aktualizované 02-10-2024

### KMA/PAS & KMA/PST

Pravděpodobnost & statistika

ZS 2024/2025

© Martin Bod'a martin.boda@outlook.com https://orcid.org/0000-0002-7503-6898

## Poznámka k historickému chápaniu štatistiky

Slovo štatistika vo svojich lexikálnych variáciách (štatistik, štatistický) je odvodené od latinského slova status vo význame politického štátu. V roku 1770 bola chápaná ako "náuka, ktorá nás učí, aké je politické usporiadanie moderných štátov poznaného sveta" (Yule, 1924, s. 1). Spočiatku išlo o prevažne verbálne charakterizovanie vlastností štátu a jeho fungovania, až neskôr asi v druhej polovici 19. storočia prevládol numerický prístup k hodnoteniu súvislostí minulého a súdobého politického a spoločenského života – za pomoci číselných a tabuľkových prehľadov. I keď sa štatistika začala používať aj v iných oblastiach (fyzika, biológia, meteorológia, psychológia, lingvistika ap.), mala najprv výlučne deskriptívny charakter a až neskôr v posledných desat'ročiach 19. storočia sa začala používat' aj na zovšeobecňovanie (indukciu) a v širších kontextoch dnešného chápania.

## Pojem štatistiky

Naše základné chápanie pojmu štatistika:

- (Aplikovaná) štatistika: praktická činnosť súvisiaca so zberom a vyhodnocovaním dát, ako aj s interpretáciou súvislostí v nich.
- <u>Cieľ štatistiky</u>: identifikovať vlastnosti a súvislosti skúmaného javu obsiahnuté v dátach.

Ďalšie poňatia pojmu štatistika:

- (Matematická) štatistika: vedná disciplína orientovaná na (vedecké) zovšeobecňovanie výsledkov z náhodného výberu alebo experimentu zahrňujúca odhadovanie neznámych parametrov a testovanie štatistických hypotéz (tzn. štatistickú inferenciu/indukciu).
- · Štatistika: funkcia náhodného výberu (nejaký vzťah alebo vzorec definovaný nad náhodným výberom).

### Pojem štatistiky (./.)

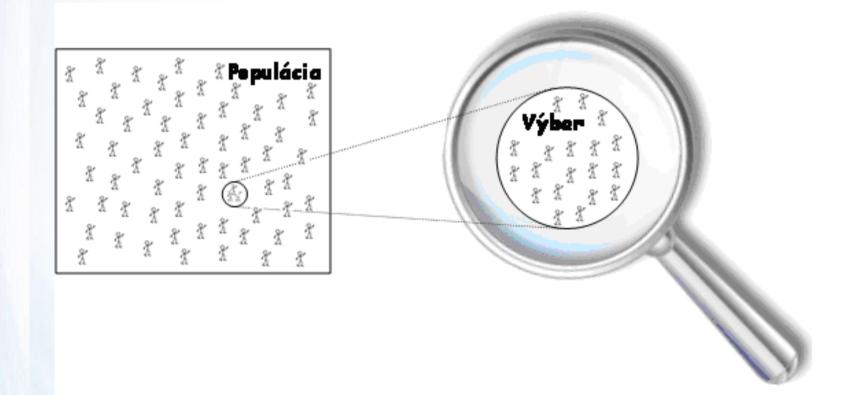
Kým aplikovaná štatistika sa zaoberá samotným použitím štatistických metód a iných metód analýzy údajov v praktických situáciách nad reálnymi dátami s cieľom

- a) poznania a deskripcie týchto dát a/alebo
- b) zovšeobecnenia v nich obsiahnutých súvislostí (teda inferenciou, indukciou) a/alebo
- c) hľadaním kauzálnych súvislostí týkajúcich sa týchto dát,

matematická štatistika má rýdzo teoretický charakter a orientuje sa na návrh a vývoj metód určených na zovšeobecňovanie (inferenciu, indukciu) a na poznanie kauzálnych väzieb.

## Základné pojmy

- · Štatistické zisťovanie: vyčerpávajúce & výberové.
- · Populácia, výber, náhodný výber, experiment
- Súbor (tzv. štatistický súbor), jednotka (tzv. štatistická jednotka), premenná (tzv. štatistický znak).
- · Štatistika: deskriptívna & inferenčná (induktívna).



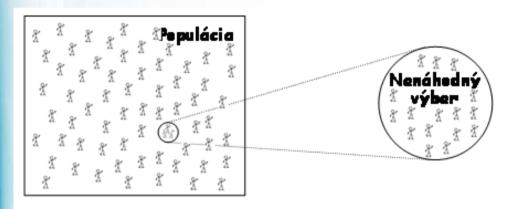


#### Vyčerpávajúce štatistické zisťovanie:

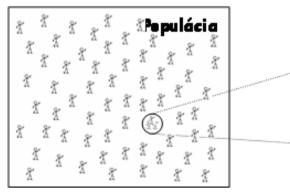
\*\* štatistický súbor = populácia \*\*

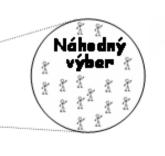
Všetka dostupná informácia je teda obsiahnutá v súbore a použije sa deskriptívna štatistika.

Výberové štatistické zisťovanie: \* štatistický súbor < populácia \*

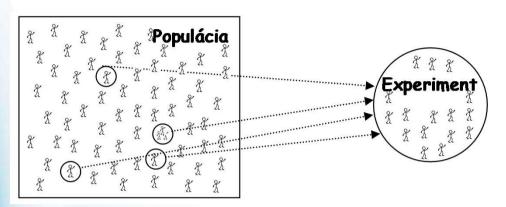


Výber je nenáhodný (a nereprezentatívny) a možno ho ale štatisticky popísať. Použije sa **deskriptívna štatistika**.





Keď je výber je náhodný (a reprezentatívny), možno zovšeobecňovať (indukovať) a použije sa **induktívna (inferenčná)** štatistika.



Experiment je ľubovoľná procedúra, ktorú možno nekonečnekrát opakovať a má dobre
definovanú množinu výsledkov.
Ak sa tieto výsledky nedajú
dopredu dobre predikovať,
ide o náhodný experiment.

Možno rozlíšiť tri typy náhodných experimentov:

- \*\* laboratórne experimenty (v umelom prostredí pod plnou kontrolou experimentátora),
- \*\* poľné experimenty (prirodzené prostredie sčasti kontrolované experimentátorom),
- \*\* prirodzené experimenty (prirodzené prostredie bez kontroly realizovanej experimentátorom).

Pozorovania majú charakter náhodnosti, a tak sa použije sa induktívna (inferenčná) štatistika.

7

Pri <u>deskriptívnej štatistike</u> iba (p)opisujeme vlastnosti príslušného súboru a výsledky nezovšeobecňujeme, čomu zodpovedá aj **opatrná** interpretácia výsledkov:

V danom súbore je ...

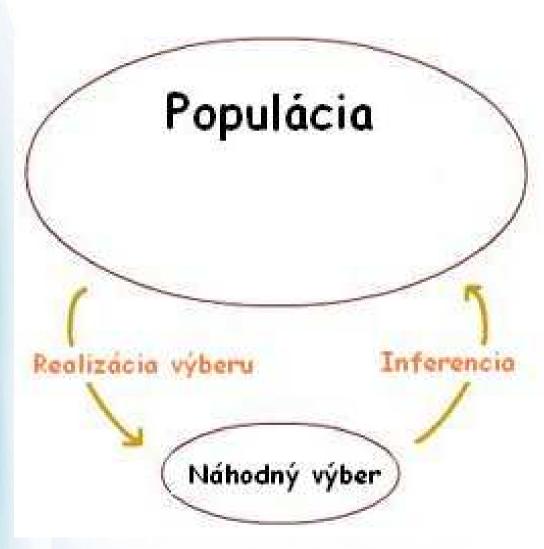
Medzi sledovanými jednotkami je ...

Keď disponujeme náhodným (a reprezentatívnym) výberom či experimentálnymi dátami, je možné zovšeobecňovať (v rozumných prípadoch) výsledky zistené pre daný súbor na úroveň populácie, odhadovať neznáme hodnoty pre populáciu a testovať o nich štatistické hypotézy (teda robiť štatistickú indukciu/inferenciu). Tomuto použitiu zodpovedá induktívna/inferenčná štatistika. Interpretácia je abmicióznejšia:

Sledovaná populácia má...

Ukazuje sa, že vo všeobecnosti platí...

### Induktívna/inferenčná štatistika



Induktívna/inferenčná štatistika sa kriticky spolieha na vykonanie náhodného výberu z vhodne definovanej populácie s cieľom bližšie poznať túto populáciu. Jej princíp je potom v tom, že poznatky získané z analýzy dát v náhodnom výbere sa potom zovšeobecnia naspäť na populáciu aplikáciou induktívneho postupu logického myslenia.

V prírodných vedách sa ale dáta pre indukciu/inferenciu získavajú obvykle formou ná-hodného experimentu v laboratórnych alebo takmer laboratórnych podmienkach.

Zdroj: https://onlinecourses.science.psu.edu/stat414/node/18

Premenná: číselná alebo nečíselná charakteristika meraná na sledovanej jednotke. (Na sledovaných jednotkách sa obvykle sleduje viacero premenných súčasne.) Pre premennú sa používa tiež pojem (štatistický) znak, veličina. Niekde sa terminologicky zavádza všeobecné označenie znak a pojem premenná sa používa iba pre znak v číselnom formáte.

Premenná: (nie nevyhnutne) číselná charakteristika alebo vlastnosť, ktorá nadobúda viaceré hodnoty alebo obmeny.

Náhodná premenná: <u>čísel-</u> <u>ná</u> premenná, ktorej hodnota je výsledkom náhodného pokusu, resp. je určená ako výsledok náhody. Dáta: merania hodnôt (realizácií), ktoré nadobúdajú premenné.

Pozorovanie (observácia): konkrétna (jedna) meraná hodnota (realizácia).

Rozsah súboru (resp. dátovej vzorky): počet meraní sledovanej premennej.

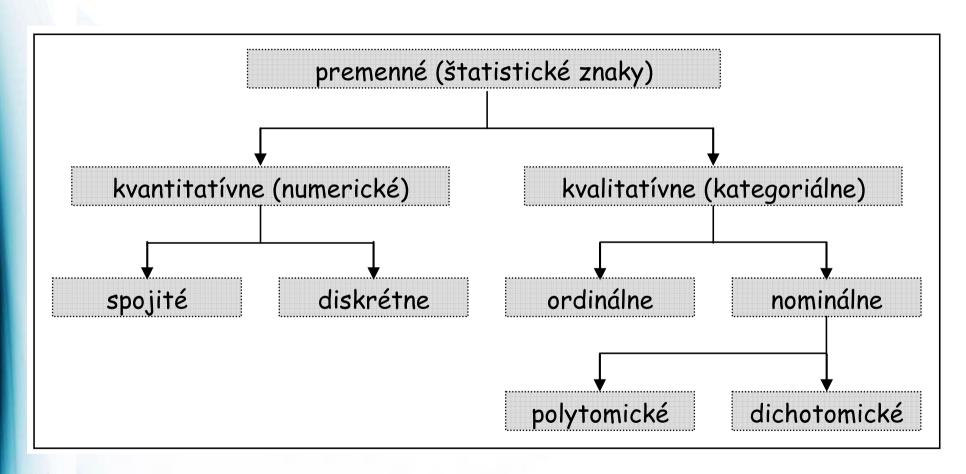
### Klasifikácia premenných (základná)

- \* kvantitatívne (numerické, číselné)
  spojité (ľubovoľná hodnota z nejakého intervalu)
  diskrétne (spočítateľný počet obmien)
- \* kvalitatívne (kategoriálne, slovné)

### Klasifikácia premenných (podľa škály meraní)

- \*\* nominálne (slovné kategórie & neusporiadateľné)
  - --> alternatívne/dichotomické/binomické/binárne
  - --> multinomické/polytomické
- \*\* ordinálne (poradové)
- \*\* numerické (kardinálne, číselné)

### Klasifikácia premenných



Na jednej strane je klasifikovanie premenných (štatistických znakov) zdanlivo nezaujímavou teóriou, na druhej strane je potrebné si uvedomiť, že pri praktickej analýze je možné uplatniť výlučne iba postupy navrhnuté (a správne) pre analýzu príslušného typu premennej. Nemožno ordinálnu premennú (napr. stupeň dosiahnutého vzdelania alebo výsledné umiestnenie v nejakej súťaži) nekriticky analyzovať ako numerickú ("plnohodnotnú") premennú, hoci môže byť vyjadrená skrze číselné hodnoty (napr. 1 - základná škola, 2 - stredná škola, 3 - vysoká škola Bc. &c.).

Slovo "plnohodnotný" pri numerickej premennej na predošlej snímke bolo použité schválne vo význame, že numerická premenná dáva najviac informácie a je možné jej vlastnosti dôslednejšie analyzovať a poznať. Z hľadiska informácie, ktorá sa typicky sleduje (úroveň, variabilita, zošikmenie a ďalšie vlastnosti rozdelenia), je "najštedrejší" prípad spojitej numerickej premennej a zase "najchudobnejší" prípad nominálnej dichotomickej premennej.

Za cenu strany časti informácie sa dá zo spojitej numerickej premennej postupne vyrobiť diskrétna numerická premenná, ordinálna premenná, nominálna polytomická premenná a nominálna dichotomická premenná. Opačne to neplatí.



## Typy dát

Štatistické dáta môžu byť primárne štvorakého druhu:

- \* prierezové dáta,
- \* časové rady,
- \* panelové (longitudinálne) dáta,
- \* opakované merania.

Kombinuje sa pritom rozlišovanie dvoch dimenzií pozorovania dát: priestor & čas.

Priestorové kritérium určuje, či sú dáta napozorované za jeden objekt (jednotku) alebo sa týkajú viacerých objektov (jednotiek).

Časové kritérium určuje, či sa dáta vzťahujú na jeden okamih alebo vznikli v rozličných obdobiach, resp. či boli zisťované iba opakovane.

Prierezové dáta sú pozorovania jednej alebo viacerých premenných uskutočňované V TOM ISTOM ČASE NA RÔZ-NYCH OBJEKTOCH (napr. domácností, firiem, medveďoch vo voľnom priestore, rastlín, celestiálnych telies). Na rozlíšenie jednotlivých pozorovaní sa používa index označujúci príslušnosť k danému objektu, tzn. priestorový index (typicky pre nejaký nešpecifický objekt i).

#### Príklad:

- priemerný mesačný príjem v októbri 2024 u študentov zapísaných na predmet KMA/PAS,
- priemerná teplota v jednotlivých sídlach Ústeckého kraja 2.10.2024 o 8.00 h,
- efektívnosť retailových pobočiek České spořitelny, a. s., v Českej republike na základe výsledkov za rok 2022.

Časové rady sú pozorovania jednej alebo viacerých premenných zisťované NA TOM ISTOM OBJEKTE V RÔZ-NYCH ČASOVÝCH OBDOBIACH (typicky konzekutívne bez prerušenia, napr. každý mesiac od I/2000 do XII/2023 alebo anuálne od 1993 do súčasnosti). Na rozlíšenie jednotlivých pozorovaní je teraz potrebný časový index určujúci datovanie daného pozorovania (typicky pre nejaký nešpecifikovaný čas t).

#### Príklad:

- nezamestnanosť od Q1/1993 do Q3/2024 v Ústeckom kraji,
- teplota na Klíše (GPS 50.668759, 14.017285) každé ráno o 6.00 h počas piatich posledných rokov,
- úhrnná mesačná spotreba plynu v spolkovej krajine Bádensko-Württembersko.

Panelové dáta sú pozorovania jednej alebo viacerých premenných uskutočňované v ROZLIČNÝCH ČASOVÝCH OKAMIHOCH NA RÔZNYCH OBJEKTOCH, pričom datovanie jednotlivých pozorovaní má svoje opodstatnenie a je dôležité z hľadiska interpretácie. Na rozlíšenie jednotlivých pozorovaní je potrebná kombinácia indexu určujúceho objekt (typicky pre nešpecifikovaný objekt i) a datovanie (typicky pre nešpecifikovaný čas t). Môže ísť o vyrovnaný (vybalansovaný) alebo nevyrovnaný (nevybalansovaný) panel.

#### Príklad:

- dáta získané z Českého panelového šetření domácností za roky 2015 - 2019 (CHPS, https://www.promenyceskespolecnosti.cz),
- Penn World Table (https://www.rug.nl/ggdc/productivity/pwt),
- údaje o zločinnosti v 90 okresoch v Severnej Karolíne a jej determinatoch za roky 1981 až 1987 (http://qed.econ.queensu.ca/ jae/ datasets/baltagi003).

18

Opakované merania vznikajú, keď NA RÔZNYCH OBJEK-TOCH sa uskutočňujú merania či zisťovania sledovanej premennej alebo sledovaných premenných OPAKOVANE, pričom konkrétne datovanie merania nemá opodstatnenie, maximálne je dôležitá CHRONOLÓGIA JEDNOTLIVÝCH MERANÍ (napr. pri experimentoch postupne slabnúci efekt). Potrebujú sa formálne dva indexy, na odlíšenie objektu (typicky pre nejaký objekt i), potom na odlíšenie merania (typicky pre nejaké ľubovoľné meranie j).

#### Príklad:

- opakované nezávislé merania úrovne koncentrácie amoniaku na rôznych poľnohospodárskych políčkach v sledovanom regióne,
- opakované nezávislé merania kvality podzemnej vody na rôznych miestach Českej republiky,
- merania postupného nabiehania efektov novej liečby proti diabetesu zisťované po týždni, mesiaci, dvoch mesiacoch.

Okrem toho existujú aj ďalšie typy dát ako funkcionálne dáta, kompozičné dáta, direkčné dáta a i.

- Funkčné dáta sú merania zisťované na krivkách, povrchoch alebo iných podkladoch variujúcich na kontinuu.
- Kompozičné dáta určujú štruktúru nejakého celku (teda percentuálne, resp. relatívne zloženie celku, ako pripadá na jednotlivé časti).
- Direkčné (cirkulárne či sférické) dáta reprezentujú merania smerových vektorov, osí alebo rotácií vo viacrozmernom priestore (napr. smery sú reprezentované ako jednotkové vektory).

### 3 úrovne a ciele štatistických analýz

Úroveň 1: deskripcia

Úroveň 2: inferencia (zovšeobecňovanie)

Úroveň 3: skúmanie kauzálnych vzťahov

\*\* Cieľom deskriptívnej analýzy (úrovne 1) je sumarizácia dát, identifikácia nezvyčajných pozorovaní a stanovenie základných súvislostí v dátach. Jej závery sa týkajú **IBA da-ného súboru** (resp. dátovej vzorky).

\*\* Cieľom inferenčnej analýzy (úrovne 2) je zovšeobecňovanie zistení z daného súboru (resp. dátovej vzorky) pre celú populáciu. Jej závery sa vzťahujú na celú populáciu.

\*\* Cieľom kauzálnej analýzy (úrovne 3) je zistenie vzájomných príčinno-následných súvislostí medzi skúmanými premennými, teda stanovenie, ktoré premenné spôsobujú variáciu cieľovej premennej. Jej závery sa týkajú celej **populácie**.

### 3 úrovne a ciele štatistických analýz (./.)

Úroveň 1: deskripcia

Úroveň 2: inferencia (zovšeobecňovanie)

Úroveň 3: skúmanie kauzálnych vzťahov

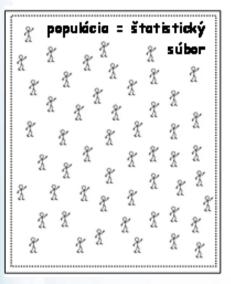
Čím vyššia je úroveň, tým vyššie sú ambície štatistickej analýzy a aj vyššie požiadavky na kvalitu dát.

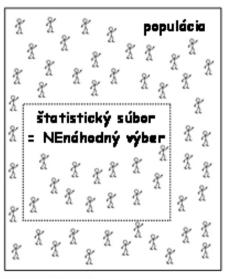
\*\* Na deskriptívnu analýzu (úroveň 1) postačujú ľubovoľné dáta (dáta celej populácie, nenáhodný výber, náhodný výber, experimentálne dáta).

\*\* Na inferenčnú analýzu (úroveň 2) je potrebný náhodný výber z dobre definovanej populácie alebo dáta získané (náhodným) experimentom.

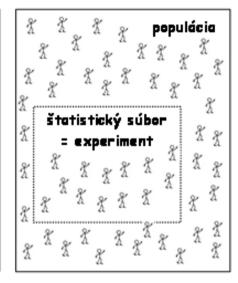
\*\* Na kauzálnu analýzu (úroveň 3) sú potrebné dáta získané (náhodným) experimentom.

## 3 úrovne a ciele štatistických analýz (./.)









	deskripcia pre celú populáciu a skúmanie súvislostí pre celú populáciu	deskripcia pre daný súbor	deskripcia pre daný súbor	deskripcia pre daný súbor
			inferencia pre populáciu	inferencia pre populáciu
				analýza kauzality pre populáciu

### Podstata štatistiky

Štatistika sa zamestnáva napospol s hromadnými javmi a poskytuje ich opis, resp. umožňuje úsudky o nich. Pod hromadným javom sa rozumie (obyčajne náhodná) udalosť, ktorá sa opakuje a vyskytuje (nezávisle) "mnohokrát" a vedie k realizácii číselnej alebo nečíselnej hodnoty. Hromadnosť javu spôsobuje, že tieto hodnoty sa vyskytujú s určitou zákonitosťou / frekvenciou (a opakujú sa, resp. takmer sa opakujú). Následne sa hovorí o frekvenčnom rozdelení alebo o rozdelení početnosti. Štatistika sa potom sústreďuje na frekvenčné rozdelenie sledovaných premenných a jej úlohou je ho v nejakej podobe opísať, stanoviť či identifikovať, a to bez ohľadu na to, či ide o deskriptívnu alebo induktívnu štatistiku.

### Príklad - hromadnosť javov

Príkladom pre numerickú premennú je "cena jedného kopčeka obyčajnej zmrzliny v Ústí nad Labem v letnej sezóne roku 2020". Táto sa mohla pohybovat', povedzme, od 15 do 25 Kč a závisela od prevádzky. Každá prevádzka uplatňovala samostatnú cenu a tieto ceny sa s určitou pravidelnosťou opakovali.

hromadnost' = to, že bolo viacero prevádzok s vlastnou cenou frekvencia = koľkokrát sa opakovali tie isté ceny

Príkladom pre nominálnu premennú je "preferovaná príchuť (preferovaný druh) zmrzliny kupovaný v Ústí nad Labem v letnej sezóne roku 2020". Závisí od každého konzumenta, či uprednostnil čokoládovú, šmolkovú, nutellovú, ananásovú, jogurtovú alebo inú príchuť. Tieto príchute sa objavovali medzi konzumentmi niekoľkokrát.

hromadnost' = to, že bolo viacero konzumentov s preferenciou frekvencia = koľkokrát sa opakovali tie isté preferencie chute

### Frekvenčné rozdelenie a frekvenčná krivka

O frekvenčnom rozdelení sa hovorí pri numerických premenných aj kategoriálnych premenných, ale o frekvenčnej krivke sa hovorí *iba* pri numerických premenných na spojitej škále.

#### Numerické premenné (kvantitatívne znaky):

Frekvenčné rozdelenie (rozdelenie početnosti) je zákonitosť (pozorovaná pravidelnosť), ktorá určuje, koľkokrát sa hodnoty premennej opakujú (pri diskrétnych premenných), alebo ktorá určuje rozdelenie hodnôt cez postupné rovnako široké číselné intervaly (pri spojitých premenných).

Frekvenčná krivka je ideálne matematické vyjadrenie (matematický funkčný predpis) frekvenčného rozdelenia spojitej numerickej premennej.

#### Ordinálne/nominálne premenné (kvalitatívne znaky):

Frekvenčné rozdelenie (rozdelenie početnosti) je zákonitosť (pozorovaná pravidelnosť), ktorá určuje, koľkokrát sa hodnoty sledovanej premennej opakujú.

#### Príklad 1

Denné výnosy záverečných cien akcie spoločnosti General Electric Company obchodovanej na NYSE od 03-01-1969 do 31-12-1998. Celkovo ide o 2528 pozorovaní uvedených v percentách p.d.

- ? hromadnost'
- ? frekvenčné rozdelenie
- ? frekvenčná krivka

```
Time Series:
Start = c(1969, 1)
End = c(1975, 338)
Frequency = 365

[1] -1.6760  1.7045 -0.2793  0.0000  0.0000 -0.5602  0.0000  1.4085

[9]   0.0000 -0.2778  0.8357  1.3812 -0.2725 -0.5464 -0.5495  1.3812

[17]   0.0000  2.1798  1.0667  1.3193  0.7812 -0.2584 -0.2591 -0.5195

[25]   0.2611  0.7812 -1.8088 -1.0526 -1.5957 -0.2703 -1.0840  1.3699

[33] -0.5405  1.9022  0.2667 -2.9255  1.0959 -1.8970  1.6575 -0.5435
```

### Príklad 1 (./.)

2528 denných výnosov akcie spoločnosti General Electric Company roztriedených do vhodne zvolených intervalov rovnakej šírky:

Frekvenčné rozdelenie sa charakterizuje početnosťami a rozlišujú sa štyri typy početností:

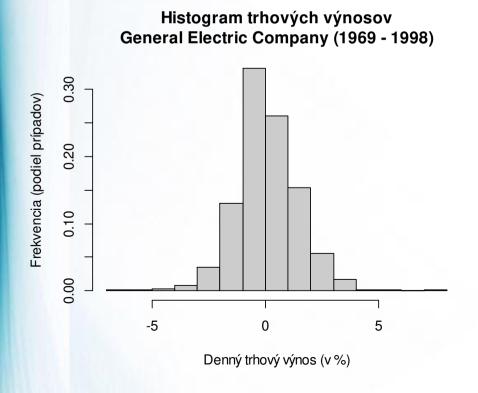
- \*\* n(i) bežná absolútna početnosť (koľkokrát sa vyskytlo)
- \*\* N(i) kumulatívna absolútna početnosť (koľkokrát sa vyskytlo "do")
- \*\* f(i) bežná relatívna početnosť (v koľko percent prípadov sa vyskytlo)
- \*\* F(i) kumulatívna relatívna početnosť (v koľko percent prípadov sa vyskytlo "do")

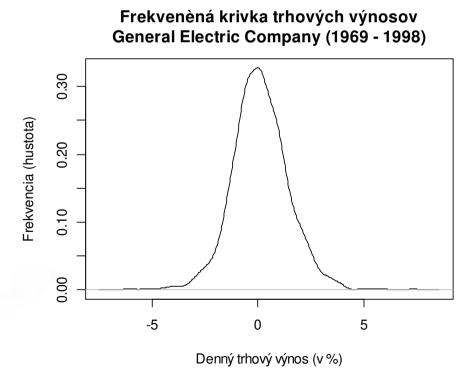
1.000

2528

### Príklad 1 (./.)

Na grafické znázornenie frekvenčného rozdelenia spojitej premennej slúži histogram (používajú sa bežné absolútne alebo relatívne početnosti). Jeho vyhladením (a zmenšovaním šírky triediaceho intervalu) vzniká frekvenčná krivka.





#### Príklad 2

Ročné tržby obchodov so značkovým oblečením v Holandsku zaznamenaných v roku 1990. K dispozícii je 400 hodnôt ročných tržieb v 10000-násobkoch holandských guldenov.

- ? hromadnost'
- ? frekvenčné rozdelenie
- ? frekvenčná krivka

```
40.0000 130.0000
     75.0000 192.6395 125.0000
                               69.4227
                                         75.0000
 \lceil 1 \rceil
     49.5340 120.0000 49.5340 91.1315
                                        23.1000
                                                  49.2033 150.0000
 [8]
             87.0000 69.4227 24.5228 37.5000
[15]
     33.0000
                                                  33.0000 30.1133
                                                  97.6817 85.0000
[22]
     97.6817 150.0000 7.7369 47.1000
                                        60.1000
[29]
     49.5340 69.4227 15.6168 49.5340 100.0000
                                                 69.4227
                                                           97.6817
[36]
     15.6168 40.0000 17.9655 87.6000
                                        30.1133
                                                  49.5340
                                                           30.1133
[43]
     85.3000 97.6817 189.4931
                                97.6817
                                         69.4227
                                                  49.5340
                                                           69.4227
[50]
     97.6817 97.6817 81.5000 49.5340 171.8000
                                                  87.5000
                                                           79.0000
```

### Príklad 2 (./.)

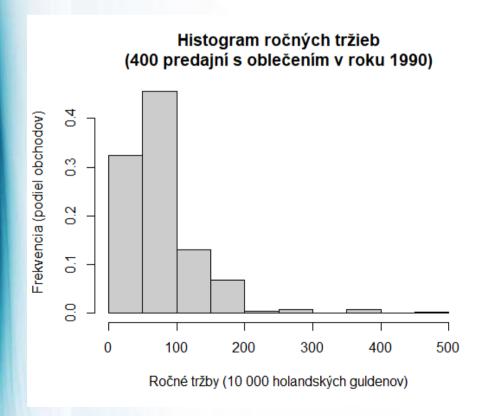
ročné tržby 400 predajní so značkovým oblečením vytriedené do intervalov rovnakej šírky:

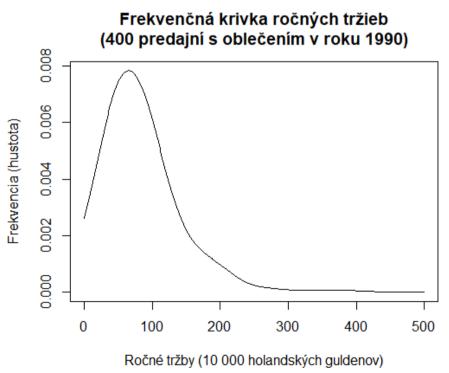
```
n(i) N(i) f(i) F(i)
          130
               130 0.325 0.325
(0,50]
(50,100]
          182 312 0.455 0.780
(100,150]
           52 364 0.130 0.910
           27 391 0.068 0.978
(150,200]
(200,250) 2 393 0.005 0.983
(250,300]
            3 396 0.008 0.991
(300,350]
            0 396 0.000 0.991
(350,400]
            3 399 0.008 0.999
(400,450]
            0 399 0.000 0.999
(450,500]
               400 0.001 1.001
          400
                   1.000
```

- \*\* n(i) bežná absolútna početnosť (koľkokrát sa vyskytlo)
- \*\* N(i) kumulatívna absolútna početnosť (koľkokrát sa vyskytlo "do")
- \*\* f(i) bežná relatívna početnosť (v koľko percent prípadov sa vyskytlo)
- \*\* F(i) kumulatívna relatívna početnosť (v koľko percent prípadov sa vyskytlo "do")

### Príklad 2 (./.)

Frekvenčné rozdelenie / frekvenčná krivka ukazuje pre dáta o ročných tržbách iný charakteristický tvar a iné vlastnosti ako v predošlom príklade.





#### Príklad 3

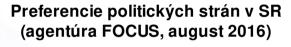
Volebné preferencie politických strán - august 2016

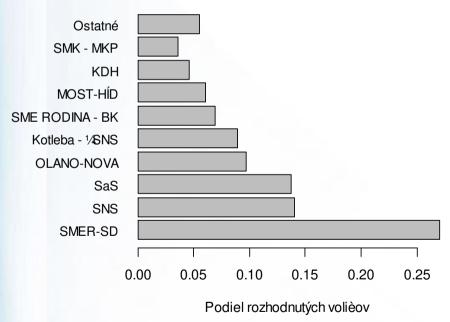
Agentúra FOCUS uskutočnila v dňoch 11.8. – 16.8.2016 prieskum verejnej mienky formou osobného opytovania. Výberovú vzorku tvorilo 1002 respondentov, ktorí reprezentujú populáciu SR vo veku nad 18 rokov z hľadiska pohlavia, veku, vzdelania, národnosti, veľkostných kategórií sídiel a krajského členenia.

Respondentom bola položená nasledujúca otázka: "Teraz si, prosím, predstavte, že by sa parlamentné voľby konali nasledujúci víkend a kandidovali by nasledujúce politické strany a hnutia. Ktorej strane alebo hnutiu by ste dali svoj hlas?"

http://www.focus-research.sk/files/215\_Volebne%20preferencie%20politickych%20stran\_august%202016.pdf

## Príklad 3 (./.)





# Preferencie politických strán v SR (agentúra FOCUS, august 2016) Podiel rozhodnutých volièov

