INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

VYUŽITÍ STATISTICKÝCH METOD PŘI HODNOCENÍ EPIDEMIOLOGICKÝCH STUDIÍ

Pátek 1.2.2013

8:00 – 15:30

Základní zpracování dat a hodnocení statistických hypotéz

Ing. Hana Tomášková, Ph.D.

Biostatistika

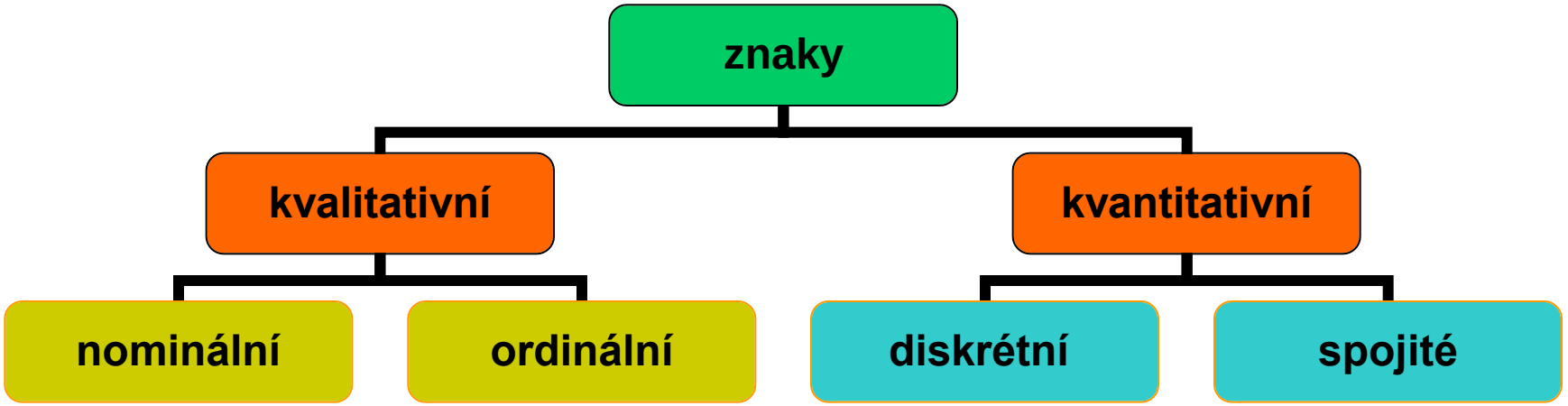
- Biostatistika představuje soubor statistických metod a postupů, které se používají v biomedicínckých oborech při studiu a pochopení biologických jevů.
- Objektem studia statistiky je soubor, kolektiv, skupina, populace. Jevy, kterými se statistika zabývá, jsou vždy nějakým způsobem vázány na určitý soubor, označují se proto jako ***jevy hromadné***.
- ***Induktivní (matematická) statistika*** nezůstává jen u popisu hromadných jevů. Soubory, které má při svém studiu k dispozici, chápe jako dílčí, omezené ***výběry*** ze souborů podstatně rozsáhlejších, tzv. ***populací.***

Biostatistika

- Induktivní statistika umožňuje zobecňovat poznatky, získané studiem výběru, na příslušné populace. Taková zobecňování jsou vždy spojena s určitým rizikem omylu. Hlavní úlohou statistické indukce je měření těchto rizik a hledání způsobu jak je redukovat.
- Aplikovat statistiku znamená shromažďovat data o studovaných jevech a zpracovávat je – třídit, počítat, interpretovat.

Znaky

Pokud sledujeme určitý soubor objektů, vždy se na těchto objektech sledují (měří) určité ***znaky***, jedná se o vlastnosti těchto objektů.



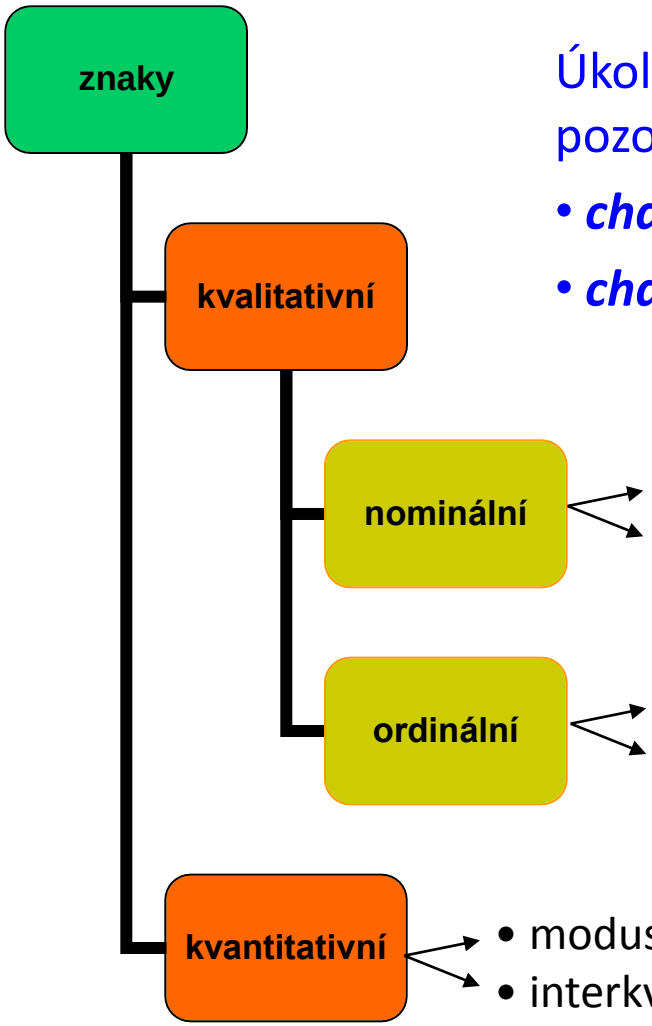
Nelze seřadit
- pohlaví
- krevní skupina
- etnikum

Seřazené
- vzdělání
- závažnost úrazu
- míra spokojenosti

Celočíselné
- počet porodů
- počet dní PN
- počet léků

Jakékoliv hodnoty
- BMI
- TK
- hodnota cholesterolu

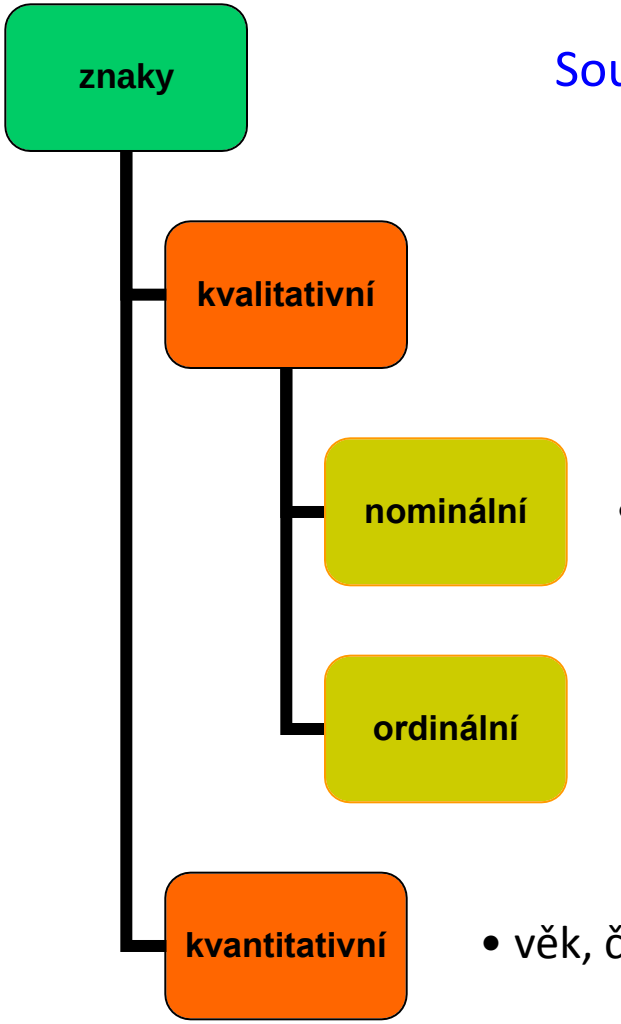
Popisná statistika



Úkolem popisné statistiky je provést popis souboru pozorovaných subjektů jako celku.

- ***charakteristiky polohy*** – střední hodnota
- ***charakteristiky variability (rozptýlení).***

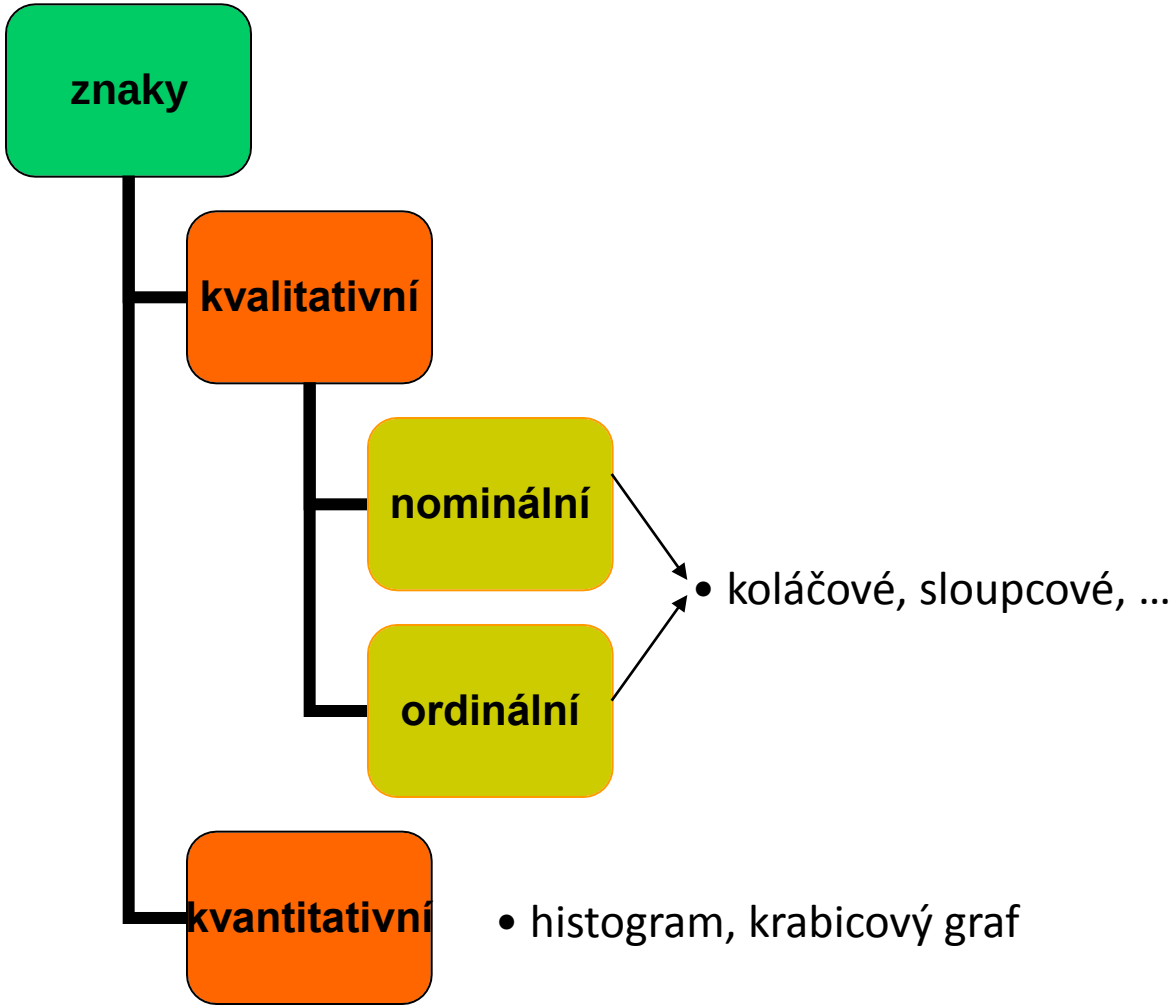
Příklady - popisná statistika

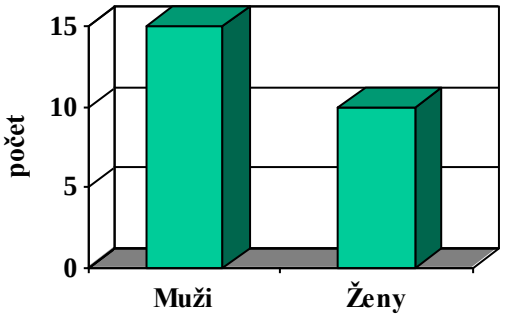


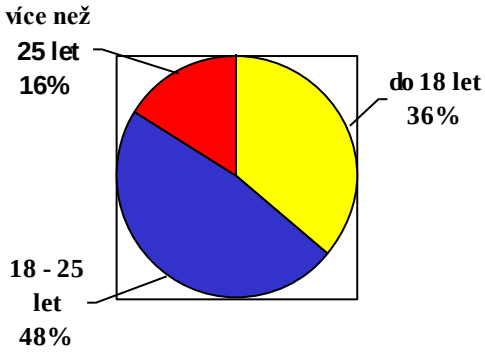
Soubor Lázně_data

- pohlaví, sociální postavení, fyzicky namáhavá práce,…
- vzdělání, zdravotní stav, …
- věk, částka

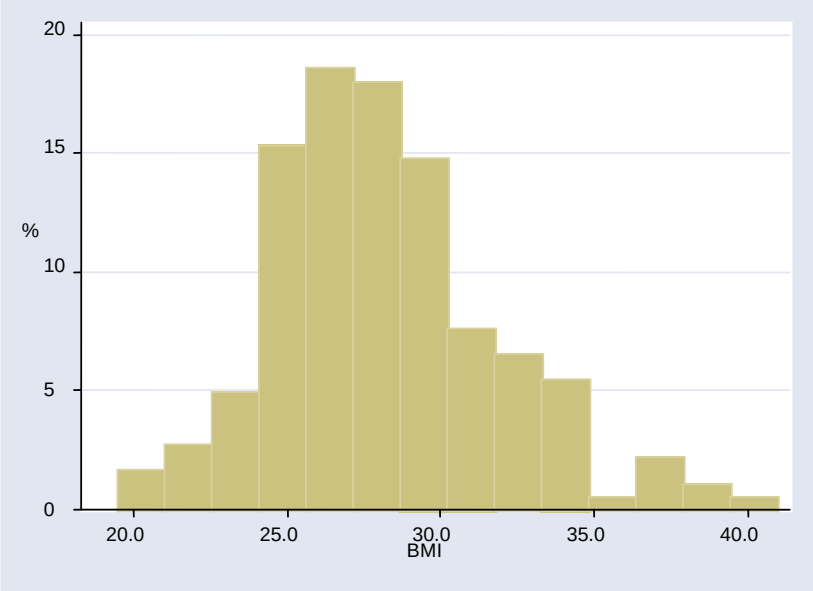
Příklady – grafické znázornění

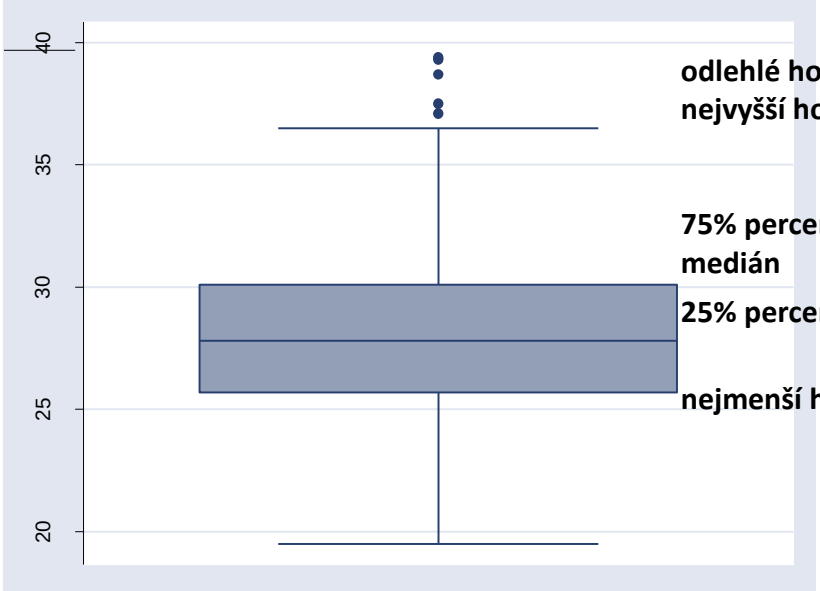






Příklady – grafické znázornění





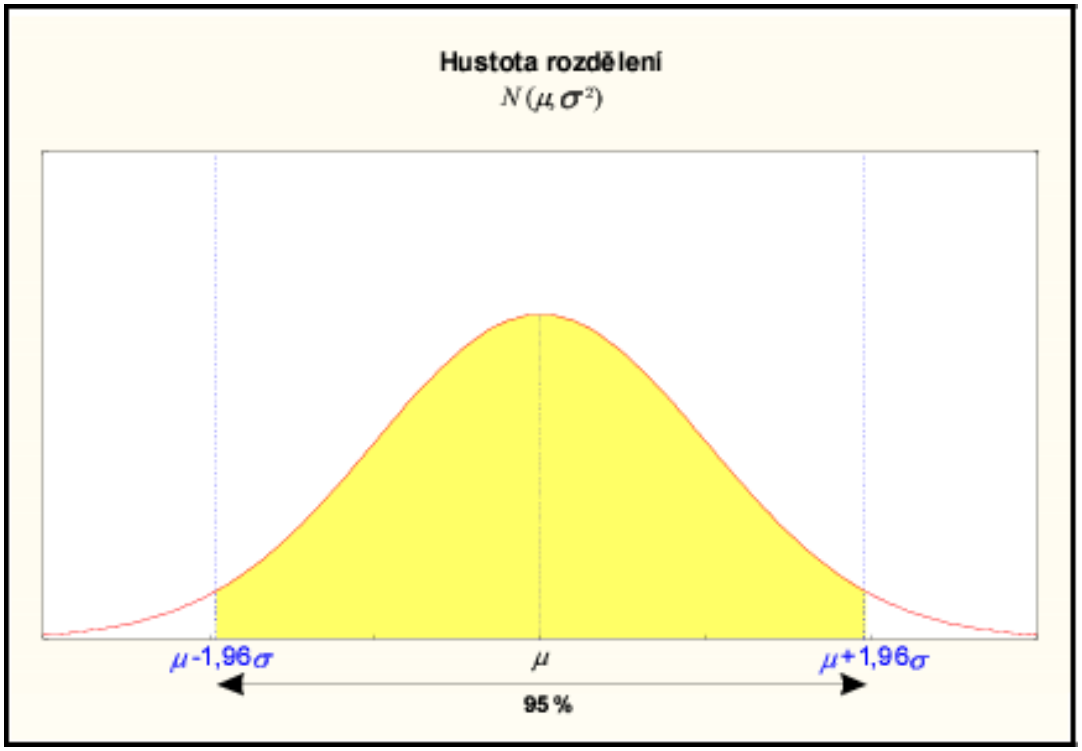
Histogram

Krabicový graf (Box plot)

Charakteristiky populace a výběru

charakteristika		populace	výběr (vzorek)
ar. průměr	mean	μ	$\bar{x}$
směrodatná odchylka	SD	σ	s
proporce	proportion	π	p

Normální rozložení dat – Gaussova křivka



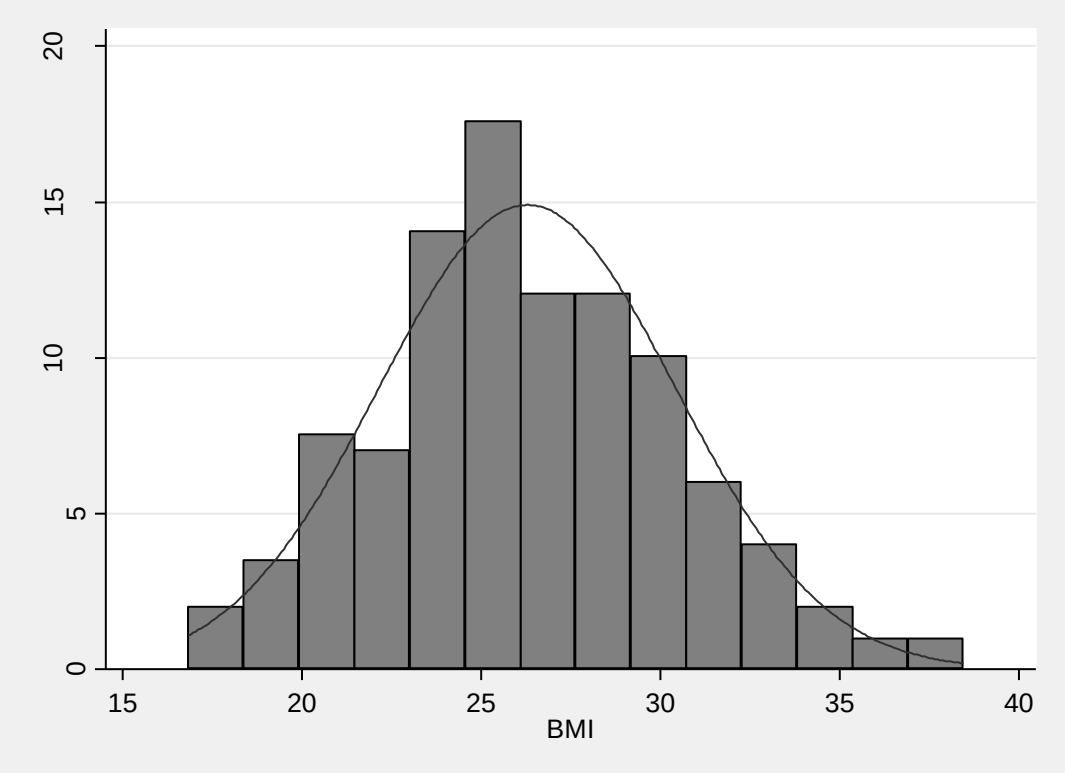
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

68 % plochy pod křivkou je v intervalu $(\mu - \sigma, \mu + \sigma)$

95 % plochy pod křivkou je v intervalu $(\mu - 1{,}96\ \sigma, \mu + 1{,}96\ \sigma)$

99 % plochy pod křivkou je v intervalu $(\mu - 2{,}58\ \sigma, \mu + 2{,}58\ \sigma)$

Příklad - Normální rozložení dat



Interval		
68%	22,2	30,4
95%	18,3	34,3
99%	15,7	36,9

sum BMI if pohlavi==1

Variable	Obs	Mean	Std. Dev.	Min	Max
BMI	199	26.3	4.1	16.8	38.5

Bodové a intervalové odhady

- **Statistické odhady** - na základě statistik získaných z dat v náhodném výběru (vzorku) odhadujeme populační parametry
- **Reprezentativní výběr** je takový, který odráží strukturu celého zkoumaného souboru (populace).
- Reprezentativnost lze dosáhnout vhodnou metodou výběru prvků do souboru. Za reprezentativní metodu výběru se považuje například **prostý náhodný výběr**.
- Statistické charakteristiky počítané na výběrových souborech se nazývají ***bodové odhady*** příslušných charakteristik populace. Odpovídající charakteristiky v základním souboru se nazývají ***parametry populace*** a značí se písmeny řecké abecedy.

Bodové a intervalové odhady

- Výběrová charakteristika se liší od skutečné neznámé hodnoty příslušného parametru v definovaném základním souboru o ***výběrovou chybu***. Výběrová chyba vzniká vlivem náhodných výkyvů závisejících na tom, které prvky ze základního souboru budou do výběru zařazeny.
- Odečtením a přičtením výběrové chyby k bodovému odhadu se stanoví takzvaný ***intervalový odhad***. Intervalový odhad je interval, ve kterém je s velkou pravděpodobností populační parametr obsažen. Tento interval se nazývá ***interval spolehlivosti*** (*IS*).

Rozsah výběru

kuřáci	N	% kuřáků	D - výběrová chyba	95% IS	
5	**30**	17%	13%	**3%**	**30%**
17	**100**	17%	7%	**9%**	**24%**
50	**300**	17%	4%	**12%**	**21%**
83	**500**	17%	3%	**13%**	**20%**
167	**1 000**	17%	2%	**14%**	**19%**
1667	**10 000**	17%	1%	**15,9%**	**17,4%**

Příklady - Intervalový odhad

Intervalový odhad průměru populace μ a neznáme populační směrodatné odchylky s se stanoví na základě výběrového průměru ($\bar{x}$) a střední chyby tohoto průměru (s_e), krajní body pro 95% interval spolehlivosti se stanoví:

$$95\%IS = \bar{x} \pm t_{0,975} s_e$$

95% IS: $\bar{x} \pm 1{,}96 s_e$

99% IS: $\bar{x} \pm 2{,}58 s_e$

Příklady – Excel – popisná statistika

Soubor: Data_lazně

Podle typu veličin vypočtěte základní charakteristiky

Využijte:

- Kontingenční tabulka (Vložení – Kontingenční tabulka)
- Doplněk – Analýza dat (Data – Analýza dat – Popisná statistika, Histogram)

Testování statistický hypotéz

1. Stanovení **nulové hypotézy H_0 a alternativní hypotézy H_a (H_1)**
2. Stanovení **hladiny významnosti** α
3. Výběr **testového kritéria** - statistického testu
4. Výpočet testového kritéria a **p-hodnoty**
5. Rozhodnutí o Ho na základě p-hodnoty
6. Interpretace výsledků

Testování statistický hypotéz – H_0, H_a

1. Stanovení ***nulové hypotézy H_0*** a ***alternativní hypotézy H_a (H_1)***

Př. H_0 – není rozdíl v průměrném BMI u osob s Ca jícnu a bez Ca

Př. H_a – je rozdíl v průměrném BMI u osob s Ca jícnu a bez Ca – ***oboustranná hypotéza***

Př. H_a – osoby s Ca jícnu mají nižší průměrné BMI než osoby bez Ca – ***jednostranná hypotéza***

Testování statistický hypotéz – hladina významnosti α

2. Stanovení ***hladiny významnosti*** α

Hladina významnosti odpovídá riziku chyby I typu (zamítnutí nulové hypotézy, když tato platí) stanovíme předem (obvykle 0,05 neboli 5%)

Chyby při testování statistických hypotéz

Rozhodnutí výzkumníka	**Skutečnost**	
	Ho platí	**Ho neplatí**
Ho nezamítáme	**správné rozhodnutí** $P=1-\alpha$ (hladina spolehlivosti)	**chyba II. druhu** $P=\beta$
Ho zamítáme	**chyba I. druhu** $P=\alpha$ (hladina významnosti)	**správné rozhodnutí** $P=1-\beta$ (síla testu)

Testování statistický hypotéz – výběr testového kritéria

3. Výběr *testového kritéria – statistický test*

- typ ***Ho*** *– co chceme zjistit*
- typ znaků (proměnných)
- rozložení dat
- počet výběrů

Testovací metody parametrické a neparametrické

Parametrické metody – vyžadují normalitu sledovaných náhodných veličin, používají se pro testování hypotéz o výběrových parametrech. Příklad - *t*-test.

Neparametrické metody – normalitu nevyžadují, jsou numericky velmi jednoduché, využívají pouze menší část informace. Mají obvykle pro zvolenou alternativní hypotézu větší pravděpodobnost chyby II. druhu než příslušná metoda parametrická. Příklad - χ^2 test.

Příklady statistických testů

proměnná		1 proměnná					2 prom.	k prom.
data \ výběr		nezávislé			závislé			
		1 výběr	2 výběry	k výběrů	2 výběry	k výběrů		
spojitá	poloha	Jedno-výběrový t test	Nepárový t test	ANOVA	Párový t test	ANOVA opak. měření	Pearson korelační koefic.	Mnohoná-sobný korelační koeficient
	variabilita	IS	F test	Bartlett				
ordinální	poloha	Jedno-výběrový Wilcoxon	Mann-Whitney test	Kruskal – Walis test	Párový Wilcoxon	Friedman test	Spearman korelační koefic.	Vícerozměrná analýza
	variabilita							
nominální	pravd. výskytu	χ^2	χ^2 2x2 tab Fisher	χ^2 rxs tab			Kontingen-ční korel. koefic.	

Testování statistický hypotéz – p-hodnota

4. Výpočet ***testového kritéria*** *a* ***p-hodnoty***

Testové kritérium - je matematický vzorec, do kterého se dosazují numerické hodnoty charakteristik, zjištěné ze souboru dat. Protože výběrové charakteristiky (aritmetický průměr, směrodatná odchylka, relativní četnost apod.) jsou náhodné veličiny, jejich hodnoty závisí na tom, které prvky byly zahrnuty do výběru, je i hodnota testového kriteria náhodnou veličinou.

p-hodnota (p-value) **-** udává pravděpodobnost s jakou bychom za předpokladu platnosti hypotézy dostali náš výsledek (nebo výsledek ještě extrémnější) pokud bychom experiment mnohokrát opakovali.

Testování statistický hypotéz - Rozhodnutí o H_0

5. Rozhodnutí o H_0 na základě p-hodnotě

Srovnání p-hodnoty a hladiny významnosti α

$p < \alpha$ (0,05; 0,01) – Ho zamítáme
$p \geq \alpha$ (0,05; 0,01) – Ho nezamítáme

Závěr - H_0 zamítáme/nezamítá na hladině významnosti α.

Testování statistický hypotéz - Interpretace

6. Interpretace výsledků

Př. H_0 – není rozdíl v průměrném BMI u osob s Ca jícnu a bez Ca

Př. H_a – je rozdíl v průměrném BMI u osob s Ca jícnu a bez Ca – ***oboustranná hypotéza***

1. *p < α (0,05; 0,01) – Ho zamítáme a přijímáme Ha*
Interpretace: Byl zjištěn statisticky významný rozdíl (p = …) v průměrné hodnotě BMI u osob s Ca jícnu a bez Ca.

2. *p ≥ α (0,05; 0,01) – Ho nezamítáme*
Interpretace: Nebyl zjištěn statisticky významný rozdíl (p = …) v průměrné hodnotě BMI u osob s Ca jícnu a bez Ca.

Vybrané statistické test

- Test homogenity pro kvalitativní data
- T-test pro dva výběry

Pro výpočet použijeme MS Excel a program OpenEpi
http://openepi.com/OE2.3/Menu/OpenEpiMenu.htm