2. Bodové a intervalové rozložení četností

(Jak získat informace z datového souboru?)

Po prostudování této kapitoly budete umět:

- konstruovat diagramy znázorňující rozložení četností
- vytvářet tabulky četností
- sestrojit grafy četnostní funkce, empirické distribuční funkce, hustoty četnosti a empirické intervalové distribuční funkce

Nejprve se seznámíme s bodovým rozložením četností a ukážeme si, jak pomocí různých diagramů graficky znázornit bodové rozložení četností. Pro datový soubor známek z matematiky a angličtiny pak vytvoříme několik typů diagramů.

2.1. Definice

Nechť je dán jednorozměrný datový soubor. Jestliže počet variant znaku X není příliš velký, pak přiřazujeme četnosti jednotlivým variantám a hovoříme o *bodovém rozložení četností*.

2.2. Definice

Existuje několik způsobů, jak graficky znázornit bodové rozložení četností.
Tečkový diagram: na číselné ose vyznačíme jednotlivé varianty znaku X a nad každou variantu nakreslíme tolik teček, jaká je její absolutní četnost.
Polygon četnosti: je lomená čára spojující body, jejichž x-ová souřadnice je varianta znaku X a y-ová souřadnice je absolutní četnost této varianty.
Sloupkový diagram: je soustava na sebe nenavazujících obdélníků, kde střed základny je varianta znaku X a výška je absolutní četnost této varianty.
Výsečový graf: je kruh rozdělený na výseče, jejichž vnější obvod odpovídá absolutním četnostem variant znaku X.
Dvourozměrný tečkový diagram: na vodorovnou osu vyneseme varianty znaku X, na svislou varianty znaku Y a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dané dvojice.

2.3. Příklad

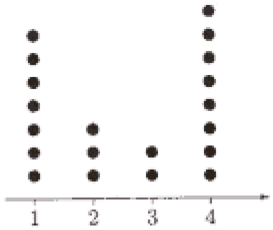
Pro datový soubor z příkladu 1.5 sestrojte

a) jednorozměrné tečkové diagramy pro znak X a znak Y
b) polygony četností pro znak X a znak Y
c) sloupkové diagramy pro znak X a znak Y
d) výsečové diagramy pro znak X a znak Y

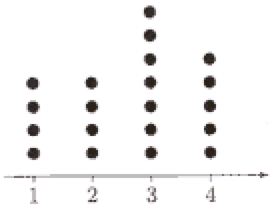
Řešení:

ad a)

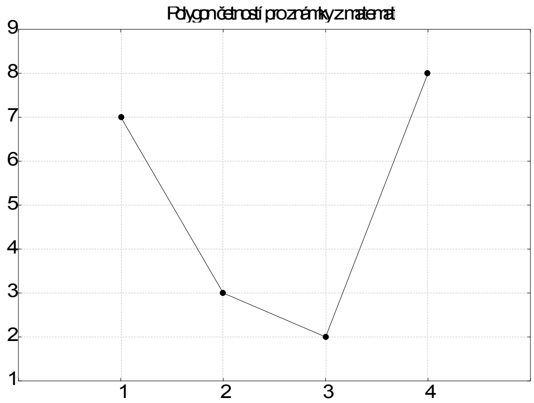
Známka z M

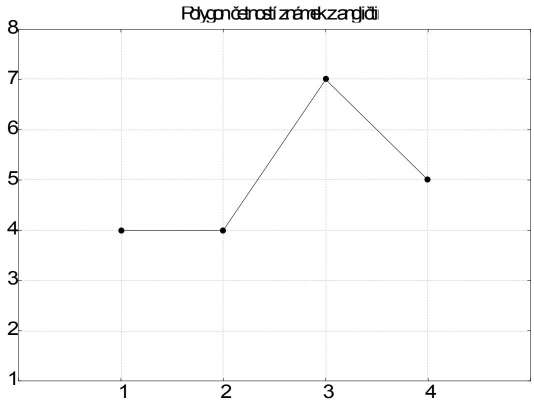


Známka z A

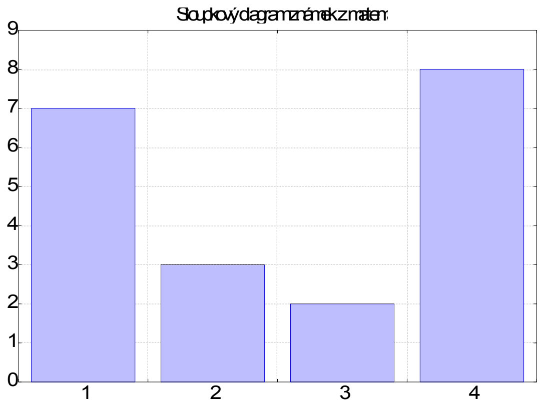


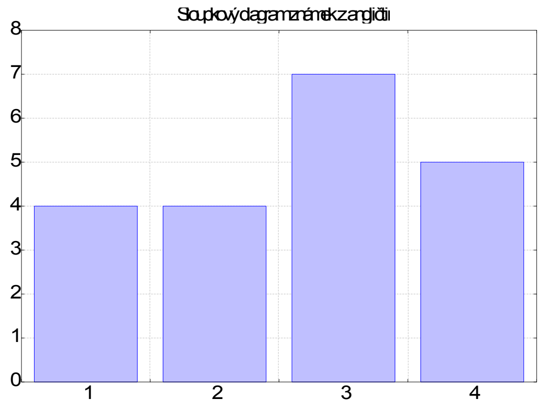
ad b)



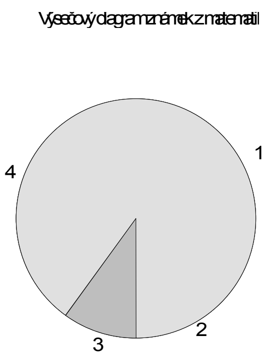


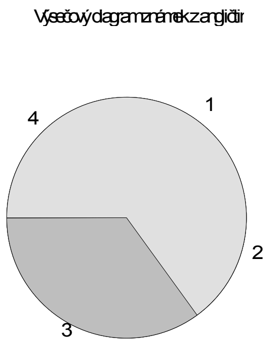
ad c)





ad d)





Ze všech těchto diagramů je vidět odlišný přístup zkoušejících ke studentům. Matematik nešetří jedničkami, ale místo trojky raději rovnou dává čtyřku. Naproti tomu angličtinář považuje trojku za typickou studentskou známu.

2.4. Definice

Nechť je dán jednorozměrný datový soubor, v němž znak X nabývá r variant. Pro j = 1, ..., r definujeme:

$n_j = N(X = x_{[j]})$ – *absolutní četnost varianty $x_{[j]}$ ve výběrovém souboru*

$p_j = \frac{n_j}{n}$ – *relativní četnost varianty $x_{[j]}$ ve výběrovém souboru*

$N_j = N(X \le x_{[j]}) = n_1 + ... + n_j$ – *absolutní kumulativní četnost prvních j variant ve výběrovém souboru*

$F_j = \frac{N_j}{n} = p_1 + ... + p_j$ – *relativní kumulativní četnost prvních j variant ve výběrovém souboru*

Tabulka typu

$x_{[j]}$	n_j	p_j	N_j	F_j
$x_{[1]}$	n_1	p_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{[r]}$	n_r	p_r	N_r	F_r

se nazývá *variační řada*.

Funkce $p(x) = \begin{cases} p_j \text{ pro } x = x_{[j]}, j = 1, \ldots, r \\ 0 \text{ jinak} \end{cases}$ se nazývá *četnostní funkce*.

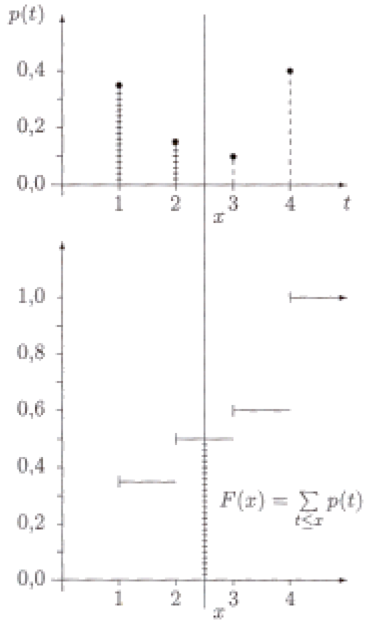
Funkce $F(x) = \begin{cases} 0 \text{ pro } x < x_{[1]} \\ F_j \text{ pro } x_{[j]} \le x < x_{[j+1]}, j = 1, \ldots, r-1 \\ 1 \text{ pro } x \ge x_{[r]} \end{cases}$ se nazývá *empirická distribuční funkce*.

2.5. Příklad

Pro datový soubor z příkladu 1.5 sestavte variační řadu pro znak X. Nakreslete grafy četnostní funkce a empirické distribuční funkce.

Řešení:

$x_{[j]}$	n_j	p_j	N_j	F_j
1	7	0,35	7	0,35
2	3	0,15	10	0,50
3	2	0,10	12	0,60
4	8	0,40	20	1,00
-	20	1,00	-	-



V některých datových souborech je počet variant znaku příliš veliký a použití bodového rozložení četností by vedlo k nepřehledným a roztříštěným výsledkům. V takových situacích používáme intervalové rozložení četností. Definujeme třídicí interval a jeho absolutní a relativní četnost, absolutní a relativní kumulativní četnost. Nově zavádíme četnostní hustotu třídicího intervalu. Uvedené četnosti zapisujeme do tabulky rozložení četností. Počet třídicích intervalů stanovujeme např. podle Sturgesova pravidla. Intervalové rozložení četností použijeme v příkladu s datovým souborem obsahujícím údaje o mezích plasticity a pevnosti 60 vzorků oceli.

2.6. Definice

Nechť je dán jednorozměrný datový soubor. Jestliže počet variant znaku X je blízký rozsahu souboru, pak přiřazujeme nikoliv jednotlivým variantám, ale celým intervalům hodnot. Hovoříme pak o *intervalovém rozložení četnosti*.

2.7. Definice

Číselnou osu rozložíme na intervaly typu $(-\infty, u_1)$, $\langle u_1, u_2)$, ..., $\langle u_r, u_{r+1})$, $\langle u_{r+1}, \infty)$ tak, aby okrajové intervaly neobsahovaly žádnou pozorovanou hodnotu znaku X. Užíváme označení:

$\langle u_j, u_{j+1})$ – *j-tý třídicí interval znaku X*, j = 1, ..., r.

$d_j = u_{j+1} - u_j$ – *délka j-tého třídicího intervalu znaku X*

$x_{[j]} = \frac{u_j + u_{j+1}}{2}$ – *střed j-tého třídicího intervalu znaku X*

Třídicí intervaly volíme nejčastěji stejně dlouhé. Jejich počet určíme např. pomocí Sturgesova pravidla: $r = 1 + 3{,}3 \times \log_{10} b$, kde b je počet variant znaku X.

2.8. Definice

Nechť je dán jednorozměrný datový soubor rozsahu n. Hodnoty znaku X roztřídíme do r třídicích intervalů. Pro j = 1, ..., r definujeme:

$n_j = N(u_j < X \le u_{j+1})$ – *absolutní četnost j-tého třídicího intervalu ve výběrovém souboru*

$p_j = \dfrac{n_j}{n}$ – *relativní četnost j-tého třídicího intervalu ve výběrovém souboru*

$f_j = \dfrac{p_j}{d_j}$ – *četnostní hustota j-tého třídicího intervalu ve výběrovém souboru*

$N_j = N(X \le u_{j+1}) = n_1 + \dots + n_j$ – *absolutní kumulativní četnost prvních j třídicích intervalů ve výběrovém souboru*

$F_j = \dfrac{N_j}{n} = p_1 + \dots + p_j$ – *relativní kumulativní četnost prvních j třídicích intervalů ve výběrovém souboru*.

Tabulka typu

$(u_j, u_{j+1}\rangle$	d_j	n_j	p_j	f_j	N_j	F_j
$(u_1, u_2\rangle$	d_1	n_1	p_1	f_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(u_r, u_{r+1}\rangle$	d_r	n_r	p_r	f_r	N_r	F_r
Součet		n	1			

se nazývá *tabulka rozložení četností*.

2.9. Příklad

Z fiktivního základního souboru všech vzorků oceli odpovídajících „všem myslitelným tavbám“ bylo do laboratoře dodáno 60 vzorků a zjištěny a hodnoty znaku X – mez plasticity a Y – mez pevnosti. Datový soubor má tvar:

154	178	83	98	73	76
133	164	106	111	77	85
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

a) Pro znak X stanovte optimální počet třídicích intervalů dle Sturgesova pravidla.
b) Sestavte tabulku rozložení četností.

Řešení:

ad a) Znak X má 50 variant, tedy podle Sturgesova pravidla je optimální počet třídicích intervalů r = 7. Budeme tedy volit 7 intervalů stejné délky tak, aby v nich byly obsaženy všechny pozorované hodnoty znaku X, z nichž nejmenší je 33, největší 160; volba u_1 = 30, ..., u_8 = 170 splňuje požadavky.

ad b)

(u_j, u_{j+1})	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$(30, 50)$	20	40	8	0,13333	8	0,13333	0,00666
$(50, 70)$	20	60	4	0,06667	12	0,20000	0,00333
$(70, 90)$	20	80	13	0,21667	25	0,41667	0,01083
$(90, 110)$	20	100	15	0,25000	40	0,66667	0,01250
$(110, 130)$	20	100	9	0,15000	49	0,81667	0,00750
$(130, 150)$	20	140	7	0,11667	56	0,93333	0,00583
$(150, 170)$	20	160	4	0,06667	60	1,00000	0,00333
Součty			$n = 60$	1,00000			

Ke grafickému znázornění intervalového rozložení četností slouží histogram. S jeho pomocí lze dobře vysvětlit, co znamená hustota četnosti, což je funkce zavedená pomocí četnostních hustot jednotlivých třídicích intervalů. S hustotou četnosti úzce souvisí intervalová empirická distribuční funkce (je všude spojitá, protože je funkcí horní meze integrálu z hustoty četnosti). Pro údaje o mezi plasticity oceli vytvoříme histogram a graf intervalové empirické distribuční funkce. Seznámíme se rovněž s vlastnostmi obou výše zmíněných funkcí.

2.14. Definice

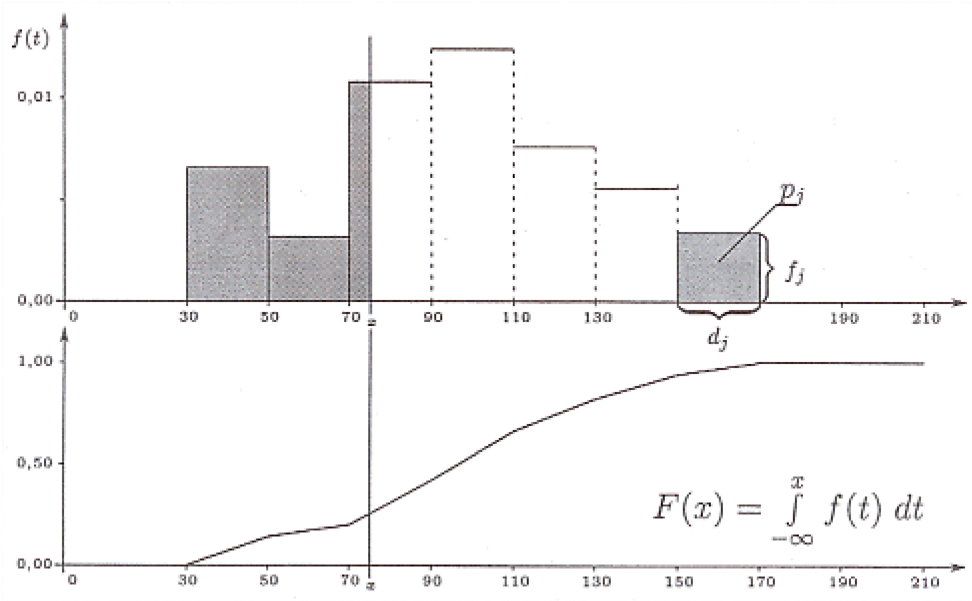
Intervalové rozložení četností graficky znázorňujeme pomocí *histogramu*. Je to graf skládající se z r obdélníků, sestrojených nad třídicími intervaly, přičemž obsah j-tého obdélníku je roven relativní četnosti p_j j-tého třídicího intervalu, j = 1, ..., r. Histogram je shora omezen schodovitou čarou, která je grafem funkce zvané *hustota četnosti*: f(x) = $\begin{cases} f_j \text{ pro } u_j < x \leq u_{j+1}, j = 1, \ldots, r \\ 0 \text{ jinak} \end{cases}$

Pomocí hustoty četnosti zavedeme *intervalovou empirickou distribuční funkci*: F(x) = $\int_{-\infty}^{x} f(t)$.

2.15. Příklad

Pro datový soubor z příkladu 2.12 nakreslete histogram pro znak X a pod histogram nakreslete graf intervalové empirické distribuční funkce.

Řešení:



Shrnutí

Není-li v jednorozměrném datovém souboru počet variant znaku příliš velký, pak přiřazujeme četnosti jednotlivým variantám znaku a hovoříme o **bodovém rozložení četností**. To lze znázornit graficky pomocí různých **diagramů** (např. tečkový diagram, sloupkový diagram atd.). Pokud zapíšeme četnosti do tabulky, dostaneme **variační řadu**. Pomocí relativních četností zavedeme **četnostní funkci**, pomocí kumulativních relativních četností **empirickou distribuční funkci**, která má schodovitý průběh.

Pracujeme-li s dvourozměrným datovým souborem, zavádíme **simultánní četnosti** a zapisujeme je do **kontingenční tabulky**. Na okrajích kontingenční tabulky jsou uvedeny **marginální četnosti**, které se vztahují jen k jednomu znaku. Pomocí simultánních kumulativních relativních četností zavádíme simultánní četnostní funkci. Simultánní a marginální četnosti či četnostní funkce nám snadno umožní ověřit **četnostní nezávislost** dvou znaků v daném výběrovém souboru.

Je-li se počet variant znaku srovnatelný s rozsahem souboru, použijeme raději **intervalové rozložení četností**, při němž přiřazujeme četnosti nikoli jednotlivým variantám, ale třídicím intervalům. Jejich počet určíme např. pomocí **Sturgesova pravidla**. Četnosti třídicích intervalů zapisujeme do **tabulky rozložení četností**. Relativní četnosti třídicích intervalů znázorňujeme pomocí **histogramu**. Schodovitá čára shora omezující histogram je grafem **hustoty četnosti**. Spojitým protějškem schodovité empirické distribuční funkce je **intervalová empirická distribuční funkce** zavedená jako funkce horní meze integrálu z hustoty četnosti.

Kontrolní otázky a úkoly

1. Jaké grafy znázorňující rozložení četností znáte? Popište způsob jejich konstrukce.
2. Jak vzniká variační řada?
3. Jaké četnosti zapisujeme do kontingenční tabulky?
4. Kdy jsou v daném výběrovém souboru znaky četnostně nezávislé?
5. K čemu slouží Sturgesovo pravidlo?
6. (S) U 50 náhodně vybraných posluchačů a posluchaček VŠE v Praze byla zjišťována jejich hmotnost v kg (znak X) a jejich výška v cm (znak Y).

58	178	60	168	56	172
68	173	68	173	52	165
56	170	63	171	72	185
60	170	72	177	75	170
61	173	90	192	52	163
71	181	57	176	63	184
85	184	51	168	63	172
65	170	81	190	58	162
80	170	73	177	64	174
52	172	75	179	52	168
72	182	71	180	55	164
57	169	66	178	67	173
65	169	67	182	60	178
60	170	72	191	55	160
54	162	57	174	62	172
52	169	57	160	70	171
83	182	56	170		

a) Pro znak X stanovte optimální počet třídicích intervalů podle Sturgesova pravidla, sestavte tabulku rozložení četnosti, nakreslete histogram a graf intervalové empirické distribuční funkce.

b) Pro znak Y rovněž stanovte optimální počet třídicích intervalů podle Sturgesova pravidla. Pro vektorový znak (X, Y) sestavte kontingenční tabulku absolutních četností a nakreslete dvourozměrný tečkový diagram.