Kapitola 1

Základy popisné statistiky

Všude kolem nás se setkáváme se shromažďováním velkého počtu údajů o nejrůznějších objektech. Mohou to být národohospodářské údaje o vývoji ekonomiky dané země sbírané v pravidelných časových intervalech, údaje o klientech dané banky, údaje o příjmech a výdajích pojišťovny, údaje o zdravotním stavu pacientů ošetřených ve vybrané nemocnici, personální údaje o studentech a údaje o jejich prospěchu na určité univerzitě v daném roce, údaje o výrobcích daného podniku a podobně. Souhrnně, tyto údaje vytvářejí rozsáhlé datové soubory, které obsahují velké množství informace. Informace obsažená v takových rozsáhlých datových souborech se může lidskému pozorovateli jevit jako nepřehledná a pro její utřídění se zavádí speciální aparát – **popisná statistika**. Cílem popisné statistiky je informaci z datových souborů zhuštěně a přehledně popsat tak, aby byla snadněji vnímatelná. K přehlednému popisu rozsáhlých datových souborů se v popisné statistice často používají různé typy tabulek, grafů, diagramů a různé funkcionální charakteristiky jednoduše stanovené pomocí elementárních matematických prostředků. Ukázat základní metody popisné statistiky je cílem této kapitoly.

1.1 Základní pojmy

Datové soubory obvykle pořizujeme pozorováním, měřením, nebo jiným zjišťováním hodnoty sledovaného ukazatele či proměnné na množině k tomu účelu vybraných prvků. Tyto prvky nazýváme **statistické jednotky** a jejich množinu, která je předmětem prováděného sledování nazýváme **statistický soubor**. Statistický soubor je obvykle dobře vymezen z hlediska věcného, prostorového a časového. Např. při průzkumu názorů studentů na placení školného může být statistický soubor tvořen všemi studenty studujícími první ročník Masarykovy univerzity v Brně v roce 2005. Vymezení statistického souboru musí být jednoznačné a neměly by vznikat pochybnosti, zda daný prvek do statistického souboru patří či nikoliv. V uvedeném příkladě

by tedy mělo být řečeno, zda se soubor vztahuje také na studenty distančního studia nebo jenom na studenta řádného studia apod. Zjišťovaný ukazatel, který na jednotlivých prvcích statistického souboru pozorujeme nebo měříme, se nazývá **statistický znak**. Statistické znaky budeme označovat velkými písmeny z konce abecedy, např. X může být na výše zmíněném souboru ukazatel „názor na placení školného", Y může být „finanční situace jeho rodiny" zakódovaná následujícím způsobem: 1 – výborná, 2 – dobrá, 3 – uspokojivá, 4 – neuspokojivá; Z může být „měsíční výše kapesného v 1000 Kč" apod.

Na příkladu předchozích tří znaků X, Y a Z je dobře patrné, že se tyto znaky od sebe svým charakterem mohou značně lišit. Zatímco znak X může nabývat pouze 3 hodnot z množiny $\{1, 2, 3\}$ a znak Y může nabývat pouze čtyř hodnot z množiny $\{1, 2, 3, 4\}$, může znak Z nabývat při dostatečně přesném sledování znaku, kterékoliv hodnoty v intervalu $(0, \infty)$. (Příliš vysokých hodnot nabývá prakticky s nulovou pravděpodobností.) Možné hodnoty znaku se nazývají **varianty**, nebo též **obměny** znaku a tvoří množinu, kterou označíme V. Je-li množina V konečná, nebo spočetná (tj. její prvky lze uspořádat do posloupnosti), mluvíme o **diskrétním znaku**. Je-li množina variant diskrétního znaku X konečná, označíme je $V_x = \{x_{[1]}, x_{[2]}, \ldots, x_{[r]}\}$. Číslo r je počet možných variant diskrétního znaku X. V opačném případě, když je množina V tvořena intervalem, mluvíme o **spojitém znaku**. V uvedeném příkladě jsou tedy znaky X a Y diskrétní a znak Z je spojitý.

Jiné dělení znaků dostaneme podle stupně jejich kvantifikace. Vyjdeme ze statistického souboru, který obsahuje n statistických jednotek. Číslo n budeme nazývat **rozsahem statistického souboru** a hodnoty znaku X zjištěné na jednotlivých statistických jednotkách označíme $x_1, x_2, \ldots, x_n$. Potom podle obsahové kvantifikace hodnot znaku rozdělujeme znaky na:

a) **nominální**, které připouštějí mezi hodnotami $x_1, x_2, \ldots, x_n$ pouze relaci rovnosti, např. $x_1 = x_2, x_2 \neq x_3$ apod. Jednotlivé hodnoty znaků představují pouze číselné kódy kvalitativních pojmenování. Např. znak X – názor na placení školného je nominální znak. Jiným příkladem může být očíslování městských tramvají, zakódování profese zaměstnance apod. Nominální znak, který může nabývat pouze dvou hodnot se nazývá **alternativní** v opačném případě **množinný**.

b) **ordinální**, které připouštějí kromě relace rovnosti také obsahovou interpretaci relace uspořádání $x_1 < x_2$ (nebo $x_1 > x_2$). Uspořádání vyjadřuje větší nebo menší intenzitu popisované vlastnosti. Např. znak Y je ordinální, pro hodnoty znaku $y_1 = 1, y_2 = 3$ a $y_3 = 1$ platí, že první a třetí student uvažovaného statistického souboru mají stejně ohodnocenou finanční situaci rodiny, ale finanční situace rodiny prvního studenta je lepší než finanční situace rodiny druhého studenta ($y_1 = y_3$, ale $y_1 < y_2$).

c) **kardinální** znaky připouštějí obsahovou interpretaci nejen relací rovnosti a

uspořádání ale také operací součtu $x_1 + x_2$ a rozdílu $x_1 - x_2$. To znamená, že v případě kdy $x_1 - x_2 = x_2 - x_3$, je interval (x_2, x_1) stejně dlouhý jako interval (x_3, x_2) a tato stejná délka obou intervalů představuje u obou dvojic x_1, x_2 a x_2, x_3 také stejný rozdíl v extenzitě zkoumané vlastnosti. Např. znak Z – měsíční výše kapesného studenta je kardinální znak, je-li $z_1 = 1.8$, $z_2 = 2$ a $z_3 = 2.2$, je stejný rozdíl mezi kapesným studentů 2 a 1 jako mezi kapesným studentů 3 a 2.

Má-li u kardinálního znaku smysluplnou obsahovou interpretaci také operace podílu, tj. x_1/x_2, pak se kardinální znak nazývá **poměrový**. V případě, že operace podílu nemá smysluplnou obsahovou interpretaci, nazývá se tento kardinální znak **intervalový**. Příkladem poměrového znaku je znak Z – měsíční výše kapesného studenta, kdy pro $z_1 = 3.2$ a $z_2 = 6.4$ lze smysluplně prohlásit, že druhý student dostává 2x vyšší kapesné něž první. Příkladem intervalového znaku může být např. teplota měřená ve stupních Celsia, kde nula na dané stupnici vznikla pouhou konvencí. Lze proto u teploty naměřené ve třech dnech ve stupních Celsia $t_1 = 2, t_2 = 4, t_3 = 6$ říci, že z prvního na druhý den teplota vzrostla o 2 stupně Celsia a že rovněž ze druhého na třetí den teplota vzrostla o 2 stupně Celsia. Chybná interpretace těchto údajů by byla, kdybychom řekli, že teplota z prvního na druhý den vzrostla dvakrát, kdežto ze druhého na třetí den pouze jedenapůlkrát.

1.2 Rozdělení četností statistického znaku

Budeme uvažovat statistický znak X, který na daném statistickém souboru nabyl hodnot $x_1, x_2, \ldots, x_n$. Předpokládejme, že množina jeho variant je konečná, tedy $V_X = \{x_{[1]}, x_{[2]}, \ldots, x_{[r]}\}$. Pak zavedeme následující pojmy:

n_j ... absolutní četnost varianty $x_{[j]}$ v daném souboru

$p_j = n_j/n$ ... relativní četnost varianty $x_{[j]}$ v daném souboru

Je-li znak X ordinální nebo kardinální a varianty $x_{[j]}$ lze uspořádat, tj. když $x_{[1]} < x_{[2]} < \ldots < x_{[r]}$ můžeme zavést kumulativní četnosti

$N_j = \sum_{i=1}^{j} n_i$ ... absolutní kumulativní četnost do varianty $x_{[j]}$ v daném souboru

$P_j = \sum_{i=1}^{j} p_i$ ... relativní kumulativní četnost varianty $x_{[j]}$ v daném souboru

Uvedené četnosti lze uspořádat do tabulky, která má 3 nebo 5 sloupců podle typu znaku.

Tabulka Tab. 1.1 zhuštěně popisuje na daném statistickém souboru naměřený statistický znak X a nazývá se **tabulka rozdělení četností** znaku X. Pro ještě lepší představu o naměřeném znaku X se data z uvedené tabulky znázorňují graficky.

Varianta	Absolutní četnost	Relativní četnost	Absolutní kumulativní četnost	Relativní kumulativní četnost
$x_{[1]}$	n_1	p_1	N_1	P_1
$x_{[2]}$	n_2	p_2	N_2	P_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{[r]}$	n_r	p_r	N_r	P_r
Součet	n	1	–	–

Tabulka 1.1: Tabulka rozdělení četností znaku X

Podle způsobu grafického znázornění tabulky rozdělení četnosti můžeme mluvit o sloupcovém diagramu absolutních (relativních) četností, polygonu absolutní (relativních) četností, kruhovém diagramu absolutních (relativních) četností. Příslušná grafická znázornění jsou na Obr. 1.1. Podobně lze pro kardinální nebo ordinální znak získat sloupcový diagram nebo polygon kumulativních četností (absolutních nebo relativních).

Příklad 1.1 Na náhodně vybraném souboru studentů rozsahu $n = 100$ byly zjišťovány statistické znaky X– názor na placení školného, Y – finanční situace rodiny a Z – měsíční výše kapesného, které byly detailněji popsány v odstavci 1.1. Výsledkem je tabulka hodnot Tab. 1.2. V uvedené tabulce je ve sloupci PČ uvedeno pořadové číslo vybraného studenta. Dále v tabulce Tab. 1.3 je uvedeno rozdělení četnosti znaku X a v tabulce Tab. 1.4 je uvedeno rozdělení četností znaku Y.

Grafické znázornění znaku X pomocí kruhového diagramu je na obrázku Obr. 1.2. Na Obr. 1.3 jsou uvedena vybraná grafická znázornění znaku Y.

Z grafických znázornění rozdělení četností je dobře patrné, že je přehledné a lze jej s výhodou užít v případě, kdy uvažovaný znak může nabývat menšího počtu variant. V případě, kdy diskrétní znak může nabývat velkého počtu variant nebo pro spojitý statistický znak se častěji místo rozdělení četností používá tzv. skupinové rozdělení, které uvažovaný znak lépe popisuje. Bude o něm pojednáno v následujícím odstavci.

1.3 Skupinové rozdělení četností

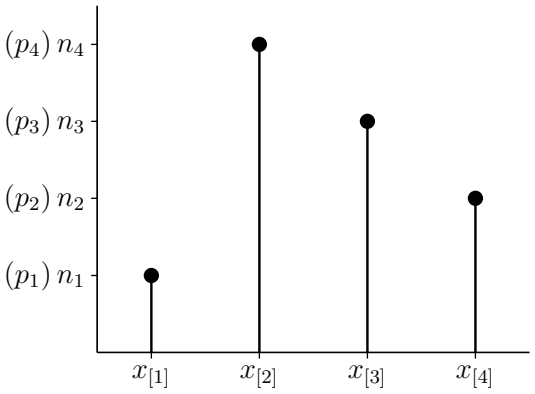
Rozdělení četností diskrétního statistického znaku, který může nabývat velkého počtu variant nebo spojitého statistického znaku již není výhodné znázorňovat pomocí rozdělení četností, protože absolutní četnosti bývají velmi nízké, často rovny 1 a počet variant r může být blízký rozsahu souboru n. V tomto případě se možné hodnoty znaku rozdělí do intervalů (někdy se říká do tříd nebo do třídních intervalů) a do tabulky rozdělení četnosti se vypisují četnosti příslušné těmto intervalům.

PČ	X	Y	Z	PČ	X	Y	Z	PČ	X	Y	Z	PČ	X	Y	Z
1	3	2	7	26	2	3	2.4	51	3	4	1	76	2	2	6.8
2	3	3	3	27	2	2	6.8	52	2	3	1.9	77	2	3	2
3	3	2	6.8	28	2	3	2.5	53	2	3	2.6	78	1	2	6.9
4	2	3	2.9	29	2	2	7.3	54	2	3	2.7	79	2	3	2.4
5	1	1	8.9	30	2	3	3.2	55	2	3	2.3	80	1	2	6.6
6	2	3	3.8	31	1	3	3	56	3	2	7.6	81	2	2	5.2
7	3	2	4.2	32	2	4	0.5	57	2	3	2	82	2	2	8.4
8	2	3	4.1	33	3	2	7.4	58	2	3	3.4	83	3	3	3.1
9	1	3	2	34	2	3	3.9	59	1	2	7.5	84	2	2	7.1
10	2	3	2.5	35	2	3	2.4	60	2	3	3.5	85	2	2	7
11	1	2	7.6	36	1	1	7.4	61	2	3	3	86	2	3	3
12	2	4	1.2	37	2	3	1.8	62	1	1	11.2	87	3	4	1.2
13	2	3	2.6	38	1	1	10.1	63	2	2	7.3	88	3	3	3
14	3	1	9.1	39	1	2	7.5	64	2	2	7.2	89	2	2	6.9
15	3	3	1.9	40	3	2	8	65	2	2	7.1	90	2	2	6.8
16	2	2	7.3	41	2	3	2.3	66	2	3	3.3	91	2	3	3.1
17	2	3	0.8	42	2	2	4.2	67	2	3	3.2	92	1	2	7.2
18	2	3	1.9	43	2	3	2.1	68	1	1	8.4	93	3	2	7.3
19	2	2	5.9	44	2	1	5.2	69	2	3	3.1	94	3	3	3.2
20	3	3	3.2	45	2	3	3.3	70	2	3	2.8	95	2	3	2.9
21	2	2	6.4	46	2	3	3.4	71	2	3	2.9	96	1	1	8.9
22	2	3	2.9	47	2	2	3.9	72	3	2	7	97	1	1	9
23	2	2	6.5	48	2	3	3.5	73	2	3	1.9	98	3	3	3
24	2	2	6.6	49	3	1	9.6	74	2	3	2.8	99	2	2	7.1
25	2	4	0.9	50	2	3	4.1	75	1	3	2.5	100	2	2	7

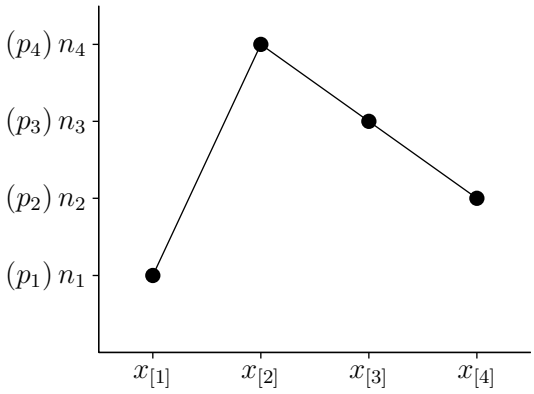
Tabulka 1.2: Datový soubor rozsahu $n = 100$ se třemi zjišťovanými znaky X, Y, Z. Ve sloupci PČ je uvedeno pořadové číslo statistické jednotky (studenta)

Slovní varianta znaku X	Zakódované varianty $x_{[i]}$	Absolutní četnosti n_i	Relativní četnosti p_i
ANO	1	15	0.15
NE	2	65	0.65
NEVÍM	3	20	0.20
Součet		100	1

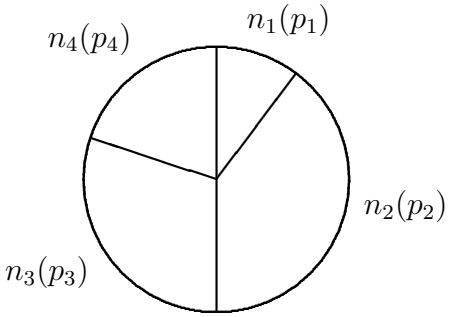
Tabulka 1.3: Rozdělení četností znaku X



Obr 1.1a) Sloupcový diagram absolutních (relativních) četností pro $r = 4$



Obr 1.1b) Polygon rozdělení absolutních (relativních) četností pro $r = 4$



Obr 1.1c) Kruhový diagram absolutních (relativních) četností pro $r = 4$

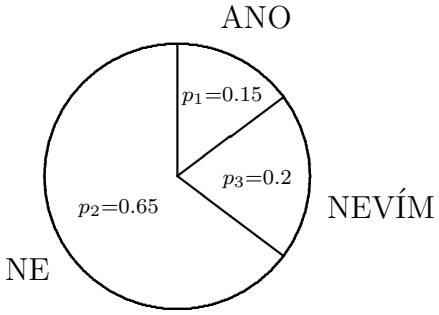
Obrázek 1.1: Grafická znázornění rabulky rozdělení četnosti

Mluvíme pak o **skupinovém rozdělením četností**.

Předpokládejme, že množina variant kardinálního znaku X (obor hodnot znaku X) je interval (a, b), $-\infty < a < b < \infty$. Tento interval můžeme zapsat jako sjednocení k podintervalů $I_1 = (a_0, a_1\rangle, I_2 = (a_1, a_2\rangle, \ldots, I_k = (a_{k-1}, a_k), a_0 = a, a_k = b$, které se nepřekrývají a jejichž sjednocením je interval (a, b). Tedy $I_i \cap I_j = \varnothing, i \neq j, \bigcup_{j=1}^{k} I_j = (a, b), a = a_0 < a_1 < a_2 < \ldots < a_k = b$. Dále označíme $d_i = a_i - a_{i-1}$ délku intervalu I_i a $s_i = \frac{1}{2}(a_i - a_{i-1})$ střed intervalu $I_i, i = 1, 2, \ldots, k$.

Nabývá-li znak X na daném statistickém souboru hodnoty $x_1, x_2, \ldots, x_n$, můžeme stanovit četnosti jednotlivých intervalů a vynést je do tabulky rozdělení četností, kde v prvním sloupci budou místo variant znaku třídní intervaly. Dostaneme tak **tabulku skupinového rozdělení četnosti** znaku X. Označení četnosti ponecháme stejné jako v předchozím odstavci. Tedy značíme

n_i ... absolutní četnost i-tého intervalu (tj. počet těch hodnot z $x_1, \ldots, x_n$, které



Obrázek 1.2: Kruhový diagram relativních četností znaku X

Slovní varianta znaku Y	Zakódované varianty znaku Y $y_{[i]}$	Absolutní četnosti n_i	Relativní četnosti p_i	Kumulativní absolutní četnosti N_i	Kumulativní relativní četnosti P_i
Výborné	1	10	0.10	10	0.10
Dobré	2	35	0.35	45	0.45
Uspokojivé	3	50	0.50	95	0.95
Neuspokojivé	4	5	0.05	100	1.00
Součet		100	1.00	–	–

Tabulka 1.4: Tabulka rozdělení četností znaku Y.

padnou do intervalu $I_i, i = 1, \ldots, k$)

$p_i = n_i/n$ … relativní četnost i-tého intervalu

$N_i = \sum_{j=1}^{i} n_j$ … absolutní kumulativní četnost intervalu I_i

$P_i = \sum_{j=1}^{i} p_j$ … relativní kumulativní četnost intervalu I_i

Kromě toho zavádíme ještě tzv. **četnostní hustotu**. Označíme

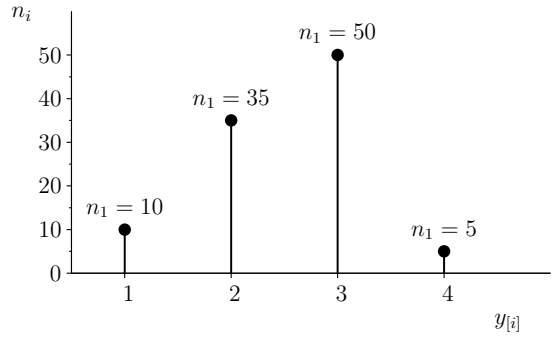
$f_i = p_i/d_i$ … četnostní hustota i-tého intervalu I_i

a funkci

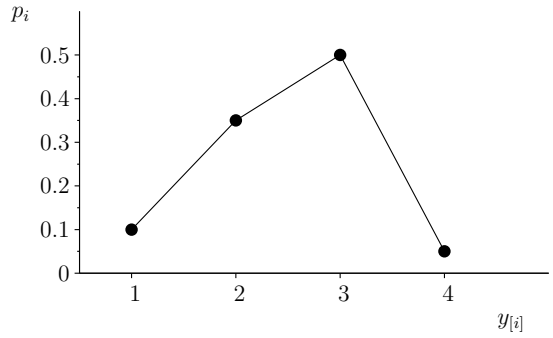
$$f^*(x) = \begin{cases} f_i & \text{pro } a_{i-1} < x \leq a_i, i = 1, 2, \ldots, k-1, \\ f_k & \text{pro } a_{k-1} < x \leq a_k, \\ 0 & \text{jinak.} \end{cases}$$

nazveme četnostní hustotou.

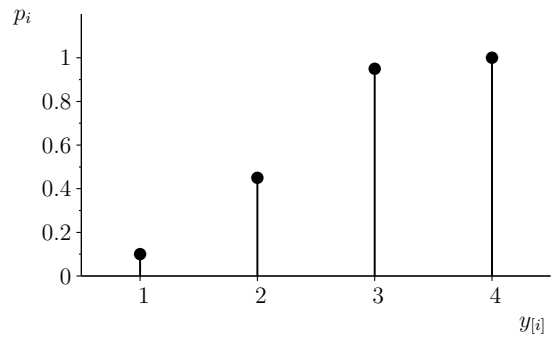
Uvedené četnosti a četnostní hustotu lze uspořádat do tabulky Tab. 1.5. Tato tabulka se pak nazývá tabulkou skupinového rozdělení četností znaku X.



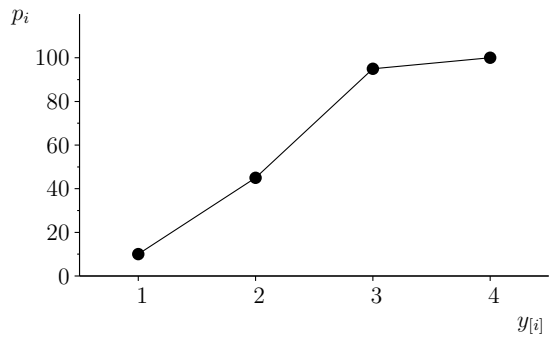
Obr 1.3a) Sloupcový diagram absolutních četností znaku Y



Obr 1.3b) Polygon relativních četností znaku Y



Obr 1.3c) Sloupcový diagram kumulativních relativních četností znaku Y



Obr 1.3d) Polygon kumulativních absolutních četností znaku Y

Obrázek 1.3: Grafické znázornění četností znaku Y

Obecně není třeba volit třídní intervaly stejné délky. V případě, že $d_1 = d_2 = \ldots = d_k$, mluvíme o **ekvidistantních intervalech**.

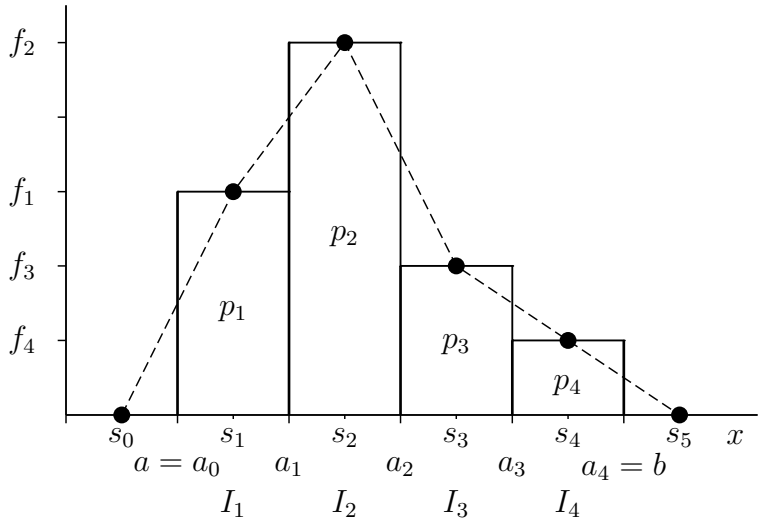
Grafickým znázorněním tabulky skupinového rozdělení je histogram nebo polygon.

Polygon skupinového rozdělení četností (absolutních, relativních, absolutních kumulativních nebo relativních kumulativních) konstruujeme stejně jako polygon rozdělení četností, jenom na osu x se místo variant znaku $x_{[1]}, \ldots, x_{[r]}$ vynášejí středy třídících intervalů $s_1, \ldots, s_k$. Polygon četnostní hustoty získáme tak, že úsečkami spojíme body o souřadnicích $[s_i, f_i], i = 0, 1, 2, \ldots, k, k+1$, přičemž klademe $f_0 = f_{k+1} = 0$ a $s_0 = a_0 - \frac{1}{2}d_1, s_{k+1} = a_k + \frac{1}{2}d_k$.

Histogramem rozumíme graf, který získáme, když na osu x vyneseme hranice třídních intervalů a nad každým třídním intervalem znázorníme úsečku rovnoběžnou s osou x ve výšce f_i nad intervalem I_i. Když potom svislými úsečkami spojíme hranice třídních intervalů s krajními body úseček, které jsme získali vynesením četnostní hustoty, získáme obdélníky a obsah i-tého z takto získaných obdélníků je p_i. Schodovitá čára, která shora omezuje histogram je grafem četnostní hustoty $f^*(x)$ a obsah plochy pod četnostní hustotou je 1, protože $p_1 + p_2 + \cdots p_k = 1$. Příklad histogramu je pro $k = 4$ na Obr. 1.4.

Třídní interval	Střed intervalu	Četnosti				Četnostní hustota
		n_i	p_i	N_i	P_i	
$I_1 = (a_0, a_1\rangle$	s_1	n_1	p_1	N_1	P_1	f_1
$I_2 = (a_0, a_2\rangle$	s_2	n_2	p_2	N_2	P_2	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$I_k = (a_{k-1}, a_k)$	s_k	n_k	p_k	N_k	P_k	f_k
Součet		n	1	–	–	–

Tabulka 1.5: Tabulka skupinového rozdělení četností a četnostní hustoty

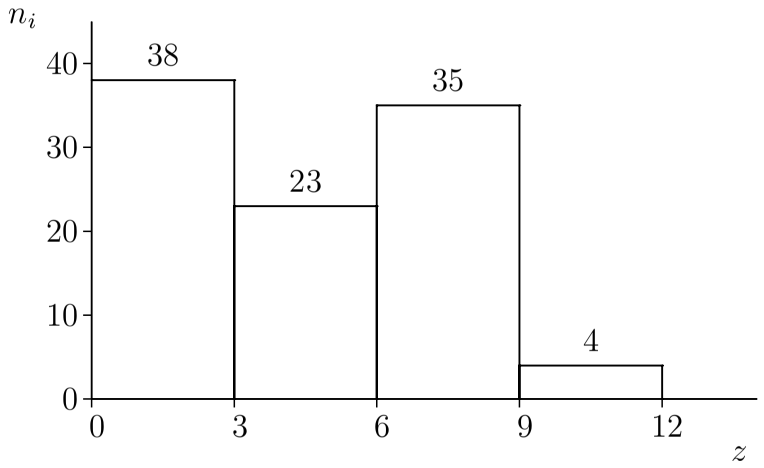


Obrázek 1.4: Histogram rozdělení četnosti je vynesen plnými čarami. Polygon četnostní hustoty je znázorněn přerušovanou čarou.

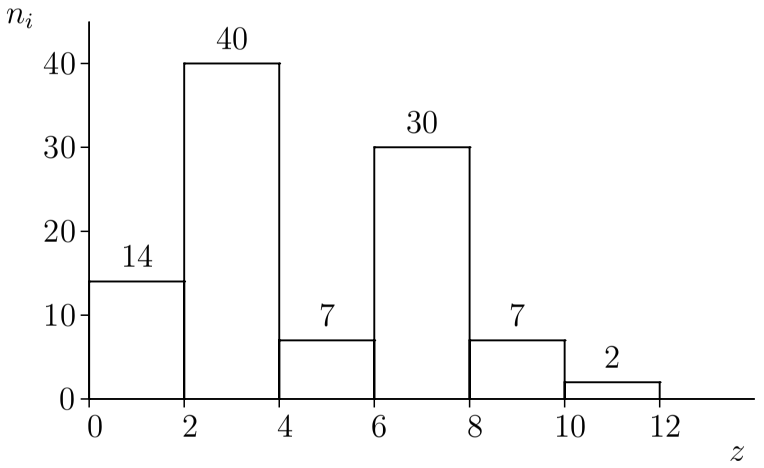
Poznamenejme ještě, že v mnoha praktických situacích se kromě uvedeného histogramu používá také histogram absolutních nebo relativních četností, případně histogram absolutních kumulativních četností nebo histogram relativních kumulativních četností. Tyto varianty histogramu se získají tak, že se při konstrukci histogramu na osu y vynáší místo četnostní hustoty f_i některá z četností n_i, p_i, N_i nebo P_i. Takto konstruované histogramy také dávají dobrou představu o skupinovém rozdělení sledovaného znaku, ovšem již neplatí, že obsah plochy pod takovým histogramem je 1.

Při stanovení skupinového rozdělení četností se ve většině praktických situací volí třídní intervaly ekvidistantní, tedy o stejné délce. Pro ekvidistantní třídní intervaly pak histogram konstruovaný pomocí četnostní hustoty a histogram konstruovaný pomocí absolutních nebo relativních četností liší pouze stupnicí na svislé ose. Při vhodné volbě této stupnice je jejich celkový vzhled shodný. Otázkou zůstává, jak

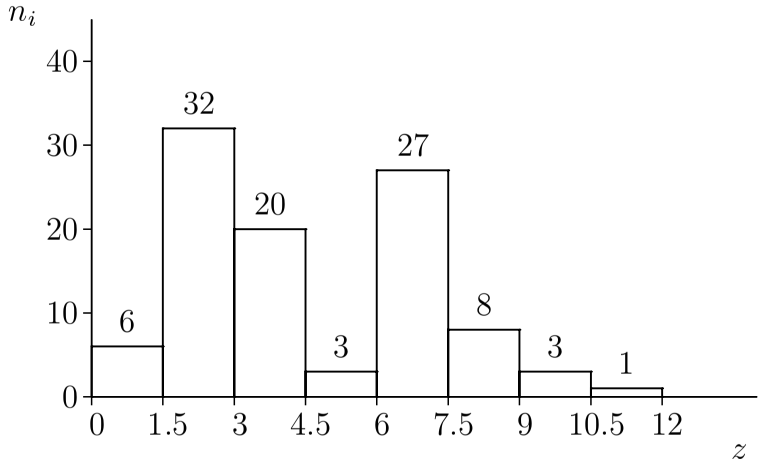
volit počet třídních intervalů k, který může vzhled histogramu podstatně ovlivnit. Názorně je tato situace demonstrována na Obr. 1.5 pro znak Z z příkladu 1.1 (Tab. 1.2).



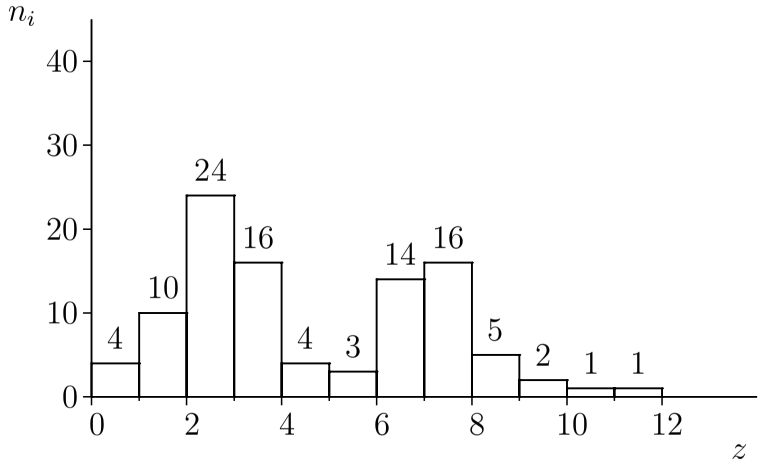
Obr 1.5a) Histogram znaku Z ekvidistantní délka tříd $d = 3$, počet tříd $k = 4$



Obr 1.5b) Histogram znaku Z ekvidistantní délka tříd $d = 2$, počet tříd $k = 6$



Obr 1.5c) Histogram znaku Z ekvidistantní délka tříd $d = 1.5$, počet tříd $k = 8$



Obr 1.5d) Histogram znaku Z ekvidistantní délka tříd $d = 1$, počet tříd $k = 12$

Obrázek 1.5: Vliv počtu třídních intervalů k na vzhled histogramu znaku Z

Počet různých hodnot znaku podle Sturgersova pravidla	Počet různých hodnot znaku podle Yulleova pravidla	Optimální počet tříd
3–5	3–6	3
6–11	7–16	4
12–22	17–33	5
23–45	34–61	6
46–90	62–104	7
91–181	105–167	8
182–362	168–256	9
363–724	257–374	10
...	...	...

Tabulka 1.6: Optimální počet tříd podle Sturgersova a Yulleova pravidla

V literatuře se pro volbu počtu tříd doporučují různé postupy. Nejčastěji se užívá tzv. **Sturgersovo pravidlo**, které doporučuje volit optimální počet tříd podle vzorce (viz. [**?**])

$$k \doteq 1 + 3.332 \log_{10}(n),$$

kde k je počet třídních intervalů a n je zde počet různých hodnot sledovaného znaku.

Jiné pravidlo pro volbu počtu tříd je tzv. **Yulleovo pravidlo**

$$k \doteq 2.5 \sqrt[4]{n}.$$

Podle jiného přístupu se pro kardinální znak doporučuje volit délku ekvidistantních tříd d od $0.08R$ do $0.12R$, kde R je tzv. rozpětí definované vztahem $R = x_{(n)} - x_{(1)}$, přičemž $x_{(1)}$ je nejmenší a $x_{(n)}$ největší pozorovaná hodnota znaku X v souboru. Pak se počet tříd k stanoví podle přibližného vzorce $k = \frac{R}{d}$.

Optimální počet tříd stanovený podle Sturgersova a Yulleova pravidla lze najít v závislosti na n v tabulce Tab. 1.6.

1.4 Empirická distribuční funkce a empirické kvantily

V předchozích odstavcích jsme se zabývali popisem rozdělení četností statistického znaku na daném statistickém souboru. V tomto odstavci zavedeme další možný přístup k popisu rozdělení četností daného statistického znaku. Budeme předpokládat,

že uvažovaný znak X je ordinální nebo kardinální a na daném souboru rozsahu n nabývá hodnot $x_1, x_2, \ldots, x_n$, které lze uspořádat do konečné neklesající posloupnosti $x_{(1)} \leq x_{(1)} \leq \cdots \leq x_{(n)}$. Tedy $x_{(1)}$ je nejmenší a $x_{(n)}$ největší hodnota mezi pozorováním $x_1, x_2, \ldots, x_n$.

Nejdříve zavedeme charakteristickou funkci množiny A (tzv. indikátor množiny A) vztahem

$$I_A(x) = \begin{cases} 1 & \text{když } x \in A, \\ 0 & \text{když } x \notin A. \end{cases}$$

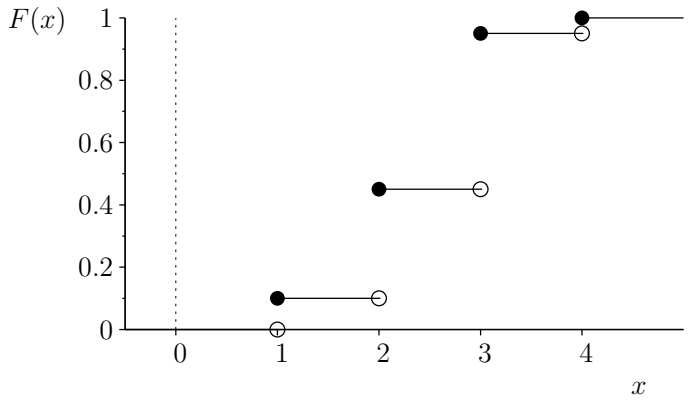
Proto pro libovolné $x \in (-\infty, \infty)$ položíme $A = (-\infty, x\rangle$ a snadno stanovíme $I_{(-\infty,x\rangle}(x_i) = 1$, když $x_i \leq x$ a $I_{(-\infty,x\rangle}(x_i) = 0$, když $x_i > x, i = 1, 2, \ldots, n$. Potom funkce

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x\rangle}(x_i)$$

pro dané x udává počet pozorování v souboru $x_1, x_2, \ldots, x_n$, která jsou nejvýše rovna x dělený rozsahem souboru n. Funkce $F_n^*(x)$ se nazývá **empirická distribuční funkce**. Pro daný statistický soubor dává o rozdělení četností podobnou informaci jako tabulka rozdělení četností nebo tabulka skupinového rozdělení četností.

Graf empirické distribuční funkce $F_n^*(x)$ snadno získáme tak, že na vodorovnou osu naneseme uspořádané hodnoty znaku $x_{(1)} \leq x_{(1)} \leq \cdots \leq x_{(n)}$. Tím získáme tzv. **diagram rozptýlení**. $F_n^*(x)$ je po částech konstantní neklesající, zprava spojitá funkce, v každém bodě $x_{(i)}$ má skok velikosti $\frac{1}{n}$ (platí-li, že hodnota $x_{(i)}$ je v daném souboru zastoupena n_i-krát, je skok v bodě $x_{(i)}$ roven velikosti $\frac{n_i}{n}$). Čtenář jistě vidí souvislost mezi $F_n^*(x)$ a kumulativními relativními četnostmi $N_i, i = 1, 2, \ldots, k$.

Empirická distribuční funkce znaku Y z příkladu 1.1 je znázorněna na Obr. 1.6.



Obrázek 1.6: Empirická distribuční funkce znaku Y

Empirickou distribuční funkci lze také konstruovat pro spojitý kardinální znak s

velkým počtem hodnot. Často se ale v této situaci používá její aproximace pomocí četnostní hustoty $f^*(x)$ tvaru

$$F_A^*(x) = \int_{-\infty}^{x} f^*(t)\mathrm{d}t.$$

Aproximace $F_A^*(x)$ závisí na zvolených třídních intervalech, zatímco empirická distribuční funkce $F_n^*(x)$ nikoliv.

Jsou-li data rozdělena do tabulky skupinového rozdělení četnosti, pak aproximace $F_A^*(x)$ empirické distribuční funkce $F_n^*(x)$ lze vyjádřit ve tvaru

$$F_A^*(x) = \begin{cases} 0 & \text{pro } x < a_0 \\ P_{j-1} + (x - a_{j-1})f_j = & \\ \quad = \sum_{i=1}^{j-1} p_i + p_j(x - a_{j-1})\frac{1}{d_j} & \text{pro } x \in I_j = (a_{j-1}, a_j\rangle, j = 1, \ldots, k, \\ 1 & \text{pro } x \geq a_k \end{cases}$$

Pomocí empirické distribuční funkce lze zavést tzv. **kvantilovou funkci**, kterou si lze představit jako zobecněnou inverzní funkci k empirické distribuční funkci $F_n^*(x)$. Zavádí se pro $p \in (0, 1)$ vztahem

$$F_{-1}^*(p) = \inf\{x : F_n^*(x) \geq p\},$$

kde $\inf\{A\}$ značí tzv. infimum číselné množiny A (viz. [**?**]) (Připomeňme, že pro konečnou množinu A značí $\inf\{A\}$ její nejmenší prvek a pro nekonečnou množinu se jedná o zobecnění pojmu minimálního prvku na nekonečnou množinu.)

Pro dané číslo $p \in (0, 1)$ se potom číslo $x_p = F_{-1}^*(p)$ nazývá **p-kvantilem** znaku X na souboru $x_1, \ldots, x_n$. Ze zavedené kvantilové funkce je dobře patrné, že p-kvantil x_p je číslo, které rozděluje uspořádanou řadu pozorování $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ na dvě části. První část hodnot obsahuje alespoň $100p\%$ hodnot z celého souboru, které jsou nejvýše rovné kvantilu x_p a druhá část obsahuje alespoň $100(1 - p)\%$ hodnot, které jsou větší nebo rovné než kvantil x_p. Kvantil x_p je důležitou charakteristiku statistického souboru a pro různá p poskytuje o statistickém souboru podobnou informaci jako tabulka rozdělení nebo skupinového rozdělení četností. Dříve než uvedeme ilustrativní příklad poznamenejme, že poslední slovní charakteristikou není kvantil x_p určen jednoznačně. Zavedení kvantilu pomocí kvantilové funkce už je jednoznačné.

Příklad 1.2 Určete kvantily $x_{0.1}, x_{0.25}, x_{0.50}$ a $x_{0.75}$ pro znak Y z příkladu 1.1.

Z grafu na Obr. 1.6 vidíme, že „nejmenší číslo x", pro které platí, že $F_{100}^*(x) \geq 0.1$ je číslo $x_{0.1} = 1$. Podobně stanovíme $x_{0.25} = 2$, $x_{0.5} = 3$ a $x_{0.75} = 3$. Zároveň vidíme, že procento hodnot znaku Y, které jsou nejvýše rovny $x_{0.25} = 2$ je 35%, což je více než $100p\%$=25% a zároveň procento hodnot znaku Y, které jsou větší nebo rovny

než kvantil $x_{0.25} = 2$ tvoří 90% hodnot souboru a to je procento větší nebo rovno než $100(1-p)\%$=75%. Také je dobře patrné, že kdybychom místo kvantilu $x_{0.25} = 2$ zvolili libovolné číslo z intervalu $\langle 2, 3\rangle$, pořád by platilo, že před $x_{0.25}$ a včetně $x_{0.25}$ leží alespoň 25% hodnota a za $x_{0.25}$ včetně $x_{0.25}$ také leží alespoň 75% hodnot. To je příklad nejednoznačnosti ve volbě kvantilu zmíněné v předchozím odstavci.