Popisná /deskriptivní/ statistika

- úvod
- rozdělení hodnot
- míry polohy
- míry variability
- grafy

Úvod

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

Úvod

- techniky deskriptivní statistiky pomáhají redukovat větší množství dat do zvládnutelné podoby
- 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
celkem	423	100

Rozdělení hodnot

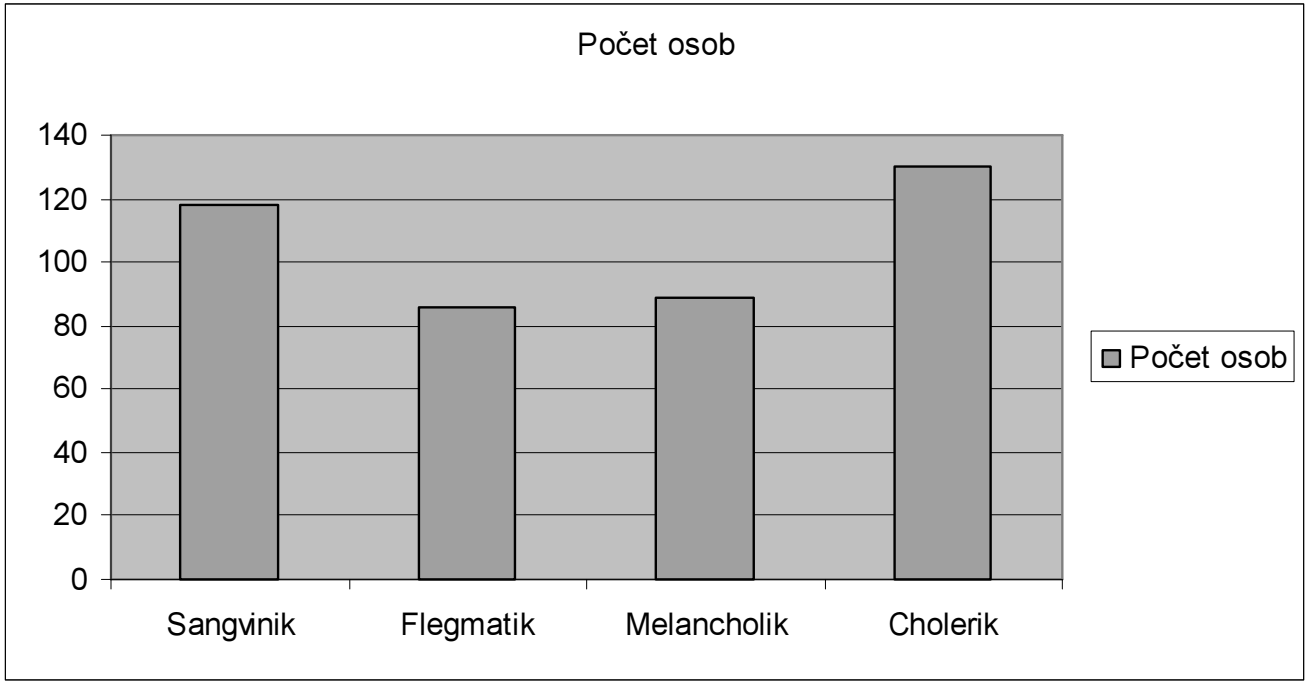
- vždy je třeba uvést celkový počet osob (N)
- relativní četnosti mohou být uvedeny buď jako procenta (8%) nebo podíly (0.08)
- může jít rovněž o poměr (*ratio*) dvou kategorií (např. poměr dívek a chlapců s ADHD 1:4 (nebo 0,25))

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

Rozdělení hodnot

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

Míry polohy

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

Míry polohy

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

Míry polohy

- 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 polohy

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

Míry polohy

- **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 polohy

- 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 polohy

- **aritmetický průměr** – součet všech hodnot znaku dělený jejich počtem
- jen pro proměnné, u nichž je možno hodnoty smysluplně dělit (kardinální)
- vzorec: $\mu = \Sigma \mathbf{X/N}$ (pro populaci)
- nebo $\mathbf{m} = \Sigma \mathbf{x/N}$ (pro výběr)

Míry polohy

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

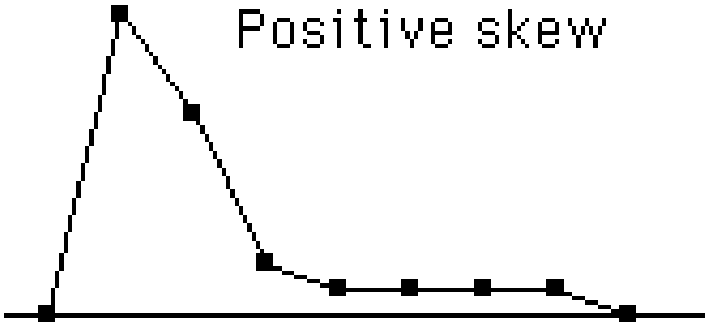
Míry polohy

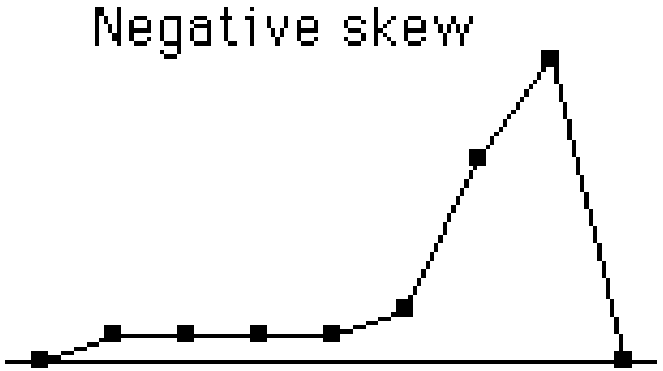
- průměr špatně reprezentuje nehomogenní skupiny
- příklad – 30 osob v parku, průměrný věk 12.5 roku, průměrná výška 130 cm: nemusí jít o školní děti, ale o 15 matek se 4-letými dětmi

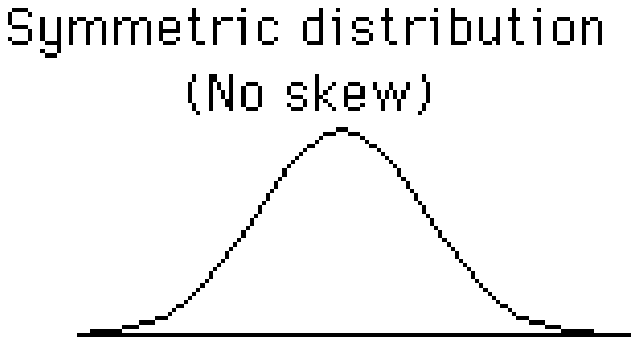
Míry polohy

- porovnáním hodnoty průměru a mediánu získáme představu o **šikmosti** rozdělení hodnot
- pokud je průměr větší než medián – kladně (doprava) zešikmeno
- průměr menší než medián – záporně (doleva) zešikmeno
- průměr = medián – symetrické rozdělení

Míry polohy


Positive skew


Negative skew


Symmetric distribution
(No skew)

Míry polohy

- 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 polohy

- kromě aritmetického průměru se v psychologii někdy používá i **harmonický průměr** – pro znaky měřené jako podíly, např. rychlost v km/h, podíly osob atd.)
- vzorec

$$m_h = \left(n \Big/ \sum_{i=1}^{n} (1/ x_i) \right)$$

Míry polohy

- kterou statistiku 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

Míry polohy

- **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
- 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

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**
- oba ukazatele slouží jako vhodné doplnění průměru – získáme představu o jeho věrohodnosti, tj. jak dobře reprezentuje všechny hodnoty

Míry variability

- příklad – porovnejte variabilitu u těchto dvou rozložení hodnot (jde o počet správně vyřešených úloh v didaktickém testu u výběru osob ze dvou tříd ZŠ)

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
- c = (s / m) .100%

Míry variability

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

Míry variability

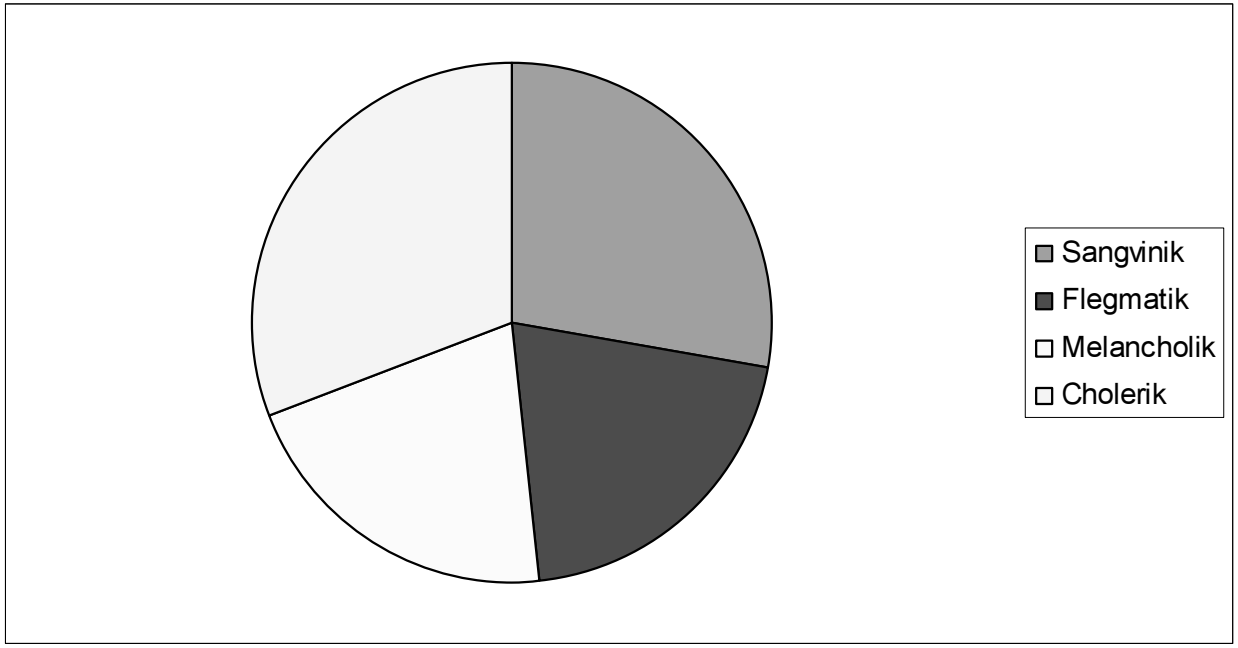
- řešení příkladu – větší variabilita je v britských platech (60%) než v českých (30%)

Grafy

- pouze základní typy
- pro kategoriální data - sloupcový diagram, výsečový graf
- pro intervalová data – histogram, frekvenční polygon, krabicový diagram, stromkový diagram
- grafy je možno znázornit v kategorizované formě – pro jednotlivé kategorie další proměnné (např. pro muže a ženy)
- grafy pro vztah dvou a více proměnných budou probrány později

Výsečový graf

- koláčový diagram, pie chart – užívá se více v populárních publikacích než v odborných

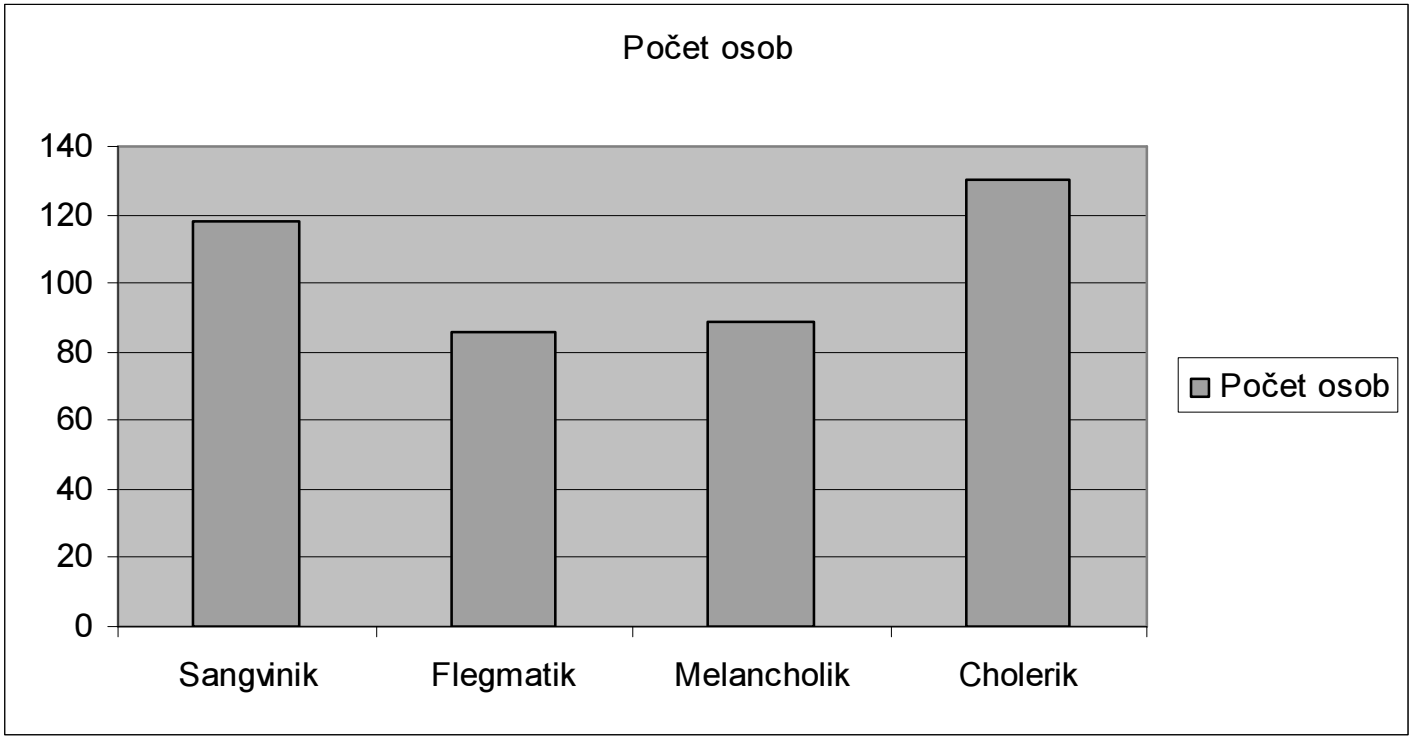


Výsečový graf

- každá výseč by měla být označena % a uveden celkový počet případů
- ideální pro 3-7 kategorií
- **výhody**: srozumitelný
- **nevýhody**: jen pro kategoriální data; neukazuje přesné údaje (pokud nejsou vyznačeny); srovnání více skupin osob problematické

Sloupcový diagram

- bar chart



Sloupcový diagram

- pro kategoriální data, může být orientován horizontálně či vertikálně
- jednotlivé sloupce odděleny mezerou
- **výhody**: srozumitelný, je možno v jednom grafu porovnat četnosti pro více skupin osob

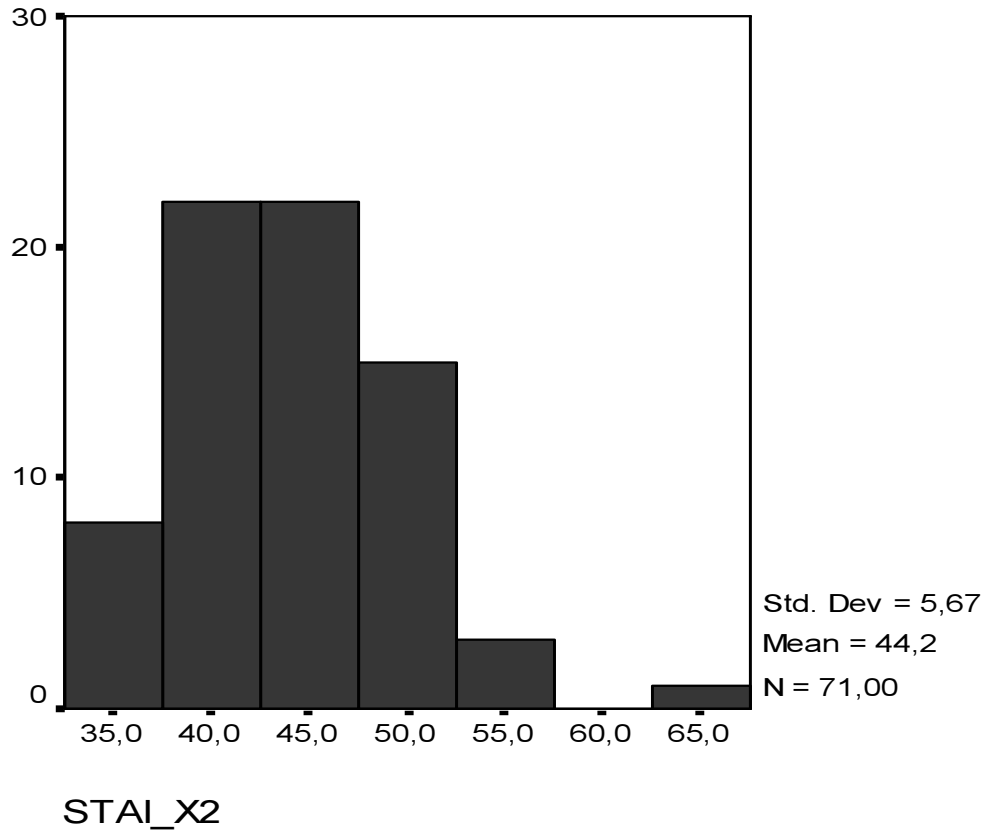
Histogram

- často užívaný
- podobný sloupcovému diagramu, ale je pro intervalová data
- jednotlivé sloupce reprezentují nikoliv jednotlivé kategorie, ale intervaly hodnot (sloupce jsou bez mezer)
- tvar histogramu závisí také na šířce intervalů

Histogram

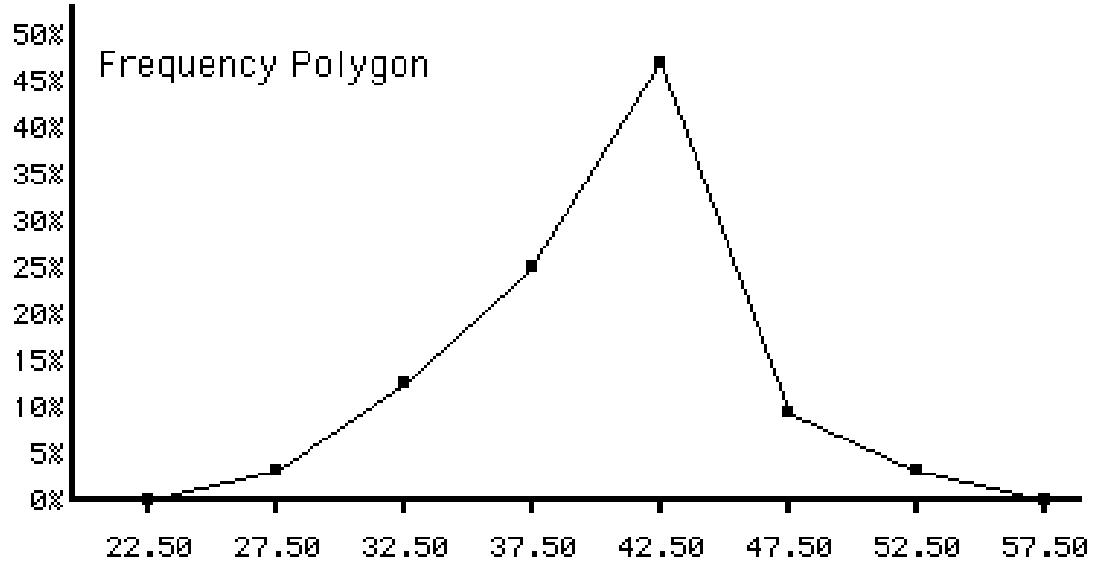
- **výhody**: umožňuje detekovat odlehlá pozorování, srovnání s normálním rozdělením
- **nevýhody**: nezjistíte přesné hodnoty jednotlivých případů, obvykle se nezobrazují data pro více skupin případů

Histogram


30
20
10
0
35,0
40,0
45,0
50,0
55,0
60,0
65,0
Std. Dev = 5,67
Mean = 44,2
N = 71,00
STAI_X2

Grafy

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



Stromkový diagram

- stem-and-leaf plot; stonek a list – podobný histogramu (naležato), ale obsahuje informace o každém případu
- konstrukce diagramu – hodnoty jsou rozděleny např. na desítky (stonek) a jednotky (list)
- např. hodnota 85 = 8x10 + 5x1
- pokud je hodnot pro některé desítky více, rozdělí se na další listy

Stromkový diagram

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)

Stromkový diagram

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)

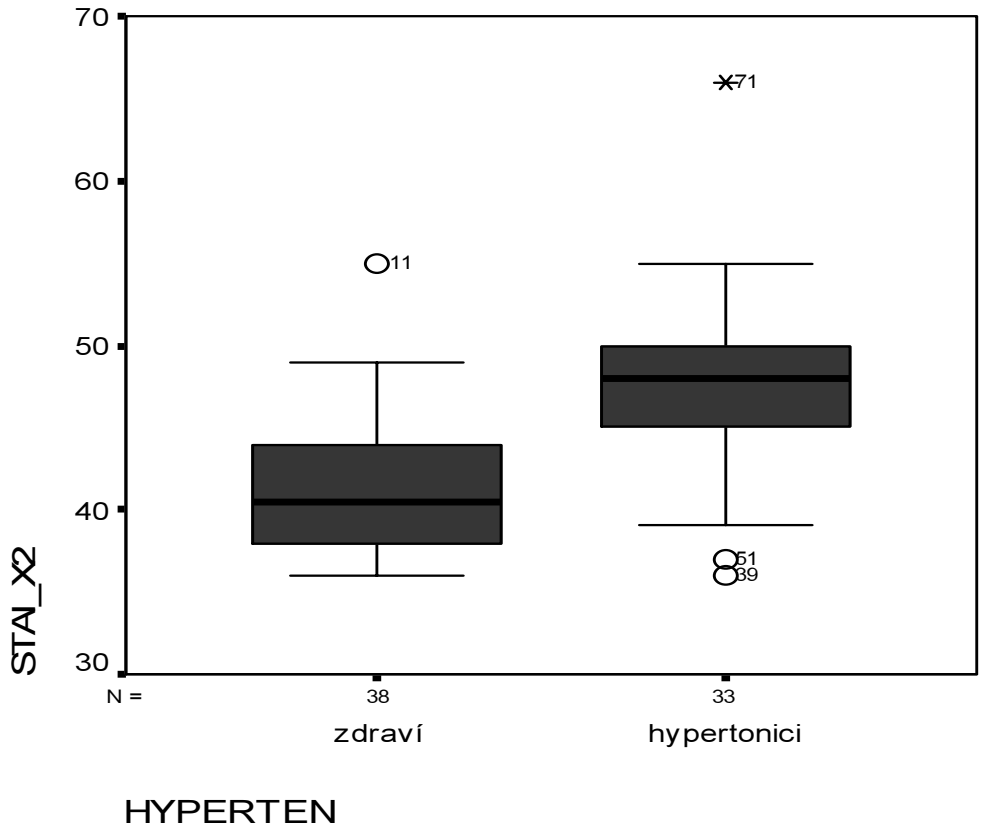
Stromkový diagram

- **výhody**: ukazuje údaje pro každý případ; je možné snadno identifikovat minimum, maximum, shluky případů, odlehlá pozorování; můžeme porovnat dvě skupiny případů zobrazením dvou přilehlých diagramů;
- **nevýhody**: nevypadá zajímavě; vhodnější spíše pro menší datové soubory (N<100);

Krabicový diagram

- boxplot, vousatá krabička
- poskytuje bohaté zobrazení důležitých aspektů rozdělení hodnot
- délka krabice odpovídá interkvartilové odchylce; uvnitř krabice je vyznačen medián
- v některých variantách grafu jde např. o směrodatnou odchylku a průměr
- „vousy“ je ohraničeno rozmezí hodnot

Krabicový diagram



Odlehlá pozorování

- zvlášť jsou u boxplotu vyznačena tzv. **odlehlá pozorování** (outliers – obvykle hodnoty vzdálené více než 1.5 délky krabice od jejího okraje) a extrémní pozorování (obvykle více než 3x délky krabice)
- odlehlá pozorování mohou zkreslit výsledky některých statistik a statistických testů

Odlehlá pozorování

- je proto důležité je v datech hledat; pokud je najdeme, musíme se rozhodnout, zda se jedná o ojedinělý výskyt (který by se v jiném vzorku nevyskytl) nebo výsledek chyby měření; nebo zda je tak reprezentována určitá část populace
- pokud jde o ojedinělý výskyt, je možno je z další analýzy vyloučit
- jinak je nutno se rozhodnout mezi dvěma možnostmi: buď je vyloučit s vědomím, že výsledky budou jejich nepřítomností zkresleny, nebo použít neparametrický test (vhodnější přístup)

Krabicový diagram

- **výhody**: užitečný pro detekci odlehlých pozorování, šikmosti rozdělení; vhodný pro porovnání více skupin případů
- **nevýhody**: složitější

Grafy – obecná doporučení

- každý graf by měl mít stručný a výstižný **název**
- obě **osy** grafu by měly být označeny názvy proměnných a jednotkami měření (závislá proměnná je obvykle na svislé ose)
- **počátek os** by měl být v nule – pokud není, je třeba to vyznačit
- **velikost** grafu a **rozsah** os by měl být takový, aby většina dat zabírala celý graf

Kontrolní otázky

- rozdíly mezi absolutními a relativními četnostmi, poměrem a mírou
- 3 základní míry polohy (+ u jakých dat použijeme průměr, modus či medián)
- základní míry variability, výpočet rozptylu
- základní typy grafů, výhody/nevýhody
- odlehlá pozorování

Doplňující literatura

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