

# ZPRACOVÁNÍ STATISTICKÝCH DAT

---

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

# 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 softwarovým vybavením
  - Např. nejrozšířenější je program Excel – data lze vkládat, roztrždit, zpracovat do tabulek a grafů a provést základní výpočty
  - Software Statistika – pořízení je finančně náročné

# 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

## 2. Zásady pro sestavování tabulek

- Pokud máme data roztríděná 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					CELKEM
	velmi dobrý	spíše dobrý	tak napůl	spíše špatný	velmi špatný	
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 = dle 2 a více znaků
  - Korelační tabulky = závislost 2 kvantitativních znaků
- Kontingenční tabulky = závislost 2 kvalitativních znaků

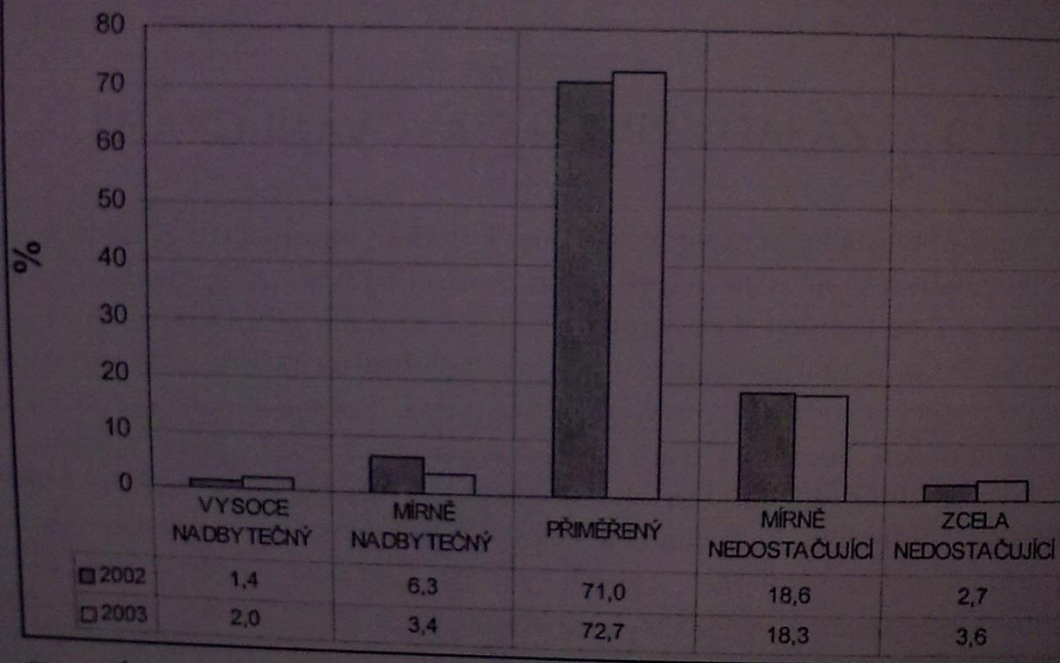
## 3. Sestavování grafů

- Názornější a rychlejší vyjádření dat
- Zpravidla horizontální osa (x) a vertikální osa (y)
- Nejčastěji se používají sloupcové grafy
- Umožňuje 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.*

**Graf č. 1 - Názory občanů na množství lůžkových zařízení v regionu**

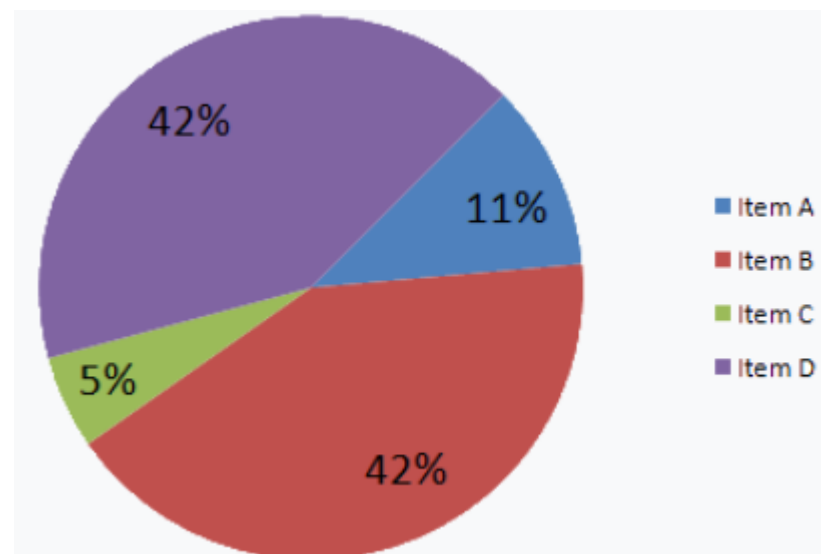


*Pozn. Údaje z reprezentativních výzkumů občanů ČR v roce 2002 (1565 respondentů) a 2003 (1776 respondentů).*



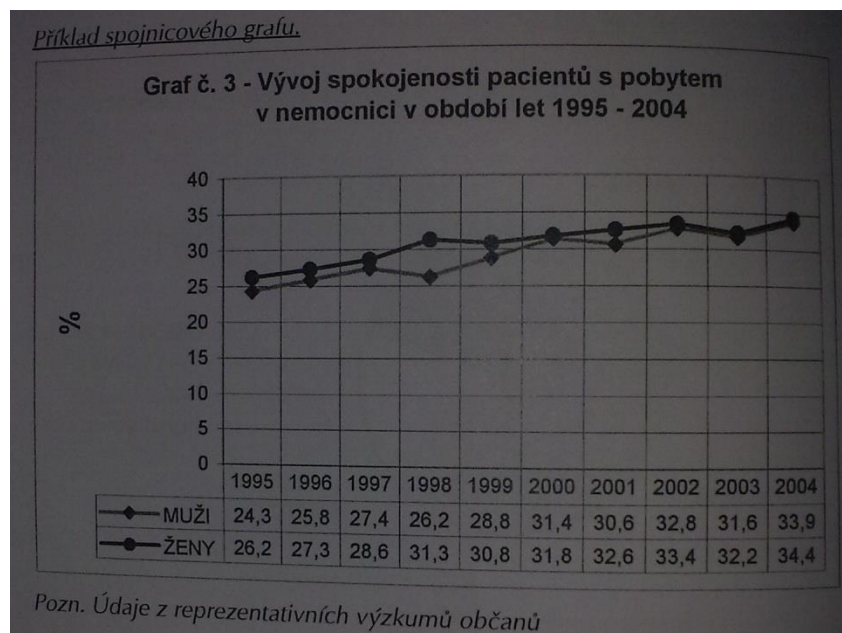
# Graf výsečový (kruhový, sektorový)

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



# Spojnicový graf

- Vyjádření vývoje jevu v čase



# 4 Statistické charakteristiky

- Statistické ukazatele umožňují podat informaci o souboru
- Parametry = pokud se ukazatele vztahují k celému 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 kvalit. 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

# Ukazatel extenzity

- Vyjadřují podíl části k celku
- Počet nositelů znaku násobíme 100 a dělíme počtem jednotek souboru
- Pokud vyjadřujeme v procentech volíme  $k = 2$ , v promilích  $k = 3$
- Znázorňujeme kruhovými grafy

**Podíl počtu chlapců v celkovém počtu narozených dětí,  
který činí  $49\,405/96\,097 = 0,51$**

# Ukazatel intenzity

- Vyjadřuje frekvenci četnosti znaku v souboru
- Dělíme je na:
  - Hrubé = počet na 1000 exponovaných případů
  - Specifické (čisté) = počet na 100 000
- Patří sem např. porodnost, úmrtnost, počet lůžek na 1 lékaře apod.

# Indexní čísla

- Indexy vyjadřujeme změny, které nastaly ve vývoji ve sledovaném čase
- Jednotkami indexu mohou být relativní i absolutní čísla
- Mohou být vyjádřena dvojím způsobem:
  - Indexy s pevným základem (bazické) = první hodnotu v čase stanovíme ze 100%, ostatní adekvátně přirovnáme
  - Indexy s pohyblivým základem (řetězové) = vyjadřují tempo přírůstku v čase, údaj dělíme hodnotou předcházející a násobíme 100

# Bazické vs. řetězové indexy - příklad

$7,7/23,7*100\% = 32,5\%$ , řetězový index v témže roce jako  $7,7/10,8*100\% = 71,3\%$ . V roce 1995 byla kojenecká úmrtnost tedy o 67,5 % nižší než v roce 1965 a o 28,7 % nižší než v roce 1990.

**Tabulka:** Kojenecká úmrtnost na 1000 živě narozených dětí  
v ČR v letech 1965-2000

rok	1965	1970	1975	1980	1985	1990	1995	2000
úmrtnost (na 1000)	23,7	20,2	19,4	16,9	12,5	10,8	7,7	4,1
bazický index (%)	100,0	85,2	81,9	71,3	52,7	45,6	32,5	17,3
řetězový index (%)	--	85,2	96,0	87,1	74,0	86,4	71,3	53,2



## B. Ukazatele polohy

- Sledujeme, zda a jak se náhodné veličiny kupí kolem střední hodnoty
- Střední hodnota = číslo, které v určitém smyslu zastupuje jednotlivé hodnoty zkoumaného souboru, umožňuje reálné srovnání souborů
- SH dělíme na:
  - Průměry = stanoveny výpočtem, bereme v úvahu všechny hodnoty v souboru
  - Ostatní SH = neparametrické, stanoveny dle hodnoty některé statistické proměnné

# Aritmetický průměr

- Charakterizuje normální rozložení
- = úhrn hodnot kvantitativního znaku dělený rozsahem souboru
- $x_i$  = jednotlivé hodnoty
- = aritmetický průměr prostý
- Používáme, když se nám znaky často opakují

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

# Vážený aritmetický průměr

- Pokud je soubor větší a hodnoty se často opakují

$f_i$  = počet opakování znaku

$$\bar{x} = \frac{\sum x_i \cdot f_i}{\sum f_i}$$

- V případě velkého souboru vytvoříme intervalové rozdělení četností, vypočítáme AP pro jednotlivé intervaly a pak vypočítáme AP pomocí vzorce pro VAP

- 

- Třída A — 62, 67, 71, 74, 76, 77, 78, 79, 79, 80, 80, 81, 81, 82, 83, 84, 86, 89, 93, 98

- Třída B — 80, 81, 82, 83, 84, 85, 86, 87, 87, 88, 88, 89, 89, 90, 90, 90, 90, 91, 91, 91, 92, 92, 93, 93, 94, 95, 96, 97, 98, 99, 100

$$\bar{x} = \frac{4480}{52} = 86,15385$$

$$\bar{x} = \frac{20 \cdot 80 + 32 \cdot 90}{20 + 32} = 86,15385$$

# Nevhodné použití AP

- 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
  - Použijeme pokud chybí určení dolní a horní hranice intervalu

# 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

- V případě intervalového rozdělení viz příklad
- Zkoumáme intenzitu jevu
- Soubor může mít i více vrcholů

# 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

= daná rozdílem mezi největší a nejmenší hodnotou souboru

- Bere v úvahu jen extrémní hodnoty
- Pouze orientační

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

## 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

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

# 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 vymežit 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 souboru se doporučuje variabilitu vyjadřovat pomocí kvantilů (decil, percentil)

# 5 Pravděpodobnosti

= teorie pravděpodobnosti studuje a formuluje zákonitosti náhody a jejího využívání

- Náhoda = souhrn drobný 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)

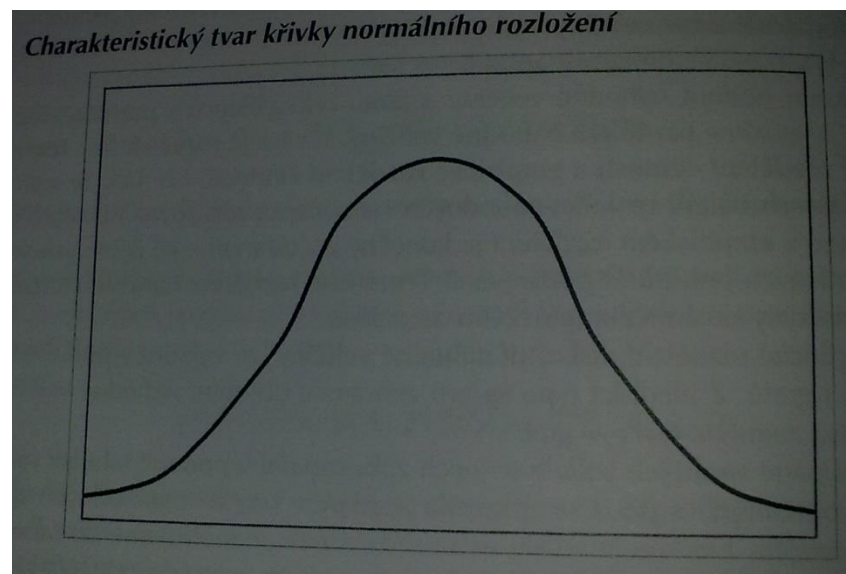
# Rozdělení náhodné veličiny

= 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  $\mu$ , směrodatná odchylka  $\sigma$  a rozptyl  $\sigma^2$
- Výběrový soubor charakterizuje výběrová střední hodnota  $\bar{x}$ , výběrová směrodatná odchylka  $s$  a výběrový rozptyl  $s^2$

# 7 Testování hypotéz

- Pomocí testovací statistiky určíme, zda se dva ukazatele od sebe liší reálně nebo náhodně
- V praxi potřebujeme odlišit, zda dva náhodné výběry pochází ze stejného základního souboru
- K tomu slouží statistické testy
- Statistická hypotéza

# Statistická hypotéza

= tvrzení o základním souboru (chybné nebo správné)

- Rozdělení hypotéz:

- Parametrické – odlišují se pouze hodnotami parametrů
- Neparametrické – odlišuje se ještě i tvarem rozložení funkce

- Další rozdělení hypotéz:

- Jednoduché
- Složené – z více jednoduchých

# Hypotézou ověřujeme především tyto 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

# Kroky při stavení hypotézy

- Nulová hypotéza  $H_0$
- Alternativní hypotéza  $H_1$  = s jakou hypotézou počítáme, pokud  $H_0$  neplatí
- Statistický test významnosti = rozhodovací postup mezi nulovou a alternativní hypotézou založený na datech

# Testování hypotézy – možné situace

SKUTEČNOST	TEST	
	$H_0$ je správná +	$H_0$ je nesprávná -
$H_0$ je správná +	test odhaduje skutečnost správně +	test odhaduje skutečnost nesprávně - (chyba 1. druhu $\alpha$ )
$H_0$ je nesprávná -	test odhaduje skutečnost nesprávně - (chyba 2. druhu $\beta$ )	test odhaduje skutečnost správně +

Hladina významnosti = chyba 1. druhu

Síla, mohutnost test = chyba 2. druhu

$\alpha = 0.05$ , pro 5% je úsudek nesprávný



# Volba hladiny významnosti

Pravděpodobnost chyby	$>0,05$	$\leq 0,05$	$\leq 0,01$	$\leq 0,001$
Slovní vyjádření	nesignifikantní	signifikantní	velmi signifikantní	velmi vysoce signifikantní
Písmenová symbolika	n.s.	s.	v.s.	v.v.s.
Grafická symbolika		*	**	***

# Obecný postup při testování hypotéz

1. Formulace nulové a alternativní 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

- Alespoň přibližně normální rozložení
- Porovnáváme např. průměr nebo rozptyly souborů
- Používáme F-test nebo T-test

# Fischerův test

- = parametrický test významnosti, kde testujeme hypotézy o rozptylu
- Určujeme signifikantnost odlišnosti na určité hladině pravděpodobnosti

# 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 NPT nepotřebujeme znát parametry souboru, bývají jednodušší, ale poskytují méně informací

# Neparametrické testy - rozdělení

1. Kolmogorovův-Smirnovův test pro jeden výběr – hodnocení kumulativního rozdílu četností, spojitě rozdělení souboru
2. Kolmogorovův-Smirnovův test pro dva závislé 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

# Testy chí-kvadrát

- Přejechod mezi parametrickými a neparametrickými testy
- Pro hodnocení kvalitativních znaků
- Podrobněji později (závislost kvalitativních znaků)



# 8 Měření závislostí

- Závislost kvantitativních znaků
- Závislost kvalitativních znaků

# Závislost kvantitativních znaků

- Např. věk, tělesná váha, výška, výsledky laboratorních vyšetření
- Stanovujeme druh a těsnost sledovaných závislostí
- Závislost většinou vyjadřujeme pomocí křivky (regresní křivka)
- Závislost může být lineární, logaritmická, parabolická, exponenciální
- Těsnost (korelace) vyjadřujeme obvykle korelačním koeficientem (0 až 1)

# Závislost kvalitativních znaků

- Pohlaví, krevní skupina, rodinný stav atd. mají kvalitativní charakter
- Chí-kvadrát test ( $\chi^2$ ) je zaměřen na ověřování shody rozdělení, porovnáváme empirické a teoretické hodnoty
- Východiskem jsou tzv. kontingenční tabulky

# Kontingenční tabulka - Test chí-kvadrát

$$\chi^2 = \sum_{i=1}^k \frac{(X_i - Np_i)^2}{Np_i}$$

Chceme ověřit, zda je hrací kostka pravidelná. Hodíme kostkou 60krát a budeme sledovat četnosti jednotlivých hodnot. Při pravidelné kostce je pravděpodobnost každého čísla 1/6, tedy všechny hodnoty od 1 do 6 mají očekávanou četnost 10.

Následující tabulka uvádí skutečné a očekávané četnosti jednotlivých hodnot.

Hodnota	1	2	3	4	5	6	$\sum_{i=1}^k$
Skutečné četnosti	5	14	4	10	14	13	60
Očekávané četnosti	10	10	10	10	10	10	60
$\chi^2$	2.5	1.6	3.6	0	1.6	0.9	10.2

Dosadíme-li do vzorce, je výsledná hodnota  $\chi^2$  je 10,2 a kritická hodnota chí-kvadrát s 5 stupni volnosti na nejpoužívanější 5% hladině významnosti je 11,07. Nelze tedy prohlásit, že by předpoklad byl porušen, kostka může být pravidelná. Je však možné, že provedení většího počtu pokusů by již mohlo odchytku od pravidelnosti prokázat.

# 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