

# Popisná statistika

---

- úvod
  - rozdělení hodnot
  - míry centrální tendence
  - míry variability
  - míry šikmosti a špičatosti
  - grafy
-

# Úvod

---

- užívá se k popisu základních vlastností dat
  - poskytuje jednoduché shrnutí hodnot proměnných ve výběrovém souboru
  - předchází indukční statistiku (která odvozuje zjištění ze vzorku na populaci)
-

# Úvod

---

- techniky deskriptivní statistiky pomáhají redukovat větší množství dat do zvládnutelné podoby – grafické, tabulkové, do jednoho ukazatele
  - touto redukcí např. údajů o rychlosti čtení u 200 žáků na jeden ukazatel, např. na hodnotu průměru samozřejmě část informací ztratíme
-

# Úvod

---

- pro každou proměnnou obvykle popisujeme 3 charakteristiky
  - rozdělení hodnot (i graficky), středovou hodnotu a míru rozptýlení hodnot kolem tohoto středu
-

# Rozdělení hodnot

---

- rozdělení (distribuce) hodnot - souhrn četností jednotlivých kategorií nebo intervalů hodnot proměnné
  - jednou z možností, jak zobrazit rozložení hodnot proměnné je **tabulka četností** – seznam kategorií proměnné a u nich počet osob, které do každé kategorie spadají
-

# Rozdělení hodnot

---

příklad  
**tabulky  
četností**

	počet osob	%
Sangvinik	118	28
Flegmatik	86	20
Melancholik	89	21
Cholerik	130	31
<i>celkem</i>	423	100

---

# Rozdělení hodnot

---

- vždy je třeba uvést celkový počet osob (N)
  - relativní četnosti mohou být uvedeny buď jako procenta (8%) nebo podíly (0.08)
  - může jít rovněž o poměr (*ratio*) dvou kategorií (např. poměr dívek a chlapců s ADHD 1:4 (nebo 0,25))
-

# Rozdělení hodnot

---

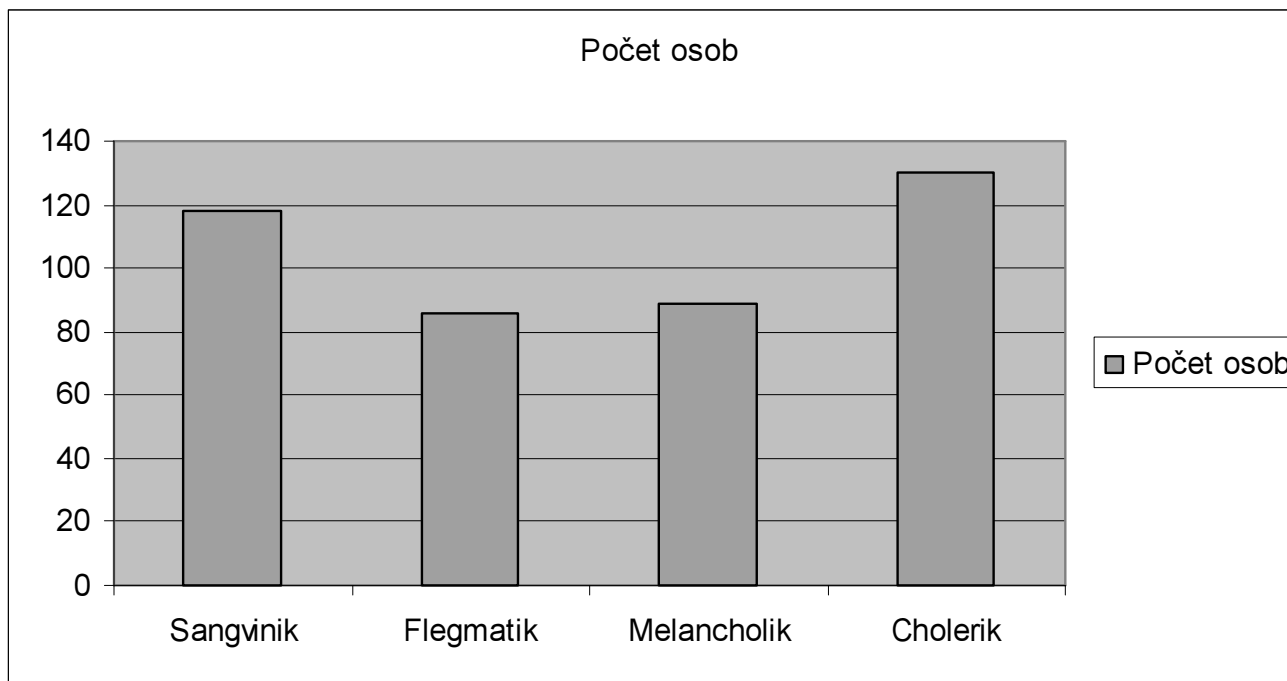
- jako míra (*rate*) se označuje počet výskytů nějakého jevu dělený počtem možných výskytů v nějakém čase
  - např. míra úmrtnosti = počet mrtvých za rok / počet obyvatel x 1000
  - získáme hrubou míru úmrtnosti na 1000 obyvatel
-



# Rozdělení hodnot

---

- stejná data je možno zobrazit i **graficky** (v příkladu sloupcový diagram – barchart)



# Rozdělení hodnot

---

- pokud proměnná nabývá mnoha hodnot, je vhodnější je **sloučit do kategorií (intervalů)**
  - počet intervalů by měl být přiměřený počtu hodnot
  - někdy se používá tzv. Sturgesovo pravidlo  $k = 1 + 3,3 \log_{10}(n)$
  - podle něj by pro 200 hodnot byl vhodný počet intervalů 9
  - záleží i na počtu osob – pro menší výběry raději histogram s menším počtem sloupců
-

# Rozdělení hodnot

<b>IQ</b>	počet	%	<i>kumul.%</i>
<b>méně než 86</b>	11	10	10
<b>86 – 100</b>	36	34	44
<b>101 – 115</b>	34	32	76
<b>116 – 130</b>	20	19	95
<b>131 a více</b>	5	5	100
<i>celkem</i>	<i>106</i>	<i>100</i>	

# Míry centrální tendence

---

- míry centrální tendence (středu, polohy) jsou výsledkem snahy najít typickou hodnotu pro daný znak
  - nejčastěji používané modus, medián, aritmetický průměr, méně často např. harmonický a geometrický průměr
-

# Míry centrální tendence

---

- **modus** – nejčastěji se vyskytující hodnota (např. u příkladu s temperamentem to byl *choleric*)
  - jediná použitelná charakteristika polohy pro nominální data; u pořadových a kardinálních jsou většinou více typickými charakteristikami medián nebo průměr
-

# Míry centrální tendence

---

- pokud je v rozdělení více modů, jde o rozdělení vícevrcholové (obvykle bimodální) – může odhalit nehomogenitu výběru
  - např. rozdělení hodnot tělesné výšky může mít dva mody – pro muže a pro ženy
-

# Míry centrální tendence

---

- modus není užitečnou statistikou pro zobecňování ze vzorku na populaci – dá se očekávat, že různé vzorky z téže populace budou mít různé mody
-

# Míry centrální tendence

---

- **medián** - prostřední hodnota v řadě hodnot uspořádaných podle velikosti (50% percentil)
  - je jen pro data, která je možno podle velikosti uspořádat, tj. pořadová a kardinální
  - dělí soubor na dvě poloviny (pro sudý počet hodnot je medián průměrem dvou prostředních pozorování)
-



# Míry centrální tendence

---

- vzorec pro výběr s lichým počtem hodnot:

$$Me = x_{(n+1)/2}$$

- vzorec pro výběr se sudým počtem hodnot:

$$Me = (x_{n/2} + x_{n/2+1})/2$$

---

# Míry centrální tendence

---

- používá se především, pokud chceme eliminovat vliv extrémních hodnot
  - příklad – průměrný plat 20 tisíc může u 10 osob znamenat, že 9 z nich má 10 tisíc a jeden 110 tisíc; použijeme-li medián – 10 tisíc, získáme více typickou hodnotu
  - můžeme ho vyčíst z tabulky četností, pokud jsou uvedeny kumulativní četnosti
-

# Míry centrální tendence

---

- **aritmetický průměr** – součet všech hodnot znaku dělený jejich počtem
  - jen pro proměnné, u nichž je možno hodnoty smysluplně dělit (kardinální)
  - vzorec:
    - $\mu = \Sigma_i X_i / N$  (pro populaci)
    - $m = \Sigma_i x_i / n$  (pro výběr)
  - součet odchylek od průměru = 0
-

# Míry centrální tendence

---

- průměr zahrnuje každou hodnotu znaku – což je jak výhoda, tak nevýhoda (citlivý na extrémní hodnoty)
  - to je možno vyřešit použitím tzv. seříznutého průměru (*trimmed mean*), který se počítá tak, že se vynechá určité % hodnot z obou stran rozdělení, např. 5% nejnižších a 5% nejvyšších
-

# Míry centrální tendence

---

- průměr špatně reprezentuje  
nehomogenní skupiny
  - příklad – 30 osob v parku, průměrný věk 12.5 roku, průměrná výška 130 cm: nemusí jít o školní děti, ale o 15 matek se 4-letými dětmi
-

# Míry centrální tendence

---

- pro znaky s normálním rozdělením hodnot je průměr **nejúčinnější** charakteristikou (tj. nejvíce stabilní pro různé výběrové soubory) – dá se nejlépe použít pro odhad parametru populace z charakteristik výběru
  - je nejčastěji užívanou mírou polohy
-

# Míry centrální tendence

---

- kterou statistiku použít a uvádět?
  - průměr** – pokud může být spočítán a pokud není rozdělení příliš šikmé
  - modus** – pokud je rozdělení multimodální (neexistuje jediná typická hodnota)
  - medián** – pokud je rozdělení šikmé a unimodální, pokud obsahuje odlehlé hodnoty
-

# Míry centrální tendence

---

- **příklad** – spočítejte modus, medián a aritmetický průměr následujícího rozdělení hodnot

18 5 128 2 14 87 50 87 70

---



# Míry variability

---

- míry variability popisují kolísání  
v rozdělení hodnot
  - označují se i jako **míry**  
**rozptýlenosti**
  - užívá se rozpětí, mezikvartilové  
rozpětí, rozptyl, směrodatná  
odchylka, variační koeficient
-

# Míry variability

---

- **rozpětí** (variační šíře, variační rozpětí) – rozdíl mezi nejvyšší a nejnižší hodnotou
  - značně ovlivněno extrémními hodnotami, není dobrým odhadem parametru populace
  - používá se zřídka
-

# Míry variability

---

- **mezikvartilové rozpětí**  
(interkvartilová odchylka) – rozdíl mezi hodnotou horního kvartilu a dolního kvartilu
  - **kvartily** – dělí soubor na 4 stejné části; horní kvartil odděluje 25% nejvyšších hodnot, dolní 25% nejnižších
-

# Míry variability

---

- mezikvartilové rozpětí udává rozpětí pro středních 50% hodnot (=délka obdélníku v krabicovém diagramu)
  - není (podobně jako medián) citlivé na extrémní hodnoty
-

# Míry variability

---

□ **rozptyl** (střední kvadratická odchylka průměru) - ukazuje, jak jsou hodnoty rozptýleny kolem průměru

□ v populaci

$$\sigma^2 = (1/(N)) \sum_{i=1}^n (x_i - \mu)^2$$

□ výběr

$$s^2 = (1/(n-1)) \sum_{i=1}^n (x_i - m)^2$$

---

# Míry variability

---

- více než rozptyl se používá jeho odmocnina – **směrodatná odchylka průměru** (je ve stejném měřítku jako původní hodnoty)
  - oba ukazatele slouží jako vhodné doplnění průměru – získáme představu o jeho věrohodnosti, jak dobře reprezentuje všechny hodnoty
-

# Míry variability

---

- příklad – porovnejte variabilitu u těchto dvou rozložení hodnot (jde např. o počet správně vyřešených úloh v didaktickém testu ve 2 třídách)

a) 4 5 4 3 5 5 3 4 3

b) 8 2 12 1 4 3 5 0 1

---

# Míry variability

---

□ řešení příkladu

□  $m_a = 4, s_a = 0.87$

□  $m_b = 4, s_b = 3.87$

□ u prvního rozdělení je průměr lepší reprezentací hodnot; u druhého jsou hodnoty kolem průměru hodně rozptýleny

---



# Míry variability

---

- **variační koeficient** – pro porovnání míry variability u různých souborů
  - pokud se u různých souborů měřené hodnoty výrazně liší svou úrovní anebo jsou dokonce v různých jednotkách, nelze podle rozptylu či standardní odchylky porovnávat přímo, který ze souborů má větší variabilitu - je třeba srovnávat relativní variabilitu
-

# Míry variability

---

- jde o podíl směrodatné odchylky a průměru
  - většinou se udává v procentech
  - $VK = ( s / m ) * 100\%$
-

# Míry variability

---

- příklad – porovnejte variabilitu průměrného platu v ČR (v korunách) a v GB (v librách) (*fiktivní data*)
  - $m_{GB} = 1000$  liber,  $s_{GB} = 600$
  - $m_{CZ} = 10\ 000$  Kč,  $s_{CZ} = 3000$
-

# Míry variability

---

- řešení příkladu – větší variabilita je v britských platech (60%) než v českých (30%)
-

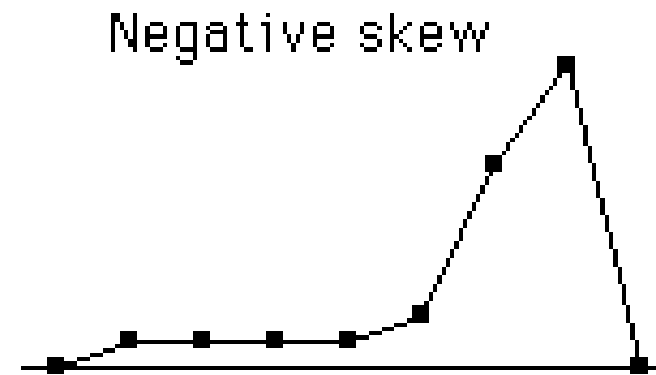
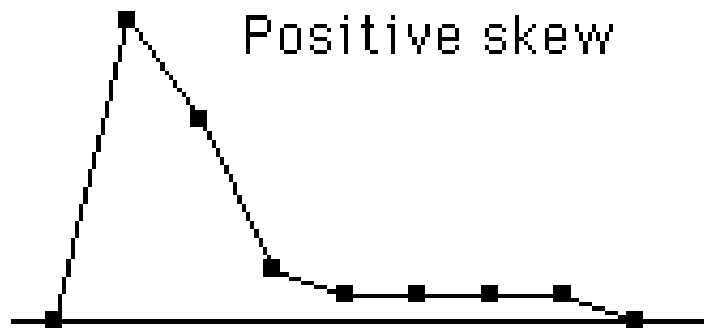
# Míry šikmosti a špičatosti

---

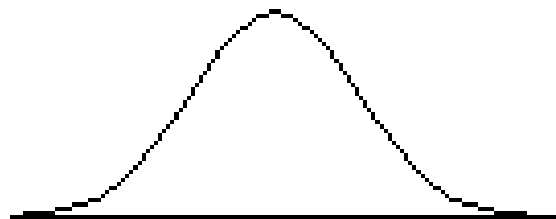
- hodnotíme, jak se rozdělení dat podobá normálnímu (Gaussovu) rozdělení
  - **šikmost** (skewness) měří nesymetrii vzhledem k podélné ose
    - pro symetrické rozdělení se koeficient šikmosti = 0
    - pokud je  $> 0$ , je rozdělení s prodlouženým pravým koncem (doprava, kladně šikmé)
    - pokud je  $< 0$ , je rozdělení s prodlouženým levým koncem (doleva, záporně šikmé)
-

# Míry šikmosti a špičatosti

---



Symmetric distribution  
(No skew)



# Míry šikmosti a špičatosti

---

- i porovnáním hodnoty průměru a mediánu získáme představu o šikmosti rozdělení hodnot
    - pokud je průměr větší než medián – kladně zešikmeno
    - průměr menší než medián – záporně zešikmeno
    - průměr = medián – symetrické rozdělení
  - Pearsonův vzorec pro koeficient šikmosti na základě srovnání hodnot průměru a mediánu
    - $SK = 3 * (m - Me) / s$
-

# Míry šikmosti a špičatosti

---

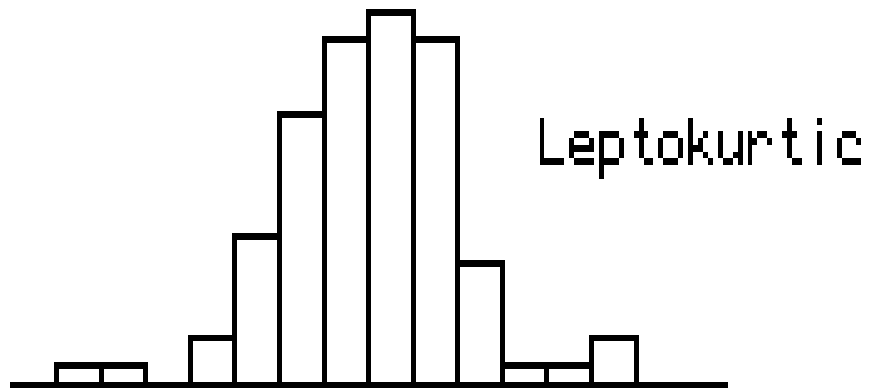
- koeficient **špičatosti** (kurtosis)
    - pro normální rozdělení = 0
    - pokud je  $> 0$ , je rozdělení tzv. leptokurtické (více špičaté než normální)
    - pokud je  $< 0$ , je rozdělení tzv. platykurtické (plošší než normální)
-



# Míry šikmosti a špičatosti

---

Kurtosis = 1.25



Kurtosis = -1.23



# Grafy

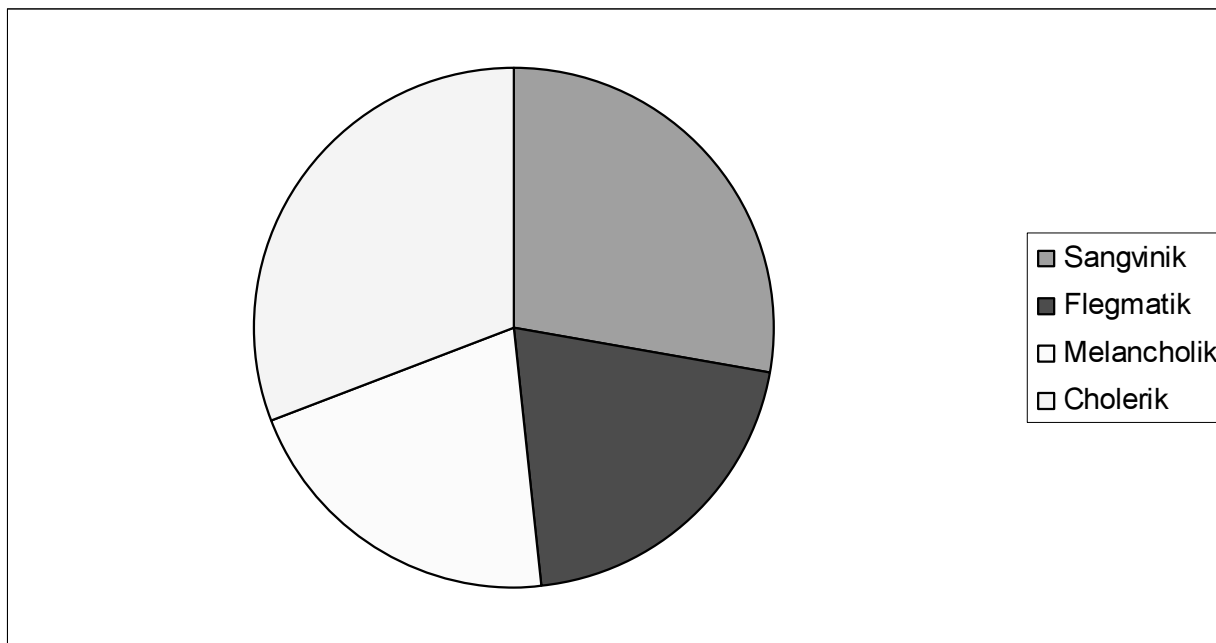
---

- pouze základní typy
  - pro kategoriální data - sloupcový diagram, výsečový graf
  - pro spojitá data – histogram, frekvenční polygon, krabicový diagram, stromový diagram
  - grafy je možno znázornit v kategorizované formě – pro jednotlivé kategorie další proměnné (např. pro muže a ženy)
-

# Grafy

---

- **výsečový graf** (koláčový diagram, pie chart) – užívá se více v populárních publikacích než v odborných



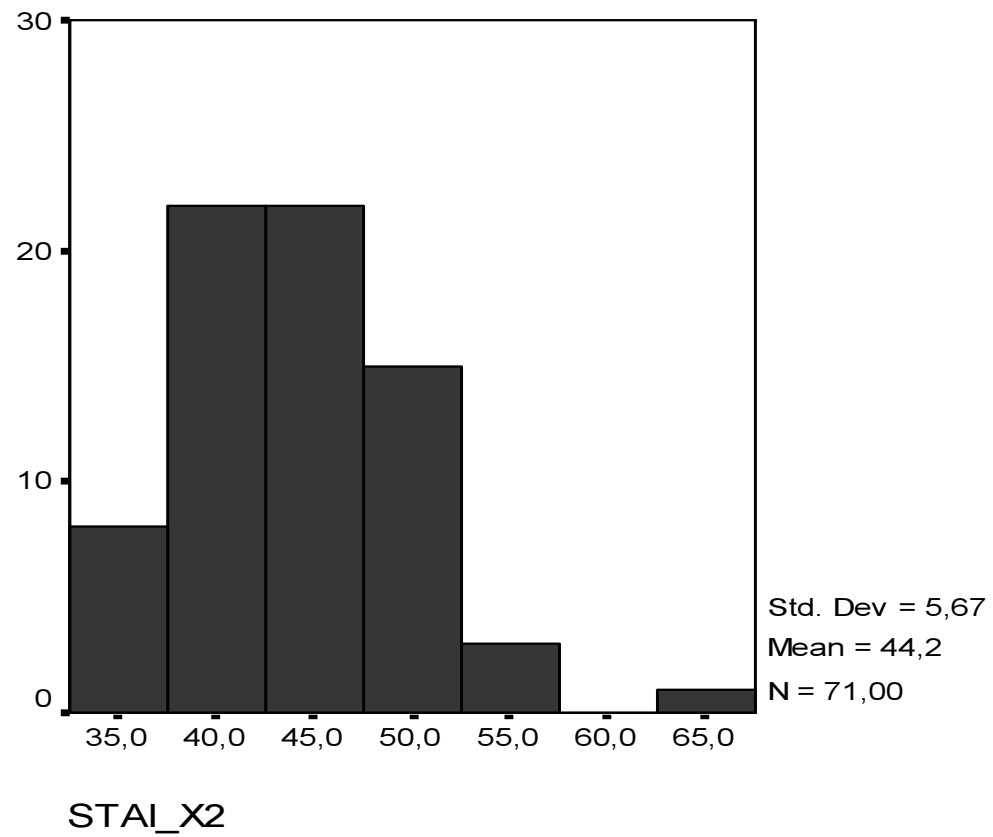
# Grafy

---

- **histogram** – podobný sloupcovému diagramu, ale je pro spojitá data
  - jednotlivé sloupce reprezentují nikoliv jednotlivé kategorie, ale intervaly hodnot
  - tvar histogramu závisí do jisté míry na šířce intervalů
-

# Grafy

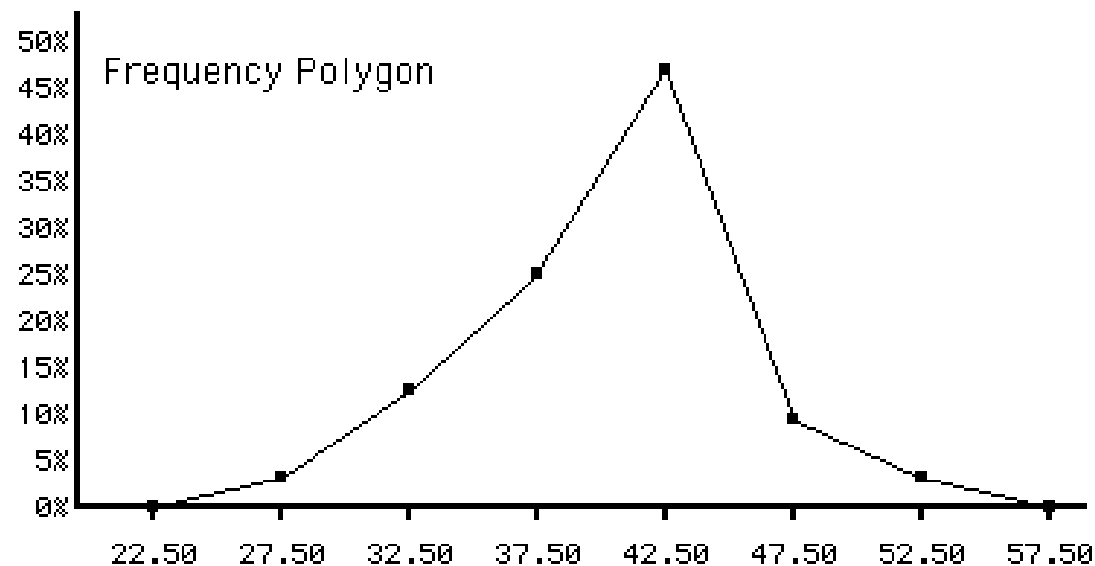
---



# Grafy

---

- **frekvenční polygon** – konstruován podobně jako histogram, jen místo sloupců jsou tečky spojené čarou



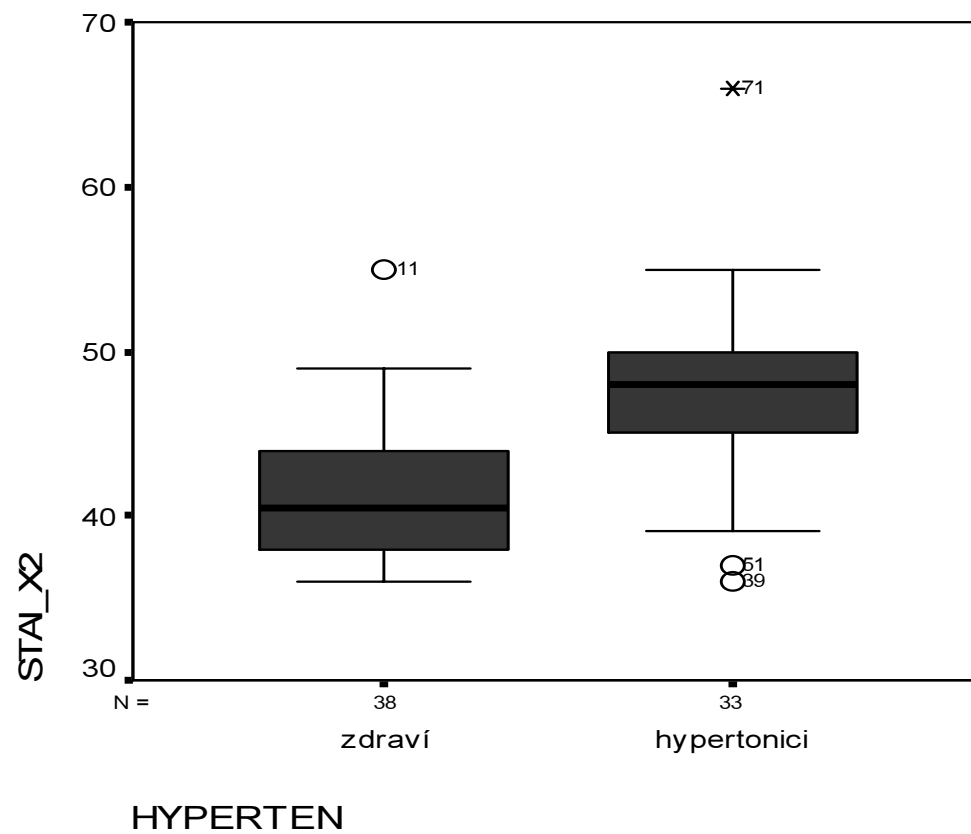
# Grafy

---

- **krabicový diagram** (boxplot, vousatá krabíčka) – poskytuje bohaté zobrazení důležitých aspektů rozdělení hodnot
  - délka krabice odpovídá interkvartilové odchylce; uvnitř krabice je vyznačen medián
  - „vousy“ nebo „anténami“ je ohraničeno rozmezí hodnot bez odlehlých hodnot (outliers) a extrémních hodnot (více než 3x délky krabice od jejího konce)
-

# Grafy

---





# Grafy

---

- **stromkový diagram** (stem-and-leaf plot; stonek a list) – podobný histogramu (naležato), ale obsahuje informace o každém případě
  - konstrukce diagramu – hodnoty jsou rozděleny např. na desítky (stonek) a jednotky (list)
  - např. hodnota  $85 = 8 \times 10 + 5 \times 1$
  - pokud je hodnot pro některé desítky více, rozdělí se na další listy
-

# Grafy

---

Frequency	Stem & Leaf
3,00	1 . 468
7,00	2 . 0225588
9,00	3 . 011234449
10,00	4 . 3455567799
3,00	5 . 344
7,00	6 . 0111389
4,00	7 . 1234
2,00	8 . 34
1,00	9 . 1

Stem width: 10,00  
Each leaf: 1 case(s)

---

# Grafy

---

Frequency	Stem & Leaf
,00	3 .
6,00	3 . 667777
8,00	3 . 88889999
9,00	4 . 000001111
5,00	4 . 22333
5,00	4 . 44455
3,00	4 . 667
1,00	4 . 9
1,00	Extremes (>=55)

Stem width: 10  
Each leaf: 1 case(s)

---

# Grafy

---

- čeho si v grafu všímat?
    - tvaru rozdělení
    - míst s největší četností hodnot (zhuštění, shluky)
    - mezer
    - odlehlých hodnot
-

# Kontrolní otázky

---

- rozdíly mezi absolutními a relativními četnostmi, poměrem a mírou; kumulativní četnosti
  - 3 základní míry centrální tendence (+ u jakých dat použijeme průměr, modus či medián)
  - základní míry variability, výpočet rozptylu
  - typy grafů
-

# Literatura

---

## □ **Hendl – kapitola 3**

### □ doplňující (v IS):

- Wainer, H., & Velleman, PF (2001).  
Statistical graphics: Mapping the  
pathways of science. *Annual  
Review of Psychology*, 52, 305-  
335.
-