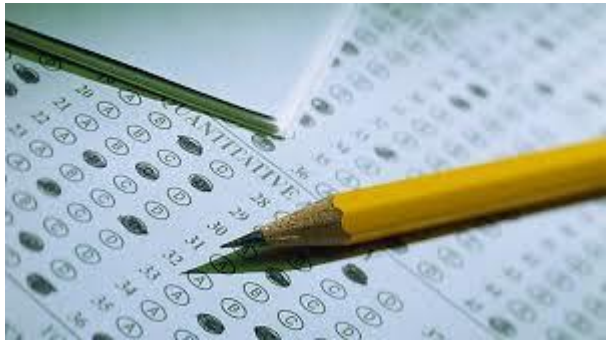
STATISTIKA

VĚDA O USUZOVÁNÍ NA ZÁKLADĚ DAT



Patrícia Martinková
Ústav informatiky AV ČR
martinkova@cs.cas.cz
www.cs.cas.cz/martinkova

1.LF UK, 22. a 30. března 2017

Motivace

- Velké množství (medicínských i jiných) dat je sbíráno každý den
- Více je uplatňovaná tzv. „Evidence based medicine“
- Medik tohoto století se neobejde bez znalosti metod analýzy dat:
  - Pro porozumění odborným článkům
  - Pro kvalitní sběr vlastních dat
  - Pro dobrou komunikaci se statistikem, který pomůže data analyzovat
  - Pro samostatnou analýzu dat a interpretaci výsledků

- Znalost statistických metod je pro Vás **důležitá**!

Obsah (požadavky k zápočtu)

Popisná statistika

- kategorická a numerická data
- četnosti (absolutní, relativní, kumulativní)
- aritmetický průměr, rozptyl, směrodatná odchylka
- medián, modus, kvartil, percentil, krabicový graf
- grafická a tabulková prezentace statistických dat, histogram

Induktivní statistika

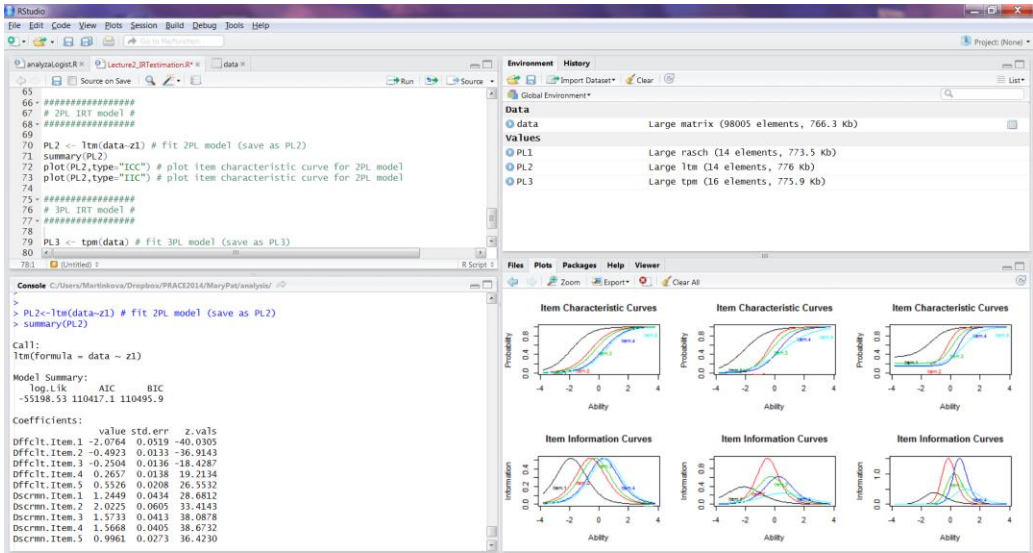
- náhodný výběr, náhodný jev
- rozdělení pravděpodobnosti (binomické, normální /Gaussovo)
- intervalový odhad
- testování hypotéz (princip a význam).

Zdroje pro tuto přednášku

- Jana Zvárová: Biomedicínská statistika I. Základy statistiky pro biomedicínské obory. Karolinum, 2016.
- Karel Zvára: Biostatistika. Karolinum, 2008.
- Karel Zvára: Biomedicínská statistika IV. Základy statistiky v prostředí R. Karolinum, 2013.

Software:

- Leccos lze spočítat v Excelu
- $tatistica, $P$$, A, ...
- **Statistické prostředí R**



1.

Popisná statistika

Popisná statistika

Účel: popsat daný soubor dat

Číselně:

- průměr, směrodatná odchylka
- minimum, maximum
- medián, kvartily
- absolutní a relativní četnosti

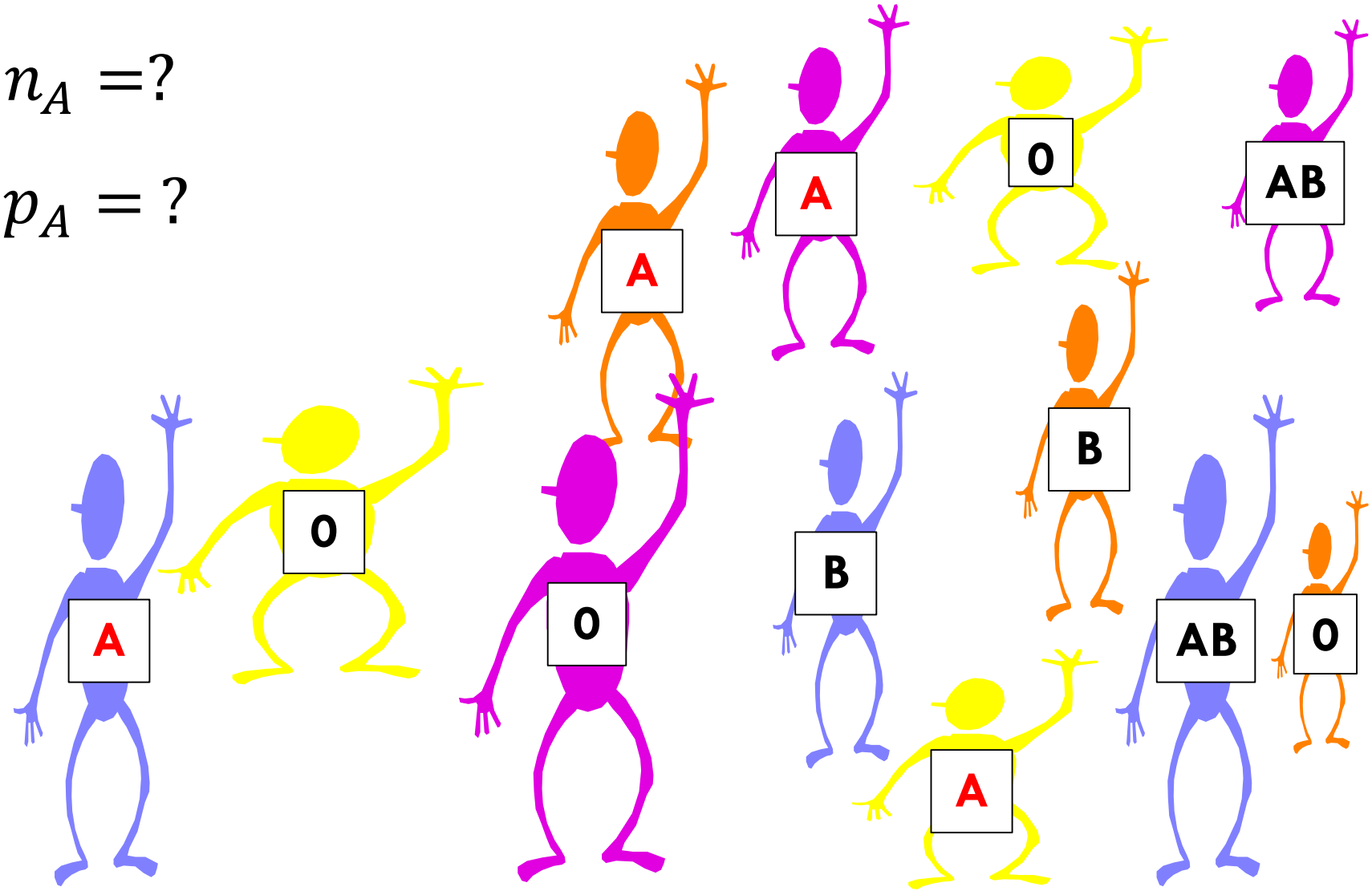
Graficky:

- výsečový („koláčový“) graf
- sloupcový graf
- histogram
- dvourozměrný (x-y) graf (scatter plot) …

Absolutní a relativní četnost (krevní skupina)

$n_A = ?$

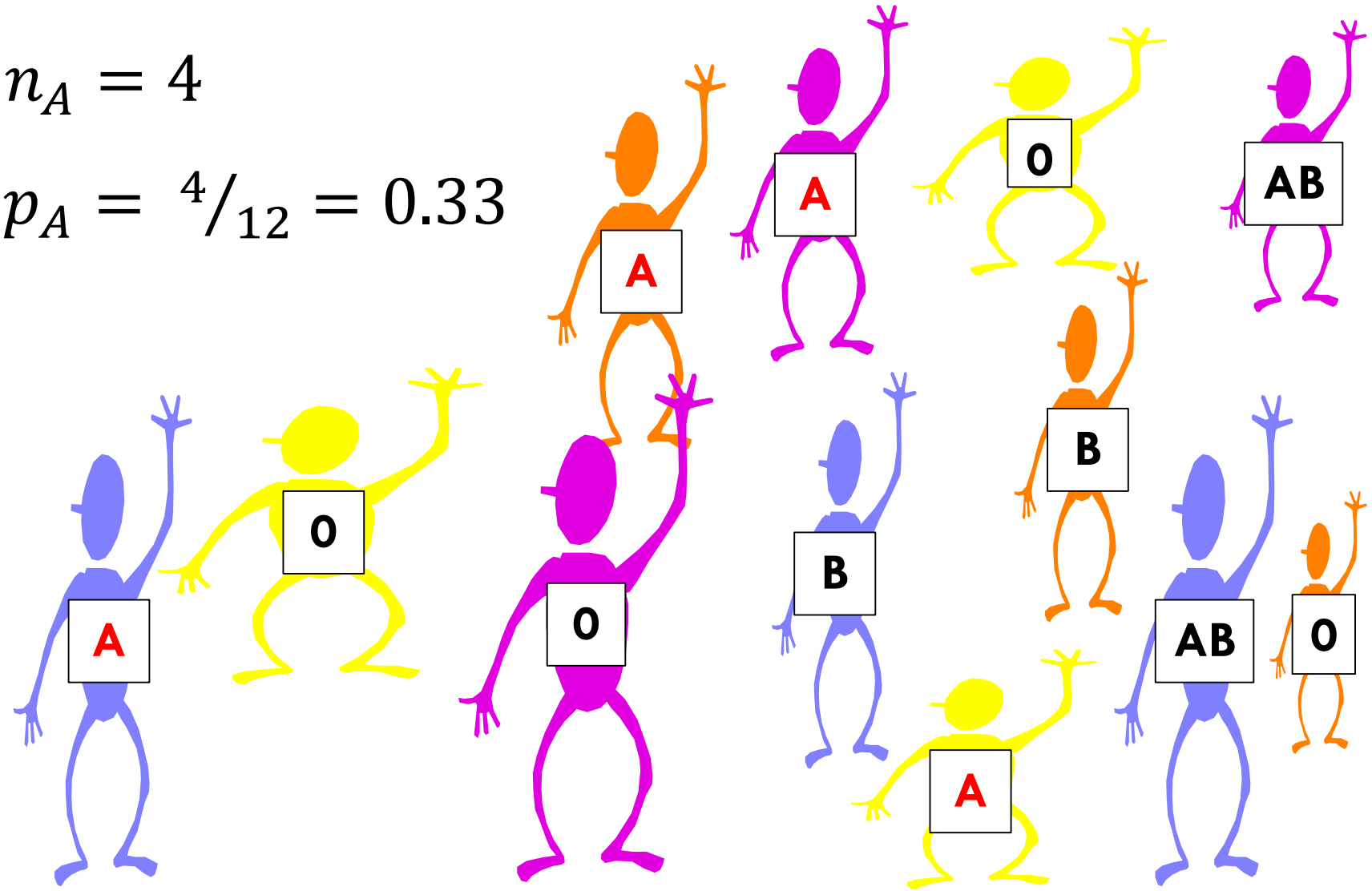
$p_A = ?$



Absolutní a relativní četnost (krevní skupina)

$n_A = 4$

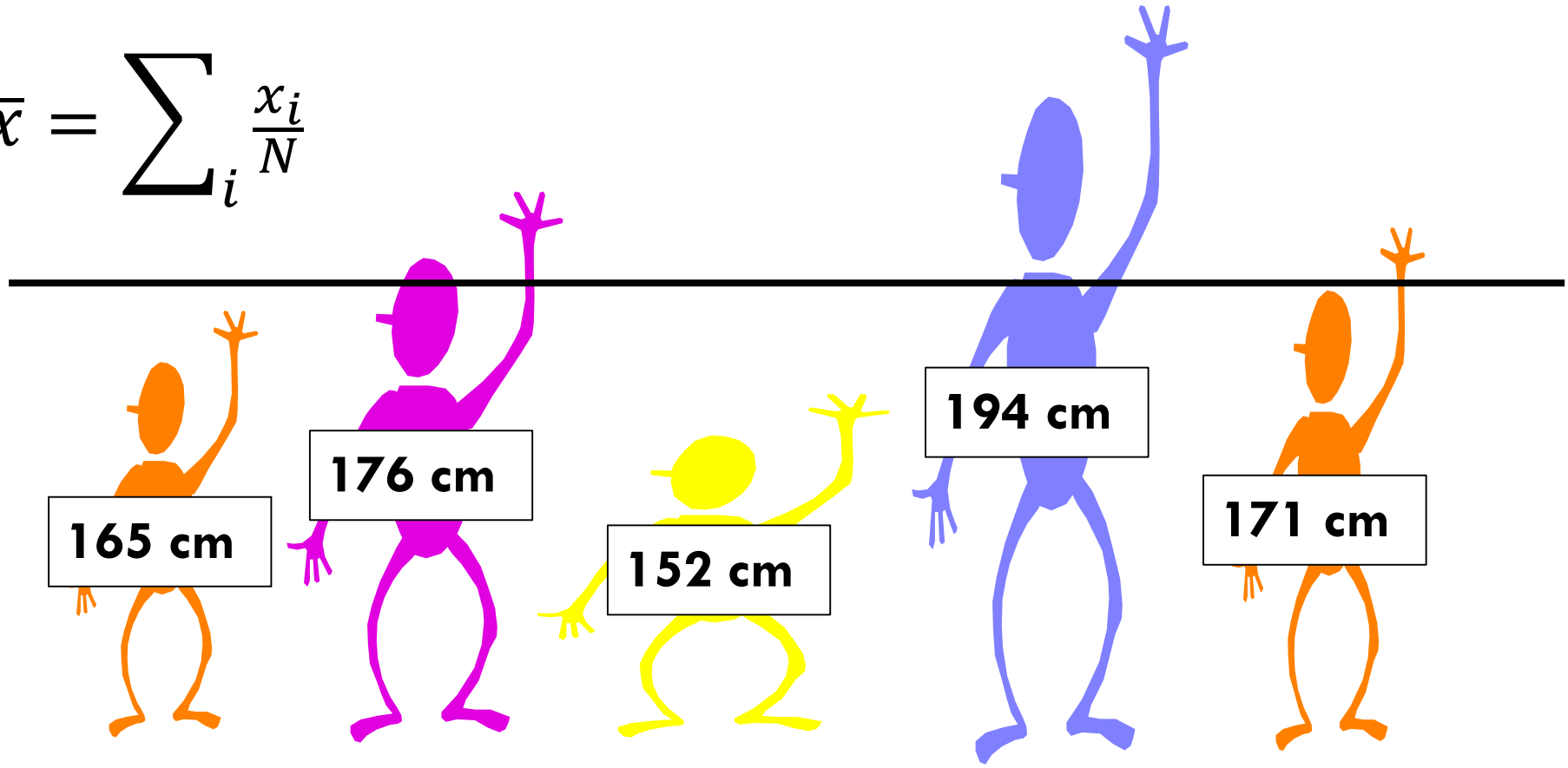
$p_A = {}^{4}/_{12} = 0.33$



Aritmetický průměr (výška)

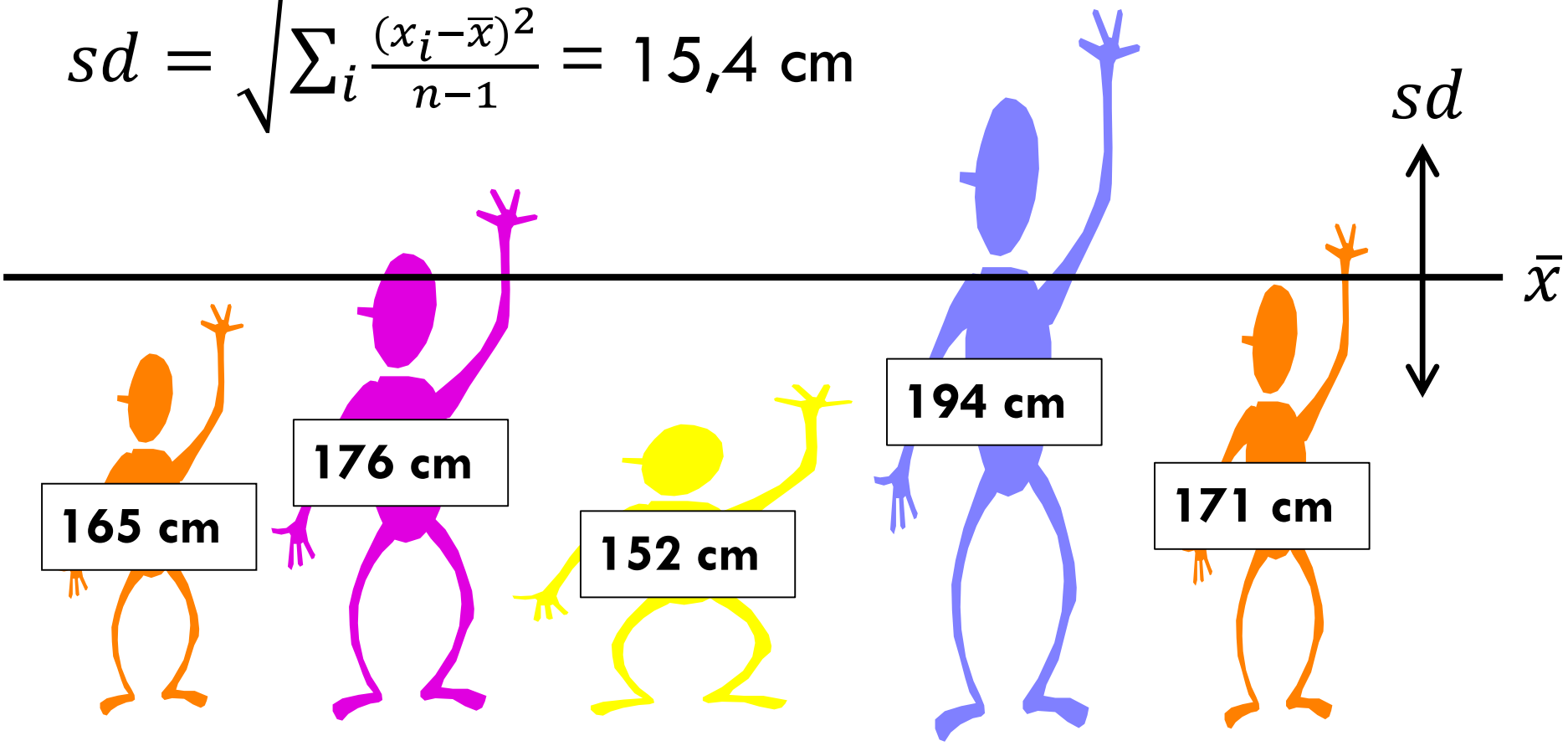
$$\bar{x} = (165 + 176 + 152 + 194 + 171) / 5 = 171{,}6$$

$$\bar{x} = \sum_i \frac{x_i}{N}$$

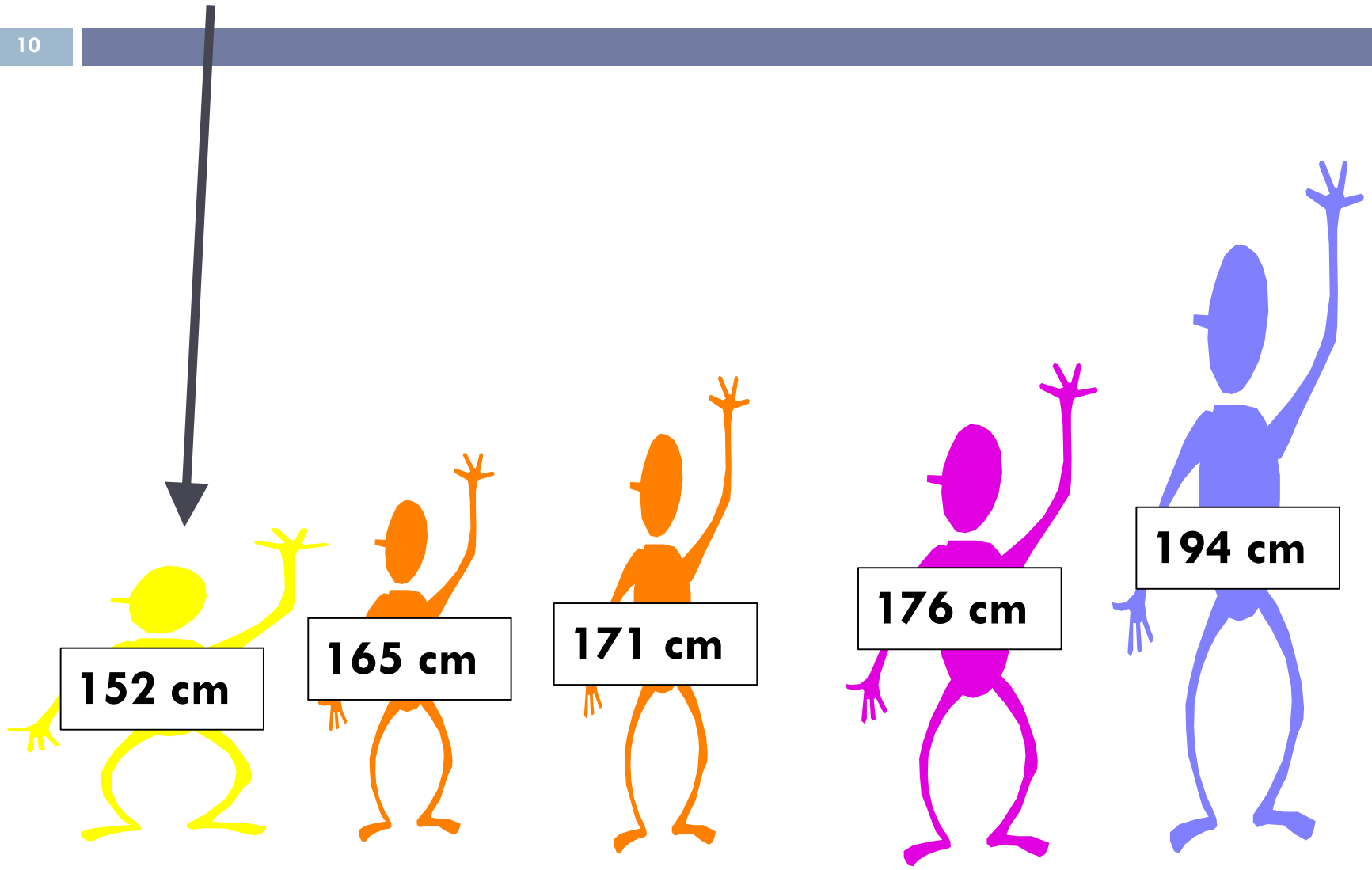


Směrodatná odchylka (výška)

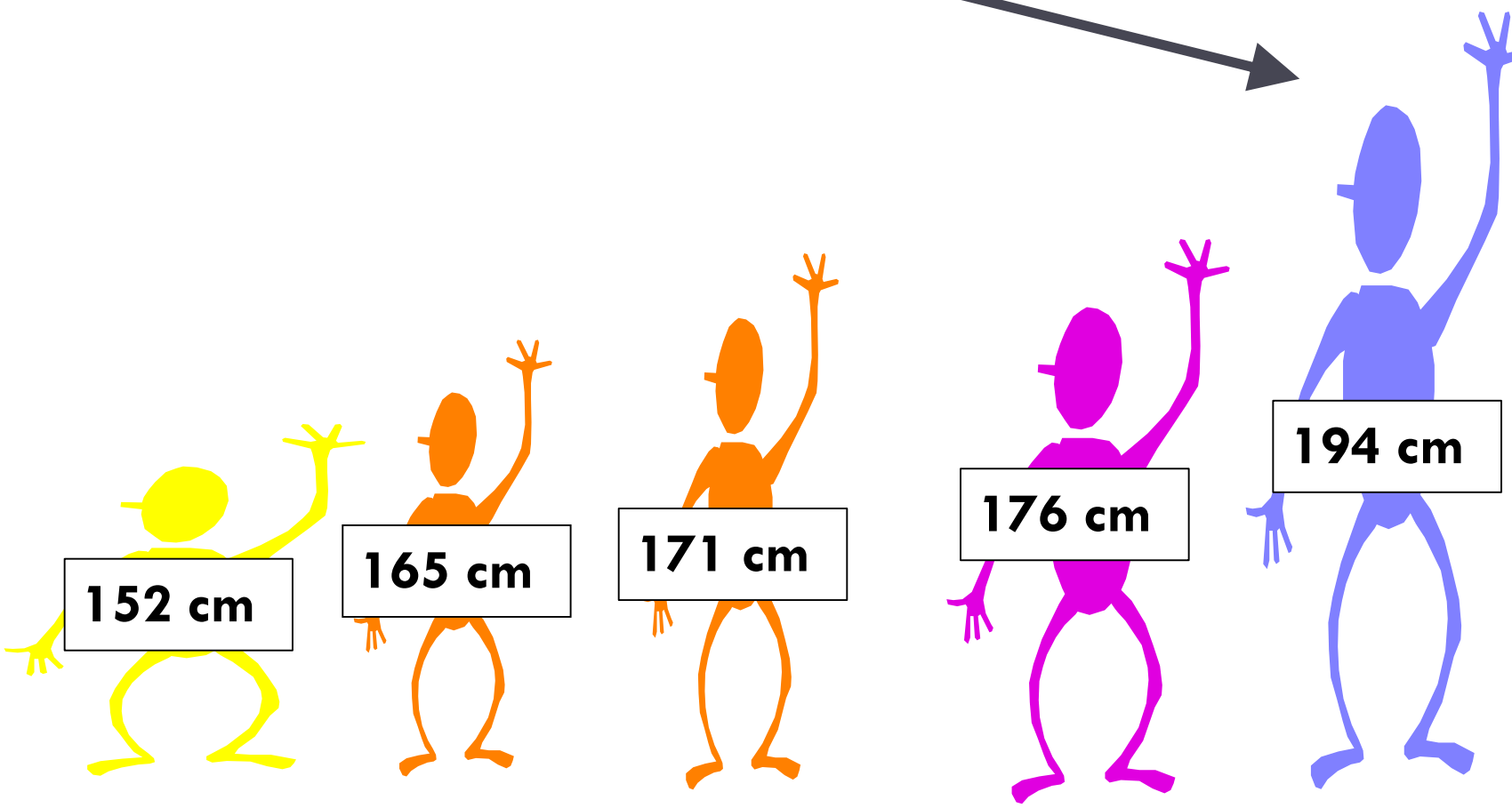
$$sd = \sqrt{\sum_i \frac{(x_i - \bar{x})^2}{n-1}} = 15{,}4 \text{ cm}$$



Minimum (výška)

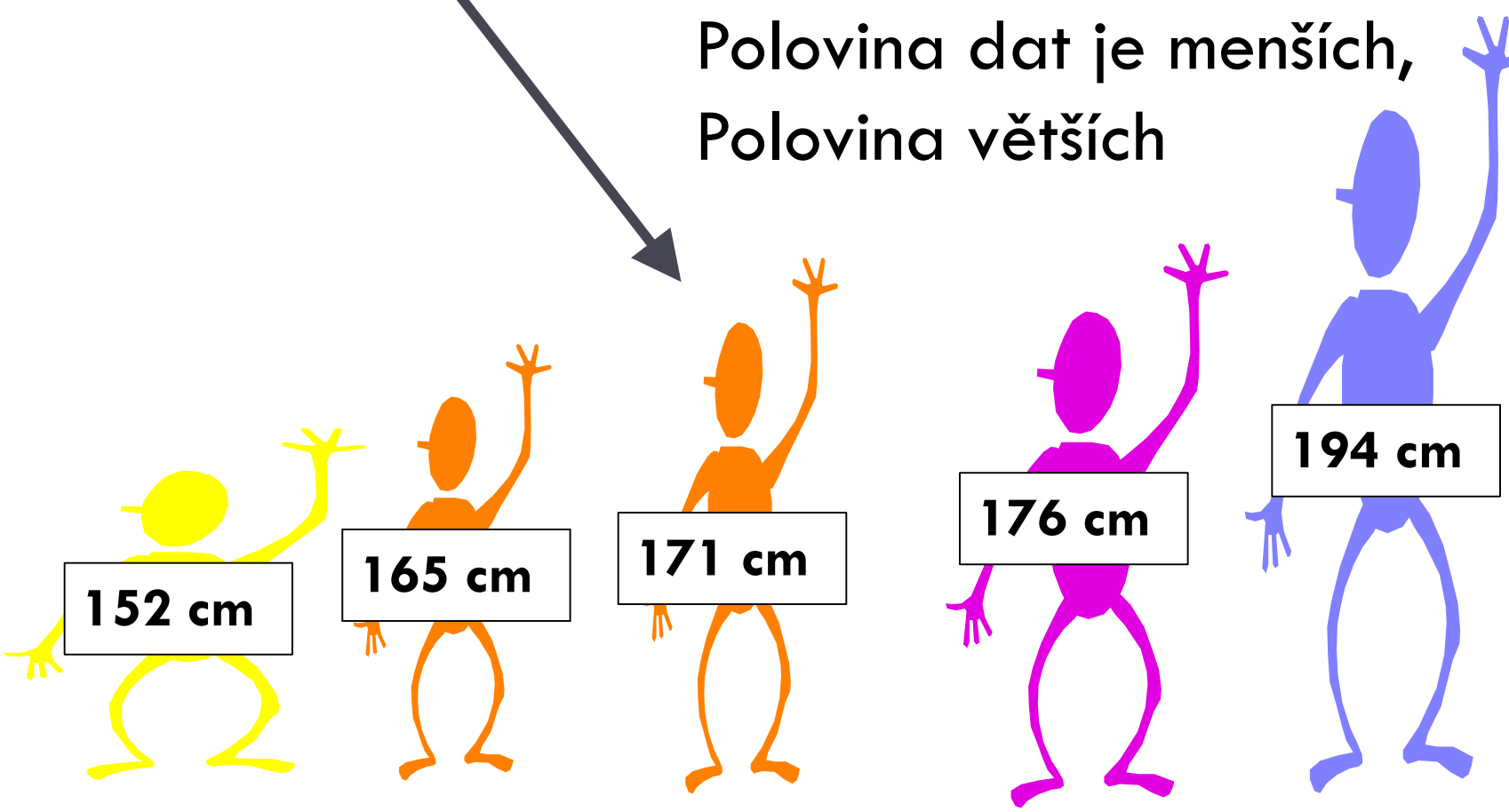


Maximum (výška)



Medián (výška)

Prostřední pozorování:
Polovina dat je menších,
Polovina větších



Pozn.: Pokud sudý počet, uvažuje se průměr prostředních dvou hodnot

Příklad: Pět mužů ze Seattlu

Pět mužů popíjí kávu v kavárně v Seattlu.

Jejich roční příjmy jsou 19, 20, 24, 25 a 27 tisíc USD.

- Jaký je jejich průměrný příjem?
- Jaký je medián jejich příjmu?

Příklad: Pět mužů ze Seattlu

Pět mužů popíjí kávu v kavárně v Seattlu.

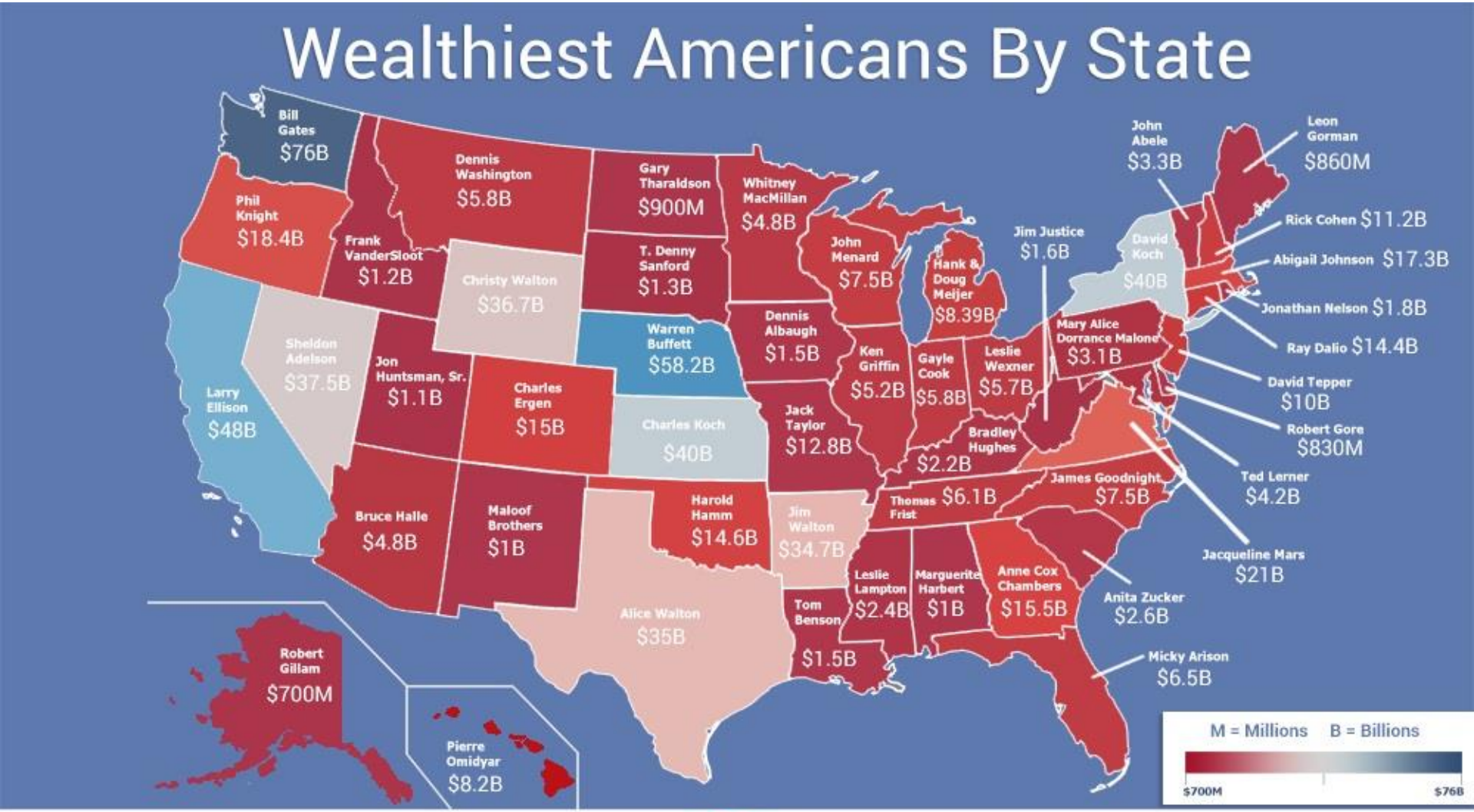
Jejich roční příjmy jsou 19, 20, 24, 25 a 27 tisíc USD.

- Jaký je jejich průměrný příjem?

  $(19+20+24+25+27)/5 =$ **23 tisíc USD**

- Jaký je medián jejich příjmu? **24 tisíc USD**

Příklad: Pět mužů ze Seattlu



MOVOTO

Příklad: Pět mužů ze Seattlu... a Bill Gates

Pět mužů popíjí kávu v kavárně v Seattlu.

Jejich roční příjmy jsou 19, 20, 24, 25 a 27 tisíc USD.

Vejde Bill Gates, jehož roční příjem je 719885 tisíc USD.

- Jaký je nyní průměrný příjem těchto **šesti** mužů?

- Jaký je medián jejich příjmu?

Příklad: Pět mužů ze Seattlu... a Bill Gates

Pět mužů popíjí kávu v kavárně v Seattlu.

Jejich roční příjmy jsou 19, 20, 24, 25 a 27 tisíc USD.

Vejde Bill Gates, jehož roční příjem je 719885 tisíc USD.

- Jaký je nyní průměrný příjem těchto **šesti** mužů?

(19+20+24+25+27+ 719885)/6= 72000/6 =
=120 000 tisíc USD

- Jaký je medián jejich příjmu? 24 tisíc USD

- Kdy dáte přednost mediánu před průměrem?

Grafická interpretace (věk při prvním porodu)

Příklad Uvažujme věky 11 matek při prvním porodu:
19, 25, 23, 28, 20, 24, 23, 21, 26, 22, 22

Krabicový graf znázorňuje:

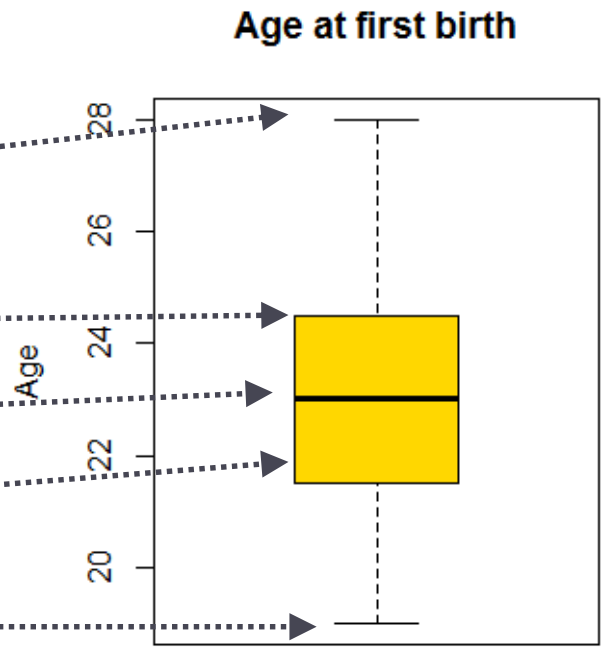
Maximum:

Horní kvartil:

Medián:

Dolní kvartil:

Minimum:



Pozn. Kvartil odděluje čtvrtinu dat od zbylých tří čtvrtin

Grafická interpretace (věk při prvním porodu)

Věky 11 matek při prvním porodu seřazené vzestupně:
19, 20, 21, 22, 22, 23, 23, 24, 25, 26, 28

Krabicový graf znázorňuje:

Maximum: 28 let

Horní kvartil: 24,5 let

Medián: 23 let

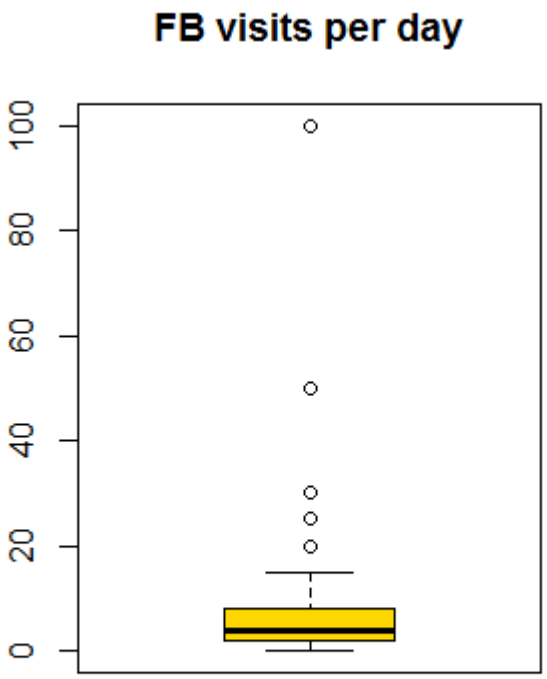
Dolní kvartil: 21,5 let

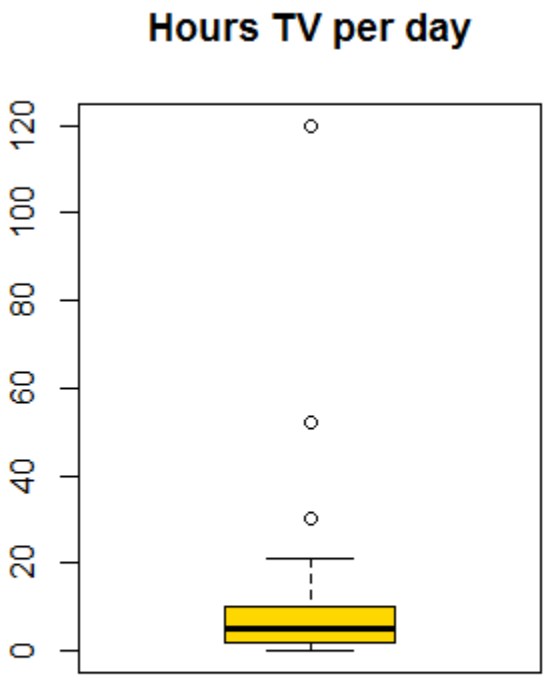
Minimum: 19 let

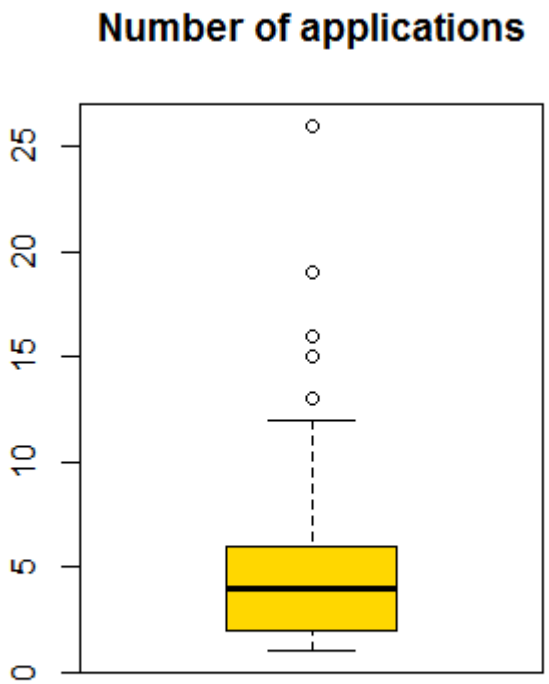
Pozn. Kvartil odděluje čtvrtinu dat od zbylých tří čtvrtin

Krabicový graf

Odhalte chybu v datech

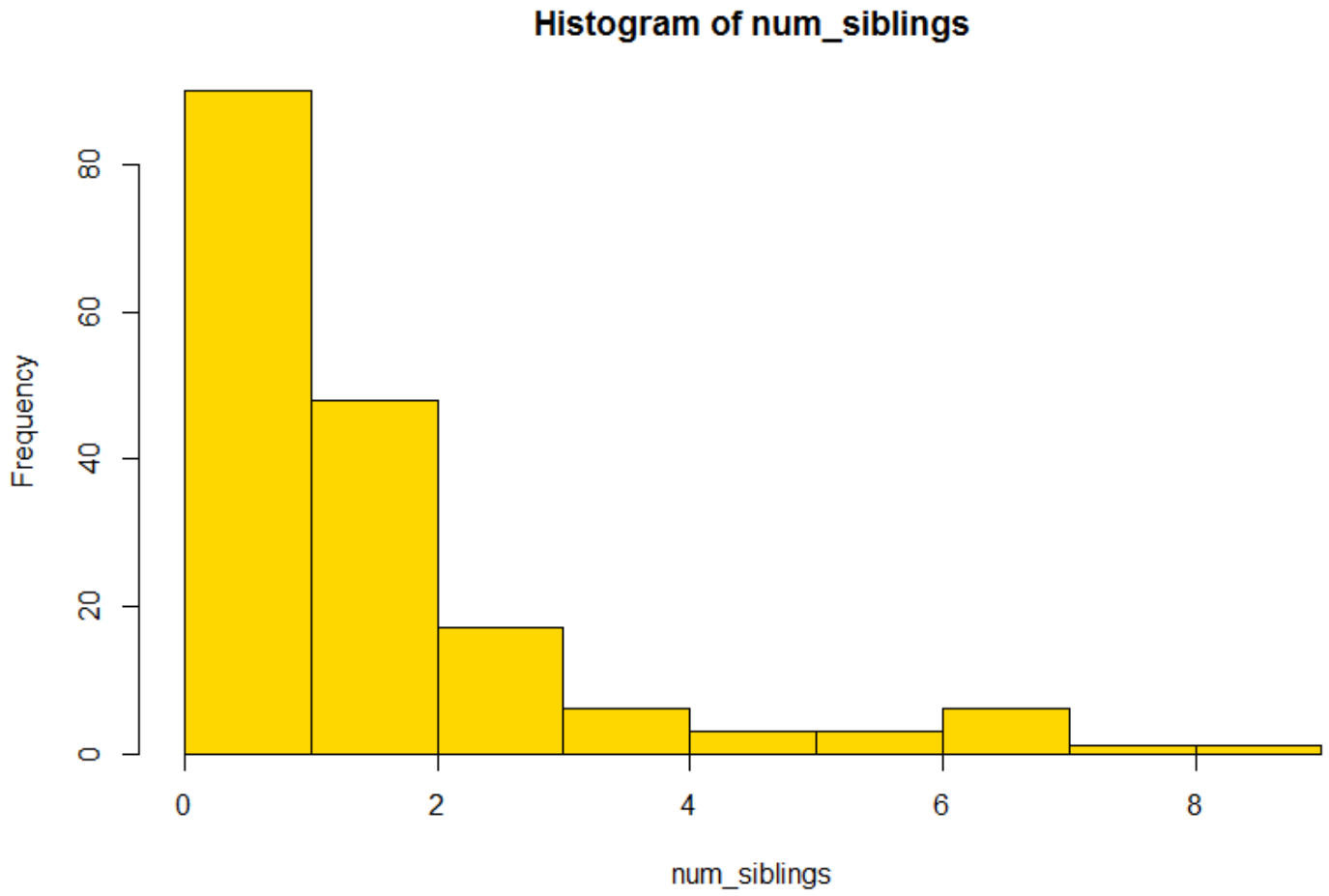






Pozn.: Body jsou tzv. **odlehlá pozorování**

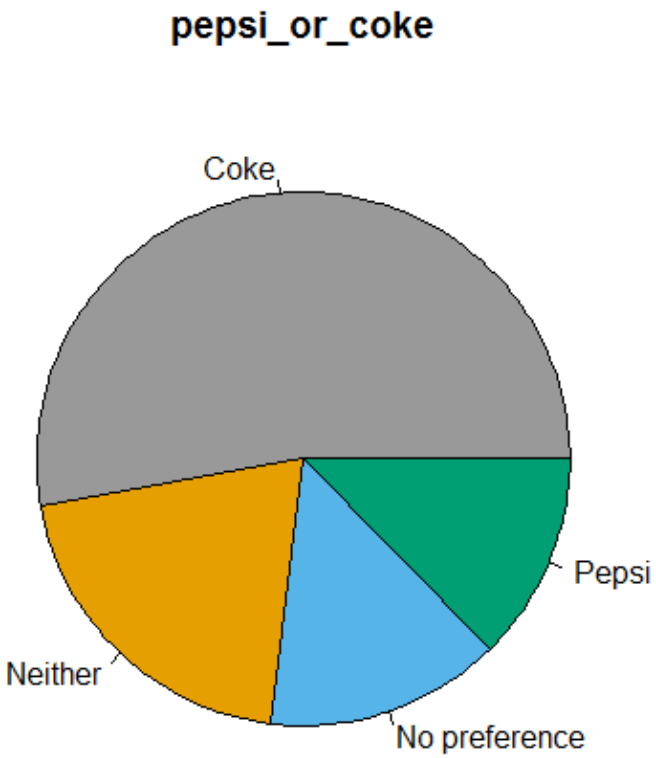
Histogram (počet sourozenců)



Jak by vypadal histogram porodní délky dětí?

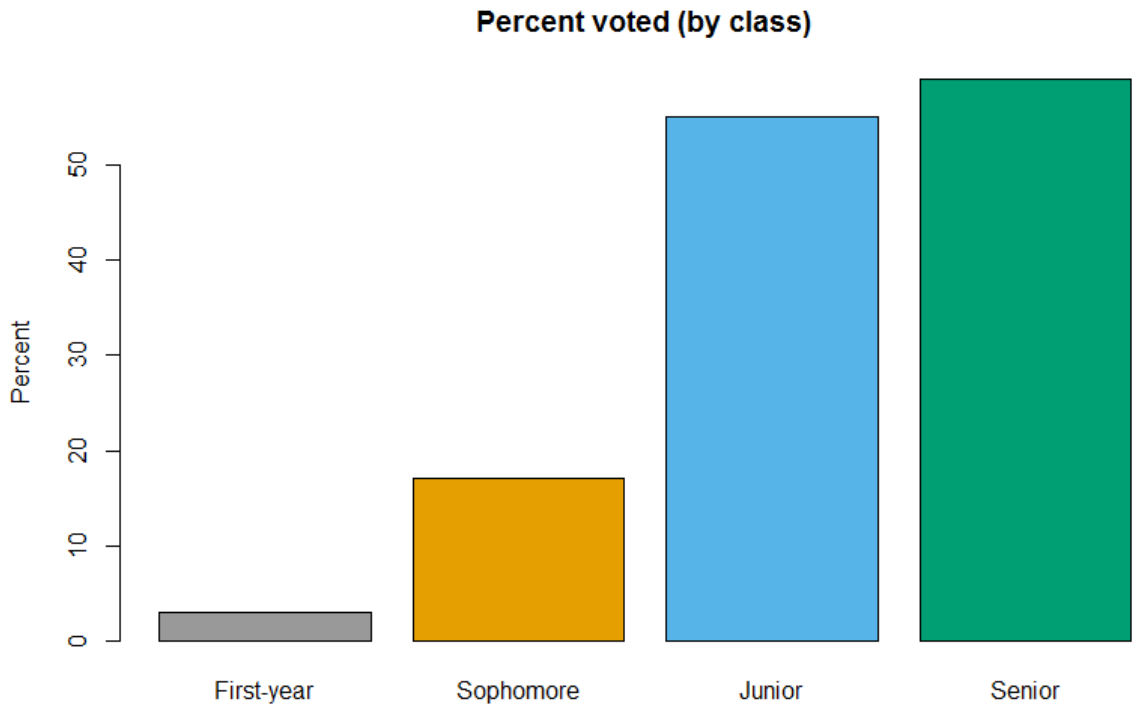
Koláčový graf

Vhodné pro znázornění frekvence kategoriálních dat (procenta se sčítají na 100%)

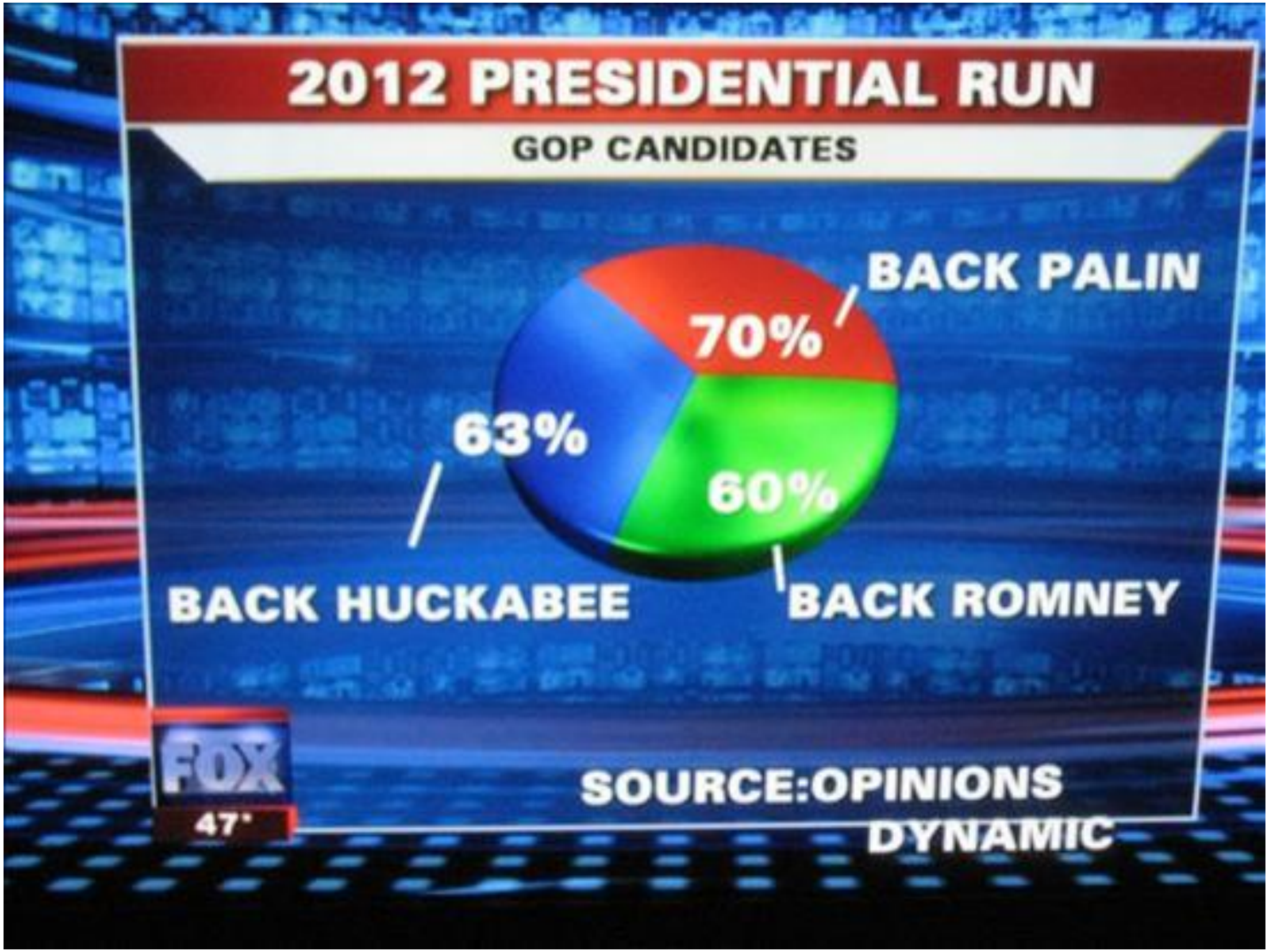


Bar plot

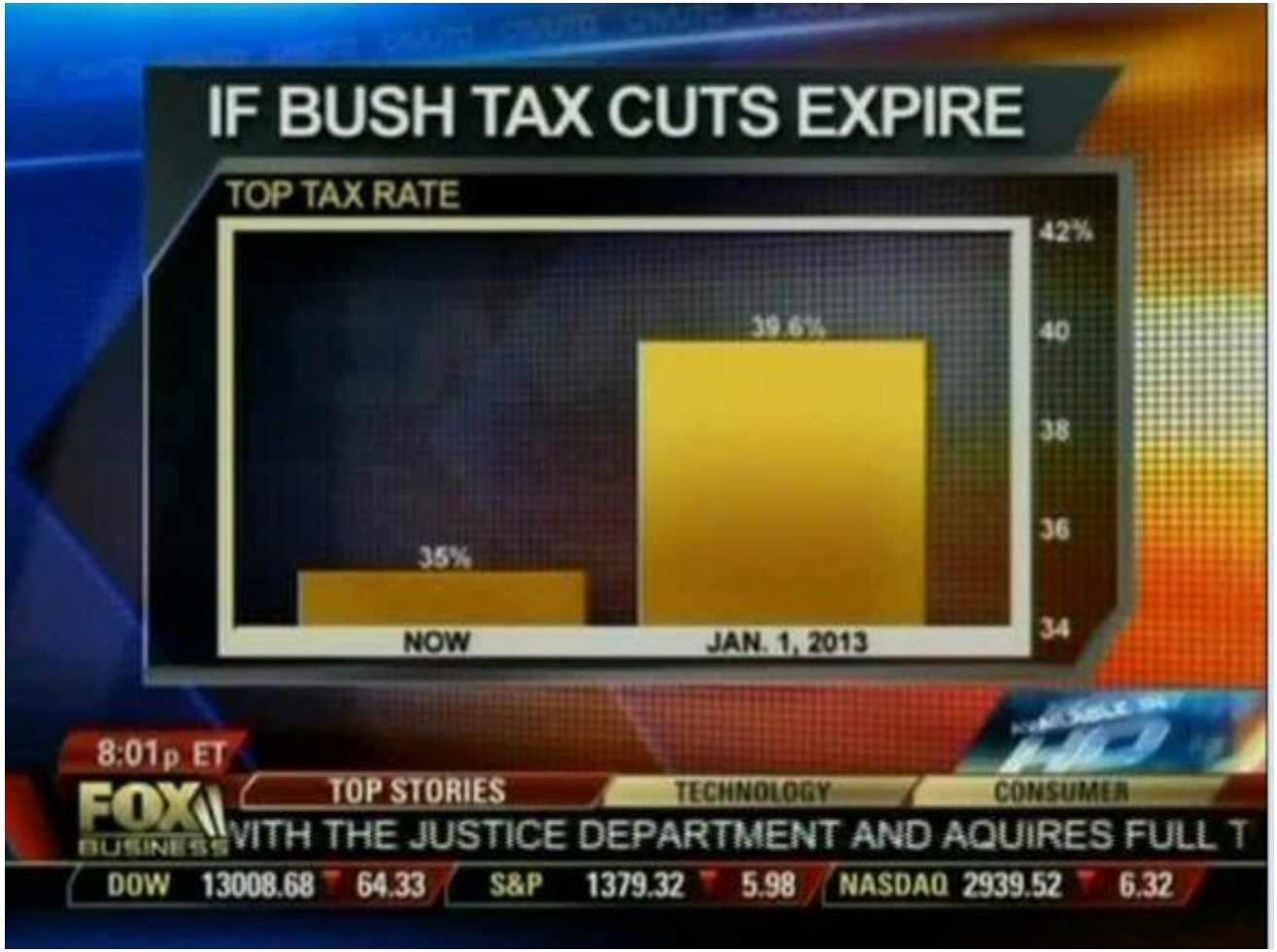
Vhodný i když se procenta nesčítají na 100%



Příklady špatných grafů



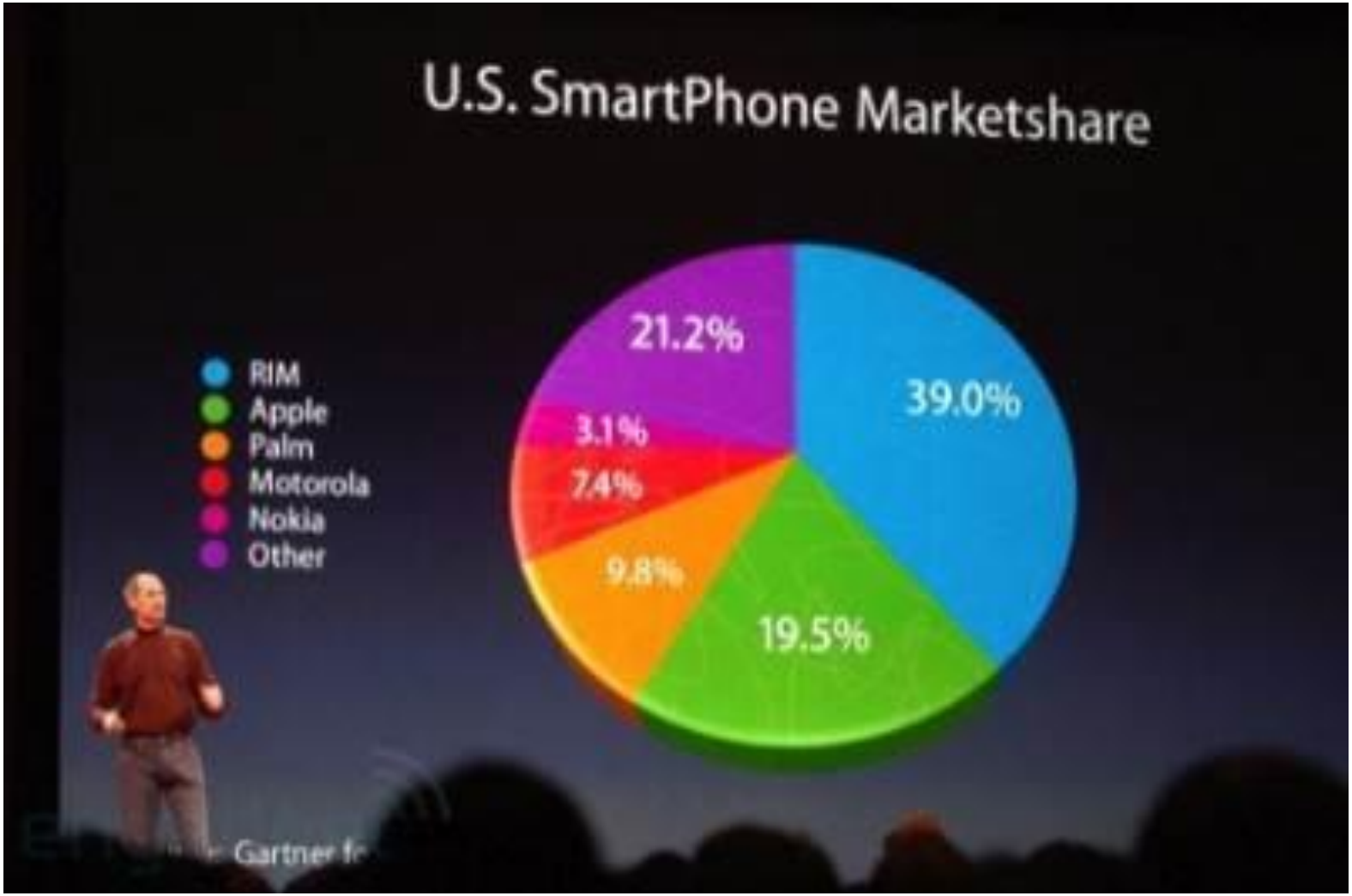
Příklady špatných/klamných grafů



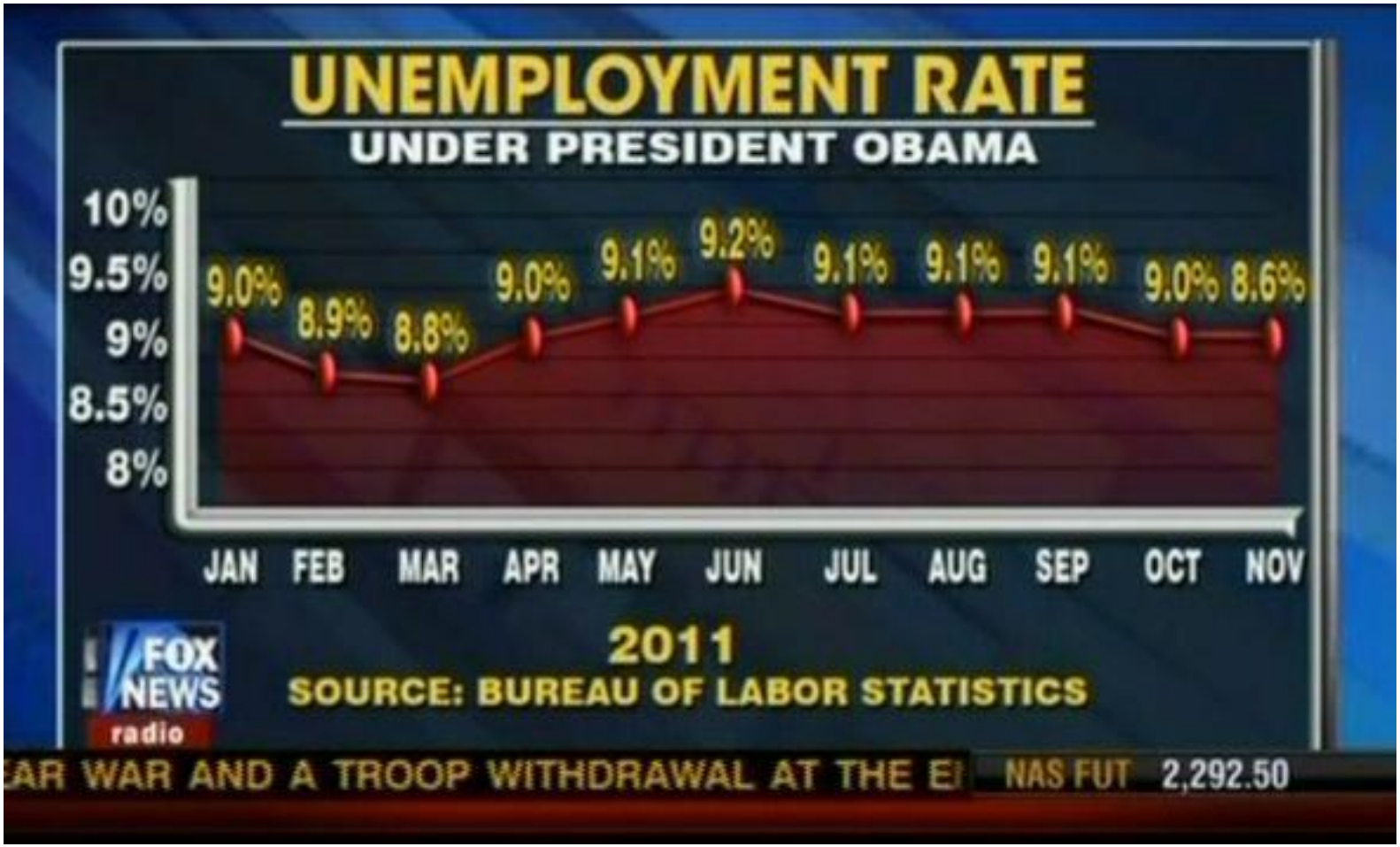
Příklady špatných/klamných grafů

Visits to the 391 sites in the national park system, including the 58 major parks, peaked in 1987 at 287.2 million.

Příklady špatných/klamných grafů



Příklady špatných/klamných grafů



Popisná statistika - Shrnutí

- Úkolem je popsat daný datový soubor
- Důležité je zvolit vhodný způsob popisu dat
  - **Kategoriální data:**
    - Tabulka absolutních a relativních četností
    - Koláčový graf
  - **Numerická data:**
    - Průměr, směrodatná odchylka
    - Minimum, maximum, kvartily, medián
    - Krabicový graf
    - Histogram

DOTAZY?

2.

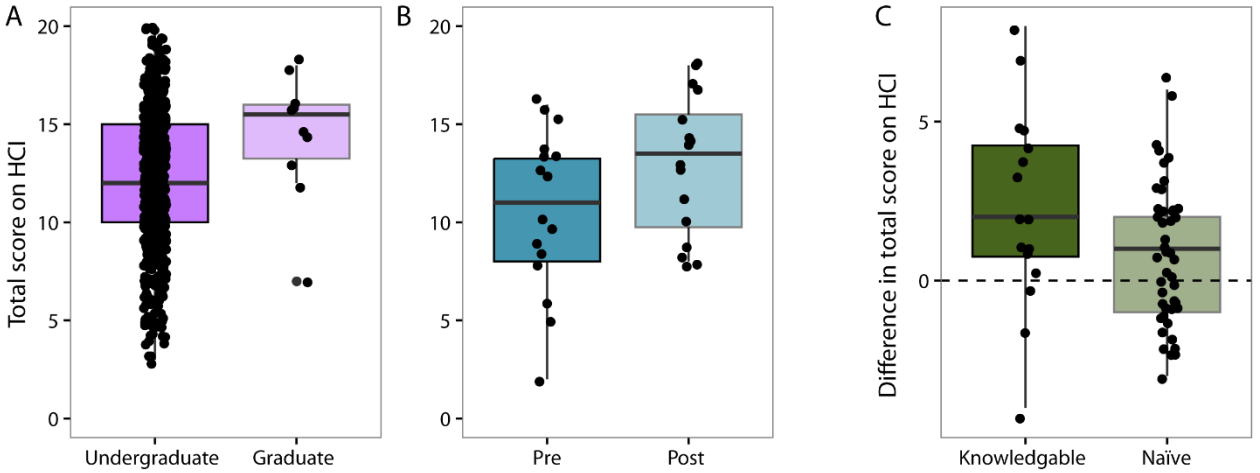
Induktivní statistika

POZOR!!!

O DOST NÁROČNĚJŠÍ!!!

Motivace: Validizační studie HCI testu

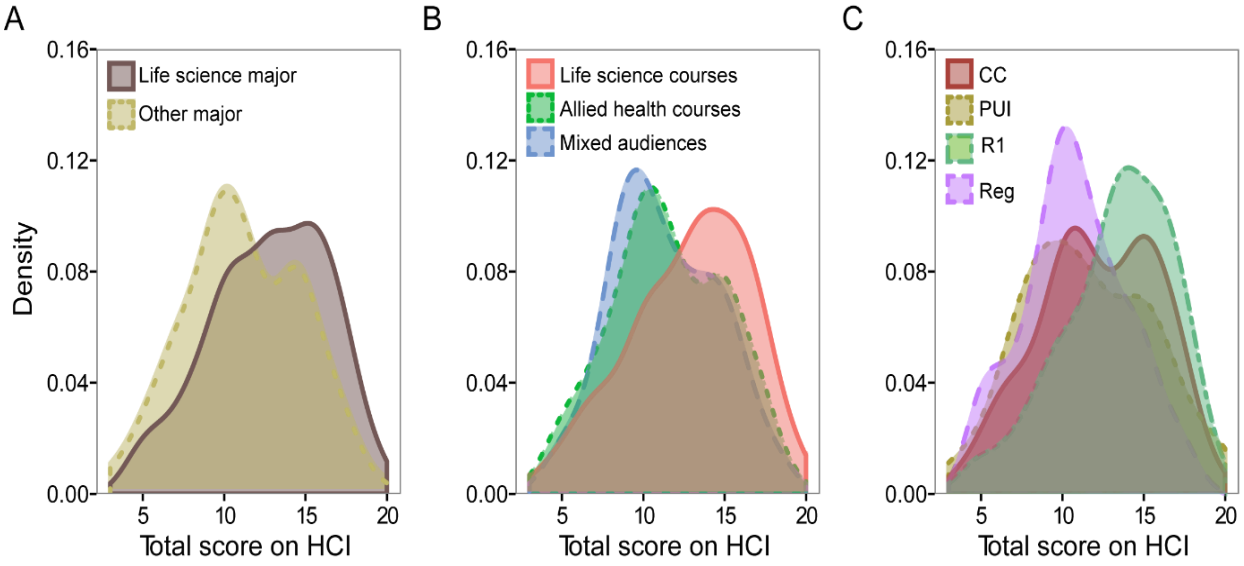
- Dosahují starší studenti (graduate) lepších výsledků než mladší?
- Zlepší se studenti po výuce homeostázy?
- Je toto zlepšení větší, než u studentů bez výuky homeostázy?



(McFarland, Price, Wenderoth, Martinkova et al, *CBE LSE*, in press)
http://physiologyconcepts.org/

Motivace: Validizační studie HCI testu

- Dosahují studenti biologických oborů lepších výsledů?
- Jaký vliv má typ instituce?

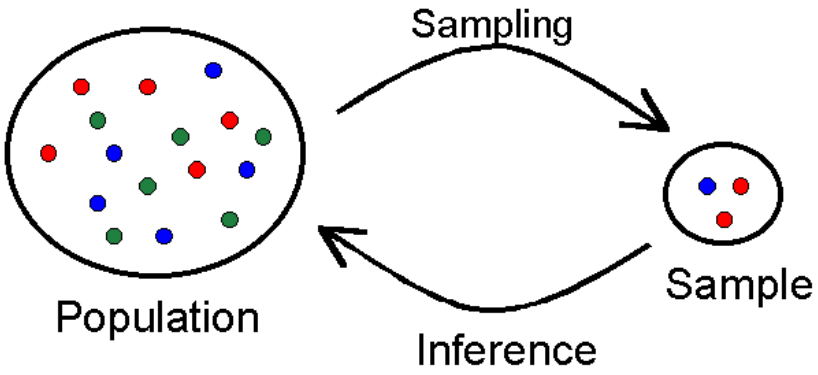


(McFarland, Price, Wenderoth, Martinkova et al, *CBE LSE*, in press)

Induktivní statistika (také „statistická inference“)

Cíl: Zobecnění závěrů z výběru na celou populaci

- **Populace** - soubor všech jednotek (osob, zvířat, …) splňujících dané podmínky
  - zpravidla není celý dostupný
  - bývá příliš velký
- **Výběr** - část populace,
- u níž se „sbírají“ data
  - měl by být reprezentativní



POZOR: Každý zobecňující závěr o populaci učiněný na základě výběrových dat je nutně zatížen nejistotou

2.

Induktivní statistika

2a: Výběr

Výběr – měl by být reprezentativní

Odstrašující příklad 1: Hodnocení lékaře

ma • 2012-08-06 • ★★★★★

Paní doktorku mohu vřele doporučit - je milá a vstřícná (poskytuje i telefonickou konzultaci). Také sestřička je vlídná a je vidět, že to s malými dětmi opravdu umí. Oceňuji objednací systém, který rozděluje nemocné a zdravé pacienty. Dalším pozitivem je pěkně vymalovaná a zařízená čekárna i ordinace.

Gabriela • 2017-02-23 • ★☆☆☆☆

Velmi nedoporučuji, paní doktorka odmítá názor matky, léčila nás na něco, co dítě nemělo. Léčila je silné slovo, dostali jsme letáčky, co si máme koupit, že se dítě nelepší = nekoupili jsme to z letáčku. CRP se platí. Právě jsme se vrátili z nemocnice, léčba od paní doktorky šla proti tomu, co našemu dítěti skutečně je. A to jsem neustále konzultovala zdravotní stav a

Tzv. „Voluntary response sample“
Řešení: např. zadat anketu všem pacientům daného lékaře

Výběr – měl by být reprezentativní

Odstrašující příklad 2: Porovnání skupin

- **Experimentální skupina:**
  - Náhodný výběr nemocných nemocí XY

    (Pozn.: jeden lékařův příbuzný má také nemoc XY)

- **Kontrolní skupina:**
  - Zdraví jedinci, většinou příbuzní a známí lékaře

Výběr – měl by být reprezentativní

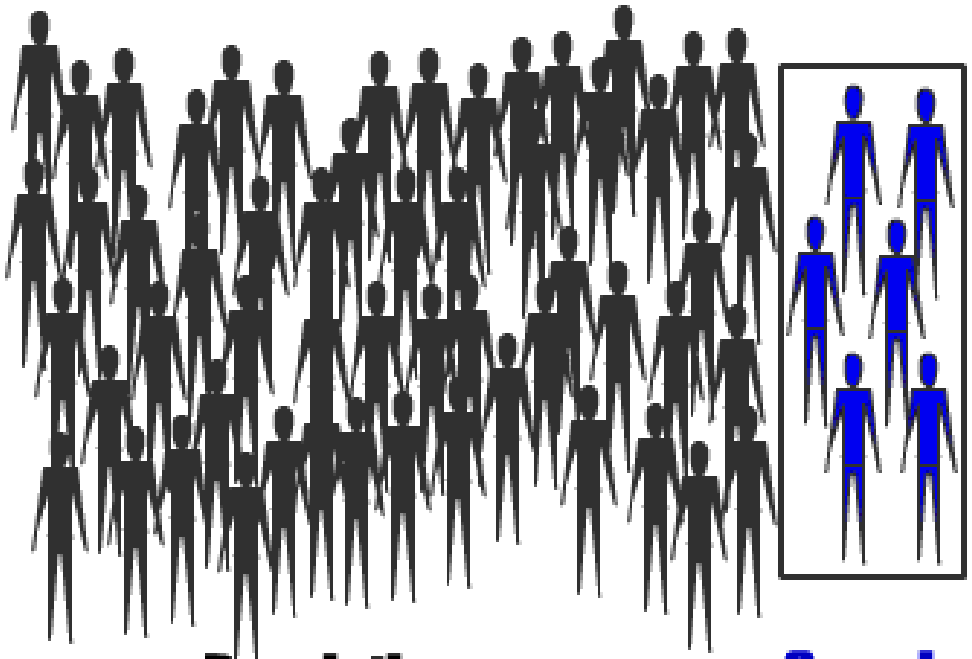
Strategie náhodného výběru:

- Prostý náhodný výběr (simple random sample)
- Stratifikovaný výběr (stratified sample)
- Klastrový výběr (cluster sample)

- **Zajistěte náhodný výběr,** pokud to jen trochu je možné!

- Pokud to není možné, popište výběr, nenáhodnost uveďte jako limitaci studie.

Charakteristiky populace a výběru



Population	Sample
quantity (count) = N	quantity (count) = n
mean = μ	mean = $\overline{x}$
variance = σ^2	variance = s^2
standard deviation = σ	standard deviation = s

2.

Induktivní statistika

2b: Rozdělení výběrového průměru

Rozdělení výběrového průměru

Abychom mohli činit z **výběrového** průměru závěry o **populačním** průměru, potřebujeme vědět, jak se výběrový průměr chová (jaké má rozdělení).

Rozdělení výběrového průměru

Abychom mohli činit z **výběrového** průměru závěry o **populačním** průměru, potřebujeme vědět, jak se výběrový průměr chová (jaké má rozdělení).

Platí (tzv. Centrální limitní věta):

Průměr z náhodného výběru o (dostatečné) velikosti *n* má přibližně **normální rozdělení**, a to

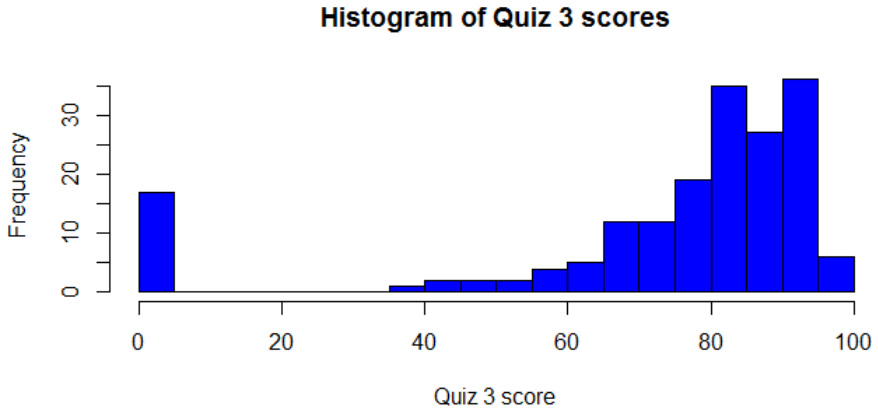
- se stejným průměrem
- s rozptylem n-krát menším

Pozn.: Původní rozdělení ani nemusí být normální!

Skvělá věc, jelikož o **normálním rozdělení** toho hodně víme!

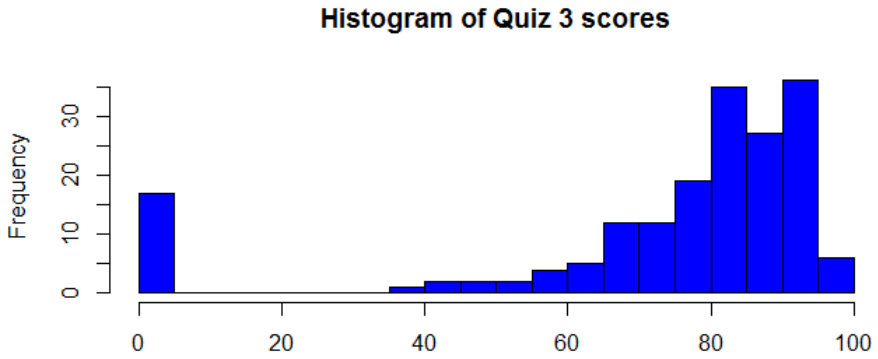
Příklad: průměrné skóre z Quizu č. 3

- Výběr o velikosti n = 1
  - Histogram podobný histogramu původních skóre

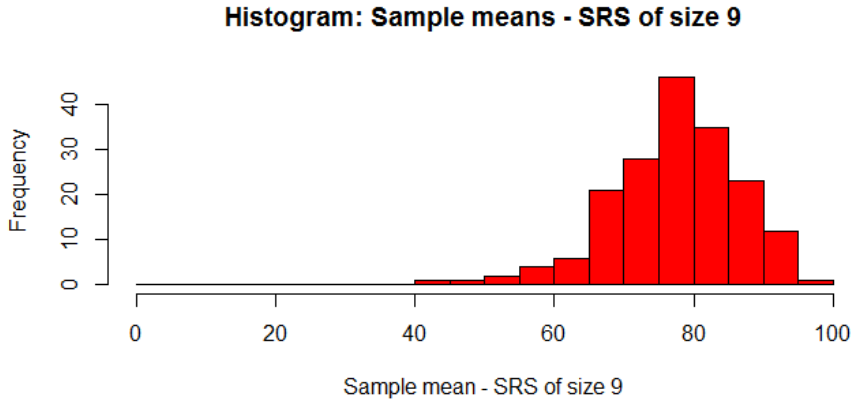


Příklad: průměrné skóre z Quizu č. 3

- Výběr o velikosti n = 1
  - Histogram podobný histogramu původních skóre

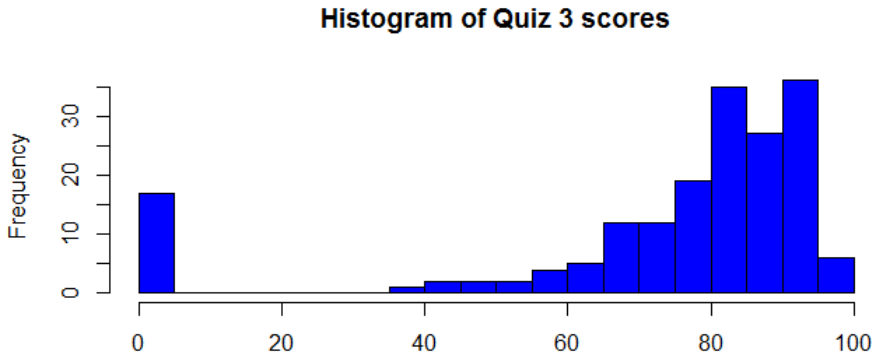


- Výběr o velikosti n = 9
  - Histogram je
    - Více symetrický
    - S menším rozptylem

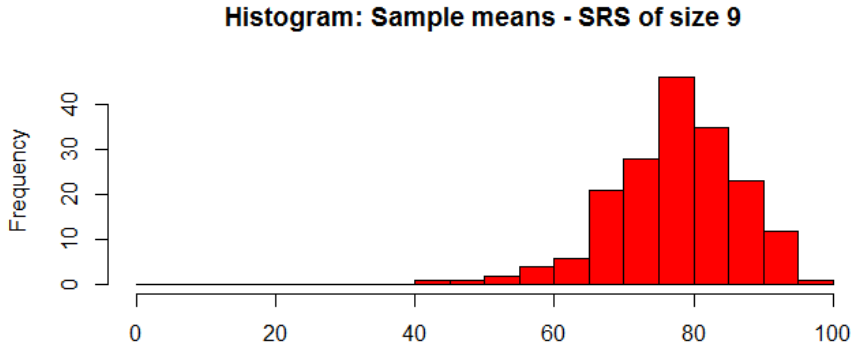


Příklad: průměrné skóre z Quizu č. 3

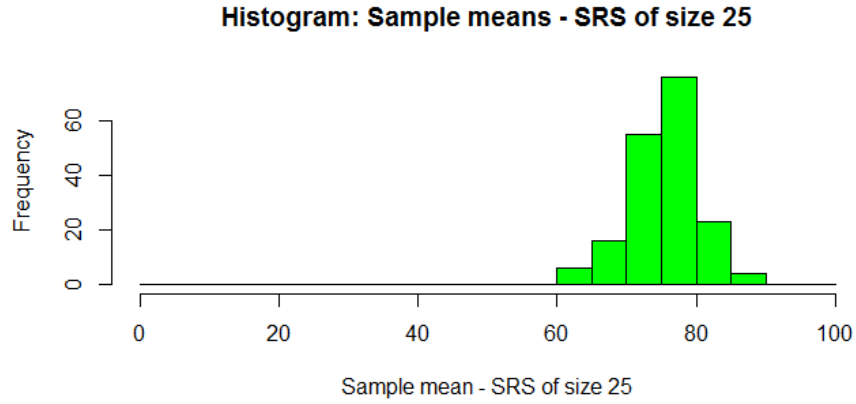
- Výběr o velikosti n = 1
  - Histogram podobný histogramu původních skóre



- Výběr o velikosti n = 9
  - Histogram je
    - Více symetrický
    - S menším rozptylem



- Výběr o velikosti n = 25
  - Histogram je
    - Ještě více podobný normálnímu
    - S ještě menším rozptylem

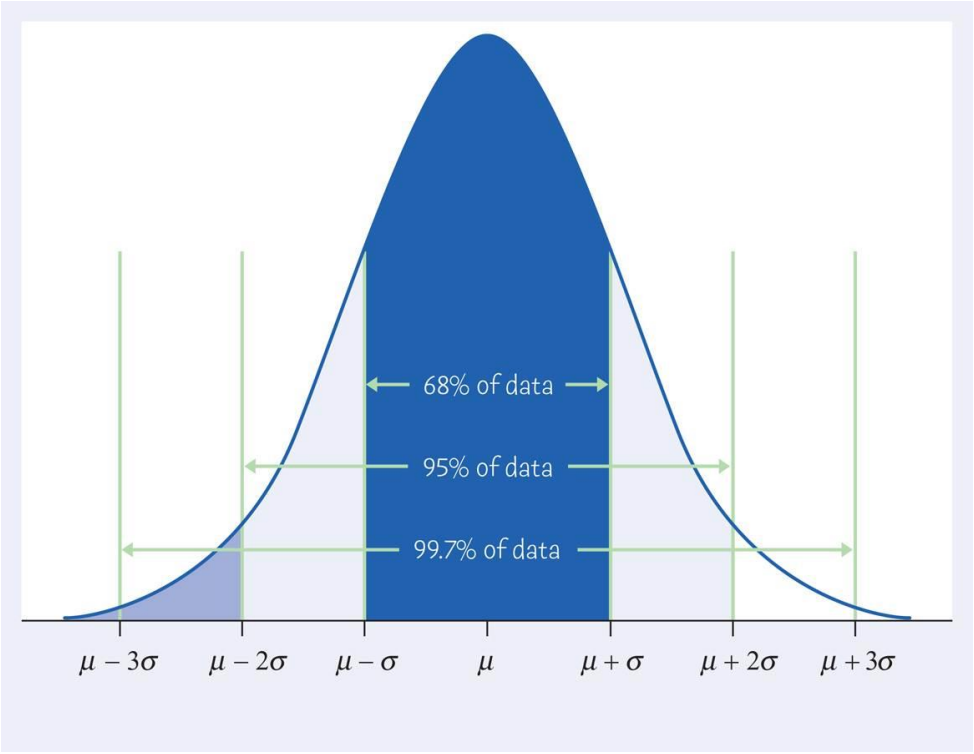


Rozdělení výběrového průměru

Platí (tzv. Centrální limitní věta):

Průměr náhodného výběru o (dostatečné) velikosti *n* má přibližně **normální rozdělení:** $\overline{x} \sim N(\mu, \frac{\sigma^2}{n})$

- se stejným průměrem
- s rozptylem n-krát menším



To je skvělá věc, jelikož o **normálním rozdělení** toho hodně víme:

2.

Induktivní statistika

2c: Konfidenční interval

Bodový odhad vs. Intervalový odhad

Jak odhadnout populační průměr?

Pomocí výběrového průměru:

(Bodový odhad)

- Jako bychom házeli šíp
- Nejspíš mineme

Bodový odhad vs. Intervalový odhad

Jak odhadnout populační průměr?

Pomocí výběrového průměru:

(Bodový odhad)

- Jako bychom házeli šíp
- Nejspíš mineme

Pomocí konfidenčního intervalu:

(intervalový odhad)

- Jako bychom házeli síť
- Nejspíš rybu chytíme!

Teď se naučíme házet síť!

Intervalový odhad (Skóre z testu č. 3)

Úkol: Na základě náhodného výběru o velikosti n=9 studentů spočtěte 95% konfidenční interval pro populační průměr.

Prozradím Vám, že populační směrodatná odchylka je $\sigma = 27$.

1. Střed intervalu: $\bar{x} =$ _____ (Váš výběrový průměr)
2. Poloměr $2 \cdot \frac{\sigma}{\sqrt{n}} =$ _____ (pro všechny stejný!)
3. Dolní mez intervalu: $\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}} =$ _____

   Horní mez intervalu: $\bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}} =$ _____

4. Závěr: Jsem si z 95% jistý, že skutečná hodnota populačního průměru skóre z testu č. 3 leží v intervalu _____ .

Intervalový odhad (Skóre z testu č. 3)

Úkol: Na základě náhodného výběru o velikosti n=9 studentů spočtěte 95% konfidenční interval pro populační průměr.

Prozradím Vám, že populační směrodatná odchylka je $\sigma = 27$.

1. Střed intervalu: $\bar{x} =$ 85.0 (Váš)
2. Poloměr $2 \cdot \frac{\sigma}{\sqrt{n}} = 2 \cdot \frac{27}{\sqrt{9}} = 18$ (Pro všechny stejný!)
3. Dolní mez intervalu: $\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}} = 85.0 - 18 = 67$

   Horní mez intervalu: $\bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}} = 85.0 + 18 = 103$

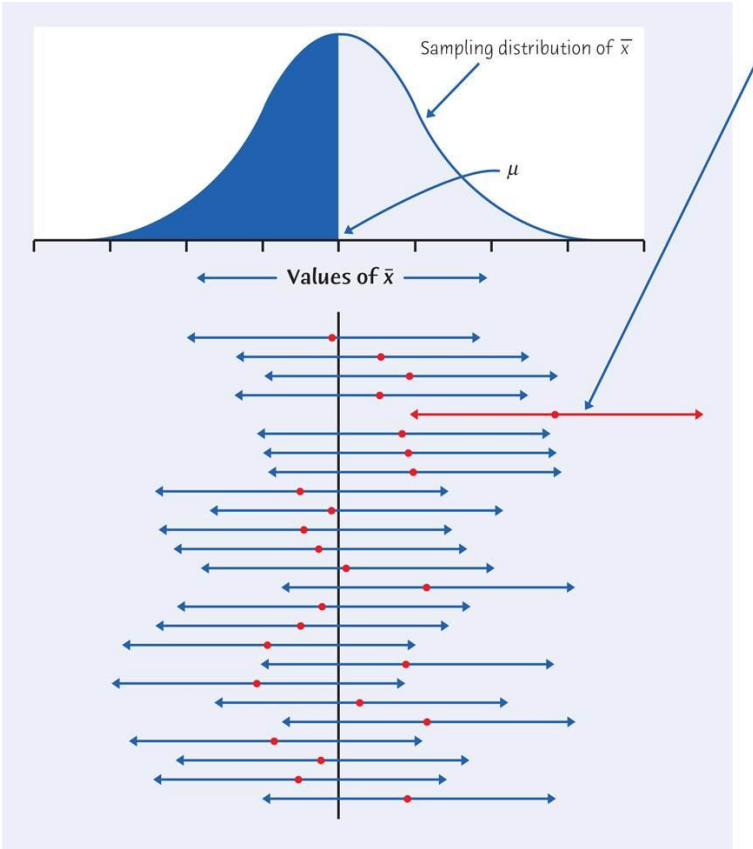
4. Závěr: Jsem si z 95% jistý, že skutečná hodnota populačního průměru skóre z testu č. 3 leží v intervalu (67, 103).

„na 95% jistý…“

Správně řečeno:

V 95% případů, kdy budeme dělat náhodný výběr o stejné velikosti (n = 9) z této populace, náš interval pokryje skutečnou hodnotu populačního průměru:

- Představte si, že mnohokrát provedeme výběr o velikosti n = 9 a spočítáme interval s využitím rovnice $\bar{x} \pm 2 \cdot \frac{\sigma}{\sqrt{n}}$
- Pak přibližně v 95% případů tyto intervaly pokryjí skutečný populační průměr μ



Zkuste si: https://jrnold.shinyapps.io/connt/

Intervalový odhad (Skóre z testu č. 3)

... a nyní Vám prozradím, že

skutečný populační průměr z testu byl **75.5**

Pokrývá Váš konfidenční interval tuto hodnotu?

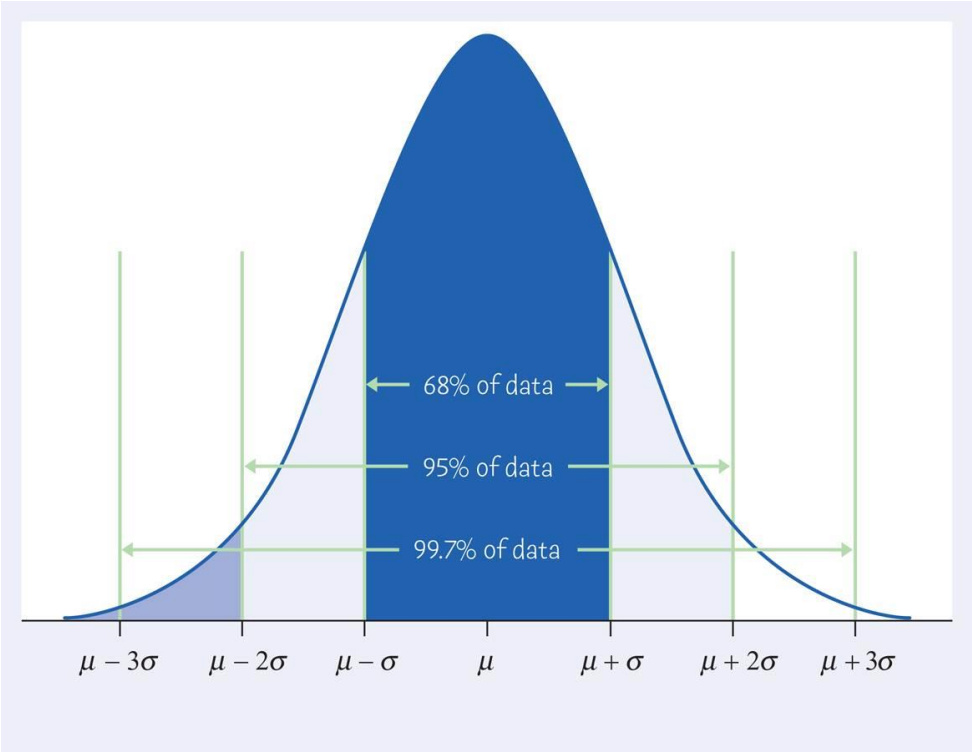
1. Ano, můj konfidenční interval pokrývá hodnotu 75.5
2. Ne, můj konfidenční interval nepokrývá hodnotu 75.5

Kolik z vás odpoví „Ne“?

Intervalový odhad

- Jaký bude konfidenční interval při větším vzorku?
  - Užší nebo širší?

- Který konfidenční interval bude užší
  - 95% nebo 99%?



2.

Induktivní statistika

2d: Testování hypotéz

Testování hypotéz (Příklad „Skóre z testu“)

Někdo Vám tvrdí, že (populační) průměrné skóre z testu je 89bodů. Chcete to otestovat.

Hypotéza (tzv. **nulová hypotéza**) $H_0: \mu = 89$

Testování hypotéz (Příklad „Skóre z testu“)

Někdo Vám tvrdí, že (populační) průměrné skóre z testu je 89bodů. Chcete to otestovat.

Hypotéza (tzv. **nulová hypotéza**) $H_0: \mu = 89$

Chcete, aby tzv „chyba I. druhu“, tj. to že hypotézu zamítnete, když přitom platí, byla malá, <0.05

Testování hypotéz (Příklad „Skóre z testu“)

Někdo Vám tvrdí, že (populační) průměrné skóre z testu je 89bodů. Chcete to otestovat.

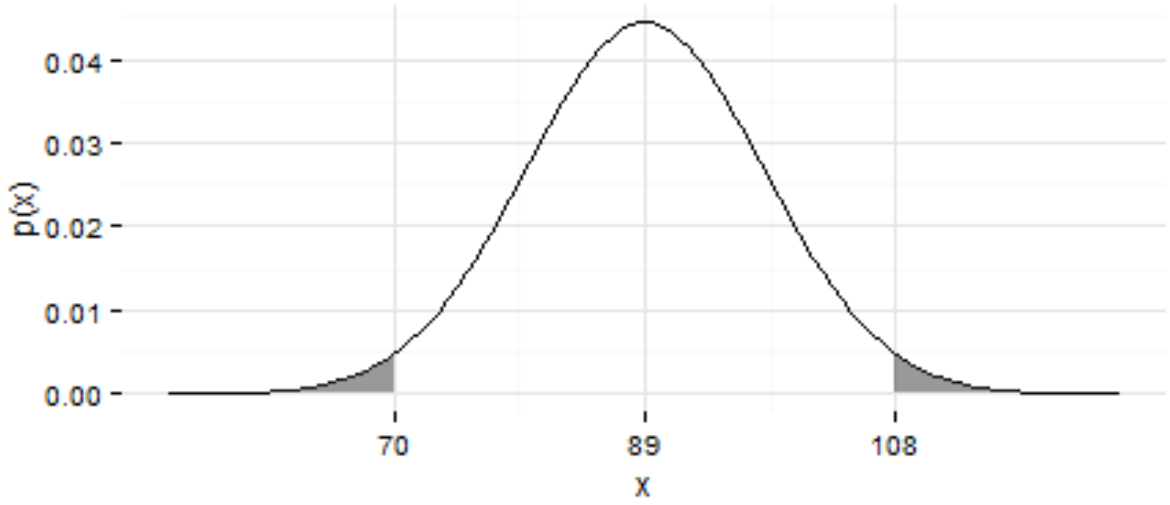
Hypotéza (tzv. **nulová hypotéza**) $H_0: \mu = 89$

Chcete, aby tzv „chyba I. druhu“, tj. to že hypotézu zamítnete, když přitom platí, byla malá, <0.05

Za tím účelem jste se dotázali náh. výběru 9 studentů, průměr jejich skóre je $\bar{x} = 70$. Již víte, že populační směrodatná odchylka je 27. Co lze o populačním skóre říci?

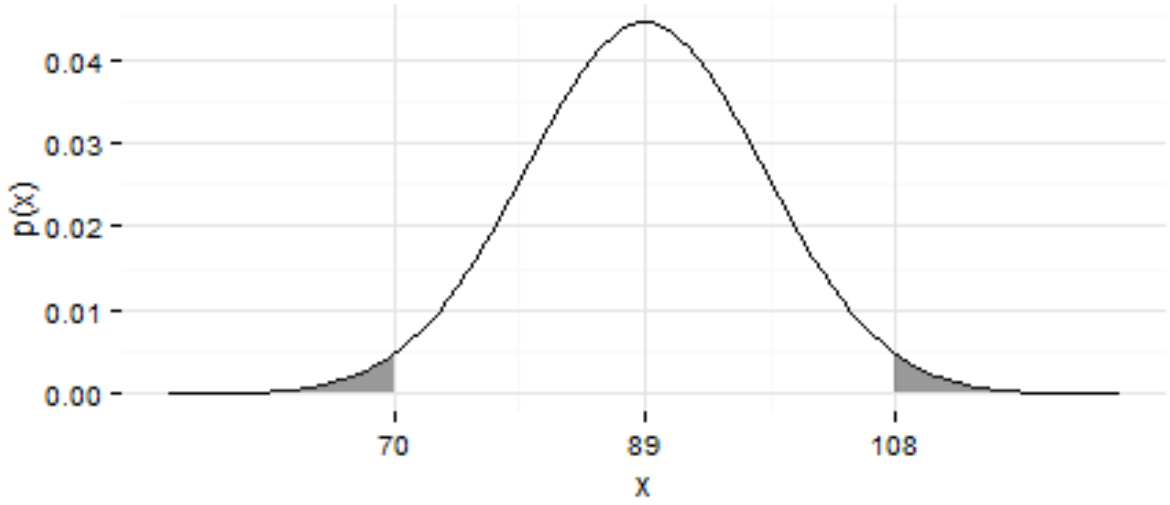
Testování hypotéz (Příklad „Skóre z testu“)

Pokud toto tvrzení (tzv. nulová hypotéza: H_0: $\mu = 89$) platí, pak $\bar{x}$ má rozdělení:



Testování hypotéz (Příklad „Skóre z testu“)

Pokud toto tvrzení (tzv. nulová hypotéza: H_0: $\mu = 89$) platí, pak $\bar{x}$ má rozdělení:



p hodnota: Jak je pravděpodobný tento výsledek (70)

nebo výsledek víc svědčící proti H_0 („horší“ výsledek)?

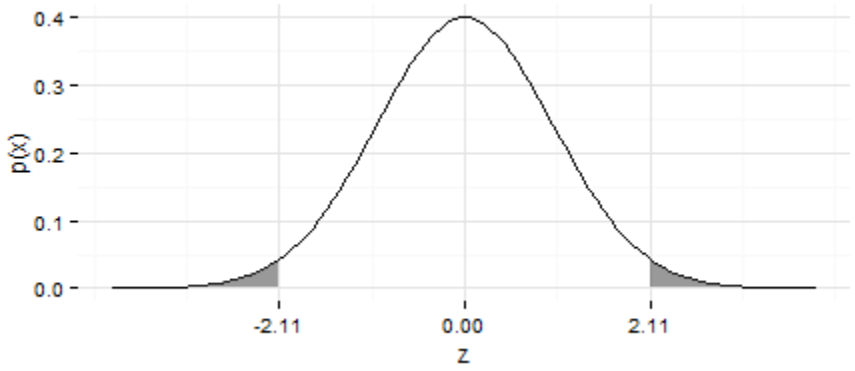
p<0.05, takže zamítáme H_0.

Testování hypotéz („Skóre z testu“)... jinak

1. Zformulujte tzv. nulovou hypotézu: $H_0: \mu = 89$
   tzv. alternativní hypotézu: $H_1: \mu \neq 89$

2. Spočtěte z dat **testovou statistiku** $Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{70-89}{27/\sqrt{9}} = -2.11$

ta má za platnosti H_0
normální rozdělení
s nulovou střední hodnotou
a rozptylem rovným jedné.

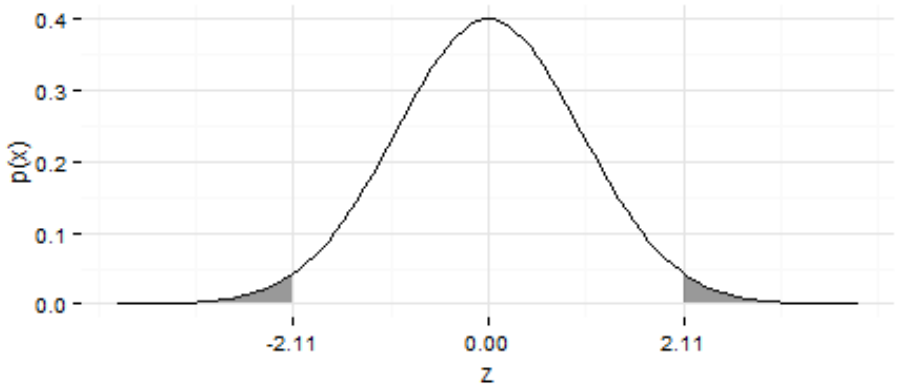


Testování hypotéz („Skóre z testu“) … jinak

1. Zformulujte tzv. nulovou hypotézu: $H_0: \mu = 89$
   tzv. alternativní hypotézu: $H_1: \mu \neq 89$

2. Spočtěte z dat **testovou statistiku** $Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{70-89}{27/\sqrt{9}} = -2.11$

   ta má za platnosti H_0
   normální rozdělení
   s nulovou střední hodnotou
   a rozptylem rovným jedné.



3. Porovnejte s kritickou hodnotou, zde 1.96 (přibližně 2).
   Hodnota testové statistiky je v absolutní hodnotě
   větší než 1.96, tudíž na 5% hladině zamítáme H_0.

Testování hypotéz („Skóre z testu“)

Ještě jinak:

- Zamítnutí H_0: $\mu = 89$ na hladině 5%
  souvisí s tím, že 95% konfidenční interval je (52, 88),
  tj. nepokrývá hodnotu 89.

Poznámka:

- Pro větší n by byl konfidenční interval užší
  a při stejném $\bar{x}$ by evidence proti H_0 byla silnější.
  (*p* hodnota by byla ještě menší)

Další testy

- **(Jednovýběrový) t test** v případě neznámé populační směrodatné odchylky
- **Dvouvběrový t test** v případě porovnávání průměru dvou skupin
- **F test** analýzy rozptylu pro porovnání průměru více skupin, aj.

… PRINCIP TESTOVÁNÍ JE ALE STÁLE STEJNÝ

3.

Závěr

Obsah (požadavky k zápočtu)

Popisná statistika

- kategorická a numerická data
- četnosti (absolutní, relativní, kumulativní)
- aritmetický průměr, rozptyl, směrodatná odchylka
- medián, modus, kvartil, percentil, krabicový graf
- grafická a tabulková prezentace statistických dat, histogram

Induktivní statistika

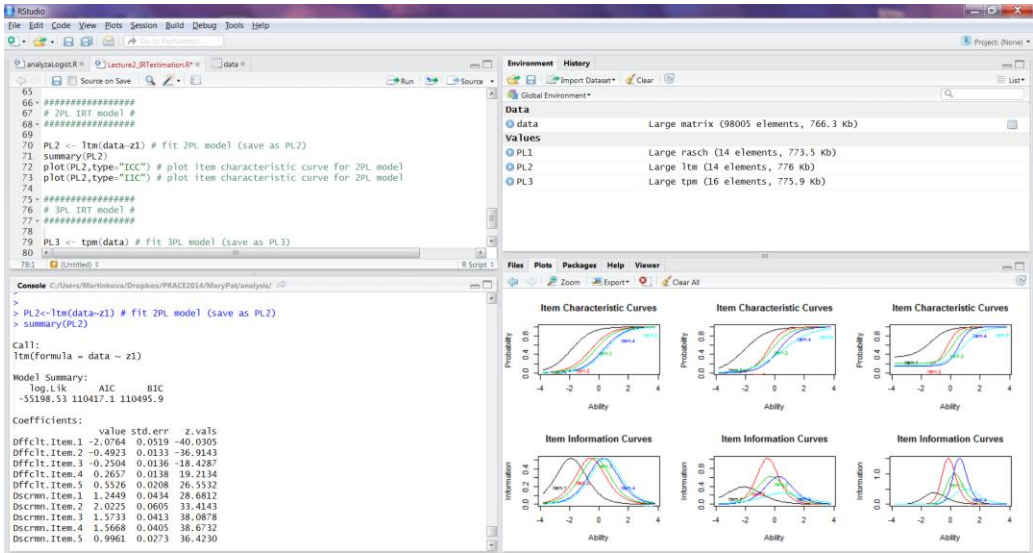
- náhodný výběr, náhodný jev
- rozdělení pravděpodobnosti (binomické, normální /Gaussovo)
- intervalový odhad
- testování hypotéz (princip a význam).

Zdroje pro tuto přednášku

- Jana Zvárová: Biomedicínská statistika I. Základy statistiky pro biomedicínské obory. Karolinum, 2016.
- Karel Zvára: Biostatistika. Karolinum, 2008.
- Karel Zvára: Biomedicínská statistika IV. Základy statistiky v prostředí R. Karolinum, 2013.

Software:

- Leccos lze spočítat v Excelu
- $tatistica, $P$$, A, ...
- **Statistické prostředí R**



Děkuji za pozornost!

martinkova@cs.cas.cz

www.cs.cas.cz/martinkova/