Popisná statistika

- grafy
- z-skóry

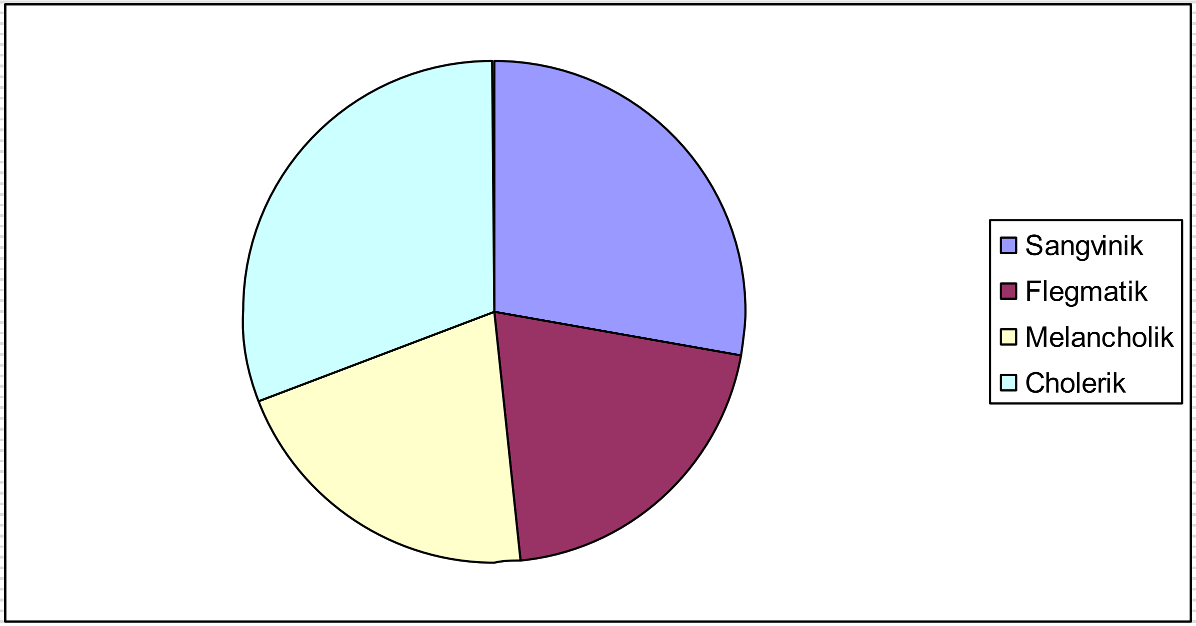
- pravděpodobnost
  (jako příprava pro úvod do induktivní statistiky)

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

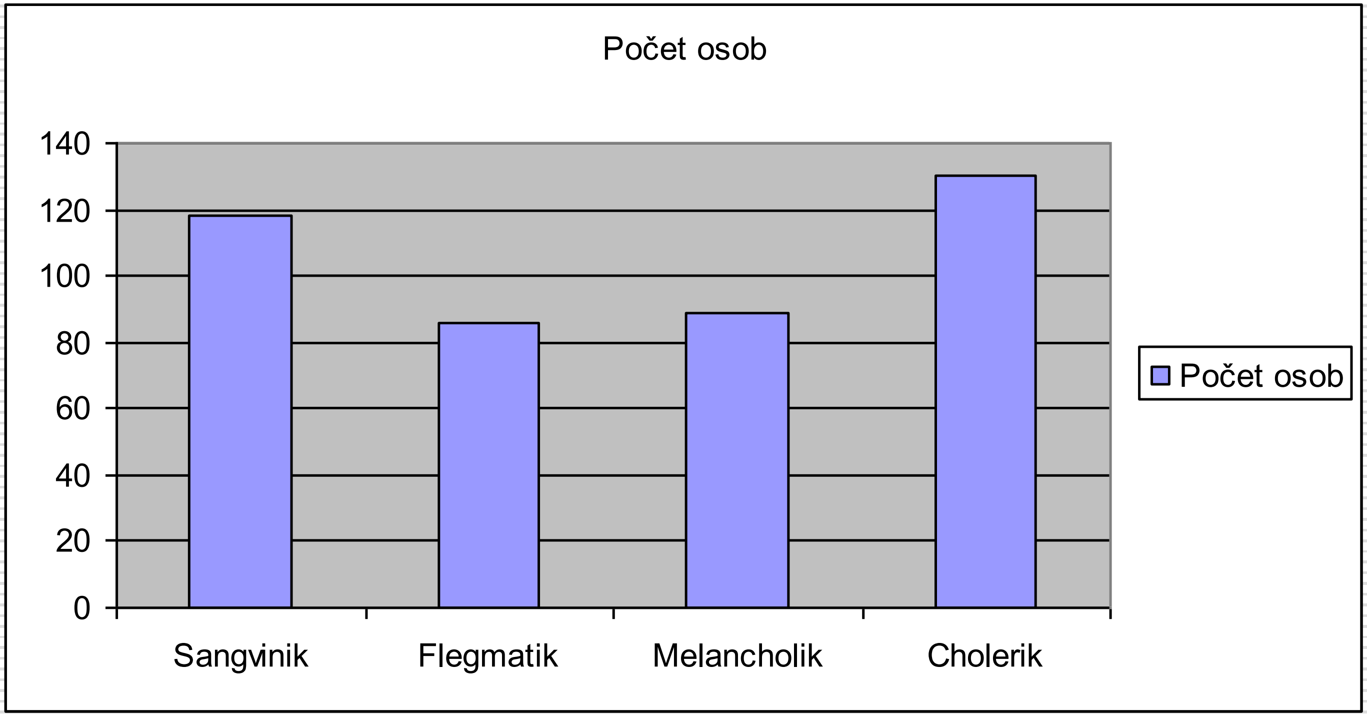


Výsečový graf

- 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

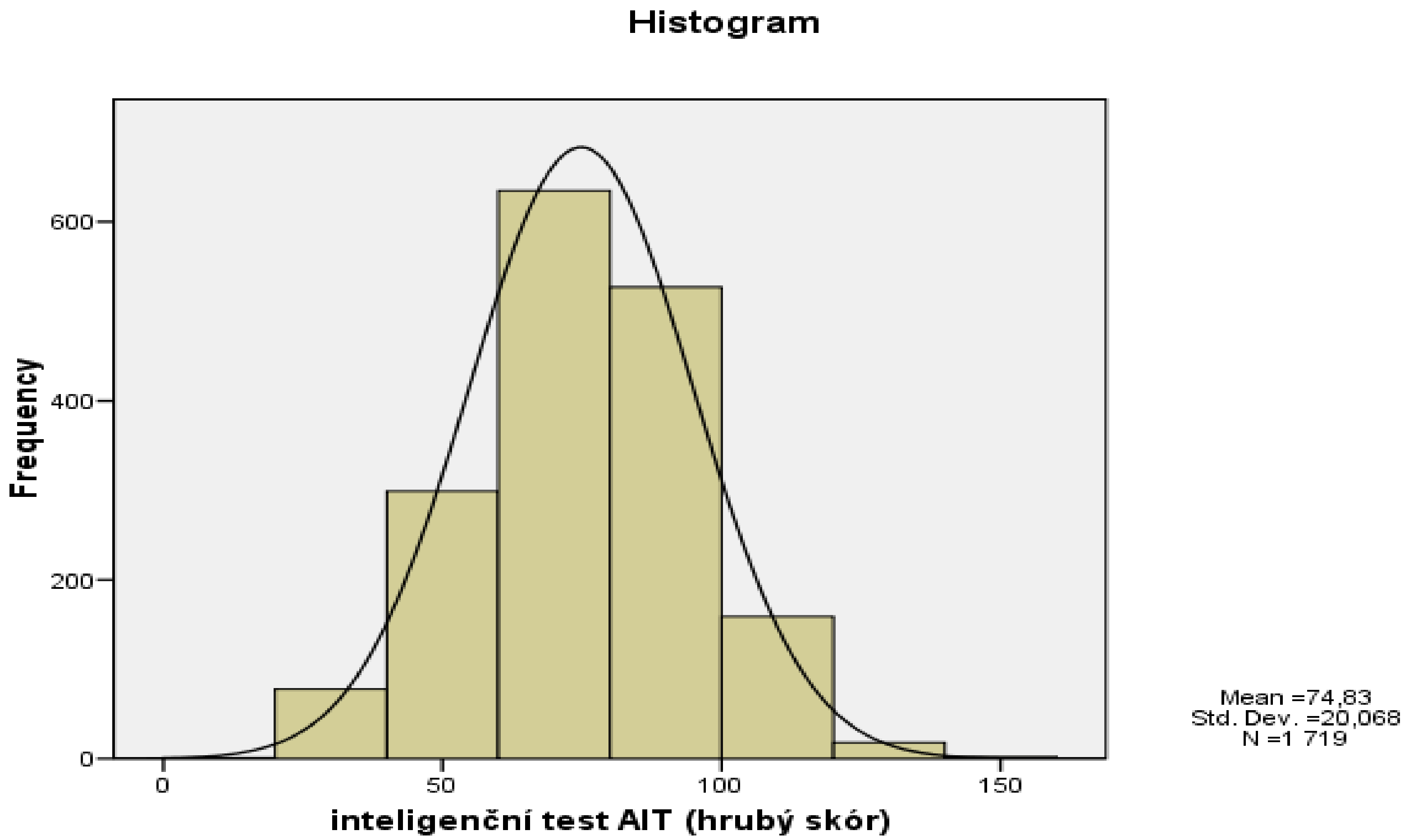
Histogram

- 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

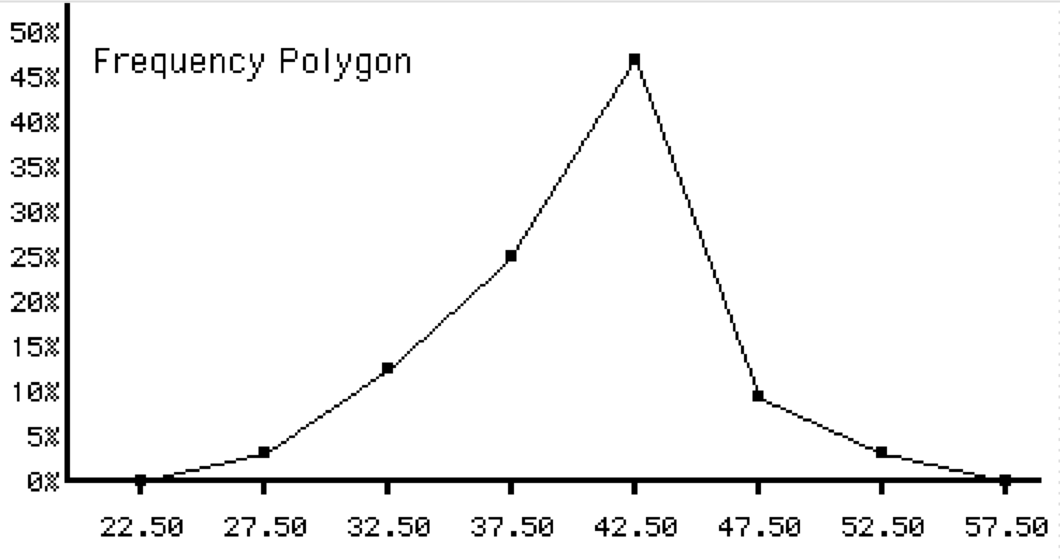
- **výhody**: umožňuje detekovat odlehlá pozorování, srovnání s normálním rozdělením
- **nevýhody**: nezjistíte přesné hodnoty jednotlivých případů, obvykle se nezobrazují data pro více skupin případů

Histogram


Histogram
Frequency
600
400
200
0
0
50
100
150
inteligenční test AIT (hrubý skór)
Mean =74,83
Std. Dev. =20,068
N =1 719

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 stonky

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)
```

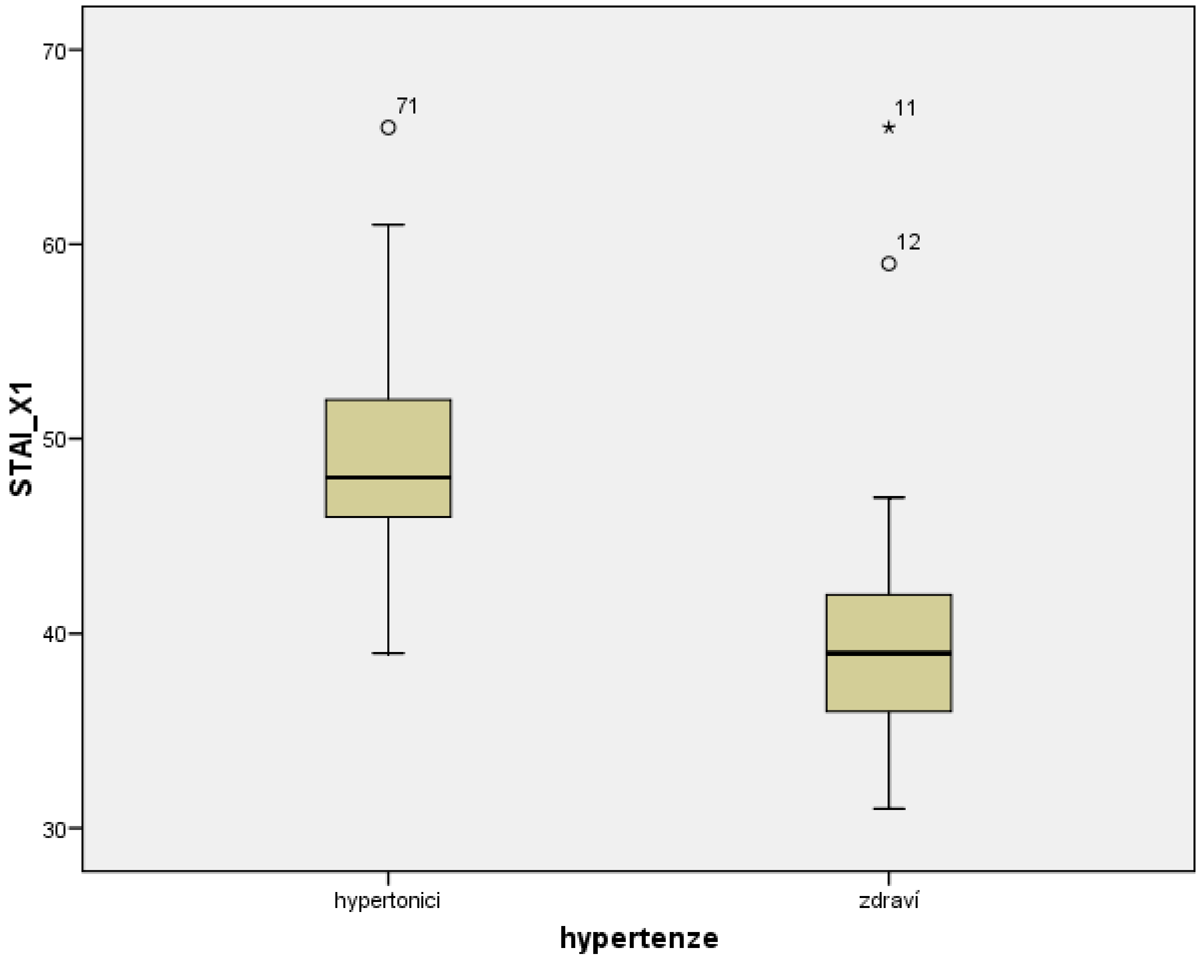
Stromkový diagram

- **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 mezikvartilové odchylky od hodnoty kvartilů – v grafu kolečka) a extrémní pozorování (obvykle více než 3x mezikvartilové odchylky – v grafu hvězdičky)
- 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 <u>jednotkami měření</u> (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

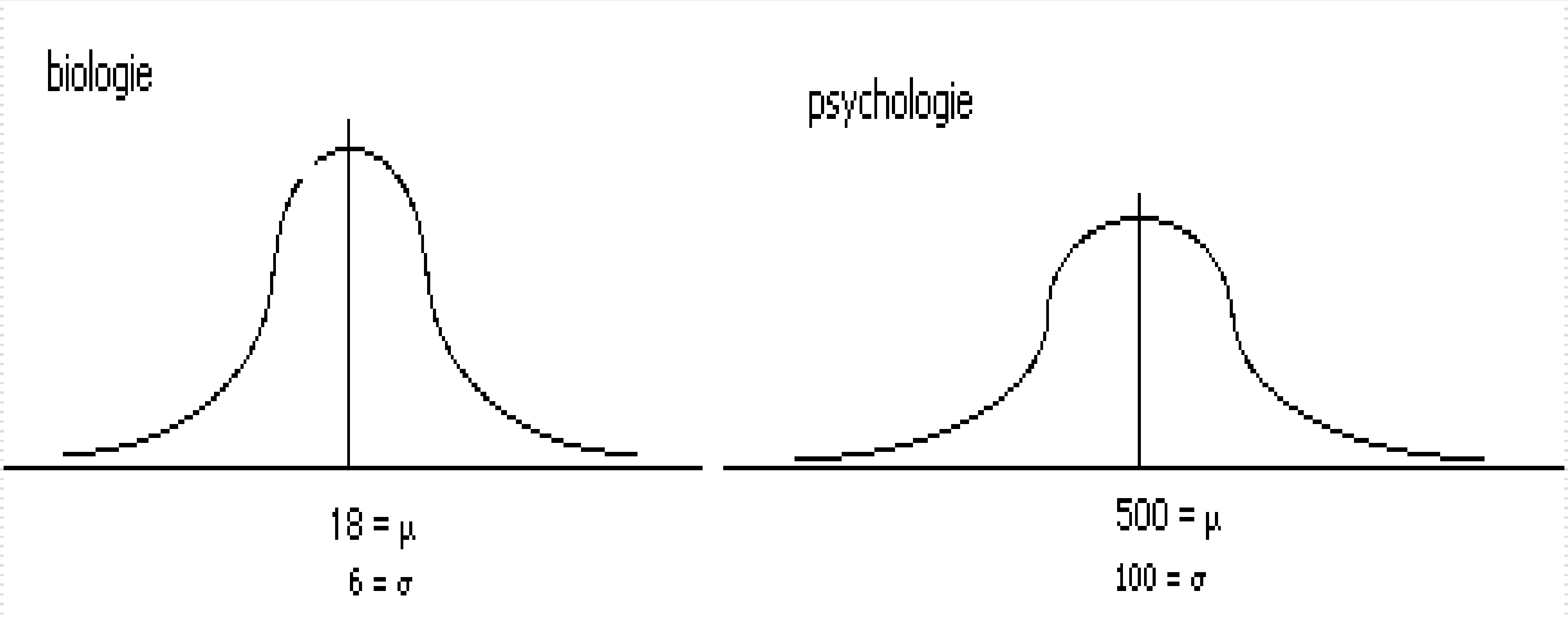
Z-skóry

- umožňují najít a popsat **pozici každé hodnoty** v rámci rozdělení hodnot
- a také **srovnávání hodnot** pocházejících z měření **na rozdílných stupnicích**
- hrubé skóry jsou převedeny na **standardizovanou stupnici** (jednotkou je směrodatná odchylka)

Z-skóry - příklad

- např. skóry ze dvou testů – biologie a psychologie
- student získal 26 bodů z biologie a 620 z psychologie. Ve kterém předmětu byl lepší?

Z-skóry - příklad



Z-skóry

- přímé porovnání není snadné – skóry z obou testů mají rozdílné průměry i směrodatné odchylky
- z skór =odchylka skóru od průměru vzhledem k velikosti směrodatné odchylky
- z = odch. od průměru/směr. odch.

Z-skóry - příklad

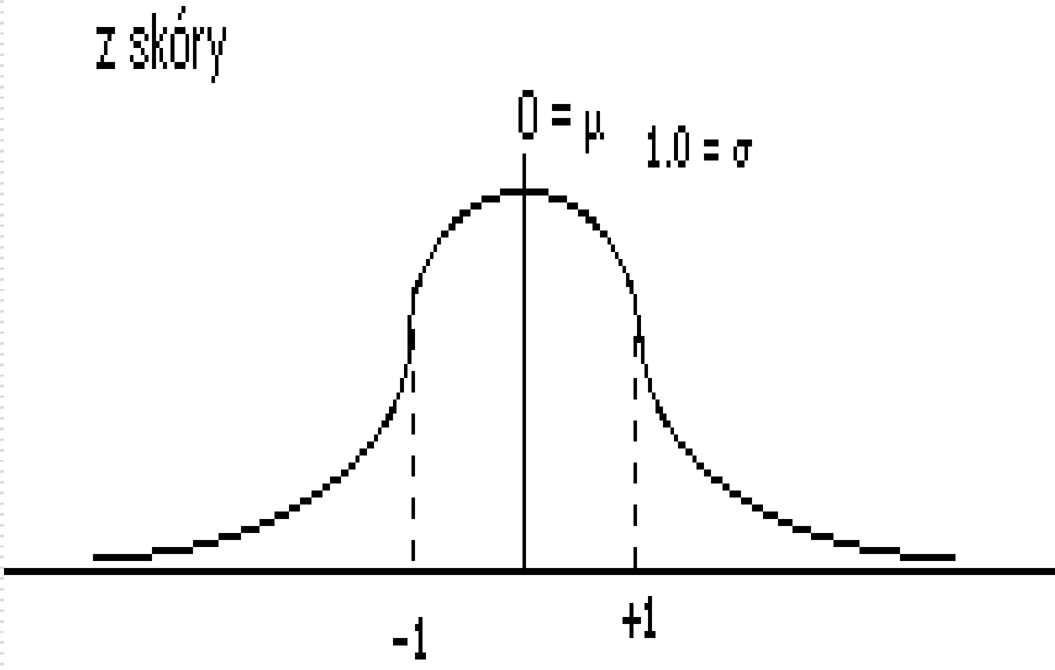
- skór z biologie: (26-18)/6 = 1,33
- skór psychologie: (620-500)/100=1,2
- v biologii byl student lepší – 1,33 směrodatné odchylky nad průměrem

Z-skóry

- z-skór přesně udává pozici každé hodnoty vzhledem k ostatním hodnotám
- znaménko (+ nebo -) ukazuje, zda je hodnota nad nebo pod průměrem rozdělení
- hodnota z-skóru upřesňuje, kolik směrodatných odchylek byla hodnota od průměru vzdálena

Z-skóry

- průměr rozdělení z-skórů je vždy 0
- směrodatná odchylka je 1



Z-skóry

vzorec pro výpočet z-skóru hodnoty X

- <u>u populace</u>: $z = (X - \mu) / \sigma$
- <u>u vzorku</u>: $z = (X - m) / s$

Z-skóry

- podobně můžeme i z-skór převést na hrubý skór, známe-li průměr a směrodatnou odchylku

Z-skóry

- např. u stupnice IQ
- $\mu = 100, \sigma = 15$
- pro osobu se z=-3 (3 směrodatné odchylky pod průměrem) bude IQ ?

Z-skóry

- např. u stupnice IQ $\mu = 100$, $\sigma = 15$
- pro osobu se z=-3 (3 směrodatné odchylky pod průměrem) bude IQ

$$X = Z \cdot \sigma + \mu$$

$$X = -3 \cdot 15 + 100$$

$$X = 55$$

Rozdělení z-skórů

- **tvar** rozdělení z-skórů je **stejný** jako tvar původního rozdělení hrubých skórů
- průměr je 0, směrodatná odchylka 1
- transformace změní jen označení hodnot na ose X

Pravděpodobnost

- postupy induktivní statistiky vycházejí z teorie pravděpodobnosti
- **pravděpodobnost**, že nastane určitý výsledek, **definujeme** jako podíl

$$P(A) = \frac{\text{počet pokusů, kdy nastal jev A}}{\text{celkový počet jevů}}$$

Pravděpodobnost - příklady

- jaká je pravděpodobnost, že si z balíčku 52 karet vytáhneme určitou kartu (např. pikovou dámu) ?

Pravděpodobnost - příklady

- jaká je pravděpodobnost, že si z balíčku 52 karet vytáhneme určitou kartu (např. pikovou dámu) ?

P (piková dáma) = f/N = 1/52 = 0,019= 1,9%

Pravděpodobnost - příklady

- jaká je pravděpodobnost, že při hodu kostkou padne trojka nebo šestka ?

Pravděpodobnost - příklady

- jaká je pravděpodobnost, že při hodu kostkou padne trojka nebo šestka ?

P (3 n. 6) = f/N = 2/6 = 0,333= 33,3%

Pravděpodobnost

- pravděpodobnost bývá uváděna nejčastěji jako **podíl** (0,33), **zlomek** (1/3) nebo **procento** (33,3%)
- pravděpodobnost určitého jevu nebo třídy jevů můžeme odhadnout z rozdělení hodnot (četností)

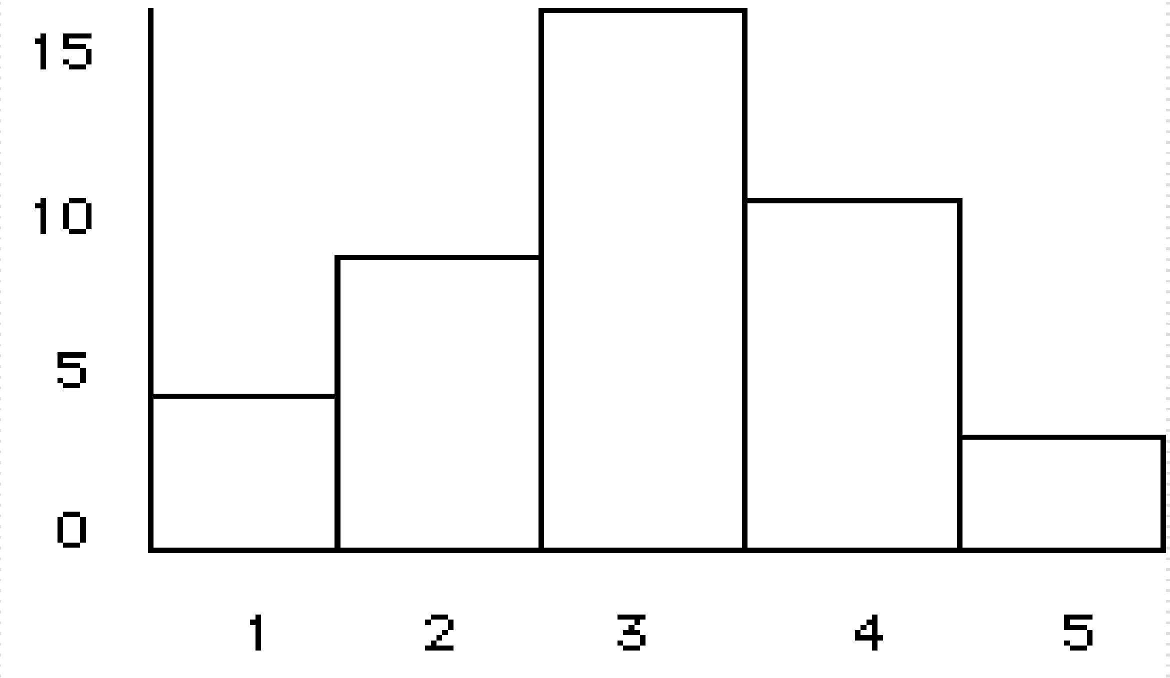
Pravděpodobnost - příklady

- představme si, že máme krabici se 40 očíslovanými žetony s čísly 1 – 5
- v tabulce jsou uvedeny absolutní i relativní četnosti jednotlivých čísel žetonů

Pravděpodobnost

X	*f*	*p*
5	2	0,05
4	10	0,25
3	16	0,40
2	8	0,20
1	4	0,10

Pravděpodobnost



Pravděpodobnost - příklady

- vaším úkolem je vytáhnout 1 žeton
- **jaká je pravděpodobnost, že vytáhnete žeton s číslem 3?**

Pravděpodobnost

X	f	p
5	2	0,05
4	10	0,25
3	16	0,40
2	8	0,20
1	4	0,10

Pravděpodobnost

- vaším úkolem je vytáhnout 1 žeton
- jaká je pravděpodobnost, že vytáhnete žeton s číslem 3?
- **p (3) = f/N = 16/40 =0,40**

  nebo 2/5 či 40%

Pravděpodobnost

- **Jaká je pravděpodobnost, že vytáhnete žeton s číslem vyšším než 2?**

Pravděpodobnost

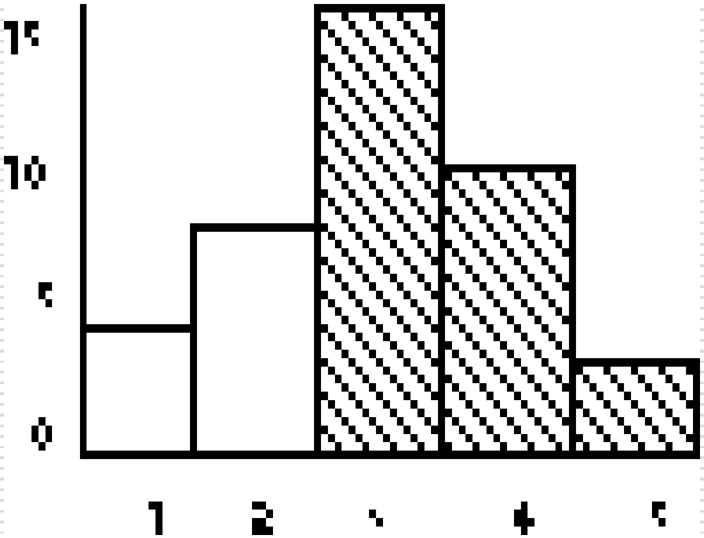
X	f	p
5	2	0,05
4	10	0,25
3	16	0,40
2	8	0,20
1	4	0,10

Pravděpodobnost

- Jaká je pravděpodobnost, že vytáhnete žeton s číslem vyšším než 2?

  $p(X > 2) = ?$

  0,05 + 0,25 + 0,40 = **0,70**



Pravděpodobnost

- **Jaká je pravděpodobnost, že vytáhnete žeton s číslem nižším než 5?**

Pravděpodobnost

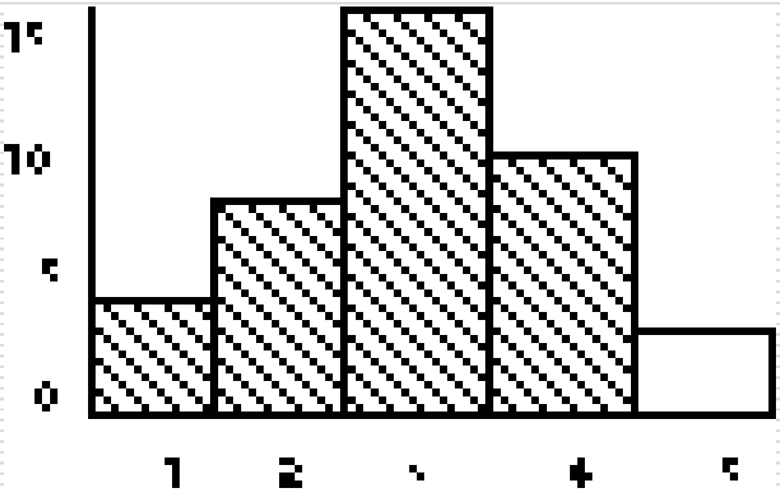
X	*f*	*p*
5	2	0,05
4	10	0,25
3	16	0,40
2	8	0,20
1	4	0,10

Pravděpodobnost

- Jaká je pravděpodobnost, že vytáhnete žeton s číslem nižším než 5?

  $p(X < 5) = ?$

  0,10 + 0,20 + 0,40 + 0,25 = **0,95**



Pravděpodobnost

- **Jaká je pravděpodobnost, že vytáhnete žeton s číslem nižším než 4 a vyšším než 1?**

Pravděpodobnost

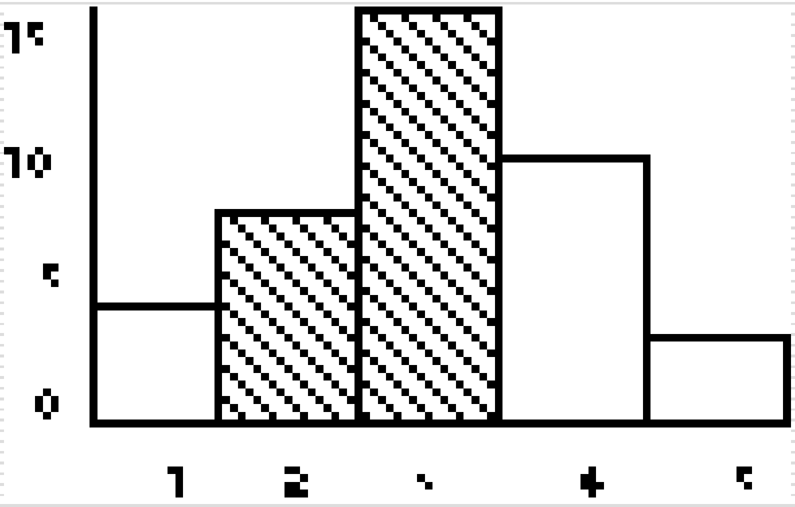
X	*f*	*p*
5	2	0,05
4	10	0,25
3	16	0,40
2	8	0,20
1	4	0,10

Pravděpodobnost

- Jaká je pravděpodobnost, že vytáhnete žeton s číslem nižším než 4 a vyšším než 1?

  $p(4 > X > 1) = ?$

  0,20 + 0,40 = **0,60**



Pravděpodobnost

- pravděpodobnost odpovídá hustotě **oblasti pod křivkou** pro daný interval

Kontrolní otázky

- základní typy grafů, výhody/nevýhody
- odlehlá pozorování
- výpočet a interpretace z-skóru

Doplňující literatura

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