

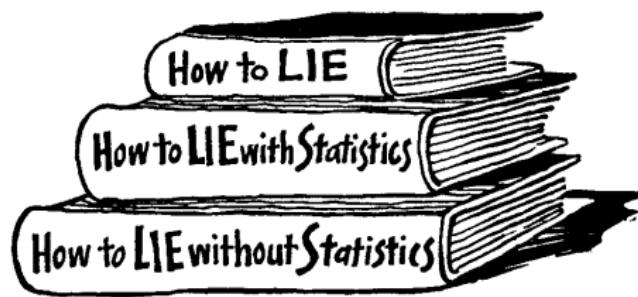
Úvod do (popisnej) štatistiky a spracovania dát (I. časť)

... alebo nenechajme sa oklamáť!

Andrej Gajdoš

Ústav matematických vied, PF UPJŠ, Košice

ZS 2021/2022



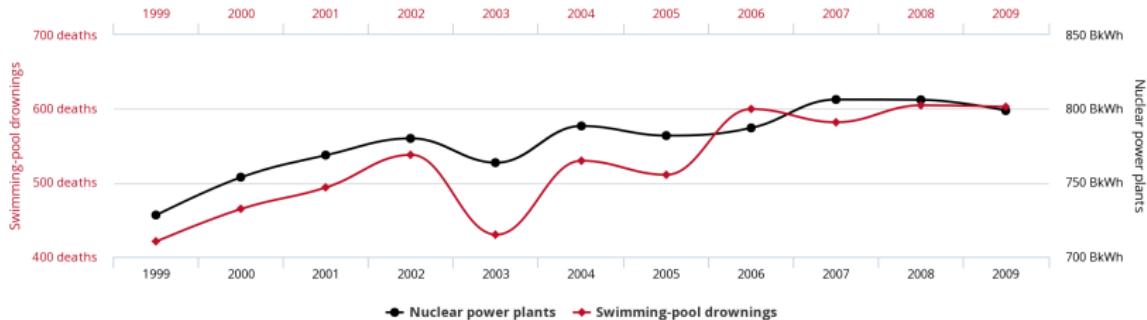
Nie je 50% ako 50%

"There are three types of lies – lies, damn lies, and statistics."
(Benjamin Disraeli)

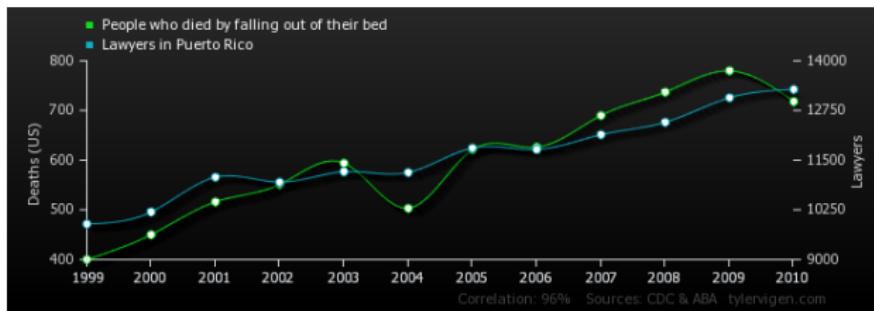


Klamlivá závislosť?!

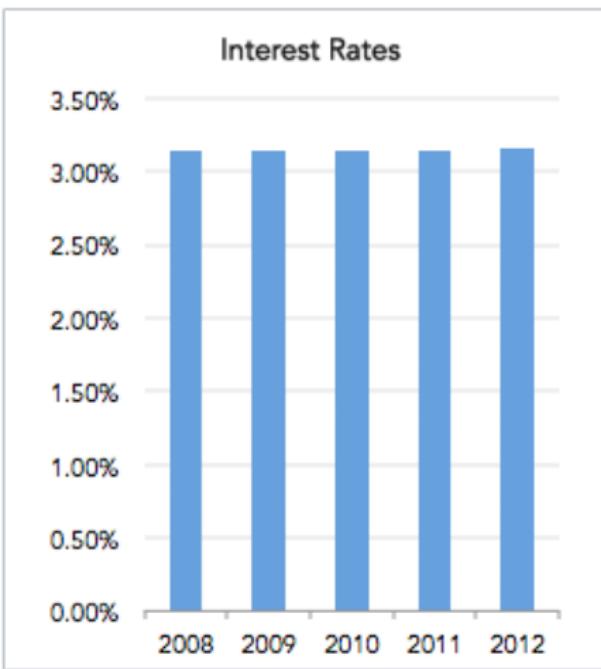
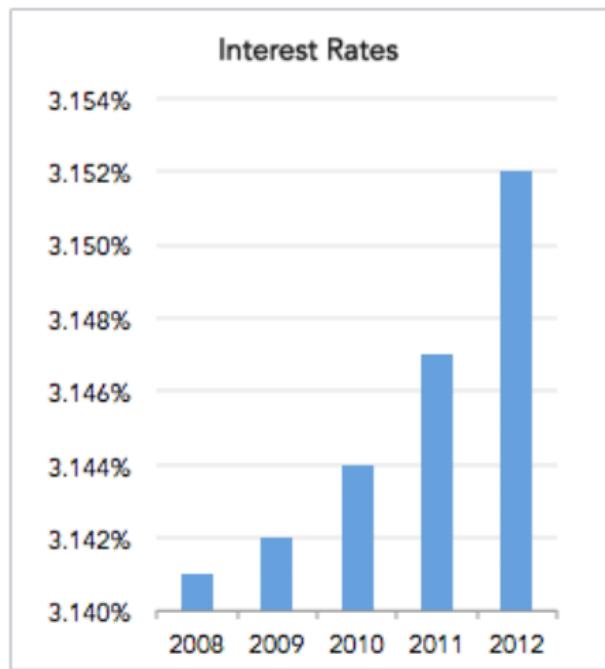
Number people who drowned while in a swimming-pool
correlates with
Power generated by US nuclear power plants



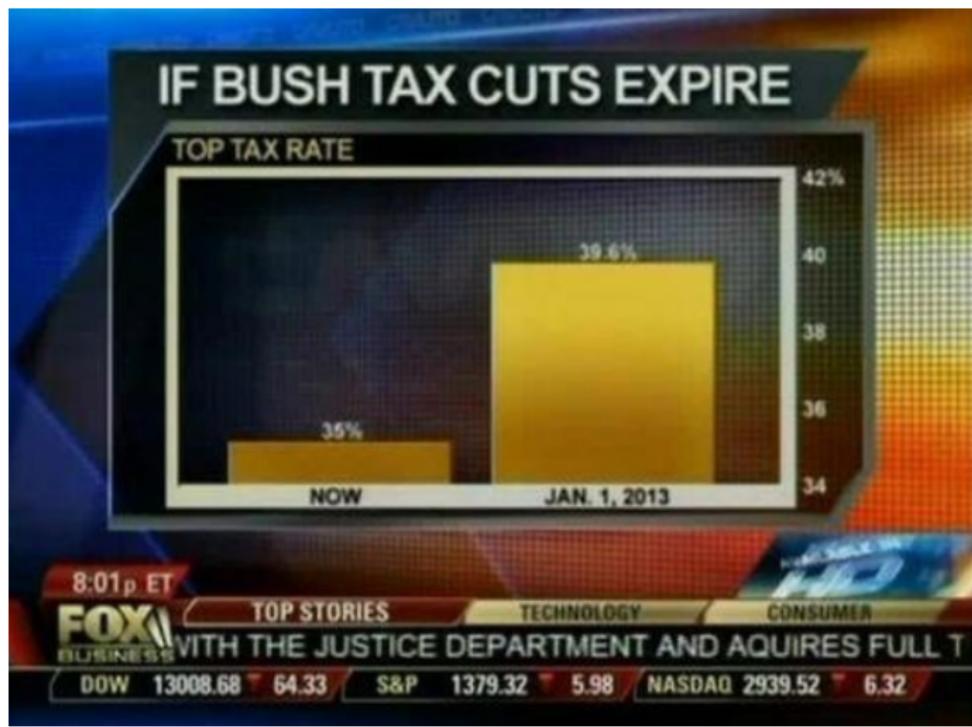
tylervigen.com



Rovnaké dáta, rôzne grafy?!



Nič nie je také ako vyzerá.



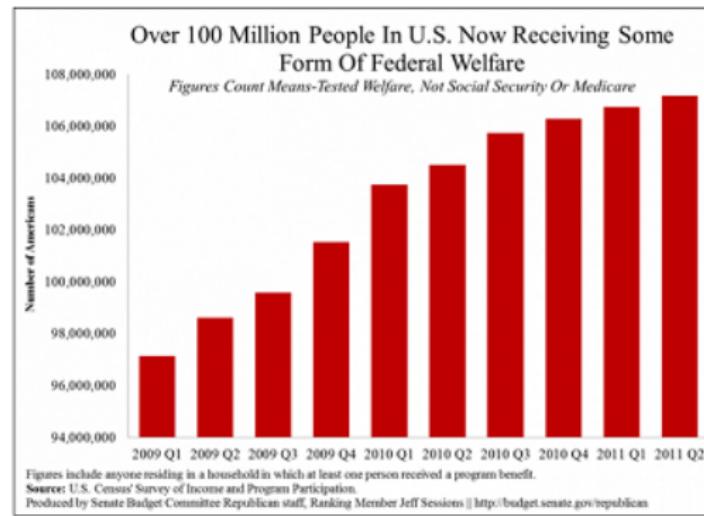
THE BLOG

Over 100 Million Now Receiving Federal Welfare

2:40 PM, AUG 6, 2012 • BY DANIEL HALPER 

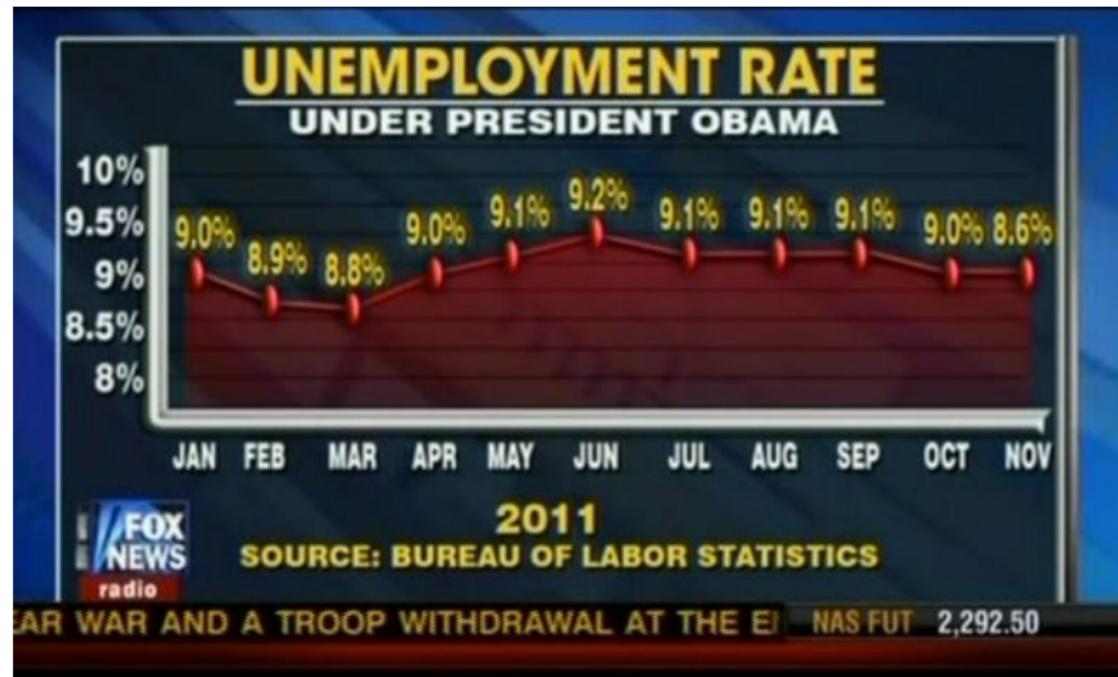
A new chart set to be released later today by the Republican side of the Senate Budget Committee details a startling statistic: "Over 100 Million People in U.S. Now Receiving Some Form Of Federal Welfare."



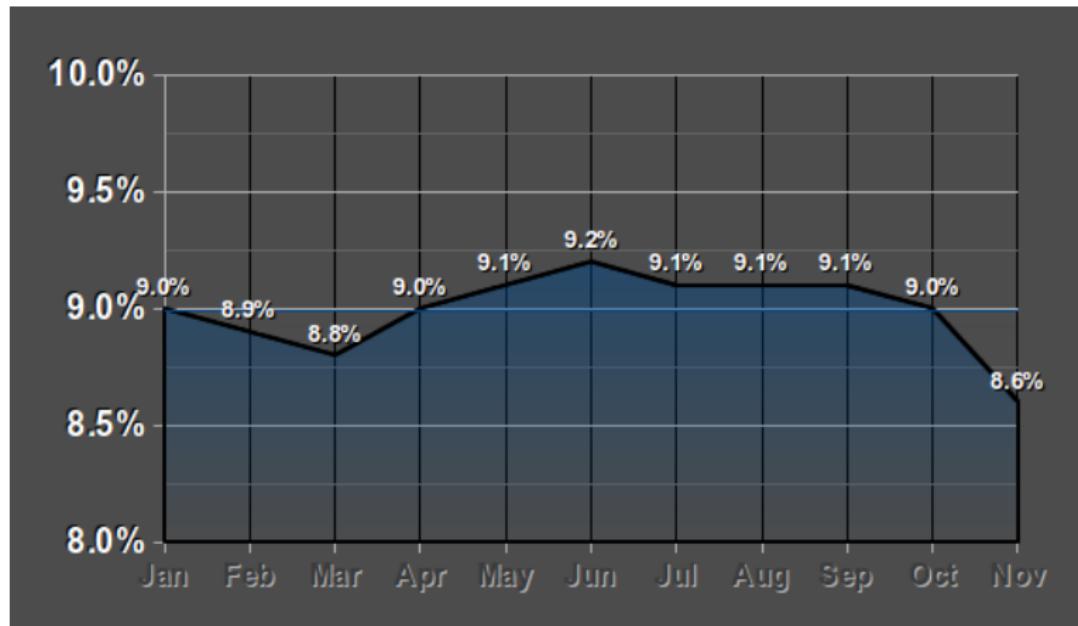
Ako mohol vzniknúť taký graf?!



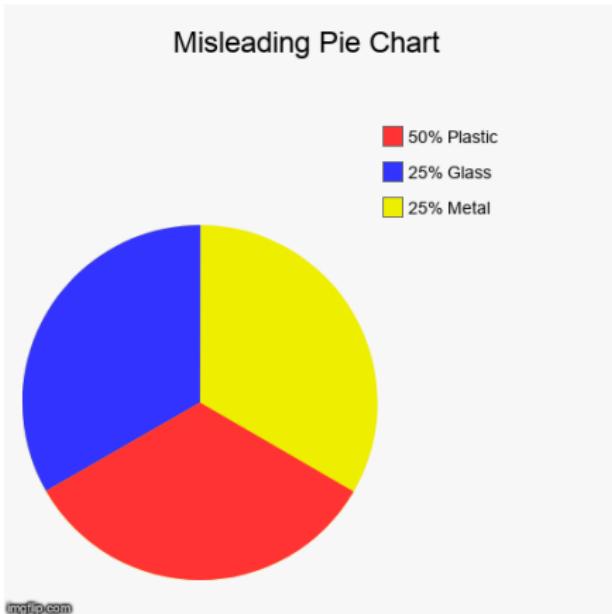
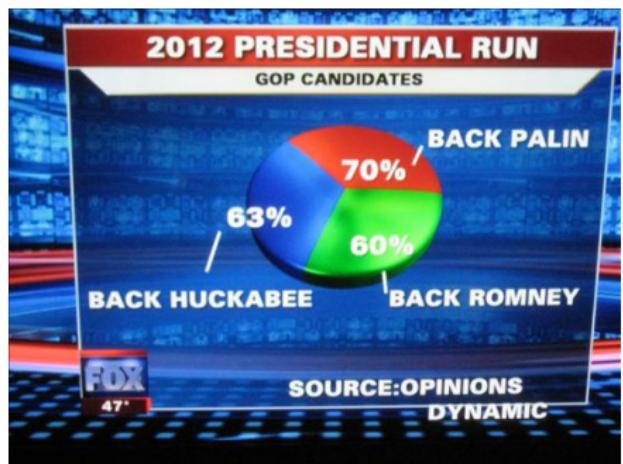
Kde je chyba?



Správny graf.

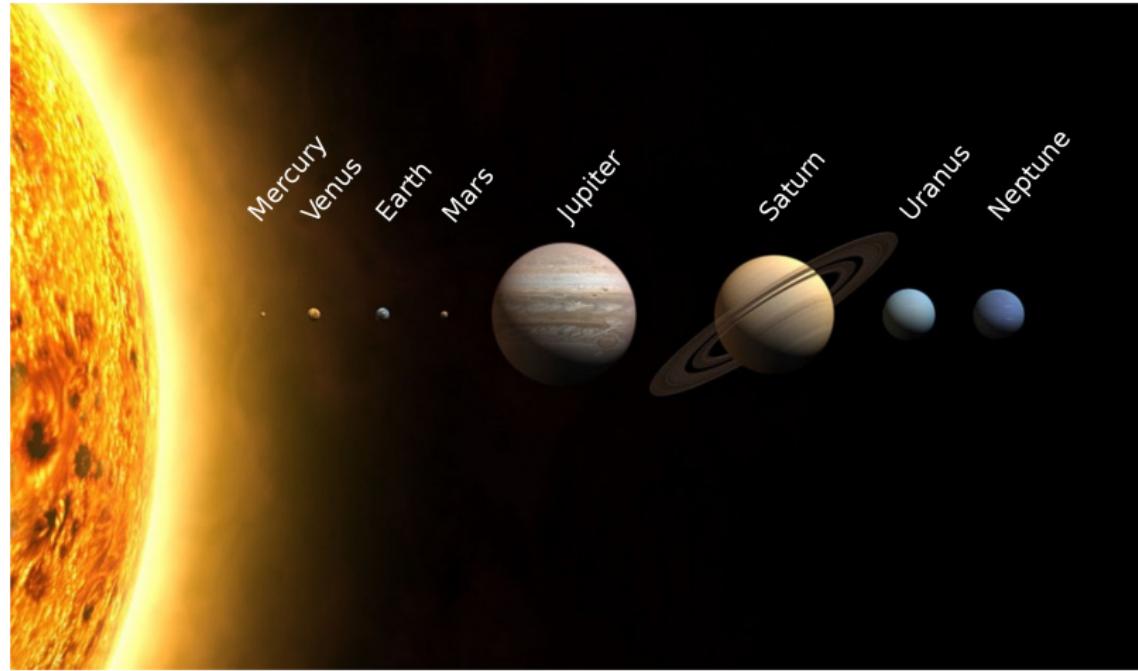


Ked' koláče nechutia tak, ako vyzerajú.

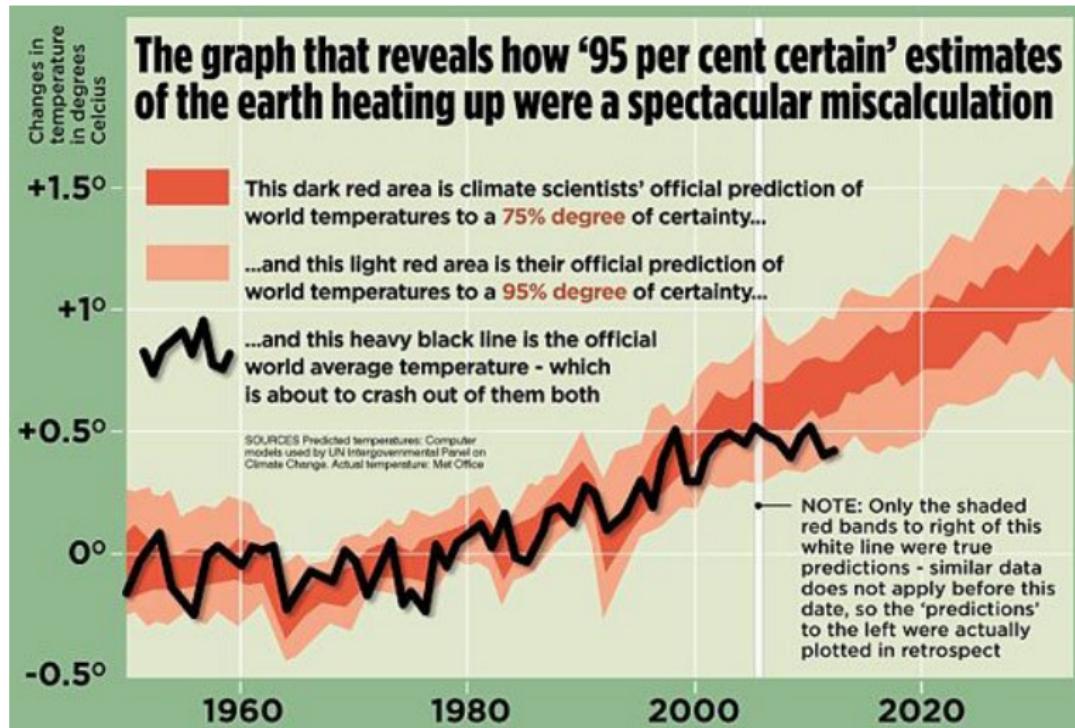


Relativita.

Planéta Mars má polomer 3389 km. Je to veľa alebo málo?



Koniec globálneho otepľovania?



Viac ako 80% stomatológov odporúča čokoľvek.



Prečo sme tu?

- pripomienutie resp. oboznámenie sa so základnými pojмami (matematickej) štatistiky, spracovania dát
- objasnenie vzťahov medzi dôležitými konceptmi
- získanie softvérových zručností pre základné spracovanie dát a robenie jednoduchých dátových analýz
- kritické mysenie
- predchádzanie situáciám z predoších príkladov 😊



Vybrané aplikácie štatistiky.

- reprezentáca, spracovanie a vizualizácia dát za účelom **zefektívnenia výrobných procesov v priemysle**
- **analýza tvarov** v prírode pri skumaní evolúcie (zistovanie účinku rôznych vplyvov) alebo pri rozpoznávaní ľudských tvári (bezpečnostné sýstemy)
- **prieskumy** - sledovanosť TV, preferencie politických strán
- **modelovanie vývoja a predikcie cien** na trhoch, poistovníctvo
- **klasifikácia spamov** (v mailoch), Google prekladač
- **v strojovom (i hlbokom) učení** pri diagnostike ochorení, pri rozpoznávaní dopravných značiek až po samojazdiace autá
- veda, experimenty, testovanie liekov a liečebných postupov
- ...

Na aké konkrétné otázky môže dať odpovede štatistika?

- Ako je možné vyhodnotiť dôkazy o globálnom otepľovaní?
- Sú mobilné telefóny nebezpečné pre naše zdravie?
- Aká je pravdepodobnosť, že vyhráte v lotérii?
- Existuje zaujatosť voči ženám pri menovaní manažérov?
- Koľko "horúcich sérií" ("hot streaks") by ste mohli očakávať v basketbale?
- Ako môžete zistiť, či diéta skutočne funguje?
- Ako sa dá predpovedať predajná cena domu?



Čo je to vlastne tá štatistika?

- štatistika - z latinského slova "*status*"(štát, stav)
- **vedná disciplína**
 - ◊ zaoberá sa vysvetľovaním metód skúmania a vyhodnocovania štatistických dát
 - ◊ súbor matematických metód, ktoré významným spôsobom pomáhajú robiť rozhodnutia v situáciách, kde vzniká neistota kvôli náhode
 - ◊ náhodu pritom popisuje štatistika pomocou pojmu pravdepodobnosť
- **súhrn** demografických a ekonomických **údajov** (čísel), napr. obraz o hospodárskom, politickom stave štátu/firmy ...
- **konkrétny ukazovateľ** (náhodná veličina) - aritmetický priemer, testovacia štatistika, ...

Niekoľko základných pojmov.

- **hromadný jav** - každý prírodný alebo spoločenský jav, ktorý skúmame na veľkom počte objektov (prípadov). Ide o také javy (predmety, udalosti, procesy), ktoré sú výsledkom pôsobenia veľkého množstva príčin a ich vlastnosti sa neprejavujú v jednotlivých javoch, ale v ich súbore. Napr. platové rozdiely (muži vs ženy, medzi mužmi, medzi ženami), úroveň dosiahnutého vzdelania, tvarová rozmanitosť kvetov, ...
- **štatistická jednotka** – je základnym objektom štatistického skúmania; na nej sledujeme konkrétnie **znaky (premenné)**, charakteristiky alebo javy; štatistická jednotka môže byť napr. človek, zviera, technické zariadenie/prístroj, auto, kvet, ...

Niekoľko základných pojmov.

- **statistický súbor** – množina štatistických jednotiek, z ktorých každá splňa určité vlastnosti spoločné pre všetky jednotky v súbore.

Vymedzujeme ho z hľadiska:

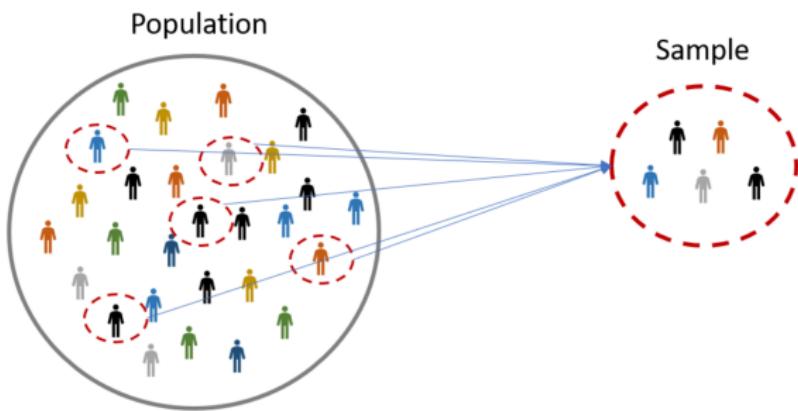
- ◊ **priestorového** - príslušnosť štatistickej jednotky k určitému miestu alebo územiu, napr. obyvateelia istého štátu, či mesta a pod.
- ◊ **časového** - určenie časového úseku, v ktorom sa jednotky zahŕňajú do skúmania, napr. všetky deti narodené v roku 1990 na Slovensku a pod.
- ◊ **vecného** - stanovenie určitých vlastností, ktorými musí disponovať každá štatistická jednotka, napr. novorodenec, zamestnanec istej firmy a pod.

Niekoľko základných pojmov.

- **základný súbor (populácia)** – množina objektov (napr. osôb, zvierat, obcí atď.), ktoré sú predmetom nášho záujmu a o ktorých sa majú robiť závery. Je to súbor všetkých štatistických jednotiek ktoré z hľadiska vecného, časového a priestorového vymedzenia do súboru patria. Môže mať konečný rozsah (ako tomu je zvyčajne u demografických populácií), ale i nekonečný rozsah (hypotetické populácie vymedzené vlastnosťami svojich prvkov).
- obvyklé prekážky pri skúmaní celej populácie:
 - ◊ obmedzenosť zdrojov – zistovanie verejnej mienky: nie je čas ani peniaze pýtať sa všetkých
 - ◊ vzácnosť – zistovanie dedičnosti - jednovaječné dvojčatá sú ideálne, ale vzácné
 - ◊ deštruktívnosť testovania – životnosť žiarovky sa zistí len jej zničením - nedá sa testovať celá populácia

Niekoľko základných pojmov.

- **výberový súbor (náhodný výber/vzorka)** – výber určitej veľkosti n zo základného súboru veľkosti N (t.j. podmnožina základného súboru (populácie)). n - počet statistických jednotiek zaradených do výberu t.j. **veľkosť vzorky** [► simulácia](#)
- **reprezentatívnosť náhodného výberu** – základná požadovaná vlastnosť, aby sa závery štatistického spracovania dali zovšeobecniť na celú populáciu, nie len na jej časť; znamená, že všetky štatistické jednotky z celej populácie majú rovnakú šancu byť vybrané do náhodného výberu; napr. môže byť chybou oslovovať na ulici ľudí (budeme si vyberať len dobre vyzerajúcich), príp. oslovovať len mladých ľudí a závery zovšeobecňovať na všetkých



Základný súbor vs výberový súbor.

Otázka 1

Vykonávate prieskum medzi ľudími v Spojenom kráľovstve, aby ste zistili, ako populárne sú raketové športy. Náhodne si vyberiete ľudí, ktorým zavoláte, a uskutočníte 1 000 telefonických hovorov s ľudími roztrúsenými po celej krajine. Aký je v tejto štúdii štatistický výraz pre ľudí v celej Británii a aký je štatistický výraz pre ľudí, ktorým ste volali?

- A) Ľudia v celom Spojenom kráľovstve aj ľudia, ktorým ste volali, je možné označiť pojmom POPULÁCIA.
- B) Ľudia vo Veľkej Británii sú POPULÁCIA a ľudia, ktorým ste volali, sú VZORKA.
- C) Ľudia vo Veľkej Británii sú VZORKA a ľudia, ktorým ste volali, sú POPULÁCIA.
- D) Ľudí z celej Veľkej Británie i ľudí, ktorým ste volali, môžeme označiť pojmom VZORKA.

Základný súbor (populácia) vs výberový súbor (vzorka).

Otázka 2

Výrobca telefónov si stanobil, že linka je v poriadku, ak menej ako 3% telefónov vyrobených za deň je vadných. Aby si výrobca overil kvalitu dennej produkcie, rozhodol sa náhodne vybrať 30 telefónov na otestovanie poruchovosti. Populáciou v tomto výskume bude:

- A) všetky telefóny vyrobené počas dňa testovania;
- B) 30 vybraných a testovaných telefónov;
- C) 30 telefónov - vadných alebo dobrých;
- D) 3% telefónov, ktoré sú vadné.

Základný súbor (populácia) vs výberový súbor (vzorka).

Úloha 1

Rozhodnite, či v daných prípadoch pôjde o základný alebo výberový štatistický súbor.

- 1) Skúmame výskyt malárie na 1 000 respondentoch.
- 2) Skúmame návštevnosť krúžku v 2. A triede istej základnej školy v Košiciach za prítomnosti všetkých žiakov triedy.
- 3) Skúmame množstvo vynaloženej práce na vypracovanie určitého výrobku u pracovníkov istého závodu.
- 4) Skúmame vzdelanostnú úroveň obyvateľov kraja SR podľa získaných údajov o počte 20% respondentov v každom okrese uvedeného kraja.

Rozcvička s Excelom

ExcelRozcvicka.xls

Deskriptívna vs induktívna štatistiká.

Úloha 2

Sud je naplnený dvoma odrodami hrozna: hroznom so zelenými a fialovými bobuľami. Zaujíma nás, koľko zelených a koľko fialových bobúľ je v sude.



Deskriptívna vs induktívna štatistiká.

Dve možnosti ako riešiť Úlohu 2:

- ① **Spočítať všetky bobule v sude**, a takto zistiť presný počet zelených i fialových bobúľ v sude. ALE! Ak by v sude bolo napr. 250 000 bobúľ a za 1 sekundu sme schopní spočítať 4 bobule, potom by nám počítanie zabralo viac ako x (koľko?) hodín!
 - ② Druhá možnosť je **odobrat' zo suda hrst' bobúľ a spočítať počty zelených a fialových bobúľ len v tejto hrsti**. Zaberie to podstatne kratší čas, ALE! bobule v sude musia byť dôkladne premiešané. To znamená, aby vybratá hrst' reprezentovala celý sud a aby pomer zelených a fialových bobúľ bol približne taký ako v celom sude.
- ✓ Excel: *Reprezentativny_vyber.xlsx*

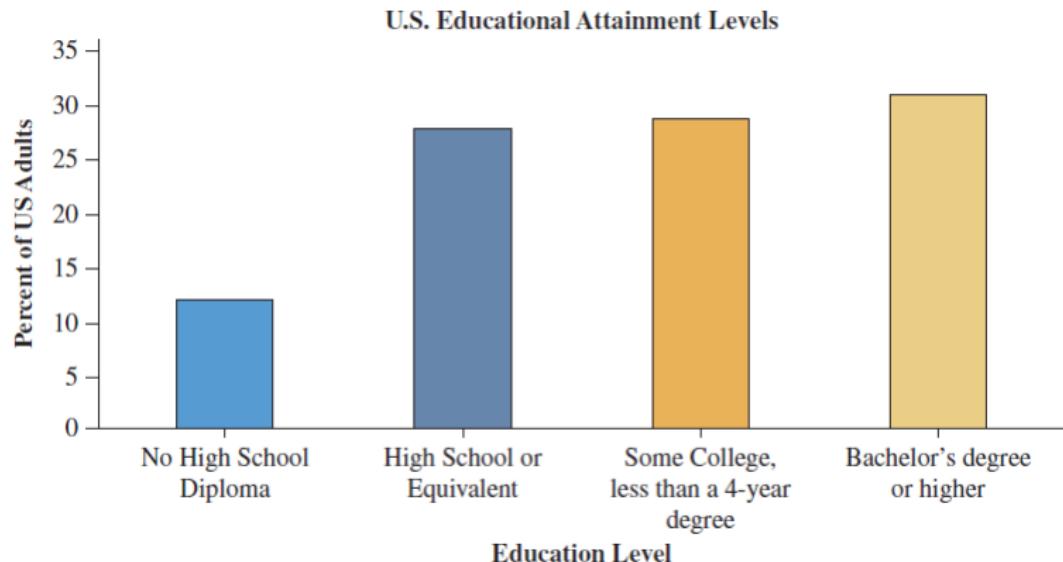
Deskriptívna vs induktívna štatistika.

① prvý prístup - deskriptívna (popisná) štatistika

- ◊ cieľ: presné zachytenie skutočnosti, vystihnutie dôležitých aspektov často extrémne veľkej množiny údajov niekoľkými číslami či obrázkami
- **vyčerpávajúci výskum**
- ◊ usporiadanie údajov, ich popis a účelná summarizácia (miery polohy/rozptylenosti)
- ◊ závery nie je možné zovšeobecňovať na objekty, ktoré neboli súčasťou výskumu
- ◊ využitie: voľby, referendum, sčítanie ľudu
- ◊ **výhody:** jednoduchá matematika (+, −, ·, :, √); vyčerpávajúca, zachytáva takmer vždy daný stav presne
- ◊ **nevýhody:** veľký základný súbor; zdĺhavé alebo nákladné zisťovanie údajov; skúmanie prvkov môže mať za následok ich znehodnotenie; principálne nemožné, ak máme príliš veľký (nekonečný) počet prvkov

Deskriptívna vs induktívna štatistiká.

Sumár prieskumu 78 000 domácností amerického úradu pre sčítanie ľudu z roku 2013



Deskriptívna vs induktívna štatistika.

② druhý prístup - **induktívna štatistika**

- ◊ základ skúmania: **náhodný výber** (vzorka)
- ◊ pomocou údajov (vzorky) odpovedať na všeobecnejšie otázky, závery zovšeobecniteľné aj na objekty, ktoré sami o sebe neboli súčasťou výskumu (vzorky)
- ◊ prepojenie popisných štatistik s teóriou pravdepodobnosti - umožňuje formulovať všeobecné závery s udaním stupňa spoľahlivosti
- ◊ využitie: prieskumy verejnej mienky, kontrola kvality, klinické testy
- ◊ **výhody:** časovo menej náročná; finančne nenáročná
- ◊ **nevýhody:** menej presná, možnosť vzniku chýb; zložitejšia matematika (znalosť teórie pravdepodobnosti)

Chyby (nepresnosti) je už dnes možné lepšie regulovať (niekedy dokonca eliminovať) aj vďaka počítačom.

Deskriptívna vs induktívna štatistika.

Prieskum verejnej mienky ohľadom kontroly predaja zbraní

Predpokladajme, že by sme chceli vedieť, čo si ľudia myslia o kontrolách predaja ručných zbraní. Pozrime sa, ako sa to vidia ľudia na Floride, v štáte s relatívne vysokou mierou násilných zločinov. Záujmovou populáciou je súbor viac ako 10 miliónov dospelých obyvateľov Floridy. Pretože nie je možné diskutovať o probléme so všetkými týmito ľuďmi, môžeme študovať výsledky nedávneho prieskumu verejnej mienky 834 obyvateľov Floridy, ktorý uskutočnil Inštitút pre výskum verejnej mienky na Floridskej medzinárodnej univerzite. V tomto prieskume 54.0% ľudí zaradených do vzorky uviedlo, že uprednostňuje kontrolu predaja ručných zbraní. Novinový článok o prieskume uvádza, že chyba (odchýlka) toho, ako blízko je toto číslo k skutočnému percentu obyvateľov uprednostňujúcich kontrolu predaja zbraní, je 3.4%. Neskôr uvidíme, že to znamená, že môžeme s vysokou dôverou (asi 95% istotou) predpovedať, že percento všetkých dospelých Floridčanov, ktorí uprednostňujú kontrolu nad predajom ručných zbraní, sa pohybuje v rozmedzí $\pm 3.4\%$ od hodnoty 54.0% z prieskumu, to znamená medzi 50.6% a 57.4%.

- otázka na zamyslenie: **Čo je v uvedenej štúdii popisná štatistika a čo je induktívna štatistika?**



Klamstvá a formy komunikácie.

Deception and Design: The Impact of Communication Technology on Lying Behavior (Computer-Human Interaction [2009]: 130–136)

Článok popisuje štúdiu, ktorej cieľom bolo zistiť, či vysokoškoláci klamú pri osobnej komunikácii ("tvárou v tvár") menej často ako pri iných formách komunikácie, ako sú telefonické rozhovory alebo e-mail. Účastníkmi tejto štúdie bolo 30 študentov komunikačného kurzu vyšej divízie na Cornell University, ktorí za účasť získali kredit za kurz. Účastníci boli požiadani, aby týždeň zaznamenávali všetky svoje sociálne interakcie a všímali si všetky klamstvá, ktoré povedali. Na základe údajov z týchto záznamov autori príspevku dospeli k záveru, že študenti klamú častejšie v telefonických rozhovoroch než v rozhovoroch z očí do očí a častejšie v rozhovoroch z očí do očí ako v e-mailoch.



- Aká je cieľová populácia? Kto/čo je štatistická jednotka?
- Kto/čo tvorí vzorku? Bola vzorka vybraná korektne?
- Je možné aby vybraná vzorka bola reprezentatívna vzhľadom k cieľovej populácii?
- Sú bádateľné systematické chyby pri výbere vzorky, ktoré by mohli viest' k skresleniu výsledkov štúdie?

Porozmýšľajte, než si objednáte ten hamburger!

What People Buy from Fast-Food Restaurants: Caloric Content and Menu Item Selection (Obesity [2009]: 1369–1374)

V článku sa uvádza, že priemerný počet kalórií spotrebovaných na obed v reštauráciach rýchleho občerstvenia v New Yorku bol 827. Vedci náhodne vybrali 267 miest rýchleho občerstvenia. Príspevok uvádza, že v každom z týchto miest "boli pri vstupe do reštaurácie oslovení dospelí zákazníci a požiadani aby pri odchode poskytli informácie ohľadom svojho jedla a vyplnili krátke dotazník".

- Aká je cieľová populácia? Kto/čo je štatistická jednotka?
- Kto/čo tvorí vzorku? Bola vzorka vybraná korektne?
- Je možné aby vybraná vzorka bola reprezentatívna vzhľadom k cieľovej populácii?
- Sú bádateľné systematické chyby pri výbere vzorky, ktoré by mohli viest' k skresleniu výsledkov štúdie?

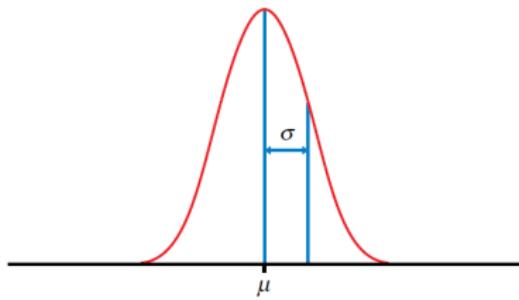
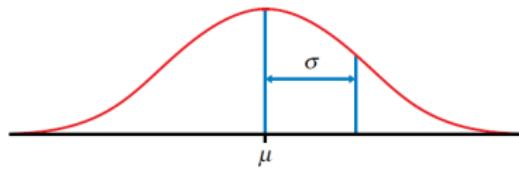
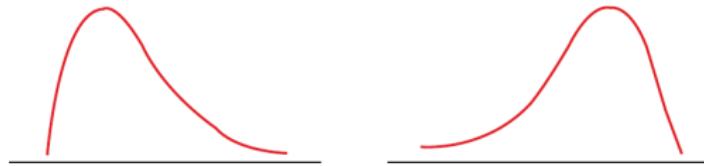


Čo je dobré uvedomiť si pred samotným spracovaním dátovej vzorky ...

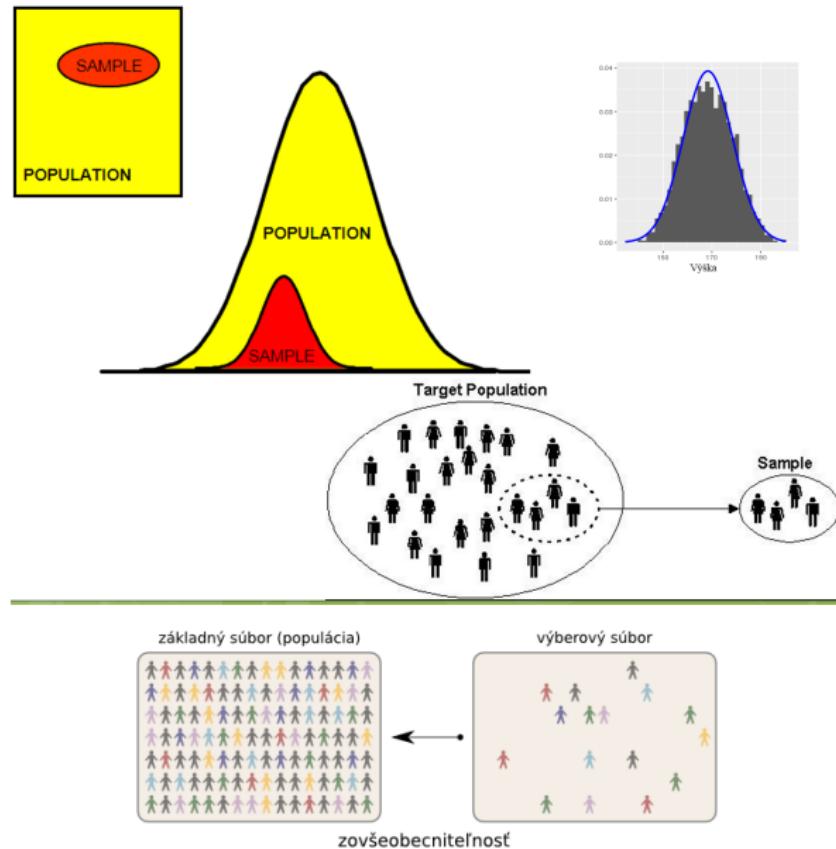
- kto podnietil (prípadne financoval) výskum
- kto a ako zozbieral údaje (ako bola vybraná vzorka)
- ako boli namerané údaje
- za akých podmienok bol robený výskum
- porovnatelnosť vzoriek



Populácia vs výberový súbor (vzorka) ešte raz.



Populácia vs výberový súbor (vzorka) ešte raz.

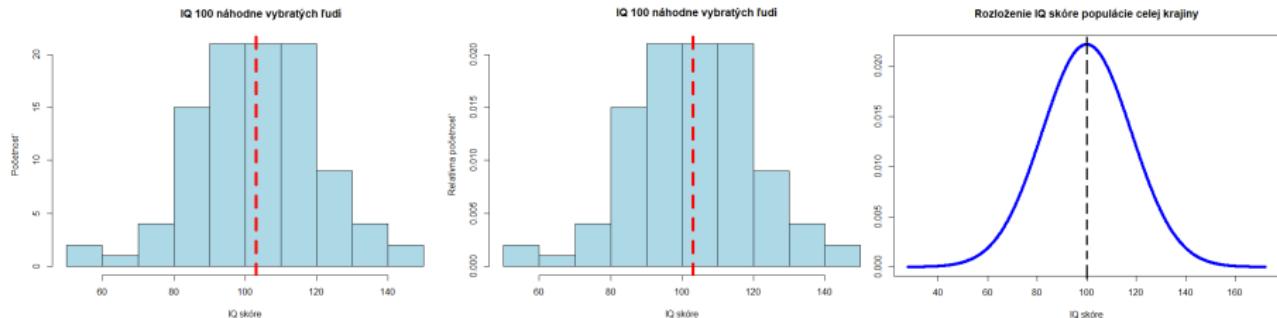


Populácia vs výberový súbor (vzorka) ešte raz.

- **populácia** → (neznáme) parametre ako napr. **stredná hodnota μ , smerodajná odchýlka σ**
- **vzorka** (výberový súbor) → **odhady** (neznámych) parametrov:
napr. odhad pre μ je **výberový priemer \bar{x}** ,
odhad pre σ je **výberová smerodajná odchýlka s**
- $\bar{x} = \frac{x_1 + x_2 + \dots + x_{n-1} + x_n}{n},$
$$s = \sqrt{\frac{(\bar{x} - x_1)^2 + (\bar{x} - x_2)^2 + \dots + (\bar{x} - x_{n-1})^2 + (\bar{x} - x_n)^2}{n}}$$
- \bar{x}, s sa niekedy nazývajú aj (popisné) štatistiky výberového súboru

► simulácia

Populácia vs výberový súbor (vzorka) ešte raz.



Základný súbor (populácia)	Výberový súbor (vzorka)
parameter	odhad
populačný priemer (stredná hodnota)	výberový priemer
populačná disperzia (rozptyl)	výberová disperzia (rozptyl)
populačná smerodajná odchýlka	výberová smerodajná odchýlka
populačná štandardná chyba	výberová štandardná chyba
pravdepodobnosť	relatívna početnosť

Kofein a reakčný čas.

Otázka 3

V experimente na určenie ako kofein ovplyvňuje ľudské telo boli respondenti požiadani stlačiť tlačidlo čo najrýchlejšie, pričom jednej skupine bola podaná tabletka obsahujúca kofein a druhej skupine bola podaná placebo tabletka. Priemerný reakčný čas "kofeinovej" skupiny bol 158 milisekúnd a "placebo" skupiny 197 milisekúnd. Čísla 158 a 197 sú:

- A) parametre;
- B) štatistiky;
- C) 158 je parameter, 197 je štatistika;
- D) 158 je štatistika, 197 je parameter.



Malé upozornenie na záver.

- klasická induktívna štatistika - klúčová ingrediencia je väčšinou náhodný výber
- práca s nezávislými meraniami/pozorovaniami (neovplyvňujú sa vzájomne)
- prípad väčšiny uvedených a diskutovaných príkladov
- existencia (uvidíme neskôr) iných situácií, procesov a s tým súvisiacich typov dát i metodológií ich spracovania/vyhodnocovania
- sledovanie nejakej veličny v čase (cena určitej komodity) - vznik postupnosti údajov v čase a tiež vznik časovej závislosti (nasledujúca resp. budúca hodnota závisí od súčasnej hodnoty prípadne aj od minulých hodnôt)
- priestorová závislosť - geologické dáta (ložiská nerastných surovín)
- závislosť aj v obidvoch zmysloch ...

Zdroje.

-  A. Agresti, Ch. A. Franklin, B. Klingenberg (2017). *Statistics : The Art and Science of Learning from Data*. Boston: Pearson.
- ▶ D. Klein (2016). *Štatistické metódy v geografii*. PF UPJŠ - ÚMV.
-  D. S. Moore, G. P. McCabe & B. A. Craig (2009). *Introduction to the practice of statistics*. New York, W.H. Freeman.
-  J. M. Utts, R. F. Heckard (2014). *Mind on Statistics*. Cengage Learning.
- ▶ M. Hančová (2021). *Úvod do analýzy dát*. PF UPJŠ - ÚMV.
-  R. Peck, T. Short (2018). *Statistics: Learning from Data*. Cengage Learning.
- ▶ S. Glen (2021). *Misleading Graphs: Real Life Examples*. From StatisticsHowTo.com: Elementary Statistics for the rest of us!
<https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/>
- ▶ S. Glen (2021). *Misleading Statistics Examples in Advertising and The News*. From StatisticsHowTo.com: Elementary Statistics for the rest of us!
<https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/>

andrej.gajdos@upjs.sk

