Matematická štatistika v geológii

Peter Blišťan [1]

Mathematical Statistics in the Geology

During the last period is modern geology oriented toward intensive utilisation of mathematical methods. Utilisation of these methods was conditioned by complicated structure of geological bodies, which resulted from interaction of a couple of factors. In the period of projection, realisation and evaluation of geological works one meet many problems of description of a character of geological data. These problems – very often trivial – arise from the poor knowledge of the principles of statistical methods. Transformation of real geological object into the form of abstract mathematical model is the basic and usually also the most complicated step of mathematical solution. There is also to be mentioned that there is no unitary approach to the modelling or uniform direction for a method of data processing selection. Complicacy of geological objects needs rational simplification of the model, otherwise the solution would be too complicated or even impossible.

***Key words**: histogram, normal distribution, lognormal distribution, exponential distribution, mean value, median, mode, sample variance, standard deviation, correlation analysis.*

1. Úvod

Široké uplatnenie matematických metód v geovedách vyplýva zo skutočnosti, že v procese formovania geologických telies dochádzalo interakciou celého radu faktorov k vzniku veľmi zložitých systémov. Tieto sa vyznačujú určitou priestorovou štruktúrou a zastúpením náhodnej zložky. Pri projektovaní, vykonávaní a vyhodnocovaní geologických prác sa stretávame s problémami popisu charakteru priestorových geologických dátových veličín, ktoré môžu byť spojité, diskrétne, skalárne alebo vektorové.

Základným a často aj najťažším krokom pri matematickom riešení geologických problémov je prevod reálnych geologických objektov do formy abstraktných matematických modelov. V zásade môžeme odlíšiť:

- priestorové modelovanie morfológie geologických telies,
- modelovanie vnútorných atribútov.

Tu je potrebné zdôrazniť, že neexistuje jednotný postup modelovania ani jednoznačné pravidlá výberu vhodných metód spracovania dát. Mnoho matematicko-štatistických postupov vyžaduje splnenie striktných predpokladov pre ich správnu aplikáciu, tie sa však u prírodných objektov nie vždy dajú overiť. Zložitosť geologických objektov teda vyžaduje rozumné zjednodušenie modelu, ináč by riešenie bolo príliš komplikované alebo dokonca nemožné. Tento jav vyjadruje aj princíp inkompatibility: " Ak rastie zložitosť systému, klesá naša schopnosť formulovať presné závery o jeho správaní, až dosiahneme hranice, za ktorými sú presnosť a relevantnosť prakticky vzájomne vylučujúce sa charakteristiky. „ (Zádech, 1973). Skutočnosťou je aj to, že reálny efekt z aplikovania zložitých modelov nie je často úmerný vynaloženej práci.

2. Princípy štatistickej analýzy dát

Cieľom štatistickej analýzy je stanovenie zvláštností dát z kvantitatívneho hľadiska a overenie predpokladov pre ďalšie štatistické úvahy o dátach. Z hľadiska aplikácie matematicko-štatistických metód je možné vyčleniť určité typy úloh a postupy, ktoré sú spoločné pre rôzne druhy riešených problémov. Riešenie týchto typových úloh spracovania dát je možné rozdeliť do niekoľkých chronologicky za sebou nasledujúcich krokov:

- popis typu štatistickej distribúcie veličín, resp. hustoty rozdelenia pravdepodobnosti ich výskytu,
- najpravdepodobnejší odhad stredných hodnôt veličín,
- určenie variability veličín,
- stanovenie spoľahlivosti odhadu matematickej nádeje,
- stanovenie spoľahlivosti odhadu disperzie,
- riešenie problému odľahlých – extrémnych veličín,
- hodnotenie nezávislosti výberov,
- ocenenie a popis vzájomných vzťahov veličín,
- analýza štruktúry polí veličín,
- určenie priemerných charakteristík polí veličín,
- testovanie zhody alebo rozdielu geologických objektov (Schejbal, 1980).

[1] *Ing. Peter Blišťan*, Katedra geológie a mineralógie F BERG Technickej univerzity, Park Komenského 15, 043 84 Košice, e-mail: blistan@tuke.sk

(Recenzovali: *Doc. Ing. Ľubica Floreková, CSc.* a *Ing. Anton Grinč)*

Poslednými tromi krokmi sa zaoberá predovšetkým špeciálna matematická disciplína – geoštatisika, vychádzajúca z priestorových vzťahov meraných dát.

Okrem uvedených metód existuje ešte celý rad ďalších štatistických, geoštatistických, deterministických, stochastických a iných metód, ktoré majú špeciálne zameranie alebo sú súčasťou iných odborných disciplín.

2.1. Prieskumová analýza dát

Účelom prieskumovej analýzy (predspracovania) je odhaliť zvláštnosti získaných dát a overiť predpoklady pre ich ďalšie štatistické spracovanie. Opodstatnenie má predovšetkým pri analýze zložitých, nákladných alebo unikátnych meraní, pretože môže zabrániť vzniku chýb ešte pred začatím detailnejších štatistických výpočtov, alebo na základe svojich záverov neodporučiť realizáciu ďalších meraní. Z týchto dôvodov sa úspešne používa aj v geológii, kde môže odhaliť, alebo do istej miery predpokladať negatívny výsledok ďalších prieskumných prác ešte v ich počiatku. Zabráni sa tým v konečnom dôsledku finančným stratám, ktoré by vznikli pri realizácii vyšších etáp prieskumu.

Pri tejto analýze sa používajú hlavne rýchle a nenáročné metódy a postupy. Medzi najčastejšie patria predovšetkým grafické metódy, ktoré sú vhodné na zjednodušenie popisu dát, identifikáciu typu rozdelenia dát a konštrukciu takzvaného empirického rozdelenia dát. Sú jednoduché a relatívne najrýchlejšie, preto majú iba orientačný charakter, a to hlavne pri zmiešaných distribúciách (zvláštny typ rozdelenia dát). Sú založené na používaní tzv. pravdepodobnostných grafických papierov, do ktorých sa vynášajú napríklad triedne znaky a zodpovedajúce kumulatívne relatívne početnosti, respektíve iné parametre súboru dát, takže dávajú okamžitý vizuálny obraz.

Grafy identifikácie štatistických zvláštností dát

Z grafických metód sa pri prieskumovej analýze využívajú predovšetkým diagramy rozptýlenia hodnôt a kvantilová charakteristika súboru, umožňujúca sledovať lokálne správanie sa dát (koncentrácia dát, extrémne hodnoty a pod.).

Diagram rozptýlenia hodnôt

Diagram rozptýlenia hodnôt (obr.1a) predstavuje jednorozmernú projekciu na X-ovú alebo Y-ovú os. Je vhodný na rýchle a prehľadné určenie lokálnej koncentrácie dát, ktorá sa prejaví lokálnymi zhlukmi bodov v grafe. Výskyt tohto javu pri spracovaní dát poukazuje na koncentráciu meraných dát okolo jednej, prípadne viacerých hodnôt, ktoré môžu byť v ideálnom prípade očakávanými hodnotami. Ak teda namerané hodnoty korešpondujú s očakávanými, znamená to, že merania boli správne, resp., že sa potvrdil výskyt očakávaných hodnôt.

Diagram dokáže ľahko odhaliť aj vybočujúce a extrémne hodnoty, ležiace na koncoch grafu. Tieto môžu predstavovať náhodné chybné merania (chybné vzorky, preklepy pri spracovaní), alebo naopak, vysokú (nízku) koncentráciu sledovaného javu, ktorá poukazuje na prípadné objavenie hľadaného ložiska (napríklad vysoko pozitívne šlichové vzorky pri šlichovej prospekcii na zlato), alebo na jeho neexistenciu (vykliňovanie, „hluché“ – jalové úseky).

Kvantilový graf

Predstavuje dvojrozmernú projekciu variačného radu (na Y-ovú os) a poradovej pravdepodobnosti (na X-ovú os). Hodnoty poradovej pravdepodobnosti P_i sa určia podľa vzťahu [1], v ktorom ***n***=počet hodnôt výberu a ***i*** je poradové číslo príslušnej hodnoty vo variačnom rade(Meloun a Militký, 1995). Variačný rad vznikne vzostupným usporiadaním nameraných hodnôt. Graf môže byť symetrický, zošikmený k vysokým alebo nízkym hodnotám. Umožňuje prehľadne znázorniť dáta a na základe krivky ľahšie určiť tvar rozdelenia (obr.1b) ktorý je dôležitý pre výber správnej metódy určenia základných charakteristík súboru.

$$P_i = \frac{i - \frac{1}{3}}{n + \frac{1}{3}} \qquad [1]$$

Pre symetrické rozdelenia má kvantilová funkcia sigmoidálny tvar. Pre rozdelenia zošikmené k vyšším hodnotám je konvexne rastúca a pre rozdelenia zošikmené k nižším hodnotám konkávne rastie. Je z neho možné identifikovať lokálne koncentrácie dát a vybočujúce, respektíve extrémne hodnoty podobným spôsobom ako z diagramu rozptýlenia hodnôt.

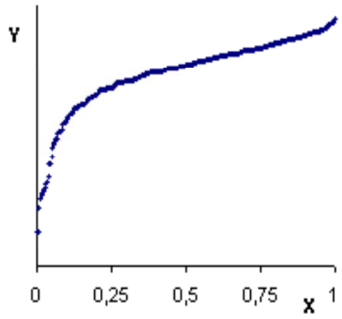
Diagram rozptýlenia s kvantilmi

Vyniká jednoduchou konštrukciou a univerzálnosťou posudzovania štatistických zvláštností súboru dát. Jeho základom je kvantilový graf, ktorý sa získa spojením bodov $\{X_i , P_i\}$ lineárnymi úsekmi (obr.1c). Hodnota poradovej pravdepodobnosti P_i sa počíta podľa vzťahu [1].

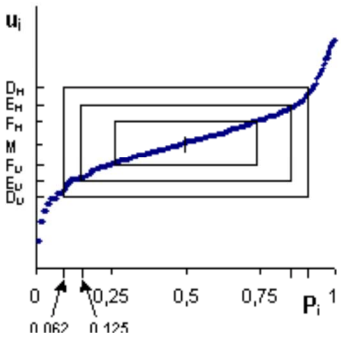
Pre uľahčenie interpretácie sa do grafu vynášajú tri obdĺžniky F, E a D.

- *Kvartilový obdĺžnik F* – má na súradnici Y vrcholy dané hodnotami kvartilov $F_D \equiv X_{25}$ a $F_H \equiv X_{75}$. Odpovedajúce súradnice na osi X sú poradové pravdepodobnosti $P_2=2^{-2}=0{,}25$ a $1-2^{-2}=0{,}75$.
- *Oktilový obdĺžnik E* – má na súradnici Y oktily $E_D \equiv X_{12,5}$ a $E_H \equiv X_{87,5}$ a na osi X poradové pravdepodobnosti $P_3=2^{-3}=0{,}125$ a $1-2^{-3}=0{,}875$.
- *Sedecilový obdĺžnik D* – má na súradnici Y sedecily $D_D \equiv X_{6,25}$ a $D_H \equiv X_{93,75}$ a na osi X poradové pravdepodobnosti $P_4=2^{-4}=0{,}0625$ a $1-2^{-4}=0{,}9375$.

Obr.1a. Diagram rozptýlenia hodnôt.



Obr.1b. Kvantilový graf.



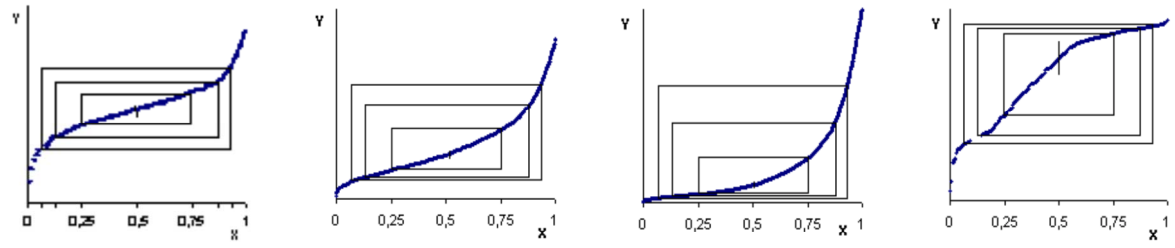
Obr.1c. Diagram rozptýlenia s kvantilmi.

$$\tilde{x}_{0,5} \pm 1{,}75 R_F / \sqrt{n} \qquad [2]$$

Do kvantilového obdĺžnika sa vo výške mediánu $\tilde{x}_{0,5} \equiv X_{50}$ zakresľuje horizontálna čiara a na kolmici k nej sa na os X vynáša robustný odhad konfidenčného intervalu mediánu kde $R_F=F_H-F_D=IQR$ (inter quartil range). Na základe tohoto grafu, rozpätia R_L, relatívnej šikmosti S_L a relatívnej dĺžky koncov variačného radu T_L je možné identifikovať zvláštnosti výberu (obr.2):

- *Symetrické unimodálne rozdelenie výberu* obsahuje jednotlivé kvantilové obdĺžniky symetricky vo vnútri seba a hodnoty šikmosti $S_L \approx 0$. Ak sú dĺžky koncov T_L približne rovné svojím teoretickým dĺžkam, je možné rozlíšiť rozdelenie normálne, Laplaceovo (dlhé konce) a rovnomerné (krátke konce).
- *Nesymetrické rozdelenia* majú pre rozdelenia zošikmené k vyšším hodnotám vzdialenosti medzi dolnými hranami obdĺžnikov F, E a D výrazne kratšie ako medzi ich hornými hranami. Hodnoty koeficientov šikmosti sú potom záporné (normálne rozdelenie asymetrické vpravo), zatiaľ čo rozdelenia zošikmené k nižším hodnotám (logaritmickonormálne a exponenciálne rozdelenie) majú koeficienty šikmosti kladné (obr.3a,b a obr.5).
- *Odľahlé pozorovania* sú indikované tým, že na kvantilovej funkcii sa mimo obdĺžnika F objaví náhly vzrast, hodnota smernice výrazne vzrastie. *Viacmodálne rozdelenia* sú indikované tým, že kvantilová funkcia vo vnútri obdĺžnika F má v niekoľkých úsekoch nulovú smernicu (Meloun a Militký, 1995).

Zostrojením diagramu rozptýlenia s kvantilmi získame jeden z prvých ukazovateľov charakteru rozdelenia dát v súbore. Na príkladoch grafov z obrázku 2 je možné podľa tvaru krivky a obdĺžnikov jednoznačne určiť, či sa jedná o symetrické rozdelenie a v prípade nesymetrického aj smer vychýlenia.



a.) b.) c.) d.)

Obr.2. Diagramy rozptýlenia s kvantilmi pre výbery z rozdelenia: a) normálneho, b) lognormálneho, c) exponenciálneho a d) dvojvrcholového normálneho.

Overenie homogenity súboru

Pri posudzovaní výsledkov pozorovaní sa často stretávame s výskytom hodnôt, nápadne sa odlišujúcich od ostatných. Sú to hodnoty extrémne a v štatistike sa často označujú ako odľahlé pozorovania. V ložiskovej

geológii tieto hodnoty spôsobujú veľké problémy pri výpočte zásob variabilných ložísk, predovšetkým farebných a drahých kovov.

Spôsoby spracovania extrémnych hodnôt je možné rozdeliť na :

- *empirické* - kedy sú extrémne hodnoty nahradzované napr. strednou hodnotou, priemerom susedných hodnôt, alebo najvyššou prijateľnou normálnou hodnotou, poradovými – znamienkovými hodnotami, mediánovými hodnotami, resp. robustnou štatistikou,
- *matematicko-štatistické* - kde sú extrémne hodnoty na základe rôznych kritérií a testov vyčleňované zo súboru hodnôt a ďalej sa s nimi nepracuje. Z nich najznámejšie sú parametrický *Grubbsov test* a neparametrický *Dixonov test*, pre testovanie malých súborov.

Konštrukcia usporiadania dát pre identifikáciu vhodného typu rozdelenia dát

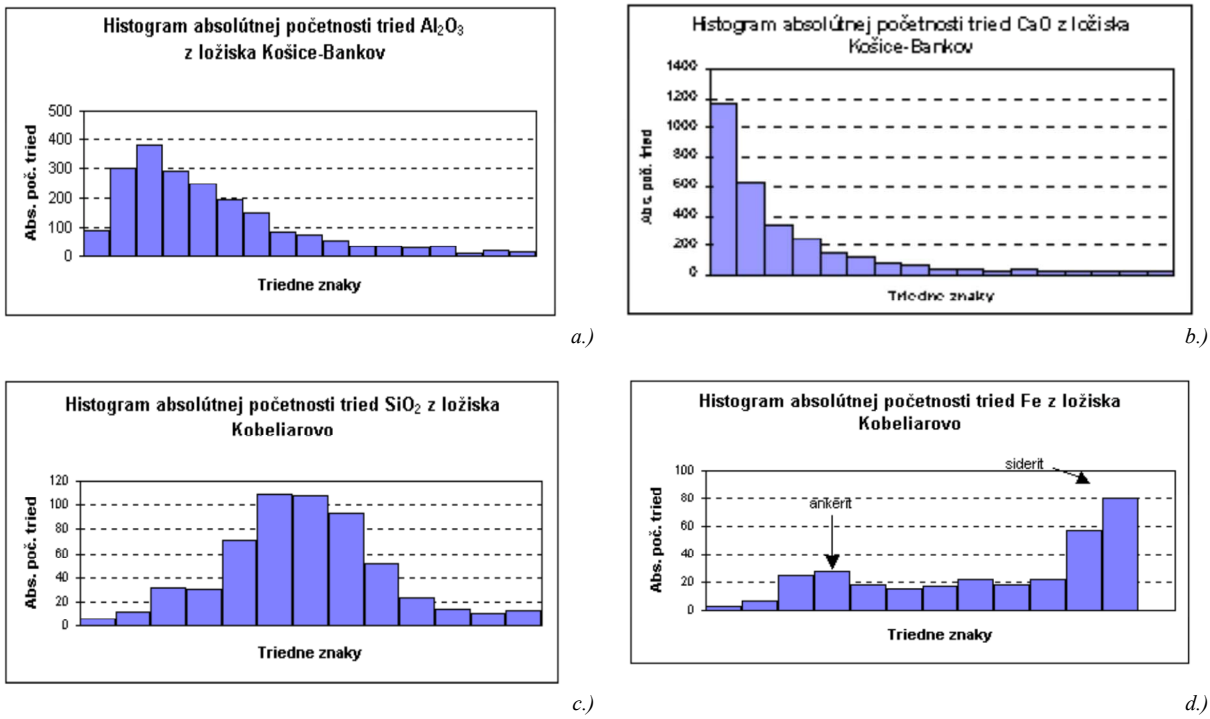
Histogram

Pre vizuálne posúdenie možného rozdelenia dát je v prípade dostatočne veľkého súboru vhodné vytvoriť histogram. Je jedným z najstarších klasických odhadov hustoty pravdepodobnosti, slúžiacich predovšetkým na vizualizáciu dát. Pre zjednodušenie výpočtov sa súbor rozdelí na intervaly, obyčajne rovnako veľké, nazývané *triedy*. Triedne rozdelenie početností sa potom graficky zobrazuje ich *histogramom* – stĺpcovým grafom (obr.3), alebo *polynómom* – čiarovým grafom. Počet stĺpcov v grafe je rovný počtu tried ***k***, šírka stĺpca predstavuje šírku triedy ***h*** a výška predstavuje absolútnu alebo relatívnu početnosť triedy. Vypovedacia schopnosť histogramu je silne závislá na správnej voľbe počtu tried, ktorý by mal ležať v intervale 7 až 20 (Schejbal, 1980), resp. 8 až 15 (Bučko, 1981). Pre stanovenie šírky a počtu tried sa používajú napríklad nasledujúce vzťahy (Bučko, 1981; Riečanová, 1987):

- na základe variačného rozpätia $R=u_{max}-u_{min}$, $h \cong 0{,}08.R$ [3], $k = \frac{R}{h}$ [4],

- na základe rozsahu súboru

$k = 5.log(n)$ [5], $k \doteq 1 + 3{,}30.log(n)$ [6], $k \doteq \sqrt{n}$ [7], potom $h = \frac{R}{k}$ [8].



Obr. 3. Histogramy absolútnej početnosti tried: a) lognormálneho, b) exponenciálneho, c) normálneho rozdelenia a d) dvojvrcholového normálneho rozdelenia.

Ďalším krokom triedenia je nahradenie dát v každej triede hodnotou *triedneho znaku* u_j (stred tridy), určeného podľa vzťahu $u_j = \frac{(u_d + u_h)}{2}$ [9],

kde: u_d je dolná hranica a u_h horná hranica triedy, pričom počet hodnôt v triede vyjadruje *absolútna početnosť triedy* $\boldsymbol{n_j}$ (počet hodnôt, ktoré sa nachádzajú v úseku vymedzenom dolnou a hornou hranicou triedy), pričom $\sum_j n_j = n$. *Kumulatívna absolútna početnosť* triedy $\boldsymbol{N_j}$ je potom

$$N_j = \sum_{j=1}^{k} n_j \qquad [10], \qquad N_k = n \ .$$

Relatívna početnosť $\boldsymbol{f_j}$ a *kumulatívna relatívna početnosť* triedy $\boldsymbol{F_j}$ sa určia nasledovne

$$f_j = \frac{n_j}{n} \qquad [11], \qquad \sum_j f_j = 1 \ , F_j = \frac{N_j}{n} \qquad [12], \qquad F_k = n \ .$$

Všeobecne sa dá povedať, že pre výbery v ktorých sa predpokladá normálne rozdelenie, stačí použiť histogram s menším počtom tried a naopak, pre zložité rozdelenia dát je potrebné zväčšiť počet triednych intervalov, alebo použiť niektoré špeciálne postupy hľadania triednych intervalov. Odporúča sa tiež voľba nepárneho počtu tried, aby sa dala rýchlo z histogramu $\boldsymbol{n_j}$ určiť poloha $\overline{u}$, prípadne $\widetilde{u}$ a $\hat{u}$ (Floreková a Benková, 1999). Na obrázku 3 sú príklady troch základných typov unimodálnych, čiže jednovrcholových rozdelení (lognormálneho, exponenciálneho a normálneho). Ak pri konštrukcii histogramu získame graf s dvomi vrcholmi, je rozdelenie bimodálne, dvojvrcholové, ktoré patrí k špeciálnym typom rozdelení dát. Takýto graf, zostrojený napríklad z chemických analýz zásekových vzoriek z ložiska (obr. 3 – *d*), môže poukazovať na dve generácie mineralizácie, respektíve na viazanie sa toho istého prvku v dvoch rôznych mineráloch (Fe v ankerite a siderite).

Transformácia dát

V praxi sa často stáva, že rozdelenie dát je značne odlišné od normálneho rozdelenia, čím vzniká problém pri jeho vyhodnocovaní. Tento problém je možné v mnohých prípadoch odstrániť vhodnou transformáciou, ktorá zaistí stabilizáciu rozptylu, symetrickosť rozdelenia a niekedy aj normalitu. Vychádza z predstavy, že spracované dáta sú nelineárnou transformáciou normálne rozdelenej náhodnej veličiny *x*. Úlohou transformácie je teda nájsť inverznú funkciu *g(u)*. Ak sa podarí nájsť vhodnú transformáciu, ktorá vedie k približnej normalite, je možné určiť $\overline{u}$, $s^2(\overline{u})$ a intervaly ich spoľahlivosti. Typickým príkladom je lognormálne rozdelenie, ktorého logaritmickou transformáciou získame normálne rozdelenie. Transformácia je veľmi jednoduchá, spočíva v zlogaritmovaní každej hodnoty, čím získame nový súbor dát s normálnym rozdelením. S takto upraveným súborom je potom možné pracovať ako s bežným súborom s normálnym rozdelením.

Určenie minimálnej veľkosti súboru dát

Rozsah súboru do značnej miery ovplyvňuje presnosť odhadu parametrov polohy a rozptýlenia. Z toho vyplýva, že ak chceme získať čo najpresnejšie a najreprezentatívnejšie výsledky, musíme realizovať dostatočné množstvo pozorovaní. Stanovenie rozsahu súboru je možné vykonať niekoľkými spôsobmi. Jedným z nich môže byť splnenie požadovaných kritérií spoľahlivosti niektorých vybraných charakteristík súboru. Minimálny rozsah výberu je potom možné určiť tak, aby napríklad relatívna chyba smerodajnej odchýlky δ *(s)* mala predpísanú hodnotu. Potom je minimálny rozsah výberu

$$n_{\min} = \frac{g_2(u) - 1}{4\delta(s)^2} + 1 \qquad [13],$$

kde g_2 je špicatosť rozdelenia dát. Relatívna chyba smerodajnej odchýlky sa obyčajne volí δ(s)=0,1 (t.j. 10%). Podľa tohoto vzťahu platia pre rôzne typy rozdelení dát nasledovné minimálne rozsahy súborov (tab. č.1).

Tab.1. Tabuľka vybraných rozdelení s charakteristickou špicatosťou a odporúčaným minimálnym rozsahom.

Rozdelenie	***Špicatosť***	***Minimálny rozsah*** $\boldsymbol{n_{min}}$
rovnomerné	1,8	21
normálne	3,0	51
exponenciálne	6,0	126
lognormálne	15,0	351

Druhý, oveľa jednoduchší spôsob je odvodený z intervalu spoľahlivosti pre matematickú nádej

$$\overline{u} \pm \delta = \overline{u} \pm 1{,}96\frac{s}{\sqrt{n}} \qquad [14], \text{ resp. } \overline{u} \pm 2{,}58\frac{s}{\sqrt{n}} \qquad [15], \text{ takže pre}$$

α=0,05 je $n_{\min} = \left(\frac{1{,}96s}{\delta}\right)^2$, resp. α=0,01 je $n_{\min} = \left(\frac{2{,}58s}{\delta}\right)^2$.

2.2 Podrobná štatistická analýza jednorozmerných dát

Po prieskumovej analýze nameraných dát nasleduje ďalšia etapa spracovania, ktorej úlohou je určiť štatistické charakteristiky súboru pozorovaní. Matematická štatistika rozlišuje pri aplikácii svojich metód :

- *malý súbor hodnôt,* zahŕňajúci menej ako 30 meraní,
- *veľký súbor hodnôt* zahŕňajúci viac ako 30 meraní.

Vzhľadom na fakt, že v geologickej praxi je najčastejšie potrebné spracovať veľké počty dát, budeme sa ďalej venovať hlavne princípom spracovania veľkého súboru dát.

Štatistické odhady parametrov veľkého súboru hodnôt

Vlastnosti odhadovaných parametrov

Bodový odhad $\boldsymbol{Q}$ parametru $\boldsymbol{\theta}$ súboru je číslo, ktoré sa považuje za aproximáciu neznámej hodnoty parametru. Je zaťažené istou chybou a dá sa očakávať, že jeho presnosť bude stúpať s veľkosťou súboru dát. Jeho hodnota sa pohybuje v takzvanom konfidenčnom intervale (intervale spoľahlivosti) <a,b>.

Odhad $\boldsymbol{Q}$ môžeme považovať za dobrý odhad parametru $\boldsymbol{\theta}$, ak spĺňa nasledujúce podmienky:

- *nestrannosť odhadu a asymptoticita nestrannosti*,
- *výdatnosť odhadu* (s veľkosťou súboru rastie pravdepodobnosť odhadu $\boldsymbol{Q}$),
- *robustnosť odhadu* (odhad je stabilný i pri nie úplne statických podmienkach, je odolný voči extrémom a chybám pri získavaní údajov).

Odhad parametrov polohy

Štatistický súbor je konečná množina pozorovaných hodnôt náhodnej premennej *(u_1 , u_2 , ... , u_{n-1} , u_n)*. Hodnoty súboru sa pri spracovaní usporiadajú do rastúcej postupnosti - variačného radu. *Variačné rozpätie* $\boldsymbol{R}$ je dĺžka intervalu < $u_{(1)}$, $u_{(n)}$ >, inak povedané, je to rozdiel najväčšej a najmenšej nameranej hodnoty. Pri spracovaní veľkého súboru sa namerané hodnoty rozdelia do jednotlivých tried podľa vypočítaných hraníc (výpočet sa realizuje podľa vzťahov 3 - 8). Určia sa hodnoty triednych znakov, relatívna a absolútna početnosť a jednotlivé kumulatívne početnosti (podľa vzťahov 9 - 12) a nakoniec sa zostrojí histogram početností. Tento postup je identický s postupom spracovania dát v prieskumovej analýze.

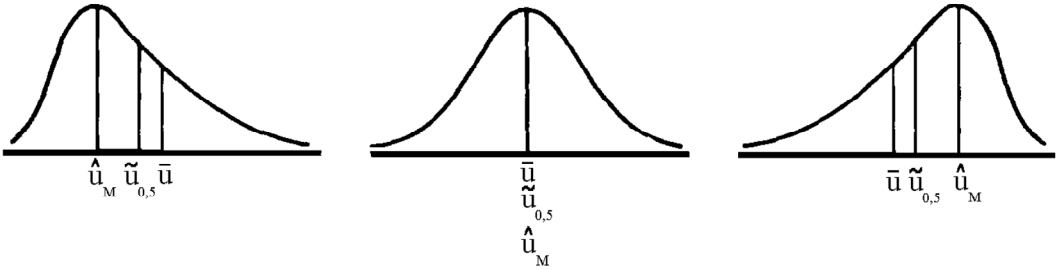
Potom nasleduje určenie základných parametrov polohy. *Aritmetický priemer*, ktorý je zároveň maximálnym vierohodným odhadom strednej hodnoty normálneho rozdelenia, ťažiskom štatistického súboru, sa určí podľa vzťahu

$$\bar{u} = \frac{\sum_{j=1}^{k} n_j . u_j}{\sum_{j=1}^{k} n_j} \qquad [16],$$

- *medián* $\tilde{u}$ je definovaný ako stredný prvok variačného radu, v prípade párneho počtu hodnôt je to priemer dvoch stredných prvkov variačného radu.
- *modus* je najčastejšie sa opakujúci prvok variačného radu.

Pre *malý súbor dát* s normálnym rozdelením platia rovnaké vzťahy, v ktorých n_j=*1,* u_j sa zamieňa za pôvodné dáta u_i. Aritmetický priemer rozsiahleho výberu je súčasne tzv. *vážený aritmetický priemer*, v ktorom sú váhy v_j rovné početnostiam n_j. Váhou môže byť napríklad aj dĺžka zásekovej vzorky, dĺžka vrtného jadra, alebo iné vhodne definované číslo, ktoré charakterizuje každú vzorku. Je nevyhnutné, aby súčet všetkých váhových konštánt bol rovný 1 (Floreková, 1986).

Vzťah medzi jednotlivými charakteristikami polohy vyjadruje obrázok č. 4. Akékoľvek vychýlenie rozdelenia od normality (normálneho rozdelenia dát) spôsobuje zníženie reprezentatívnosti strednej hodnoty - aritmetického priemeru, prejavujúce sa podhodnotením, alebo v horšom prípade nadhodnotením nameraných dát. Akceptovanie nereprezentatívnych výsledkov spracovania dát môže znamenať značné skreslenie očakávaných záverov (napríklad vypočítané zásoby v ložisku môžu byť niekedy aj o 100% väčšie ako reálne).



Obr.4. Vzťah medzi aritmetickým priemerom $\bar{u}$, mediánom $\tilde{u}_{0,5}$ a modusom $\hat{u}_M$ v prípade symetrického a nesymetrického rozdelenia (Curwin a Slater, 1996).

Správny odhad strednej hodnoty musí teda spĺňať definované podmienky pre dobrý odhad, inak ho nemôžeme považovať za zástupcu nameraných dát. Dôležitým znakom pre posúdenie správnosti, resp. významnosti odhadu strednej hodnoty súboru dát $\bar{u}$, vyplývajúcim z charakteru normálneho rozdelenia dát, je rovnosť mediánu, modusu a strednej hodnoty. Inak povedané, ak určíme aritmetický priemer, medián a modus v testovanom súbore dát a získame približne rovnakú hodnotu (Mn v tab. č.2), môžeme prehlásiť, že z hľadiska polohy ťažiska dát, je štatistický súbor symetrický, dáta majú normálne, alebo iné symetrické rozdelenie a vypočítaný priemer spĺňa podmienky dobrého odhadu (príklad odhadu niektorých parametrov súboru dát z ložiska Rožňava-Strieborná žila je v tabuľke č.2). V prípade, že sa vypočítané hodnoty od seba výrazne líšia (priemer, medián a modus sú zoradené za sebou, podobne ako na obr. 4 a tab. č.2) rozdelenie nie je normálne a priemer nemôžeme považovať za dobrý odhad strednej hodnoty. Potom je potrebné vhodnou transformáciou dát, alebo inou úpravou súboru (napr. odrezaním dolnej alebo hornej časti súboru), zaistiť normalitu rozdelenia, resp. hľadať iné vhodné rozdelenie, vizuálne zodpovedajúce histogramu n_j / N_j.

Tab. 2. Príklad určenia základných charakteristík súboru dát (Blišťan, 1995).

	Mocnosť [m]	*Ag [g/t]*	*Mn [%]*	*Fe [%]*
Vážený aritmetický priemer	-	193,47	1,55	-
Jednoduchý aritmetický priemer	2,01	172,9	1,50	31,89
Medián	1,6	102	1,56	37
Modus	0,4	0,1	1,82	40,1
Štandardná odchýlka	1,72	216,44	0,38	10,16
Rozptyl	2,95	46848	0,15	103,40
Špicatosť	3,29	6,302	3,59	-0,12
Šikmosť	1,72	2,32	-1,61	-1,03
Rozpätie	9,2	1386,8	2,17	44,8
Počet vzoriek	314	314	314	586

Odhad parametrov rozptýlenia

Okrem odhadu strednej hodnoty je potrebné poznať aj rozptyl hodnôt, ktorý vyjadruje do akej miery sú dáta rozptýlené okolo strednej hodnoty. Táto informácia je potrebná pre posúdenie variability dát a výpočet intervalov spoľahlivosti. Využíva sa aj pri vyčleňovaní anomálnych alebo odľahlých pozorovaní. Najvhodnejšou mierou rozptýlenia hodnôt je *rozptyl*, ktorý je v prípade normálneho rozdelenia

$$S^2 = \frac{1}{n}\sum_{j=1}^{k}\left(u_j - \bar{u}\right)^2 .n_j \quad \text{- pre n>30} \qquad [17].$$

Je to, ako vyplýva zo vzťahu, priemerná kvadratická odchýlka merania od strednej hodnoty (v ideálnom prípade =0). Čím je hodnota S^2 väčšia, tým je štatistický súbor variabilnejší, dáta sú nehomogénne, alebo zahŕňajú príliš veľa extrémnych vychýlených hodnôt.

Štandardná odchýlka $$S = \pm\sqrt{S^2} \qquad [18]$$

je odmocnina z odhadu rozptylu a keďže je v „rozmere“ dát, slúži za základ intervalu spoľahlivosti M(u) D(u) príslušného základného súboru. Pre porovnanie variability rôznorodých veličín sa používa relatívna charakteristika rozptýlenia hodnôt, nazvaná *koeficient variácie*

$$V = \frac{S}{\bar{u}} * 100 \quad [\%] \quad \text{- v prípade normálneho rozdelenia} \qquad [19],$$

Odhad parametrov tvaru

Parametre tvaru charakterizujú tvar hustoty pravdepodobnosti rozdelenia dát. Každý typ rozdelenia má svoje charakteristické parametre tvaru a preto je ich odhad dôležitým znakom pri určovaní typu rozdelenia získaných dát. Na obrázku č.5 sú príklady vzťahu základných parametrov tvaru, koeficientov šikmosti a špicatosti.

Koeficient šikmosti (asymetrie) – vyjadruje vychýlenie rozdelenia dát od symetrického rozdelenia (obr.5). Hodnota koeficienta sa určí podľa vzťahu

$$g_1 = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(u_i - \bar{u}\right)^3 .n_j}{S^{3/2}} \qquad [20].$$

Rozdelenia vychýlené vľavo majú kladný koeficient šikmosti a rozdelenia vychýlené vpravo majú záporný koeficient šikmosti. Na základe veľkosti a znamienka koeficienta šikmosti môžeme jednoznačne povedať, na ktorú stranu je experimentálne rozdelenie vychýlené a tak predpokladať jeho typ.

a.) *b.)*

Obr.5. Rozdelenie hodnôt s rôznymi koeficientmi šikmosti g_1 a špicatosti g_2 (Meloun a Militký, 1995).

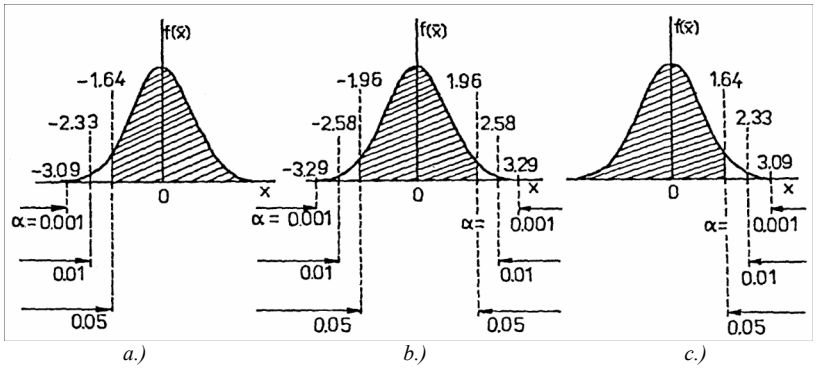
Koeficient špicatosti (ostrosti, excesu) – vyjadruje ostrosť krivky hustoty pravdepodobnosti (normálne rozdelenie má špicatosť g_2=3). Špicaté rozdelenia majú koeficient väčší ako 3 a ploché rozdelenia menší ako 3 (*obr.5*). Hodnota koeficienta špicatosti sa určí podľa vzťahu

$$g_2 = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(u_i - \bar{u}\right)^4 .n_j}{S^2} - 3 \qquad [21].$$

Výber niektorých typov rozdelení a ich koeficientov šikmosti je v tabuľke č.1.

Spoľahlivosť odhadu parametrov polohy

Bodové odhady stredných hodnôt a rozptylu sú vo svojej podstate náhodné veličiny a ako také kolíšu okolo zodpovedajúcich skutočných hodnôt. Odhad parametrov je teda zaťažený určitou chybou, ktorej veľkosť je potrebné pre posúdenie spoľahlivosti odhadov poznať. V súčasnej dobe je už dokonca v niektorých prípadoch by sa



a.) *b.)* *c.)*

Obr.6. Intervaly spoľahlivosti aritmetického priemeru normálneho rozdelenia pre rôzne hladiny významnosti α : a) ľavostranný, c) pravostranný, b) obojstranný interval (Meloun a Militký, 1995).

odhadovaná veličina mala pohybovať s požadovanou pravdepodobnosťou výskytu. Ich určenie má veľký význam pri hodnotení spoľahlivosti výpočtu zásob, technicko-ekonomických hodnoteniach alebo projekčných úlohách. Je možné konštatovať, že intervalové odhady práve v oblasti geologických vied sú veľmi užitočné, vzhľadom k značnému stupňu neurčitosti poznania prírodných javov a procesov.

S ohľadom na praktické potreby je asi najzaujímavejší intervalový odhad stredných hodnôt. Tento odhad je závislý na frekvencii výskytu hodnôt v súbore a stanovenej hladine významnosti α (obr.6). Môže byť jednostranný alebo ohraničený z ľavej či pravej strany. Obojstranný intervalový odhad matematickej nádeje súboru dát s normálnym rozdelením na hladine významnosti $\alpha = 0{,}05$ je

$$\hat{\mu} \in \left\langle \overline{u} \pm 1{,}96.\frac{s}{\sqrt{n}} \right\rangle \qquad [22].$$

Konštrukcia intervalového odhadu disperzie je analogická ako v predchádzajúcom prípade a určí vzťahom

$$\hat{\sigma}^2 \in \left\langle (n-1).\frac{S^2}{\chi_L^2}; (n-1).\frac{S^2}{\chi_P^2} \right\rangle \qquad [23].$$

Pri určení hraníc intervalu sa vychádza z frekvenčnej funkcie $f_r(\chi^2)$, pričom χ_L^2 je jej dolný kvantil $\chi^2_{1-\alpha/2}$ a χ_P^2 horný kvantil $\chi^2_{\alpha/2}$.

Overenie normality rozdelenia dát

Normalita rozdelenia dát patrí k základným predpokladom, pretože je na nej založená celá deskriptívna štatistika. Testy normality, slúžiace na určenie charakteru rozdelenia dát v súbore, používajú spomínané rýchle grafické metódy, alebo presnejšie numerické metódy. Jednou z numerických metód sú tzv. momentové testy, založené na špicatosti a šikmosti. Medzi ďalšie numerické metódy patrí porovnanie frekvenčných a distribučných funkcií. Toto sú najčastejšie používané metódy predovšetkým pri testovaní veľkého počtu dát. Medzi najznámejšie a najpoužívanejšie numerické metódy patria nasledovné testy :

- *Pearsonov test* χ^2 - používa sa na určenie zhody empirickej a teoretickej frekvencie výskytu hodnôt. Vypočítaná hodnota testovania $\boldsymbol{\chi^2}$ sa porovná s kritickou tabuľkovou hodnotou a v prípade, že je menšia, alebo rovná je možné prijať hypotézu o zhode empirickej a teoretickej frekvencie, H_0: n_j-$f(u_j)$=0.
- *Kolmogorov - Smirnovov test* D_1 - je založený na porovnaní distribučnej funkcie teoretického a empirického rozdelenia. Vypočítanú hodnotu $\boldsymbol{D_1}$ porovnáme s kritickou tabuľkovou hodnotou a v prípade, že je menšia alebo rovná je možné hypotézu o zhode distribúcií prijať, H_0: N_j-$F(u_j)\neq 0$.

Numerické testy sú podľa niektorých autorov menej citlivé na odchýlky od normality než diagnostické grafy. Odchýlka spôsobená napríklad vybočujúcimi hodnotami je potom ľahšie identifikovateľná v diagnostických grafoch (Blišťan, 1998).

2.3 Diskusia

Tento článok by svojím obsahom mal prispieť k uvedomovaniu si podmienok správneho používania matematických metód a postupov pri riešení geologických problémov. Ten kto správne pochopí možnosti a princípy štatistiky, získa na svoju stranu silnú zbraň, ponúkajúcu obrovské možnosti jej použitia. Vzhľadom na náročnosť a rozsiahlosť celej problematiky spracovania dát je nemožné na niekoľkých stranách popísať všetky princípy a postupy. Preto je toto len veľmi stručný a v mnohých smeroch až schematický popis danej problematiky. Ďalším zámerom je preto pripraviť pokračovanie článku, ktoré bude obsahovať ďalšie metódy a princípy štatistiky.

Literatúra

Blišťan, P.: Analýza kvalitatívnych a kvantitatívnych parametrov ložiska Rožňava-Strieborná žila. *Manuskript, archív KGaM, F BERG-TU, Košice, 1995*, *61 s.*

Blišťan, P.: Matematicko-štatistické a grafické spracovanie chemizmu na ložisku Nižná Slaná. *Písomná časť dizertačnej skúšky. Manuskript - archív KGaM, Košice, 1998*, *45 s.*

Bučko, M.: Počet pravdepodobnosti a matematická štatistika. *ALFA, Bratislava, 1981*, *245 s.*

Curwin, J. and Slater, R.: Quantitative Methods for Business Decisions. *International Thomson Business Press, London, 1996, pp. 667.*

Floreková, Ľ.: Matematické modelovanie. *ALFA, Bratislava, 1986, 272 s.*

Floreková, Ľ. a Benková, M.: Štatistické metódy. *FPP-F BERG TU Košice, 1999, 120 s.*

Meloun, M. a Militký, J.: Statistické zpracování experimentálních dat. *Plus, Praha, 1995, 839 s.*

Riečanová, Z. et al.: Numerické metódy a matematická štatistika. *ALFA, Bratislava, 1987,496 s.*

Schejbal, C.: Matematická geologie. *Ediční středisko VŠB, Ostrava, 1980,125 s.*