Statistika v biomedicínském výzkumu a ve zdravotnictví

©Prof. RNDr. Jana Zvárová, DrSc.
EuroMISE Centrum
Ústav informatiky AV ČR v.v.i.

1

Literatura

Edice Biomedicínská statistika vydávaná na Univerzitě Karlově v Praze, Karolinum

Zvárová, J.: *Základy statistiky pro biomedicínské obory I. Karolinum, Univerzita Karlova v Praze, Praha 2001*

http://www.euromise.cz/

2

Proč se používají statistické metody v biomedicíně

- Biomedicínské obory se rychle kvantifikují
- Plánování, provádění a interpretadce výsledků biomedicínského výzkumu se stále více stávají závislými na statistické metodologii
- Statistika proniká do literatury v biomedicínských oborech.

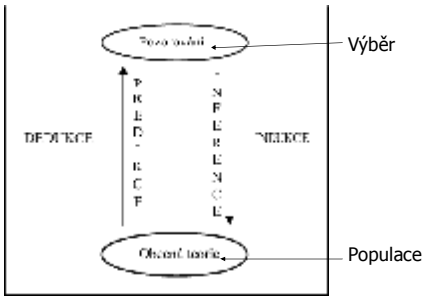
3

Statistika

- ***popisná statistika***
  - shromažďování, uspořádání a popis **souborů dat**
  - přehledná **sumarizace informací** (např. ČSÚ)
- ***induktivní statistika***
  - na základě vzorku (výběru) se snažíme odvodit obecná tvrzení o celku - závěry jsou zatíženy *statistickou chybou*
  - spolehlivost závěrů je stanovena **objektivně**

4

Obsah a význam statistiky - schéma



5

Populace

- ***populace (základní soubor)***

  přesně určena výčtem nebo vlastnostmi prvků
  - **určené výčtem** (např. *demografické populace*)

    - konečný rozsah
  - **vymezené vlastnostmi** (např. *všechny možné výsledky pokusu v daných experimentálních podmínkách*)

    - nekonečný rozsah

6

Výběr

výběr část **populace**, kterou sledujeme v rámci výzkumu

Druhy výběru:

- *reprezentativní výběr*
  výběr, který svojí strukturou odpovídá struktuře populace
- *selektivní výběr*
  výběr, jehož struktura neodpovídá struktuře populace - - zkreslení výsledků

rozsah výběru ..... počet prvků **populace** zahrnutých do **výběru**

7

Konstrukce výběru

- *záměrný výběr*
  - expertní hledisko, často subjektivní
- náhodný výběr
  - prostý
  - mechanický
  - oblastní (stratifikovaný)

G Data tvoří základ každého výzkumu, při němž jsou použity statistické metody. Jejich kvalita určuje kvalitu výsledků.

8

Znaky

Předmětem statistického výzkumu jsou *znaky*, tj. určité vlastnosti objektů sledovaných.

kvantitativní znak ..... sledovanou vlastnost je možné změřit a vyjádřit číslem

Př.: tělesná výška, počet sourozenců

kvalitativní znak ..... vlastnost je vyjádřena slovně

Př.: **rodinný stav, stupeň bolesti**

9

Kvalitativní znaky

- *nominální* …… rozdělení do tříd bez uspořádání

  Př.: rodinný stav:{svobodný, ženatý, vdovec, rozvedený, druh}
- *ordinální* …… rozdělení do tříd s uspořádáním

  Př.: stupeň bolesti: {žádná bolest, malá bolest, středně silná bolest, velmi silná bolest}

Kvalitativní znaky se pro účely analýzy mohou kódovat.

10

Kvalitativní pozorování

Jméno pacienta	Rodinný stav	Rod. stav - kód	Riziko povolání	Kód rizika
Novák	svobodný	0	střední	1
Kubíček	ženatý	1	velké	2
Bláhová	rozvedená	2	malé	0
Roubal	rozvedený	2	střední	1
Kratochvílová	svobodná	0	malé	0
Zemanová	ovdovělá	3	malé	0
Novotná	vdaná	1	střední	1
Žitný	svobodný	0	velké	2

11

Absolutní a relativní četnosti

Absolutní četnost ..... počet pozorování v dané skupině

Př.: počet svobodných osob ve výběru (v našem příkladu 3)

Relativní četnost $\text{relativníčetnost} = \frac{m}{n}$

n ..... počet pozorování

m ..... absolutní četnost

Př.: relativní četnost svobodných ve výběru je 3/8 = 0,375

12

Kvantitativní znaky

- *spojité*

  Př.: výška, hmotnost, koncentrace roztoku
- *diskrétní*

  Př.: počet úmrtí, počet sourozenců

13

Kvantitativní znaky - míry polohy

- hodnoty charakterizující "střední" hodnotu znaku

Aritmetický průměr $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Př.: Vypočtěte průměr následujících výsledků vyšetření: 39, 42, 73, 67, 24, 55.

Řešení:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{6} x_i = \frac{1}{6}\cdot 300 = 50$$

14

Kvantitativní znaky - míry polohy

Modus (výběrový) ..... hodnota, která se v souboru dat vyskytuje nejčastěji

Př.: Co je modus v následujících výsledcích zjišťování krevních skupin: A, 0, 0, B, B, AB, A, A, 0, 0, 0, AB, B, 0, B, A, 0, AB, 0, 0, B, 0, A?

Řešení: Výsledky měření shrneme do tabulky:

krevní skupina	četnost výskytu
A	5
B	5
AB	3
0	10

Modem je tedy krevní skupina 0.

15

Kvantitativní znaky - míry polohy

Medián (výběrový) ..... v seřazeném souboru je to taková hodnota, která soubor rozděluje na dvě stejně velké části

Př.: Co je mediánem v následujících měřeních výšek dětí: 115, 117, 119, 122, 124, 128, 150?

Řešení: 122

16

Kvantitativní znaky - míry rozptýlenosti

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

Rozptyl

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

Směrodatná odchylka $s = \sqrt{s^2}$

17

Kvantily

100*P*% kvantil x_P je číslo, které odděluje 100*P*% nejmenších hodnot znaku.

Známé kvantily:

- Percentily
- Decily
- Kvartily
- Medián

18

Normální rozdělení

19

Prostá tabulka

Tělesná výška v cm								
130	140	136	141	139	133	149	151	139
136	138	142	127	147	139	135	141	143
132	146	151	146	141	141	131	142	141

Skupinová tabulka

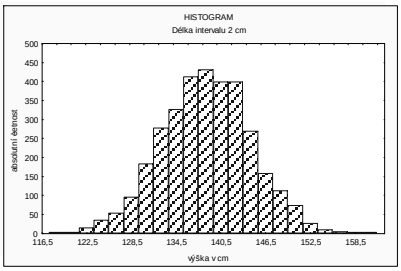
Střed třídního intervalu	Počet dětí
125	1
130	3
135	4
140	12
145	4
150	3
Celkem	27

20

Rozdělení chlapců ve věku 9,5 - 10 let podle tělesné výšky
(délka třídního intervalu 5 cm)

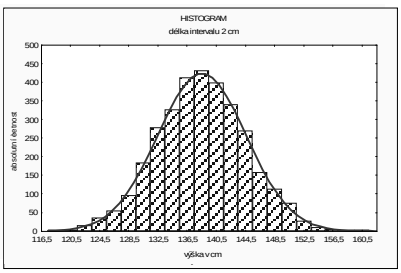
Číslo třídy i	Střed třídy x_i	Absolutní četnost n_i	Součin $n_i x_i$	Relativní četnost n_i/n	Relativní kumulativní četnost
1	120	13	1 560	0,0040	0,0040
2	125	95	11 875	0,0294	0,0334
3	130	414	53 820	0,1281	0,1615
4	135	880	118 800	0,2724	0,4339
5	140	1013	141 820	0,3135	0,7474
6	145	582	84 390	0,1801	0,9275
7	150	199	29 850	0,0616	0,9891
8	155	29	4 495	0,0090	0,9981
9	160	6	960	0,0019	1,0000
celkem	-	3231	447 570	1,0000	-

Histogram - výška chlapců



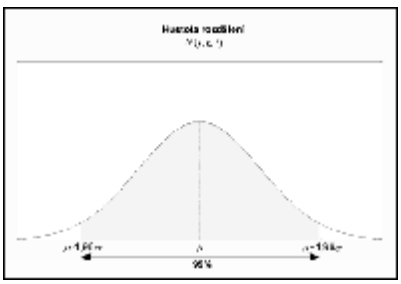
22

Histogram - výška chlapců



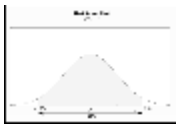
23

Normální rozdělení



24

Normální rozdělení



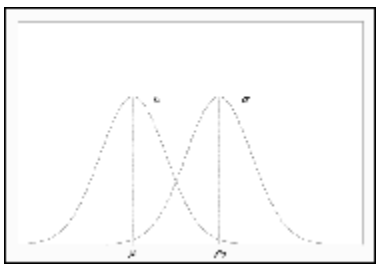
Normální rozdělení se týká populace, nikoli výběru.

V případě normálního rozdělení průměr, modus a medián splývají.

Normální rozdělení je plně určenou dvěma parametry:
- populačním průměrem μ a
- populační směrodatnou odchylkou σ.

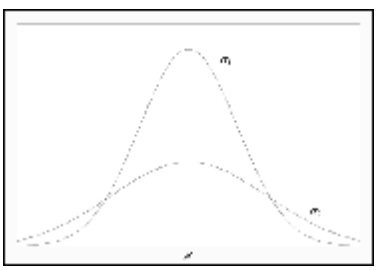
25

Vliv populačního průměru na tvar normálního rozdělení



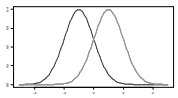
26

Vliv populační směrodatné odchylky na tvar normálního rozdělení

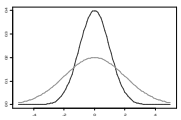


27

Význam parametrů



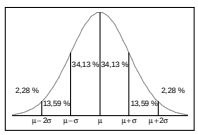
- μ střední hodnota (poloha, populační průměr, těžiště), **modus i medián**
- $\mu_1 < \mu_2, \quad \sigma_1 = \sigma_2$



- σ směrodatná odchylka (míra variability, koncentrace)
- σ^2 rozptyl (populační rozptyl)
- $\mu_1 = \mu_2, \quad \sigma_1 < \sigma_2$

28

Obecné normální rozdělení



$$X \sim N(m, s^2) \Rightarrow$$

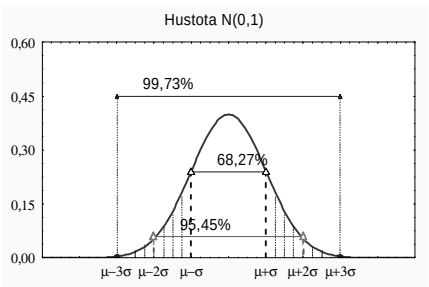
- (standardizace)

$$Z = \frac{X - m}{s} \sim N(0,1)$$

$$P(X \le x) = P\left(\frac{X - m}{s} \le \frac{x - m}{s}\right) = \Phi\left(\frac{x - m}{s}\right)$$

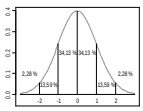
29

Hustota N (μ, σ^2)



30

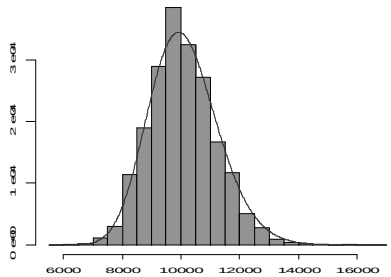
Normální rozdělení



- součet spousty nepatrných nezávislých příspěvků
- nepřesnost měření
- délkové rozměry částí lidského těla
- objemy, koncentrace … zpravidla až po transformaci
- relativní četnost při velkém počtu pokusů

31

Hustota logaritmicko-normálního rozdělení



32

Odhad populačního průměru

Statistická teorie:

"Nejlepším odhadem **populačního** průměru μ je **výběrový** průměr."

? Nakolik odhad vystihuje skutečnou hodnotu μ?

Vypovídací hodnotu odhadu posuzujeme podle délky *intervalu spolehlivosti*.

Interval spolehlivosti ... interval, v němž s 95% pravděpodobností leží neznámý populační průměr μ.

Příklad: Průměrná výška dětí:
- pro 5 pozorování můžeme tvrdit, že průměrná výška leží s 95% pravděpodobnostní v intervalu (128; 148)
- pro 1285 pozorování získáme interval (138,1; 138,9)

33

Rozdělení výběrového průměru

- **náhodný výběr**: za stálých podmínek nezávisle provedená opakovaná měření stejné vlastnosti
- výšky náhodně vybraných desetiletých hochů
- $x_1, x_2, \ldots, x_n$ mají stejné rozdělení:
- μ - populační průměr, σ^2 - populační rozptyl
- pro velké n potom přibližně

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \sim N\left(m, \frac{s^2}{n}\right) \qquad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

34

Interval spolehlivosti pro μ

- pro velké n má průměr $\bar{X}$ přibližně normální rozdělení $N\left(m, s^2/n\right)$

$$1 - a = P\left(\left|\frac{\bar{x} - m}{s/\sqrt{n}}\right| < z_{1-a/2}\right)$$

$$= P\left(\bar{x} - \frac{s}{\sqrt{n}} z_{1-a/2} < m < \bar{x} + \frac{s}{\sqrt{n}} z_{1-a/2}\right)$$

- při opakovaném pořizování výběrů obsahuje asi
- 100(1-α) % intervalů populační průměr μ

35

Bodový a intervalový odhad pro μ u rozdělení $N(\mu,\sigma^2)$

•Bodový odhad ... $\bar{X}$ $\qquad \bar{X} \sim N\left(m, \frac{s^2}{n}\right)$

•95% interval spolehlivosti ... $\left\langle \bar{X} - 1{,}96\frac{s}{\sqrt{n}}; \bar{X} + 1{,}96\frac{s}{\sqrt{n}} \right\rangle$

délka intervalu: přesnost odhadu

36

Rozsah výběru

Příklad: Chceme konstruovat 95 % interval spolehlivosti pro průměrnou hodnou cholesterolu s délkou +- 0,2 mmol/l, rozptyl hladiny cholesterolu je $1{,}25^2$.

Řešení:

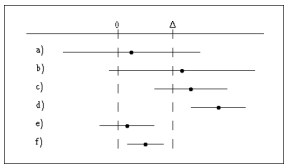
$$\left\langle \overline{X} - 1{,}96\,\frac{1{,}25}{\sqrt{n}};\ \overline{X} + 1{,}96\,\frac{1{,}25}{\sqrt{n}} \right\rangle$$

$$1{,}96\,\frac{1{,}25}{\sqrt{n}} = 0{,}2$$

$$n = (1{,}96 * 1{,}25 / 0{,}2)^2 = 150$$

37

Statistická a klinická významnost



Možnost	Statistická významnost	Klinická významnost
a	ne	možná
b	ne	možná
c	ano	možná
d	ano	ano
e	ne	ne
f	ano	ne

38