ZPRACOVÁNÍ STATISTICKÝCH DAT

Osnova

1. Statistické třídění
2. Sestavování tabulek
3. Sestavování grafů
4. Statistické charakteristiky
5. Pravděpodobnosti
6. Statistické odhady
7. Testování hypotéz

Úvod do zpracování statistických dat

- Zpracování statistických dat je využitelné výhradně v kvantitativním výzkumu – lze využít pouze u velkého množství dat
- Data musí být pořizována dle určitých pravidel (dodržení podmínek a postupů platných pro sběr dat)
- Popis dat a jejich analýzu je možné zpracovat na standardním pc se sofwarovým vybavením
  - Např. nejrozšířenější je program Excel – data lze vkládat, roztřídit, zpracovat do tabulek a grafů a provést základní výpočty
  - Software Statistika – pořízení je finančně náročné, využívají zejména pracoviště, která se zabývají výzkumem profesionálně

1. Statistické třídění

= rozdělení souboru dat do skupin (do tříd) dle třídících znaků

- podle počtu znaků rozdělujeme třídění:
  - jednostupňové – třídíme jen podle 1 znaku
  - vícestupňové (třídíme současně dle 2-4 znaků, používáme max. 4 znaky – pak je třídění nepřehledné)
- 1 třídící znak: klasickým příkladem je rozdělení souboru na muže a ženy – dle pohlaví.
- Více stupňové: např. rozdělení souboru podle pohlaví, věku, refrakční vady…apod.
- produktem třídění je tzv. rozdělení (rozložení) četnosti (počet v určité skupině)

2. Zásady pro sestavování tabulek

- Pokud máme data roztříděná a známe jejich četnosti, zapisujeme je do tabulek (tabulka rozdělení četnosti)
- Číselné údaje jsou uspořádány do vodorovných řádků
- Obsah sloupců uvádí „hlavička“ (zdravotní stav)
- Obsah řádků vyjadřuje legenda (pohlaví)

POHLAVÍ	ZDRAVOTNÍ STAV					
	velmi dobrý	spíše dobrý	tak napůl	spíše špatný	velmi špatný	CELKEM
muži	10,8	21,5	11,9	3,3	0,7	48,2
ženy	9,3	22,5	14,4	5,1	0,5	51,8
CELKEM	20,1	44,0	26,3	8,4	1,2	100,0

Třídění statistických tabulek

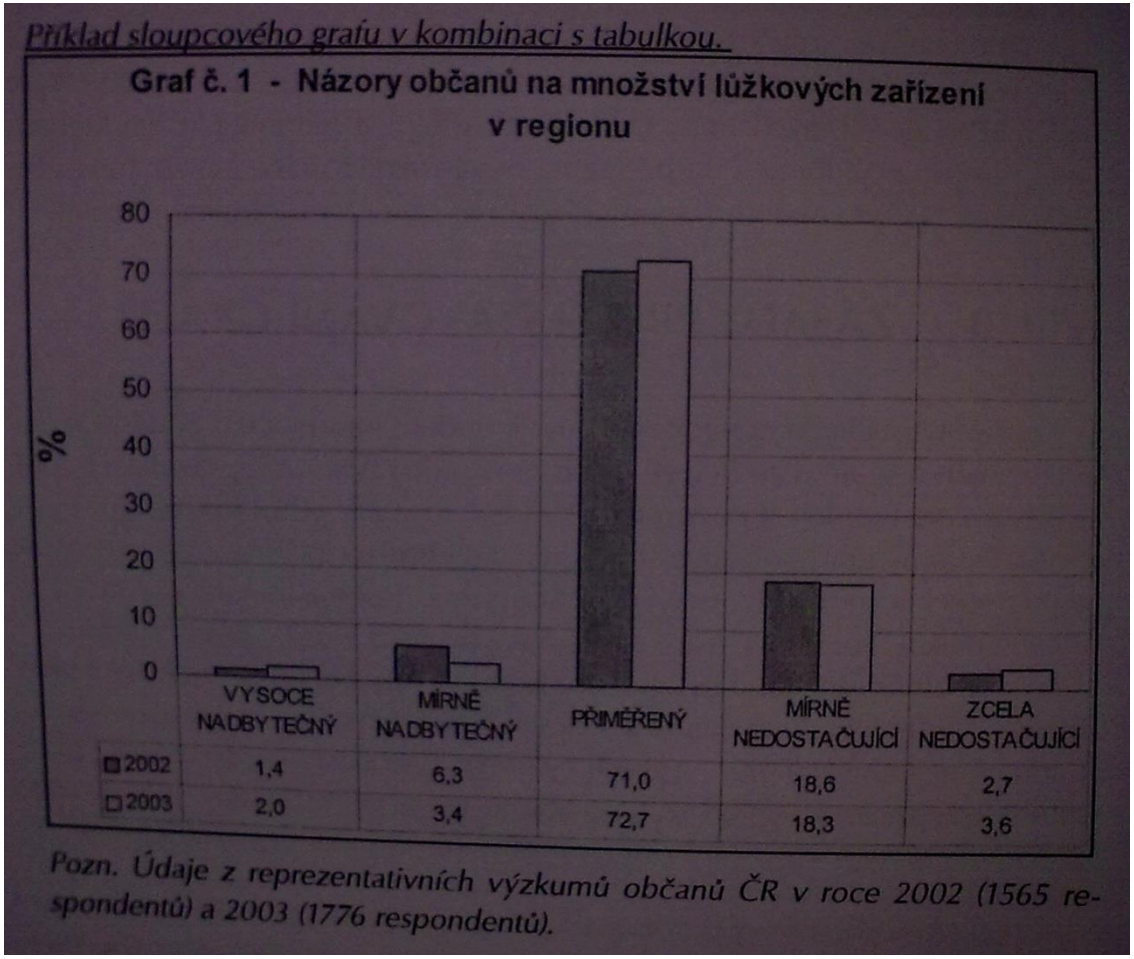
- Prosté tabulky = uvádějí data bez třídění
- Skupinové tabulky = třídění podle jednoho znaku
- Kombinační tabulky = soubor je roztříděn dle 2 a více znaků
  - Korelační tabulky = druh kombinační tabulky, užíváme ke studium závislosti 2 kvantitativních znaků
- Kontingenční tabulky = závislost 2 kvalitativních znaků

3. Sestavování grafů

- Názornější a rychlejší vyjádření dat
- Zpravidla horizontální osa (x) a vertikální osa (y)
- Nejčastěji se používají sloupcové grafy
- Umožňují zhodnocení sledovaného znaku (jednovrcholové, vícevrcholové)
- V rámci grafů využíváme i slovní charakteristiky – název, podtitul, poznámky nebo vysvětlivky ke grafu

Příklad sloupcového grafu s tabulkou

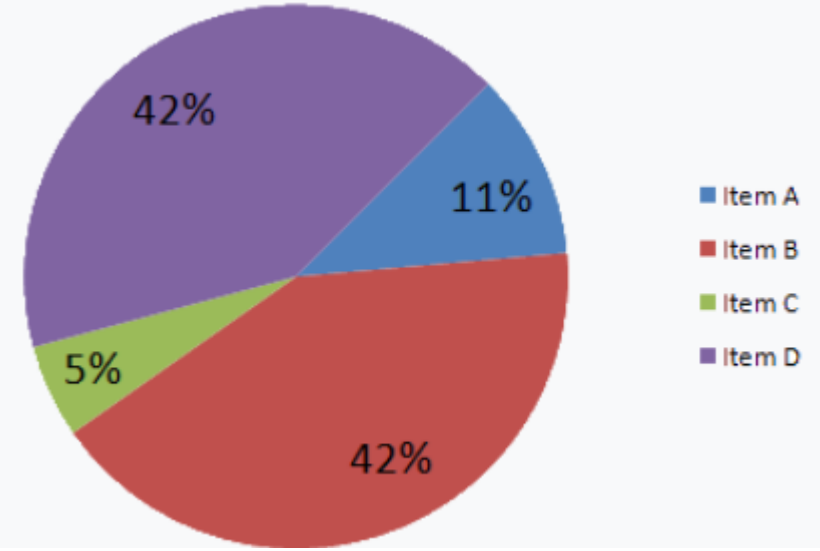
Příklad sloupcového grafu v kombinaci s tabulkou.



Pozn. Údaje z reprezentativních výzkumů občanů ČR v roce 2002 (1565 respondentů) a 2003 (1776 respondentů).

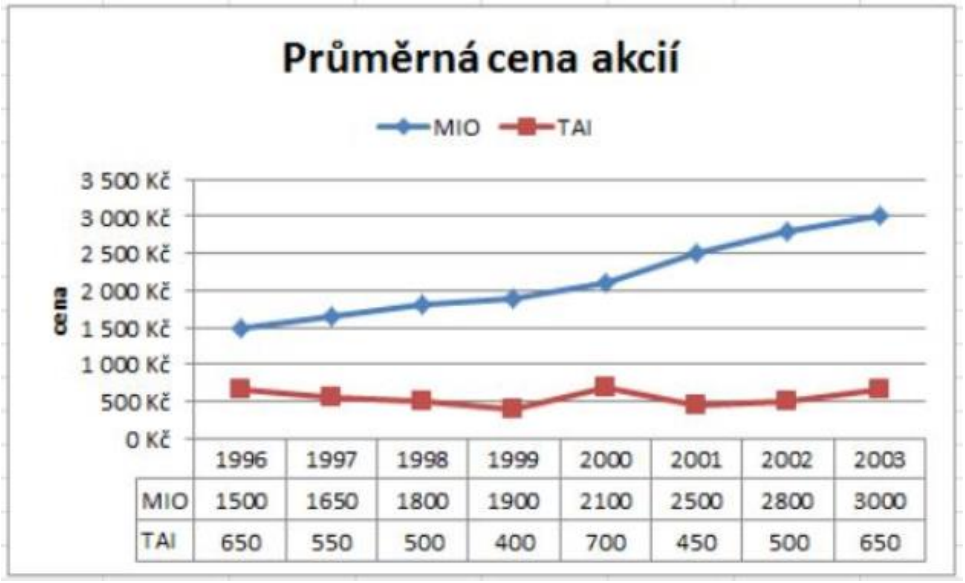
Graf výsečový (kruhový, sektorový)

- Plocha celého kruhu je rovna 100%
- Odlišení šrafováním nebo barvou



Spojnicový graf

- Vyjádření vývoje jevu v určitém čase



	1996	1997	1998	1999	2000	2001	2002	2003
MIO	1500	1650	1800	1900	2100	2500	2800	3000
TAI	650	550	500	400	700	450	500	650

4. Statistické charakteristiky

- Statistické ukazatele umožňují podat informaci o souboru
- Výběrové charakteristiky = souvisí s výběrovým souborem, umožňují ověřování hypotéz

- Základní statistické ukazatele jsou:

  A. Relativní ukazatele (poměrná čísla)

  B. Ukazatele polohy (střední hodnoty)

  C. Ukazatele variability

A. Relativní ukazatele, poměrná čísla

- v případě hodnocení kvalitat. znaků
- data jsou vyjádřena v relativních číslech
  - vznikají podílem 2 absolutních čísel
  - např. spokojenost pacientů v roce 2000 (1590 pacientů) a v roce 2001 (12890 pacientů)

- Dle povahy absolutních čísel rozlišujeme tyto relativní ukazatele:
  - Ukazatel extenzity – vyjadřuje podíl části k celku (v %)
  - Ukazatel intenzity, četnosti – vyjadřují frekvenci výskytu jevu v daném souboru
  - Indexní čísla – relativní ukazatelé, používáme k vyjádření změn, které nastaly ve vývoji sledovaného jevu v čase

B. Ukazatele polohy (střední hodnoty)

- Sledujeme, zda a jak se náhodné veličiny kupí kolem střední hodnoty
- Střední hodnota = číslo, které v určitém smyslu zastupuje jednotlivé hodnoty zkoumaného souboru
- Význam středních hodnot: umožňuje reálné a jednoduché srovnání úrovně zkoumaného jevu v několika souborech

- Rozdělení středních hodnot:
  - **Aritmetický průměr – prostý, vážený**
  - **Medián**
  - **Modus**

Aritmetický průměr prostý

- nejčastěji používanou statistickou charakteristikou
- charakterizuje střed normálního rozdělení
- vzorec aritmetického průměru:

$$\bar{x} = \frac{\sum x_i}{n}$$

- součet všech naměřených hodnot $x_1, x_2 \ldots\ldots x_n$ dělíme počtem měření n
- používáme, když se nám znaky často opakují

Vážený aritmetický průměr

- se užívá v případě, že jednotlivé hodnoty mají různou důležitost (váhu), tu je nutno přiřadit každé hodnotě.
- v případě, že mají všechny hodnoty stejnou váhu, je vážený průměr totožný s průměrem aritmetickým
- můžeme použít tehdy, pokud známe velikosti a průměry dílčích podsouborů

Krok 1

Známka:	**5**	**3**	**1**	*Průměr je:* $9 : 3 = \underline{\mathbf{3}}$
Váha:	*6*	*2*	*2*	

Krok 2

$$\frac{5 \times 6 + 3 \times 2 + 1 \times 2}{6 + 2 + 2}$$

Krok 3

$$\frac{38}{10} = 3{,}8$$

Vážený průměr je: 3,8

Nevhodné použití aritmetického průměru

- Malý soubor s extrémními hodnotami
- Asymetrické rozdělení souboru
- Není vymezen dolní nebo horní interval četností

Medián

= hodnota prostředního člena souboru, který je uspořádán dle velikosti (u sudého počtu je to průměr ze dvou středních)

- není ovlivněn robustností, tj. hodně malými nebo hodně velkými chybami
- pomocí mediánu vyjadřujeme střední hodnotu statistického souboru, ve kterém dochází k extrémnímu výkyvu – výhoda oproti aritmetickému průměru

Percentil (kvantil)

= další pořadový ukazatel

P10 = desátý percentil, tj. 10% hodnot je menších a 90% je větších

Modus

= takový člen, který se ve sledované skupině vyskytuje nejčastěji

- Zkoumáme intenzitu jevu, soubor může mít i více vrcholů
- Př. : údaje jsou roztříděny do nějakých intervalů (v našem případě dle věku) a zkoumáme interval s nejvyšším počtem

Věk nemocných - počet nemocných

- 5-14 5
- 15-24 33
- 25-34 458
- 35-44 1123
- 45-54 1136
- 55-64 746
- 64-74 233
- 75-84 24

C. Ukazatele variability

= proměnlivost vlastností v čase a v prostoru se projevuje rozptýlením znaků k určitým hranicím

- K základním charakteristikám variability patří:
  - I. Variační šíře (rozpětí)
  - II. Průměrná odchylka
  - III. Rozptyl
  - IV. Směrodatná odchylka
  - V. Variační koeficient

I. Variační šíře (rozpětí)

= daná rozdílem mezi největší a nejmenší hodnotou souboru

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

- Je závislá pouze na krajních hodnotách souboru, proto pouze orientační
- Za vyhovující míru variability lze rozpětí považovat pouze u malých souborů (do 10 pozorování)

II. Průměrná odchylka

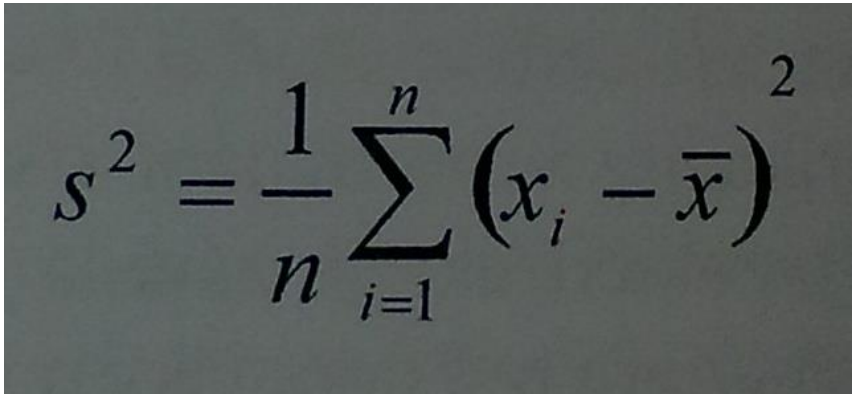
- Dokonalejší mírou variability - je závislá na všech hodnotách souboru
- Rozlišujeme:
  - PO prostá
  - Vážená PO = pokud se hodnoty opakují
- Bereme absolutní hodnoty rozdílu od průměru, ty mohou negativně ovlivnit posunutí hodnot od průměru

$$\bar{d} = \frac{\sum |x_i - \bar{x}|}{n}$$

$$\bar{d} = \frac{\sum |x_{i-} \bar{x}| \cdot f_i}{\sum f_i}$$

III. Rozptyl

= je to průměr čtverců odchylek jednotlivých pozorování od aritmetického průměru



IV. Směrodatná odchylka

= odmocnina z rozptylu

- Udává rozptýlenost dat ve stejných jednotkách jako jsou původní data a průměr
- Používá se i název standardní odchylka
- Umožňuje vymezit hranice, ve kterých se nachází určité množství jednotek

V. Variační koeficient

= podíl směrodatné odchylky a průměru

- Je to relativní míra variability souboru a uvádí se často v %
- Používá se při srovnání variability dvou a více souborů s odlišným průměrem nebo se znaky v odlišných jednotkách

$$V = \frac{s}{\bar{x}} \times 100(\%)$$

- Pozn.: Při asymetrickém rozdělení soboru se doporučuje variabilitu vyjadřovat pomocí kvantilů (decil, percentil)

5. Pravděpodobnosti

= teorie pravděpodobnosti studuje a formuluje zákonitosti náhody a jejího využívání

- Náhoda = souhrn drobných špatně zjistitelných vlivů, které ovlivňují výsledky
- Náhodný jev = lze určit, zda nastane, či nikoliv (orel/panna)
- Náhodný pokus = situace, která vede k výskytu náhodného jevu (hod korunou)

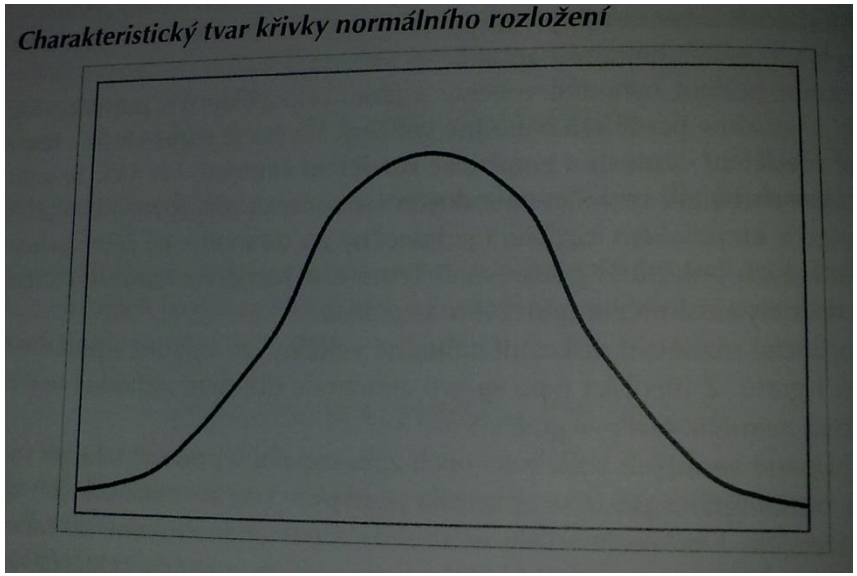
Rozdělení náhodné veličiny

= soubor hodnot náhodné veličiny

- Rozeznáváme RNV:
  - Teoretické = rovná se nekonečnu
  - Empirické = je dáno výsledky pokusů, pozorování
    - Diskrétní = lze vyjádřit pomocí tabulek a grafů
    - Spojité = vyjádření je složité, používáme intervaly
- RNV se řídí těmito zákony:
  - Normální (Gaussovo)
  - Binomické
  - Poissonovo

Normální - Gaussovo rozdělení

- U kvantitativních znaků
- Spojité rozložení
- Biologické jevy, které se týkají člověka
- Je definováno standardní odchylkou a průměrem



Binomické rozdělení

- Vyskytuje se u diskrétní veličiny
- Pokud děláme pokus, kde existují jen dvě možnosti (např. hod mincí)
- Náhodná veličina nabývá pouze celočíselných hodnot (1 nebo 0)

Poissonovo rozdělení

- Týká se diskrétních veličin, např. počet zákazníků v obchodě, počet automobilů projetých určitým úsekem
- Může mít tyto vlastnosti:
  - Pravděpodobnost výskytu události v intervalu je přímo úměrná délce intervalu
  - Události se vyskytují nezávisle jak ve stejném intervalu, tak mezi po sobě jdoucími intervaly

6. Statistické odhady

- Základní versus výběrový soubor
- Teorie odhadů = na základě kvalifikovaných odhadů provádíme statistickou indukci
- Náhodný výběr umožní, aby se jakákoliv jednotka ze základního souboru dostala do výběrového
- Základní soubor charakterizuje střední hodnota μ, směrodatná odchylka σ a rozptyl σ2
- Výběrový soubor charakterizuje výběrová střední hodnota $\bar{x}$, výběrová směrodatná odchylka s a výběrový rozptyl s2

7. Testování hypotéz

- Pomocí testovací statistiky (rozptyl, střední hodnota) určujeme, zda se dva ukazatele od sebe liší reálně nebo náhodně
- Dle statistických testů zjišťujeme, jestli hypotéza platí nebo neplatí
- Rozdělení hypotéz dle statistiky:
  - Parametrické – odlišují se pouze hodnotami parametrů
  - Neparametrické – odlišuje se ještě i tvarem rozložení funkce

Hypotézou ověřujeme především tato tvrzení:

1. Zkoumaný výběr pochází z populace, která má určité teoretické rozdělení
2. Dva zkoumané výběry vychází ze stejného základního souboru
3. Existuje lineární závislost mezi dvěma nebo více veličinami souboru
4. Jedna nezávisle proměnná ovlivňuje sledovanou závislou více než druhá

Volba hladiny významnosti

Pravděpodob-nost chyby	>0,05	≤0,05	≤0,01	≤0,001
Slovní vyjádření	nesignifikantní	signifikantní	velmi signifikantní	velmi vysoce signifikantní
Písmenová symbolika	n.s.	s.	v.s.	v.v.s.
Grafická symbolika		*	**	***

Při statistickém testování si musíme zvolit hladinu významnosti. Např. 0,05 je 5% pravděpodobnost chyby a zbylých 95% případů bude mít statisticky významný výsledek (signifikantní výsledek). Je možno si zvolit i přísnější kritérium – 0,01 - kde 1% bude pravděpodobnost chyby a 99% případů bude mít statisticky významný výsledek.

Obecný postup při testování hypotéz

1. Formulace hypotézy
2. Zvolíme hladinu významnosti
3. Některou metodou náhodného výběru nasbíráme data
4. Vybereme vhodný statistický test
5. Z dat vypočítáme hodnotu testovacího kritéria
6. Pro zvolenou hladinu významnosti v tabulce vyhledáme kritickou hodnotu
7. Provedeme statistické rozhodování: jestliže hodnota testového kritéria překročí kritickou hodnotu, zamítáme nulovou hypotézu ve prospěch alternativní. V opačném případě prohlásíme odchylku za nevýznamnou na této hladině.

Parametrické testy

- Používají se v případě, jestliže výběrové soubory, se kterými pracujeme, mají alespoň přibližně normální rozložení
- Poskytují zpravidla více informací o zkoumaných souborech než neparametrické testy
- Porovnáváme např. průměr nebo rozptyly souborů
- Při testování záleží na volbě hladiny významnosti - zpravidla 0,05
- Používáme parametrický test:
  - F-test (Fischerův)
  - T-test (Studentův)

Fischerův test

= parametrický test významnosti, kde testujeme hypotézy o rozptylu

- Umožňuje nám určit signifikantnost odlišnosti na určité hladině pravděpodobnosti
- Můžeme ho použít např. při měření účinnosti určitého léku ve dvou různých souborech

Studentův test

- Testujeme významnost rozdílu dvou průměrů
- Používáme je v situacích:
  - Dva výběrové soubory s odlišnými průměry, testujeme zda odlišnost je náhodná nebo podstatná
  - Sledování vlivu dvou faktorů (např. léčiv)
  - Zda náleží výběrové soubory témuž základnímu souboru

Neparametrické testy

- Parametrické testy nemůžeme použít za těchto okolností:
  - Teoretické rozložení proměnné v populaci je příliš malé
  - Rozložení nelze převést na normální
  - Údaje mají povahu pořadových čísel, tj. jsou řazeny vzestupně nebo sestupně
- U neparametrických testů nepotřebujeme znát parametry souboru, bývají jednodušší, ale poskytují méně informací

Neparametrické testy - rozdělení

1. **Kolmogorovův-Smirnovův test** pro jeden výběr – lze jej použít, pokud rozdělení zákl. souboru je spojité
2. **Kolmogorovův-Smirnovův test** pro dva nezávislé výběry – hodnocení shody četností dvou srovnávaných výběrů
3. **Znaménkový test** – pro soubory uspořádané ve dvojicích, znaménka ukazují změnu, pokud se znaménka vyskytují se stejnou pravděpodobností, znamená to, že mezi soubory není rozdíl
4. **Wilcoxonův test** pro párové hodnoty – nebere pouze +/- ale bere konkrétní hodnoty rozdílu, ověřuje rozdíly při opakovaném měření

Děkuji za pozornost

Studijní literatura:

- **Gerylová, Anna. Úvod do statistiky. Brno: Masarykova univerzita, Lékařská fakulta, 2009, 1. vydání, ISBN 978-80-210-4223-0.**
- Bártlová, Sylva. Výzkum a ošetřovatelství. Brno: NCONZO, 2009, ISBN 57-851-08.
- Kratochvíl, Jiří. Získávání a zpracování vědeckých informací. Brno: Masarykova univerzita, Lékařská fakulta, 2011, ISBN 978-80-210-5535-3.
- Kuchynková, Zdeňka. Medicína založená na důkazech. In. Kuchynka, Pavel. Oční lékařství: Praha: Grada 2007