1 Základy štatistiky

Štatistika je vedecká disciplína založená na empirických skúsenostiach, ktorej cieľom je zbierať dáta, skúmať ich a pomocou nich vytvárať všeobecné závery pre sledovanú štatistickú populáciu. Trochu inými slovami, **štatistika** je „*odbor zaoberajúci sa kvantitatívnou charakteristikou rozličných javov*" [5].

1.1 Štatistické jednotky a štatistické znaky

Pod pojmom **štatistická populácia** rozumieme súbor alebo množinu, ktorý/ktorá obsahuje všetky jednotky, ktoré sú predmetom daného skúmania. Odborným výrazom **štatistický znak** (alebo štatistická premenná) označujeme istú vlastnosť jednotiek štatistickej populácie, ktorá ich charakterizuje. **Základné typy štatistických znakov** sú nasledovné:

- časové štatistické znaky,
- priestorové štatistické znaky,
- vecné štatistické znaky,
  - kvantitatívne štatistické znaky,
  - kvalitatívne štatistické znaky.

Príklady štatistických jednotiek a štatistických znakov:

štatistická jednotka („objekt", ktorý skúmame)	*štatistický znak* (*štatistická premenná*)	*príklad hodnoty štatistického znaku*
IQ test	bodový zisk dosiahnutý v teste	120 bodov
obyvateľ Popradu	národnosť	slovenská
zamestnanec UK	mesto trvalého bydliska	Šamorín
pacient nitrianskej nemocnice	krvný tlak	135/89
študent PriF UK	výška	181 cm
ESET, spol. s r.o.	DIČ	2020317068
Fakultná nemocnica Trenčín	počet lôžok k 1.1.2022	808
osobný automobil	typ motoru	benzínový
kuchynský stôl	hlavný použitý materiál	drevo
zimný semester na UK	dĺžka v týždňoch	13
plynomer	miesto montovania	suterén rodinného domu
Slovenská republika	dátum vstupu do EÚ	1.5.2004

Podľa podstaty a spôsobu merania rozlišujeme nasledovné **druhy štatistických znakov** [3]:

- kardinálne – môžu byť číselné alebo intervalové,
- nominálne (nazývame ich aj ako kategorické znaky) – vyznačujú sa tým, že nie je možné medzi nimi vytvoriť usporiadanie,
- ordinálne (nazývame ich aj ako poradové znaky),

- absolútne – napr. počty jedincov, počty udalostí, počty jednotiek a pod.,
- relatívne – napr. počty jedincov, počty udalostí, počty jednotiek a pod., ktoré sú vydelené celkovým počtom skúmaných objektov alebo celkovou veľkosťou skúmaného súboru.

Príklady štatistických znakov a ich charakterizácia:

(a) *celková spotreba vody v Bratislave v roku 2021* (vyjadrená v m^3) – vecný, kvantitatívny, kardinálny (číselný), absolútny štatistický znak;

(b) *stav elektromeru v konkrétnej domácnosti k 1.9.2022* (vyjadrený v kWh) – vecný, kvantitatívny, kardinálny (číselný), absolútny;

(c) *typ motoru osobného automobilu* – vecný, kvalitatívny, kategorický (nominálny);

(d) *výrobné číslo elektromeru v konkrétnej domácnosti* – vecný, kvalitatívny, ordinálny (*výrobné čísla sú zvyčajne prideľované podľa poradia vyrobeného zariadenia*);

(e) *miesto umiestnenia vodomeru* – priestorový;

(f) *dátum zavedenia eura na Slovensku* – časový;

(g) *počet novoprijatých študentov na študijný program Geografia, kartografia a geoinformatika na PriF UK za rok 2021* – vecný, kvantitatívny, kardinálny (číselný), absolútny;

(h) *bodový zisk dosiahnutý v IQ teste* – vecný, kvantitatívny, kardinálny (číselný), absolútny;

(i) *národnosť vybraného obyvateľa Popradu* – vecný, kvalitatívny, kategorický (nominálny);

(j) *mesto trvalého bydliska vybraného zamestnanca UK* – priestorový;

(k) *daňové identifikačné číslo (DIČ) spoločnosti ESET* – vecný, kvantitatívny, kategorický (nominálny);

(l) *dátum vstupu Slovenska do EÚ* – časový;

(m) *podiel počtu obyvateľov Slovenska vzhľadom na celkový počet obyvateľov EÚ k 1.1.2022* – vecný, kvantitatívny, kardinálny (číselný), relatívny;

(n) *umiestnenie Slovenska podľa počtu obyvateľov k 1.1.2022 v zozname členských štátov EÚ* – vecný, kvantitatívny, ordinálny (poradový), absolútny.

1.2 Dáta, údaje, typy údajov

Dáta reprezentujú informáciu o prvkoch zvoleného súboru. Podľa charakteru rozlišujeme **kvantitatívne** (číselné) dáta a **kvalitatívne** (nečíselné) dáta. Z trochu iného aspektu dáta môžeme deliť podľa nasledovne:

- **kardinálne (číselné) dáta**,
  - *absolútne*: počty jedincov, počty udalostí, počty jednotiek, a pod., napr.:
    - počet slnečných hodín v Hurbanove počas roku 2021,
    - počet študentov PriF UK k 1.10.2022,
    - výška hladiny rieky Dunaj v centre Bratislavy dňa 20.7.2022 o 8:15 hod.,
  - *relatívne* (proporcionálne): počty jedincov, počty udalostí, počty jednotiek a pod. vyjadrené vzhľadom na celkovú veľkosť súboru jedincov, udalostí alebo jednotiek, napr.:
    - relatívne zastúpenie hnedej pôdy na území Slovenskej republiky,
    - počet novoprijatých študentov PriF UK vzhľadom k celkovému počtu prihlásených uchádzačov v akademickom roku 2022/2023,
    - podiel počtu obyvateľov Nitrianskeho samosprávneho kraja v populácii SR k 1.1.2022,
  - *intervalové*: nadobúdajú hodnotu z nejakého intervalu, napr.:
    - teplota vzduchu v stupňoch Celzia podľa intervalového delenia: $\ldots$, $\langle 10{,}0; 14{,}9\rangle$, $\langle 15{,}0; 19{,}9\rangle$, $\langle 20{,}0; 24{,}9\rangle$, $\langle 25{,}0; 29{,}9\rangle$, $\ldots$, napríklad „*teplota vzduchu v centre Trnavy bola dňa 1.7.2022 o 10:00* $\langle 20{,}0; 24{,}9\rangle$ *stupňa Celzia*",
    - hrubá mesačná mzda zamestnancov v eurách podľa intervalového delenia: $\ldots$, $\langle 500; 999\rangle$, $\langle 1000; 1499\rangle$, $\langle 1500; 1999\rangle$, $\langle 2000; 2499\rangle$, $\ldots$, napríklad „*hrubá mesačná mzda Jozefa Mrkvičku v januári roku 2022 bola* $\langle 1000; 1499\rangle$ *eur*",
    - počet obyvateľov obcí k 1.1.2021 podľa intervalového delenia: $\langle 0; 499\rangle$, $\langle 500; 999\rangle$, $\langle 1000; 1999\rangle$, $\langle 2000; 2999\rangle$, $\langle 3000; 4999\rangle$, $\langle 5000; 9999\rangle$, $\langle 10000; 14999\rangle$, $\langle 15000; 19999\rangle$, $\langle 20000; 24999\rangle$, $\ldots$, napríklad „*počet obyvateľov Borinky k 1.1.2021 bol* $\langle 500; 999\rangle$ *ľudí*",
- **kategorické dáta**: máme vopred určený počet kategórií, do ktorých dáta rozdelíme,
  - *nominálne*: nie je možné ich usporiadať do poradia, napr.
    - slovné druhy (podstatné mená, prídavné mená, $\ldots$),
    - typ pôdy (hnedozem, kambizem, černozem, podzol, $\ldots$),
    - farba očí (hnedá, modrá, zelená, $\ldots$),
  - *ordinálne*: ide o také dáta, ktoré je možné usporiadať do poradia,
    - prieskum verejnej mienky s odpoveďami: *súhlasím / skôr súhlasím / nie som rozhodnutý / skôr nesúhlasím / nesúhlasím*,

    - veľkosť podniku podľa počtu zamestnancov (veľký podnik / stredný podnik / malý podnik / mikropodnik),
    - hodnosti v armáde (vojak / slobodník / desiatnik / ... / generál),
  - špeciálny typ kategorických dát: *binárne* (dichotomické) dáta – nadobúdajú iba dve možné hodnoty, napr. áno/nie, 0/1 a pod.

2 Opisné charakteristiky dátového súboru

Opisné charakteristiky slúžia na to, aby sme o skúmaných dátach získali základné informácie, ktoré sa dajú ľahko interpretovať a prezentovať. Nazývame ich aj ako **deskriptívne štatistiky** (*descriptive statistics*). Najčastejšie používame nasledovné tri typy opisných charakteristík [10]:

- **charakteristiky polohy** (*measures of location*):
  - zvyčajne popisujú akýsi „stred" štatistického súboru, no význam slova „stred" v tomto prípade nie je jednoznačný a závisí od typu dát resp. zvolenej charakteristiky,
  - medzi charakteristiky polohy patria najmä: výberový priemer, modus, medián, kvartily, decily a percentily (kvantily),
- **charakteristiky variability** (*measures of variability*):
  - poskytujú informácie o heterogenite dát, opisujú mieru rozptýlenosti údajov,
  - medzi charakteristiky variability patria najmä: výberový rozptyl, štandardná odchýlka, variačné rozpätie, medzikvartilové rozpätie a variačný koeficient,
- **charakteristiky tvaru** (*measures of shape*):
  - opisujú najmä mieru symetrie/asymetrie údajov, ako aj tvar rozdelenia dát okolo „stredu dátového súbor",
  - medzi charakteristiky tvaru patria najmä: koeficient šikmosti (koeficient asymetrie), koeficient špicatosti (koeficient strmosti, koeficient excesu).

Predpokladajme, že máme k dispozícii súbor absolútnych kvantitatívnych dát, t. j. merania o nejakom kardinálnom štatistickom znaku. Označme ich ako $x_1, x_2, x_3, \ldots, x_n$, kde $n \in \mathbb{N}$ značí počet dátových bodov (veľkosť súboru, prípadne rozsah výberu). Skúmaným kardinálnym štatistickým znakom môže byť napríklad výška 20-ročných slovenských mužov, vyjadrená v centimetroch. Potom veličina x_1 by označovala výšku prvého 20-ročného slovenského muža (napr. $x_1 = 191{,}4$ cm), ktorý sa zúčastnil štatistického experimentu. Ďalej, x_2 by označovala výšku druhého 20-ročného slovenského muža, ktorého údaj bol zaradený do nášho dátového súboru (napr. $x_2 = 180{,}4$ cm). Všeobecne, veličina x_i by označovala výšku i-teho 20-ročného slovenského muža (pre $i = 1, 2, 3, \ldots, n$), ktorý sa zúčastnil procesu merania telesnej výšky. Zaveďme aj ďalšie označenie: $X = (x_1, x_2, x_3, \ldots, x_n)$, kde X značí celý súbor údajov (hovoríme, že X je dátový vektor).

2.1 Charakteristiky polohy

Výberový priemer (*sample average*, *sample mean*) je zrejme najčastejšie používanou charakteristikou polohy. Označujeme ho $\bar{x}$ a počítame ho podľa vzorca:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{1}$$

Je dôležité si poznamenať, že výberový priemer je možné použiť len pre dáta číselného charakteru. Zo vzťahu (1) môžeme vidieť, že výberový priemer sa počíta ako jednoduchý aritmetický priemer údajov. Súvisiaca funkcia v MS Exceli: `AVERAGE`.

Príklad 1. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` máme prehľad o hrubých mesačných platoch v malej kartografickej firme pred a po prijatí špičkového odborníka. Vypočítajme výberové priemery hrubých mesačných platov.

Riešenie tohto príkladu je prezentované v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

Modus (v angličtine *mode*) je ďalšou charakteristikou polohy dátového súboru. Opisuje tú hodnotu, ktorá sa v štatistickom súbore **vyskytuje najčastejšie**. Súvisiace funkcie v MS Exceli: `MODE.SNGL`, `MODE.MULT`, prípadne aj `MODE` (v dávnejších vydaniach MS Excelu).

Vlastnosti a možnosti využitia modusu:

- modus dá sa použiť pre skoro všetky typy dát: číselné aj kvalitatívne, ako aj pre kardinálne, nominálne a ordinálne dáta,
- ide vlastne o jedinú charakteristiku polohy, ktorá sa dá „poctivo" použiť aj pre nominálne dáta,
- na rozdiel od iných charakteristík polohy nemusí byt určený jednoznačne; existujú totiž aj bimodálne alebo multimodálne štatistické súbory (v takýchto prípadoch hovoríme o viacnásobnom moduse),
- neodporúča sa použiť ako charakteristiku intervalových dát, a ani v prípadoch, keď máme relatívne veľa kategórií a relatívne málo dát,
- modus nie je citlivý na malý počet extrémnych či odľahlých hodnôt,
- Pozor! – MS Excel hľadá modus len pre dáta číselného charakteru,
- Pozor! – MS Excel má viacero funkcií, ktoré hľadajú modus (z dát číselného charakteru), odporúča sa používať funkciu `MODE.MULT`; existujú totiž aj dve iné funkcie: `MODE` (v starších verziách MS Excelu) a `MODE.SNGL`, ktoré ale vo výstupe poskytujú len jedinú hodnotu modusu, a to aj v prípade, že ide o multimodálny dátový súbor (v takýchto prípadoch teda dávajú nepresné a neúplné riešenie – vypíšu iba prvý modus, ktorý našli v súbore).

Príklad 2. Vo vybratej skupine ľudí 12 majú modré oči, 24 majú hnedé oči a 5 ich majú zelené. Za modus pokladáme hnedú farbu, pretože ide o najčastejšie sa vyskytujúci prvok v dátovom súbore. Môžeme si všimnúť, že v tomto prípade sú slová „poloha" a „stred" skôr obrazným vyjadrením, keďže ide o kvalitatívne (nečíselné) dáta (modré oči, hnedé oči, zelené oči), a nedá sa hovoriť o strede v geometrickom zmysle slova.

Príklad 3. Označme základoškolské vzdelanie číselným kódom 1, stredoškolské číslom 2, stredoškolské s maturitou číslom 3 a vysokoškolské vzdelanie číslom 4. Uvažujme aj vybranú skupinu 14 respondentov, u ktorých sme zisťovali ich najvyššie dosiahnuté vzdelanie. Štatistický súbor s týmito údajmi je možné nájsť v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. Vypočítajme modus tohto dátového súboru.

Riešenie tohto príkladu je prezentované v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

Príklad 4. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` máme prehľad o hrubých mesačných platoch v malej kartografickej firme pred a po prijatí špičkového odborníka. Vypočítajme modus hrubých mesačných platov (pred a po prijatí špičkového odborníka).

Riešenie tohto príkladu je prezentované v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

Medián (v angličtine *median*) je **prostredná hodnota v usporiadanom súbore údajov**. To znamená, že medián dátového súboru môžeme počítať tak, že dáta najprv usporiadame vzostupne (od najmenšej hodnoty po najväčšiu), a potom si zoberieme prostredný prvok usporiadaného súboru. Súvisiaca funkcia v MS Exceli: `MEDIAN`.

- Ak v dátovom súbore máme párny počet údajov (t. j. keď n je párne číslo), tak to znamená, že v usporiadanom súbore budeme mať až dva „prostredné prvky". Medián dátového súboru potom počítame ako ich aritmetický priemer.
- Ak v dátovom súbore máme nepárny počet údajov (t. j. keď n je nepárne číslo), tak v usporiadanom súbore budeme mať práve jeden „prostredný prvok". Medián dátového súboru sa potom rovná hodnote spomínaného prostredného prvku usporiadaného súboru.

Vlastnosti a možnosti využitia mediánu:

- medián je možné použiť pre všetky také typy dát, ktoré sa dajú usporiadať,
- nie je citlivý na malý počet extrémnych či odľahlých hodnôt,
- je určený jednoznačne,
- štatistické programy niekedy medián aproximujú (odhadujú), teda nie vždy sa jeho číselná hodnota zhoduje s vyššie uvedenou definíciou; rozdiely sú však minimálne.

Príklad 5. Označme základoškolské vzdelanie číselným kódom 1, stredoškolské číslom 2, stredoškolské s maturitou číslom 3 a vysokoškolské vzdelanie číslom 4. Uvažujme aj vybranú skupinu 14 respondentov, u ktorých sme zisťovali ich najvyššie dosiahnuté vzdelanie. Štatistický súbor s týmito údajmi je možné nájsť v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. Vypočítajme medián tohto dátového súboru.

Riešenie tohto príkladu je prezentované v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

Príklad 6. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` máme prehľad o hrubých mesačných platoch v malej kartografickej firme pred a po prijatí špičkového odborníka. Vypočítajme medián hrubých mesačných platov.

Riešenie tohto príkladu je prezentované v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

Usporiadané dátové súbory sa používajú nielen pri určení mediánu, ale pri niekoľkých ďalších charakteristikách polohy. Uvažujme teda súbor pozorovaní, v ktorom sú údaje usporiadané vzostupne (od najmenšej hodnoty po najväčšiu). Často používanými štatistickými ukazovateľmi polohy sú takzvané **kvartily** (*quartiles*), ktoré usporiadaný súbor delia na štvrtiny. Slová kvartil či quartile pochádzajú z latinského *quartus*: ‚*štvrtý, štvrtá časť*', resp. z latinského *quattuor*: ‚*štyri*' [2].

Prvý kvartil (*first quartile*) udáva tú hodnotu, pre ktorú platí, že zhruba jedna štvrtina údajov je od nej menšia alebo rovná, a zároveň zvyšné tri štvrtiny hodnôt sú od nej väčšie alebo rovné. Inými slovami: **prvý kvartil** je taký dátový bod, ktorý usporiadaný súbor rozdeľuje približne v pomere 25 % : 75 %.

Poznámka. Častice „*zhruba*" a „*približne*" sme v predchádzajúcom popise použili preto, pretože v reálnych situáciách sa iba málokedy stáva, že usporiadaný súbor je možné rozdeliť presne v pomere 1 : 3. Mierne technické problémy môžu spôsobiť viacnásobné pozorovania (rovnaké, opakujúce sa hodnoty v dátovom súbore), ako aj veľkosť súbor $n \in \mathbb{N}$ (nie v každom prípade je možné dosiahnuť rozdelenie na požadované veľkosti). □

Analogicky, druhý kvartil je taký dátový bod usporiadaného súboru, ktorý ho rozdeľuje v pomere 50 % : 50 %, t. j. prvá polovica údajov je od neho menšia alebo rovná, a zároveň druhá polovica hodnôt je od neho väčšia alebo rovná. **Druhý kvartil je totožný s mediánom**: veď oba tieto ukazovatele usporiadaný súbor rozdeľujú na dve rovnako veľké (rovnako početné) časti a je možné ich interpretovať ako „prostrednú hodnotu" usporiadaného súboru údajov. Kvôli tejto skutočnosti slovné označenie *druhý kvartil* bežne nepoužívame a prostredný prvok usporiadaného súboru nazývame len ako **medián**.

Tretí kvartil (*third quartile*) je ten údaj v usporiadanom dátovom súbore, pre ktorý platí, že zhruba tri štvrtiny údajov sú od neho menšie alebo rovné, a zároveň zvyšná jedna štvrtina hodnôt je

od neho väčšia alebo rovná. Platí teda, že tretí kvartil usporiadaný súbor rozdeľuje približne v pomere 75 % : 25 %.

Minimum je najmenší údaj v dátovom vektore, t. j. prvý prvok vzostupne usporiadaného dátového súboru. Technicky by sme mohli povedať, že minimum je vlastne nultý kvartil dátového súboru, pretože platí, že nula štvrtín dát je od neho menšia, a zároveň štyri štvrtiny údajov (t. j. všetky dáta) sú od neho väčšie alebo rovné. Podobným spôsobom môžeme definovať aj **maximum** ako najväčší údaj v dátovom vektore, t. j. posledný prvok vzostupne usporiadaného dátového súboru. Maximum môžeme pomenovať aj ako štvrtý kvartil dátového súboru (no toto slovné označenie zvyčajne nepoužívame).

Takzvané **decily** (v angličtine *deciles*) usporiadaný dátový súbor rozdeľujú na desať rovnako početných častí. Slovo *decil* pochádza z latinského *decimus*: ‚*desiaty*', resp. z latinského *decem*: ‚*desať*'. Pre prvý decil (*first decile*) platí, že jedna desatina údajov je od neho menšia alebo rovná, a zároveň zvyšných deväť desatín hodnôt je od neho väčších alebo rovných. Analogicky, druhý decil je taký dátový bod, ktorý usporiadaný súbor rozdeľuje v pomere 2 : 8 (t. j. v pomere 20 % : 80 %), atď. Minimum dátového súboru je totožný s nultým decilom, kým maximum je rovný desiatemu decilu.

Najvšeobecnejšími ukazovateľmi polohy, ktoré sa používajú pri charakterizácii číselných (kardinálnych) dátových súborov, sú takzvané **percentily** (*percentiles*). Kým kvartily usporiadaný súbor rozdeľujú na štvrtiny a decily na desatiny, percentily ho (teoreticky) delia na sto rovnako početných častí. Aj výraz *percentil* pochádza z latinčiny, a to z pojmu *per centum*: ‚*na stotiny*'. Štatistický odborný výraz *percentil* bol, samozrejme, vytvorený až neskôr (na konci 19. storočia). [2]

Kvartily, decily aj percentily poskytujú informácie o rozložení údajov v usporiadanom dátovom súbore (no robia to s inou „jemnosťou"). Súvisiace funkcie v MS Exceli sú nasledovné: `QUARTILE.EXC`, `QUARTILE.INC`, `PERCENTILE.EXC`, `PERCENTILE.INC`. Všimnime si, že pre výpočet decilov softvér MS Excel neponúka osobitné procedúry, no treba rovno dodať, že aj funkcie pre určenie kvartilov (`QUARTILE.EXC` a `QUARTILE.INC`) sú tak trochu nadbytočné. Kvartily a decily sú totiž iba špeciálnymi prípadmi percentilov, a teda všetky kvartily a decily je možné vypočítať pomocou percentilov (a teda v MS Exceli prostredníctvom funkcií `PERCENTILE.EXC` resp. `PERCENTILE.INC`):

- prvý kvartil sa rovná 25. percentilu dátového súboru,
- medián (druhý kvartil) je totožný s 50. percentilom, a zároveň sa rovná aj 5. decilu,
- tretí kvartil sa rovná 75. percentilu dátového súboru,
- prvý decil je totožný s 10. percentilom, druhý decil je rovný 20. percentilu atď.

Vo všeobecnosti môžeme zaviesť takzvanú **percentilovú funkciu**, ktorú v niektorých zdrojoch nazývajú aj ako **kvantilová funkcia** (*quantile function*; pozor, nie *kvartilová funkcia*, ale kva**n**tilová). Označujeme ju ako $q(X;\alpha)$ pre $\alpha \in \langle 0;1\rangle$, kde X dátový vektor (množina údajov). Funkčná hodnota $q(X;\alpha)$ nám určí číslo, ktoré rozdelí súbor tak, že $100 \times \alpha$ % dát je od neho menších alebo rovných, a zároveň zvyšných $100 \times (1-\alpha)$ % údajov je väčších alebo rovných. Napríklad:

- 25. percentil dátového súboru, $q(X;\alpha = 0{,}25) \triangleq q(X;0{,}25)$, rozdeľuje vektor údajov v pomere 25 % : 75 %; dátový bod $q(X;0{,}25)$ nazývame aj ako prvý kvartil,
- 50. percentil dátového súboru, $q(X;\alpha = 0{,}50) \triangleq q(X;0{,}50)$, rozdeľuje vektor údajov v pomere 50 % : 50 %; hodnotu $q(X;0{,}50)$ nazývame aj ako medián alebo 5. decil,
- 75. percentil dátového súboru, $q(X;\alpha = 0{,}75) \triangleq q(X;0{,}75)$, rozdeľuje vektor údajov v pomere 75 % : 25 %; dátový bod $q(X;0{,}75)$ nazývame aj ako tretí kvartil,
- 20. percentil (druhý decil) označujeme ako $q(X;\alpha = 0{,}20) \triangleq q(X;0{,}20)$; tento dátový bod vektor údajov rozdeľuje v pomere 20 % : 80 %, atď.,
- minimum dátového vektoru X je totožný s nultým percentilom, t. j. $\min(X) = q(X;0)$, kým maximum vektoru X sa dá formálne zapísať v tvare stého percentilu: $\max(X) = q(X;1)$.

Príklad 7. V tabuľke, ktorá je prezentovaná v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`, sú uvedené výšky príplatkov na stravovanie a cestovanie, ktoré firma ABC Geo s. r. o. vyplatila svojim zamestnancom počas kalendárneho roka 2021. Počet údajov je 137. Vypočítajme výberový priemer, modus, medián, prvý kvartil, tretí kvartil, všetkých 10 decilov, a aj 5., 25., 50., 75. a 95. percentil dátového súboru.

Riešenie tohto príkladu je prezentované v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

2.2 Charakteristiky variability

Charakteristiky polohy nie sú samy o sebe dostatočne dobrým popisom štatistického súboru. Napríklad súbory $X = (-3,-1,0,1,3)$ a $Y = (-128,-50,0,50,128)$ majú rovnaký aritmetický priemer, ale zjavne sa od seba líšia (z hľadiska rozptýlenosti údajov). O rozdieloch spôsobených mierou heterogenity dát hovoria charakteristiky variability.

Výberový rozptyl (*sample variance*) a **štandardná odchýlka** (*sample standard deviation*) sú najznámejšími charakteristikami variability, ktoré popisujú „priemerné vzdialenosti dát" od výberového priemeru. Výberový rozptyl sa označuje zápisom s^2 a je definovaný nasledovne:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\,. \tag{2}$$

Štandardnú odchýlku zvyčajne označujeme písmenom s a definujeme ako druhú odmocninu z výberového rozptylu:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{3}$$

Štandardnú odchýlku niekedy nazývame aj ako **smerodajná odchýlka**. Súvisiace funkcie v MS Exceli: `VAR.P`, `VAR.S`, `STDEV.P`, `STDEV.S`.

Vlastnosti výberového rozptylu a štandardnej odchýlky:

- nadobúdajú len nezáporné hodnoty,
- dajú sa použiť pre všetky také typy dát, pre ktoré sa dá vypočítať výberový priemer,
- obe tieto charakteristiky sú citlivé aj na malý počet extrémnych či odľahlých hodnôt,
- existujú aj vzorce, kde je v menovateli použitá veličina n namiesto výrazu $n-1$; ide ale len o technický (a málo podstatný) rozdiel,
- čím väčší je rozptyl (resp. štandardná odchýlka), tým menšia je homogenita dát (inými slovami: tým „ďalej" sú dáta od aritmetického priemeru).

Príklad 8. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` máme prehľad o hrubých mesačných platoch v malej kartografickej firme pred a po prijatí špičkového odborníka. Vypočítajme výberový rozptyl a štandardnú odchýlku, ktoré charakterizujú úroveň variability v dátových súboroch (pred a po prijatí špičkového odborníka). Ako by sa zmenili hodnoty výberového rozptylu a štandardnej odchýlky, keby firma namiesto špičkového odborníka prijala bežného zamestnanca, ktorý bude mať hrubý mesačný plat vo výške 1100 €?

Riešenie tohto príkladu je prezentované v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

Variačné rozpätie (*range*, RNG) je definované ako rozdiel medzi maximálnou a minimálnou hodnotou v štatistickom súbore. Formálne môžeme písať: $\mathrm{RNG}(X) = \max(X) - \min(X) \triangleq q(X;1) - q(X;0)$. Variačné rozpätie nadobúda len nezáporné hodnoty a je citlivé aj na malý počet extrémnych či odľahlých hodnôt.

Medzikvartilové rozpätie (*interquartile range*, IQR) je definované ako rozdiel medzi tretím a prvým kvartilom v štatistickom súbore. Formálne: $\mathrm{IQR}(X) = q(X;0{,}75) - q(X;0{,}25)$. Aj medzikvartilové rozpätie nadobúda len nezáporné hodnoty a je citlivé aj na malý počet extrémnych či odľahlých hodnôt. Pomocou percentilov môžeme definovať aj ďalšie charakteristiky variability, napr. medzidecilové rozpätie ($q(X;0{,}9) - q(X;0{,}1)$), medzipercentilové rozpätie ($q(X;0{,}99) - q(X;0{,}01)$) a ďalšie.

Variačný koeficient (*coefficient of variation*, CV) je definovaný ako podiel štandardnej odchýlky a výberového priemeru:

$$\mathrm{CV}(X) = \frac{s}{\bar{x}}. \tag{4}$$

Môže nadobúdať kladné aj záporné hodnoty, jeho znamienko závisí od znamienka priemeru (vo výnimočných prípadoch sa môže rovnať aj nule).

Príklad 9. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` máme prehľad o hrubých mesačných platoch v malej kartografickej firme pred a po prijatí špičkového odborníka.

Vypočítajme variačné rozpätie, medzikvartilové rozpätie a variačný koeficient, ktoré charakterizujú úroveň variability v dátových súboroch (pred a po prijatí špičkového odborníka). Ako by sa zmenili tieto hodnoty, keby firma namiesto špičkového odborníka prijala bežného zamestnanca, ktorý bude mať hrubý mesačný plat vo výške 1100 €?

Riešenie tohto príkladu je prezentované v excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

2.3 Charakteristiky tvaru

Koeficient šikmosti (*skewness*) označujeme symbolom g_1 (prípadne skratkou Skew) a definujeme predpisom:

$$g_1 \triangleq \mathrm{Skew}(X) = \sqrt{n}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}}. \tag{5}$$

Niekedy ho nazývame aj **koeficientom asymetrie**, pretože meria úroveň symetrie/asymetrie dát. Pre symetrické dáta nadobúda nulovú hodnotu; inak hovoríme o kladne resp. záporne zošikmených dátach. Dáta, ktoré sú rozdelené podľa Gaussovho normálneho rozdelenia, majú nulový koeficient šikmosti. Súvisiaca funkcia v MS Exceli: `SKEW`.

Koeficient špicatosti (*kurtosis*) označujeme symbolom g_2 (prípadne skratkou Kurt) a definujeme vzťahom:

$$g_2 \triangleq \mathrm{Kurt}(X) = n\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} - 3. \tag{6}$$

Koeficient špicatosti sa v niektorých zdrojoch nazýva aj ako koeficient strmosti alebo koeficient excesu. Slúži predovšetkým na porovnanie s normálne rozdelenými dátami. Údaje, ktoré sú rozdelené podľa Gaussovho normálneho rozdelenia, majú totiž nulový koeficient špicatosti. Ak v nejakom dátovom súbore vypočítame koeficient špicatosti, a ten padne do intervalu $\langle -3; 0)$, hovoríme, že dané dáta sú menej špicaté ako normálne rozdelené údaje. Ak vypočítaný koeficient špicatosti je väčší ako 0, hovoríme, že skúmané dáta sú viac špicaté ako normálne rozdelené údaje. Súvisiaca funkcia v MS Exceli: `KURT`.

Príklad 10. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` máme prehľad o hrubých mesačných platoch v malej kartografickej firme pred a po prijatí špičkového odborníka. Vypočítajme koeficient šikmosť a koeficient špicatosti (pred a po prijatí špičkového odborníka). Ako by sa zmenili hodnoty koeficientu šikmosti a koeficientu špicatosti, keby firma namiesto špičkového odborníka prijala bežného zamestnanca, ktorý bude mať hrubý mesačný plat vo výške 1100 €?

Riešenie je prezentované v súbore `Zaklady statistiky - Priklady a ulohy.xlsx`. □

3 Možnosti vizualizácie dát

3.1 Vizualizácia číselných premenných

Pri zobrazovaní rozdelenia číselných (kardinálnych) premenných sa často používajú takzvané **histogramy** (*histogram chart*). Ide o taký typ grafu, v ktorom sa zobrazujú frekvencie údajov v rámci rozdelenia. Jednotlivé stĺpce histogramu sa nazývajú rozsahy. Hodnoty znázornené na vodorovnej osi môžu mať bodový alebo intervalový charakter, na zvislej osi sú uvedené početnosti údajov.

Takzvaný **x-y graf** (*scatter plot*) je taký druh grafu, ktorý je vhodný na znázornenie vzťahu dvoch kvantitatívnych (číselných) premenných v 2-dimenzionálnom priestore. Používa sa aj na zobrazenie funkčného vzťahu medzi nezávislou premennou (x) a závislou premennou (y).

Pod slovným označením **škatuľový graf** (krabicový graf, *box plot*) rozumieme taký graf, ktorý zobrazuje minimálnu hodnotu, maximálnu hodnotu, medián, prvý kvartil a tretí kvartil dátového súboru. Jeho výhodou je, že pomocou neho môžeme zobraziť viacero opisných charakteristík kvantitatívnej premennej v jednom obrázku.

3.2 Vizualizácia realizácií časových radov

Stĺpcový diagram (*column chart*) je taký typ grafu, ktorý je vhodný na zobrazovanie hodnôt kvalitatívnej premennej za určitú dobu alebo na porovnávanie hodnôt viacerých premenných. Stĺpcové grafy je odporúčané používať najmä vtedy, ak potrebujeme **ilustrovať zmeny hodnoty premennej v čase** (v prípade realizácie časového radu) alebo porovnať hodnoty vo viacerých kategóriách (v prípade kategorickej premennej, pozri ďalšiu podkapitolu). V stĺpcových grafoch sú časové indexy resp. kategórie zvyčajne znázornené na vodorovnej osi a hodnoty sledovanej premennej (sledovaných premenných) na zvislej osi.

Poznámka. Obdobou stĺpcového grafu, v ktorom časové indexy resp. kategórie sú znázornené na zvislej osi a číselné hodnoty premennej na vodorovnej osi, nazývame **pruhový graf** (*bar chart*). □

Čiarový graf (*line chart*) je vo väčšine prípadov ideálny na zobrazenie trendov vývoja údajov v rovnakých intervaloch, napríklad mesiacoch, štvrťrokoch alebo rokoch. Na vodorovnej osi sú obvykle uvedené časové údaje alebo kategórie, kým na zvislej osi hodnoty sledovanej premennej.

3.3 Možnosti vizualizácie kategoriálnych premenných

Frekvenčná tabuľka (*frequency table*) je metodický nástroj, ktorý zachytáva rozdelenie početnosti hodnôt premennej a poskytuje informáciu o rozdelení **kategorickej premennej** alebo číselnej premennej rozdelenej do kategórií (prípadne intervalov). Frekvenčná tabuľka môže obsahovať absolútne a/alebo relatívne početnosti.

Na **kruhovom diagrame** (nazývame ho aj ako koláčový graf; v angličtine *pie chart*) je možné

zobraziť proporcionálnu veľkosť položiek k súčtu všetkých položiek. **Stĺpcový diagram** (*column chart*) je taký druh grafu, ktorý je vhodný na porovnávanie hodnôt kategoriálnych premenných. Stĺpcový graf zvyčajne zobrazuje kategórie pozdĺž vodorovnej osi (x = kategória) a hodnoty pozdĺž zvislej osi (y = hodnota).

Takzvaná **kontingenčná tabuľka** (*pivot table*) sa používa na prehľadné zobrazenie vzájomného vzťahu dvoch alebo viacerých štatistických premenných. Riadky kontingenčnej tabuľky zodpovedajú možným hodnotám prvej štatistickej premennej, stĺpce možným hodnotám druhej štatistickej premennej. V príslušnej bunke kontingenčnej tabuľky je uvedený počet takých prípadov, kedy zároveň mala prvá štatistická premenná hodnotu zodpovedajúcu príslušnému riadku a druhá štatistická premenná hodnotu zodpovedajúcu príslušnému stĺpcu. Kontingenčná tabuľka sa obvykle používa pri zobrazení vzájomného vzťahu kvalitatívnych alebo kategorických premenných.

4 Úlohy a cvičenia

Úloha 4.1. Uvažujme Slovenskú republiku (SR) ako štatistickú jednotku. V nižšie uvedenej tabuľke sú uvedené rôzne štatistické znaky (štatistické premenné) vzťahujúce sa k nami zvolenej štatistickej jednotke: Slovenskej republike. Charakterizujte uvedené štatistické znak podľa ich typu a podstaty.

štatistický znak	*charakterizácia*
rozloha územia	
počet obyvateľov k 1.1.2021	
počet novorodencov v roku 2021	
počet novorodencov v roku 2021	
rozloha Bratislavského kraja	
adresa Úradu vlády SR	
dátum vzniku SR	
IČO Štatistického úradu SR	
meno prezidentky SR k 1.1.2021	
výška hrubého domáceho produktu SR za rok 2021	

Úloha 4.2. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` sú uvedené výšky 40 faktúr, ktoré boli náhodne vybraté zo základnej množiny všetkých faktúr istej geoinformatickej firmy. Zobrazte tieto údaje na vhodných typoch grafov. Vypočítajte hodnoty nasledujúcich ukazovateľov a interpretujte vypočítané výsledky:

- výberový priemer, modus, medián,
- kvartily a decily,
- výberový rozptyl a štandardnú odchýlku,
- variačné rozpätie a medzikvartilové rozpätie,
- variačný koeficient,
- koeficient šikmosti
- koeficient špicatosti.

Úloha 4.3. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` sú uvedené údaje o počtoch zamestnancov (k dátumu 1.7.2022) 50 stavebných a geodetických firiem, ktoré boli náhodne vybraté z databázy takýchto spoločností. Zobrazte tieto údaje na vhodných typoch grafov, vypočítajte hodnoty nasledujúcich ukazovateľov a interpretujte vypočítané výsledky:

- výberový priemer, modus, medián,
- kvartily a decily,
- výberový rozptyl a štandardnú odchýlku,
- variačné rozpätie a medzikvartilové rozpätie,
- variačný koeficient,
- koeficient šikmosti,
- koeficient špicatosti.

Úloha 4.4. V excelovom súbore `Zaklady statistiky - Priklady a ulohy.xlsx` sú uvedené dáta o telesnej výške 180 mužov a 180 žien v dokončenom veku medzi 20 a 22 rokov. Údaje sú uvedené v centimetroch. Zobrazte tieto údaje na vhodných typoch grafov. Pre oba súbory osobitne vypočítajte hodnoty nasledujúcich ukazovateľov a interpretujte vypočítané výsledky:

- výberový priemer, modus, medián,
- kvartily a decily,
- výberový rozptyl a štandardnú odchýlku,
- variačné rozpätie a medzikvartilové rozpätie,
- variačný koeficient,
- koeficient šikmosti,
- koeficient špicatosti.

Zoznam použitej a odporúčanej literatúry

[1] Anděl, J.: *Základy matematické statistiky*. 2. vydanie, Matfyzpress, Praha, 2007, ISBN 80-7378-001-1.

[2] Etymonline.com [online]: *Online etymology dictionary*. Douglas Harper. [cit. 04.08.2022] Dostupné na adrese: `https://www.etymonline.com/`.

[3] Gavora, P. a kol. [online]: *Elektronická učebnica pedagogického výskumu*. Univerzita Komenského v Bratislave, Bratislava, 2010, ISBN 978–80–223–2951–4. [cit. 08.08.2021] Dostupné na adrese: `http://www.e-metodologia.fedu.uniba.sk/`.

[4] Chajdiak, J.: *Štatistika jednoducho v Exceli*. Statis, Bratislava, 2013, ISBN 978-80-85659-76-4.

[5] Králik, Ľ.: *Stručný etymologický slovník slovenčiny*. Vydavateľstvo VEDA, Jazykovedný ústav ĽŠ SAV, Bratislava, 2015, ISBN 978-80-224-1493-7.

[6] Kurzy UKF [online]: *Kurzy elektronického vzdelávania Univerzity Konštantína Filozofa v Nitre*. [cit. 08.08.2022] Dostupné na adrese: `https://amos.ukf.sk/mod/book/view.php?id=8404&chapterid=3151`.

[7] Microsoft Corporation [online]. *Excel help & learning*. [cit. 08.08.2022] Dostupné na adrese: `https://support.microsoft.com/en-us/excel`.

[8] Somorčík, J., Teplička, I.: *Štatistika zrozumiteľne*. Enigma, Slovensko, 2015, ISBN 978-80-8133-042-1.

[9] Wimmer, G.: *Štatistické metódy v pedagogike*. Gaudeamus, Hradec Králové, 1993, ISBN 80-7041-864-8.

[10] Zvára, K., Štěpán, J.: *Pravděpodobnost a matematická statistika*. 4. vydanie, Matfyzpress, Praha, 2006, ISBN 80-86732-71-1.