Statistika pro gymnázia

Pracovní verze učebního textu

ZÁKLADNÍ POJMY

Statistika zkoumá jevy (společenské, přírodní, technické) ve velkých **statistických souborech**. Prvky statistických souborů se nazývají **statistické jednotky**. Počet jednotek souboru se nazývá **rozsah souboru**.

Vlastnosti statistických jednotek se vyjadřují pomocí **statistických znaků**. Shodnými znaky je vymezena příslušnost jednotky k danému statistickému souboru, předmětem šetření jsou znaky proměnlivé.

Dělení statistických znaků:
a) **kvantitativní znaky** – jsou vyjádřeny číselnou velikostí; *spojitý* znak nabývá (v určitých mezích) jakékoliv reálné hodnoty, zatímco *nespojitý (diskrétní)* znak může nabýt jen některých (např. celočíselných) hodnot,
b) **kvalitativní znaky** – liší se kvalitou, popsány slovně; pokud může znak nabýt pouze dvou obměn, nazývá se *alternativní* znak. (Alternativní znaky lze ovšem snadno kvantifikovat přidělením logických hodnot 0 resp. 1.)

Statistické zkoumaní se rozděluje do čtyř etap:

(1) **Plán zkoumání** – úkoly a obsah akce. Stanovení statistických jednotek a znaků. Příprava rozpočtu, pracovníků, formulářů, techniky.
(2) **Šetření** – úplné, nebo výběrové (stanovení výběrového souboru).
(3) **Zpracování výsledků** – kontrola, roztřídění, shrnutí údajů. Tvorba tabulek.
(4) **Rozbor** získaných výsledků.

POPIS JEDNOROZMĚRNÝCH ROZDĚLENÍ ČETNOSTÍ

Popis četnosti kvantitativního znaku

Nejprve se budeme zabývat pouze *jediným kvantitativním* statistickým znakem. Předpokládejme, že jsme šetřením statistického znaku x u n jednotek zjistili n hodnot znaku x_i, kde $i = 1, 2, \ldots, n$. Tento znak – jak již víme – může být buďto diskrétní, nebo spojitý.

1. Diskrétní znak

Předpokládejme, že znak může nabýt právě r různých hodnot, které označíme V_1, V_2, ... , V_r. Výsledky zkoumání můžeme tedy uspořádat do tabulky, která bude mít r řádků a dva sloupce; v prvním sloupci budou podle hodnoty uspořádané varianty znaku (tj. V_j, kde $j = 1, 2, \ldots, r$), ve druhém sloupci pak číslo n_j udávající, kolikrát se ve statistickém souboru daná hodnota znaku vyskytla. Toto číslo se nazývá **absolutní četnost** hodnoty V_j. **Relativní četností** se rozumí

číslo

$$\nu_j = \frac{n_j}{n}; \tag{1}$$

relativní četnost se často násobí 100, pak je vyjádřena v procentech.

Součet četností všech možných hodnot znaku se rovná počtu všech jednotek souboru, tedy

$$\sum_{j=1}^{r} n_j = n. \tag{2}$$

Součet relativních četností všech možných hodnot znaku se rovná jedné, tedy

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

Vedle uvedené absolutní resp. relativní četnosti se zavádějí absolutní resp. relativní **kumulativní četnosti**, které informují kolik resp. jaký podíl jednotek souboru má hodnotu $x_i \leq V_j$.

Grafické znázornění: Na vodorovnou osu naneseme jednotlivé hodnoty znaku, na svislou osu četnosti. Poté vztyčíme kolmice k vodorovné ose, jejichž délka je úměrná příslušné četnosti. Spojením jejich koncových bodů vznikne **polygon četnosti** (spojnicový diagram).

2. Spojitý znak

Spojitý znak může nabývá nekonečně mnoha hodnot z jistého intervalu omezeného nejmenší a největší hodnotou. Tento interval rozdělíme na několik **dílčích intervalů** (zpravidla stejné šíře), v nichž spočítáme příslušné četnosti podobně jako v předchozím případě. Doporučuje se, aby dílčích intervalů bylo 5–20; podle Sturgesova pravidla by počet intervalů měl být zhruba dán výrazem $1 + 3,3 \log n$.

Grafické znázornění: Zpravidla se užívá **histogram**. Jde o typ sloupcového diagramu, kde sloupce (obdélníky) tvořící diagram mají šířku rovnou šířce dílčích intervalů; výška odpovídá četnosti zjištěných hodnot v daném intervalu.

Popis četnosti kvalitativního znaku

Zjištěné údaje se uspořádají do tabulky, v níž se jednotlivým variantám znaku přiřadí jejich četnosti.

Grafické znázornění: Kruhový (výsečový) diagram – v tomto diagramu různým hodnotám znaku odpovídají kruhové výseče, jejichž plošné obsahy jsou úměrné četnostem.

V další části textu se budeme zabývat již jen kvantitativními znaky.

CHARAKTERISTIKY POLOHY

Charakteristikami polohy se snažíme vystihnout úroveň, na níž se zhruba pohybují hodnoty kvantitativního znaku v daném souboru.

Aritmetický průměr, modus

Jednoduchou charakteristikou polohy je **aritmetický průměr** definovaný vztahem:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{4}$$

Výpočet aritmetického průměru z tabulky četností je rychlejší než podle vzorce (4); musíme přitom každou hodnotu V_j násobit její četností n_j:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{r} V_j n_j. \tag{5}$$

Výpočet aritmetického průměru z dílčích souborů. Předpokládejme, že se soubor skládá z dílčích souborů A, B, C, které mají počty jednotek n_A, n_B, n_C a průměry $\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$. Potom

$$\bar{x} = \frac{\bar{x}_A n_A + \bar{x}_B n_B + \bar{x}_C n_C}{n_A + n_B + n_C}. \tag{6}$$

Takto počítaný průměr se nazývá vážený průměr. Obecně je **vážený průměr** čísel u_1, u_2, … u_n s váhami $v_1 > 0$, $v_2 > 0$, … , $v_n > 0$ dán vztahem

$$\bar{u} = \frac{\sum_{j=1}^{n} u_j v_j}{\sum_{j=1}^{n} v_j}; \tag{7}$$

vztah užívají studenti dr. Voršilkové a dr. Hrnčířové k výpočtu svých známek z matematiky. □

Vedle aritmetického průměru existuje ještě další, jednodušší charakteristika polohy: modus. **Modus** $\mathrm{Mod}(x)$ znaku x je hodnota V_j tohoto znaku s největší četností.[1])

Kvantily

Kvantil $\tilde{x}_\vartheta$ (čteme: ϑ-procentní kvantil) je hodnota kvantitativního znaku x, pro kterou platí, že nejméně ϑ % statistických jednotek má hodnotu tohoto znaku menší nebo rovnou $\tilde{x}_\vartheta$, a alespoň $(100 - \vartheta)$ % jednotek nabývá hodnoty větší nebo rovné $\tilde{x}_\vartheta$. Nejčastěji užívané kvantily mají speciální názvy:

dolní kvartil	$\tilde{x}_{25}$
medián	$\mathrm{Med}(x) := \tilde{x}_{50}$
horní kvartil	$\tilde{x}_{75}$
decily	$\tilde{x}_{10}, \tilde{x}_{20}, \ldots, \tilde{x}_{90}$
percentily	$\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{99}$

Na základě výše uvedeného můžeme říci, že **medián** $\mathrm{Med}(x)$ znaku x je prostřední hodnota znaku, jsou-li hodnoty znaku uspořádány podle velikosti, neboť podle obecné definice kvantilu má nejméně 50 % statistických jednotek hodnotu znaku menší než medián a alespoň 50 % jednotek nabývá hodnoty větší nebo rovné mediánu. Přesněji: Je-li počet hodnot znaků n lichý, rovná

[1]) V některé literatuře se pro modus používá označení $\hat{x}$.

se medián prostřednímu, tj. $\frac{n+1}{2}$-tému členu posloupnosti hodnot seřazených podle velikosti; je-li počet hodnot znaků n sudý, rovná se medián aritmetickému průměru dvou hodnot „kolem středu“, tzn. průměru $\frac{n}{2}$-té hodnoty a $\frac{n+1}{2}$-té hodnoty.

Medián vhodnější charakteristika polohy než aritmetický průměr zejména v takovém souboru, kde některé hodnoty „extrémně vybočují“, a tím aritmetický průměr příliš zvyšují resp. snižují.

CHARAKTERISTIKY VARIABILITY

Charakteristika polohy je číslo, vyjadřující úroveň, kolem které jednotlivé hodnoty znaku kolísají. Velikost tohoto kolísání („jak moc se jednotlivé hodnoty liší od průměru resp. mediánu“) vyjadřují **charakteristiky variability** (měnivosti, kolísavosti, rozptýlenosti).

Charakteristiky variability související s aritmetickým průměrem

Odchylkou i-té hodnoty znaku od aritmetického průměru rozumíme rozdíl $x_i - \bar{x}$. Jak je patrné, odchylka je kladná resp. záporná podle toho, zda je i-tá hodnota znaku větší resp. menší než průměr. Sečteme-li odchylky všech hodnot daného znaku, musí – protože jsou počítány z aritmetického průměru – vyjít nula.

Rozptyl s^2 je průměr druhých mocnin odchylek, tedy

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{8}$$

Směrodatná odchylka s je odmocnina z rozptylu:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{9}$$

Poznámka pro všetečného čtenáře. Čtenář možná položí otázku, proč jsou uvedené veličiny zavedeny tak prapodivným způsobem. Pokusme se to vysvětlit: Bylo by přirozené vyjít od odchylek a spočítat jejich průměr. Jenže – jak již víme – součet odchylek je roven nule, proto i jejich průměr je roven 0. Je tedy třeba „zbavit se záporných znamének“. To je možné učinit vložením absolutní hodnoty do vzorce; absolutní hodnoty jsou však pro teoretické úvahy dosti „nepohodlné“. Byla proto dána přednost druhým mocninám odchylek, které jsou – podobně jako absolutní hodnoty – nezáporné. Tak byl zaveden pojem rozptyl. Nevýhodou rozptylu však je, že má jiný fysikální rozměr, než vyhodnocovaný statistický znak. (Zkoumáme-li např. délku L, má její rozptyl rozměr L^2, tedy rozměr obsahu.) Vše se však spraví, pokud vypočítaný rozptyl opět odmocníme. Tím jsme již došli k pojmu směrodatná odchylka. Problematikou chyb fysikálních měření se podrobně zabývá jiný autorův učební text.

Charakteristiky variability související s mediánem

Je-li poloha charakterizována mediánem, není dobré popisovat variabilitu souboru směrodatnou odchylkou, zavedenou pomocí aritmetického průměru. Raději užijeme mezikvartilovou odchylku.

Mezikvartilová odchylka je definována pomocí dolního a horního kvartilu vztahem

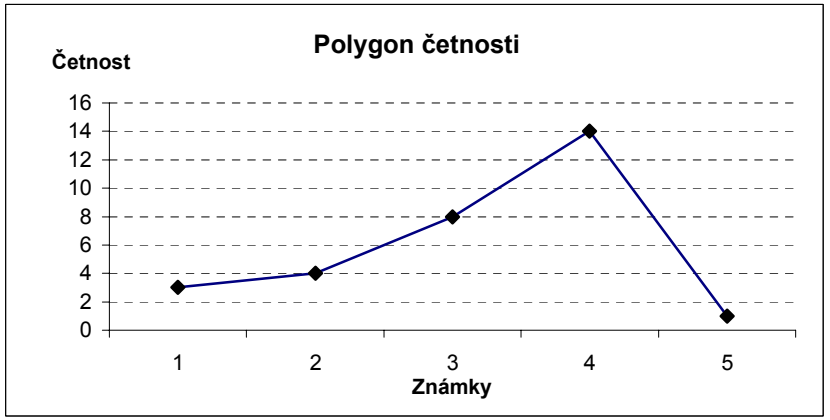
$$Q = \frac{1}{2}(\tilde{x}_{75} - \tilde{x}_{25}). \tag{10}$$

Tab. 1: Základní pojmy

Vyšší kuchařská M. D. Rettigové, Praha-Bubny Žáci třídy 4. B				
	Jméno	Body F	Známka M	Dobromila
1	Adéla Bernardová	163,5	1	0
2	Andrea Cvrkalová	136	2	0
3	Anna Čurdová	90	4	1
4	Dominika Futerová	122,5	3	0
5	Adam Hájek	85	4	0
6	Eliška Jackmannová	109	3	1
7	Jan Kalát	135	2	1
8	Eva Kneysová	137,5	2	1
9	Ivana Kristenová	95	4	0
10	Jiří Kuc	106,5	3	0
11	Lukáš Kučera	77,5	4	0
12	Jana Kurelová	84,5	4	0
13	Jana Le Ha Hai	88	4	0
14	Jana Ličíková	92,5	4	0
15	Jitka Loumová	101,5	3	1
16	Kateřina Pavlíčková	90,5	4	0
17	Markéta Petružálková	145,5	2	0
18	Martina Piršelová	90	4	0
19	Michael Preisler	134	3	1
20	Pavel Pytloun	117	3	0
21	Michaela Romová	85,5	4	0
22	Michaela Rutová	56,5	5	1
23	Monika Srpová	107,5	3	0
24	Nikol Svobodová	100	3	0
25	Veronika Šulcová	88	4	1
26	Nikola Švecová	163	1	0
27	Petr Taibr	183	1	0
28	Phuong Třešňáková	79	4	0
29	Sandra Vondráčková	76	4	0
30	Petr Voženílek	96,5	4	1

Tab. 2: Četnosti diskrétního znaku

Četnosti jednotlivých známek					
Známka	*Četnost*	*Rel. č.*	*Rel. č. %*	*KRČ*	ke (5)
1	3	0,10	10,00	10,00%	3
2	4	0,13	13,33	23,33%	8
3	8	0,27	26,67	50,00%	24
4	14	0,47	46,67	96,67%	56
5	1	0,03	3,33	100,00%	5
Celkem	**30**	**1,00**	**100**	**100,00%**	**96**
Průměr					**3,2**



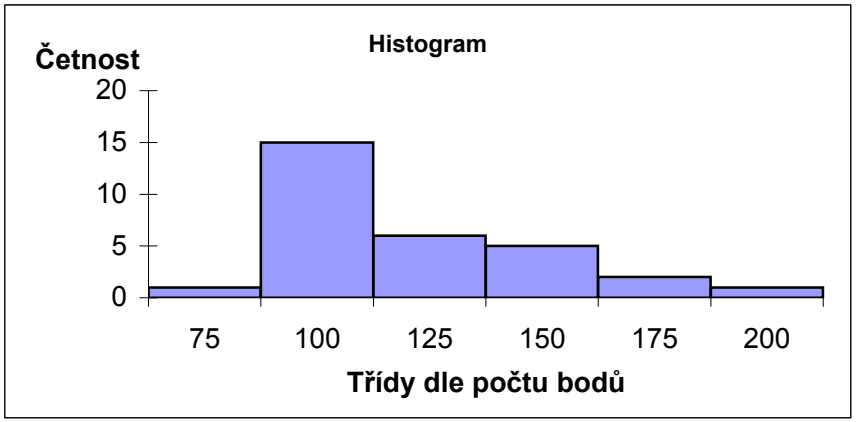
Tab. 3: Četnosti spojitého zn.

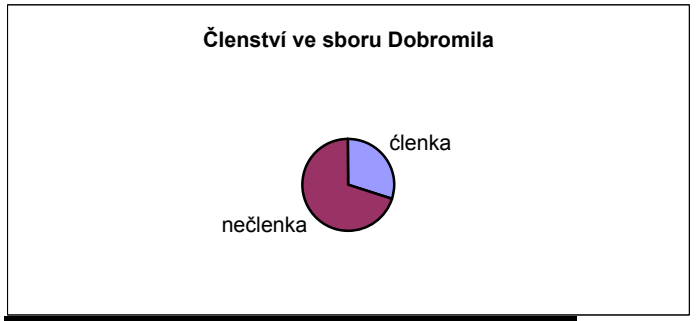
Četnosti bodů v intervalech		
Třídy	*Četnost*	*Rel. č.*
75	1	0,03
100	15	0,50
125	6	0,20
150	5	0,17
175	2	0,07
200	1	0,03
Součet	**30**	**1,00**

Dolní kvartil:	3
Medián:	3,5
Horní kvartil:	4
20. percentil	2
Čtvrtý decil:	3
Modus	4

Tab. 4: Seřazení

1	1
2	1
3	1
4	2
5	2
6	2
7	2
8	3
9	3
10	3
11	3
12	3
13	3
14	3
15	3
16	4
17	4
18	4
19	4
20	4
21	4
22	4
23	4
24	4
25	4
26	4
27	4
28	4
29	4
30	5





Vyhodnocení kvalitativního znaku

Graf k tab. 3

Tab. 6: Seřazení

1	80
2	90
3	90
4	90
5	90
6	90
7	90
8	100
9	100
10	100
11	100
12	100
13	100
14	110
15	110
16	110
17	110
18	110
19	120
20	890

Medián:	100
Dolní kvartil:	90
Horní kvartil:	110
Q	10

Tab. 5: Příklad užití mediánu

Roční příjem pracovníků JZD v tis. Kč								
Roční příjem	80	90	100	110	120	890	Součet	Průměr
Četnost	1	6	6	5	1	1	20	
k (5)	80	540	600	550	120	890	2780	139

Tab. 7: Charakteristiky variability související s průměrem

Opakované měření délky			
Číslo měření	*x*	*Odchylka*	*Čtv. odch.*
1	2,09	0,03	0,0009
2	2,01	-0,05	0,0025
3	2,11	0,05	0,0025
4	2,02	-0,04	0,0016
5	2,03	-0,03	0,0009
6	2,11	0,05	0,0025
7	2,1	0,04	0,0016
8	2,03	-0,03	0,0009
9	2,05	-0,01	0,0001
10	2,05	-0,01	0,0001
Součet:	20,6	0	0,0136
Průměr:	2,06		
Rozptyl:			0,0014
Směr. o.			0,0369