



## (二) 均值的稳定性

张 强

上一讲我们引进了两个基本特征，一组数据  $x_1, \dots, x_n$  的均值与方差。本讲是要进一步阐明均值与方差的联系，揭示均值的变异规律。

现以表 1.1 (见第一讲) 所给的 80 个数据为例，如果每 5 年取一个均值，80 年的资料共得到 16 个数据 (见表 2.1)。计算这 16 个数据的

表 2.1

年份	均值	年份	均值
1860—1864	14.0	1865—1869	17.4
1870—1874	18.2	1875—1879	17.2
1880—1884	11.4	1885—1889	11.2
1890—1894	18.0	1895—1899	14.6
1900—1904	16.6	1905—1909	11.4
1910—1914	14.4	1915—1919	13.2
1920—1924	18.8	1925—1929	15.8
1930—1934	16.4	1935—1939	19.8

均值和方差，得

$$\bar{x} = 16.025, s^2 = 5.88, s = 2.425.$$

如果每 10 年取一个均值，共得到 8 个数据 (见表 2.2)。8 个数据的均值和方差是

$$\bar{x} = 16.025, s^2 = 2.84, s = 1.685.$$

表 2.2

年份	均值	年份	均值
1860—1869	15.7	1870—1879	17.7
1880—1889	12.8	1890—1899	16.3
1900—1909	14.0	1910—1919	13.3
1920—1929	17.3	1930—1939	13.1

为了便于对照，我们把这些结果列成表 2.3，从表上明显地看到，把原始资料分组取均值后，总的均值是不变的，但是方差和标准差随着组内资料个数的增多而逐渐下降。如果用标准差的大小来反映数据之间的差异程度，80 个数据的差异最大， $s = 4.76$ ，每 5 年取一个

均值，共有 16 个均值，这 16 个均值之间的差异小一些， $s = 2.425$ ；每 10 年取一个均值，共有 8 个均值，

表 2.3

资料分法	总数	均值	方值	标准差
每年 1 个	80	16.025	22.7	4.76
每 5 年的均值	16	16.025	5.88	2.425
每 10 年的均值	8	16.025	2.84	1.685

8 个均值之间的差异就更小， $s$  只有 1.685。可以设想，如果每 20 年取一个均值，共有 4 个数据，4 个数据的差异就更小。这种现象绝不是偶然的，它是有内在的规律的，它表明了均值的稳定性。所谓稳定，当然是相对的。从概率论的数学推导可以得到这样的公式：如果原来的单个数据的标准差是  $s$ ，把这些数据分成  $m$  个组，每组用均值来代替，则相应的这些均值的标准差就是  $\frac{1}{\sqrt{m}}s$ 。用这个公式算得的数与前面实际的  $s$  值进行比较，见表 2.4。单个数据 相应的标准差  $s = 4.76$ 。

前面实际的  $s$  值进行比较，见表 2.4。单个数据 相应的标准差  $s = 4.76$ 。

表 2.5 北京降水量 (毫米)

降水量 年代	年份	0	1	2	3	4	5	6	7	8	9
		1840	719	711	659	627	811	605	497	796	731
	1850	593	570	607	987	334	610				
	1860	462	545								
	1870	585	1054	691	372	670	132	621	491	814	710
	1880	377	602	620	984						
	1890	392	169	863	1034	1000	370	684	674	557	351
	1900						482	664	497	677	
	1910	628	752			721	718	417	781	513	561
	1920	277	256	838	380	1059	985	362	583		
	1930	450	537	687	762	661	385	406			
	1940	572	355	478	498	476	513	719	602	537	921

表 2.4

分组	m	公式算得的。	实际的
5 个	5	$1/\sqrt{5}$ 4.76 (= 2.13)	2.425
10 个	10	$1/\sqrt{10}$ 4.76 (= 1.54)	1.685

均值的稳定性对我们处理气象资料有很重要的意义。例如我们要比较两个地区的气温情况，不应该把一年内 365 天的气温逐日加以比较，这是没有意义的。因为两地区每日的差异不一定就能反映它们总的差异。如果比较月平均气温就会有意义些，更常见的是比较年平均气温。我们知道，夏季我国的月平均气温的等温线自东向西，有平行于海岸线的趋势，这说明了海洋对气候的影响；冬季的等温线却自南向北，有平行于纬度的趋势，反映了地理纬度对气候的影响。这些都是利用了均值的稳定性，它能更好地反映总的情况。

了解了均值的特性之后，就可以用它来处理一些资料，下面我们以北京各年降水量 (表 2.5) 为例

来进行分析。

现在从1840—1940年这110年的资料中分析北京降水的趋势。如果将逐年的降水数值点在图上，然后依时间顺序连成曲线，那就是一条锯齿形的折线(图略)，出现忽高忽低的现象，看不出明显的趋势。如果将110年的数据，每10年求一个均值，可得到11个数据(见表2.6)。将11个数值点绘成图2.1，可看出曲

表 2.6

年	平均降水量	资料个数
1840—1849	684	9
1850—1859	617.7	6
1860—1869	416.6	3
1870—1879	695	10
1880—1889	645.8	4
1890—1899	675.8	10
1900—1909	580	4
1910—1919	636.4	8
1920—1929	592.5	8
1930—1939	555.4	7
1940—1949	567.1	10

注 缺资料的年份不取，以有资料的年份来平均，因此资料不等(下同)。

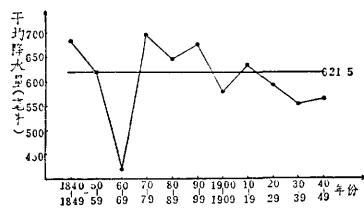


图 2.1

线的波动比较小，其整个趋势比较明显。在1900年以前的60年中，只有1860—1869年的平均降水量在

600毫米以下，其余的都在600毫米以上；然而，从1900年以后，50年中只有1920—1929年这10年的平均降水量在600毫米以上，其余都在600毫米以下，其整个图形呈下降趋势。

说到这里，自然会提出一个问题，究竟用多少年的资料来平均才合适呢？这个问题是很不容易回答的。如果我们的目的是分析110年内降水量的趋势，这可以作如下的考虑：

1. 首先将110年资料取一个总平均，得平均值621.5，仅从这个数值是看不出趋势的。

2. 如把前60年和后50年分开算平均，得两个均值652.5和586.5，相差66，似乎是下降趋势。但是，只有两个数值，一般总有大小，还不能说一定有下降趋势，进一步的理由见下一讲。

3. 如果每10年取一个均值，就得图2.1，从图上看出在1900年以前，6个点中有4个点在总均值621.5线上之上，2个点在总均值线下；而1900年以后，5个点中有4个点在总均值线下，一个点在总均值线上，从整个图形看，似乎是有下降的趋势。

4. 如果每5年和每3年取一个平均值，分别点绘成图2.2、图2.3。

从图2.2和图2.3上看出，每5年平均和每3年平均都大些，其趋势还不如每10年平均(图2.1)清楚。从图2.1可看出，如果1860—1869年的降水值不那么低的话，其趋势会更清楚些。事实上，1860

—1869年一共只有3年的资料，而且降水量还都比较小，用3年的资料来反映10年的情况，代表性就很差，因此，这个数值可以不必重视，这样再来看图2.1的曲线，其整个下降趋势就更明显了。

通过上面的讨论和分析比较，是可以找到一个较好的期限，在这个期限内求平均值，大致上可以看出趋势的情况。今后我们还会看到许多不同的方法都可以来寻找趋势的变化，然而，它们也都离不开均值和方差这两个基本特征，在以下的各讲中要逐步地来介绍这些方法。

现在，再回来讨论一下，为什么均值的方差会比单个数据的方差小，而且规律很好，n个数据的均值的方差与单个数据的方差之比恰好是 $1/n$ ，现来说明一些理由。

首先，我们看两个数据之和的方差与单个数据的方差之间有什么关系。

设  $x_1, \dots, x_n$  与  $y_1, \dots, y_n$  是两组不同的数据， $\{x_i\}$  的方差是  $S_x^2$ ， $\{y_i\}$  的方差是  $S_y^2$ ，现在看  $z_i = x_i + y_i$  的方差是多少。注意到

$$z_i = x_i + y_i, \quad i = 1, 2, \dots, n.$$

所以  $\sum_{i=1}^n z_i = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$ ，两边都除以 n 得均值， $\bar{z} = \bar{x} + \bar{y}$

$$\therefore \sum_{i=1}^n (z_i - \bar{z})^2 = \sum_{i=1}^n [(x_i + y_i) - (\bar{x} + \bar{y})]^2$$

$$= \sum_{i=1}^n [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2]$$

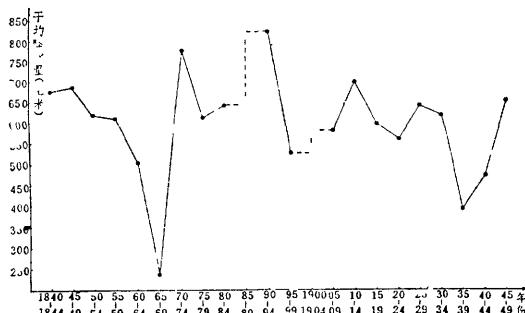


图 2.2

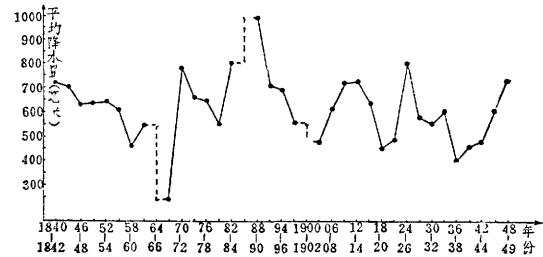


图 2.3

$$= \sum_{i=1}^n (x_i - \bar{x})^2 + 2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n (y_i - \bar{y})^2$$

如果上式右端第二项为 0，两边除以 n 后，就得到  $\{z_i\}$  的方差  $s_z^2$  是  $s_x^2 + s_y^2$ ，就有等式：

$$s_z^2 = s_x^2 + s_y^2 \quad (2.1)$$

即两个数据之和的方差等于各个数据方差之和。只要这两组数据彼此是不相关的，(2.1)式就能成立。以此类推，可以看到 n 个数据之和的方差就是各个数据的方差之和，如果每个数据相应的方差都一样，n 个数据之和的方差就是单个数据方

差的 n 倍。

另一方面，从方差的公式就可看出，如果数据  $x_1, \dots, x_n$  的方差是  $s_x^2$  的话，那么把每个数据都乘以 c，即令  $y_i = cx_i, i = 1, 2, \dots, n$ ， $y_1, \dots, y_n$  的方差  $s_y^2$  就是  $s_x^2$  的  $c^2$  倍。下面来验证一下：

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (cx_i - c\bar{x})^2 = \\ &c^2 \sum_{i=1}^n (x_i - \bar{x})^2, \end{aligned}$$

两边除以 n，就得

$$s^2 = c^2 s_x^2 \quad (2.2)$$

把 (2.1) 和 (2.2) 两式合起来，就可以说明  $1/\sqrt{n}$  的规律。设  $x_1,$

$\dots, x_m$  这 m 个数据每一个的方差都一样，都是  $s^2$ ，它们彼此是不相关的，用 (2.2) 知道每一个  $x_i/m$  的方差都是  $\frac{1}{m^2}s^2$ ，用 (2.1) 知道  $\frac{1}{m}x_1 + \frac{1}{m}x_2 + \dots + \frac{1}{m}x_m$  的方差  $s_{\bar{x}}^2$  确实是单个数据的方差  $s^2$  的  $\frac{1}{m}$ 。从标准差来看，将方差开方，就得  $1/\sqrt{m}$  的因子。上面这些只是一个说明，并不是数学上公式的推导，要推导这些公式，用概率论的方法要清楚、简便得多，这里就不作介绍了。