Lineární korelace a regrese

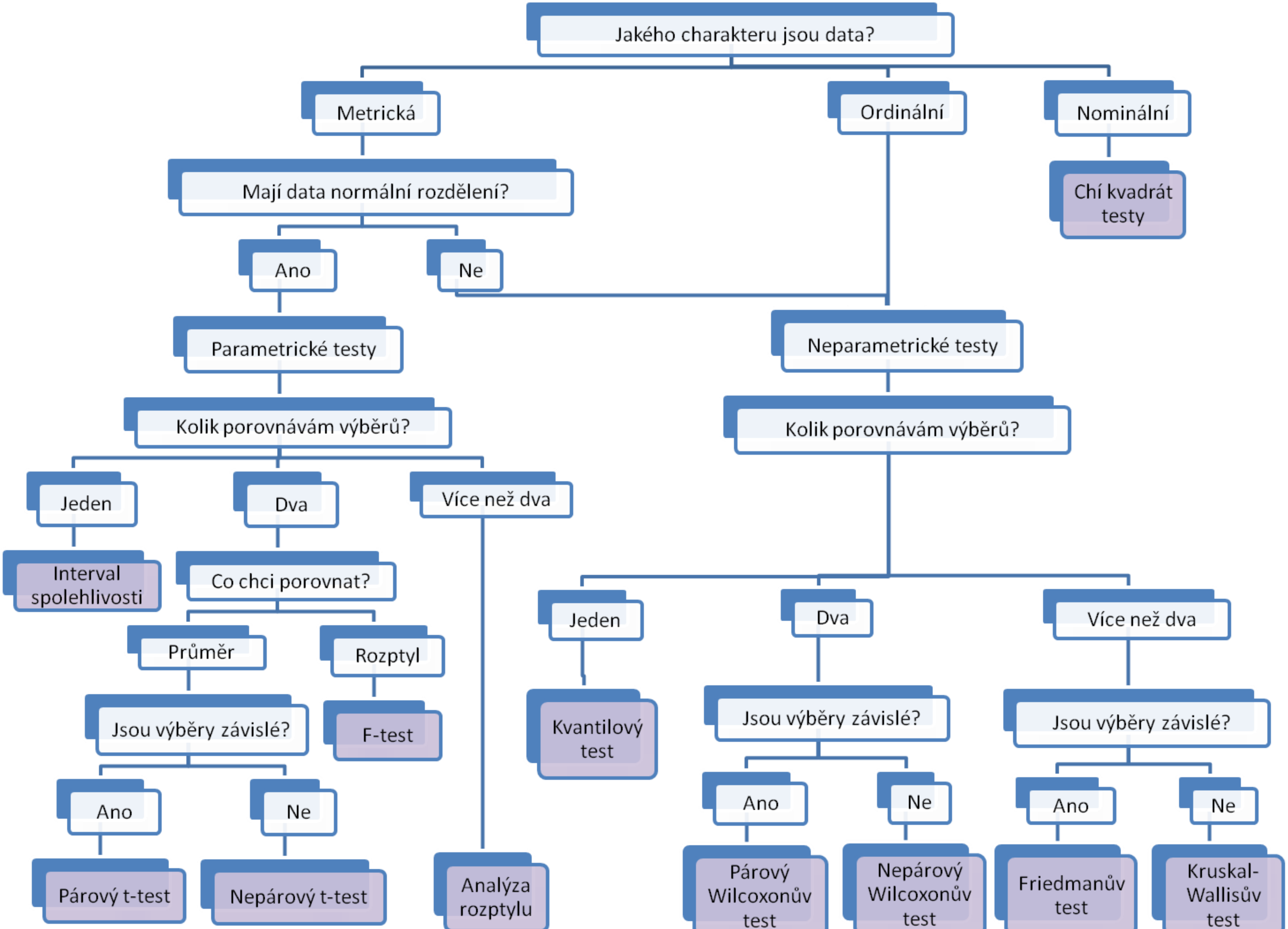
B02907 Informační a komunikační technologie

Lubomír Štěpánek,
Ústav biofyziky a informatiky
1. LF UK

Upozornění!

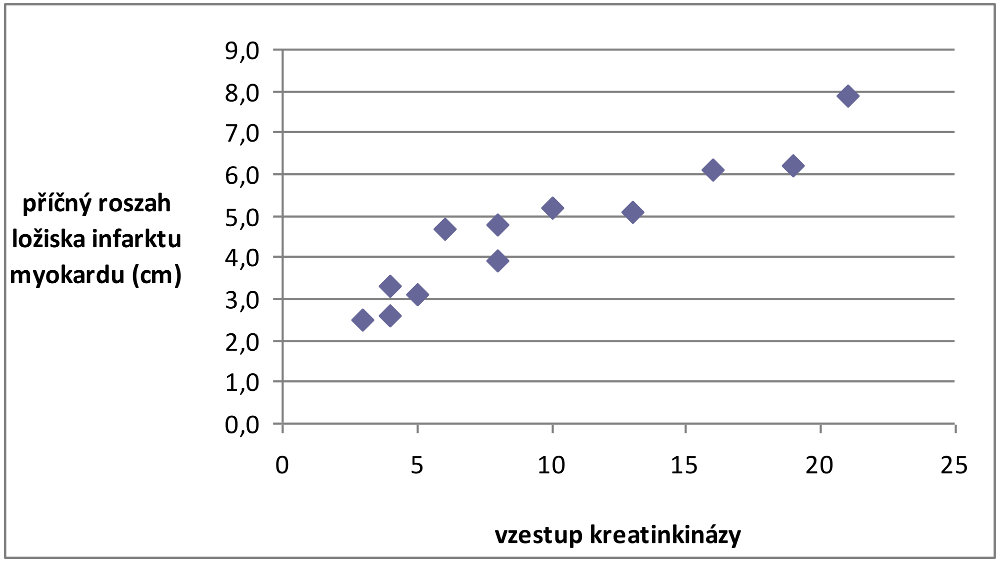
- dole v poznámkách jsou u většiny snímků rozšiřující a vysvětlující komentáře
- u některých statistických metod budete odkazováni na statistické tabulky, které jsou volně přístupné online na adrese http://new.euromise.org/czech/tajne/ucebnice/html/html/node15.html
- (obvykle bude ještě na příslušném snímku odkaz zopakován; autor vynaložil značné úsilí, aby se symbolika v prezentacích shodovala se symbolikou v tabulkách, proto by neměla být orientace v tabulkách problémem)
- z předložených prezentací se můžete učit, můžete je kopírovat či jinak měnit, ale bez dovolení autora/autorů je nesmíte použít do svých publikací ☺
- předložené prezentace nejsou bezchybnou statistickou kuchařkou, proto ne zcela doporučuji se na ně ve svých pracích odkazovat, nebo je dokonce citovat ☺
- pokud se budu sám odkazovat na vhodnou literaturu, myslím tím nejspíše následující dvě knihy:
  - Zvára: Biostatistika. Karolinum, Praha 1988
  - Zvárová et al.: Biomedicínská statistika I. Základy statistiky pro biomedicínské obory
- dotazy a konzultace možné a vlastně i doporučeny

(Lubomír Štěpánek, stepanek.lub@seznam.cz)


Jakého charakteru jsou data?
Metrická
Ordinální
Nominální
Mají data normální rozdělení?
Chí kvadrát testy
Ano
Ne
Parametrické testy
Neparametrické testy
Kolik porovnávám výběrů?
Kolik porovnávám výběrů?
Jeden
Dva
Více než dva
Interval spolehlivosti
Co chci porovnat?
Průměr
Rozptyl
Jsou výběry závislé?
F-test
Ano
Ne
Párový t-test
Nepárový t-test
Analýza rozptylu
Jeden
Dva
Více než dva
Kvantilový test
Jsou výběry závislé?
Jsou výběry závislé?
Ano
Ne
Ano
Ne
Párový Wilcoxonův test
Nepárový Wilcoxonův test
Friedmanův test
Kruskal-Wallisův test

Korelace

- lineární závislost mezi hodnotami prvního a druhého znaku zkoumaných na každém prvku výběru, je-li se hodnota prvního znaku, je větší hodnota i druhého znaku
- *např. o každém pacientovi zjistíme jeho hmotnost a tělesnou výšku, nebo např. plazmatickou hladinu PSA a klinické stádium karcinomu prostaty; obě hodnoty v obou příkladech rostou a klesají spolu atd.*
- znaky mají metrické či ordinální hodnoty (tělesná výška/cm, hmotnost/kg; klinické stádium/stupeň atd.)
- po vynesení hodnoty prvního znaku na vodorovnou a hodnoty druhého znaku na svislou osu pro každý prvek můžeme zjistit náznak závislosti

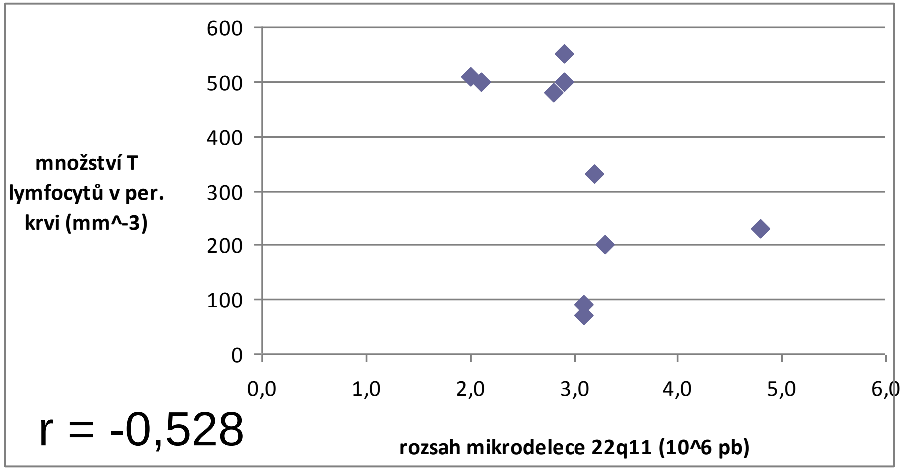


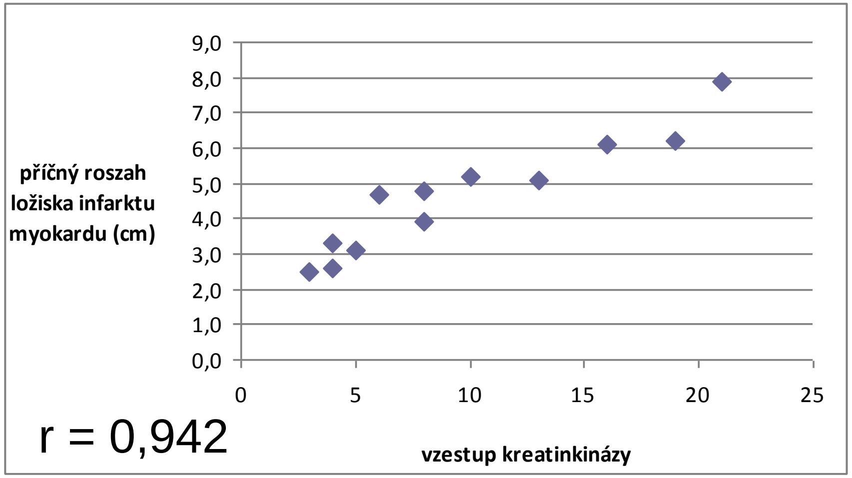
Korelační koeficient

- míru korelace mezi dvěma znaky (veličinami) stejných prvků popisuje korelační koeficient
- jsou-li oba znaky vyjádřeny daty metrickými (ideálně spojitými), používáme *Pearsonův korelační koeficient*
- jsou-li oba znaky vyjádřeny daty ordinálními, používáme *Spearmanův korelační koeficient*
- je-li jeden znak vyjádřen daty metrickými, druhý ordinálními, obvykle data metrická uspořádáme dle velikosti a opět užijeme *Spearmanův korelační koeficient*

Pearsonův korelační koeficient

- číselně udává míru korelace mezi oběma znaky, značen r, hodnoty $-1 \leq r \leq 1$, 0 znamená nezávislost nebo nelineární závislost
- nutný předpoklad – oba znaky vyjádřeny spojitými metrickými daty (nespojitost dat lze tolerovat)
- možné příklady např. *výška (cm) a váha (kg) pacienta, počet dětí v rodině (1) a příjmy rodiny* atd.
- matematický vzorec viz vhodná literatura
- MS Excel:
  - vložíme funkci PEARSON, ta vyžaduje pouze hodnoty obou znaků (Pole 1 a 2)
  - kladně vzatá ($|r|$) výsledná hodnota značí těsnost korelace, není tabelizovaná; čím větší $|r|$, tím větší korelace mezi oběma znaky





Intermezzo

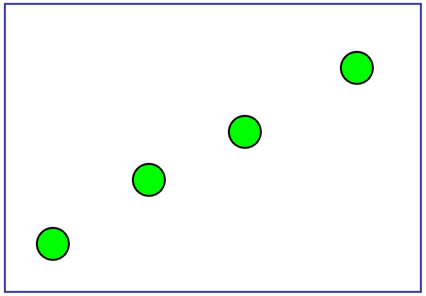
Přiřaďte typy závislostí a korelačních koeficientů k pravděpodobným grafickým reprezentacím daných závislostí.

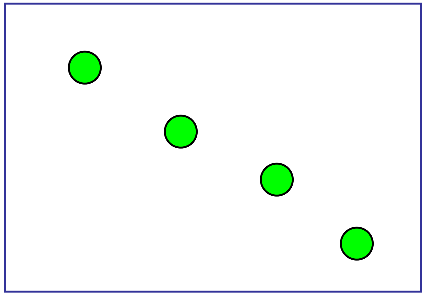
Nezávislost r = 0

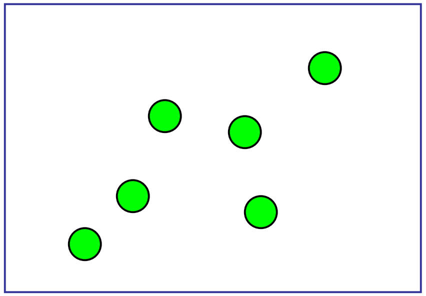
Přímá úměra r = +1

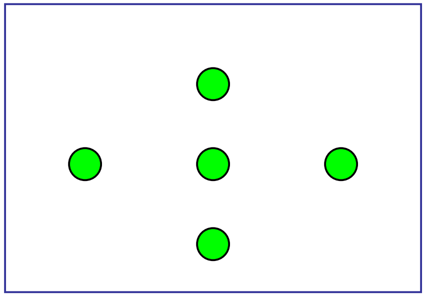
Částečná závislost r = 0,3

Nepřímá úměra r = -1



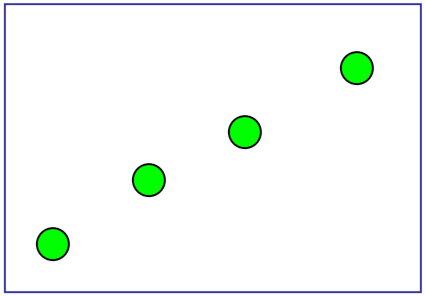




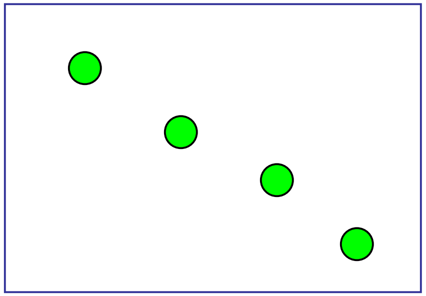


Korelace – míra závislosti dvou porovnávaných proměnných

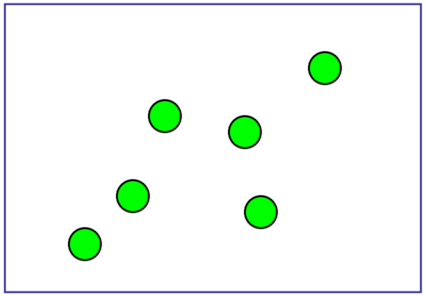
Pearsonův korelační koeficient (r) – kritérium závislosti,
nabývá hodnot od -1 do +1



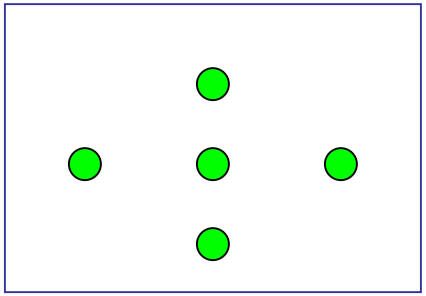
Přímá úměra r = +1



Nepřímá úměra r = -1



Příklad částečné závislosti r = 0,3



Nezávislost r = 0

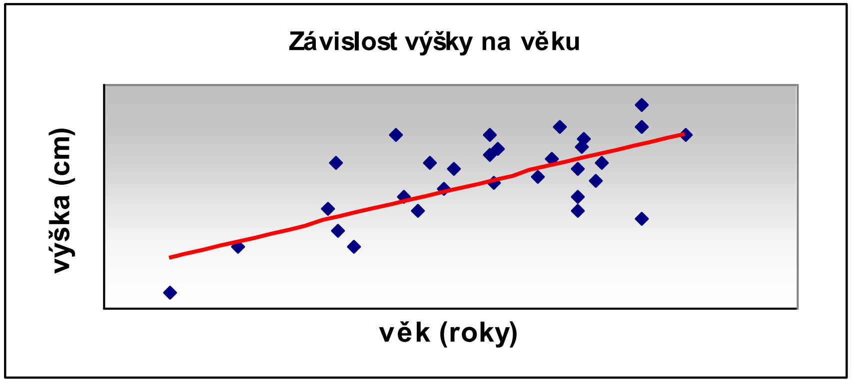
Spearmanův korelační koeficient

- číselně udává míru korelace mezi oběma znaky, značen r_S, hodnoty $-1 \leq r_S \leq 1$
- nutný předpoklad – oba znaky vyjádřeny ordinálními daty (pokud má jeden znak data metrická, uspořádáme je, získáme tak vlastně data ordinální)
- např. *stupeň vážnosti nádoru (I, II, III) na rozsahu nádoru (1, 2, 3, 4), klinický průběh zánětu (subklinický, mírný, střední, závažný) na sedimentaci (jednotek/hodina)* atd.
- MS Excel nemá vhodnou funkci
- postup:
  - nejdříve každému z n prvků přiřadíme pořadí, které by získal při vzestupném řazení hodnot první znaku, poté podobně s druhým znakem (každý prvek má tak dvě pořadí – pro první a druhý znak)
  - nyní pořadí prvku pro první znak a pro druhý znak odečtěme (získáme tím d_i)
  - Spearmanův korelační koeficient je

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

  - je-li následně $|r_S| \geq$ kritická hodnota(n, alfa), jsou oba znaky na sobě závislé, korelují
  - (kritické hodnoty v tabulkách)

Regrese



Nezávisle proměnná –
výchozí, ovlivňující znak (kupř. věk dítěte), vodorovná osa

Závisle proměnná –
odvozený, ovlivňovaný znak (kupř. výška dítěte), svislá osa

Regresní rovnice – popisuje typ závislosti
Regresní odhad – dosazením do rovnice lze odhadnout hodnotu závisle proměnné

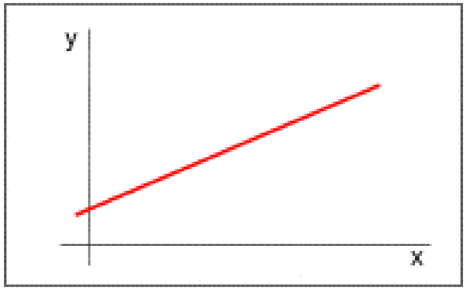
Lineární regrese

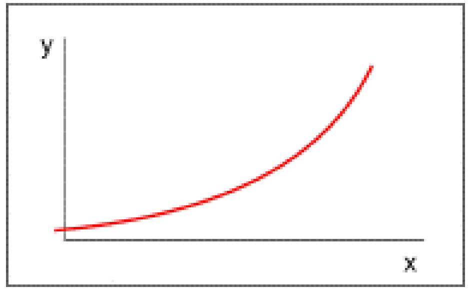
- lineární matematický popis závislosti dvou znaků, které spolu dostatečně korelují (podle Pearsonova koeficientu)
- znaky musí být popsány metrickými hodnotami (ideálně spojitými)
- výsledkem lineární regrese je graficky přímka, jejíž rovnice určuje vztah mezi hodnotami obou znaků

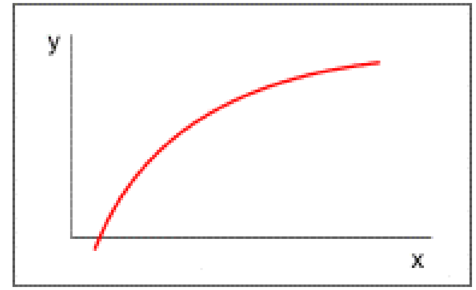
$$y = a \cdot x + b$$

- *y* je hodnota druhého znaku, *x* je hodnota prvního znaku, *a* je *směrnice*, *b* je *absolutní člen*
- známe-li hodnoty *x* a *y* obou korelujících znaků každého prvku výběru, můžeme sestavit regresní rovnici
- známe-li regresní rovnici, můžeme z hodnoty *y*, nebo x, dopočítat zbylou hodnotu (*x*, nebo *y*) pro každý prvek

Regrese – matematický vztah mezi dvěma proměnnými







$y = a + b * x$
(př. viz níže)

$y = a * e^{bx}$
(př. zátěž - TF)

$y = a * \ln (b * x)$
(př. Saturace Hb – O2)

výška = 80 cm+ 5 * věk

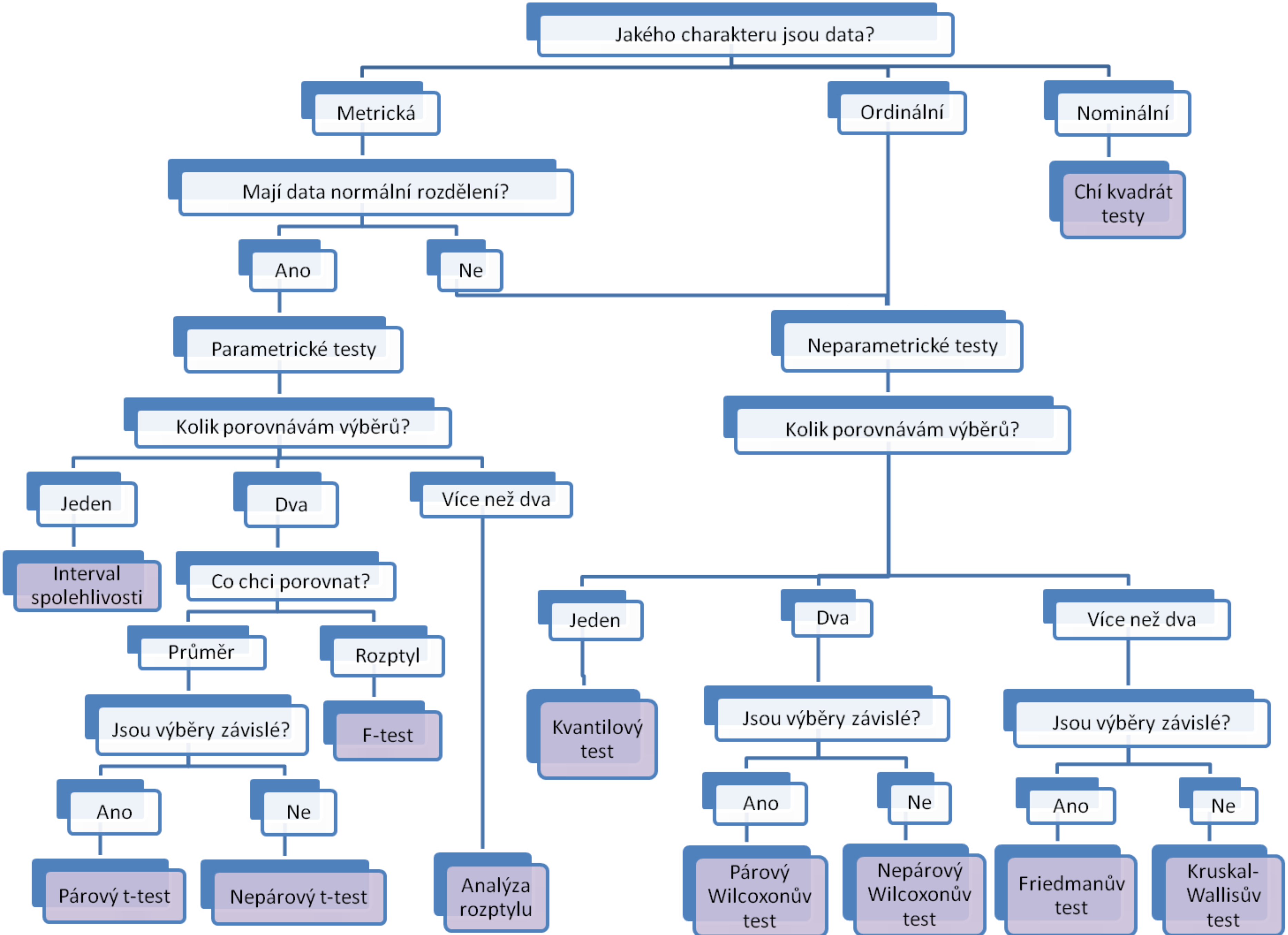
váha = 8 kg+ 2 * věk

Příklady rovnic užívaných v pediatrické praxi pro odhad přiměřené výšky a váhy v předškolním věku, kdy je závislost zhruba lineární

Výpočet regresní rovnice

- nutné určit směrnici *a* a absolutní člen *b*
- výpočet matematicky je relativně obtížný, pomůže nám MS Excel
  - vložíme funkci SLOPE, která požaduje Pole_y (závislých) a Pole_x (nezávislých) hodnot obou znaků, vrátí směrnici *a*
  - vložíme funkci INTERCEPT, která požaduje Pole_y (závislých) a Pole_x (nezávislých) hodnot obou znaků, vrátí absolutní člen *b*
- protože pracujeme s hodnotami prvků výběru, nikoliv s populací, je směrnice ovlivněna náhodností výběru, užíváme pro ni interval spolehlivosti, ve kterém leží s určitou pravděpodobností; jeho vyjádření viz vhodná literatura, např.

Zvárová et al.:Biomedicínská statistika I. Základy statistiky pro biomedicínské obory, Karolinum, Praha, 1988


Jakého charakteru jsou data?
Metrická
Ordinální
Nominální
Mají data normální rozdělení?
Chí kvadrát testy
Ano
Ne
Parametrické testy
Neparametrické testy
Kolik porovnávám výběrů?
Kolik porovnávám výběrů?
Jeden
Dva
Více než dva
Interval spolehlivosti
Co chci porovnat?
Jeden
Dva
Více než dva
Průměr
Rozptyl
Kvantilový test
Jsou výběry závislé?
F-test
Jsou výběry závislé?
Jsou výběry závislé?
Ano
Ne
Ano
Ne
Ano
Ne
Párový t-test
Nepárový t-test
Analýza rozptylu
Párový Wilcoxonův test
Nepárový Wilcoxonův test
Friedmanův test
Kruskal-Wallisův test

lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz