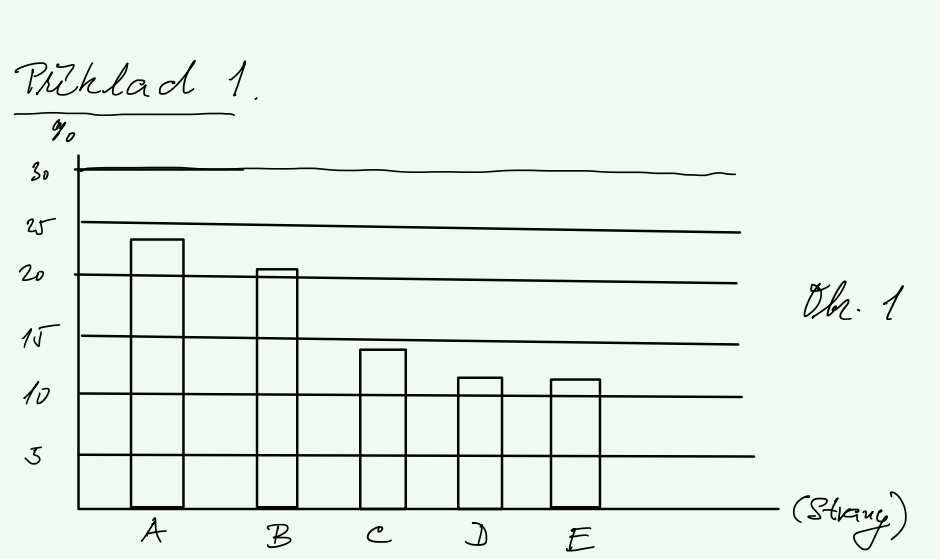
L1_Popisna_statistika

neděle 27. září 2020 15:21

Příklad 1.



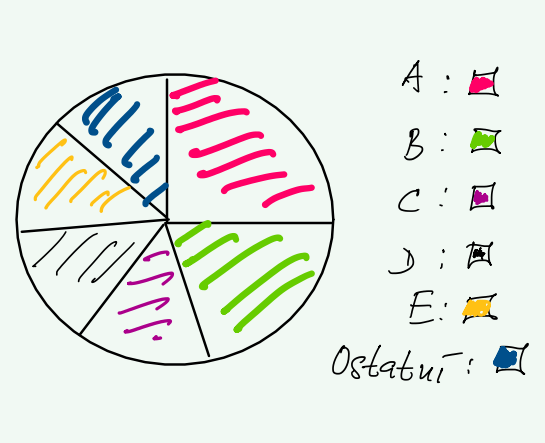
Předpokládejme, že ve volbách do parlamentu bylo odevzdáno celkem 2,521,879 hlasů a stranu A volilo 612,632 voličů, stranu B volilo 534,852 voličů, stranu C volilo 369,236, stranu D volilo 316,456 voličů a stranu E volilo 312,839 voličů.
Pak lze výsledek prezentovat pomocí sloupcového grafu (Obr. 1).

Výsledky v procentech:

A: 24,3%, B: 21,3%, C: 14,6%

D: 12,5%, E: 12,4%

Jinou možností je „koláčový graf":



Příklad 2. Sociologická firma realizovala předvolební průzkum ve kterém oslovila 1000 náhodně vybraných voličů. V průzkumu se voličů ptala, kterou stranu by volili, kdyby volby proběhly dnes. Zjistilo se, že stranu A by volilo 203 z dotázaných, stranu B by volilo 201 dotázaných, stranu C by volilo 160 z dotázaných, stranu D by volilo 134 z dotázaných a stranu E by volilo 120 z dotázaných.

Úkol: Načrtněte sloupcový a koláčový graf prezentující výsledky výzkumu.

Komentář: Situace v předchozích příkladech se zásadně liší. Zatímco v prvním případě máme kompletní informaci o statistickém souboru všech voličů z hlediska jejich voličských preferencí, ve druhém příkladu máme informaci o voličských preferencích pouze u vzorku 1000 voličů. Vzniká tedy otázka jak by se lišily výsledky průzkumu, kdybychom oslovili jiný vzorek 1000 voličů. Jinak řečeno, do jaké míry lze výsledky průzkumu aplikovat na celý statistický soubor.
Toto je zásadní otázkou statistiky. Zobecňování zjištěných vlastností vzorku na celý statistický soubor se bude nazývat statistickou inferencí.

Tzv. statistická šetření mají sloužit jako podklad pro rozhodovací proces ve firmách, státních úřadech apod. Celý proces lze vidět ve čtyřech fázích:

1. výběr statistického vzorku;
2. zvolení otázky;
3. statistická inference;
4. rozhodnutí na základě interpretace statistických dat.

MEDIÁN

Příklad 3. Předpokládejme, že máme zakoupeny podílové listy 7 podílových fondů a máme data o ročním výnosu těchto fondů (v procentech):

Fond	A	B	C	D	E	F	G
Výnos	2,7	9,2	11,4	4,6	5,2	5,6	-2,4

Komentář. Jak je vidět, tak nejmenší výnos dosahoval fond G (-2,4%) a největší výnos v daném roce dosahoval fond C (11,4%).
Vidíme tedy v jakém rozsahu se pohybuje vývoj výnosnosti sledovaných podílových fondů. Je pochopitelné, že tyto extrémní hodnoty nám nedávají obrázek o rozložení hodnot. Potřebujeme tedy znát další parametry našeho vzorku. Jedním z těchto parametrů je tzv. medián.
Medián je zhruba řečeno hodnota, pro kterou platí, že polovina pozorovaných hodnot je menších než je hodnota mediánu a polovina pozorovaných hodnot je větší než je hodnota mediánu.

Definice. Mediánem souboru n hodnot/pozorování seřazených vzestupně budeme rozumět:

- hodnotu s pořadovým číslem $\frac{n+1}{2}$ je-li n liché.
- střed mezi $\frac{n}{2}$-tou hodnotou a $\frac{n}{2}+1$-ní hodnotou v souboru.

Příklad 4. Najděme medián souboru hodnot:
1,5%; 2,3%; -3,4%; -5,6%; 0,3%; -3,4%; 3,2%; 2,2%.

Řešení. Za prvé vidíme, že počet hodnot v souboru je $n = 8$. Tedy n je sudé.
Seřadíme tyto hodnoty vzestupně: -5,6%; -3,4%; -3,4%; 0,3%; 1,5%; 2,2%; 2,3%; 3,2%.

Jelikož je počet hodnot sudé číslo, je mediánem střed mezi 4. a 5. hodnotou, tj.

$$\text{Medián} = \frac{0,3+1,5}{2}\,\% = \boxed{0,9\%}.$$

Komentář. Užitečnost mediánu se jeví užitečným, pokud pracujeme s mnohem početnějším souborem hodnot. Pokud bychom měli např. k dispozici hodnoty o změně ceny 2000 akcií nebudeme zobrazovat všechny hodnoty. Např. se ukáže, že nejmenší hodnotou je -10,4%, největší hodnotou je 12,5% a medián je hodnota -9,5%. Pak lze usoudit, že nejméně polovina akcie ztratily svoji původní hodnotu, dokonce o více než -9,5%. Bohužel nebudeme vidět, kolik akcií zdražily nebo třeba kolik akcií zdražilo o více než 10%. Abychom mohli odpovědět na tyto otázky, budeme potřebovat další informace kromě znalosti mediánu.

KVARTILY A MODUS

Jednoduše řečeno, kvartily souboru n hodnot najdeme pokud vzestupně seřazený soubor rozdělíme na čtyři stejně početné části. Hodnota dělící první a druhou část bude prvním kvartilem. Hodnota dělící prostřední skupiny hodnot bude vlastně medián. Hodnota oddělující poslední dvě skupiny hodnot bude nazvána třetím kvartilem.

Speciálně, pokud bude počet hodnot n takové přirozené číslo, že $n+1$ je dělitelné číslem 4 pak hodnota s pořadovým číslem $\frac{n+1}{4}$ prvním kvartilem a třetím kvartilem bude hodnota s pořadovým číslem $\frac{3(n+1)}{4}$.

Vzdálenost mezi prvním a třetím kvartilem nazveme mezikvartilním rozsahem. Je-li tato hodnota malá, lze říci že polovina hodnot (dat) je blízko sebe. Lze tedy tuto hodnotu považovat za míru toho jak jsou data rozptýleny.

Hodnota, která se v daném souboru vyskytuje nejčastěji se nazývá modus.

Příklad 5. Dejme tomu, že jsme sbírali data o návštěvnících hotelu a zajímalo nás kolik nocí strávili v hotelu. Záznamy jsou souhrnně uvedeny v následující tabulce:

dny	1	2	3	4	5	6	7	8	9	10	x_i
četnost	419	609	305	204	177	156	103	105	62	35	n_i

Najděte modus, medián, první a třetí kvartil tohoto souboru dat.

Řešení. $n = \sum_{i=1}^{10} n_i = 2175$.

1. Nejčastější pozorovanou hodnotou jsou 2 dny. Tedy nejpravděpodobněji zůstane host v hotelu na dvě noci. Tudíž modus je roven 2 dnům.
2. Mediánem je hodnota s pořadovým číslem $\frac{n+1}{2} = \frac{2175+1}{2} = 1088$.
Suma $n_1+n_2 = 419+609 = 1028$. Tudíž mediánem je hodnota $x_3 = 3$ dny.
3. Určeme ještě 1. a 3. kvartil:

$\frac{n+1}{4} = 544$, $3 \times 544 = 1632$.

Hodnota s pořadovým číslem 544 leží ve druhé skupině a tedy 1. kvartil je roven 2 dnům.
Dále pokud sečteme počet prvků v prvních 4 skupinách dostaneme $1537 = 419+609+305+204$.
To znamená, že 3. kvartil je roven 5 dnům (leží v 5. skupině.)

Popišme nyní podrobněji postup jak určíme 1. kvartil (dolní kvartil) a 3. kvartil (horní kvartil):

a) určíme počet n hodnot/pozorování v souboru;
b) vypočítáme čísla $k_1 = \frac{n+1}{4}$ a $k_3 = \frac{3(n+1)}{4}$. Po vydělení dostaneme jeden ze zbytků: $0, \frac{1}{4}, \frac{1}{2}$ nebo $\frac{3}{4}$.
(i) je-li zbytek 0 použijeme přímo hodnotu s pořadím k.
(ii) Vyjde-li nám zbytek $\frac{1}{4}$, startujeme z hodnoty s pořadím $k-\frac{1}{4}\ (\in \mathbb{Z})$ a tuto hodnotu zvětšíme o 25% vzdálenosti k následující hodnotě pozorování.
(iii) Vyjde-li nám zbytek $\frac{1}{2}$, startujeme z hodnoty s pořadím $k-\frac{1}{2}\ (\in \mathbb{Z})$ a tuto hodnotu zvětšíme o 50% vzdálenosti k následující hodnotě.
(iv) Vyjde-li nám zbytek $\frac{3}{4}$, startujeme z hodnoty s pořadím $k-\frac{3}{4}\ (\in \mathbb{Z})$ a tuto hodnotu zvětšíme o 75% vzdálenosti k následující hodnotě.

Příklad 6. Najděte 1. a 3. kvartil pro následující posloupnosti pozorování:
(a) 2, 6, 10, 14, 18, 22

(b) 2, 6, 10, 12, 14, 18, 22

Řešení. (a) $n = 6$, $\frac{n+1}{4} = \frac{7}{4} = 1+\frac{3}{4}$
Tedy z pozorování číslo 1, tj. od hodnoty 2 a posuneme se o 75% vzdálenosti od čísla 6:
1. kvartil $= 2 + 0,75\cdot(6-2) = 5.$
Dále $\frac{3(n+1)}{4} = \frac{21}{4} = 5+\frac{1}{4}$. Tedy od hodnoty číslo 5, tj. od hodnoty 18 se posuneme o 25% vzdálenosti k hodnotě 22.
3. kvartil $= 18 + 0,25(22-18) = 19.$

(b) $n = 7$; $\frac{n+1}{4} = \frac{8}{4} = 2$; $\frac{3(n+1)}{4} = 6$.

Tedy 1. kvartil = ~~2~~ 6
2. kvartil = 18.

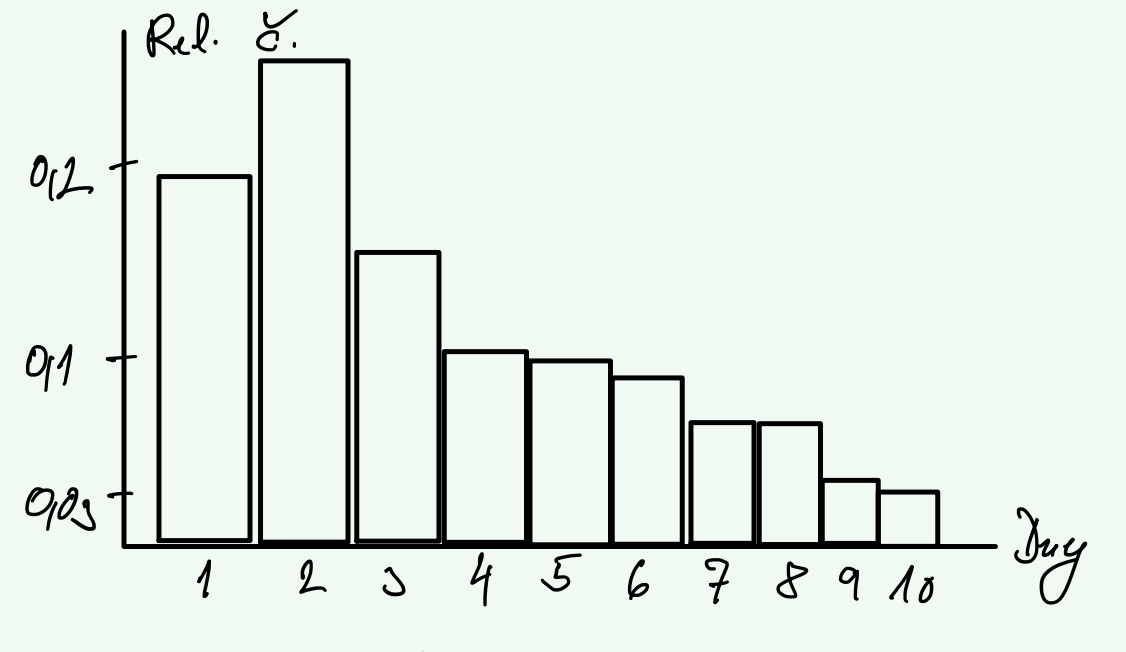
RELATIVNÍ ČETNOST A HISTOGRAMY

Využijme nyní příklad 5 k ukázce toho jak se prakticky počítají relativní četnosti jednotlivých skupin:

dny	1	2	3	4	5	6	7	8	9	10
četnost	419	609	305	204	177	156	103	105	62	35

$$\text{Relativní četnost kategorie"} = \frac{\text{Počet hodnot v kategorii}}{\text{Celkový počet hodnot}}.$$

Dny	1	2	3	4	5	6	7	8	9	10
Rel. četnost	0,19	0,28	0,14	0,09	0,08	0.07	0,05	0,05	0,03	0,02



PRŮMĚR

Nechť X označuje obecně hodnotu nějakého znaku.

Představme si, že X označuje počet nocí, které strávil v hotelu pět po sobě přicházejících hostů. Máme:

$X_1 = 2,\ X_2 = 3,\ X_3 = 2,\ X_4 = 1,\ X_5 = 2.$

Průměr je roven číslu:

$\bar{X} = \frac{1}{5}(2+3+2+1+2) = 2.$

Definice průměru: Máme-li dáno n hodnot náhodné veličiny X (= hodnota jistého statistického znaku), pak tzv. průměr $\bar{X}$ je definován vztahem:

$$\boxed{\bar{X} = \frac{1}{n}(X_1+X_2+\dots+X_n).}$$

VÝBĚROVÁ ODCHYLKA, VÝBĚROVÁ SMĚRODATNÁ ODCHYLKA

Definice výběrového rozptylu:

Uvažujme n hodnot: $X_1,\dots,X_n$ a nechť $\bar{X}$ označuje průměr tohoto souboru. Pak hodnotu:

$$S_X^2 = \frac{1}{n-1}\left((X_1-\bar{X})^2+\dots+(X_n-\bar{X})^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

nazveme výběrovým rozptylem souboru hodnot.

Výběrová směrodatná odchylka je pak dána vztahem:

$$\boxed{S_X = \sqrt{S_X^2}}$$

Příklad 7: Mějme dána (jednorozměrná) data šesti pozorování:
$X_1 = 1,\ X_2 = 8,\ X_3 = 10,\ X_4 = 4,\ X_5 = 7,\ X_6 = 12.$ Výpočty uspořádejme do tabulky:

i	X_i	$(X_i-\bar{X})$	$(X_i-\bar{X})^2$
1	1	-6	36
2	8	1	1
3	10	3	9
4	4	-3	9
5	7	0	0
6	12	5	25
Součet	42	0	80

Tedy $S_X^2 = \frac{1}{5}\sum_{i=1}^{6}(X_i-\bar{X})^2 = \frac{1}{5}\cdot 80 = 16.$
Pak
$S_X = \sqrt{16} = 4.$
$(\bar{X} = \frac{42}{6} = 7.)$

Komentář: Z tabulky je patrné, že
$\sum_{i=1}^{6}(X_i-\bar{X}) = 0.$
To platí obecně pro jakýkoli soubor dat:

$$\sum_{i=1}^{n}(X_i-\bar{X}) = \sum_{i=1}^{n}\left(X_i-\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \sum_{i=1}^{n}X_i - \sum_{i=1}^{n}X_i = 0.$$

Výběrovou odchylku S_X interpretujeme jako typickou hodnotu odchylky i-tého měření X_i znaku X od průměru $\bar{X}$.

Cvičení: Uvažujme 5členný soubor dat:
$X_1 = 5,\ X_2 = 2,\ X_3 = 4,\ X_4 = 4,\ X_5 = 10.$

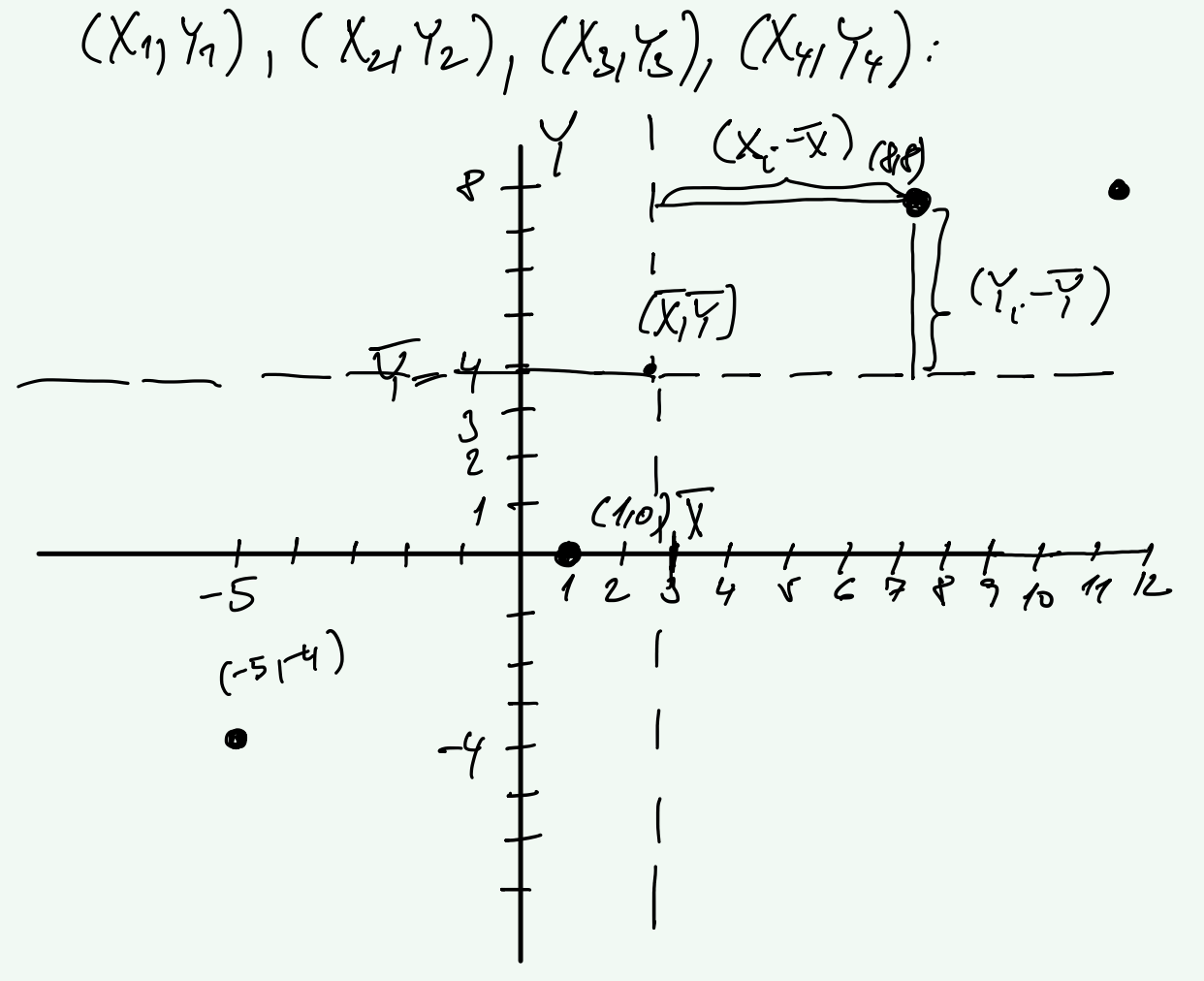
Najděte průměr $\bar{X}$, výběrovou odchylku S_X^2 a výběrovou směrodatnou odchylku S_X.

Výsledky: $\bar{X} = 5,\ S_X^2 = 9,\ S_X = 3.$

POROVNÁNÍ DVOU NÁHODNÝCH VÝBĚRŮ

Příklad 8. Mějme dány dva soubory dat:
$X_1 = 1,\ X_2 = 8,\ X_3 = -5,\ X_4 = 12,$
$Y_1 = 0,\ Y_2 = 8,\ Y_3 = -4,\ Y_4 = 8$

Nyní vynesme do roviny body:
$(X_1,Y_1), (X_2,Y_2), (X_3,Y_3), (X_4,Y_4)$:



Z obrázku je vidět, že hodnoty X_i a Y_i mají stejnou tendenci, tj. je-li X_i malé, je malá i hodnota Y_i. Podobně je-li hodnota X_i velká, je i hodnota Y_i velká.

Výběrová kovariance S_{XY} charakterizuje blízkost obou výběrů:

$$S_{XY} = \frac{1}{n-1}\left[(X_1-\bar{X})(Y_1-\bar{Y})+\dots+(X_n-\bar{X})(Y_n-\bar{Y})\right] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}).$$

Tedy máme: $\bar{X} = \frac{1}{4}(1+8-5+12) = 4$; $\bar{Y} = \frac{1}{4}(0+8-4+8) = 3.$

$$\begin{aligned} S_{XY} &= \frac{1}{3}\left[(X_1-\bar{X})(Y_1-\bar{Y})+(X_2-\bar{X})(Y_2-\bar{Y}) + (X_3-\bar{X})(Y_3-\bar{Y})+(X_4-\bar{X})(Y_4-\bar{Y})\right] = \\ &= \frac{1}{3}\left[(1-4)(0-3)+(8-4)(8-3) + (-5-4)(-4-3)+(12-4)(8-3)\right] = \\ &= \frac{1}{3}(9+20+63+40) = \frac{132}{3} \\ &= 44. \end{aligned}$$

Cvičení: Uvažujme dva soubory hodnot:
$X_1 = 242,\ X_2 = 266,\ X_3 = 218,\ X_4 = 234$ a
$Y_1 = 363,\ Y_2 = 399,\ Y_3 = 327,\ Y_4 = 351.$

Najděte S_{XY}. [Výsledek: $S_{XY} = 600$

Komentář. Jak jsme již řekli, hodnota S_{XY} vyjadřuje míru s jakou souhlasí data v obou výběrech. Platí-li, že vysoká hodnota veličiny X (což znamená, že $X > \bar{X}$) je následována s vysokou hodnotou veličiny Y (tj. $Y > \bar{Y}$), pak je výběrová kovariance kladná.
Vyjde-li nám kladná hodnota kovariance znamená to, že hodnoty X a Y mají stejnou tendenci být vyšší než je průměr a nebo být nižší než je průměr výběrového souboru hodnot.
Při záporné hodnotě S_{XY} pak mají hodnoty veličin X a Y opačnou tendenci.
Vyjde-li $S_{XY} > 0$ řekneme, že X a Y jsou kladně korelované. Vyjde-li $S_{XY} < 0$ řekneme, že X a Y jsou negativně korelované.
Pokud je S_{XY} blízké k nule řekneme, že X a Y jsou nekorelované.

Interpretovat absolutní velikost hodnoty S_{XY} je obtížné a záleží na konkrétním kontextu.

K měření (lineární) míry kovariance lze využít tzv. korelační koeficient:

$$\boxed{R_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}.}$$

Pak platí: $-1 \le R_{XY} \le 1.$

Spočítejte koeficient R_{XY} pro soubory hodnot z předchozího cvičení.

[Výsledek: $R_{XY} = 1.$]

Cvičení. Následující tabulka obsahuje jednotkové ceny jistého zboží a poptávky po tomto zboží v jednotkách zboží.

Cena	$X_1 = 71$	$X_2 = 47$	$X_3 = 23$	$X_4 = 27$
Poptávka	$Y_1 = 58$	$Y_2 = 106$	$Y_3 = 154$	$Y_4 = 146$

Najděte S_X, S_Y, S_{XY} a R_{XY}. Jaký je vztah mezi veličinami X a Y vyjadřující cenu za jednotku zboží a poptávkou uvedenou v počtu jednotek zboží.

[Výsledek: $S_X = 22,\ S_Y = 44,\ S_{XY} = -968,$
$R_{XY} = -1.$]