



数理统计在化探中的应用简介(三)

二、统计推断

1. 分布型式的推断

根据实测数据(子样)制成的频率分布图往往是不很规则的,它可以近似于某种理论上的概率分布,但多少会有某种偏离。所谓分布型式的推断,就是用统计的方法检验这种偏离是偶然原因引起的,还是所研究的对象(母体)本来就不服从那种理论频率分布。确定地质对象中某种元素的含量服从什么理论分布,对化探工作是很有意义的,这在上一章中已介绍过了。

在数理统计中已制定出多种检验方法,现以月山花岗闪长岩中Cu含量(稀盐酸提取,比色法测定)为例,介绍偏度峰度检验法。表1前三列为100个测定数据按对数间隔分组统计的结果。图1为据此作出的直方图。现在要检验其是否服从正态分布。方法的根据是计算偏度 γ_1 及峰度 γ_2 二个统计量。

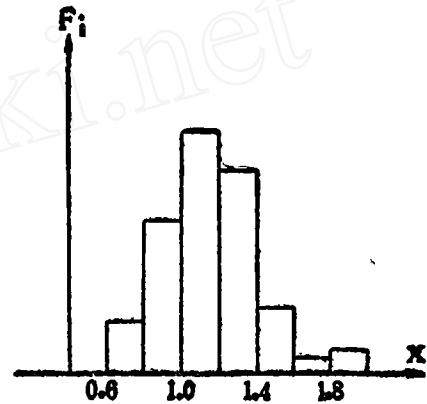


图1

前者表示直方图的对称性,后者描述直方图的陡缓情况。它们的计算公式为:

$$\gamma_1 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^3}{S^3} \dots\dots\dots (1)$$

偏度峰度法检验计算表

表1

对数间隔	组中值 x_i	频数 f_i	μ_i	$f_i \mu_i$	$f_i \mu_i^2$	$f_i \mu_i^3$	$f_i \mu_i^4$	
0.6-0.8	0.7	7	-2	-14	28	-56	112	$\mu = 0.29$ $\mu^2 = 0.084$
0.8-1.0	0.9	20	-1	-20	20	-20	20	$\mu^3 = 0.025$ $\mu^4 = 0.007$
1.0-1.2	1.1	32	0	0	0	0	0	$\mu^2 = 1.77$
1.2-1.4	1.3	27	1	27	27	27	27	$\mu^3 = 2.69$
1.4-1.6	1.5	9	2	18	36	72	144	$\mu^4 = 12.33$
1.6-1.8	1.7	2	3	6	18	54	162	$\gamma_1 = \frac{2.69 - 3 \times 1.77 \times 0.29 + 2 \times 0.02}{(1.77 - 0.084)^{3/2}} = 0.54$
1.8-2.0	1.9	3	4	12	48	192	768	
$n=100$			$\Sigma = 29$	177	269	1233	1233	$\gamma_2 = \frac{12.33 - 4 \times 2.69 \times 0.29 + 6 \times 1.77 \times 0.084 - 3 \times 0.007}{(1.77 - 0.084)^2} = 0.65$
			$\Sigma N = 0.29$	1.77	2.69	12.33		

$$\gamma_2 = \frac{\sum_{i=1}^n f_i(x_i - \bar{x})^4}{S^4} - 3 \dots\dots (2)$$

式中S为均方根差,等于 $\sqrt{\sum_{i=1}^n f_i(x_i - \bar{x})^2}$

$\bar{x}$ 为平均值, f_i, x_i 为各组的频率和组中值。

根据正态分布的公式,可以算出 γ_1, γ_2 的理论值均等于零。显然实际计算结果很难碰巧都等于零。但如X值确实服从正态分布,则 γ_1, γ_2 不应与零偏离太多。从理论上已经证明,当母体服从正态分布时,若子样容量较大($N > 100$),算出的 γ_1, γ_2 均为正态变量,其平均值皆为零,而方差分别为 $\sqrt{\frac{6}{N}}$

及 $\sqrt{\frac{24}{N}}$ 。根据正态分布的性质可知, γ_1 落在 $\pm 2\sqrt{\frac{6}{N}}$ 和 γ_2 落在 $\pm 2\sqrt{\frac{24}{N}}$ 之间的

概率为95%,而超出上述范围的可能性只有5%。这个“5%”在统计学中称为“信度”,

与之相应的 $\pm 2\sqrt{\frac{6}{N}}$ 和 $\pm 2\sqrt{\frac{24}{N}}$ 则称为 γ_1 和 γ_2 在5%信度上的临界值。信度高低

决定于所研究问题的性质。在地质科学中一般用5%或1%信度不同,临界值也不同。如计算结果未超出临界值,则认为母体服从正态分布。

在上面的例子中,为了简化计算,令 $\mu_i = \frac{x_i - x_0}{l}$ 。式中 x_0 一般取频数最大的组中值, l 为组距。本例中 $l = 0.2, x_0 = 1.1$ 。当 $x_i = x_0$ 时, $\mu = 0$; x_i 在 x_0 左边的各组, μ 取 $-1, -2, -3, \dots$; x_i 在 x_0 右边的各组取 $\mu = 1, 2, 3, \dots$ 。据此代入式(1)、(2)中整理而得:

$$\gamma_1 = \frac{\overline{\mu^3} - 3\overline{\mu^2}\overline{\mu} + 2\overline{\mu}^3}{(\overline{\mu^2} - \overline{\mu}^2)^{3/2}}$$

$$\gamma_2 = \frac{\overline{\mu^4} - 4\overline{\mu^3}\overline{\mu} + 6\overline{\mu^2}\overline{\mu}^2 - 3\overline{\mu}^4}{(\overline{\mu^2} - \overline{\mu}^2)^2} - 3.$$

式中各量的意义见表1(注意 $\overline{\mu^n}$ 与 $\overline{\mu}^n$ 不同)。

我们计算得 $\gamma_1 = 0.54, \gamma_2 = 0.65$ 。N = 100, 5%信度时的临界值为

$$\pm 2\sqrt{\frac{6}{100}} = 0.49, \pm 2\sqrt{\frac{24}{100}} = 0.98.$$

偏度 γ_1 已超过其临界值,所以这100个样品总的说来不服从正态分布。但由表1可见,最后一行频数为3,使 $\sum f_i \mu^3$ 增加192,而这三个样品受到矿化影响,如舍去这些样品,计算结果将落在临界值范围内。这里必须指出上述运算都是以含量的对数进行的,所以岩体未受矿化影响时Cu含量呈对数正态分布。如以含量真数进行计算, γ_1, γ_2 将大大超过其临界值。

偏度峯度法只适用于检验正态分布,而且要求子样容量相当大。这类方法叫参数性方法。与其对应的另一类方法是非参数性方法,可以不限母体的分布类型,因此在地质方面更适用。下面介绍一种柯尔莫洛夫-斯米尔诺夫检验法。此法的根据是最大累积频率差 D_n 的理论分布。据此可以算出不同子样容量在各种信度上的临界值。图2为子样容量在30~300之间时二种信度的 D_n 临界值曲线。具体方法如下:首先将子样自小而大顺序排列,统计出含量相同的样品数,并计算出各含量的累积频率,如表2。如果含量所取数值很分散,也可分间隔统计。

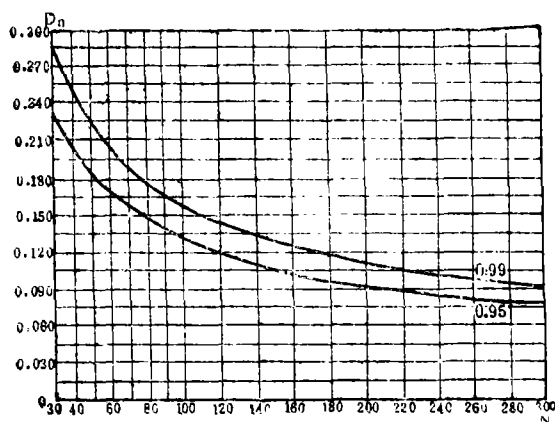


图2

表 2

含量	2	4	6	8	10	12	15	16	18	20	25	27	30	45	50	60	100	150	200	500	总计
子样数	1	1	0	4	6	8	4	11	13	25	5	0	8	1	10	4	5	9	8	2	$N_1=125$
频率	0.8	0.8	0	3.2	4.8	6.4	3.2	8.8	10.4	20	4	0	6.4	0.8	8	3.2	4.0	7.2	6.4	1.6	
1 累积频率	0.8	1.6	1.6	4.8	9.6	16	19.2	28	38.4	58.4	62.4	62.4	68.8	69.6	77.6	80.8	84.8	92	98.4	100	
子样数	0	0	1	3	3	0	3	2	0	24	2	5	3	2	5	2	10	10	4	1	$N_2=80$
频率	0	0	1.25	3.75	3.75	0	3.75	2.5	0	30	2.5	6.25	3.75	2.5	6.25	2.5	12.5	12.5	5	1.25	
2 累积频率	0	0	1.25	5	8.75	8.75	12.5	15	15	45	47.5	53.75	57.5	60	66.25	68.75	81.25	93.75	98.75	100	
累积频率差	0.8	1.6	0.35	-0.20	0.85	7.25	6.7	13	23.4	13.4	14.9	8.65	11.3	9.6	11.35	12.05	3.55	-1.75	-0.35	0	

最大频率差为23.4%。因 $N_1 \neq N_2$ ，应取 $N = \frac{N_1 + N_2}{2} = \frac{125 + 80}{2} = 102.5$ 。由图 2 查得Dn的5%临界值为19.5%。两个子样的最大累积频率差已超出此值，故认为它们来自不同分布的母体。

如果要检验某一子样是否服从正态分布，则需算出平均值及均方根差，代入理论

公式，算出理论累积频率，及其与实测累积频率之差的极大值，再与表 2 中的临界值对比。

2. 平均数的对比

在化探工作中常遇到含量对比的问题。例如含矿与不含矿岩体中某些元素的含量有无差别，同一岩体的东支与西支有无差别等等。解决这类问题的统计方法是根据子样的观察值计算出一个统计量 t ，

司都顿变量的临界值

表 3

自由度	$t_{0.01}$	$t_{0.05}$	自由度	$t_{0.01}$	$t_{0.05}$
1	63.65	12.71	18	2.88	2.10
2	9.93	4.30	19	2.86	2.09
3	5.84	3.18	20	2.85	2.09
4	4.60	2.78	21	2.83	2.08
5	4.32	2.57	22	2.82	2.07
6	3.71	2.45	23	2.81	2.07
7	3.50	2.37	24	2.80	2.06
8	3.36	2.31	25	2.79	2.06
9	3.25	2.26	26	2.78	2.06
10	3.17	2.23	27	2.77	2.05
11	3.11	2.20	28	2.76	2.05
12	3.06	2.18	29	2.76	2.05
13	3.01	2.16	30	2.75	2.04
14	2.97	2.15	40	2.70	2.02
15	2.95	2.13	60	2.66	2.00
16	2.92	2.12	120	2.62	1.98
17	2.89	2.11	∞	2.58	1.96

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sum f_i(X_i - \bar{X}_1)^2 + \sum f_i(X_i - \bar{X}_2)^2}{n_1 + n_2 - 2}}} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \dots\dots\dots (3)$$

式中 $\bar{x}_1$ 、 $\bar{x}_2$ 分别为两个子样的平均值， n_1 、 n_2 是它们的容量。

当 X_1 、 X_2 取自两个参数相同的正态母体，而 n_1 、 n_2 均 >50 时， t 服从标准正态分布。因此其5%信度的临界值为2。此时称为大子样推断。如 n_1 、 n_2 中有一个较小，则 t 不服从标准正态分布，而服从一种“司都顿 t 分布”（或“学生氏 t 分布”）。这种分布只有一个参数，即自由度 γ 。 $\gamma = n_1 + n_2 - 2$ 。根据 t 分布的理论公式可算出各种自由度在不同信度下的临界值。如表3。计算所得 t 值如未超过表中所列临界值，则可推断两个母体的平均值无显著差别。

表 4

对数间隔	组中值	子 样 1 (1号岩体)				子 样 2 (2号岩体)			
		频数 f_i	μ_i	$f\mu$	$f\mu^2$	频数 f_i	μ_i	$f\mu$	$f\mu^2$
0.6-0.8	0.7	4	-2	-8	16	2	-2	-4	8
0.8-1.0	0.9	10	-1	-10	10	1	-1	-1	1
1.0-1.2	1.1	14	0	0	0	8	0	0	0
1.2-1.4	1.3	8	1	8	8	18	1	18	18
1.4-1.6	1.5	2	2	4	8	9	2	18	36
1.6-1.8	1.7	2	3	6	18	1	3	3	9
1.8-2.0	1.9	0	4	0	0	1	4	4	16
		$\Sigma = 40$		0	60	40		38	106
		$\bar{\mu} = 0 \quad \bar{\mu}^2 = 0$				$\bar{\mu} = 0.95 \quad \bar{\mu}^2 = 0.9$			

下面举例说明具体方法。为简化计算，仍应用变量替换的方法。当两个子样的间隔划分及 μ_0 一致时，式3可化为：

$$t = \frac{\bar{\mu}_1 - \bar{\mu}_2}{\sqrt{\frac{\sum f\mu_1^2 - n\bar{\mu}_1^2}{n_1 + n_2 - 2} + \frac{\sum f\mu_2^2 - n_2\bar{\mu}_2^2}{n_1 + n_2 - 2}}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

设有两个岩体，各取40个样品，测得某元素含量后分组统计结果如下（表4）：

$$t = \frac{0 - 0.95}{\sqrt{\frac{60 - 40 \times 0 + 88 - 40 \times 0.9}{40 + 40 - 2}}} \cdot \sqrt{\frac{1}{40} + \frac{1}{40}} = -3.54。$$

由表8查 $\gamma = 40 + 40 - 2 = 78$ 时5%信度的临界值为 ± 2 ，计算所得 t 值超过此临界值，由此推断这两个岩体中被测元素含量有显著差异。

必须指出的是 t 分布要求两个母体的方

差相等。如不知方差是否相等时，需用相等数目的样品进行对比。至于检验方差是否相等，也有相应的统计推断方法，需要时可参阅有关书籍。

