χ^2 testy

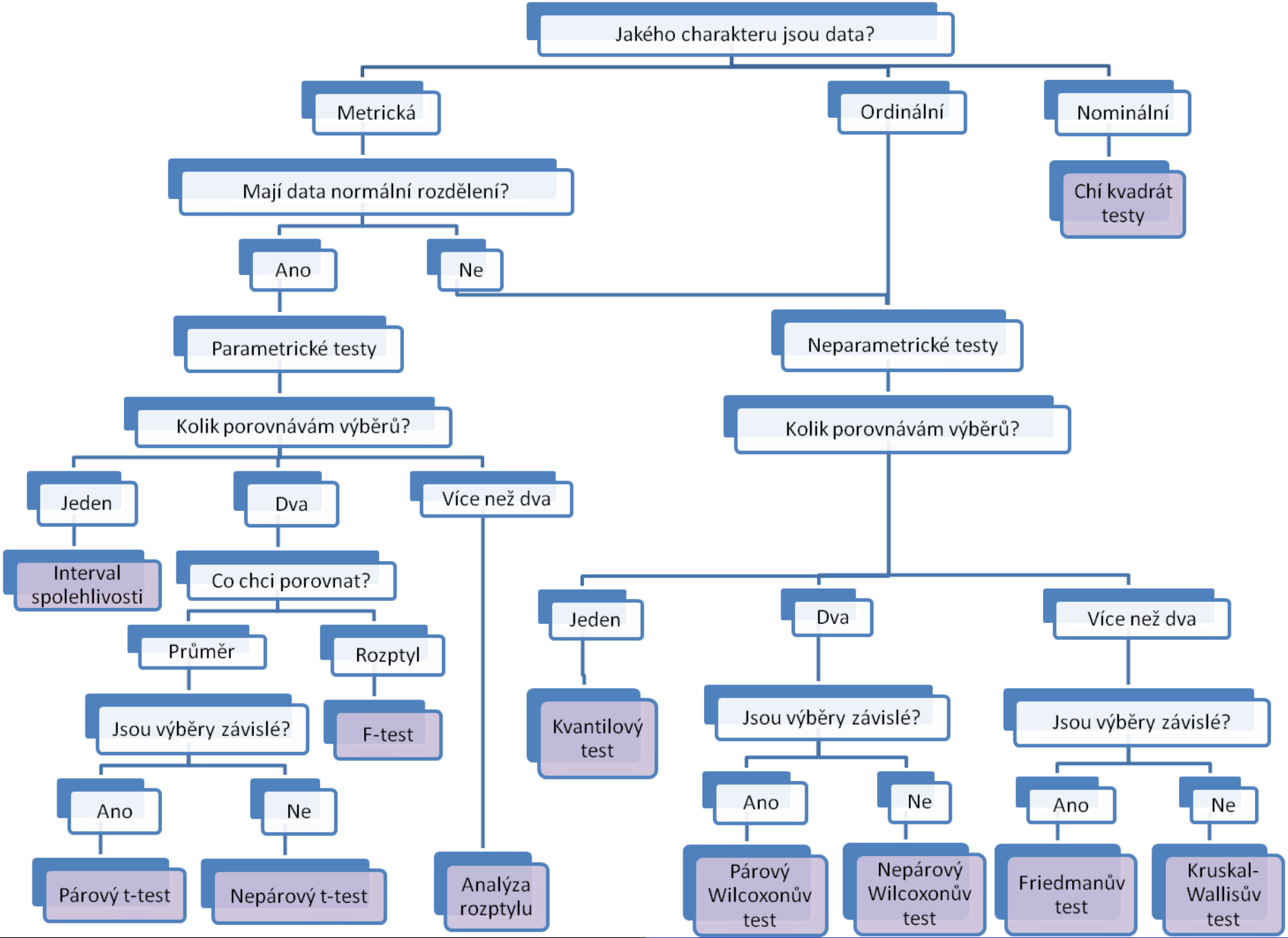
B02907 Informační a komunikační technologie

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?
Jeden
Dva
Více než dva
Průměr
Rozptyl
Kvantilový test
Jsou výběry závislé?
Jsou výběry závislé?
Jsou výběry závislé?
F-test
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

χ^2 testy

(především pro nominální data, dále pro ordinální, diskrétní metrická a spojitá metrická kategorizovaná data)

Chí kvadrát test dobré shody

- pro jeden výběr s nominálními, příp. ordinálními, diskrétními nebo kategorizovanými spojitými metrickými daty
- H0: zkoumaná veličina má stejné pravděpodobnostní rozdělení jako předpokládaný model, H1: zkoumaná veličina má jiné rozdělení než předpokládaný model
- lze užít k ověření normálnosti dat, potvrzení Poissonova rozdělení, binomického, ale i zcela jiného nespojitého rozdělení
- např. H0: „tělesná výška člověka má na vybraném souboru normální rozdělení“
- zkoumané rozdělení je nespojité, má k hodnot četností, předpokládaná četnost i-té hodnoty je π_i, pozorovaná četnost i-té hodnoty je n_i, testová statistika χ^2 je:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(\text{pozorovaná četnost - očekávaná četnost}\right)^2}{\text{očekávaná četnost}} = \sum_{i=1}^{k} \frac{\left(n_i - n\pi_i\right)^2}{n\pi_i}$$

- pokud je $\chi^2 \geq \chi_{1-\alpha}(df)^2$, zamítáme nulovou hypotézu H0; $df = k-1$ je počet stupňů volnosti
- kritické hodnoty viz tabulky

Chí kvadrát test 2x2

- též kontingenční tabulka 2x2
- pro dva výběry s nominálními, příp. ordinálními, diskrétními nebo kategorizovanými spojitými metrickými daty
- obecně H0: oba výběry mají stejné nebo podobné pravděpodobnostní rozdělení (jsou tedy ze stejné populace, je mezi nimi vztah), H1: výběry mají různá rozdělení
- rozeznáváme tři typy běžně testovaných hypotéz:
  - hypotéza o shodnosti struktur
  - hypotéza o symetrii

Chí kvadrát test 2x2

- první výběr má a prvků první hodnoty znaku, b prvků druhé hodnoty, celkem $n_1 = a+b$, druhý výběr má c prvků první hodnoty, d prvků druhé hodnoty, celkem $n_2 = c+d$, $n = n_1+n_2$
- pro čtyřpolní tabulku je $df = 1$ a tedy $\chi_{0,95}(1)^2=3{,}84$, lze dokázat, že χ^2 statistika je rovna:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{pozorovaná četnost - očekávaná četnost})^2}{\text{očekávaná četnost}} = \frac{(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}n$$

	1. hodnota	2. hodnota	
první výběr	a	b	$n_1=a+b$
druhý výběr	c	d	$n_2=c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Chí kvadrát test 2x2

- očekávaná četnost = součet v řádku X součet ve sloupci / celkový součet

	1. hodnota	2. hodnota	
první výběr	a	b	$n_1=a+b$
druhý výběr	c	d	$n_2=c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Chí kvadrát test 2x2

- hypotéza o shodnosti struktur
  - H0: např. „pacienti očkovaní a pacienti neočkovaní trpí na danou infekci se stejnou četností“, H1: „ pacienti očkovaní a pacienti neočkovaní trpí na danou infekci s různou četností“

	infekce ano	infekce ne	
očkovaní	a	b	$n_1=a+b$
neočkovaní	c	d	$n_2=c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

- hypotéza o symetrii (McNemarův test)
  - H0: např. „úspěšnost obou léků je stejná“, H1: „úspěšnost obou léků je různá“
  - pokud použijeme soubor stejných pacientů, kterým dáme nejdřív lék A a pak B, jedná se o závislé soubory

	efekt ano	efekt ne	
lék A	a	b	$n_1=a+b$
lék B	c	d	$n_2=c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Poměr šancí, relativní riziko

- poměr šancí (odds ratio, OR)
  - užívá se u studií případů a kontrol (case-control study)
  - poměr šancí následku v exponované a neexponované skupině
- relativní riziko (relative risk, RR)
  - užívá se u kohortových a intervenčních studií
  - poměr incidencí následku v exponované a neexponované skupině

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a(c+d)}{c(a+b)}$$

	efekt ano	efekt ne	
expozice ano	a	b	$a+b$
expozice ne	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

Poměr šancí, relativní riziko

- je-li OR nebo RR:
- =1 => expozice nemá na výskyt efektu žádný vliv (pouze náhodný)
- >1 => expozice zvyšuje pravděpodobnost efektu; např. expozice tabákovému kouři zvyšuje riziko/incidenci karcinomu plic
  - typické pro rizikové expozice
- <1 (ale >0) => expozice snižuje pravděpodobnost efektu, např. expozice slunečnímu záření snižuje riziko/incidenci křivice
  - typické pro protektivní expozice (faktory)

Hodnocen 1 faktor v různých situacích nebo 2 faktory (závislost)

Nezávislé – různé výběry
Závislé (př. tíž pacienti v různých situacích)

Srovnání s populací, dva nebo více výběrů

Proměnné / Výběry / Data	1 faktor					2 faktory	
	NEZAVISLE			ZAVISLE			
	1 výběr	2 výběry	k výběrů	2 výběry	k výběrů		
Metrická	Interval spolehlivosti, u-test	t-test	ANOVA při jednoduchém třídění	Párový t-test	Analýza rozptylu s opakování	Pearsonův korelační koeficient	Poloha
	Interval spolehlivosti	F-test	Bartlett	Fergusonův			Variabilita
Ordinální	Kvantilový test	Wilcoxon 2výběrový Mann-Whitney	Kruskal-Wallis (-H test)	Wilcoxon 2výběrový pro závislé	Friedman	Spearmanův korelační koeficient	Poloha
	Siegel - Tukey			Shorac			Variabilita
Alternativní	Test dobré shody	χ-kvadrát 2*2, Fisher	χ-kvadrát, k*m tabulka	MC Nemar	Q-test	Kontingenční korelační koeficient	Cetnosti výskytu

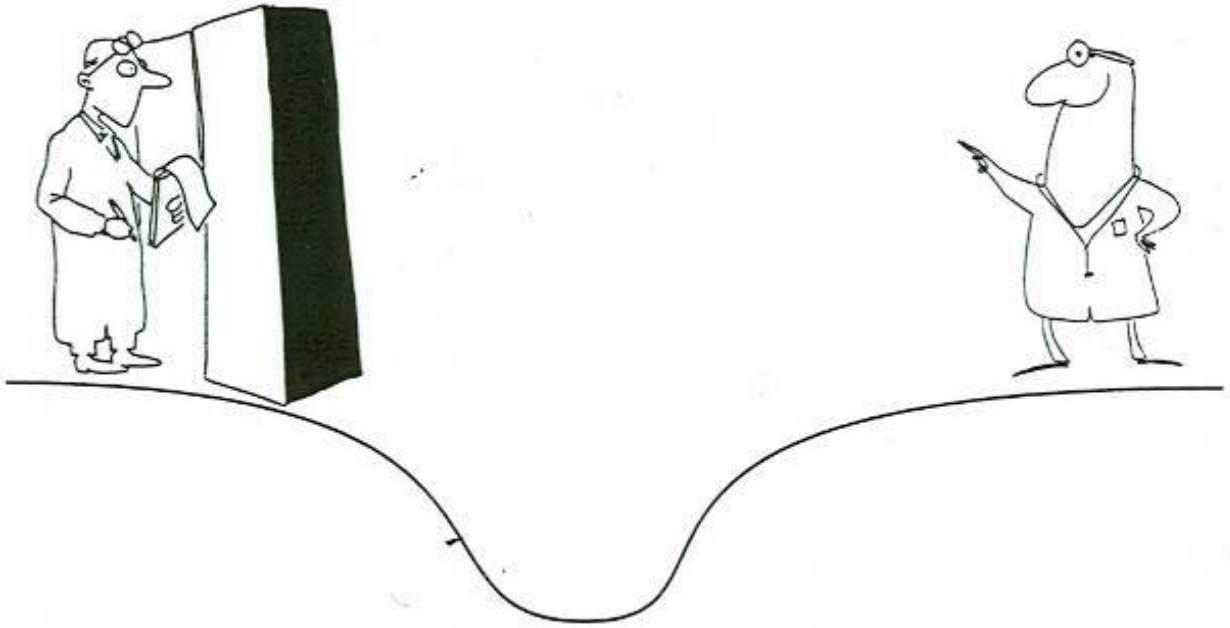
Data metrická (měřitelná) symetrická ordinální (pořadí) nebo asymetrická alternativní (ano-ne)

Srovnáváme střední hodnoty nebo variability

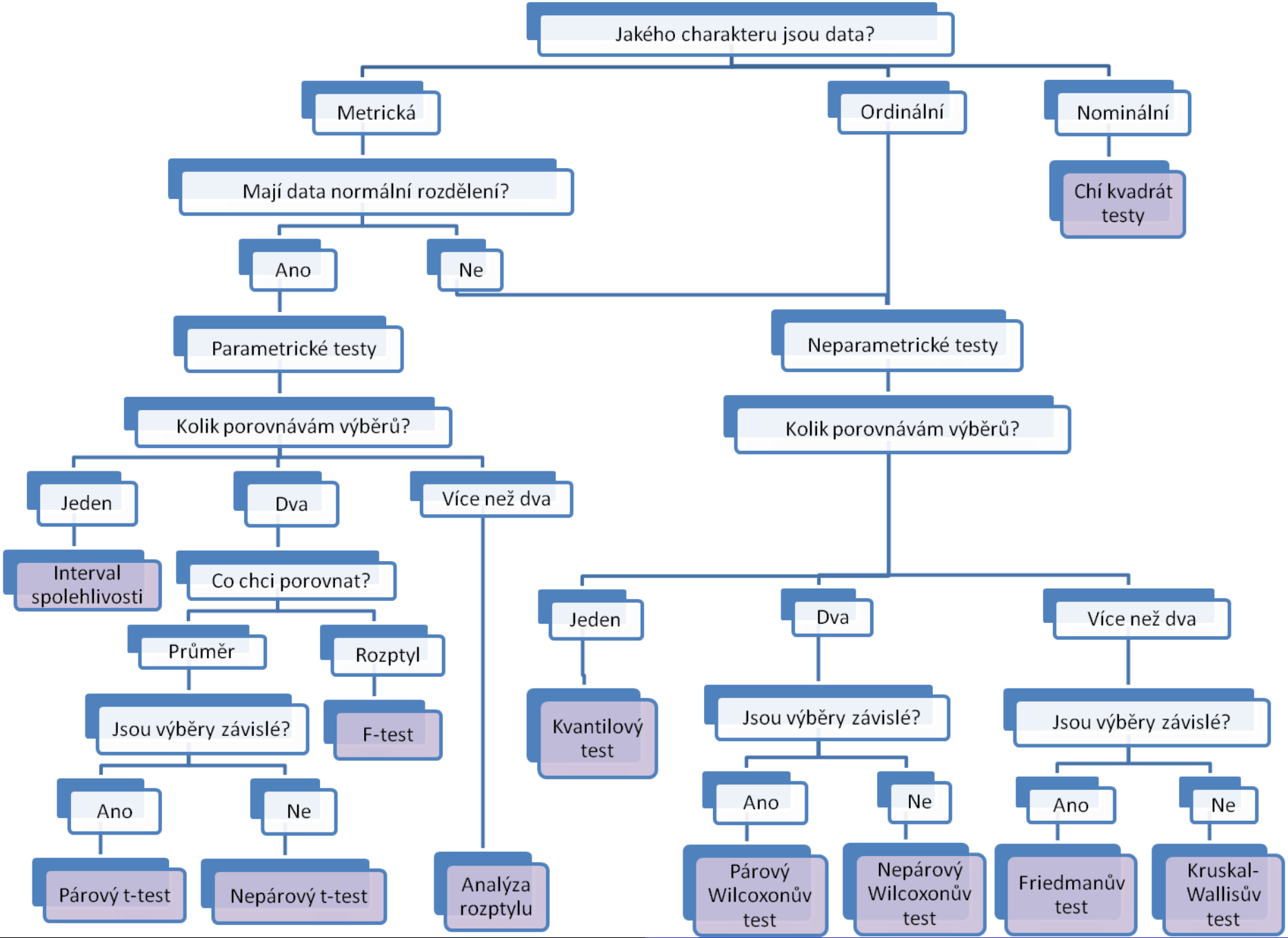
Jak správně zvolit test hypotézy?

- vyžaduje zkušenost a know-how
- nejlépe s pomocí statistika
- dobrá zpráva: většina uvedených testů hypotéz existuje v user-friendly online podobě např. na adrese (anglicky) http://vassarstats.net/
- oproti Statistice či R nám ale nenakreslí graf atd.

Spolupráce se statistikem ☺



- Těžko dostupná konzultace
- Obtížně se s ním domluvím
- Někdy ještě zpochybní mé výsledky


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?
Jsou výběry závislé?
Jsou výběry závislé?
F-test
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