1. STATISTIKA

Statistika zkoumá jevy na dostatečně rozsáhlém souboru případů a hledá ty vlastnosti jevů, které se projeví až v souboru případů a ne na jednom případě.

Typickým příkladem je průměr známek ve škole z daného předmětu, průměrná volební účast ve volbách, procentuální zastoupení homosexuálně orientovaných jedinců v populaci, …

1.1 Statistický soubor, statistické jednotky, znak

Základním pojmem je **statistický soubor** a jeho prvky, které se nazývají **statistické jednotky**. Tento soubor vyšetřujeme z hlediska zvoleného **znaku** (nebo více znaků). Hodnota znaku musí být vždy jednoznačně stanovena. Existují přitom dva druhy znaků:

1. kvantitativní znak – jeho hodnota je určena číselnou hodnotou (výška postavy v cm, známka z matematiky, počet sourozenců, příjem uvedený v Kč, …);
2. kvalitativní znak – jeho hodnoty jsou dány kvalitou (povolání dané osoby, barva očí, pohlaví, rodinný stav, …), nejjednodušší kvalitativní znaky jsou přitom ty, které jsou dány určitým jevem a jeho opakem (muž - žena, prospěl - neprospěl, voják - nevoják, …); tyto znaky se nazývají alternativní znaky.

Příkladem statistického souboru je databáze žáků školy, ve které je o každém žákovi uložena řada informací – jméno, příjmení, datum narození, místo narození, adresa bydliště, číslo ISIC karty, známky na pololetních vysvědčeních, uvolnění z předmětů (např. z TV), … Jednotlivými statistickými jednotkami jsou pak jednotliví žáci a jejich „vlastnosti“. Znakem, z hlediska kterého statistický soubor vyšetřujeme, může být „narozen v květnu“, „bydlí v Praze“, „muž“, „průměr známek na všech vysvědčeních z matematiky je menší než 1,5“, …

1.2 Rozdělení četností a jeho grafické znázornění

1.2.1 Zavedení pojmů a jejich vysvětlení

Při statistickém šetření se zpravidla vyšetřuje více znaků, které nás zajímají jak každý zvlášť, tak i ve vzájemném vztahu.

Při předvolením průzkumu zajímá politické strany, jak bude kdo volit (jeden znak - „strana, kterou volím“). Je ale zajímavé také vědět, jak volí lidé z venkova a z větších měst („strana, kterou volím“ a „bydliště“), jak volí lidé různého vzdělání („strana, kterou volím“ a „nejvyšší dosažené vzdělání“) nebo náboženství („strana, kterou volím“ a „vyznání“), …

Pro další výklad se omezíme jen na šetření, ve kterých nás bude zajímat jen jeden znak. Výsledkem šetření tedy je seznam jednotek s udáním hodnoty znaku u každé z nich. Jsou-li jednotky v seznamu očíslovány 1, 2, …, n ($n \in \mathbb{N}$), pak jim odpovídající hodnoty znaku x označíme symboly $x_1, x_2, ..., x_n$. Často může znak nabývat jen určitého počtu r různých hodnot; tyto hodnoty znaku označíme symboly $x_1^*, x_2^*, ..., x_r^*$.

Ve statistickém souboru známek z matematiky na výročním vysvědčení 30 žáků jedné třídy nebude 30 různých hodnot ($n = 30$) ale jen pět ($r = 5$) - tj. jedna z pěti známek používaných ke klasifikaci.

Pro každou možnou hodnotu x_j^* zjistíme, kolikrát se vyskytla mezi znaky $x_1, x_2, ..., x_n$. Takto zjištěný počet n_j se nazývá **četnost** hodnoty x_j^*. Součet četností všech možných hodnot znaku se rovná počtu všech jednotek souboru, tj. platí:

$$\sum_{j=1}^{r} n_j = n \,. \tag{1}$$

Relativní četnost je pak definována vztahem

$$\nu_j = \frac{n_j}{n} \tag{2}$$

a udává, jaká část souboru má hodnotu znaku x_j^*. Na základě vztahů (1) a (2) je zřejmé, že platí:

$$\sum_{j=1}^{r} \nu_j = 1 \,. \tag{3}$$

Relativní četnost je možné vyjadřovat také v procentech; součet relativních četností je pak roven 100 %.

Rozdělení četností daného znaku statistického souboru můžeme graficky znázornit různými způsoby:

1. tabulkou rozdělení četností – zpravidla dvouřádková (resp. dvousloupcová) tabulka, v jejímž prvním řádku (resp. sloupci) jsou uvedeny hodnoty x_j^* znaku a ve druhém řádku (resp. sloupci) je uvedena četnost nebo relativní četnost (udaná desetinným číslem nebo v procentech);

2. spojnicovým diagramem (polygonem četností) – spojení bodů, jejichž první souřadnice je hodnota x_j^* znaku a druhá souřadnice je odpovídající četnost;
3. sloupkový diagram (histogram) – používá se zejména tehdy, jsou-li hodnoty znaku sdruženy do intervalů; tyto intervaly pak tvoří základny sloupků a odpovídající četnosti udávají výšky sloupků;

Do intervalů lze sdružit např. tělesnou výšku postavy udanou v centimetrech (vytvoříme intervaly 150 - 154, 155 - 159, 160 - 164, …), počty obyvatel ve městech (0 - 5000, 5001 - 10000, 10001 - 15000, …) a podobně.

4. kruhový diagram – různým hodnotám znaku odpovídají kruhové výseče, jejichž plošné obsahy (a tedy středové úhly) jsou úměrné četnostem (v tomto grafu se velmi často používají relativní četnosti vyjádřené v procentech).

Histogram se používá většinou pro znázornění rozdělení četností kvantitativních znaků a kruhový diagram se používá většinou pro znázornění rozdělení četností kvalitativních znaků.

1.2.2 Konkrétní ukázka

Máme k dispozici seznam známek z matematiky na pololetním vysvědčení v jedné třídě:

1,1,1,2,2,3,2,4,3,3,4,5,3,4,5,1,2,2,1,2,3,4,3,3,3. (4)

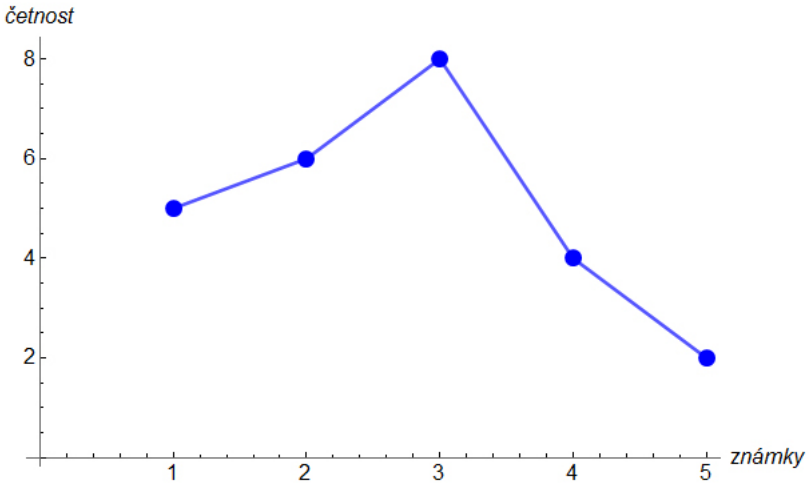
Pro některé statistické charakteristiky (uvedené v odstavci 1.3) je nutné řadu známek (4) seřadit vzestupně. Dostaneme tak řadu

1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,3,3,4,4,4,4,5,5. (5)

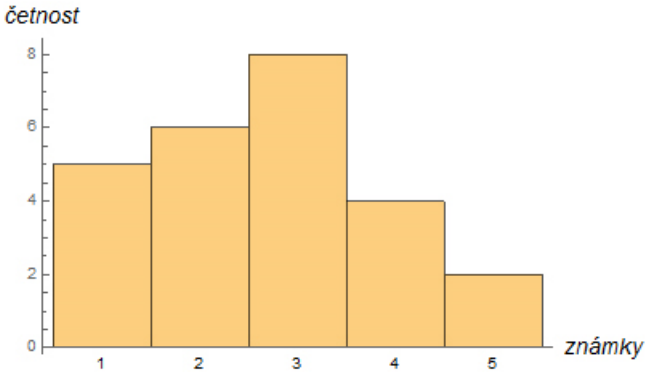
Tabulka rozdělení četností je zobrazena na obr. 1. Na základě ní je pak sestrojen polygon četností zobrazený na obr. 2. Jednotlivé body tohoto grafu odpovídají jednotlivým známkám a jejich četnostem.

Známka	Četnost
1	5
2	6
3	8
4	4
5	2

obr. 1



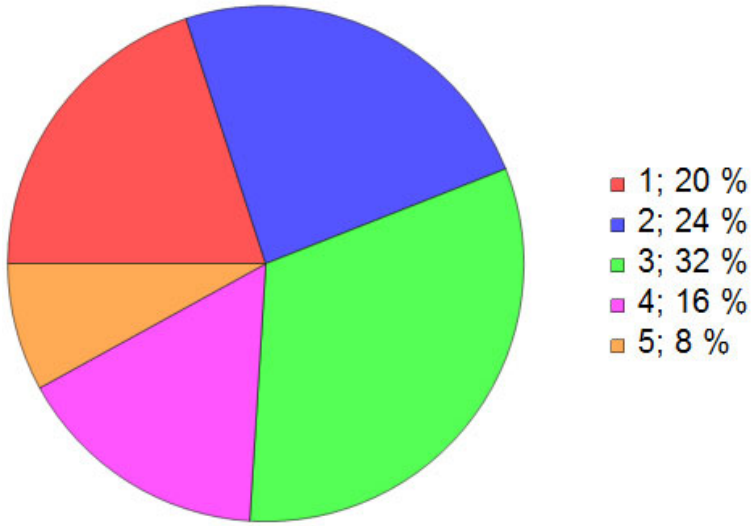
obr. 2



obr. 3

Na obr. 3 je pak zobrazen histogram se stejnou vypovídací hodnotou, jakou má polygon četností.

S využitím vztahu (2) je možné dopočítat relativní četnosti a zobrazit je v kruhovém diagramu (viz obr. 4).



obr. 4

1.3 Charakteristiky polohy a variability

Pro kvantitativní znaky lze zavést další charakteristiky, které jsou udány samostatnými čísly. Úplnou statistiku daného znaku podává rozdělení četnosti (viz odstavec 1.2), ale i čísla charakterizující polohu a variabilitu znaku jsou pro řadu aplikací důležitá.

1.3.1 Zavedení pojmů a jejich vysvětlení

Nejčastěji používanou **charakteristikou polohy** znaku x je **aritmetický průměr**.

ARITMETICKÝ PRŮMĚR $\bar{x}$ ZNAKU x JE SOUČET HODNOT ZNAKU ZJIŠTĚNÝCH U VŠECH JEDNOTEK SOUBORU DĚLENÝ POČTEM VŠECH JEDNOTEK SOUBORU:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{6}$$

Počítáme-li aritmetický průměr z tabulky rozdělení četností, je nutné každou hodnotu x_j^* násobit její četností. Vztah (6) je proto nutné upravit do tvaru:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{r} x_j^* n_j \,. \tag{7}$$

Určování průměru má smysl pouze v těch statistických souborech, ve kterých jsou individuální odchylky jednotlivých prvků souboru nahodilé. Ve statistických souborech, ve kterých jsou individuální odchylky jednotlivých prvků souboru systematické, je vhodnější určovat **průměrný přírůstek**.

Takovými soubory jsou typicky data získaná během fyzikálního měření. Jednotlivé odchylky jsou dány nepřesným odečtením hodnoty z měřidla, nepřesným nastavením hodnoty elektrického proudu, … Systematická chyba (špatná metoda měření, řádová chyba při čtení údaje z měřidla, …) se velmi rychle v případě fyzikálního měření pozná a lze ji tedy eliminovat (změnou metody měření, pečlivějším čtením z přístroje, …).

V některých případech se statistický soubor skládá z více dílčích souborů A, B, C a D. Přitom známe počty jednotek n_A, n_B, n_C a n_D v dílčích souborech a průměry $\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$ a $\bar{x}_D$ znaku x v dílčích souborech. Na základě znalosti těchto dat chceme určit průměr $\bar{x}$ v celém souboru.

Dílčími soubory mohou být např. žáci dvou paralelních tříd v ročníku střední školy a znakem x známky z matematiky na pololetním vysvědčení v jednotlivých třídách. Za dílčí soubory lze považovat známky z jednoho předmětu (např. matematika), které mají různé váhy (domácí úkol, zkoušení, test, pololetní práce, …).

Hledaný průměr určíme tak, že součet hodnot sledovaného znaku v celém souboru vydělíme počtem všech jednotek v souboru. Dostaneme tak vztah:

$$\bar{x} = \frac{\bar{x}_A \cdot n_A + \bar{x}_B \cdot n_B + \bar{x}_C \cdot n_C + \bar{x}_D \cdot n_D}{n_A + n_B + n_C + n_D} \,. \tag{8}$$

Vztah (8) se nazývá **vážený průměr** čísel $\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$ a $\bar{x}_D$ s váhami n_A, n_B, n_C a n_D. Tento vztah je možné zobecnit i pro větší počet dílčích souborů, než jsou uvedené čtyři.

Nyní se pokusíme vysvětlit, proč je právě *průměr* dobrou charakteristikou polohy znaku. Uvažujme, že každá zjištěná hodnota znaku je součtem dvou složek:

1. složka charakteristická pro celý soubor obsahující určitou globální informaci o tomto souboru;
2. individuální odchylka dané jednotky souboru, která má víceméně náhodný charakter.

Vypočítáme-li průměr, tak první složka vynikne, protože individuální odchylky, které jsou kladné i záporné, se navzájem téměř odečtou.

Právě uvedené můžeme ilustrovat na fyzikálním měření. Uvažujme měření průměru válcové součástky (viz tab. 1), Při měření ve fyzice se běžně uvádí první a druhý sloupec, třetí sloupec je doplněn pro ilustraci faktu, že individuální odchylky (vznikající chybou měření) se navzájem mají tendenci odečíst.

Číslo měření	$\frac{d}{\text{mm}}$	$\frac{\overline{d} + \triangle d}{\text{mm}}$
1	25,3	25,4 − 0,1
2	25,7	25,4 + 0,3
3	24,9	25,4 − 0,5
4	24,8	25,4 − 0,6
5	25,7	25,4 + 0,3
6	25,5	25,4 + 0,1
7	25,6	25,4 + 0,2
8	24,6	25,4 − 0,8
9	26,0	25,4 + 0,6
10	25,8	25,4 + 0,4

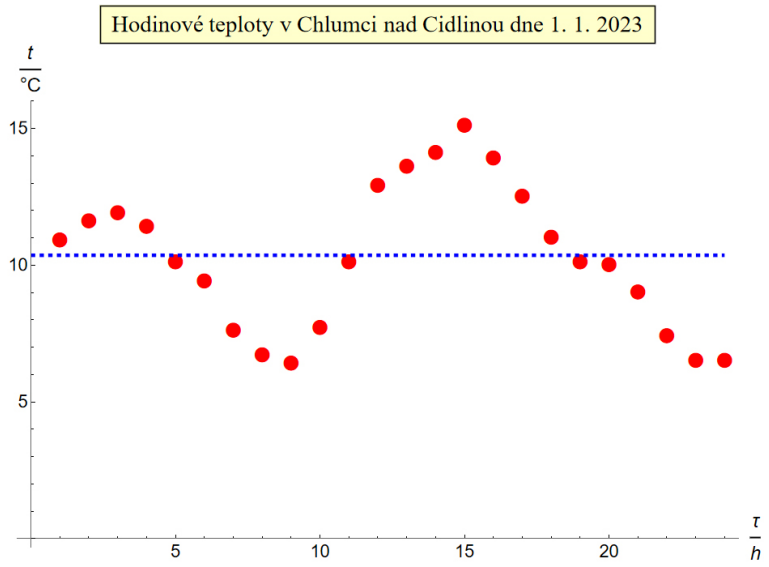
tab. 1

Z výše uvedeného důvodu není průměr vhodnou charakteristikou polohy v těch případech, v nichž individuální odchylky nejsou nahodilé, ale systematické. To je případ v časových řadách, v nichž data vykazují určitý trend, vývoj v čase; v tomto případě bývá lepším (zajímavějším) ukazatelem než průměr **průměrný přírůstek** (resp. **průměrný úbytek**) vyšetřovaného znaku za jedno časové období.

Jsou-li jednotlivá období očíslována 0, 1, 2, …, n, jsou-li x_0, x_1, x_2, …, x_n jim odpovídající hodnoty znaku a jsou-li $y_1 = x_1 - x_0$, $y_2 = x_2 - x_1$, …, $y_n = x_n - x_{n-1}$ přírůstky za jednotlivá období, pak pro průměrný přírůstek platí vztah

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{x_n - x_0}{n}. \tag{9}$$

Ze vztahu (9) plyne, že k výpočtu průměrného přírůstku stačí znát celkový přírůstek za sledovaný počet období a není nutné znát přírůstky za jednotlivá období.



obr. 5

Na obr. 5 je zobrazen průběh průměrných hodinových teplot na stanici v Chlumci nad Cidlinou ze dne 1. 1. 2023. Průměrná teplota toho dne byla $10{,}4\,°\mathrm{C}$ (v grafu na obr. 5 je zobrazena modrou přerušovanou čarou), zatímco průměrný přírůstek (resp. úbytek) byl $-0{,}2\,°\mathrm{C}$.

Pro některé časové řady zobrazující zejména národohospodářské údaje (inflace, zisk, …) je vhodné udávat **průměrné tempo růstu** za jedno období. Tím je myšlen průměr podílů hodnot za dvě po sobě následující období, tedy průměr podílů $z_1 = \frac{x_1}{x_0}$, $z_2 = \frac{x_2}{x_1}$, …, $z_n = \frac{x_n}{x_{n-1}}$. V tomto případě se počítá **geometrický průměr** pomocí vztahu

$$\overline{z}_G = \sqrt[n]{z_1 \cdot z_2 \cdot \ldots \cdot z_n}\,. \tag{10}$$

Geometrický průměr se definuje pouze pro kladná čísla.

Dosazením do vztahu (10) pro průměrné tempo růstu získáme vztah $\overline{z}_G = \sqrt[n]{\frac{x_n}{x_0}}$.

Dalším typem průměru, který se pro kladná čísla x_1, x_2, …, x_n používá, je **harmonický průměr** definovaný vztahem

$$\overline{x}_H = \frac{1}{\frac{1}{n}\cdot\left(\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}\right)}. \tag{11}$$

Posledním typem průměru, který lze zavést, ale který se běžně ve statistice příliš nepoužívá, je kvadratický průměr definovaný vztahem

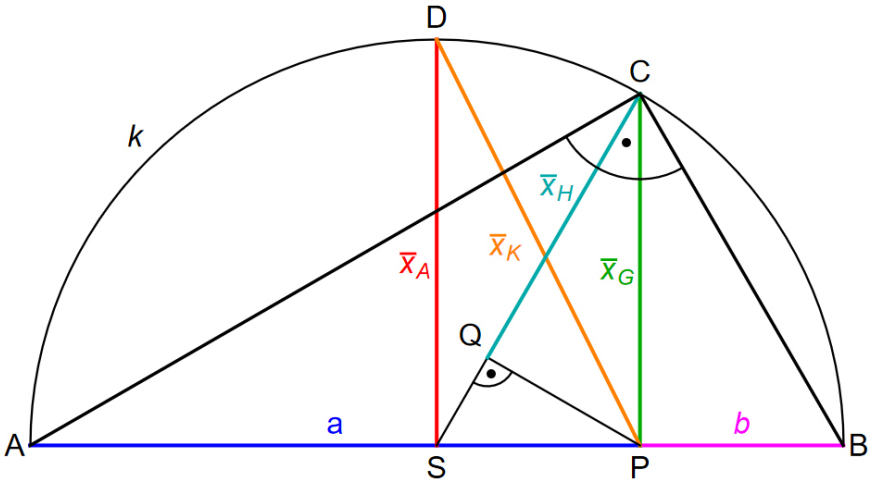
$$\overline{x}_K = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}\,. \tag{12}$$

Všechny průměry, které byly výše popsány, jsou schematicky pro dvě kladná čísla reprezentovaná úsečkami délek a a b zobrazeny na obr. 6. S využitím tohoto obrázku lze i na základě geometrických úvah a vztahů pro daná kladná čísla a a b odvodit vztahy vyplývající ze vztahů (6), (10), (11) a (12).

Pro zobrazené průměry přitom platí vztah

$$\overline{x}_K \geq \overline{x}_A \geq \overline{x}_G \geq \overline{x}_H\,, \tag{13}$$

přičemž rovnost všech průměrů nastává pouze pro situaci, kdy $a = b$.



obr. 6

Dalšími charakteristikami polohy jsou **modus** a **medián**.

MODUS ZNAKU x SE ZNAČÍ $\mathrm{Mod}(x)$ A UDÁVÁ HODNOTU ZNAKU x S NEJVYŠŠÍ ČETNOSTÍ.

MEDIÁN ZNAKU x SE ZNAČÍ $\mathrm{Med}(x)$ A UDÁVÁ PROSTŘEDNÍ HODNOTU ZNAKU x, JSOU-LI HODNOTY $x_1, x_2, \ldots, x_n$ USPOŘÁDÁNY VZESTUPNĚ PODLE VELIKOSTI. PŘITOM PLATÍ:

$\mathrm{Med}(x) = x_{\frac{n+1}{2}}$ **JE-LI n LICHÉ;**

$\mathrm{Med}(x) = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$ **JE-LI** n **SUDÉ.**

Medián se užívá jako charakteristika polohy zejména v těch souborech, ve kterých hodnoty znaku u některých jednotek extrémně vybočují z řady ostatních hodnot (viz odstavec 1.3.2).

Každá charakteristika polohy (aritmetický průměr, modus, medián) je určena číslem, kolem kterého jednotlivé hodnoty znaku kolísají. Míru tohoto kolísání vyjadřují **charakteristiky variability** (**charakteristiky proměnlivosti**) znaku.

Jako charakteristika variability k aritmetickému průměru (jakožto charakteristice polohy) se většinou používá **rozptyl**.

ROZPTYL s_x^2 **ZNAKU** x **JE DEFINOVANÝ JAKO PRŮMĚR DRUHÝCH MOCNIN ODCHYLEK OD ARITMETICKÉHO PRŮMĚRU:**

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 . \tag{14}$$

Je-li rozptyl počítán na základě tabulky četností, pak vztah (14) přejde na vztah

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{r}\left(x_j^* - \bar{x}\right)^2 n_j . \tag{15}$$

Pokud provedeme ve vztahu (14) naznačené umocnění, získáme postupně:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\cdot n\cdot\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\cdot\bar{x} + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 .$$

Analogickou úpravu můžeme provést i se vztahem (15).

Další charakteristikou je **směrodatná odchylka**.

SMĚRODATNÁ ODCHYLKA s_x **JE DEFINOVANÁ JAKO DRUHÁ ODMOCNINA Z ROZPTYLU:**

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} . \tag{16}$$

Směrodatná odchylka charakterizuje variabilitu znaku ve stejných jednotkách, v jakých byly udány hodnoty znaku, zatímco rozptyl je vyjádřen ve druhé mocnině těchto jednotek. Proto je pro technická měření (fyzika, chemie, …) vhodnější používat směrodatnou odchylku naměřených dat.

Bezrozměrným číslem (vyjadřujícím se v procentech), které charakterizuje variabilitu znaku x, je **variační koeficient**.

VARIAČNÍ KOEFICIENT v_x **JE DEFINOVÁN JAKO PODÍL SMĚRODATNÉ ODCHYLKY A ARITMETICKÉHO PRŮMĚRU:**

$$v_x = \frac{s_x}{\bar{x}} \cdot 100\,\% . \tag{17}$$

Z definice je zřejmé, že variační koeficient má smysl pouze tehdy, pokud jsou hodnoty znaku x nezáporné.

V případě, že statistický soubor bude tvořen naměřenými daty (fyzikálními, chemickými, …), představuje variační koeficient relativní chybu (relativní odchylku) daného měření.

Poslední charakteristikou variability jsou **kvantily**.

KVANTIL x_p **URČUJE HODNOTU ZNAKU, PRO KTEROU PLATÍ, ŽE NEJMÉNĚ** p **PROCENT PRVKŮ DANÉHO STATISTICKÉHO SOUBORU MÁ HODNOTU MENŠÍ NEBO ROVNOU** x_p **A** $100 - p$ **PROCENT PRVKŮ SOUBORU MÁ HODNOTU VĚTŠÍ NEBO ROVNOU** x_p**.**

V praxi se používají tyto kvantily:

1. x_{50} je medián;

Definice uvedená výše říká, že medián je „prostřední“ prvek vzestupně uspořádaného souboru. To ale znamená, že 50 % hodnot souboru je menších nebo rovných mediánu a zbývajících 50 % hodnot je větších nebo rovných mediánu.

2. x_{25} je dolní kvartil;
3. x_{75} je horní kvartil;
4. $x_1, x_2, x_3, ..., x_{99}$ jsou percentily.

Při různých soutěžích, výběrových řízeních, přijímacích zkouškách, … se používají k popisu úspěšnosti daného uchazeče právě percentily. Je-li tedy např. percentil uchazeče 89 (tj. je na 89tém percentilu), znamená to, že 89 % uchazečů je horších než on a pouze 11 % je lepších než on. Pokud je navíc předem jasné, že úspěšných v daném řízení bude nejlepších 90 % účastníků, má tento účastník jisté, že je mezi úspěšnými.

Budeme-li pracovat se vzestupně uspořádaným statistickým souborem, který bude mít n statistických jednotek, je kvantil x_p roven prvku stojícímu na k-tém místě v souboru. Přitom k získáme ze vztahu

$$k = \frac{n \cdot p}{100} \tag{18}$$

tak, že výsledek uvedeného podílu zaokrouhlíme vždy nahoru.

1.3.2 Konkrétní ukázka

Vrátíme-li se nyní ke známkám z matematiky (4) uvedených v odstavci 1.2.2, můžeme dopočítat další charakteristiky tohoto souboru.

Průměrná známka (vypočtená podle vztahu (6) je 2,68.

Medián známek je 3 - jedná se o třináctou známku ve vzestupně seřazeném seznamu známek (5), kterých je celkem 25. Pokud budeme chápat medián jako kvantil x_{50}, pak můžeme podle vztahu (18) spočítat pořadí známky, která je mediánem: $k = \frac{25 \cdot 50}{100} = \frac{1250}{100} = 12{,}5$; po zaokrouhlení nahoru dostáváme tedy $k \doteq 13$. Získali jsme tedy stejné pořadí jako s využitím definice mediánu.

Modus známek je 3, protože trojka se vyskytuje v seznamu známek nejčastěji (viz též tabulka četností na obr. 1).

Rozptyl známek je podle vztahu (14) roven 1,42.

Směrodatná odchylka je na základě vztahu (16) rovna 1,19.

Variační koeficient známek vypočtený podle vztahu (17) je roven 0,44.

Z praktického hlediska je někdy vhodnější udávat medián než průměr. V některých případech má totiž lepší vypovídací hodnotu.

Uvažujme tento soubor čísel: 1, 1, 1, 2, 3, 3, 4, 4, 5, 6, 8, 120, 150, 200, 242.

Průměr těchto čísel je 50 a medián je 4.

V případě souboru čísel 44, 45, 46, 47, 48, 48, 49, 50, 51, 52, 52, 53, 54, 55, 56, který má stejný počet statistických jednotek, je průměr 50 a medián také 50.

U druhé řady čísel mají průměr a medián stejnou vypovídací hodnotu. A to ne proto, že jsou obě hodnoty stejné, ale proto, že jsou přibližně stejné jako jednotlivá čísla v tomto souboru (směrodatná odchylka je rovna 3,56).

U první řady je ale většina čísel výrazně menší, než je průměr těchto čísel. Proto je medián lepší charakteristikou tohoto souboru, protože na základě něj lze získat představu o hodnotách čísel (víme, že na prostředním místě vzestupně seřazeného souboru čísel je číslo 4, které je typickým číslem souboru, zatímco průměr je více než 10krát vyšší; směrodatná odchylka je 80,89).

I z toho důvodu je např. údaj o průměrné mzdě obyvatel v dané zemi nevhodný. Výrazně lepší by bylo udávat medián platu obyvatel dané země.

1.4 Korelace

Minulé úvahy byly vedeny pro případ, kdy jsme popisovali:

1. jeden znak souboru;
2. více znaků, které jsme popisovali odděleně.

Dvěma oddělenými znaky mohou být známky žáků jedné třídy z matematiky a češtiny, délka a šířka školních hřišť v Praze, …

Nyní se budeme vyšetřovat popis dvojice znaků (x, y); výsledkem provedeného statistického šetření jsou přitom v souboru s délkou n data ve tvaru (x_1, y_1), (x_2, y_2), …, (x_n, y_n).

Může se jednat o:

data evropských zemí o roční spotřebě alkoholu na jednoho obyvatele a data o ročním úmrtí na cirhózu jater;

data o výsledcích státních maturitních zkoušek z matematiky a počtu žáků, kteří navštěvovali technicky zaměřené obory;

data o výskytu hrabošů na území jednotlivých krajů České republiky a počtu poničených hektarů půdy osetých obilím;

data o volební účasti obyvatelů jednotlivých krajských měst v prezidentských volbách a barvě domů dotazovaných voličů;

…

Kromě charakteristik polohy a variability, počítaných pro každý z obou znaků odděleně, je nutné v tomto případě znát i **míru statistické závislosti obou znaků**. Zvolíme-li za znaky polohy průměry $\overline{x}$ a $\overline{y}$ obou sledovaných znaků a za charakteristiky variability směrodatné odchylky s_x a s_y těchto znaků, pak za společnou charakteristiku se velmi často volí **koeficient korelace** r_{xy}, který je definován vztahem:

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right)}{s_x \cdot s_y}. \tag{19}$$

Čitatele zlomku lze postupně upravit do tvaru: $\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) =$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i y_i - x_i \overline{y} - \overline{x} y_i + \overline{x} \cdot \overline{y}\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i y_i\right) - \frac{1}{n} \cdot \overline{y} \cdot \sum_{i=1}^{n} x_i - \frac{1}{n} \cdot \overline{x} \cdot \sum_{i=1}^{n} y_i + \overline{x} \cdot \overline{y} =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i y_i\right) - \overline{x} \cdot \overline{y} - \overline{x} \cdot \overline{y} + \overline{x} \cdot \overline{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i y_i\right) - \overline{x} \cdot \overline{y}.$$

Na základě těchto úprav lze vztah (19) psát ve tvaru:

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i y_i\right) - \overline{x} \cdot \overline{y}}{s_x \cdot s_y}, \tag{20}$$

který je pro manuální výpočty vhodnější (jednodušší).

Z obou vztahů (19) i (20) vyplývá, že koeficient korelace je bezrozměrná charakteristika daného statistického souboru; přitom platí $r_{xy} \in \langle -1; 1 \rangle$. Krajních hodnot intervalu nabývá koeficient korelace tehdy, je-li mezi znaky *x* a *y* **funkční závislost** (tedy nejenom statistická), a to lineární závislost ve tvaru:

$$y = a \cdot x + b \tag{21}$$

pro $a \in \mathbb{R} \setminus \{0\}$ a $b \in \mathbb{R}$. Znaménko koeficientu *b* přitom určuje, zda koeficient korelace bude nabývat hodnoty 1 nebo -1.

Vzhledem k rozdílu v čitateli zlomku ve vztahu (20) je zřejmé, že koeficient korelace r_{xy} může nabývat kladné i záporné hodnoty:

Je-li statistická závislost mezi znaky *x* a *y* taková, že nadprůměrným hodnotám *x* zpravidla odpovídají nadprůměrné hodnoty *y*, pak bude v čitateli zlomku ve vztahu (19) většina součinů kladných, a tedy i koeficient korelace bude **kladný**.

V souboru zaměstnanců podniku znaky počet let v podniku a roční příjem; v souboru evropských zemí spotřeba cigaret na jednoho obyvatele a počet úmrtí za rok na rakovinu plic; …

Jestliže nadprůměrným hodnotám znaku *x* odpovídají zpravidla podprůměrné hodnoty znaku *y* a naopak, bude v čitateli zlomku ve vztahu (19) většina součinů záporná, a tedy i koeficient korelace bude **záporný**.

V souboru obyvatel České republiky míra nezaměstnanosti v jednotlivých okresech a výše vkladů obyvatel u peněžních ústavů; v souboru úmrtí osob v daném roce spotřeba cigaret a věk, kterého se zemřelá osoba dožila; …

Není-li mezi znaky *x* a *y* žádná závislost, budou mít kladné i záporné součiny v čitateli zlomku ve vztahu (19) tendenci se navzájem rušit; koeficient korelace tedy bude **blízký nule**.

V souboru volebního průzkumu: volební účast a barva domu daného respondenta; v souboru zaměstnanců podniku: tělesná hmotnost a výše měsíčního příjmu; …

Ačkoliv jsme výše uvedené úvahy prováděli pouze na základě rozboru čitatele zlomku vztahu (19), je jmenovatel uvedeného vztahu nutný proto, aby koeficient korelace nezávisel na násobku jednotky, v níž sledované znaky *x* a *y* uvádíme.

Bude-li sledovaným znakem *x* měsíční příjem obyvatel a znakem *y* spotřeba pohonných hmot na čerpacích stanicích, tak koeficient korelace musí být nezávislý na tom, zda měsíční příjem bude uveden v korunách nebo v tisících korun. Popsanou změnou jednotky se čitatel vztahu (19) 1000krát zmenší, ale současně se 1000krát zmenší jeho jmenovatel. Koeficient korelace zůstane tedy beze změny.

Mírně odlišná situace nastane, pokud sledované znaky mohou nabývat pouze určitých hodnot. Uvažujme situaci, kdy znak x nabývá r různých hodnot označených symboly x_1^*, x_2^*, …, x_r^* a znak y nabývá pouze s různých hodnot označených symboly y_1^*, y_2^*, …, y_s^*. Do této skupiny patří i případy, kdy hodnoty znaků x a y sdružují do určitých intervalů; hodnoty x_j^* a y_k^* jsou pak středy těchto intervalů.

Pro každou možnou dvojici hodnot $\left\{x_j^*; y_k^*\right\}$ zjistíme, kolikrát se vyskytla mezi daty $\left\{x_1; y_1\right\}$, $\left\{x_2; y_2\right\}$, …, $\left\{x_n; y_n\right\}$; to znamená, že určíme její četnost n_{jk} v souboru n jednotek. Získáme tím **rozdělení četností dvojice znaků** (x, y) – viz tab. 2.

x \ y	y_1^*	y_2^*	…	y_s^*
x_1^*	n_{11}	n_{12}	…	n_{1s}
x_2^*	n_{21}	n_{22}	….	n_{2s}
…	…	…	…	…
x_r^*	n_{r1}	n_{r2}	….	n_{rs}

tab. 2

Další postup při hledání koeficientu korelace je tento:

1. sečteme četnosti v jednotlivých řádcích tab. 2, čímž získáme rozdělení četností samotného znaku x, a z tohoto rozdělení vypočítáme průměrnou hodnotu $\overline{x}$ a směrodatnou odchylku s_x;
2. sečteme četnosti v jednotlivých sloupcích tab. 2, čímž získáme rozdělení četností samotného znaku y, a z tohoto rozdělení vypočítám průměrnou hodnotu $\overline{y}$ a směrodatnou odchylku s_y;
3. vypočítáme součin $\sum_{i=1}^{n}\left(x_i y_i\right)$; každý ze součinů $x_i y_i$ je přitom roven některému součinu $x_j^* y_k^*$ a to s četností n_{jk}, takže můžeme psát

   $$\sum_{i=1}^{n}\left(x_i y_i\right) = x_1^* y_1^* n_{11} + x_1^* y_2^* n_{12} + ... + x_1^* y_s^* n_{1s} + ... + x_r^* y_1^* n_{r1} + x_r^* y_2^* n_{r2} + ... + x_r^* y_s^* n_{rs};$$

4. vypočítáme výraz $= \frac{1}{n} \cdot \sum_{i=1}^{n}\left(x_i y_i\right)$ a spolu s vypočtenými hodnotami $\overline{x}$, $\overline{y}$, s_x a s_y dosadíme do vztahu (20) a získáme tak koeficient korelace.