## Popisná štatistika

*Meranie* je zisťovanie hodnoty danej veličiny porovnávaním s veličinou toho istého druhu, ktorá bola prijatá za jednotku. Pri opakovaní merania pri nezmenených podmienkach dostávame mierne sa odlišujúce hodnoty. Je to spôsobené nedokonalosťou ľudských zmyslov a reflexov, ako aj nedokonalosťou meracích prístrojov. Skutočná hodnota danej veličiny zostáva pri meraní neznáma. Tento fakt sa nazýva *neistota merania*.

**Príklad 1.** Pri meraní dĺžky súčiastky (v mm) boli získané tieto výsledky:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6.04 | 5.99 | 6.03 | 6.03 | 6.00 | 6.02 | 6.01 | 6.01 | 5.99 | 5.98 |
| 6.00 | 6.02 | 6.00 | 6.01 | 6.01 | 5.99 | 6.00 | 6.02 | 6.01 | 6.02 |

Číselné údaje zapisujeme v takom poradí, v akom sme ich získali (*raw data*). Takto vznikne tzv. *prvotná tabuľka*. Celkový počet hodnôt v prvotnej tabuľke označujeme symbolom $n$. Teda prvotná tabuľka má tvar:

$$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad \dots \quad x_n$$

Číselné údaje v prvotnej tabuľke sa označujú symbolom $x_j$.

Číselné hodnoty môžeme usporiadať podľa veľkosti od najmenšej hodnoty po najväčšiu, pričom rovnaké hodnoty zapíšeme toľkokrát, koľkokrát sa vyskytujú (*ordered data*). Pre údaje z príkladu 1 dostávame nasledujúcu tabuľku:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5.98 | 5.99 | 5.99 | 5.99 | 6.00 | 6.00 | 6.00 | 6.00 | 6.01 | 6.01 |
| 6.01 | 6.01 | 6.01 | 6.02 | 6.02 | 6.02 | 6.02 | 6.03 | 6.03 | 6.04 |

Takto vznikne tzv. *variačný rad*, ktorý má tvar:

$$x_{(1)} \leqq x_{(2)} \leqq x_{(3)} \leqq x_{(4)} \leqq x_{(5)} \leqq \dots \leqq x_{(n)}$$

Číselné údaje vo variačnom rade sa označujú symbolom $x_{(j)}$.

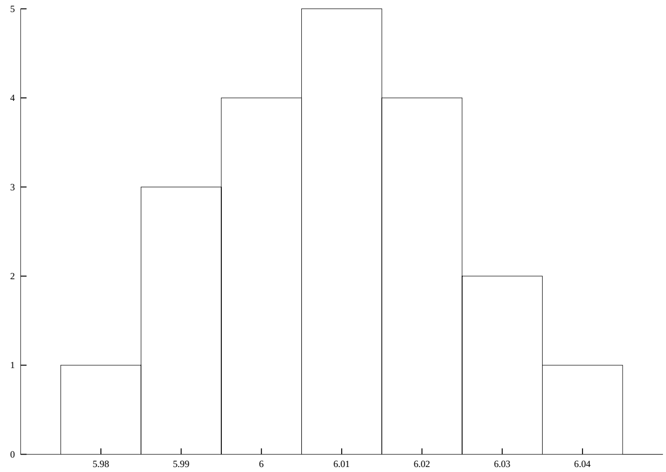
V prípade, že sa niektoré hodnoty opakujú, môžeme použiť tzv. *variačnú tabuľku*. V prvom stĺpci sú navzájom rôzne hodnoty a v druhom stĺpci im odpovedajúce *početnosti* (*frequencies*) výskytu. Pre údaje z príkladu 1 dostávame nasledujúcu tabuľku:

| $x_j$ | $n_j$ |
|---|---|
| 5.98 | 1 |
| 5.99 | 3 |
| 6.00 | 4 |
| 6.01 | 5 |
| 6.02 | 4 |
| 6.03 | 2 |
| 6.04 | 1 |

| $x_j$ | $n_j$ |
|---|---|
| $x_1$ | $n_1$ |
| $x_2$ | $n_2$ |
| $x_3$ | $n_3$ |
| $\vdots$ | $\vdots$ |
| $x_\ell$ | $n_\ell$ |

Počet navzájom rôznych hodnôt označujeme symbolom $\ell$. Početnosť (*frequency*) výskytu hodnoty $x_j$ označujeme symbolom $n_j$.

Na grafické znázornenie nameraných hodnôt používame tzv. *histogram* (*histogram*). Teoreticky pripúšťame, že je možné namerať ľubovoľnú hodnotu z istého intervalu. V skutočnosti však výsledky merania získavame vždy iba s istou presnosťou. Napríklad hodnoty, ktoré boli zaznamenané ako 6.02, sú v skutočnosti čísla ležiace niekde medzi 6.015 a 6.025. Nemáme však žiadnu možnosť zistiť ich presnú polohu v uvedenom intervale. Z tohto dôvodu zostrojíme histogram ako stĺpcový diagram. Nad každým takýmto intervalom so stredom v bode $x_j$ nakreslíme obdĺžnik, ktorého výška je úmerná početnosti $n_j$. Histogram pre variačnú tabuľku z príkladu 1 vidíme na obrázku 1.




Obr. 1

Jednou z najdôležitejších vlastností rozloženia početností je poloha (*central tendency*). Určujeme ju pomocou *stredných hodnôt*, okolo ktorých sa namerané údaje sústreďujú.

Najpoužívanejšou strednou hodnotou je *aritmetický priemer* (*arithmetic mean*). Počítame ho podľa vzorca:

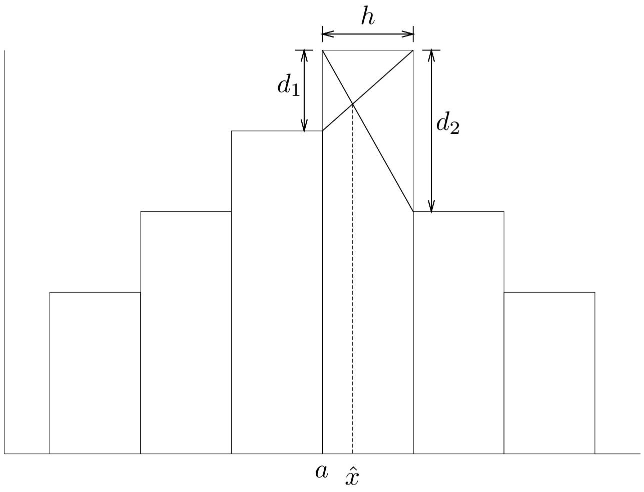
$$\text{aritmetický priemer} = \frac{\text{súčet všetkých nameraných údajov}}{\text{počet všetkých nameraných údajov}}$$

Aritmetický priemer označujeme symbolom $\bar{x}$. Pri jeho používaní musíme dávať pozor na jednu nepríjemnú vlastnosť. Je extrémne citlivý na vybočujúce údaje, ktoré sa nápadne odlišujú od ostatných.

Pre údaje z variačnej tabuľky vzorec pre aritmetický priemer má tvar:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{\ell} x_i \cdot n_i$$

*Modus* (*mode*) je charakteristika odvodená z histogramu. Predstavuje polohu jeho maxima. Pri jeho určení vychádzame z predpokladu, že namerané hodnoty sa v modálnom intervale viac koncentrujú k tej hranici, ktorej susedný interval má väčšiu početnosť (pozri obrázok 2).




Obr. 2

Z podobnosti trojuholníkov vyplýva, že platí:

$$\frac{\hat{x}-a}{d_1}=\frac{(a+h)-\hat{x}}{d_2}.$$

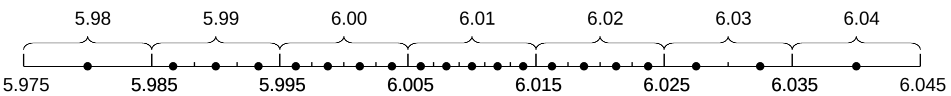
Odtiaľ dostávame:

$$\hat{x}=a+h\cdot\frac{d_1}{d_1+d_2}.$$

Pritom
- $a$ je začiatok modálneho intervalu (t. j. intervalu s najväčšou početnosťou),
- $a+h$ je koniec modálneho intervalu,
- $h$ je dĺžka modálneho intervalu,
- $d_1$ je rozdiel medzi početnosťou modálneho a predchádzajúceho intervalu,
- $d_2$ je rozdiel medzi početnosťou modálneho a nasledujúceho intervalu.

Tento výpočet využíva iba tri intervaly v blízkosti maxima. Preto modus nie je citlivý na vybočujúce údaje.

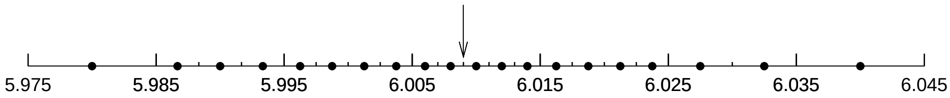
*Medián* (*median*) rozdeľuje variačný rad na dve rovnako početné časti.
Začneme konkrétnym príkladom. Pozrime sa na variačnú tabuľku z príkladu 1. Pretože výsledky merania získavame vždy iba s istou presnosťou, nemáme žiadnu možnosť zistiť ich presnú polohu. Z tohto dôvodu budeme predpokladať, že v každom intervale (t. j. medzi číslami $x_j - h/2$ a $x_j + h/2$) sú namerané hodnoty rozložené rovnomerne.




Teraz musíme identifikovať interval, v ktorom medián leží. V našom príklade máme 20 výsledkov merania. Teda 10 hodnôt by malo ležať pod mediánom a 10 nad mediánom. V prvom intervale máme 1 nameranú hodnotu, v prvých dvoch 4, v prvých troch 8, v prvých štyroch 13. Teda medián leží vo štvrtom intervale, t. j. medzi číslami 6.005 a 6.015. Dĺžka tohto intervalu je $h = 6.015 - 6.005 = 0.01$.

Pretože v mediánovom intervale leží 5 nameraných hodnôt, rozdelíme tento interval na 5 rovnakých častí. Pretože v prvých troch intervaloch leží 8 hodnôt, potrebujeme k nim pridať ešte dve hodnoty. Teda musíme ísť do 2/5 mediánového intervalu. Z tohto dôvodu k dolnej hranici mediánového intervalu pridáme $(2/5) \cdot h$. Potom platí:

$$\text{medián} = 6.005 + (2/5) \cdot 0.01 = 6.009$$




Medián označujeme symbolom $\tilde{x}$. Predchádzajúci výpočet možno zapísať vzorcom:

$$\tilde{x} = a + h \cdot \frac{\frac{n}{2} - N_{i-1}}{n_i}$$

kde

$a$ je začiatok mediánového intervalu,

$n_i$ je početnosť mediánového intervalu,

$N_{i-1} = n_1 + n_2 + \cdots + n_{i-1}$ je súčet početností zo všetkých intervalov (t. j. kumulatívna početnosť), ktoré predchádzajú mediánovému intervalu,

$h$ je dĺžka mediánového intervalu,

$n$ je počet všetkých nameraných hodnôt.

Pritom *kumulatívne početnosti* (*cummulative numbers of frequencies*) $N_j$ definujeme nasledujúcim spôsobom:

$$\begin{aligned}
N_1 &= n_1 \\
N_2 &= n_1 + n_2 \\
N_3 &= n_1 + n_2 + n_3 \\
&\vdots \\
N_\ell &= n_1 + n_2 + n_3 + \cdots + n_\ell = n
\end{aligned}$$

Ďalšou z vlastností rozloženia početností je variabilita.

Prvý nápad, totiž použiť priemernú odchýlku od aritmetického priemeru, je nepoužiteľný. Je to spôsobené tým, že platí:

$$\frac{1}{n} \cdot \sum_{i=1}^{\ell} (x_i - \bar{x}) \cdot n_i = \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{\ell} x_i \cdot n_i}_{=\bar{x}} - \frac{1}{n} \cdot \bar{x} \cdot \underbrace{\sum_{i=1}^{\ell} n_i}_{=n} = \bar{x} - \bar{x} = 0$$

Vhodnejší je priemer z druhých mocnín odchýlok, tzv. *disperzia* (*variance*):

$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^{\ell} (x_i - \bar{x})^2 \cdot n_i$$

Disperzia charakterizuje rozptyl údajov okolo aritmetického priemeru. Nevýhodou disperzie je, že jej fyzikálny rozmer je druhou mocninou fyzikálneho rozmeru meranej veličiny. Preto sa zavádza *stredná kvadratická odchýlka* (*standard deviation*):

$$S = \sqrt{S^2}$$

Pri výberovom skúmaní sa používa tzv. *výberová disperzia* (*sample variance*):

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{\ell} (x_i - \bar{x})^2 \cdot n_i$$

a *výberová odchýlka* (*sample deviation*):

$$s = \sqrt{s^2}$$

Mierou rozptylu je aj *kvartilová odchýlka* (*quartile deviation*), ktorá je definovaná ako polovica vzdialenosti medzi prvým a tretím kvartilom:

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

Medián $\tilde{x} = Q_2$ oddeľuje dolných 50% hodnôt od horných 50% hodnôt. *Prvý kvartil* (*first quartile*) $Q_1$ oddeľuje dolných 25% hodnôt od horných 75% hodnôt (používa sa aj názov *dolný kvartil*). Počítame ho podľa vzorca:

$$Q_1 = a + h \cdot \frac{\frac{n}{4} - N_{i-1}}{n_i}$$

kde

$a$ je začiatok intervalu obsahujúceho prvý kvartil,
$n_i$ je početnosť intervalu obsahujúceho prvý kvartil,
$N_{i-1} = n_1 + n_2 + \cdots + n_{i-1}$ je súčet početností zo všetkých intervalov (t. j. kumulatívna početnosť), ktoré predchádzajú intervalu obsahujúcemu prvý kvartil,
$h$ je dĺžka intervalu obsahujúceho prvý kvartil,
$n$ je počet všetkých nameraných hodnôt.

*Tretí kvartil* (*third quartile*) $Q_3$ oddeľuje dolných 75% hodnôt od horných 25% hodnôt (používa sa aj názov *horný kvartil*). Počítame ho podľa vzorca:

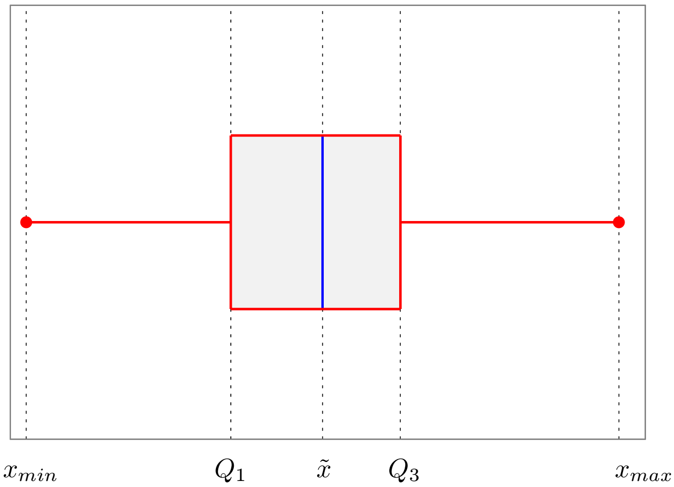
$$Q_3 = a + h \cdot \frac{\frac{3n}{4} - N_{i-1}}{n_i}$$

kde

$a$ je začiatok intervalu obsahujúceho tretí kvartil,
$n_i$ je početnosť intervalu obsahujúceho tretí kvartil,
$N_{i-1} = n_1 + n_2 + \cdots + n_{i-1}$ je súčet početností zo všetkých intervalov (t. j. kumulatívna početnosť), ktoré predchádzajú intervalu obsahujúcemu tretí kvartil,
$h$ je dĺžka intervalu obsahujúceho tretí kvartil,
$n$ je počet všetkých nameraných hodnôt.

Všimnime si, že medzi prvým a tretím kvartilom leží 50% zo všetkých nameraných hodnôt.

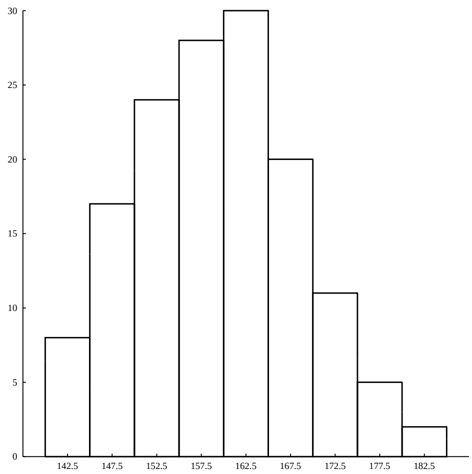
Spôsob znázornenia všetkých troch kvartilov, ako aj najmenšej hodnoty a najväčšej hodnoty, sa volá *box-and-whisker plot*:



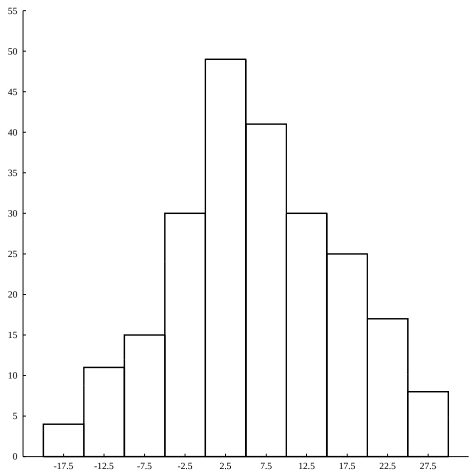

Obr. 3

V ďalšej časti sa budeme zaoberať charakteristikami koncentrácie hodnôt číselných údajov.

*Koeficient šikmosti* (*coefficient of skewness*) je založený na porovnaní stupňa koncentrácie malých hodnôt so stupňom koncentrácie veľkých hodnôt. Rovnaký stupeň koncentrácie malých a veľkých hodnôt sa prejavuje symetričnosťou histogramu početností. Väčší stupeň koncentrácie malých hodnôt v porovnaní so stupňom koncentrácie veľkých hodnôt sa prejavuje kladne zošikmeným tvarom histogramu početností (pozri obrázok 4). Väčší stupeň koncentrácie veľkých hodnôt v porovnaní so stupňom koncentrácie malých hodnôt sa prejavuje záporne zošikmeným tvarom histogramu početností (pozri obrázok 5).




Obr. 4




Obr. 5

Koeficient šikmosti počítame podľa vzorca:

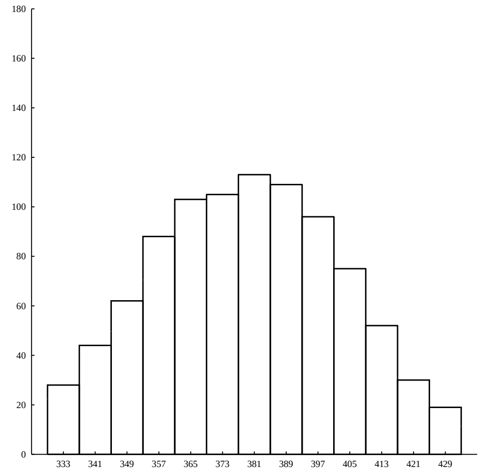
$$\gamma_3 = \frac{\mu_3}{S^3}$$

kde

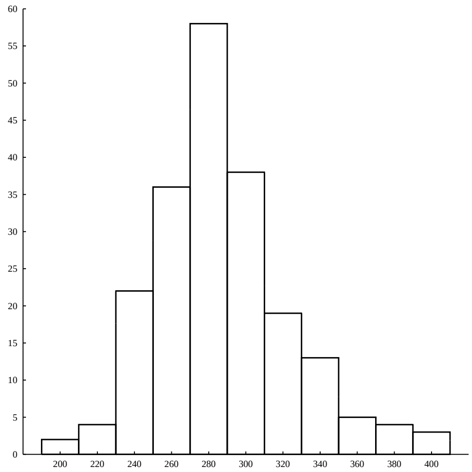
$$\mu_3 = \frac{1}{n} \cdot \sum_{i=1}^{\ell} (x_i - \bar{x})^3 \cdot n_i$$

je tzv. *tretí centrálny moment.*

*Koeficient špicatosti* (*coefficient of kurtosis*) je založený na porovnaní stupňa koncentrácie hodnôt strednej veľkosti so stupňom koncentrácie ostatných hodnôt. Rovnaký stupeň koncentrácie prostredných hodnôt a ostatných hodnôt sa prejavuje plochosťou histogramu početností (pozri obrázok 6). Vyšší stupeň koncentrácie prostredných hodnôt v porovnaní s ostatnými hodnotami sa prejavuje špicatejším tvarom histogramu početností (pozri obrázok 7).




Obr. 6




Obr. 7

Koeficient špicatosti počítame podľa vzorca:

$$\gamma_4 = \frac{\mu_4}{S^4} - 3$$

kde

$$\mu_4 = \frac{1}{n} \cdot \sum_{i=1}^{\ell} (x_i - \bar{x})^4 \cdot n_i$$

je tzv. *štvrtý centrálny moment.*

Pri veľkých súboroch (viac ako 50 hodnôt) bude mať histogram „hrebeňovitý“ tvar, v ktorom sú niektoré intervaly úplne prázdne alebo len veľmi málo obsadené.

**Príklad 2.** Nasledujúca tabuľka obsahuje výsledky merania:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 561 | 555 | 569 | 555 | 567 | 559 | 567 | 555 | 566 | 557 | 580 | 568 | 561 | 572 | 563 |
| 574 | 542 | 562 | 542 | 572 | 564 | 560 | 569 | 543 | 560 | 565 | 568 | 558 | 539 | 550 |
| 566 | 563 | 562 | 546 | 570 | 582 | 568 | 565 | 561 | 554 | 548 | 558 | 586 | 562 | 559 |
| 558 | 545 | 563 | 557 | 574 | 550 | 562 | 557 | 566 | 559 | 576 | 560 | 554 | 552 | 541 |
| 534 | 574 | 560 | 548 | 573 | 562 | 556 | 577 | 554 | 564 | 567 | 546 | 571 | 563 | 557 |
| 552 | 562 | 550 | 551 | 566 | 576 | 572 | 542 | 569 | 556 | 557 | 555 | 569 | 571 | 575 |
| 556 | 540 | 557 | 549 | 577 | 562 | 552 | 568 | 554 | 568 | 554 | 531 | 568 | 567 | 545 |
| 566 | 547 | 571 | 558 | 555 | 550 | 555 | 562 | 550 | 561 | 552 | 571 | 559 | 556 | 558 |
| 554 | 580 | 571 | 560 | 553 | 549 | 544 | 565 | 557 | 562 | 580 | 546 | 538 | 553 | 541 |
| 572 | 544 | 556 | 542 | 552 | 571 | 555 | 560 | 564 | 565 | 538 | 552 | 552 | 563 | 577 |
| 566 | 560 | 544 | 548 | 560 | 549 | 543 | 560 | 552 | 570 | 560 | 549 | 567 | 543 | 542 |
| 538 | 552 | 549 | 553 | 561 | 566 | 549 | 543 | 561 | 547 | 547 | 587 | 576 | 576 | 563 |
| 547 | 548 | 556 | 562 | 537 | 554 | 548 | 572 | 569 | 568 | 550 | 558 | 574 | 560 | 545 |
| 560 | 545 | 536 | 557 | 561 | | | | | | | | | | |

Namerané hodnoty najskôr usporiadame do variačného radu. Zapíšeme ich skráteným spôsobom, ktorý sa nazýva *stem-and-leaf display*:

| | |
|---|---|
| 53 | 146 788 89 |
| 54 | 011 222 223 333 444 555 566 677 778 888 899 999 9 |
| 55 | 000 000 122 222 222 233 344 444 445 555 555 666 666 777 777 778 888 889 999 |
| 56 | 000 000 000 000 111 111 122 222 222 223 333 334 445 555 666 666 677 777 888 888 899 999 |
| 57 | 001 111 112 222 234 444 566 667 77 |
| 58 | 000 267 |

Napríklad v prvom riadku máme zapísané čísla 531, 534, 536, 537, 538, 538, 538, 539.

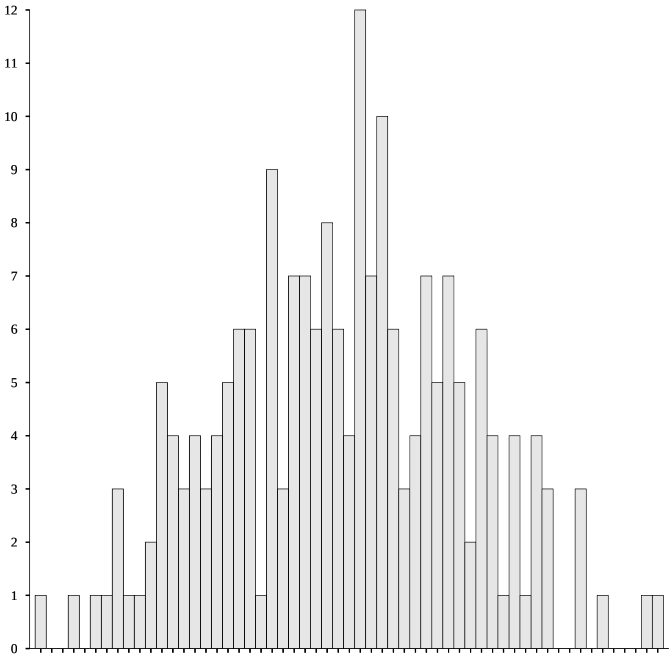
Príslušný histogram vidíme na obrázku 8. Takéto histogramy sa používajú napr. v digitálnej fotografii. V tomto histograme sú niektoré intervaly prázdne, alebo len veľmi málo obsadené. Je to spôsobené tým, že intervalov je príliš veľa. Oblasť výskytu hodnôt preto rozdelíme na primeraný počet intervalov rovnakej dĺžky. Pri veľkom počte intervalov bude mať histogram ešte stále „hrebeňovitý“ tvar, pri malom počte intervalov bude mať histogram malú informačnú hodnotu.

V literatúre možno nájsť množstvo rôznych vzorčekov na určenie počtu intervalov $k$. Asi najjednoduchší je

$$k \doteq \sqrt{n}$$

Niektorí autori odporúčajú tzv. *Sturgesovo pravidlo*:

$$k \doteq 1 + \log_2 n$$




Obr. 8

Tento vzorec vďačí za svoj vznik nasledujúcemu pozorovaniu. Ak $n = 2^{k-1}$, potom číslo $n$ možno vyjadriť (pomocou binomickej vety) v tvare súčtu $k$ kombinačných čísel. Napríklad:

$$n = 2^6 = (1+1)^6 = \underbrace{\binom{6}{0} + \binom{6}{1} + \binom{6}{2} + \binom{6}{3} + \binom{6}{4} + \binom{6}{5} + \binom{6}{6}}_{k=7 \text{ sčítancov}}$$

Ak z rovnosti $n = 2^{k-1}$ vyjadríme $k$, dostávame $k = 1 + \log_2 n$.

Žiadny z takýchto vzorcov však nemôže mať univerzálnu platnosť, pretože optimálny výber počtu intervalov záleží nielen od počtu nameraných údajov, ale aj od spôsobu ich rozloženia v oblasti výskytu. Túto skutočnosť zohľadňuje vzorec:

$$k = A \cdot n^{0.4}$$

pričom číslo $A$ závisí od spôsobu rozloženia nameraných údajov. Bolo zistené, že číslo $A$ kolíše medzi 0.55 až 1.25.

Uvedené vzorce považujeme iba za orientačné. Niektorí autori dokonca namiesto vzorcov iba uvádzajú, že je vhodné zvoliť $k$ niekde medzi 7 až 15.

Namerané hodnoty spracujeme do tzv. *tabuľky variačného triedenia* (*frequency distribution*). Pre údaje z príkladu 2 bude mať tvar:

| Stated limits | $z_j$ | $n_j$ |
|---|---|---|
| 530 – 534 | 532 | 2 |
| 535 – 539 | 537 | 6 |
| 540 – 544 | 542 | 15 |
| 545 – 549 | 547 | 22 |
| 550 – 554 | 552 | 26 |
| 555 – 559 | 557 | 31 |
| 560 – 564 | 562 | 38 |
| 565 – 569 | 567 | 28 |
| 570 – 574 | 572 | 18 |
| 575 – 579 | 577 | 8 |
| 580 – 584 | 582 | 4 |
| 585 – 589 | 587 | 2 |

Pritom symbolom $z_j$ sme označili stred (*mid-value*) intervalu $I_j$.

V skutočnosti však výsledky merania získavame vždy iba s istou presnosťou. Napríklad hodnoty, ktoré boli zaznamenané ako 562, sú v skutočnosti čísla ležiace niekde medzi 561.5 a 562.5. Nemáme však žiadnu možnosť zistiť ich presnú polohu v uvedenom intervale. Predchádzajúcu tabuľku teda upravíme do tvaru:

| True limits | $z_j$ | $n_j$ |
|---|---|---|
| 529.5 – 534.5 | 532 | 2 |
| 534.5 – 539.5 | 537 | 6 |
| 539.5 – 544.5 | 542 | 15 |
| 544.5 – 549.5 | 547 | 22 |
| 549.5 – 554.5 | 552 | 26 |
| 554.5 – 559.5 | 557 | 31 |
| 559.5 – 564.5 | 562 | 38 |
| 564.5 – 569.5 | 567 | 28 |
| 569.5 – 574.5 | 572 | 18 |
| 574.5 – 579.5 | 577 | 8 |
| 579.5 – 584.5 | 582 | 4 |
| 584.5 – 589.5 | 587 | 2 |

Číselné charakteristiky z tabuľky variačného triedenia počítame takým istým spôsobom, ako z variačnej tabuľky, len namiesto nameraných hodnôt $x_j$ dosadzujeme zástupné hodnoty $z_j$. Napríklad vzorec pre aritmetický priemer bude mať tvar:

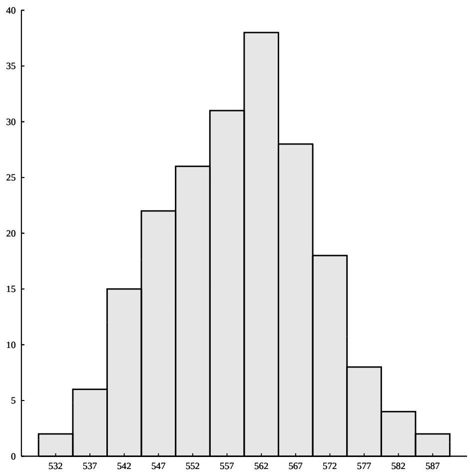
$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} z_i \cdot n_i$$

Takto vypočítané číselné charakteristiky sa líšia od číselných charakteristík vypočítaných z pôvodných údajov. Je to spôsobené tým, že namerané hodnoty nie sú v triednych intervaloch rozložené rovnomerne, ale sa zhusťujú smerom k tej hranici intervalu, ktorá leží bližšie k vrcholu histogramu. Ak teda namiesto skutočne nameraných hodnôt použijeme stredy intervalov, dopúšťame sa určitého skreslenia. Na odstránenie tohto skreslenia používame tzv. *Sheppardove korekcie*:

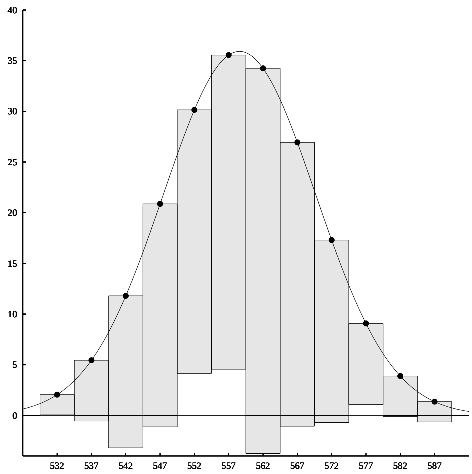
$$S^2_{kor.} = S^2 - \frac{1}{12} \cdot h$$
$$\mu_{4,\,kor.} = \mu_4 - \frac{1}{2} \cdot S^2 h^2 + \frac{7}{240} \cdot h^4$$

Dodajme, že korekcia sa neprevádza pre aritmetický priemer, ani pre tretí centrálny moment.

Histogram pre tabuľku variačného triedenia z príkladu 2 vidíme na obrázku 9. Na porovnanie histogramov s „ideálnym rozdelením“ sú vhodnejšie tzv. *visiace histogramy*, pretože sa pri nich hodnotia odchýlky od priamky, nie od krivky (pozri obrázok 10).



Obr. 9



Obr. 10

V štatistike sa môžeme stretnúť aj s číselnými údajmi, ktoré majú diskrétny charakter, nie sú teda výsledkom merania.

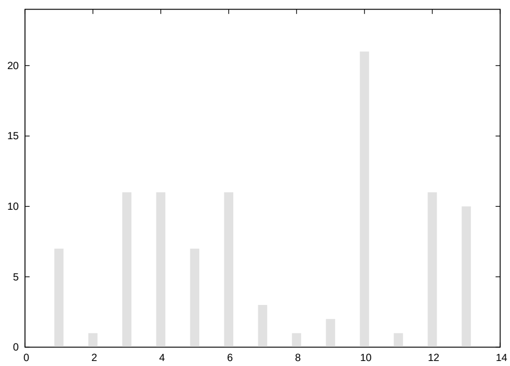
**Príklad 3.** Súčiastky sa dodávajú v balení po 100 kusoch. Zákazník zistil nasledujúce počty súčiastok v jednotlivých baleniach:

99 100 98 101 98 100 102 99 99 100

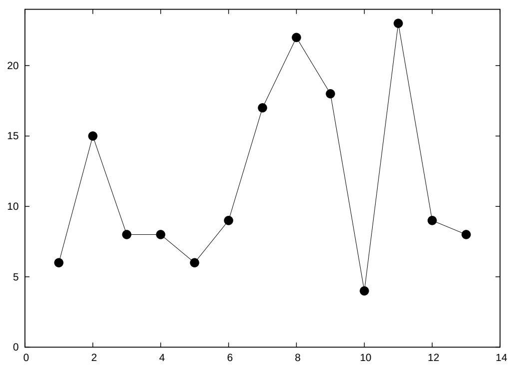
Variačná tabuľka pre tieto údaje má tvar:

| $x_j$ | $n_j$ |
|---|---|
| 98 | 2 |
| 99 | 3 |
| 100 | 3 |
| 101 | 1 |
| 102 | 1 |

Na grafické znázornenie variačnej tabuľky používame *stĺpcový diagram* (obrázok 11), prípadne *polygón* (obrázok 12).




Obr. 11




Obr. 12

Aritmetický priemer a disperziu môžeme vypočítať priamo z prvotnej tabuľky:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad\qquad S^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

alebo z variačnej tabuľky:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{\ell} x_i \cdot n_i \qquad\qquad S^2 = \frac{1}{n} \cdot \sum_{i=1}^{\ell} (x_i - \bar{x})^2 \cdot n_i$$

Pre údaje z príkladu 3 máme $\bar{x} = 99.6$ a $S^2 = 1.44$.

Avšak pri odvodení niektorých číselných charakteristík boli podstatným spôsobom použité intervaly. Pri diskrétnych údajoch musíme zvoliť inú stratégiu. Najskôr si vysvetlíme základnú myšlienku. Budeme pri tom potrebovať *empirickú distribučnú funkciu* (*empirical distribution function*). Pre variačnú tabuľku ju definujeme nasledujúcim spôsobom:

$$F(x) = \begin{cases} 0 & \text{ak} \quad x \leqq x_1 \\ F_1 & \text{ak} \quad x_1 < x \leqq x_2 \\ F_2 & \text{ak} \quad x_2 < x \leqq x_3 \\ \vdots & \qquad \vdots \\ F_{\ell-1} & \text{ak} \quad x_{\ell-1} < x \leqq x_\ell \\ F_\ell & \text{ak} \quad x > x_\ell \end{cases}$$

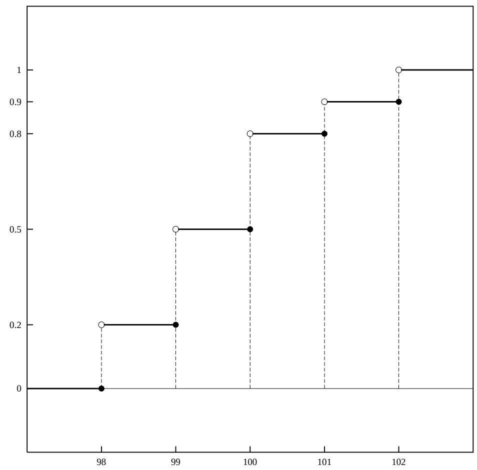
Pritom symbolom $F_j$ označujeme tzv. *relatívne kumulatívne početnosti* (*cumulative relative frequency*), ktoré definujeme takto:

$$F_j = \frac{N_j}{n}$$

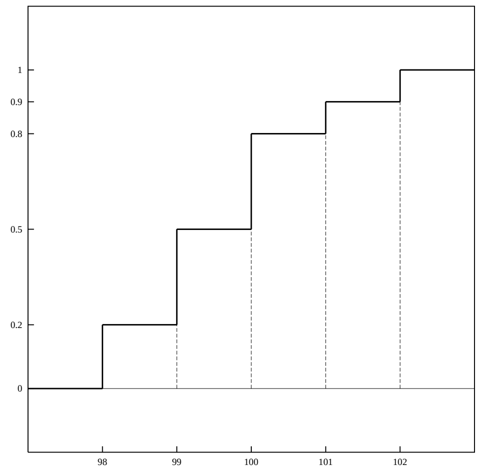
Variačnú tabuľku pre údaje z príkladu 3 doplníme o absolútne a relatívne kumulatívne početnosti:

| $x_j$ | $n_j$ | $N_j$ | $F_j$ |
|---|---|---|---|
| 98 | 2 | 2 | 0.2 |
| 99 | 3 | 5 | 0.5 |
| 100 | 3 | 8 | 0.8 |
| 101 | 1 | 9 | 0.9 |
| 102 | 1 | 10 | 1.0 |

Graf príslušnej empirickej distribučnej funkcie vidíme na obrázku 13. Z praktických dôvodov je vhodnejšie používať schodovitý graf, ktorý vidíme na obrázku 14.




Obr. 13

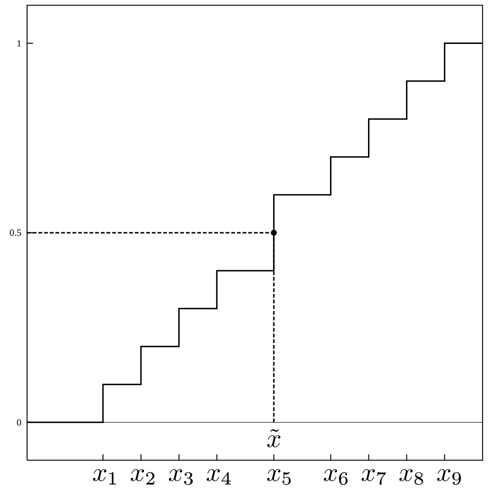



Obr. 14

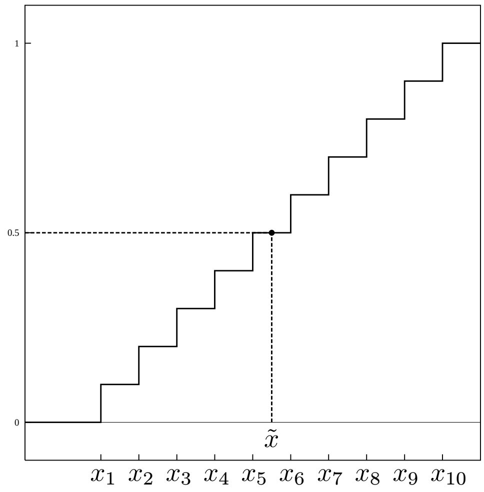
Pretože medián oddeľuje dolných 50% hodnôt od horných 50% hodnôt, určíme ho ako $x$-ovú súradnicu toho bodu na schodovitom grafe empirickej distribučnej funkcie, ktorého $y$-ová súradnica je 0.5. Ako vidíme na nasledujúcich obrázkoch, sú možné dve situácie. Ak priamka $y = 0.5$ pretína zvislú časť grafu (obrázok 15), hľadaný bod existuje jediný. V opačnom prípade (obrázok 16) je takýchto bodov nekonečne veľa – vypĺňajú celý interval. Za medián zvolíme stred tohto intervalu.

Pri praktickom počítaní mediánu z diskrétnych údajov je výhodnejšie vychádzať z variačného radu:

1) ak počet pozorovaní $n$ je nepárne číslo, potom medián je tá hodnota vo variačnom rade, ktorej poradové číslo je $\dfrac{n+1}{2}$;
2) ak $n$ je párne číslo, potom medián určíme ako aritmetický priemer z dvoch susedných hodnôt vo variačnom rade, ktorých poradové čísla sú $\dfrac{n}{2}$ a $\dfrac{n}{2} + 1$.




Obr. 15




Obr. 16

Môžeme to zapísať nasledujúcim spôsobom:

ak $n$ je nepárne číslo, potom $\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$

ak $n$ je párne číslo, potom $\tilde{x} = \frac{1}{2} \cdot \left( x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right)$

Prvotnú tabuľku z príkladu 3 usporiadame do variačného radu:

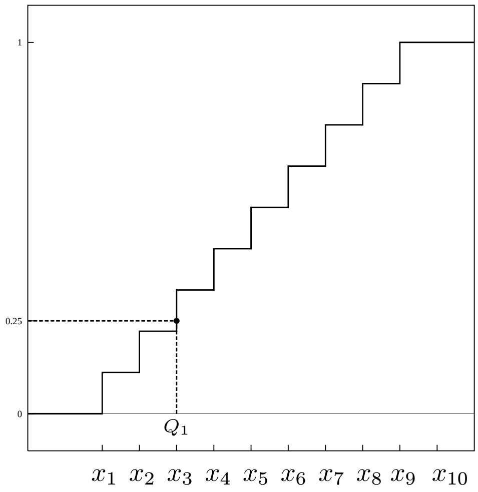
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 98 | 98 | 99 | 99 | 99 | 100 | 100 | 100 | 101 | 102 |
| $x_{(1)}$ | $x_{(2)}$ | $x_{(3)}$ | $x_{(4)}$ | $x_{(5)}$ | $x_{(6)}$ | $x_{(7)}$ | $x_{(8)}$ | $x_{(9)}$ | $x_{(10)}$ |

Pretože počet pozorovaní $n = 10$ je párne číslo, medián vypočítame ako aritmetický priemer z čísel $x_{(5)} = 99$ a $x_{(6)} = 100$, teda $\tilde{x} = \frac{1}{2} \cdot (99 + 100) = 99.5$.

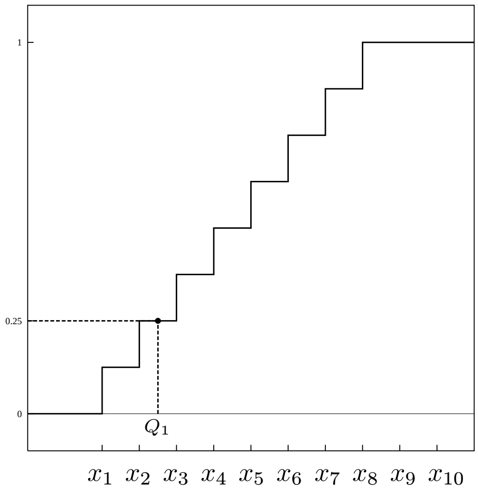
Pretože dolný kvartil oddeľuje dolných 25% hodnôt od horných 75% hodnôt, určíme ho ako $x$-ovú súradnicu toho bodu na schodovitom grafe empirickej distribučnej funkcie, ktorého $y$-ová súradnica je 0.25. Ako vidíme na nasledujúcich obrázkoch, sú možné dve situácie. Ak priamka $y = 0.25$ pretína zvislú časť grafu (obrázok 17), hľadaný bod existuje jediný. V opačnom prípade (obrázok 18) je takýchto bodov nekonečne veľa – vypĺňajú celý interval. Za dolný kvartil zvolíme stred tohto intervalu.

Pri praktickom počítaní dolného kvartilu z diskrétnych údajov je výhodnejšie vychádzať z variačného radu:

1) ak číslo $n$ nie je deliteľné štyrmi, potom v intervale $\left(\frac{n}{4}, \frac{n}{4} + 1\right)$ leží jediné prirodzené číslo $k$, ktoré je poradovým číslom dolného kvartilu $Q_1$; pritom číslo $k$ môžeme vyjadriť vzorcom $k = \left[\frac{n}{4}\right] + 1$, kde symbol $[a]$ označuje tzv. *celú časť reálneho čísla* $a$, t. j. také celé číslo $m$, pre ktoré platí: $m \leqq a < m + 1$;
2) ak číslo $n$ je deliteľné štyrmi, potom dolný kvartil $Q_1$ určíme ako aritmetický priemer z dvoch susedných hodnôt vo variačnom rade, ktorých poradové čísla sú $\frac{n}{4}$ a $\frac{n}{4} + 1$.




Obr. 17




Obr. 18

Môžeme to zapísať nasledujúcim spôsobom:

$$\text{ak } n \text{ nie je deliteľné štyrmi, potom } \quad Q_1 = x_{\left(\left[\frac{n}{4}\right]+1\right)}$$

$$\text{ak } n \quad \text{ je deliteľné štyrmi, potom } \quad Q_1 = \frac{1}{2}\cdot\left(x_{\left(\frac{n}{4}\right)} + x_{\left(\frac{n}{4}+1\right)}\right)$$

Pri určovaní horného kvartilu $Q_3$ postupujeme presne tak isto, len všade namiesto $\frac{n}{4}$ píšeme $\frac{3n}{4}$. Môžeme to zapísať nasledujúcim spôsobom:

$$\text{ak } n \text{ nie je deliteľné štyrmi, potom } \quad Q_3 = x_{\left(\left[\frac{3n}{4}\right]+1\right)}$$

$$\text{ak } n \quad \text{ je deliteľné štyrmi, potom } \quad Q_3 = \frac{1}{2}\cdot\left(x_{\left(\frac{3n}{4}\right)} + x_{\left(\frac{3n}{4}+1\right)}\right)$$

## Náhodné javy a ich pravdepodobnosť

*Náhodný pokus* je činnosť, ktorá pri dodržaní predpísaných podmienok môže viesť k rôznym výsledkom. Ak takýto pokus niekoľkokrát opakujeme, výsledky sa môžu od jedného prevedenia pokusu k druhému meniť. Výsledky závisia nielen na predpísaných podmienkach, ale aj na náhode. Náhodným pokusom je napr. meranie fyzikálnej veličiny, výroba určitého výrobku, výber výrobku na kontrolu a pod.

*Náhodný jav* je overiteľné tvrdenie o výsledku náhodného pokusu, príp. o skupine výsledkov náhodného pokusu, t.j. také tvrdenie, o ktorom môžeme po uskutočnení pokusu rozhodnúť, či je alebo nie je pravdivé. Napríklad pri hode hracou kockou môžeme za náhodný jav považovať padnutie steny s párnym číslom.

Uvažujme hod hracou kockou ako náhodný jav, ktorý mnohokrát opakujeme. Zistíme, že každé z čísel 1, 2, 3, 4, 5, 6 sa vyskytuje približne rovnaký počet krát. Teda každé z týchto čísel má rovnakú relatívnu početnosť výskytu blízku číslu 1/6.

*Relatívna početnosť* výskytu javu je podiel počtu tých prípadov, v ktorých sledovaný jav nastal, k celkovému počtu prevedených pokusov.

V teórii pravdepodobnosti nás zaujímajú také náhodné javy, ktorých relatívna početnosť výskytu pri mnohonásobnom opakovaní pokusu vykazuje istú stabilitu, t. j. so stále rastúcim počtom pokusov relatívna početnosť výskytu stále menej kolíše okolo určitého čísla, ktoré voláme *pravdepodobnosť* tohto javu.

Výsledky hodu kockou sú rovnako pravdepodobné len v prípade ideálnej kocky, t. j. ak je kocka geometricky pravidelná a vyrobená z homegénneho materiálu. Naproti tomu pravdepodobnosť padnutia šestky pri hode „falošnou“ kockou (ktorá vo vnútri obsahuje napr. kúsok olova umiestneného oproti stene so šestkou) bude väčšia ako 1/6. Ak chceme túto pravdepodobnosť určiť, musíme hod takouto kockou mnohokrát zopakovať a zistenú relatívnu početnosť výskytu šestky prijať za približnú hodnotu pravdepodobnosti. Presnú hodnotu tejto pravdepodobnosti pritom pokladáme za neznámu konštantu, ktorá leží niekde blízko zistenej relatívnej početnosti.

Pravdepodobnosť náhodného javu $A$ je číslo $P(A)$, ktoré môžeme interpretovať ako mieru možnosti nastatia náhodného javu $A$. Inými slovami, pravdepodobnosť môžeme považovať za predpoveď relatívnych početností pri mnohonásobnom opakovaní náhodného pokusu.

## Náhodné veličiny

*Náhodná veličina* je premenná veličina závislá na náhode. Náhodná veličina nadobúda pri opakovaní pokusu vplyvom náhodných činiteľov rôzne hodnoty. Teda výsledok náhodného pokusu, ktorý je vyjadrený reálnym číslom, je hodnotou takejto náhodnej veličiny.

Výsledky získané popisnou štatistikou majú obmedzenú platnosť. Aj keď vychádzajú z náhodnej povahy pozorovaní, nepoužívajú sa na ich rozbor pravdepodobnostné metódy. Preto nemôžeme rozšíriť platnosť záverov na súbory, ktoré by vznikli pri ďalšom opakovaní pozorovania pri rovnakých podmienkach, ani o predpovedanie výsledkov ďalších pozorovaní.

*Distribučná funkcia* náhodnej veličiny $X$ je reálna funkcia reálnej premennej, ktorá každému číslu $x$ priraďuje pravdepodobnosť toho, že náhodná veličina $X$ nadobudne hodnoty menšie ako $x$. Distribučnú funkciu označujeme symbolom $F$. Teda pre každé reálne číslo $x$ máme

$$F(x) = P(X < x)$$

Pri štatistickom vyhodnocovaní nameraných údajov budeme vychádzať z predpokladu, že teoreticky môžeme namerať ľubovoľnú hodnotu z istého intervalu, t. j. že meraná veličina má spojitý charakter. Kľúčom k nájdeniu matematického popisu spojitej náhodnej veličiny bude tvar histogramu relatívnych početností pre veľké počty nameraných hodnôt. *Relatívne početnosti*

$$f_j = \frac{n_j}{n}$$

pri veľkých hodnotách $n$ vyjadrujú pravdepodobnosť výskytu hodnôt danej náhodnej veličiny v intervale $I_j$. Ak s rastúcim počtom meraní $n$ budeme zväčšovať počet intervalov histogramu, šírka týchto intervalov bude veľmi malá a obrys histogramu sa bude približovať ku spojitej krivke. Táto krivka je grafom funkcie, ktorá sa nazýva *hustota pravdepodobnosti* spojitej náhodnej veličiny. Označujeme ju symbolom $f$. Medzi distribučnou funkciu a hustotou pravdepodobnosti platí nasledujúci vzťah:

$$F(x) = \int\limits_{-\infty}^{x} f(t)\,dt$$

## Normálne rozdelenie

Nasledujúce úvahy nás privedú k hustote pravdepodobnosti, ktorá sa nazýva *Gaussova krivka*. Uvažujme spojitú náhodnú veličinu, ktorej hustota pravdepodobnosti $f$ je dostatočne hladká a má ostré lokálne maximum v bode $x = m$. Položme

$$g(x) = \ln f(x)$$

Funkciu $g$ nahradíme v okolí bodu $x = m$ jej Taylorovým polynómom druhého stupňa:

$$g(x) \doteq g(m) + \frac{g'(m)}{1!} \cdot (x - m) + \frac{g''(m)}{2!} \cdot (x - m)^2 \tag{1}$$

Pretože funkcia $g$ má ostré lokálne maximum v bode $x = m$, jej prvá derivácia v tomto bode je nulová a druhá derivácia je záporná. Teda

$$g'(m) = 0$$
$$g''(m) < 0$$

Položme

$$\sigma = \frac{1}{\sqrt{-g''(m)}}$$

Odtiaľ po malej úprave dostávame:

$$g''(m) = -\frac{1}{\sigma^2}$$

Po dosadení do (1) máme:

$$g(x) \doteq g(m) - \frac{1}{2\sigma^2} \cdot (x - m)^2$$

Pretože $f(x) = \exp[g(x)]$, platí[1]

$$f(x) \doteq \exp\left[g(m) - \frac{1}{2\sigma^2} \cdot (x - m)^2\right]$$

Položme

$$A = \exp[g(m)]$$

Potom máme

$$f(x) \doteq A \cdot \exp\left[-\frac{1}{2} \cdot \frac{(x - m)^2}{\sigma^2}\right] \tag{2}$$

Konštantu $A$ určíme z podmienky

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1$$

t. j. z podmienky

$$A \cdot \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2} \cdot \frac{(x - m)^2}{\sigma^2}\right] dx = 1$$

---

[1]Z typografických dôvodov namiesto $e^x$ používame zápis $\exp(x)$.

Pomocou substitúcie

$$\frac{x-m}{\sigma} = t$$

prevedieme túto podmienku do tvaru:

$$A \cdot \sigma \cdot \int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{2}\right) dt = 1 \tag{3}$$

Zostáva nám vypočítať integrál

$$I = \int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{2}\right) dt$$

Použijeme nasledujúci postup:

$$I^2 = \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2}\right) dx \cdot \int_{-\infty}^{+\infty} \exp\left(-\frac{y^2}{2}\right) dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2+y^2}{2}\right) dxdy$$

Prechodom k polárnym súradniciam

$$\begin{aligned} x &= \varrho \cdot \cos\varphi \\ y &= \varrho \cdot \sin\varphi \\ J &= \varrho \end{aligned}$$

dostávame:

$$\begin{aligned} I^2 = \int_0^{+\infty} \int_0^{2\pi} \exp\left(-\frac{\varrho^2}{2}\right) \cdot \varrho \, d\varphi d\varrho = 2\pi \cdot \int_0^{+\infty} \exp\left(-\frac{\varrho^2}{2}\right) \cdot \varrho \, d\varrho = \\ = 2\pi \cdot \int_0^{+\infty} \exp(-t) \, dt = 2\pi \cdot \Big[\exp(-t)\Big]_0^{+\infty} = 2\pi \end{aligned}$$

Tým sme ukázali, že $I^2 = 2\pi$. Pretože sme integrovali kladnú funkciu, aj jej itegrál musí byť kladný. Teda platí:

$$I = \sqrt{2\pi}$$

Po dosadení do (3) máme:

$$A \cdot \sigma \cdot \sqrt{2\pi} = 1$$

Odtiaľ vyjadríme

$$A = \frac{1}{\sqrt{2\pi} \cdot \sigma}$$

Po dosadení do (2) dostávame:

$$f(x) \doteq \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left[-\frac{1}{2} \cdot \frac{(x-m)^2}{\sigma^2}\right]$$

Tento vzorec je základom nasledujúcej definície.

**Definícia.** Hovoríme, že spojitá náhodná veličina $X$ má *normálne rozdelenie*, ak jej hustota pravdepodobnosti $f$ má tvar

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left[-\frac{1}{2} \cdot \frac{(x-m)^2}{\sigma^2}\right]$$

pričom $m$ a $\sigma^2$ (kde $\sigma > 0$) sú *parametre* rozdelenia.

Normálne rozdelenie má mimoriadny význam v matematickej štatistike. Typickým príkladom sú náhodné chyby merania.

Význam normálneho rozdelenia je predovšetkým v tom, že za istých predpokladov dobre aproximuje celý rad iných rozdelení pravdepodobnosti.

Zápis

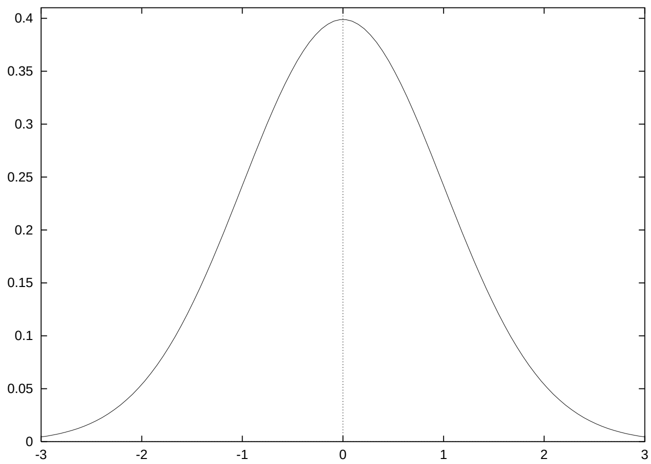
$$X \sim N(m, \sigma^2)$$

čítame: „Náhodná veličina $X$ má normálne rozdelenie s parametrami $m$ a $\sigma^2$.“

Rozdelenie $N(0,1)$ sa nazýva *normované normálne rozdelenie.* Hustota pravdepodobnosti normovaného normálneho rozdelenia sa volá *Gaussova krivka.* Na jej označovanie je vyhradený symbol $\varphi$. Teda platí:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right)$$

Graf Gausovej krivky vidíme na obrázku 19.

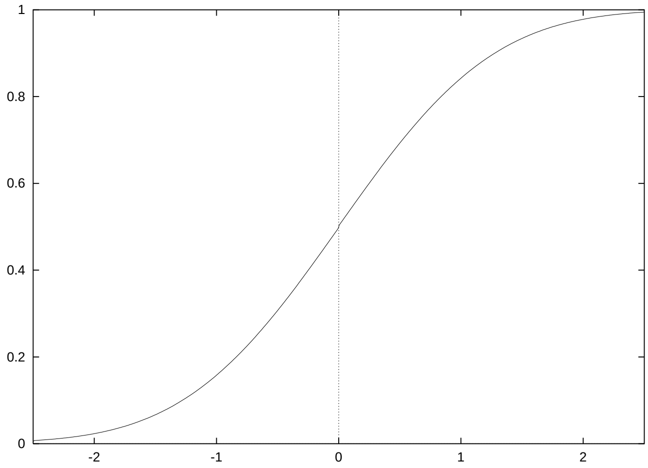



Obr. 19

Distribučná funkcia normovaného normálneho rozdelenia sa volá *Laplaceova funkcia.* Na jej označovanie je vyhradený symbol $\Phi$. Teda platí:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt$$

Graf Laplaceovej funkcie je na obrázku 20.




Obr. 20

Funkcia $\Phi$ nie je elementárnou funkciu, nedá sa vyjadriť vzorcom. Z tohto dôvodu sa v minulosti používali (a ešte niekedy aj v súčasnosti používajú) tabuľky tejto funkcie. Ďalšou možnosťou je použitie niektorej z metód numerickej matematiky.

Pomocou funkcie $\Phi$ môžeme počítať hodnoty distribučnej funkcie rozdelenia $N(m, \sigma^2)$. Skutočne, použitím substitúcie $(t-m)/\sigma = r$ dostávame:

$$F(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\cdot\frac{(t-m)^2}{\sigma^2}\right] dt =$$

$$= \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot \int_{-\infty}^{(x-m)/\sigma} \exp\left(-\frac{r^2}{2}\right)\sigma dr = \Phi\left(\frac{x-m}{\sigma}\right)$$

Tým sme odvodili vzorec:

$$F(x) = \Phi\left(\frac{x-m}{\sigma}\right)$$

K ďalším užitočným vzorcom patrí:

$$\Phi(-x) = 1 - \Phi(x)$$

Vďaka nemu stačí mať k dispozícii tabuľky funkcie $\Phi$ len pre kladné hodnoty $x$. Pozrime sa na odvodenie tohto vzorca. Zo symetrie grafu funkcie $\varphi$ vyplýva, že pre každé reálne číslo $x$ platí:

$$\varphi(-x) = \varphi(x)$$

Potom pomocou substitúcie $t = -r$ dostávame:

$$\Phi(-x) = \int_{-\infty}^{-x} \varphi(t)\,dt = -\int_{+\infty}^{x} \varphi(-r)\,dr = \int_{x}^{+\infty} \varphi(-r)\,dr = \int_{x}^{+\infty} \varphi(r)\,dr =$$

$$= \int_{-\infty}^{+\infty} \varphi(r)\,dr - \int_{-\infty}^{x} \varphi(r)\,dr = 1 - \int_{-\infty}^{x} \varphi(r)\,dr = 1 - \Phi(x)$$

Analógiou aritmetického priemeru je pri náhodných veličinách ich *stredná hodnota.* V prípade spojitej náhodnej veličiny $X$ definujeme jej strednú hodnotu pomocou hustoty pravdepodobnosti $f$ nasledujúcim spôsobom:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$$

**Veta.** *Ak $X \sim N(m, \sigma^2)$, potom $E(X) = m$.*

*Dôkaz.* Použitím substitúcie $(x - m)/\sigma = t$ dostávame:

$$\begin{aligned} E(X) &= \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \int_{-\infty}^{+\infty} x \cdot \exp\left[-\frac{1}{2} \cdot \frac{(x-m)^2}{\sigma^2}\right] dx = \\ &= \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \int_{-\infty}^{+\infty} (\sigma t + m) \cdot \exp\left(-\frac{t^2}{2}\right) \cdot \sigma dt = \\ &= \frac{\sigma}{\sqrt{2\pi}} \cdot \underbrace{\int_{-\infty}^{+\infty} t \cdot \exp\left(-\frac{t^2}{2}\right) dt}_{=0} + \frac{m}{\sqrt{2\pi}} \cdot \underbrace{\int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{2}\right) dt}_{=\sqrt{2\pi}} = m \end{aligned}$$

*Disperziu* spojitej náhodnej veličiny $X$ definujeme pomocou jej hustoty pravdepodobnosti $f$ nasledujúcim spôsobom:

$$D(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 \cdot f(x)\, dx$$

**Veta.** *Ak $X \sim N(m, \sigma^2)$, potom $D(X) = \sigma^2$.*

*Dôkaz.* Použitím substitúcie $(x - m)/\sigma = t$ dostávame:

$$\begin{aligned} D(X) &= \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \int_{-\infty}^{+\infty} (x-m)^2 \cdot \exp\left[-\frac{1}{2} \cdot \frac{(x-m)^2}{\sigma^2}\right] dx = \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} t^2 \cdot \exp\left(-\frac{t^2}{2}\right) dt = \sigma^2 \end{aligned}$$

Pritom posledný integrál sme vypočítali metódou per partes. Skutočne, ak

$$\begin{aligned} u &= t \quad & v' &= t \cdot \exp\left(-\tfrac{t^2}{2}\right) \\ u' &= 1 \quad & v &= -\exp\left(-\tfrac{t^2}{2}\right) \end{aligned}$$

potom

$$\int_{-\infty}^{+\infty} t^2 \cdot \exp\left(-\frac{t^2}{2}\right) dt = \underbrace{\left[-t \cdot \exp\left(-\frac{t^2}{2}\right)\right]_{-\infty}^{+\infty}}_{=0} + \underbrace{\int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{2}\right) dt}_{=\sqrt{2\pi}} = \sqrt{2\pi}$$

Pomocou distribučnej funkcie môžeme vypočítať pravdepodobnosť toho, že náhodná veličina nadobudne hodnoty v danom intervale.

Skutočne, ak $F$ je distribučná funkcia náhodnej veličiny $X$, potom náhodný jav $\{a \leqq X < b\}$ vyjadríme v tvare rozdielu náhodných javov:

$$\{a \leqq X < b\} = \{X < b\} - \{X < a\}$$

odkiaľ dostávame:

$$P(a \leqq X < b) = P(X < b) - P(X < a) = F(b) - F(a)$$

Dôsledok tohto vzorca (vzhľadom na jeho dôležitosť) uvedieme v tvare vety.

**Veta.** *Ak* $X \sim N(m, \sigma^2)$, *potom platí:*

$$P(a \leqq X < b) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right)$$

Náplňou nasledujúcej vety bude tzv. *pravidlo troch sigma.* Táto veta hovorí, že náhodná veličina s normálnym rozdelením nadobúda 99.73% svojich hodnôt v intervale, ktorého krajné body sú čísla $m - 3\sigma$, $m + 3\sigma$.

**Veta.** *Ak* $X \sim N(m, \sigma^2)$, *potom platí:*

$$P(m - 3\sigma \leqq X < m + 3\sigma) = 2\Phi(3) - 1 \doteq 0.99730$$

*Dôkaz.* Podľa predchádzajúcej vety máme:

$$\begin{aligned} P(m - 3\sigma \leqq X < m + 3\sigma) &= \Phi\left(\frac{(m+3\sigma)-m}{\sigma}\right) - \Phi\left(\frac{(m-3\sigma)-m}{\sigma}\right) = \\ &= \Phi(3) - \Phi(-3) = \Phi(3) - [1 - \Phi(3)] = 2\Phi(3) - 1 \end{aligned}$$

Už stačí iba vyhľadať v tabuľkách hodnotu $\Phi(3) \doteq 0.99865$ a dokončiť výpočet:
$2\Phi(3) - 1 \doteq 0.99730$.

## Weibullovo rozdelenie

V technickej praxi sa môžeme stretnúť aj s inými typmi rozdelení pravdepodobnosti náhodných veličín. Napríklad pri popise životnosti zariadení, ktorá je ovplyvnená mechanickým opotrebovaním a únavou materiálu, sa používa Weibullovo rozdelenie.

**Definícia.** Náhodná veličina $X$ má Weibullovo rozdelenie, ak jej distribučná funkcia má tvar:

$$F(x) = \begin{cases} 1 - \exp\left(-\left(\dfrac{x-c}{a}\right)^b\right) & \text{pre } x > c \\ 0 & \text{pre } x \leqq c \end{cases}$$

### Exponenciálne rozdelenie

Weibullovo rozdelenie s parametrom $b = 1$ sa volá exponenciálne rozdelenie.

**Definícia.** Náhodná veličina má *exponenciálne rozdelenie*, ak jej distribučná funkcia má tvar:

$$F(x) = \begin{cases} 1 - \exp\left(-\left(\dfrac{x-c}{a}\right)\right) & \text{pre } x > c \\ 0 & \text{pre } x \leqq c \end{cases}$$

Parameter $c$ predstavuje dolnú hranicu rozdelenia. V teórii životnosti a spoľahlivosti vyjadruje tzv. *záručnú hodnotu.*

### Binomické rozdelenie

Výrobky pri danom výrobnom procese rozdeľujeme na chybné a bezchybné. Pravdepodobnosť toho, že bude vyrobený chybný výrobok, je vyjadrená číslom $p$, pre ktoré platí $0 < p < 1$. Potom

$$P(\text{výrobok bude chybný}) = p$$
$$P(\text{výrobok bude bezchybný}) = 1 - p$$

Počet chybných výrobkov medzi $n$ nezávisle vyrobenými výrobkami je náhodná veličina $X$, ktorá nadobúda hodnoty $k = 0, 1, 2, \ldots, n$. Zápis $\{X = k\}$ označuje náhodný jav, že medzi $n$ nezávisle vyrobenými výrobkami bude presne $k$ chybných. Pravdepodobnosť javu $\{X = k\}$ vypočítame podľa vzorca:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Túto situáciu môže popísať všeobecnejšie:
Budeme realizovať $n$ nezávislých náhodných pokusov. V každom pokuse môže nastať *úspech* s pravdepodobnosťou $p$, alebo *neúspech* s pravdepodobnosťou $1 - p$. Uvažujme náhodnú veličinu $X$, ktorá vyjadruje celkový počet úspechov pri $n$ pokusoch.

**Definícia.** Hovoríme, že náhodná veličina $X$ sa riadi *binomickým rozdelením* s parametrami $n$ a $p$, ak pre každé $k = 0, 1, 2, \ldots, n$ platí:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Zápis

$$X \sim Bi(n, p)$$

čítame: „Náhodná veličina $X$ má binomické rozdelenie s parametrami $n$ a $p$."

### Hypergeometrické rozdelenie

Majme $N$ predmetov, z ktorých presne $M$ má sledovanú vlastnosť. Z týchto $N$ predmetov vyberieme náhodne (bez vrátenia) $n$ predmetov. Počet tých predmetov medzi nimi, ktoré majú sledovanú vlastnosť, je náhodná veličina $X$. Zápis $\{X = k\}$ označuje náhodný jav, že medzi týmito $n$ predmetmi bude presne $k$ predmetov majúcich sledovanú vlastnosť. Pravdepodobnosť javu $\{X = k\}$ vypočítame podľa vzorca:

$$P(X = k) = \frac{\dbinom{M}{k} \cdot \dbinom{N-M}{n-k}}{\dbinom{N}{n}}$$

**Definícia.** Hovoríme, že náhodná veličina $X$ sa riadi *hypergeometrickým rozdelením* s parametrami $N$, $M$ a $n$, ak platí:

$$P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$$

pre každé $k = 0, 1, 2, \ldots, n$, pre ktoré tento vzorec má zmysel,
t. j. ktoré spĺňa podmienky: $k \leqq M$ a $n - k \leqq N - M$.

### Poissonovo rozdelenie

Týmto typom rozdelenia sa riadi počet výskytu udalosti za časovú jednotku – napríklad počet porúch zariadenia za 100 hodín prevádzky.

**Definícia.** Hovoríme, že náhodná veličina $X$ sa riadi *Poissonovým rozdelením* s parametrom $\lambda$, ak pre každé prirodzené číslo $k$ platí:

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$