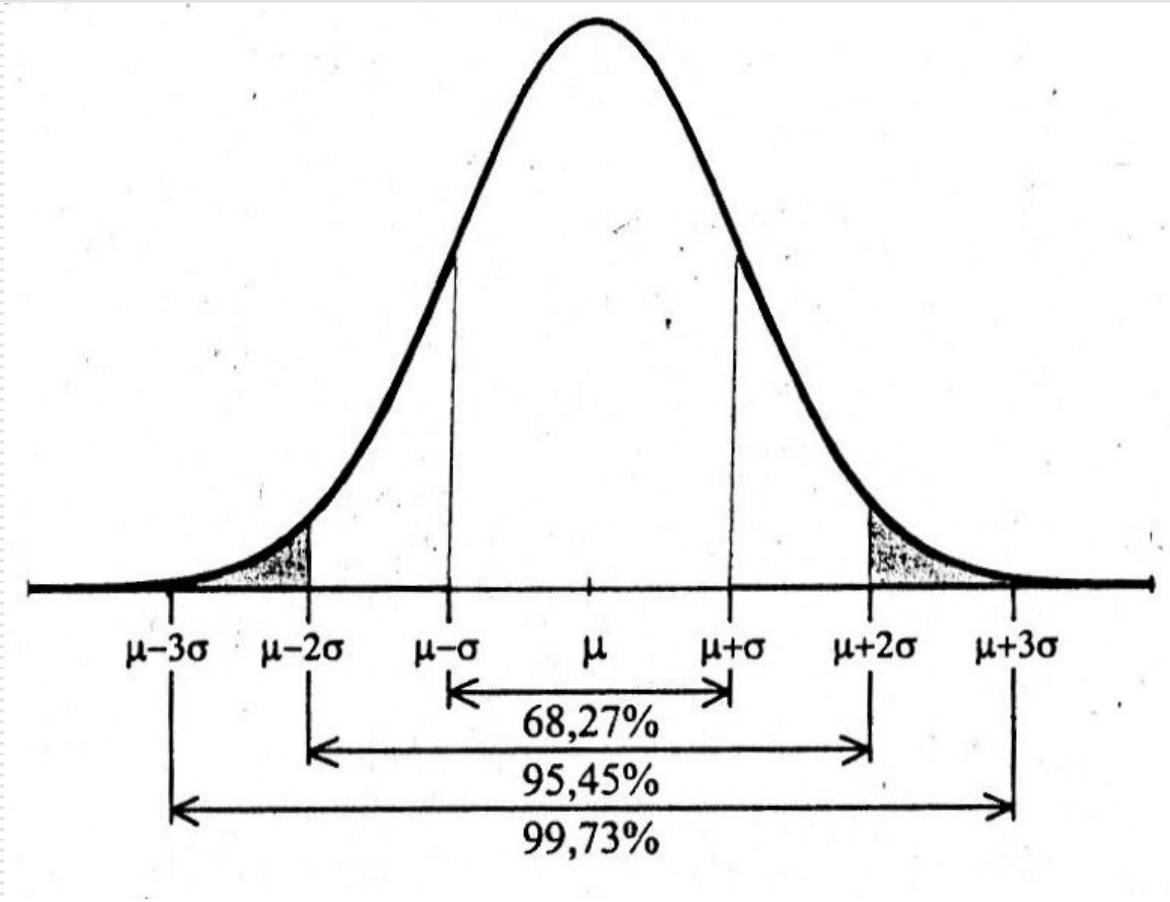
Ve skutečnosti navštěvovali divadlo dědečkové a babičky s vnoučaty.

Normální rozdělení – (GAUSSOVO)

Je nejdůležitější spojité rozdělení**. Náhodná veličina má normální rozdělení tehdy, je-li vytvářena nahromaděním velkého počtu nepatrných nezávislých příčin nahodilé povahy.** Např. tělesnou výšku lze považovat za norm. náhodnou veličinu – působí na ni řada podnětů nepříliš na sobě závislých.



Normální rozdělení – (GAUSSOVO)



Výstižný popis dat

Proč nestačí střední hodnoty k výstižnému popisu dat?

Př. počet onemocnění u pěti kojenců v 1. roce života

1. skupina: 3,4,5,6,7 m = 5
2. skupina: 0, 4, 5, 6,10 m = 5

   obě skupiny mají stejný aritmetický průměr, ale liší se kolísáním hodnot – **variabilitou**

Spolu se střední hodnotou by se měl uvádět **ukazatel variability**

Ukazatele variability - absolutní

Variační šíře (rozpětí): nejjednodušší míra variability

R = Xmax. – Xmin. Pro N menší a rovno 10

Rozptyl : průměr čtverců odchylek hodnot jednotlivých pozorování od aritmetického průměru

S² $$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

Směrodatná odchylka: odmocněný rozptyl

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

- ukazatel variability udávaný **ve stejných jednotkách jako sledovaný znak**
- vypovídá o tom, o kolik se většina hodnot sledovaného znaku odchyluje od průměru

Ukazatele variability -relativní

Variační koeficient - relativní ukazatel variability

$$v.k. = \frac{s}{m} \times 100(\%)$$

- udává, jaký podíl tvoří směrodatná odchylka z průměru
- veličina bez rozměru udávaná v %
- užití: srovnat variabilitu 2 a více souborů, jejichž průměry se značně liší nebo variabilitu znaků uváděných v různých jednotkách.

Využití variačního koeficientu

1) Je-li v. k. >50%, pak je soubor natolik nesourodý, že nemá smysl charakterizovat ho aritmetickým průměrem.
2) Slouží ke srovnání variability 2 souborů, jejichž průměry se značně liší.

   Př.: VC u mužů a žen

   M: **m**= 4,80 l **s** = 0,66 **v. k.** = 13,8%

   Ž: **m** = 3,90 l **s** = 0,42 **v. k.** = 10,8%
3) Slouží ke srovnání variability znaků uváděných v různých jednotkách.

   Př.: VC, tělesná výška, hmotnost u mužů

   VC: **m** = 4,80 l **s** = 0,66 **v. k.** = 13,8%

   Výška: **m** = 178 cm **s** = 4 **v. k.** =2,2%

   Hmotnost: **m** = 82 kg **s** = 6 **v. k.** = 7,3%

Využití variačního koeficientu

Např.1. skupina 18ti letých dívek – prům. výška 162 cm, směr. odch. je 5,2 cm

2. skupina 6ti letých dívek – prům. výška je 113 cm, směr. odch. je 4,6 cm

Využití variačního koeficientu

18. V. K. = 5,2/162 * 100 = **3,21 %**

6. V. K. = 4,6/113 * 100 = **4,07 %**

Výpočet percentilu

P_x = percentil

x = pořadí hledaného percentilu

a = dolní hranice intervalu, v němž je percentil obsažen

h = délka intervalu

n_x = absolutní četnost intervalu, v němž je percentil obsažen

n = celkový rozsah souboru

r_k = relativní kumulativní četnost předcházejícího intervalu (v %)

$$P_x = a + \frac{h \cdot n}{n_x \cdot 100} (x - rk)$$

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). Referenční hodnoty jsou z roku 1991.

Děkuji za pozornost