

# 气象资料的 统计分析方法

张尧庭 赵 溱 陈应强



# 气象资料的 统计分析方法

张尧庭 赵 溱 陈应强 编



农业出版社

30536

# **气象资料的统计分析方法**

**张尧庭 赵 溱 陈应强 编**

**农业出版社出版 新华书店北京发行所发行**

**农业出版社印刷厂印刷**

787×1092 毫米 32 开本 5.25 印张 105 千字  
1979 年 8 月第 1 版 1979 年 8 月北京第 1 次印刷  
印数 1—13,800 册

**统一书号 13144·199 定价 0.44 元**

## 序 言

1977—1978年，我们在《气象》杂志上曾连续发表了11篇介绍气象资料统计方法的短文，原想通过这个介绍，不用概率论和高等数学的知识；来说明一些统计的基本思想和方法——特别是多元分析中的一些方法。后来，感到这几篇短文，忽视了常用的回归分析，平方和分解的思想也没有很好地贯彻。现在，趁这次汇编出版的机会，对原稿的各讲作了一些修改，添上了介绍回归分析、逐步回归、判别分析和依距离聚类的几种方法，一共15讲，这样对气象资料统计分析方法的阐述可能较全面一些，平方和分解的思想也更清楚些。

本书所举的例子，着重说明方法和基本思想，因此数据很少。我们考虑到如果数据太多，势必演算的步骤就长，演算的过程就烦，反而弄不清整个的思路和技巧，抓不住重点。因此，希望读者看例子时着眼于例子所说明的概念、方法以及存在的问题。我们的目的是想把这些统计方法介绍给广大的搞实际工作的同志。只要具有高中的数学基础知识，努力一下是可以看懂的。但限于我们的水平，缺点和错误一定有不少，诚恳地希望读者批评指正。

编者 张尧庭 赵 溱 陈应强

1978年6月

## 目 录

|       |               |     |
|-------|---------------|-----|
| 第 一 讲 | 两个基本概念——均值和方差 | 1   |
| 第 二 讲 | 均值的稳定性        | 14  |
| 第 三 讲 | 显著性差异         | 23  |
| 第 四 讲 | 回归分析          | 36  |
| 第 五 讲 | 逐步回归分析        | 50  |
| 第 六 讲 | 判别分析          | 61  |
| 第 七 讲 | 极差分割法         | 68  |
| 第 八 讲 | 最优分割法         | 79  |
| 第 九 讲 | 如何使用分割法来作预报   | 91  |
| 第 十 讲 | 分类筛选预报的极差法    | 103 |
| 第十一讲  | A. I. D. 方法   | 115 |
| 第十二讲  | A. I. D. 的作图法 | 124 |
| 第十三讲  | 秩相关法          | 131 |
| 第十四讲  | 相似系数          | 142 |
| 第十五讲  | 聚类分析中的距离      | 153 |

## 第一讲 两个基本概念——均值和方差

各地气象台、站的气象资料，客观地反映了当地历史上的气象状况，如何从这些历史资料中总结出一些规律，作为指导今后天气预报的根据，这是气象资料处理的一个重要课题。

统计方法是人类长期实践活动的经验概括，它的一些基本概念已经渗透在人类活动的许多领域。例如平均数这个概念，在日常生活中常常会遇到它，然而，要真正掌握这个概念并运用它来分析和解决各种具体问题，困难还是很多的。本讲座的目的是希望读者结合学习、应用讲座中提出的方法来处理本地的气象资料，更好地掌握统计方法中的两个基本概念——均值和方差。我们会看到，这两个似乎是简单的概念，如果运用得好，就可以解决比较复杂的问题，可以引导出许多不同的方法。下面我们就从最简单的问题说起。

某地 1860—1939 年 80 年中 3 月份降水日数如表 1.1。怎样来分析这些数据呢？一般说来，我们希望从数据中分析出一些规律性的结论。比如下面几个问题：

- ① 这个地区 3 月份的降水日数正常的年景是多少？
- ② 在正常的年景，3 月份的降水日数应该在什么样的

表 1.1

| 年<br>代<br>日<br>数<br>年<br>份 | 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|----------------------------|----|----|----|----|----|----|----|----|----|----|
| 1860—1869                  | 21 | 7  | 19 | 15 | 8  | 13 | 23 | 17 | 19 | 15 |
| 1870—1879                  | 17 | 17 | 22 | 18 | 17 | 23 | 11 | 25 | 15 | 12 |
| 1880—1889                  | 5  | 8  | 7  | 21 | 16 | 16 | 23 | 13 | 9  | 10 |
| 1890—1899                  | 21 | 14 | 13 | 17 | 25 | 16 | 9  | 20 | 9  | 19 |
| 1900—1909                  | 21 | 12 | 11 | 17 | 22 | 8  | 18 | 8  | 10 | 13 |
| 1910—1919                  | 14 | 12 | 17 | 17 | 12 | 19 | 14 | 18 | 18 | 22 |
| 1920—1929                  | 21 | 20 | 20 | 15 | 18 | 11 | 18 | 17 | 21 | 12 |
| 1930—1939                  | 17 | 20 | 13 | 14 | 18 | 13 | 20 | 23 | 21 | 22 |

范围之内？

③ 哪些年是正常的？哪些年是不正常的？不正常的年是否有重现的周期？

这几个问题的回答，都可以使用平均数和方差概念来表述。

问题①就是问这个地区 3 月份的降水日数最有代表性的值是多少，而平均数就是最有代表性的值，因此，只要求出平均数就可以回答问题①。实际计算结果，该地区 3 月份平均降水日数是 16 天。

实际上，在正常年景，该地区 3 月份的降水日数不见得正好是 16 天，总是有一些出入，因此就需要考察有多大的波动，这个波动的量就是标准差，它与方差的概念是紧紧相连的，因此，只要求出了方差的值，问题②就可以回答了。

问题③实质上是与问题①、②有紧密的联系，如知道正

常年景的代表性的数值与波动的范围，就知道了正常年景降水日数的界限，利用这个界限就可以判定哪些年的降水日数是正常的或是不正常的。

从上面三个问题的分析和讨论，就明白关键是要找出这些数据的平均数和方差，下面就来引入这两个概念并说明它们的统计意义。

定义 1.1 给定了一组数据  $x_1, \dots, x_n$ ，则称  $\frac{1}{n}(x_1 + \dots + x_n)$  是它们的平均数，简称为均值，通常用  $\bar{x}$  来表示。用求和号“ $\Sigma$ ”表示， $\bar{x}$  可写成：

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

有关求和号“ $\Sigma$ ”的一些简单运算规则，主要的有三条：

$$\sum_{i=1}^n c x_i = c \sum_{i=1}^n x_i,$$

$$\sum_{i=1}^n (a_i + b_i) = \sum_{i=1}^n a_i + \sum_{i=1}^n b_i,$$

$$\left( \sum_{i=1}^n x_i \right) \left( \sum_{j=1}^m y_j \right) = \sum_{i=1}^n \sum_{j=1}^m x_i y_j.$$

为什么  $\bar{x}$  就能代表这  $n$  个数据  $x_1, \dots, x_n$  呢？为什么别的数值就不能代表这  $n$  个数据  $x_1, \dots, x_n$  呢？

当给定了一组数据  $x_1, \dots, x_n$  后，如果某一个常数  $c$  确实能代表它们，那么  $x_1 - c, x_2 - c, \dots, x_n - c$ （即  $x_1, \dots, x_n$



与  $c$  之差) 都应该很小, 也就是这些差值总起来应该小, 考虑到  $x_i - c$  的正负号没有什么意义, 因此用  $(x_1 - c)^2, (x_2 -$

$c)^2, \dots, (x_n - c)^2$  的和  $\sum_{i=1}^n (x_i - c)^2$  来反映, 很明显,  $\sum_{i=1}^n$

$(x_i - c)^2$  的值越小, 用  $c$  来代表这一组数据  $x_1, \dots, x_n$  的“代表性”就越好。因此, 最能代表  $x_1, \dots, x_n$  的数值  $c$  一定使

$\sum_{i=1}^n (x_i - c)^2$  达到最小值。我们就来寻找这个使  $\sum_{i=1}^n (x_i - c)^2$

达到最小值的  $c$ , 为此, 我们先证明两个公式, 这两个公式在今后是常常要用到的。

$$1. \quad \sum_{i=1}^n (x_i - \bar{x}) = 0 \quad (1.1)$$

公式 (1.1) 的意义是一组数据对于它们均值之差的代数和为 0。既然是均值, 那么数据中一定有比它大的, 也一定有比它小的 (因为一般地说, 一组数据全都是同一个数, 这种情况几乎是没的), 比它大的那些数据和它的差值之和, 比它小的那些数据和它的差值之和, 恰好符号相反, 绝对值相等, 因此, 全部的差值之和是 0, 这说明了均值的某种“代表性”。

证明:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$

$$n\bar{x} = \sum_{i=1}^n x_i,$$

$$\sum_{i=1}^n x_i - n\bar{x} = 0,$$

$$\sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = 0,$$

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

$$\begin{aligned} \text{又} \quad \sum_{i=1}^n (x_i - c)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - c)^2 \\ &= \sum_{i=1}^n [(x_i - \bar{x})^2 + 2(\bar{x} - c)(x_i - \bar{x}) + (\bar{x} - c)^2] \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - c) \sum_{i=1}^n (x_i - \bar{x}) + \\ &\quad + \sum_{i=1}^n (\bar{x} - c)^2, \end{aligned}$$

利用公式(1.1), 右端第二项是 0, 第三项是  $n(\bar{x} - c)^2$ , 因此, 就证明了

$$\Delta \quad 2. \quad \sum_{i=1}^n (x_i - c)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - c)^2 \quad (1.2)$$

公式(1.2)是平方和分解的最简单的形式, 它从各种角度去看, 就得到各种不同的解释和用处。

(1) 一组数据  $x_1, \dots, x_n$  与某一个常数  $c$  的差异 (1.2 式的左端) 是由两部分组成的: 一部分是  $x_1, \dots, x_n$  与它们的均值的差异 (1.2 式右端第一项), 另一部分是均值  $\bar{x}$  与常

数  $c$  的差异 (1.2 式右端第二项)。(1.2) 式左端的一个平方和分解成右端的两个平方和, 因此, 称 (1.2) 式为平方和分解的公式。

(2) 为什么  $x_1, \dots, x_n$  的均值  $\bar{x}$  能代表它们。在式 (1.2) 左端要找的最有“代表性”的  $c$  应使  $\sum_{i=1}^n (x_i - c)^2$  达到最小值, 从式 (1.2) 看出

$$\sum_{i=1}^n (x_i - c)^2 \geq \sum_{i=1}^n (x_i - \bar{x})^2$$

是永远成立的, 因此要能达到最小值, 就是要使上述不等号变成等号。等号成立的充分必要条件是  $n(\bar{x} - c)^2 = 0$ , 即  $\bar{x} =$

$c$ 。可见当  $c$  选为  $\bar{x}$  时,  $\sum_{i=1}^n (x_i - c)^2$  确实达到了最小值,

这说明了在全部可以选取的常数中, 只有均值  $\bar{x}$  与数据  $x_1, \dots, x_n$  的差异最小, 均值  $\bar{x}$  确实能“代表”这一组数据。

用图 1.1 来表示, 把数据  $x_1, \dots, x_n$  用实轴上的点来代表,  $c$  是实轴上的任一点,  $\bar{x}$  是  $x_1, \dots, x_n$  的均值。

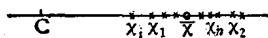


图 1.1

从上面的讨论知道: 对于一组数据  $x_1, \dots, x_n$  来说,  $\bar{x}$  是一个代表性的值,  $\sum_{i=1}^n (x_i - \bar{x})^2$  反映了这组数据偏离均值

的情况,  $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$  是  $(x_i - \bar{x})^2$  的均值, 即平均的偏差

平方和, 这在统计中是很重要的一个概念, 因此我们引入:

定义 1.2 给定了一组数据  $x_1, \dots, x_n$ , 则称  $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$  是这一组数据的方差, 通常用  $s^2$  表示, 把  $s$  称为这一组数据的标准差。

下面我们计算一下表 1.1 中数据的方差和标准差。计算结果是:  $\bar{x}=16.025$ ,  $s^2=22.7$ ,  $s=4.76$ 。如把 80 年的资料按偏离均值的程度来分析, 就得到:

第一类: 偏离均值不超过  $s$  的共有 49 年: 1860, 1862, 1863, 1865, 1867, 1868, 1869, 1870, 1871, 1873, 1874, 1878, 1879, 1884, 1885, 1887, 1891, 1892, 1893, 1895, 1897, 1899, 1901, 1903, 1906, 1909, 1910, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 1921, 1922, 1923, 1924, 1926, 1927, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936。

第二类: 偏离均值超过  $s$  而不超过  $2s$  的共有 30 年: 1861, 1864, 1866, 1872, 1875, 1876, 1877, 1881, 1882, 1883, 1886, 1888, 1889, 1890, 1894, 1896, 1898, 1900, 1902, 1904, 1905, 1907, 1908, 1919, 1920, 1925, 1928, 1937, 1938, 1939。

第三类: 偏离均值超过  $2s$  的只有一年, 即 1880 年。

由此看出, 绝大部分的降水日数偏离均值的值不会超过

$2s$ 。不超过  $s$  的，占  $49/80$ ，即占  $61.25\%$ ；不超过  $2s$  的，占  $79/80$ ，即占  $98.75\%$ 。如果以  $2s$  作为正常的波动范围，就可以判断 1880 年反常，这一年 3 月份的降水日数少得反常。

当然以  $s-2s$  为异常，而以超过  $2s$  为特大异常，这与气象普遍情况不太符合，即异常年份太多，而特大异常年份又太少，这是因为资料的分布及选定标准而出现的问题。如以  $s-1.5s$  为异常标准，则这种情况就可能好些。

在气象部门，通常划分异常不异常的方法是将全部资料从小到大排列，事先规定好一个百分比，例如一头一尾各占百分之五，这些数据就认为是异常的。这种分法不是从资料本身的波动情况来考虑，而是对各种资料千篇一律地按这个框框去套，在统计上是无法说清它的道理的，最好还是从资料本身去考虑。当然，如果除了资料本身提供的“信息”，还有其它的材料帮助我们作出判断，那就是另一回事了。

根据上面分析，对以上问题可作出如下回答：

1. 这个地区 3 月份的降水日数正常年景应是 16 天左右。

2. 3 月份降水日数范围应在  $16 \pm 9$ ，即 7—25 天为正常情况。

3. 除了 1880 年以外，其余的年份都可以看成是正常的。既然只有一个反常的年份；所以就谈不上反常情况出现的周期了。

从这看到，计算一组数据的均值和方差，对于我们分析

资料是很有帮助的。在计算均值和方差时，有一些简化计算的方法，这些方法不仅对手算是必要的，就是应用电子计算机计算时，用它也可以编出好一些的程序。

1. 将一组数据减去一个常数。

设  $x_1, \dots, x_n$  是给定的一组数据， $\bar{x}$  是它们的均值， $s_x^2$  是它们的方差。令

$$y_i = x_i - c, \quad i=1, 2, \dots, n,$$

用  $\bar{y}$  表示  $y_1, \dots, y_n$  的均值， $s_y^2$  表示  $y_1, \dots, y_n$  的方差，于是有： $\bar{x} = \bar{y} + c$ ,  $s_x^2 = s_y^2$

证明： $y_i = x_i - c$ ,

$$x_i = y_i + c,$$

$$\sum_{i=1}^n x_i = \sum_{i=1}^n y_i + nc,$$

两边除  $n$ , 得  $\bar{x} = \bar{y} + c$ . (1.3)

又  $\because x_i = y_i + c$ ,

$$\bar{x} = \bar{y} + c,$$

两式相减，得  $x_i - \bar{x} = y_i - \bar{y}$ ,  $(x_i - \bar{x})^2 = (y_i - \bar{y})^2$ ,

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

两边除  $n$ , 得  $s_x^2 = s_y^2$ . (1.4)

以表 1.1 中 1900—1909 年的数据为例，各减去 16，得：

5   -4   -5   1   6   -8   2   -8   -6   -3

$y_1$     $y_2$     $y_3$     $y_4$     $y_5$     $y_6$     $y_7$     $y_8$     $y_9$     $y_{10}$

很容易算得  $\bar{y} = -2$ ,  $s_y^2 = 24$ 。因此

$$\bar{x} = \bar{y} + 16 = -2 + 16 = 14, \quad s_x^2 = s_y^2 = 24.$$

注意到  $c$  本身可正可负, 因此一组数据同加一个常数也一样处理。

## 2. 运用公式(1.2)来计算方差。

在实际计算方差时, 不需要先算出均值, 然后将各个数据减去均值再平方求和, 应用公式(1.2), 移项就有:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - c)^2 - n(\bar{x} - c)^2 \quad (1.5)$$

$c$  是自己选取的, 选  $c$  使得  $|x_i - c|$  比较小, 这样计算量就小一些。以表 1.1 中 1900—1909 年的数据为例, 得表 1.2。

表 1.2

| $x_i$            | 21 | 12 | 11 | 17 | 22 | 8  | 18 | 8  | 10 | 13 | $\Sigma$ (总和) |
|------------------|----|----|----|----|----|----|----|----|----|----|---------------|
| $y_i = x_i - 16$ | 5  | -4 | -5 | 1  | 6  | -8 | 2  | -8 | -6 | -3 | -20           |
| $y_i^2$          | 25 | 16 | 25 | 1  | 36 | 64 | 4  | 64 | 36 | 9  | 280           |

于是,  $\bar{x} = \frac{1}{10} \sum_{i=1}^{10} y_i + 16 = -2 + 16 = 14,$

$$\begin{aligned} s_x^2 &= \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x})^2 = \frac{1}{10} \left[ \sum_{i=1}^{10} (x_i - 16)^2 - 10(16 - \bar{x})^2 \right] \\ &= \frac{1}{10} \left[ \sum_{i=1}^{10} y_i^2 - 10 \times 4 \right] = 24. \end{aligned}$$

列成上面这样的表格来计算均值和方差是手算时常用的。

在(1.5)式中, 如取  $c=0$ , 就得到常用公式

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2. \quad (1.6)$$

3. 添加或减少一个数据时, 均值和方差的计算公式。

在逐年积累的气象资料中, 每年就要增加一些资料, 于是使全部数据的均值和方差发生了改变, 如果每次重新计算就非常麻烦。例如除了表 1.1 中 80 个数据外, 增加了 1940 年 3 月份降水日数 10, 这样就有了 81 个数据, 这 81 个数据的平均数和方差是多少呢? 能不能利用原来 80 个数据的均值和方差来计算, 现在来推演这个计算公式。

设  $x_1, \dots, x_n$  的均值和方差是  $\bar{x}$  和  $s^2$ , 添加  $x_{n+1}$  后的均值和方差记为  $\bar{x}_{+1}$  和  $s_{+1}^2$ , 于是有

$$\begin{aligned} \bar{x}_{+1} &= \frac{1}{n+1} \sum_{i=1}^{n+1} x_i = \frac{1}{n+1} \left[ \sum_{i=1}^n x_i + x_{n+1} \right] \\ &= \frac{1}{n+1} \left[ \frac{n}{n} \sum_{i=1}^n x_i + x_{n+1} \right] = \frac{n}{n+1} \bar{x} + \frac{1}{n+1} x_{n+1} \\ &= \frac{1}{n+1} [n\bar{x} + x_{n+1}], \end{aligned}$$

$$\text{又 } s_{+1}^2 = \frac{1}{n+1} \sum_{i=1}^{n+1} (x_i - \bar{x}_{+1})^2, \quad \text{而}$$

$$\sum_{i=1}^{n+1} (x_i - \bar{x}_{+1})^2 = \sum_{i=1}^n (x_i - \bar{x}_{+1})^2 + (x_{n+1} - \bar{x}_{+1})^2, \quad (\text{A})$$



(A)式右端的第一项可以用公式(1.2), 把  $\bar{x}_{+1}$  看作  $c$ , 就有

$$\sum_{i=1}^n (x_i - \bar{x}_{+1})^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \bar{x}_{+1})^2 \quad (B)$$

$$\text{又} \quad \bar{x}_{+1} = \frac{1}{n+1} [n\bar{x} + x_{n+1}],$$

$$\begin{aligned} \text{由于} \quad \bar{x} - \bar{x}_{+1} &= \bar{x} - \frac{n}{n+1} \bar{x} - \frac{1}{n+1} x_{n+1} \\ &= \frac{1}{n+1} (\bar{x} - x_{n+1}), \end{aligned}$$

代入(B)式, 得

$$\sum_{i=1}^n (x_i - \bar{x}_{+1})^2 = ns^2 + \frac{n}{(n+1)^2} (\bar{x} - x_{n+1})^2 \quad (C)$$

$$\begin{aligned} \text{又} \quad x_{n+1} - \bar{x}_{+1} &= x_{n+1} - \frac{n}{n+1} \bar{x} - \frac{1}{n+1} x_{n+1} \\ &= \frac{n}{n+1} (x_{n+1} - \bar{x}), \end{aligned}$$

$$\therefore (x_{n+1} - \bar{x}_{+1})^2 = \frac{n^2}{(n+1)^2} (x_{n+1} - \bar{x})^2. \quad (D)$$

将(C)、(D)代入(A), 得到

$$\begin{aligned} \sum_{i=1}^{n+1} (x_i - \bar{x}_{+1})^2 &= ns^2 + \frac{n}{n+1} (\bar{x} - x_{n+1})^2 \\ s_{+1}^2 &= \frac{1}{n+1} \sum_{i=1}^{n+1} (x_i - \bar{x}_{+1})^2 = \frac{n}{n+1} s^2 + \\ &\quad \frac{n}{(n+1)^2} (\bar{x} - x_{n+1})^2. \end{aligned}$$