Jak vytváří statistika obrazy světa a života. I. díl

Část II. [odst. 3,1-3,16, 4,1]

In: Jaroslav Janko (author): Jak vytváří statistika obrazy světa a života. I. díl. (Czech). Praha: Jednota českých matematiků a fysiků, 1942. pp. 17–63.

Persistent URL: http://dml.cz/dmlcz/403050

ČÁST II.

(3,1) Metody k zhuštění informace vyjádřené posloupností původních dat. (Seřazení a úprava materiálu. Variační obor. Kvartily.)

Sebráním materiálu jsou prvky zkoumaného empirického souboru zastoupeny dotazníky nebo sčítacími lístky, v nichž jsou jejich vyšetřované znaky zapsány a na ně se vztahuje další zpracování. Údaj, který popisuje určitý znak se nazývá statistická proměnná (x). Je-li znak určen jednou proměnnou, nazývá se jednorozměrným, jinak vícerozměrným. Zjišťujeme-li údaje o vyšetřovaných pracích k určitému okamžiku čili studujeme stav souboru v něm, přihlížíme tedy k statické stránce problémů. Určujeme-li časové změny znaků a sledujeme tak změnu příslušnosti prvků do souboru v čase čili kinematiku souboru, přihlížíme k dynamické stránce problémů. V dalším se budeme zabývati jen statickou stránkou.

Pozorováním určitého kvantitativního znaku jsme na př. dostali v jednotkách míry (nebo váhy atd.) konkretní posloupnost hodnot odpovídajících rozsahu $r = 270$ prvků souboru

101	140	78	63	138	110	90	135	58	89
110	102	80	102	99	98	96	110	106	70
103	122	92	107	111	118	106	125	108	103
87	95	140	74	124	80	80	82	88	114
101	118	86	101	84	57	107	107	70	100
88	82	101	86	80	117	97	97	107	115
109	102	83	103	115	84	89	110	92	74
112	99	110	73	133	106	108	97	83	151
83	82	83	105	27	94	66	103	110	104
138	129	123	119	115	98	87	97	132	83
86	86	82	118	100	134	99	75	81	109

118	74	107	87	46	117	80	88	87	92
88	102	69	99	83	67	110	99	91	85
71	92	103	91	98	131	102	110	108	120
80	139	102	76	118	89	84	86	89	92
111	83	124	78	161	148	104	96	130	86
85	108	80	104	65	104	87	108	102	78
115	119	118	79	92	110	91	72	95	114
123	107	96	97	100	91	163	86	86	89
85	105	117	106	94	87	94	123	92	124
73	115	90	103	138	95	138	106	88	107
85	98	94	101	108	117	119	95	97	109
105	136	78	109	86	82	112	127	89	133
97	81	101	76	126	96	98	90	114	73
110	94	88	118	113	100	91	111	90	100
89	110	100	82	79	108	136	98	126	113
116	101	71	70	201	124	89	115	93	86

Na těchto datech uvedených ve formě neuspořádané prvotní tabulky ukážeme metody seřazení a soustředění do menšího počtu čísel. Prvním krokem v úpravě množiny pozorovaných hodnot jedné proměnné x je seřazení jejich podle velikosti, a to v pořadí hodnot neklesajících $x_1 \leqq x_2 \leqq \ldots \leqq x_r$. Při tom tedy najdeme nejmenší pozorovanou hodnotu proměnné $x_1 = 27$ a největší $x_r = 201$, jejichž rozdíl $x_r - x_1 = 174$ se nazývá variační obor nebo variační rozpětí, a snadno najdeme medián $\tilde{x} = 99$, před nímž je v seřazené posloupnosti r' členů a za ním r'' členů, při čemž $r' = r''$. Vyjmeme-li takový člen před nímž je r' a za ním r'' členů tak, že $3r' = r''$ dostáváme dolní kvartil $\tilde{x}_1 = 86$ před nímž je tudíž čtvrtina členů posloupnosti a za ním tři čtvrtiny.

Je-li $r' = 3r''$, nazýváme příslušný vyňatý člen horní kvartil $\tilde{x}_2 = 110$. Tak rozdělují oba kvartily a medián pozorované hodnoty na čtyři skupiny o stejném počtu prvků.

Seřazení hodnot x:

27	78	83	87	92	97	102	107	110	118	131
46	78	83	88	92	98	102	107	110	118	132
57	78	83	88	92	98	102	107	110*	118	133
58	79	84	88	92	98	102	107	110	118	133
63	79	84	88	92	98	102	107	111	118	134
65	80	84	88	93	98	102	107	111	118	135
66	80	85	88	94	98	102	107	111	118	136
67	80	85	89	94	99	103	108	112	119	136
69	80	85	89	94	99	103	108	112	119	138
70	80	85	89	94	99*	103	108	113	119	138
70	80	86	89	94	99*	103	108	113	120	138
70	80	86	89	95	99	103	108	114	122	138
71	81	86	89	95	100	103	108	114	123	139
71	81	86	89	95	100	104	108	114	123	140
72	82	86	89	95	100	104	109	115	123	140
73	82	86	90	96	100	104	109	115	124	148
73	82	86	90	96	100	104	109	115	124	151
73	82	86*	90	96	100	105	109	115	124	161
74	82	86	90	96	101	105	110	115	124	163
74	82	86	91	97	101	105	110	115	125	201
74	82	87	91	97	101	106	110	116	126	
75	83	87	91	97	101	106	110	117	126	
76	83	87	91	97	101	106	110	117	127	
76	83	87	91	97	101	106	110	117	129	
78	83	87	92	97	101	106	110	117	130	

(3,2) Momentové charakteristiky (obecné, kolem aritmetického průměru, momenty směrodatné proměnné). Takové uspořádání hodnot má někdy svůj význam v počátečním stadiu rozboru. Nelze však ani při poměrně nevelkém rozsahu souboru, jakým je náš příklad, zachytiti v mysli takové množství čísel v celku, proto je třeba zhušťování. K němu spějeme dvojí cestou. První cesta spočívá v tom, že si definujeme určité konstanty, které charakterisují takové posloupnosti. Snažíme se, aby byly definovány jednoduchým způsobem, aby byly snadno počitatelné a zahrnovaly všechny údaje.

Nejjednodušší charakteristikou, splňující tyto podmínky, je aritmetický průměr proměnné x, který se rovná součtu všech hodnot proměnné, dělenému jejich počtem. Označíme-li $\bar{x}$ aritmetický průměr nebo krátce průměr hodnot $x_1, x_2, \ldots, x_r$ pak tedy platí rovnice

$$\bar{x} = \frac{1}{r}(x_1 + x_2 + \ldots + x_r) = \frac{1}{r}\sum_{i=1}^{r} x_i. \qquad (1)$$

Aritmetický průměr se nazývá také prvním momentem. Zcela obdobně se pak definují další momenty, takže k-tý moment $\mu'_{x,k}$, který se také nazývá momentem k-tého řádu, jest průměrem k-tých mocnin hodnot proměnné, je tedy vyjádřen rovnicí

$$\mu'_{x,k} = \frac{1}{r}(x_1{}^k + x_2{}^k + \ldots + x_r{}^k) = \frac{1}{r}\sum_{i=1}^{r} x_i{}^k \qquad (2)$$

aritmetický průměr ovšem plyne z této rovnice pro $k = 1$, takže $\bar{x} = \mu'_{x,1}$ a další momentové charakteristiky dostáváme, klademe-li $k = 2, 3, 4, \ldots$

Vedle těchto obecných momentů mají ve statistice zvláštní význam momenty kolem aritmetického průměru.

Označíme-li odchylku jednotlivých hodnot proměnné od aritmetického průměru $\xi_i = x_i - \bar{x}$, potom momenty kolem aritmetického průměru $\mu_{x,k}$ definujeme

$$\mu_{x,k} = \frac{1}{r}\sum_{i=1}^{r} \xi_i{}^k. \qquad (3)$$

První moment kolem aritmetického průměru je roven nule, neboť pro $k = 1$

$$\mu_{x,1} = \frac{1}{r}\sum_{i=1}^{r} \xi_i = \frac{1}{r}\sum_{i=1}^{r} (x_i - \bar{x}) =$$
$$= \frac{1}{r}\sum_{i=1}^{r} x_i - \frac{r\bar{x}}{r} = \bar{x} - \bar{x} = 0. \tag{4}$$

Pro výpočet dalších momentů kolem aritmetického průměru pro $k = 2, 3, 4, \ldots$ mají význam vztahy, které platí mezi nimi a momenty obecnými. Snadno je odvodíme takto

$$\begin{array}{l}
\xi_1^2 = x_1^2 - 2x_1\bar{x} + \bar{x}^2 \\
\xi_2^2 = x_2^2 - 2x_2\bar{x} + \bar{x}^2 \\
\dots\dots\dots\dots\dots\dots \\
\xi_r^2 = x_r^2 - 2x_r\bar{x} + \bar{x}^2 \\
\hline
\sum_{i=1}^{r} \xi_i^2 = \sum_{i=1}^{r} x_i^2 - 2\bar{x}\sum_{i=1}^{r} x_i + r\bar{x}^2
\end{array}$$

Odtud plyne dělením r

$$\frac{1}{r}\sum_{i=1}^{r} \xi_i^2 = \frac{1}{r}\sum_{i=1}^{r} x_i^2 - 2\bar{x} \cdot \bar{x} + \bar{x}^2$$

čili

$$\mu_{x,2} = \mu'_{x,2} - \bar{x}^2. \tag{5}$$

Obdobně

$$\begin{array}{l}
\xi_1^3 = x_1^3 - 3x_1^2\bar{x} + 3x_1\bar{x}^2 - \bar{x}^3 \\
\xi_2^3 = x_2^3 - 3x_2^2\bar{x} + 3x_2\bar{x}^2 - \bar{x}^3 \\
\dots\dots\dots\dots\dots\dots\dots\dots \\
\xi_r^3 = x_r^3 - 3x_r^2\bar{x} + 3x_r\bar{x}^2 - \bar{x}^3 \\
\hline
\sum_{i=1}^{r} \xi_i^3 = \sum_{i=1}^{r} x_i^3 - 3\bar{x}\sum_{i=1}^{r} x_i^2 + 3\bar{x}^2\sum_{i=1}^{r} x_i - r\bar{x}^3
\end{array}$$

takže

$$\frac{1}{r}\sum_{i=1}^{r} \xi_i^3 = \frac{1}{r}\sum_{i=1}^{r} x_i^3 - 3\bar{x}\,\frac{1}{r}\sum_{i=1}^{r} x_i^2 + 3\bar{x}^2 \cdot \bar{x} - \bar{x}^3$$

a tedy

$$\mu_{x,3} = \mu'_{x,3} - 3\bar{x}\mu'_{x,2} + 2\bar{x}^3. \tag{6}$$

Stejně se odvodí

$$\mu_{x,4} = \mu'_{x,4} - 4\bar{x}\mu'_{x,3} + 6\bar{x}^2\mu'_{x,2} - 3\bar{x}^4. \tag{7}$$

Dalších momentů se užívá velmi zřídka a obecný vztah se snadno najde rozvedením $(x_i - \bar{x})^k$ podle binomické věty [11, 12].

Druhého momentu kolem průměru si zvláště povšimneme, neboť spočívá na součtu čtverců odchylek proměnné od průměru a proto charakterisuje rozptyl pozorovaných hodnot proměnné. Obyčejně se užívá k měření rozptylu čili variability jeho druhé odmocniny, která se nazývá směrodatná odchylka

$$\sigma_x = \sqrt{\mu_{x,2}}, \quad = \sqrt{\frac{1}{r}\sum \xi^2} \tag{8}$$

neboť pak je míra téhož rozměru jako pozorovaný znak. Při výpočtu vychází tedy v těchže jednotkách, v nichž jsou napozorované hodnoty proměnné a její čtverec $\sigma_x^2 = \mu_{x,2}$ nazýváme rozptyl.

Zavádíme ještě pojem „směrodatná proměnná" t_i tak, že měříme odchylky proměnné od průměru směrodatnou odchylkou čili vyjádříme je v jednotce „směrodatná odchylka", potom

$$t_i = \frac{x_i - \bar{x}}{\sigma_x} = \frac{\xi_i}{\sigma_x}, \tag{9}$$

čímž dostáváme čísla bez rozměru, můžeme říci abstraktní; poskytují však výhody při mnohých matematických operacích a usnadňují některá srovnávání.

Významné jsou některé vlastnosti momentů směrodatné proměnné

$$\mu_{t,1} = \bar{t} = \frac{1}{r}\sum_{i=1}^{r}\frac{x_i - \bar{x}}{\sigma_x} = \frac{1}{\sigma_x}\frac{1}{r}\sum_{i=1}^{r}\xi_i = 0,$$

$$\mu_{t,2} = \frac{1}{r}\sum_{i=1}^{r}\frac{(x_i - \bar{x})^2}{\sigma_x^2} = \frac{1}{\sigma_x^2}\frac{1}{r}\sum_{i=1}^{r}\xi_i^2 = \frac{\mu_{x,2}}{\sigma_x^2} = 1,$$

$$\mu_{t,3} = \frac{1}{r}\sum_{i=1}^{r}\frac{(x_i - \bar{x})^3}{\sigma_x^3} = \frac{1}{\sigma_x^3}\frac{1}{r}\sum_{i=1}^{r}\xi_i^3 = \frac{\mu_{x,3}}{\sigma_x^3}.$$

Ačkoliv tedy $\mu_{t,1} = 0$, $\mu_{t,2} = 1$, hodnota třetího momentu směrodatné proměnné $\mu_{t,3}$ závisí na hodnotách proměnné. Je známa pod názvem šikmost nebo kosost a označuje se symbolem $\alpha_{x,3}$, takže

$$\alpha_{x,3} = \frac{\mu_{x,3}}{\sigma_x^3} = \frac{\mu_{x,3}}{\mu_{x,2}\,\sigma_x}. \tag{10}$$

Postoupíme-li dále ke čtvrtému momentu, vidíme, že

$$\mu_{t,4} = \frac{\mu_{x,4}}{\sigma_x^4} \tag{11}$$

a označujeme jej pak obyčejně $\alpha_{x,4}$. Je výrazem špičatosti nebo plochosti; užívá se ho k tomu účelu ve tvaru $\alpha_{x,4} - 3$ a nazývá se koeficientem špičatosti nebo excesem.

Seznámili jsme se tedy se základními momentovými charakteristikami souboru, jimiž jsou:

1. rozsah souboru r,
2. aritmetický průměr hodnot proměnné $\bar{x}$,
3. směrodatná odchylka σ_x,
4. šikmost (kosost) $\alpha_{x,3}$,
5. špičatost (exces) $\alpha_{x,4} - 3$.

Jejich praktický výpočet vyžaduje zjednodušení výpočtu momentů. K tomu lze užíti **základního teorému o momentech**, který zní: momenty kolem aritmetického průměru se nemění, zvětší-li se nebo zmenší-li se všechny hodnoty proměnné o stejnou konstantu.

Důkaz provedeme snadno, odečteme-li na př. ode všech hodnot proměnné x hodnotu x_0, které se někdy říká předběžný průměr. Potom je $x_i - x_0 = \eta_i$ nová proměnná, pro niž jsou charakteristiky definovány stejně jako pro x, čili průměr

$$\bar{\eta} = \frac{1}{r}\sum_{i=1}^{r}\eta_i,$$

a k-tý moment kolem průměru

$$\mu_{\eta,k} = \frac{1}{r}\sum_{i=1}^{r}(\eta_i - \bar{\eta})^k.$$

Především dostáváme

$$\sum_{i=1}^{r} x_i = r x_0 + \sum_{i=1}^{r} \eta_i \quad \text{čili} \quad \bar{x} = x_0 + \bar{\eta}.$$

Můžeme tedy psáti

$$\eta_i - \bar{\eta} = (x_i - x_0) - (\bar{x} - x_0) = x_i - \bar{x} = \xi_i$$

a tudíž moment

$$\mu_{\eta,k} = \frac{1}{r}\sum_{i=1}^{r}(\eta_i - \bar{\eta})^k = \frac{1}{r}\sum_{i=1}^{r}\xi_i^k = \mu_{x,k}, \tag{12}$$

což je důkazem, že se nezměnil.

(3,3) Tabelární podávání výsledků. Rozdělení četností. Druhá cesta ke zhuštění pozorovaných dat vede přes tabulku rozdělení četností.

Statistický soubor převedeme do tabulky rozdělení četností, která má dva sloupce. V prvním jsou seřazeny podle velikosti jen různé hodnoty proměnné, které byly pozorovány. V druhém sloupci je uveden počet prvků, na nichž byla každá z těchto hodnot při šetření zjištěna. Tomuto počtu prvků s hodnotou proměnné x_i říkáme četnost n_i.*) I tato tabulka rozdělení četností je ještě málo přehledná. Proto postupujeme dále ke skupinovému rozdělení četností.

(3,4) Skupinové rozdělení četností. Tento typ tabulky vzniká tak, že několik hodnot proměnné se sdruží k utvoření jednoho intervalu a četnost se uvádí jedna pro celý interval, zvaný třída. Tak máme v prvním sloupci třídní intervaly a v druhém sloupci celkovou četnost všech hodnot proměnné,

*) Čtenář si laskavě sám napíše tuto tabulku z čísel našeho příkladu na str. 19.

které spadají do intervalu. Tím že tato tabulka již nepodává četnost každé původní pozorované hodnoty, nepředstavuje je již přesně a něco z původní informace se ztrácí v zájmu přehlednějšího obrazu o všeobecném tvaru rozdělení četností.

V našem číselném příkladu sdružíme třeba hodnoty 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, do jednoho intervalu, který bude zastupován zpravidla jejich průměrem, zde 90, a počet prvků spadajících do této třídy bude 80, což je tedy třídní četnost n_i. Tento interval má délku 15 jednotek, v nichž byl znak měřen. Postupujeme pak podél stupnice měření a zachovávajíce stálou délku intervalu, rozdělíme celé variační rozpětí na stejné intervaly a celý soubor na třídy. Při této konstrukci jak patrno, vznikají dvě důležité otázky. Jednak třeba stanoviti délku čili velikost intervalu s čímž souvisí jejich počet, jednak vymeziti hranice intervalů. S tím souvisí také otázka počátku prvního intervalu. Řešení těchto otázek se poněkud liší dle toho, jedná-li se o znak rozpojitý, kdy proměnná nabývá jen isolovaných hodnot, nebo o znak spojitý, kdy probíhá všechna reálná čísla určitého intervalu. Nejprve přihlédneme k prvnímu případu.

(3,5) Délka a hranice třídního intervalu. Není-li specielních potřeb daných přímo účelem šetření, pak určují volbu velikosti třídního intervalu dvě všeobecné podmínky:

1. hodnoty proměnné, zařazené do třídního intervalu lze pokládati s hlediska cíle šetření za zastupitelné průměrnou hodnotou třídního intervalu, která je zpravidla totožná s prostřední hodnotou. (Chceme-li sestaviti na př. tabulku úmrtnosti nějakého souboru osob podle věkových skupin, rozhodneme se pro interval 3- nebo 5-letý, podle toho, stačí-li při použití této tabulky zastoupení úmrtnosti krajních věků intervalu pětiletého či tříletého úmrtností věku prostředního.)

2. Při zachování první podmínky má býti délka intervalu co největší. V praksi bývají tyto podmínky splněny nejčastěji volbou intervalu takové délky, že se soubory podle velikosti rozsahu rozpadnou do 10 až 20 tříd. Vycházíme-li z této zkušenosti, stanovíme přibližně délku intervalu, dělíme-li variační rozpětí $(x_r - x_1)$ počtem tříd, který je zvolen tak, aby při malém rozsahu souboru byly třídy obsazeny, čili měly dostatečnou četnost a při velkém rozsahu, aby bylo rozdělení četností přehledné.

Také lze určovati velikost třídního intervalu se zřetelem ke směrodatné odchylce, kterou k tomu účelu zhruba odhadneme z předpokladu, že variační rozpětí se přibližně rovná šestinásobku směrodatné odchylky (pravidlo šesti sigma [5]). Velikost třídního intervalu h potom určíme tak, aby splňovala nerovnosti $2h < (x_r - x_1) : 6 < 4h$.

Tato libovůle v určování velikosti třídního intervalu a tím počtu tříd je ovšem pro matematickou statistiku velmi nepříjemná. Jsou proto odvozeny také způsoby určování, spočívající na porovnání s binomickým rozdělením četností [str. 66. rovnice (37)].

Druhým úkolem je stanovení hranic intervalů. Při znaku rozpojitém se snažíme stanovit dolní a horní hranici intervalu tak, aby bylo o každé hodnotě proměnné jasno, do kterého intervalu patří. Obyčejně se stanoví v polovině mezi jednotkami posledního místa, v němž byly hodnoty proměnné uvedeny. V našem příkladu se jedná o čísla celá, takže interval svrchu zmíněný vyznačíme hranicemi 82,5 až 97,5. Střed intervalu, který zastupuje všechny hodnoty do něho spadající, zůstává tak číslo celé.

Zpravidla zavádíme všechny intervaly téže délky, ač materiál si někdy vynutí výjimky, zvláště tehdy, kdy by byl rozsah stupnice příliš veliký a některé obory na př. vysokých hodnot velmi řídce obsazeny. Také někdy první interval bývá dolu neohraničený, takže jsou do něho zařazeny všechny hodnoty až do stanovené horní hranice jeho; podobně je někdy poslední interval neohraničený nahoru. Poloha intervalu stanovená dolní hranicí bývá dosti libovolná a její volba nemá velkého vlivu na hodnoty charakteristik; někdy však vyplývá z povahy materiálu.

Kupí-li se na př. hodnoty pozorované nápadně kolem určitých čísel na př. 5 nebo 10, snažíme se, aby tato čísla padla do středu intervalu, který jej zastupuje. Hodnoty, které zastupují intervaly, budeme opět značit x_i, $i = 1, 2 \ldots$, takže hranice intervalů délky h budou

$$x_i - \tfrac{1}{2}h,\ x_i + \tfrac{1}{2}h.$$

V případě spojitého znaku nemůžeme uvésti všechny jeho hodnoty. Rozdělíme zase jeho celé variační rozpětí na l stejných částí a dostaneme intervaly $(x_i - \tfrac{1}{2}h,\ x_i + \tfrac{1}{2}h)$, při čemž $x_i + \tfrac{1}{2}h = x_{i+1} - \tfrac{1}{2}h$. Zařazování prvků do tříd se může provádět trojím způsobem: do intervalu spadají hodnoty znaku x splňující nerovnosti

1. $x_i - \tfrac{1}{2}h < x \leqq x_i + \tfrac{1}{2}h$,

2. $x_i - \tfrac{1}{2}h \leqq x < x_i + \tfrac{1}{2}h$,

3. $x_i - \tfrac{1}{2}h < x < x_i + \tfrac{1}{2}h$, při čemž tam bude zařazena polovina prvků se znakem $x_i - \tfrac{1}{2}h$ a polovina prvků se znakem $x_i + \tfrac{1}{2}h$. Není-li stanovena dolní hranice prvního intervalu, zařadí se tam všechny hodnoty proměnné $x < < x_1 + \tfrac{1}{2}h$, po případě $x \leqq x_1 + \tfrac{1}{2}h$. Do posledního intervalu neohraničeného shora jsou pak zahrnuty hodnoty proměnné $x \geqq x_l - \tfrac{1}{2}h$, po případě $x > x_l - \tfrac{1}{2}h$.

Hodnoty spojité proměnné mohou býti měřeny jen s určitou přesností, takže také v tomto případě vycházejí ze statistického šetření jednotlivé hodnoty isolované. Je důležito, aby stanovené hranice ukazovaly, na které desetinné místo bylo měřeno [5]. Toho docilujeme buď tím, že jsou v hranicích přímo vyznačeny krajní hodnoty, které do intervalu spadají, nebo hranice vymezují hodnoty spadající do intervalu pomocí dalšího desetinného místa.

Objasníme si to příkladem. Představme si, že byl stanoven znak u každého prvku souboru na setiny určité jednotky míry (cm, kg, ...) a hned zaokrouhlován na desetiny; to znamená, že dostaneme výsledky, o nichž říkáme, že byly měřeny s přesností na desetiny.

Zaokrouhlení bylo prováděno třeba podle dohody, že zlomky 0,01 až 0,04 se zanedbají, 0,06 až 0,09 dávají 0,1 a 0,05 u sudého čísla na předchozím desetinném místě dává 0,1, u lichého se zanedbá.

Kdybychom tvořili na př. intervaly délky 0,5, tedy po pěti hodnotách znaku (počet lichý) dostali bychom třeba 80,0—80,4, 80,5—80,9, ... a střed intervalu čili třídní znak bude 80,2, 80,7, ...

Kdybychom tvořili intervaly délky 1,0, tedy po desíti (sudý počet), byl by střed intervalu a tedy znak 80,45.

Mohli bychom však při měření s přesností na desetiny vyznačit hranice v prvním případě čísly 79,95, 80,45, 80,95, ... což také určuje jednoznačně, že v prvém intervalu jsou hodnoty 80,0, 80,1, 80,2, 80,3, 80,4 a podobně ve druhém. Tím je současně vyjádřeno, že z hodnot stanovených na setiny spadají do prvního intervalu 79,96—80,44 a do druhého 80,45—80,95, z čehož také vyplývá střed 80,2, 80,7.

V druhém případě pak spadají do intervalu hodnoty 79,96—80,95, takže střed je 80,45, neboť 5 na dalším místě se zanedbává, když předchází liché číslo.

(3,6) Sestrojení tabulky skupinového rozdělení četností pro daný příklad. Sestrojme nyní skupinové rozdělení četností pro materiál našeho příkladu. Znak byl měřen na celé jednotky. Stanovíme hranice na př. druhým způsobem pomocí dalšího desetinného místa. Zvolíme interval délky $h = 15$. Dolní hranice prvního intervalu pak bude $x_1 - \frac{1}{2}h = 22{,}5$. Dostaneme tak tabulku o 12 třídách, kde v prvním sloupci uvedeme hranice intervalů, v druhém průměr hodnot proměnné spadajících do intervalu čili třídní znak a ve třetím sloupci příslušnou pozorovanou četnost třídní.

Tabulka 1.

Třídní		Četnost	Kumulativní četnost	Relativní četnost	Kumulativní relativní četnost
hranice	znak				
1	2	3	4	5	6
22,5					
37,5	30	1	1	0,4	0,4
52,5	45	1	2	0,4	0,8
67,5	60	6	8	2,2	3,0
82,5	75	38	46	14,1	17,1
97,5	90	80	126	29,6	46,7
112,5	105	83	209	30,7	77,4
127,5	120	39	248	14,5	91,9
142,5	135	17	265	6,3	98,2
157,5	150	2	267	0,7	98,9
172,5	165	2	269	0,7	99,6
187,5	180	0	269	0,0	99,6
202,5	195	1	270	0,4	100,0
Celkem		270		100,0	

$r = 270$

Vzhledem k poměrně malému rozsahu souboru lze provésti roztřídění do skupin stanovených třídními intervaly buď metodou skládání lístků nebo metodou čárkovací, kterou si znázorníme takto:

22,5 až 37,5	37,5 až 52,5	52,5 až 67,5	67,5 až 82,5	82,5 až 97,5	97,5 až 112,5	112,5 až 127,5	127,5 až 142,5	142,5 až 157,5	157,5 až 172,5	172,5 až 187,5	187,5 až 202,5
1	1	6	38	80	83	39	17	2	2	0	1

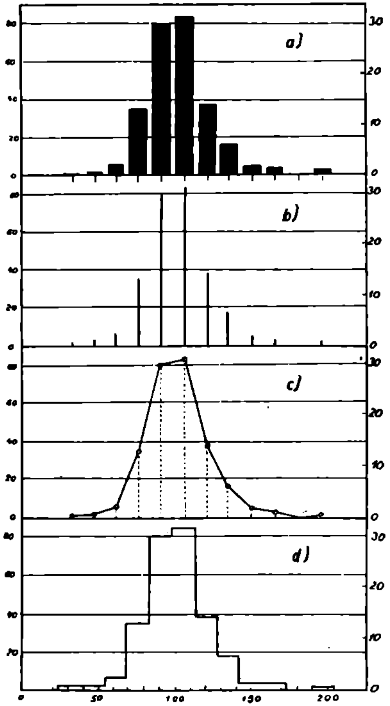
K některým účelům je důležito znáti počet prvků, jejichž hodnota znaku nepřevyšuje hranici určitého intervalu třídního; ten počet udává kumulativní čili součtové rozdělení četností, které dostáváme z rozdělení četností (sloupec 3) zapsáním četnosti první třídy a postupným načítáváním jednotlivých třídních četností k předchozímu součtu (sloupec 4) $s_i = s_{i-1} + n_i$, $s_1 = n_1$.

Tato tabulka nám udává na př. ihned počet prvků, jejichž hodnota znaku $x \leq x_i + \frac{1}{2}h$ a tedy snadno pak počet ostatních, jejichž hodnota znaku je větší čili četnosti třídění dichotomického. V našem případu, se jedná o znak kvantitativní; při něm dichotomické třídění znamená, že buď se vyskytují jeho hodnoty až do $x_i + \frac{1}{2}h$ nebo se nevyskytují. Je to obdoba alternativy při znaku kvalitativním (muž — žena). Třídíme-li podle mnoha stupňů znaku, nazýváme takové třídění polytomické. Kromě napozorované četnosti prvků, t. zv. absolutní četnosti třídní, má v dalších úvahách velký význam relativní četnost třídní, která je podílem absolutní třídní četnosti a celkového rozsahu souboru $f_i = \frac{n_i}{r}$. Velmi často se uvádějí relativní četnosti v procentech (sloupec 5).

Tabulka relativních četností, jejichž součet se rovná jedné $\sum_{i=1}^{l} f_i = 1$ je jen tehdy úplná, je-li současně uveden rozsah souboru r. Z ní se zase odvozuje kumulativní rozdělení relativních četností (sloupec 6) $F_i = F_{i-1} + f_i$.

Seznámili jsme se s hlavními metodami zhuštěného podávání výsledků jednak pomocí charakteristik, jednak formou tabelární.

(3,7) Grafické podávání statistických výsledků. K dalšímu usnadnění přehledné a jasné představy o studovaném souboru užívá statistika grafického podávání výsledků šetření. Vytváří grafické profily souboru tím, že znázorňuje jeho rozdělení četností podle vyšetřovaných znaků.

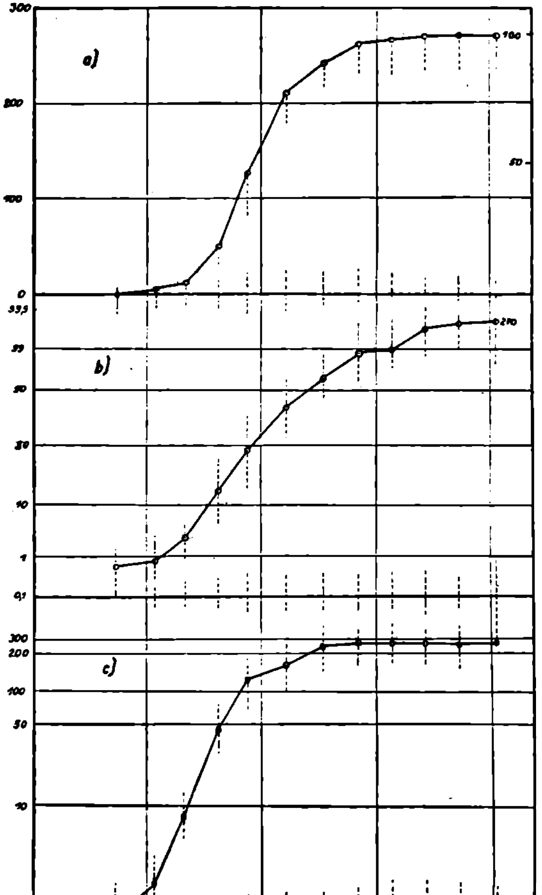


Obr. 3. Grafické znázornění rozdělení četností z tab. 1.

Používá se k tomu účelu v systému pravoúhlých souřadnic vodorovné osy úseček pro stupnici hodnot statistické proměnné a osy pořadnic pro četnosti. Znázornění lze pak provésti několika způsoby.

Diagram tyčkový vzniká vztyčením řady pruhů výšky rovnající se třídní četnosti, jejichž střed je v prostředních bodech tříd, t. j. v bodech znázorňujících třídní znak (obr. 3a). Šířka pruhů bývá různá; zvláště často se užívá úzkých úseček délky rovné třídní četnosti (obr. 3b) vztyčených v bodech třídních znaků. Velmi vhodným prostředkem grafického znázornění je mnohoúhelník (polygon) četností (obr. 3c), který dostaneme, jestliže tečky nebo kroužky ve výši třídních četností nad body třídních znaků spojíme úsečkami. Krajové tečky se spojí se středem nejbližšího intervalu na ose úseček, čímž je polygon uzavřen. Oblíbeným je znázornění pomocí histogramu četností čili sloupkový diagram sestávající z obdélníků, jejichž základna se rovná intervalu třídnímu a výška třídní četnosti, dělené délkou intervalu (obr. 3d). V histogramu představují plochy obdélníků třídní četnosti; je to jako bychom měřili stupnice třídním intervalem jako jednotkou. Kraje sloupků představují třídní hranice. V obraze a) až d) je současně patrno, že znázorňují také relativní četnosti uvedené na pravé straně obrazu, takže bývá výhodné užíti obou stupnic. Jako při tabulce, tak i při diagramu relativních četností nemá scházeti uvedení rozsahu souboru.

Znázorníme-li způsobem c) data sloupce 4. nebo 6., ale pro horní hranice příslušného intervalu, dostaneme kumulativní diagram četnosti neboli ogiv (obr. 4a), kde je vyznačena na levo aritmetická stupnice pro četnosti absolutní a napravo pro četnosti relativní, vyjádřené v procentech. Užívá se s prospěchem pro relativní četnosti také stupnice nomografické (pravděpodobnostní, obr. 4b), která převádí součtovou křivku Gaussovu (str. 84) na přímku [9], [7]. K výkladu stupnice a účelnosti její můžeme přistoupit až později.



Obr. 4. Grafické znázornění kumulativní četnosti. (Součtové křivky.)

a) V aritmetické stupnici pro absolutní i relativní četnosti.

b) V pravděpodobnostní stupnici pro relativní četnosti.

c) V logaritmické stupnici pro absolutní i relativní četnosti.

Stupnice logaritmické se užívá tam, kde je třeba diagramu méně citlivého na malé variace, kde by tedy v citlivém diagramu nevynikl celkový hlavní průběh (obr. 4c).

Úloha: Znázorněte pomocí histogramu a) úmrtnost mužů na rakovinu a b) úmrtnost obojího pohlaví na chřipku, která je uvedena v počtu případů na 10 000 žijících v letech věkové stupnice s nestejnými intervaly. Četnost v první věkové třídě se vztahuje na 10 000 živě narozených. V případě a) dostanete tak zv. J-křivku a v případě b) tak zv. U-křivku.

Věková třída	a)	b)
0—	0,1	13,7
1—	0,1	2,3
5—	0,02	0,6
15—	0,2	0,7
30—	9,4	1,5
60—	65,0	7,8
70—	100,9	19,2

Dalšího zhuštění pozorovaného materiálu dosahujeme spojením obou dříve naznačených cest, které vyžaduje, abychom upravili způsob výpočtu momentových charakteristik pro skupinové rozdělení četností.

(3,8) Základní charakteristiky a jejich výpočet pro skupinové rozdělení četností. Je-li z celé posloupnosti r hodnot x_i jen l hodnot od sebe různých, sestavujeme jednoduchou tabulku rozdělení četností

$$\frac{x_1, x_2, \ldots, x_l}{n_1, n_2, \ldots, n_l},$$

pro niž pak obecné momenty jsou vyjádřeny rovnicí

$$\mu'_{x,k} = \frac{1}{r}(n_1 x_1^k + n_2 x_2^k + \ldots + n_l x_l^k) = \frac{1}{r}\sum_{i=1}^{l} n_i x_i^k, \qquad (13)$$

při čemž $n_1 + n_2 + \ldots + n_l = r$. Vidíme totiž, že součty $\sum_{i=1}^{l} n_i x_i^k$ jsou numericky ekvivalentní součtům $\sum_{i=1}^{r} x_i^k$. Na př. pro $k = 1$ bude $\sum_{i=1}^{l} x_i n_i = \sum_{i=1}^{r} x_i$ pro $k = 2$ je $\sum_{i=1}^{l} x_i^2 n_i = \sum_{i=1}^{r} x_i^2$ atd. V součtech $\sum_{i=1}^{l} x_i^k n_i$ jsou již sečteny určité skupiny hodnot x^k, a to těch, které jsou stejné.

Pro momenty kolem aritmetického průměru platí totéž co jsme dříve odvodili a rovněž vztahy mezi nimi a momenty obecnými se nemění.

(3,9) Výpočet momentů metodou vhodně zvoleného počátku. Přistoupíme nyní k výpočtu momentů pro skupinové rozdělení četností. Výpočet se zjednoduší zpravidla metodou vhodně zvoleného počátku nebo metodou součtovou. Provedeme jej nejprve pro náš numerický příklad první metodou.

Máme l tříd četnosti a tedy l třídních znaků, které zastupují všechny hodnoty proměnné. Četnost n_i náleží celé třídě, v níž všechny hodnoty proměnné jsou zastoupeny třídním znakem x_i. Je tudíž obecný moment dán výrazem

$$\mu'_{x,k} = \frac{1}{r}\sum_{i=1}^{l} x_i^k n_i.$$

Zjednodušení výpočtu dosáhneme tím, že zavedeme novou proměnnou, pro niž zvolíme nový vhodný počátek x_0 a budeme ji měřit délkou intervalu jako novou jednotkou. Celý výpočet tedy bude proveden v jednotce h, v délce třídního intervalu. Nová proměnná bude

$$u_i = \frac{x_i - x_0}{h} = \frac{\eta_i}{h}, \text{ takže } x_i = hu_i + x_0. \qquad (14)$$

Aritmetický průměr se vyjádří takto

$$\bar{x} = \frac{1}{r}\sum_{i=1}^{l} x_i n_i = \frac{1}{r}\sum_{i=1}^{l}(hu_i + x_0)\, n_i =$$
$$= \frac{1}{r}\left[h\sum_{i=1}^{l} u_i n_i + x_0 \sum_{i=1}^{l} n_i\right],$$

z čehož je patrno, že

$$\bar{x} = h\bar{u} + x_0. \tag{15}$$

Druhý moment kolem aritmetického průměru pro proměnnou u_i odvodíme podobně.

Druhý obecný moment pro proměnnou $\eta_i = x_i - x_0$ je

$$\mu'_{\eta,2} = \frac{1}{r}\sum_{i=1}^{l}\eta_i^2 n_i = \frac{1}{r}\, h^2 \sum_{i=1}^{l} u_i^2 n_i.$$

Víme z rovnice (5), že

$$\mu_{\eta,2} = \mu'_{\eta,2} - \bar{\eta}^2$$

a vzhledem k základnímu teorému o momentech $\mu_{x,k} = \mu_{\eta,k}$ tedy také

$$\mu_{x,2} = \mu'_{\eta,2} - \bar{\eta}^2 = h^2\left[\frac{1}{r}\sum_{i=1}^{l} u_i^2 n_i - \left(\frac{1}{r}\sum_{i=1}^{l} u_i n_i\right)^2\right]$$

čili

$$\mu_{x,2} = h^2[\mu'_{u,2} - \bar{u}^2],$$

což lze psáti také

$$\mu_{x,2} = h^2 \mu_{u,2} \tag{16}$$

a směrodatná odchylka bude tudíž vyjádřena rovnicí

$$\sigma_x = h\sigma_u. \tag{17}$$

Pro třetí moment dostaneme

$$\mu'_{\eta,3} = \frac{1}{r}\sum_{i=1}^{l}\eta_i^3 n_i = \frac{1}{r}\, h^3 \sum_{i=1}^{l} u_i^3 n_i,$$
$$\mu_{\eta,3} = \mu'_{\eta,3} - 3\bar{\eta}\mu'_{\eta,2} + 2\bar{\eta}^3.$$

a podle základního teorému o momentech

$$\mu_{x,3} = \mu_{\eta,3} =$$

$$= h^3 \left[\frac{1}{r} \sum_{i=1}^{l} u_i^3 n_i - 3 \left(\frac{1}{r} \sum_{i=1}^{l} u_i n_i \right) \left(\frac{1}{r} \sum_{i=1}^{l} u_i^2 n_i \right) + 2\bar{u}^3 \right]$$

$$\mu_{x,3} = h^3 [\mu'_{u,3} - 3\bar{u}\mu'_{u,2} + 2\bar{u}^3]$$

čili

$$\mu_{x,3} = h^3 \mu_{u,3}. \tag{18}$$

Vzhledem k tomu je patrno, že šikmost nebo kosost se touto změnou proměnné nemění, neboť

$$\alpha_{x,3} = \frac{\mu_{x,3}}{\sigma_x^3} = \frac{\mu_{u,3}}{\sigma_u^3} = \alpha_{u,3}. \tag{19}$$

Stejným způsobem si čtenář ukáže, že platí pro čtvrté momenty

$$\mu_{x,4} = h^4 \mu_{u,4} \text{ a tudíž } \alpha_{x,4} = \alpha_{u,4}. \tag{20}$$

Obecné odvození pro k-tý moment nečiní potíží.

Kontrola numerického výpočtu momentů proměnné u se provádí t. zv. metodou posunutých momentů čili Charlierovým testem. Tento postup se zakládá na binomické větě

$$(u_i + 1)^3 = u_i^3 + 3u_i^2 + 3u_i + 1,$$

takže vynásobíme-li třídní četností n_i a sečteme pro všechna i dostaneme

$$\sum_{i=1}^{l} (u_i + 1)^3 n_i = \sum_{i=1}^{i} u_i^3 n_i + 3 \sum_{i=1}^{l} u_i^2 n_i + 3 \sum_{i=1}^{l} u_i n_i + \sum_{i=1}^{l} n_i. \tag{21}$$

Počítáme-li momenty až do čtvrtého řádu, pak provádíme kontrolu podle $(u_i + 1)^4$.

(3,10) Výpočet momentů metodou součtovou. Také při této metodě zvolíme pomocný počátek na příklad tak, že první hodnotě znaku, pro kterou se v tabulce vyskytuje nějaká četnost, přidělíme znak $u = 1$, znak další třídy označíme $u = 2$ atd. Sčítáme pak četnosti zdola přes celou tabulku a pro každou třídu vyznačíme příslušný mezisoučet. Tento součtový sloupec pak znovu sečítáme zdola a to opakujeme tolikrát, kolik momentů potřebujeme. Poslední součet v každém sloupci označíme postupně $S_0, S_1, S_2, \ldots, S_k \ldots$ Můžeme si odvoditi, že

$$S_0 = \sum_i n_i = N, \qquad S_1 = \sum_i i n_i,$$

$$S_2 = \sum_i \frac{i\,(i+1)}{2!} n_i, \ldots, S_k = \sum_i \binom{i+k-1}{k} n_i, \ldots$$

Označíme-li $s_k = \dfrac{S_k}{S_0}$ a tyto hodnoty s_k vyjádříme pomocí momentů proměnné u kolem počátku $u = 0$, dostaneme

$$
\begin{aligned}
s_1 &= \mu'_{u,1} = \bar{u}\\
s_2 &= \tfrac{1}{2}(\mu'_{u,2} + \bar{u})\\
s_3 &= \tfrac{1}{6}(\mu'_{u,3} + 3\mu'_{u,2} + 2\bar{u})\\
s_4 &= \tfrac{1}{24}(\mu'_{u,4} + 6\mu'_{u,3} + 11\mu'_{u,2} + 6\bar{u}).
\end{aligned}
$$

Z těchto hodnot pak plynou vzorce pro obecné momenty proměnné u

$$
\begin{aligned}
\bar{u} &= \mu'_{u,1} = s_1\\
\mu'_{u,2} &= 2s_2 - s_1\\
\mu'_{u,3} &= 6s_3 - 6s_2 + s_1\\
\mu'_{u,4} &= 24s_4 - 36s_3 + 14s_2 - s_1.
\end{aligned}
$$

Momenty kolem aritmetického průměru určíme obvyklým způsobem dříve uvedeným. Výrazy pro výpočet momentů kolem aritmetického průměru přímo z hodnot s_k jsou dosti složité a proto je neuvádíme.

Součtové metody se méně používá, protože pracuje s velkými čísly, zvláště při větším počtu tříd a vyšších momentech. Početní postup při jejím použití je nejlepé patrný z příkladu podle tab. 2.

x_i	u_i	n_i	$\Sigma(3)$	$\Sigma(4)$	$\Sigma(5)$	$\Sigma(6)$
(1)	(2)	3	4	5	6	7
30	1	1	270	1530	5345	14723
45	2	1	269	1260	3815	9378
60	3	6	268	991	2555	5563
75	4	38	262	723	1564	3008
90	5	80	224	461	841	1444
105	6	83	144	237	380	603
120	7	39	61	93	143	223
135	8	17	22	32	50	80
150	9	2	5	10	18	30
165	10	2	3	5	8	11
180	11	—	1	2	3	4
195	12	1	1	1	1	1
Σ		270	1530	5345	14723	35069
		S_0	S_1	S_2	S_3	S_4

$S_0 = 270$

$S_1 = 1530 \qquad s_1 = \dfrac{S_1}{S_0} = 5{,}6667$

$S_2 = 5345 \qquad s_2 = \dfrac{S_2}{S_0} = 19{,}7963$

$S_3 = 14723 \qquad s_3 = \dfrac{S_3}{S_0} = 54{,}5296$

$S_4 = 35069 \qquad s_4 = \dfrac{S_4}{S_0} = 129{,}8852$

$$\mu'_{u,1} = \bar{u} = s_1 = 5{,}667$$
$$\mu'_{u,2} = 2s_2 - s_1 = 33{,}926$$
$$\mu'_{u,3} = 6s_3 - 6s_2 + s_1 = 214{,}067$$
$$\mu'_{u,4} = 24s_4 - 36s_3 + 14s_2 - s_1 = 1425{,}661$$

$$\mu_{u,2} = \mu'_{u,2} - \bar{u}^2 \qquad\qquad\qquad\qquad = 1{,}815$$
$$\mu_{u,3} = \mu'_{u,3} - 3\mu'_{u,2}\bar{u} + 2\bar{u}^3 \qquad\quad = 1{,}250$$
$$\mu_{u,4} = \mu'_{u,4} - 4\mu'_{u,3}\bar{u} + 6\mu'_{u,2}\bar{u}^2 - 3\bar{u}^4 = 16{,}456$$

Je možno voliti počátek na př. poblíž třídy s největší četností, takže se dosáhne dvojími součty nahoru a dolů menších čísel, ale výrazy pro momenty jsou zase trochu složitější.

(3,11) Opravy momentů. Tím, že při skupinovém rozdělení četností zastupuje prostřední hodnota třídního intervalu všechny hodnoty znaku dotyčného intervalu, vzniká při výpočtu momentů jistá odchylka (chyba) od momentů, které by byly počítány ze všech hodnot znaku, jak byly napozorovány nebo pro spojitou proměnnou jako funkcionální momenty

$$m_{x,k} = \frac{1}{r} \int_a^l (x - \bar{x})^k \, n(x) \, \mathrm{d}x.$$

Proto se momenty, vypočítané svrchu uvedeným postupem, opravují t. zv. Sheppardovou korekcí, takže pro opravené momenty platí rovnice

$$_0\mu_{x,2} = \mu_{x,2} - \tfrac{1}{12}h^2, \tag{22}$$

$$_0\mu_{x,3} = \mu_{x,3}, \tag{23}$$

$$_0\mu_{x,4} = \mu_{x,4} - \tfrac{1}{2}h^2\mu_{x,2} + \tfrac{7}{240}h^4 \tag{24}$$

a je-li délka třídního intervalu rovna jedné, položíme $h = 1$. Tak tedy v případě proměnné u, kde je délka intervalu zvolena za jednotku, bude na př.

$$_0\mu_{u,2} = \mu_{u,2} - \tfrac{1}{12}.$$

Poněvadž

$$\mu_{x,2} = h^2\mu_{u,2}, \text{ bude tedy } {}_0\mu_{x,2} = h^2 {}_0\mu_{u,2}.$$

(3,12) Schema výpočtu. Je účelno zachovávati při výpočtu momentů určitý pořádek v zapisování výsledků; provedeme tedy podrobný výpočet pro náš numerický příklad.

Tabulka 2.

Třídní hranice	Třídní znak	Četnost n_i	u_i	u_in_i	$u_i^2n_i$	$u_i^3n_i$	$u_i^4n_i$	$(u_i+1)^4 n_i$
22,5	30	1	—5	— 5	25	—125	625	256
37,5	45	1	—4	— 4	16	— 64	256	81
52,5	60	6	—3	—18	54	—162	486	96
67,5	75	38	—2	—76	152	—304	608	38
82,5	90	80	—1	—80	80	— 80	80	0
97,5	105	83	0	0	0	0	0	83
112,5	120	39	1	39	39	39	39	624
127,5	135	17	2	34	68	136	272	1377
142,5	150	2	3	6	18	54	162	512
157,5	165	2	4	8	32	128	512	1250
172,5	180	0	5	0	0	0	0	0
187,5	195	1	6	6	36	216	1296	2401
Součty ...		270		—183 + 93		—735 +573		
Celkem ...				— 90	520	—162	4336	6718

$$x_0 = 105, h = 15.$$

O správnosti výpočtu se přesvědčíme testem Charlierovým, neboť jsme si k tomu cíli připravili poslední sloupec tabulky.

$$4336 - 4 \times 162 + 6 \times 520 - 4 \times 90 + 270 = 6718,$$

což je provedeno podle součtu rovnic

$$(u_i + 1)^4 n_i = u_i^4 n_i + 4u_i^3 n_i + 6u_i^2 n_i + 4u_i n_i + n_i$$

pro $i = 1, 2, \ldots, l$.

Momenty pomocné proměnné u,
obecné:

$$\mu'_{u,1} = \bar{u} = -0{;}3333$$

$$\mu'_{u,2} = 1{;}9259 \qquad \bar{u}^2 = 0{,}1111$$

$$\mu'_{u,3} = -0{;}6000 \qquad -3\bar{u}\mu'_{u,2} = +1{,}9257$$

$$\mu'_{u,4} = 16{;}0593 \qquad 2\bar{u}^3 = -0{,}0741$$

$$-4\bar{u}\mu'_{u,3} = -0{,}7999$$

$$+6\bar{u}^2\mu'_{u,2} = +1{,}2838$$

$$-3\bar{u}^4 = -0{,}0370$$

kolem průměru $\bar{u}$:

$$\mu_{u,2} = 1{,}8148 \qquad \sigma_u = 1{,}3471$$

$$\mu_{u,3} = 1{,}2534 \qquad \sigma_u\mu_{u,2} = 2{,}4447$$

$$\mu_{u,4} = 16{,}5062 \qquad \mu_{u,2}{}^2 = 3{,}2935$$

$$_0\mu_{u,2} = 1{,}7315 \qquad _0\sigma_u = 1{,}3159$$

$$_0\mu_{u,3} = 1{,}2534 \qquad _0\sigma_u\, {}_0\mu_{u,2} = 2{,}2785$$

$$_0\mu_{u,4} = 15{,}6280 \qquad _0\mu_{u,2}{}^2 = 2{,}9981$$

Momenty proměnné x kolem průměru

$$\bar{x} = 100{,}00$$

$\mu_{x,2} = 408{,}33$	$\sigma_x = 20{,}21$	$_0\mu_{x,2} = 389{,}59$	$_0\sigma_x = 19{,}74$
$\mu_{x,3} = 4\,230$	$\alpha_{x,3} = 0{,}51$	$_0\mu_{x,3} = 4\,230$	$_0\alpha_{x,3} = 0{,}55$
$\mu_{x,4} = 835\,600$	$\alpha_{x,4} = 5{,}01$	$_0\mu_{x,4} = 791\,200$	$_0\alpha_{x,4} = 5{,}21$

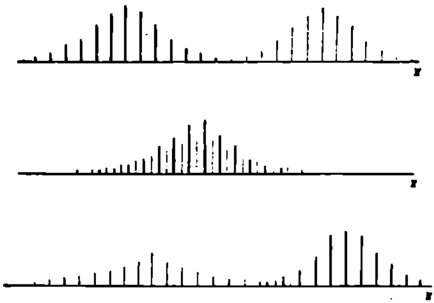
O rozdělení četností nabýváme pomocí charakteristik jisté přibližné představy. Tak průměr je charakteristikou polohy souboru na stupnici hodnot znaku, směrodatná odchylka nebo její čtverec je výrazem rozptylu, šikmost nebo kosost udává míru nesouměrnosti rozdělení četností. Na obr. 5 jsou znázorněna dvě symetrická rozdělení četnosti

a) s týmž rozptylem, ale různou polohou,

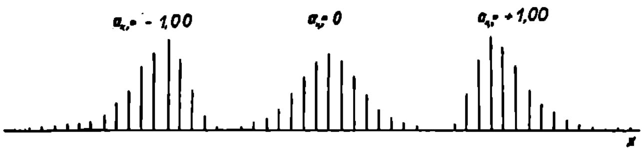
b) s různými rozptyly a touž polohou,

c) s různými rozptyly a různou polohou.

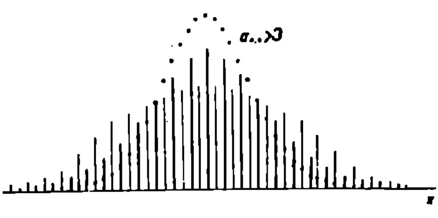
Obr. 6 ukazuje souměrné rozdělení jakož i zápornou a kladnou šikmost. Konečně obr. 7 osvětluje kladnou a zápornou špičatost (exces), která se měří srovnáváním s nor-



Obr. 5. Dvě základní charakteristiky rozdělení četností. (Poloha — Rozptyl.)



Obr. 6. Třetí charakteristika rozdělení četnosti. (Šikmost pro tři zvláštní hodnoty.)



Obr. 7. Čtvrtá charakteristika rozdělení četností. (Exces kladný a záporný přirovnán k normálnímu $\alpha_{x,4} = 3$.)

mální křivkou Laplace-Gaussovou (viz 82), jejíž špičatost je $\alpha_{x,4} - 3 = 0$ (úsečky silněji vytažené).

(3,13) Přesnost průměru a směrodatné odchylky. Otázku po přesnosti, které docílíme pro průměr a směrodatnou odchylku, můžeme zodpověděti tak, že stačí obyčejně uváděti to desetinné místo, na které jsou měřeny hodnoty znaku.

Často se také postupuje dále podle směrnice, že počet desetinných míst v numerických hodnotách charakteristik se řídí podle směrodatné odchylky. Vyjádří se tedy směrodatná odchylka třemi, resp. dvěma významnými číslicemi a charakteristiku uvedeme na týž počet desetinných míst jako směrodatnou odchylku nebo o jedno místo méně. Při tom se doporučuje dbáti, aby zaokrouhlení vzniklé vynecháním desetinných míst nepřekročilo 0,1 směrodatné odchylky.

Jinak bývá často pro usnadnění práce třeba zaokrouhlovat hodnoty znaku. Pak je nutno odhadovat největší možné hranice účinku zaokrouhlení na výsledek. Takový odhad lze prováděti vhodně na př. takto: Nahradíme-li přesné číslo x číslem x', je absolutní chyba $\vartheta = x - x'$, takže je přesné číslo $x = x' + \vartheta$, což můžeme také psáti $x = x'\left(1 + \frac{\vartheta}{x'}\right)$. Potom zlomek $\frac{\vartheta}{x'}$ se označuje symbolem ε nebo e a nazývá se relativní chybou čísla x'. Přesné číslo je pak vyjádřeno $x = x'(1 + \varepsilon)$. Máme nyní dvě čísla x' a y', jejichž relativní chyby jsou ε resp. e, pak dostáváme relativní chybu součtu a rozdílu jejich

$$x \pm y = x'(1 + \varepsilon) \pm y'(1 + e) = x' \pm y' + x'\varepsilon \pm y'e = \\ = (x' \pm y')\left(1 + \frac{x'\varepsilon \pm y'e}{x' \pm y'}\right). \qquad (25)$$

Můžeme-li považovati relativní chyby obou členů za stejné, dostaneme jednoduše $x \pm y = (x' \pm y')(1 + \varepsilon)$, takže relativní chyba součtu nebo rozdílu se rovná relativní chybě jeho členů. Tento výsledek a ovšem také rovnici (25) lze snadno rozšířiti na libovolný počet členů.

Jsou-li chyby ε a e malé, dostaneme provedením naznačených operací přibližné vyjádření chyby

součinu $\quad x \cdot y = x' \cdot y' (1 + \varepsilon)(1 + e) \doteq x'y' (1 + \varepsilon + e)$

mocniny $\quad x^m = [x' (1 + \varepsilon)]^m \doteq x'^m (1 + m\varepsilon)$

podílu $\quad \dfrac{x}{y} = \dfrac{x'}{y'} \dfrac{1 + \varepsilon}{1 + e} \doteq \dfrac{x'}{y'} (1 + \varepsilon - e).$

Uvažujeme-li o chybě průměru vzniklé zaokrouhlením hodnot znaku x_i, které jsou nahrazeny hodnotami x'_i, při čemž může býti ε největší relativní chyba ze zaokrouhlení a tu vezmeme pro všechny hodnoty, pak $x'_i (1 - \varepsilon) \leqq x_i \leqq \leqq x'_i (1 + \varepsilon)$. Utvoříme-li ze všech tří hodnot průměry vidíme, že $\bar{x}' (1 - \varepsilon) \leqq \bar{x} \leqq \bar{x}' (1 + \varepsilon)$, čili hranice chyb průměru zaokrouhlených čísel nepřekročí největší možnou chybu vzniklou zaokrouhlením jednotlivých hodnot znaku. Je zajímavo, že vypočítáme-li si pro náš příklad průměr z hodnot seřazených na str. 19, dostaneme $\bar{x} = 92{,}5$, kdežto průměr z hodnot zastoupených třídními znaky v tabulce č. 2 je 100,0. Poněvadž největší možná chyba je 7,5 vidíme, že je tato hodnota průměru právě na hranici možných chyb vzniklých seskupením do tříd, ač velmi často se chyby značně kompensují.

(3,14) Přehled charakteristik. Vzhledem k tomu, že kromě momentových charakteristik se užívá často k některým účelům také jiných, seznámíme se s těmi nejdůležitějšími.

1. Charakteristiky polohy. a) Aritmetický průměr je nejrozšířenější charakteristikou polohy nebo také mírou ústřední tendence. Jeho prvá podstatná vlastnost je, že součet odchylek hodnot znaku u všech jednotek souboru od aritmetického průměru se rovná nule. $\mu_{x,1} = 0$. Druhá vlastnost je, že součet čtverců těchto odchylek je minimem. Známe-li průměry jednotlivých částí souboru, násobíme je

jejich rozsahem a dělíme rozsahem celého souboru, abychom dostali průměr souboru. Zvětší-li se nebo zmenší-li se všechny hodnoty znaku o konstantu, zvětší nebo zmenší se o ni také průměr.

$$\mu'_{x\pm a,1} = \frac{1}{r}(x_1 \pm a + x_2 \pm a + \ldots + x_r \pm a) =$$

$$= \frac{1}{r}\sum_{i=1}^{r} x_i \pm a = \mu'_{x,1} \pm a.$$

Násobí-li se hodnoty znaku konstantou, je průměr násoben touž konstantou

$$\mu'_{ax,1} = \frac{1}{r}(ax_1 + ax_2 + \ldots + ax_r) = a\frac{1}{r}\sum_{i=1}^{r} x_i = a\mu'_{x,1}.$$

b) Poznali jsme již také medián, jehož se zvláště užívá tehdy, když jsou hodnoty znaku nesnadno měřitelné, ale lze je snadno aspoň seřadit podle velikosti. Nezávisí na krajových hodnotách znaku, které mají na př. značný vliv na průměr aritmetický. Můžeme jej určit i když jsou intervaly nekonečně velké, takže znemožňují výpočet průměru. Řekneme-li, že medián je velký, víme, že polovina pozorovaná je jistě velká, kdežto u aritmetického průměru nemůžeme ničeho říci o celém množství pozorování, neboť jeho vysoká hodnota může býti způsobena několika isolovanými případy. Proto se často dává mediánu přednost před průměrem na př. při statistice mezd. Stanovení mediánu ze skupinového rozdělení četností provedeme pro náš příklad. Ze sloupce kumulativní četnosti vidíme, že prostřední dva členy (při sudém rozsahu) jsou v intervalu 97,5 až 112,5 mezi 135. a 136. pozorováním čili mezi 9. a 10. prvkem intervalu. Budeme předpokládati, že hodnoty proměnné jsou v intervalu stejnoměrně rozloženy, takže délku intervalu s četností 83 rozdělíme úměrou a dostaneme pro devátou hodnotu

$$97{,}5 + 15\,\tfrac{9}{83} = 99{,}13 \text{ a pro desátou } 99{,}31.$$

Mezi těmito dvěma čísly leží medián, za nějž vezmeme jejich střed, takže $\tilde{x} = 99{,}22$.

Můžeme psáti obecně výraz pro medián, který má býti v intervalu $x_m - \frac{1}{2}h$, $x_m + \frac{1}{2}h$ při lichém rozsahu souboru r

$$x = x_m - \frac{h}{2} + \frac{h}{n_m}\left\{\frac{r+1}{2} - s_{m-1}\right\} \tag{26}$$

a při sudém r je mezi dvěma výrazy

$$\begin{gathered} x_m - \frac{h}{2} + \frac{h}{n_m}\left\{\frac{r}{2} - s_{m-1}\right\} < \tilde{x} < x_m - \\ - \frac{h}{2} + \frac{h}{n_m}\left\{\frac{r}{2} + 1 - s_{m-1}\right\} \end{gathered} \tag{27}$$

jak čtenář snadno sám nahlédne.

c) Modus je nejčetnější hodnota znaku. Z rozdělení četností, které zachovává jednotlivé pozorované hodnoty proměnné, se určí modus jednoduše jako hodnota, jíž přísluší největší četnost n_a. Obtíž vzniká, máme-li určiti tuto hodnotu pro skupinové rozdělení četností, které se nejčastěji v praksi vyskytuje. Udávati střed intervalu s největší třídní četností by mělo malý význam, neboť ten závisí na volbě stupnice pro třídní intervaly. Proto se určuje obyčejně modus přibližně jako hodnota, která přísluší maximu křivky proložené co nejtěsněji skutečným rozdělením četností. V případech mírně nesouměrných rozdělení četností si pomáháme proložením paraboly druhého stupně $y = c_2x^2 + c_1x + c_0$ bodem znázorňujícím největší četnost přiřazenou jejímu třídnímu znaku a obdobným bodem sousedním s každé strany. Položíme-li počátek do třídního znaku třídy s největší četností, dostaneme k určení konstant paraboly tři rovnice [10] [11]

$$\begin{aligned} n_{a-h} &= c_2h^2 - c_1h + c_0 \\ n_a &= c_0 \\ n_{a+h} &= c_2h^2 + c_1h + c_0, \end{aligned}$$

takže řešením dostáváme

$$n_{a-h} + n_{a+h} - 2n_a = 2c_2h^2$$
$$n_{a+h} - n_{a-h} = 2c_1h$$

Modus, jakožto úsečka vrcholu paraboly je stanoven podmínkou, že prvá derivace se rovná nule $y' = 2c_2x + c_1 = 0$ čili $x' = -\frac{c_1}{2c_2}$, takže dosazením z rovnic dostáváme

$$x' = -\frac{h}{2}\,\frac{n_{a+h} - n_{a-h}}{n_{a+h} - 2n_a + n_{a-h}},$$

což je tedy poloha modu měřená od počátku v třídním znaku intervalu s maximální četností, takže modus $\hat{x}$ bude pak vyjádřen vztahem

$$\hat{x} = x_a - \frac{h}{2}\,\frac{n_{a+h} - n_{a-h}}{n_{a+h} - 2n_a + n_{a-h}}. \tag{28}$$

Pro náš numerický příklad skupinového rozdělení četností z toho vyplývá toto vyjádření $x_a = 105$, $h = 15$, $n_{a+h} = 39$, $n_a = 83$, $n_{a-h} = 80$, takže modus

$$\hat{x} = 98{,}46.$$

V dokonale souměrném rozdělení četností spadá průměr $\bar{x}$, medián $\tilde{x}$ i modus $\hat{x}$ do jedné hodnoty, která je středem souměrnosti.

V nesouměrném rozdělení četností se tyto tři charakteristiky od sebe liší a je-li rozdělení mírně nesouměrné (obr. 3 nebo 6), ukazuje zkušenost, že medián leží přibližně ve třetině vzdálenosti od průměru k modu, čili osvědčuje se s překvapující přiléhavostí přibližný vztah

$$\bar{x} - \hat{x} = 3(\bar{x} - \tilde{x}), \tag{29}$$

z něhož je také možno přibližně modus určit, známe-li již průměr a medián. Zvláště u nesouměrných rozdělení četností mívá modus velkou důležitost jako hodnota znaku, která se v souboru nejčastěji vyskytuje, takže se někdy nazývá typická hodnota.

V některých odvětvích praktické statistiky mají zvláštní oprávnění ještě jiné charakteristiky polohy.

d) Geometrický průměr je r-tá odmocnina ze součinu všech r pozorovaných hodnot znaku vyšetřovaného souboru. Při skupinovém rozdělení četností, kde je n_i hodnot proměnné zastoupeno třídním znakem x_i, bude při l třídách definován geometrický průměr g rovnicí

$$g = (x_1^{n_1} \cdot x_2^{n_2} \dots x_l^{n_l})^{\frac{1}{r}}.$$

Logaritmováním dostaneme

$$\log g = \frac{1}{r} \sum_{i=1}^{l} n_i \log x_i,$$

z čehož patrno, že logaritmus geometrického průměru je aritmetickým průměrem logaritmů jednotlivých hodnot znaku, takže jeho výpočet se tím převádí na metody zavedené již pro aritmetický průměr. Spočívá na všech hodnotách znaku jako aritmetický průměr, ale je na krajové hodnoty méně citlivý. Častého užití se dostalo geometrickému průměru ve statistice cenové a při konstrukci čísel indexních. Pro danou řadu čísel je vždy menší než její aritmetický průměr. Jednoduchý důkaz pro dvě čísla x_1, x_2 od sebe různá vyplývá takto $(\sqrt{x_1} - \sqrt{x_2})^2 > 0$ tedy

$$x_1 + x_2 - 2\sqrt{x_1 x_2} > 0 \text{ a tudíž } \tfrac{1}{2}(x_1 + x_2) > \sqrt{x_1 x_2}.$$

e) Podáme ještě definici harmonického průměru, jehož používá statistická prakse poměrně zřídka. Převratná hodnota $\frac{1}{\gamma}$ harmonického průměru je aritmetickým průměrem převratných hodnot znaku

$$\frac{1}{\gamma} = \frac{1}{r} \sum_{i=1}^{l} \frac{1}{x_i} n_i.$$

Úvahy o harmonickém průměru lze tedy vhodně převésti na úvahy o aritmetickém průměru.

2. Charakteristiky rozptylu. a) Abychom vystihli způsob rozdělení prvků v mezích celkového variačního rozpětí, užíváme nejčastěji směrodatné odchylky $\sigma_x = \sqrt{\mu_{x,2}}$. Čím je při téže jednotce měření hodnot znaku σ_x menší, tím jsou hodnoty proměnné a tedy prvky souboru těsněji seskupeny kolem aritmetického průměru.

b) Pro srovnávání někdy užíváme za účelem eliminování vlivu jednotky měření poměru směrodatné odchylky k průměru, tedy míry relativního rozptylu vyjadřované v procentech. Nazývá se koeficient variační a je dán výrazem

$$v = 100 \frac{\sigma_x}{\bar{x}}. \tag{30}$$

c) Utvoříme-li průměr ze všech odchylek hodnot pozorovaného znaku od některé charakteristiky polohy, nepřihlížejíce při tom ke znaménku odchylek, dostáváme t. zv. průměrnou odchylku ϑ. Přirozeným východiskem je tu medián, takže je dána rovnicí

$$\vartheta = \frac{1}{r} \sum_{i=1}^{l} | x_i - \tilde{x} | n_i. \tag{31}$$

Lze dokázati, že průměrná odchylka je nejmenší, měří-li se odchylky od mediánu [4]. Podle empirického pravidla se obyčejně průměrná odchylka velmi blíží $\frac{4}{5}$ směrodatné odchylky pro souměrné nebo jen mírně nesouměrné rozdělení četností.

d) Také vzdálenost obou kvartilů může sloužiti za míru rozptylu. Rozdíl mezi mediánem a dolním kvartilem $\tilde{x} - \tilde{x}_1$, který se při souměrném rozdělení četností rovná rozdílu mezi horním kvartilem a mediánem $\tilde{x}_2 - \tilde{x}$, dává zřejmě možnost posouditi soustřeďování prvků kolem mediánu. Poněvadž pozorovaná rozdělení četnosti nejsou přesně souměrná, volí se za míru rozptylu poloviční součet obou uvedených hodnot, který se rovná rozdílu mezi kvartily

$$\tfrac{1}{2}(\tilde{x} - \tilde{x}_1 + \tilde{x}_2 - \tilde{x}) = \tfrac{1}{2}(\tilde{x}_2 - \tilde{x}_1) = \varrho \qquad (32)$$

a nazývá se **kvartilová odchylka**.

e) Výpočet směrodatné odchylky nebo střední odchylky spočívá na odchylkách hodnot znaku od průměru nebo od mediánu. Tím se zavádí do počtu prvek, který není nevyhnutelně nutný, neboť rozptyl členů řady statistické je možno měřiti také jejich absolutními diferencemi bez prostředkování jiné charakteristiky. Tak vznikla míra zvaná **střední diference Δ**.

Předpokládejme, že jsme ze všech hodnot znaku pozorovaných na prvcích souboru s rozsahem r vytvořili řadu neklesající

$$x_1 \leqq x_2 \leqq \ldots \leqq x_r.$$

Utvoříme pak diference

$$x_r - x_1,\ x_r - x_2,\ x_r - x_3,\ \ldots,\ x_r - x_{r-3},\ x_r - x_{r-2},\ x_{r-1} - x_{r-1}$$

$$x_{r-1} - x_1,\ x_{r-1} - x_2,\ x_{r-1} - x_3,\ \ldots,\ x_{r-1} - x_{r-3},\ x_{r-1} - x_{r-2},$$

. . . .

$$x_4 - x_1,\ x_4 - x_2,\ x_4 - x_3,$$

$$x_3 - x_1,\ x_3 - x_2$$

$$x_2 - x_1.$$

Všechny tyto diference jsou kladné nebo nula a jejich počet je zřejmě

$$(r - 1) + (r - 2) + \ldots + 3 + 2 + 1 = \frac{r\,(r - 1)}{2}.$$

Sečteme-li nyní všechny diference, zruší se část sčítanců, jak se snadno přesvědčíme, a dostaneme celkový součet S

$$S = (r - 1)\,(x_r - x_1) + (r - 3)\,(x_{r-1} - x_2) +$$
$$+ (r - 5)\,(x_{r-2} - x_3) + \ldots$$

Poslední člen pro sudé $r = 2k$ bude $1 . (x_{k+1} - x_k)$ a pro liché $r = 2k + 1$ je $2 . (x_{k+2} - x_k)$.

Střední diference tedy bude

$$\Delta = \frac{S}{\frac{1}{2}r(r-1)} = \frac{2}{r(r-1)}\sum_{i=1}^{k}(r+1-2i)(x_{r-i+1}-x_i). \quad (33)$$

Výpočet součtu se provede sestavením tabulky

$$\begin{array}{ll} r-1, & x_r - x_1, \\ r-3, & x_{r-1} - x_2, \\ r-5, & x_{r-2} - x_3, \\ \ldots & \ldots \end{array}$$

která se k-tým řádkem ukončí, kde k se určí buď ze vztahu $r = 2k$, nebo $r = 2k + 1$. Součinitel $r + 1 - 2i$ pak musí míti hodnotu buď 1 nebo 2.

f) Nejprostším odhadem rozptylu rozdělení četností je variační rozpětí; obsah jeho informace je však poměrně malý.

3. Charakteristiky šikmosti čili kososti. a) O momentové charakteristice šikmosti $\alpha_{x,3}$ jsme se již dostatečně zmínili (str. 23).

b) Míru šikmosti čili nesouměrnosti, nezávislou na jednotce, v níž je měřen pozorovaný znak, můžeme také sestrojiti, stanovíme-li poměr mezi rozdílem odchylek kvartilů od mediánu a vzdáleností obou kvartilů

$$\frac{(\tilde{x}_2 - \tilde{x}) - (\tilde{x} - \tilde{x}_1)}{\tilde{x}_2 - \tilde{x}_1} = \frac{\tilde{x}_1 + \tilde{x}_2 - 2\tilde{x}}{x_2 - \tilde{x}_1} = \tau,$$

při čemž $-1 \leqq \tau \leqq +1$.

Tato míra šikmosti charakterisuje spíše tvar rozdělení četností mezi oběma kvartily a nepřihlíží náležitě k významu hodnot znaku, ležících vně. Může se tudíž v některém případě podstatně lišiti od momentové charakteristiky $\alpha_{x,3}$. Tím se také vysvětluje v případě zvoleného rozdělení četností (41), že hodnota $\tau = -0{,}02$ nasvědčuje souměrnému rozdělení mezi oběma kvartily, kdežto momentová charakteristika udává malou kladnou šikmost.

c) Místo této míry zavedl **Pearson** výraz $\tau_1 = \dfrac{\bar{x} - \hat{x}}{\sigma_x}$.

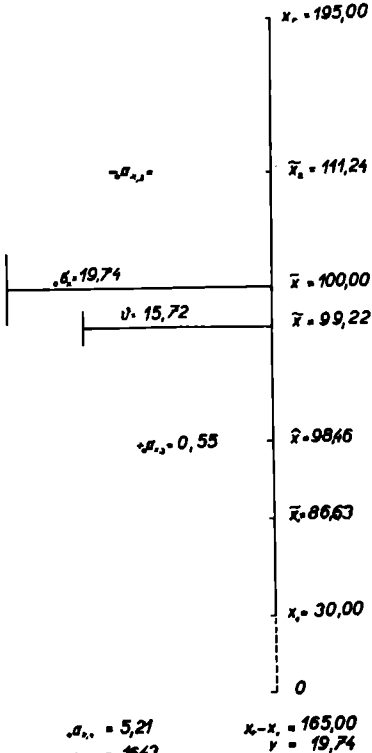
V případě souměrného rozdělení četnosti je $\bar{x} = \hat{x}$ a tudíž $\tau_1 = 0$.

Poněvadž se modus elementárním způsobem nesnadno zjišťuje, nahražuje se někdy čitatel zlomku přibližným $3(\bar{x} - \tilde{x})$ podle rovnice (29), takže touto obměnou dostáváme

$$\tau_2 = \frac{3(\bar{x} - \tilde{x})}{\sigma_x}.$$

Jako je účelno zachovávati jistý pořádek, který jsme zavedli při výpočtu charakteristik a zapisování výsledků, tak se jeví také přehledným stálé schema, na které si zvykneme pro charakterisování pozorovaného souboru. Pro náš numerický příklad je sestavíme a vyplníme.

Do tohoto schematu ovšem zapíšeme jen ty charakteristiky, které jsme potřebovali a tedy počítali. Neznáme-li při skupinovém rozdělení četností největší pozorovanou hodnotu x_r, zapíšeme si třídní znak poslední třídy x_l; x_1 značí buď nejmenší pozorovanou hodnotu znaku, nebo třídní



Obr. 8. Přehledné schema charakteristik.

znak první třídy. $\vartheta \doteq \frac{4}{5}\sigma_x$. Šikmost zapíšeme nahoru nebo dolů, podle toho, je-li záporná či kladná.

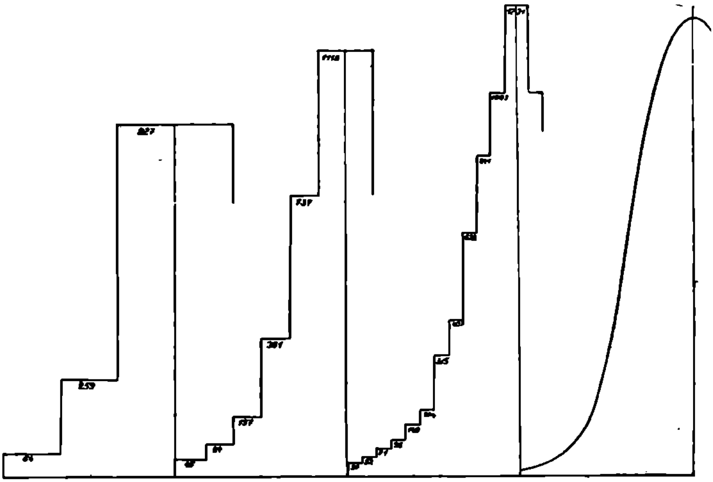
(3,15) Tři druhy řad. 1. Rozdělení četností se také nazývá statistickou řadou věcnou, která podává roztřídění pozorovaného souboru podle hodnot nějakého znaku, bez ohledu na čas nebo prostor. Vedle těchto řad se rozeznávají

2. řady časové, v nichž jsou jednotlivé hodnoty nebo četnosti uspořádány podle souslednosti časové. Mohou se také nazývati chronologické nebo historické. Časová stupnice, podle níž je řada uspořádána, je dána jednotkou, nejčastěji rok, nebo měsíc, týden, den. (Na př. hodnota dovozu nebo vývozu za každý měsíc.) Znázorňují se chronologickým diagramem.

3. řady místní, kde číselné údaje o jevu pozorovaném v určitém okamžiku jsou uspořádány podle místní příslušnosti (do obce, okresu, země, ...). Znázorňují se obyčejně kartogramem.

(3,16) Od skupinového rozdělení četností ke spojité křivce. Viděli jsme, že při volbě délky intervalu třídního máme značnou volnost. Tvar rozdělení četností pak do jisté míry závisí na této volbě. Budeme sledovati třídění spojitého znaku na pozorovaném souboru dosti velikého rozsahu $r = 10\,000$, který má při délce intervalu $h = 2$ rozdělení četnosti ve druhém sloupci tabulky 3. Provedeme-li přeřazení do tříd dvojnásobné délky intervalu a čtyřnásobné délky intervalu, dostáváme sloupce (3) a (4) tabulky 3.

Roztříděním prvků souboru a získáním rozdělení četností jsme zjistili jak jsou rozděleny prvky vzhledem k vyšetřovanému znaku v daném množství 10 000. Máme v tom však také odpověď na otázku: Vezmeme-li náhodně 10 000 předmětů druhu definovaného statistickou jednotkou a rozdělíme je do skupin, jak často vezmeme prvek do každé z těchto skupin? — Znázorníme-li si hrubé rozdělení (sloupce 4) histogramem (obr. 9), který vystupuje po stupních nahoru a pak zase sestupuje (což již pro úsporu místa a přehlednost není vykresleno), vidíme, že počet prvků v třídách blízkých průměru je největší a k oběma krajům klesá. Vzhledem k tomu očekáváme, že rozdělíme-li některou ze tříd na dvě,



Obr. 9. Vznik hladšího rozdělení četnosti zkracováním třídního intervalu.

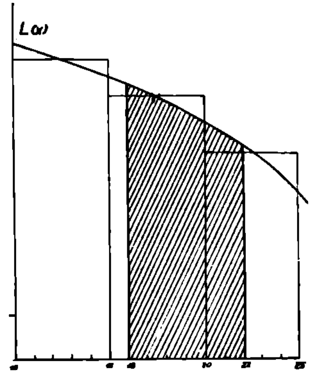
bude část bližší průměru obsahovati více prvků než část vzdálenější. Tak rozdělíme-li interval 162, 164, 166, 168 na dva 162, 164 a 166, 168, vidíme, že jeho četnost 1036 se rozpadne na četnost intervalu bližšího průměru 722 a na četnost 314 intervalu 162, 164 vzdálenějšího od průměru. Podobně je tomu v případech ostatních intervalů. Stupně jsou užší a jejich počet vzrostl, jak je patrno ze zobrazení levé části rozdělení četností histogramem, kde je tedy četnost znázorněna plochou příslušného pravoúhelníka. Opaku-

Tabulka 3.

Výška v cm	Četnost pro interval		
	h	2h	4h
(1)	(2)	(3)	(4)
154—	37		
156—	52	89	
158—	71		
160—	96	167	256
162—	140		
164—	174	314	
166—	315		
168—	407	722	1036
170—	632		
172—	841	1473	
174—	1003		
176—	1232	2235	3708
178—	1232		
180—	1003	2235	
182—	841		
184—	632	1473	3708
186—	407		
188—	315	722	
190—	174		
192—	140	314	1036
194—	96		
196—	71	167	
198—	52		
200—	37	89	256

jeme-li tento postup dále (sloupec 2), stupně se zúžují a průběh je hladší. Kdyby byl dostatečně velký rozsah souboru a hodnoty znaku pozorovány na dosti velký počet desetinných míst, zmizely by stupně a při nekonečně velkém rozsahu souboru dostali bychom spojitou křivku.

Je tedy jasno, že stupně jsou něčím umělým, neboť vznikají tím, že musíme volit do značné míry libovolně hranice tříd vzhledem k tomu, že znak je měřen různými měrami ať délky nebo váhy či věku, atd. Poněvadž tvoření tříd je libovolné, býváme nuceni rozdělení četností nahradit něčím, co nesouvisí s uspořádáním podléhajícím této libovůli. Můžeme proložit spojitou křivku vrcholky polygonu četnosti nebo jí nahradit také histogram. Spojitá křivka je nezávislá na třídách a proto je obecnější povahy než hrubý polygon. Mimochodem se zmíníme o možnosti užití spojité křivky četnosti, známe-li jen tabulku skupinového rozdělení četností v určitých třídních intervalech a potřebujeme je znáti v třídách utvořených jinak. Dostaneme na př. při sčítání lidu tabulku četností $L(x)$ jen pro pětileté nebo desetileté věkové třídy a potřebujeme k určitému účelu znáti počet osob věku 16—22 let. Z původního materiálu to již není možné, nebo by to bylo při rozsáhlosti souboru příliš nákladné. Úlohu rozřešíme potom tak, že histogram pro pětileté třídní intervaly nahradíme přiléhající křivkou, která uzavírá s osou x plochu stejnou jako s ní uzavírá obrys histogramu (obraz 10) a změříme plochu odpovídající uvedenému intervalu. Druhý je případ užití křivky četností je-li rozsah



Obr. 10. Stanovení četnosti pro změněný interval.

souboru malý, tedy pozorované četnosti třídní malé a vyznačující se nepravidelností. Proložíme tedy křivku, abychom odstranili nahodilé výkyvy a dostali celkový průběh, který by se přibližoval průběhu spojitému, jejž bychom dostali, kdyby rozsah souboru rostl nade všechny meze. Při prokládání křivky pozorovanými hodnotami volnou rukou je třeba velké opatrnosti, neboť může dáti někdy velmi nesprávný odhad ideálního výsledku. Proto se k tomu cíli užívá zvláštních metod matematických.

(4,1) Vznik hlavních typů rozdělení četností.

Typy křivek, které se uplatňují ve statistice, vyvozujeme dvojí cestou: jednak dedukcí pomocí kombinatorických úvah, zabýváme-li se problémy ryzí náhody, jako házení mincí, kostek, ... jednak indukcí, studujeme-li tvary křivek, které se obecně vyskytují při zkoumání souborů velkých rozsahů z různých oborů statistiky.

Obě cesty se doplňují a pomocí modelů sestrojených úvahami kombinatoriky [10], [11] osvětlujeme výsledky pozorování, u nichž můžeme souditi na analogické podmínky vzniku dotyčného jevu (poměr počtu narozených chlapců a děvčat), pro který však opakované provedení nějakého statistického experimentu je vyloučeno.

Uděláme si nejprve představu o spojení mezi náhodným jevem a spojitými křivkami.

Zvolíme si za pokus házení mincí a pozorovaný znak bude počet rubů a líců, které se objeví. Užijeme pracovní hypotésy, že každá mince, kterou házíme, je správná, čili vykazuje příslušnou geometrickou a mechanickou symetrii. Není tudíž důvodu k tomu, aby se rub objevoval u téže mince častěji než líc.

Zaznamenejme si všechny případy, které mohou nastati, když házíme třemi stejnými mincemi. Sestavíme si je podle počtu rubů; označíme-li písmenou „L“ líc, „R“ rub, pak vidíme:

R R R ..... 1 případ: tři ruby

R R L .....
R L R ..... } 3 případy: dva ruby
L R R .....

R L L .....
L R L ..... } 3 případy: jeden rub
L L R .....

L L L ..... 1 případ: žádný rub

Můžeme tedy sestaviti tabulku:

počet rubů	0	1	2	3	celkem
počet případů	1	3	3	1	8

Kdybychom takto postupovali dále pro čtyři mince, dostali bychom počet případů, čili četnost, vyjádřenou řadou čísel 1, 4, 6, 4, 1 a obecně jak známo řadou binomickou $(1+1)^l$.

Tato čísla tvoří pro $l = 0, 1, 2, \dots$ známý Pascalův trojúhelník [10, 11]. Když bychom prováděli skutečné pokusy, dostaneme ovšem vždy něco jiného. Tak na př. pro 14 mincí dostáváme naší úvahou řadu četností a) 1, 14, 91, 364, 1001, 2002, 3003, 3432, 3003, 2002, 1001, 364, 91, 14, 1 a v dříve již uvedeném případě, kde jsme vykonali 201 vrhů se 14 mincemi, dostáváme tabulku:

b)

počet rubů...	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	celkem
počet případů	0	0	1	3	17	23	35	49	35	20	9	8	—	1	0	201

Porovnání můžeme ovšem provésti pomocí relativních četností, čili převedením souboru na rozsah jednotkový. V našem případě pro $l = 14$ je součet absolutních četností $(1+1)^{14} = 2^{14} = 16\,384$, takže tímto číslem dělíme každý člen řady a); četnosti pozorované a sestavené právě v tabulce pak dělíme celkovým rozsahem 201.

Dostáváme tak v procentech tyto dvě řady relativních četností

	0	1	2	3	4	5	6	7
a)	0,0001	,0008	,0056	,0222	,0611	,1222	,1833	,2094
	8	9	10	11	12	13	14	
	,1833	,1222	,0611	,0222	,0056	,0008	,0001	

b)	0	1	2	3	4	5	6	7
	0	0	0,0050	,0149	,0846	,1144	,1741	,2438
	8	9	10	11	12	13	14	
	,1741	,0995	,0448	,0398	0	,0050	0	

řada b) se v jistých mezích mění, takže bychom při druhém pokusu o témž počtu vrhů dostali četnosti odlišné.

O mezích těchto odchylek budeme uvažovati později (str. 110).

Úloha:

Jest stanoviti aritmetický průměr a směrodatnou odchylku rozdělení relativních četností daného binomickou řadou

$$(1 + 1)^l \frac{1}{2^l}:$$

hodnoty znaku: $0 \quad 1 \quad 2 \ldots\ldots\ldots\ldots l;$

relat. četnosti: $1\,\frac{1}{2^l} \quad \binom{l}{1}\frac{1}{2^l} \quad \binom{l}{2}\frac{1}{2^l}\cdots\binom{l}{l}\frac{1}{2^l}.$

Násobíme-li hodnoty znaku příslušnými relativními četnostmi a sečteme, dostáváme po vytknutí l

$$l\left\{1 + \binom{l-1}{1} + \binom{l-1}{2} + \ldots + \binom{l-1}{l-1}\right\}\frac{1}{2^l} =$$
$$= l\{1 + 1\}^{l-1}\frac{1}{2^l} = l\,\frac{1}{2},$$

tedy
$$\bar{x} = \frac{l}{2}.$$

Rozptyl stanovíme, vynásobíme-li čtverce hodnot znaku příslušnými relativními četnostmi a od součtu odečteme $\bar{x}^2$.

$$\left\{1 \cdot \binom{l}{1} + 4\binom{l}{2} + 9\binom{l}{3} + \ldots + l^2\binom{l}{l}\right\}\frac{1}{2^l} - \left(\frac{l}{2}\right)^2 =$$

$$= \frac{l}{2}\left\{1 + 2\binom{l-1}{1} + 3\binom{l-1}{2} + 4\binom{l-1}{3} + \ldots + \right.$$

$$\left. + l\binom{l-1}{l-1}\right\}\frac{1}{2^{l-1}} - \left(\frac{l}{2}\right)^2 =$$

$$= \frac{l}{2}\left\{1 + \binom{l-1}{1} + \binom{l-1}{2} + \binom{l-1}{3} + \ldots + \binom{l-1}{l-1} + \right.$$

$$\left. + \binom{l-1}{1} + 2\binom{l-1}{2} + 3\binom{l-1}{3} + \ldots + (l-1)\binom{l-1}{l-1}\right\} \times$$

$$\times \frac{1}{2^{l-1}} - \left(\frac{l}{2}\right)^2 =$$

$$= \frac{l}{2}\left\{\left(\frac{1}{2} + \frac{1}{2}\right)^{l-1} + \frac{1}{2}(l-1)\left(\frac{1}{2} + \frac{1}{2}\right)^{l-2}\right\} - \left(\frac{l}{2}\right)^2 =$$

$$\frac{l}{2}\left\{1 + \frac{l-1}{2}\right\} - \left(\frac{l}{2}\right)^2 = \frac{l}{2} - \frac{l}{4} = \frac{l}{4}.$$

Je tudíž

$$\sigma_x^2 = \tfrac{1}{4}\,l$$

Zabývejme se nyní blíže rozdělením četností v případě l mincí a zkoumejme, jaké dostaneme rozdělení, jestliže l se stále zvětšuje.

Relativní četnosti v případě obecném l mincí jsou tedy vyjádřeny jednotlivými členy řady

$$\frac{1}{2^l}\left\{1 + \binom{l}{1} + \binom{l}{2} + \ldots + \binom{l}{l}\right\},$$

takže četnost x rubů čili příznivých výsledků, je

$$\frac{1}{2^l}\binom{l}{x} = \frac{1}{2^l}\,\frac{l!}{x!\,(l-x)!}.$$

Četnost $x + 1$ rubů je dána následujícím členem, a dostáváme ji z předchozího výrazu, násobíme-li jej zlomkem $\dfrac{l-x}{x+1}$. Pokud bude $l - x > x + 1$ čili $x < \dfrac{l-1}{2}$, bude následující četnost větší než předcházející. Učiníme pro zjednodušení další úvahy předpoklad, $l = 2\nu$. Při sudém l je nejčetnějším případ ν příznivých výsledků (viděli jsme na příklad, že pro $l = 4$, je největší četnost pro 2 ruby); jeho relativní četnost je dána výrazem

$$y_0 = \frac{1}{2^{2\nu}} \frac{(2\nu)!}{\nu!\,\nu!}.$$

To by byla v grafickém znázornění největší pořadnice, od níž se svažuje mnohoúhelník relativních četností na obě strany souměrně. Vezměme tedy v úvahu relativní četnost, která přísluší $\nu + x$ příznivým výsledkům, která je

$$y_x = \frac{1}{2^{2\nu}} \frac{(2\nu)!}{(\nu + x)!\,(\nu - x)!}$$

a utvoříme podíl

$$\frac{y_x}{y_0} = \frac{\nu(\nu - 1)\ldots(\nu - x + 1)}{(\nu + 1)(\nu + 2)\ldots(\nu + x)},$$

který můžeme dělením čitatele i jmenovatele ν^x uvésti na tvar

$$\frac{y_x}{y_0} = \frac{\left(1 - \frac{1}{\nu}\right)\left(1 - \frac{2}{\nu}\right)\cdots\left(1 - \frac{x-1}{\nu}\right)}{\left(1 + \frac{1}{\nu}\right)\left(1 + \frac{2}{\nu}\right)\cdots\left(1 + \frac{x-1}{\nu}\right)\left(1 + \frac{x}{\nu}\right)} \tag{34}$$

Stanovíme přibližnou hodnotu tohoto zlomku za předpokladu, že ν je veliké u porovnání s x a to tak, že můžeme zanedbati $\left(\frac{x}{\nu}\right)^2$ u srovnání s $\frac{x}{\nu}$. Poněvadž nemusíme vzhledem k pravidlu šesti sigma přihlížeti na jedné straně symetrického rozdělení k hodnotám $x > 3\sigma_x$, může býti náš předpoklad splněn. Při našem binomickém rozdělení je $\sigma_x = \sqrt{\frac{\nu}{2}}$ a tedy $\frac{x}{\nu}$ je pak $\frac{3}{\sqrt{2\nu}}$, což je při velkém ν číslo malé. Můžeme nyní použíti, za uvedeného předpokladu, rozvojů jednotlivých činitelů v čitateli i ve jmenovateli (34) v logaritmické řady (Čech sv. 20, str. 91) podle známého vztahu

$$\lg(1 + \varepsilon) = \varepsilon - \tfrac{1}{2}\varepsilon^2 + \tfrac{1}{3}\varepsilon^3 - \tfrac{1}{4}\varepsilon^4 + \ldots$$

a podržíme vždy jen první člen.

Tak dostaneme přibližné vyjádření pro logaritmus zlomku (34)

$$\lg\frac{y_x}{y_0} = -\frac{2}{\nu}[1 + 2 + 3 + \ldots + (x - 1)] - \frac{x}{\nu} =$$

$$= -\frac{x(x-1)}{\nu} - \frac{x}{\nu} = -\frac{x^2}{\nu}$$

a přejdeme-li od logaritmu k číslu

$$y_x = y_0 e^{-\frac{x^2}{\nu}}.$$

Vzhledem k tomu, že $\nu = 2\sigma_x^2$, můžeme konečně psáti

$$y_x = y_0 e^{-\frac{x^2}{2\sigma_x^2}}, \qquad (35)$$

což je výraz pro t. zv. normální funkci Laplace-Gaussovu. Pořadnice y_0, která odpovídá hodnotě $x = 0$, je maximální pořadnicí, což je zřejmo také z toho, že při $x = 0$ je

$$e^{-\frac{x^2}{2\sigma_x^2}} = \frac{1}{e^{\frac{x^2}{2\sigma_x^2}}} = 1$$

a při jakékoliv jiné hodnotě x bude

$$\frac{1}{e^{\frac{x^2}{2\sigma_x^2}}} < 1.$$

Tak jsme dospěli postupným zvětšováním počtu hodnot znaku až ke znaku spojitému a od rozpojitého rozdělení četností ke spojitému, vyjádřenému symetrickou křivkou normální.

Jsou však také jiné křivky, vyjadřující rozdělení četností jevů, které nejsou symetrické; můžeme je odvoditi podobnými úvahami. Tak nám dávají známé úvahy kombinatoriky počet případů, v nichž se objeví při házení osmi kostkami jednotky nebo dvojky binomickým rozvojem: $(1 + 2)^8$, který nám dává tato čísla:

počet jednotek nebo dvojek . .	0	1	2	3	4	5	6	7	8	celkem
četnost	256	1024	1792	1792	1120	448	112	16	1	3^8

Asymetrie je patrna; při množství pokusů bychom na př. dostali v průměru jen jednou ze $6561 = 3^8$ vrhů případ, že by všechny kostky dávaly na horní straně buď jednotky nebo dvojky.