(三)显著性差异

张　强

毛主席教导我们:"**任何运动形式,其内部都包含着本身特殊的矛盾。这种特殊的矛盾,就构成一事物区别于他事物的特殊的本质。这就是世界上诸种事物所以有千差万别的内在的原因,或者叫做根据。**"

对于气象资料的分析,也是这样,要着力于寻找正常年份与异常年份的特殊性,这才能进一步掌握灾害性天气的发生规律。在统计方法中,很多方法就是为了判别一些资料是否来自不同的总体,这一类方法统称为差异的显著性检验,简称为显著性检验。下面用一个例子来说明。

我们仍以第一讲中的问题为例,某地三月份的降水日数(表1.1)客观上就存在着两大类,一类是正常的,一类是反常的,统计上称这两类存在着差异。我们在处理这 80 年的资料时,首先要区分出两大类。怎样来区分 80 年的资料呢?通常有两种不同的考虑:

1.根据经验,认为该地区三月份的降水日数超过 15 天就是不正常,于是按"≤15"还是">15"把80年资料分成两大类。这种考虑主要依赖于经验,主观成份比较大,因各人的想法不同,其标准也会不同。

2.以资料为依据,计算均值和标准差,如以偏离均值是否超过 2S为标准,不超过就认为是正常的,超过了就是反常的。这种方法只要"原则"定了,不论谁分析这 80 年资料,结论都是一样的。这种方法具有一定的客观性。而它的弱点是:①受资料的限制,如果资料年限不同,其结论就会不同,资料年限太短了,也不好处理;②受到"原则"的限制,如果换了一个原则,例如第一讲所指出的"偏离均值是否超过1.5S"为原则,所得结论就是另一种了。

统计方法是依照第二类的考虑方法进行探索分析的,它的中心问题就是寻找一些比较好的"原则",或称为"准则"、"检验方法"等等。处理这些资料的"原则"是统计方法应该研究的。而资料的多少是受客观条件限制的,统计方法无法解决。所谓统计中的"小样理论(或方法)"是指处理资料比较少的一些"原则";"大样理论(或方法)"是指处理比较多的资料的一类"原则"。

这样我们就会知道:讨论"以偏离均值是否超过 2S或1.5S"为"原则"好呢?还是以"偏离均值是否超过S"为"原则"好呢?这是有意义的。现在就来讨论这个问题。

假定$x_1,\cdots,x_n$,这n个数据是反映正常年份降水日数的资料,x_{n+1}是一个待判断的资料,即要判断它是正常的还是反常的。我们不能只考察某一个x_i与x_{n+1}差异的大小就来判断x_{n+1},这是不可靠的,考察全部$x_1,\cdots,x_n$与x_{n+1}的差异,就是偏离x_{n+1}的偏差平方和$\sum_{i=1}^{n}(x_i-x_{n+1})^2$。这里我们引入两个名词:一组数据$x_1,\cdots,x_n$对某个常数C的偏差平方和$\sum_{i=1}^{n}(x_i-C)^2$,简称为$x_1,\cdots,x_n$对C的离差平方和;如果C就是$x_1,\cdots,x_n$的均值,$\sum_{i=1}^{n}(x_i-\bar{x})^2$就简称为$x_1,\cdots,x_n$的偏差平方和。有的常常把$x_1,\cdots,x_n$对C的离差平方和简称为对C的离差平方和,$x_1,\cdots,x_n$的偏差平方和简称

为偏差平方和。这一点一定要记住，因为今后是经常使用的，不要混淆，更不要和其他材料中类似的名词相混。把x_{n+1}看成公式（1.2）中的常数C，于是得平方和分解的公式.

$$\sum_{i=1}^{n}(x_i - x_{n+1})^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - x_{n+1})^2 \quad (3.1)$$

注意上式中的$\bar{x}$是$x_1, \cdots, x_n$的均值。(3.1)式告诉我们，$x_1, \cdots, x_n$对x_{n+1}的离差平方和（反映了$x_1, \cdots, x_n$这一组数据与x_{n+1}的总的差异情况）可以分解为两部份 一部份是$x_1, \cdots, x_n$的偏差平方和$\sum_{i=1}^{n}(x_i - \bar{x})^2$（反映了$x_1, \cdots, x_n$这组数据内部本身的差异情况）；另一部分是$x_{n+1}$与$x_1, \cdots, x_n$的均值$\bar{x}$的偏差的平方$n(\bar{x} - x_{n+1})^2$（反映了$\bar{x}$与$x_{n+1}$的差异）。很明显，前一部份与$x_{n+1}$是无关的，它说明了同一类数据本身还可能发生的差异程度；后一部份说明了x_{n+1}与$x_1, \cdots, x_n$的“真正的”差异（虽然$\sum_{i=1}^{n}(x_i - x_{n+1})^2$反映了$x_1, \cdots\cdots, x_n$与$x_{n+1}$的总的差异，然而它夸大了这个差异，因为从（3.1）式看出，它混杂了$x_1, \cdots, x_n$本身的差异$\sum_{i=1}^{n}(x_i - \bar{x})^2$在内，并不能真实地反映出$x_{n+1}$与$x_1, \cdots, x_n$的差异）。把这两部分加以比较是有意义的，即考虑比值$\dfrac{n(\bar{x} - x_{n+1})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$。如果进一步分析一下（3.1）式右端的两项，可以看出$\sum_{i=1}^{n}(x_i - \bar{x})^2$是反映了n个数据总的差异程度，平均每一个数据$x_i$偏离$\bar{x}$的情况是$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$，而$n(\bar{x} - x_{n+1})^2$是反映了一个数据$x_{n+1}$与$\bar{x}$的偏差，应该将它与平均的平方偏差$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$相比，即考虑比值$\dfrac{n(\bar{x} - x_{n+1})^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$。如果这个比值很大，说明$x_{n+1}$与$\bar{x}$的差异大大地超过了$x_1, \cdots, x_n$的内部差异，那就表明$x_{n+1}$与$x_1, \cdots, x_n$不可能属于同一个类型的，在统计上就说$x_{n+1}$与$x_1, \cdots, x_n$这两类数据存在着显著性的差异；如果这个比值不大，说明x_{n+1}与$\bar{x}$的差异比起$x_1, \cdots, x_n$内部的差异来，几乎差不多，那就表明x_{n+1}与$x_1, \cdots, x_n$这两类数据之间并不存在显著性的差异，因此，x_{n+1}可以和$x_1 \cdots, x_n$归为一类。

* 较仔细一些，读者会发现在比值$\dfrac{n(\bar{x} - x_{n+1})^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$中，分子$(\bar{x} - x_{n+1})^2$就已经体现了均值$\bar{x}$与$x_{n+1}$的差异，为什么前面还要乘一个n呢？这样不是将这个差异放大了n倍吗？实际上则不然，在第二讲介绍均值的稳定性时已指出，均值本身的波动程度就比单个数据的波动程度要小（即$\frac{1}{\sqrt{n}}$规律），因此上面比值中，$(\bar{x} - x_{n+1})^2$反映的是均值对x_{n+1}的差异，而分母$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$反映的是$x_1, \cdots, x_n$单个数据彼此之间的差异，它们之间本来就相差$\frac{1}{\sqrt{n}}$倍，考虑到平方，实际上就相差了$\frac{1}{n}$倍，于是把n乘以$(\bar{x} - x_{n+1})^2$，就使得分子、分母才可以比较，否则进行比较就不合理了。今后在许多场合还会遇到类似的情况，我们就不再一一说明了。

从上面的分析和讨论就可以知道，要判断数据x_{n+1}与$x_1, \cdots, x_n$是否属于不同的类型时，应考虑比值$\dfrac{n(\bar{x} - x_{n+1})^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$，更仔细一些的推演，可以证明分母中用$\frac{1}{n-1}$代$\frac{1}{n}$更合适一些. 许多书上都把这个比值记为F，称为F值，即

$$F = \frac{n(\bar{x} - x_{n+1})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3.2)$$

F值大到什么程度就可以认为有显著性差异呢？当$n \geqslant 50$时，F值大于4就可以认为有显著性的差异；当$10 \leqslant n < 50$时，F值大于5就可以认为有显著性的差异；当$n < 10$时，这时需要查F的临界值表（这张表以后详细介绍）。容易看出，在(3.2)式中，当$n \geqslant 10$以后，分母中$\frac{1}{n-1}$用$\frac{1}{n}$去代替，差别也不是很大的，看怎么方便就可以了。当n很小时，$\frac{1}{n-1}$与$\frac{1}{n}$的差别就比较明显了。

现举几个例子来说明（3.2）式F值的使用和计算方法。

例3.1 从表1.1中80个资料来分析，如果把1880年的5天这个数据除外，其余都看成是正常情况下的降水天数，问1880年的降水天数是否是异常情况？

解：我们用式(3.2)的F值来判断1880年的5天是不是异常的，首先需要算出79年资料的均值$\bar{x}$和相应的偏差平方和。

我们用第一讲最后介绍的减少一个数据的均值和方差的计算公式，以及第一讲中已经算得的80个数据的均值和方差，就知道除1880年以外，79年三月份的平均降水日数$\bar{x}$及偏差平方和SS是：

$$\bar{x} = \frac{1}{79}(1282 - 5) = \frac{1277}{79} \approx 16.16$$

$$SS = 1818 - \frac{80}{79}(11.025)^2 \approx 1818 - 123 = 1695$$

于是算得F值是：

$$F = \frac{79(16.16 - 5)^2}{\frac{1}{78} \cdot 1695} = \frac{79 \times 78(11.16))^2}{1695}$$

$$\approx \frac{79 \times 78 \times 125}{1695} \geqslant 125$$

F值大大地超过了5，因此，有相当地把握可以断定

1880年的降水天数确实是异常地小。

我们再仔细考察一下整个问题讨论的过程，就会发现一个问题：在全部过程中，始终没有考虑到x_{n+1}这个数值本身的可能有的波动。实际上，x_{n+1}不是一个理论值，而是一个观测值，如果x_{n+1}是理论值，上面的讨论就完全合理了。现在x_{n+1}本身是一个观测值，它除了反映不正常年份的共性外还有它自己的特性，不正常年份的降水天数也是有一个波动范围的，因此，$(\bar{x}-x_{n+1})^2$的波动就不是单个数据的$\frac{1}{n}$，而是$1+\frac{1}{n}$了（见第二讲的有关部份），而$1+\frac{1}{n}=\frac{n+1}{n}\approx 1$（当n相当大时），这样就得到了F值，应是

$$F=\frac{(\bar{x}-x_{n+1})^2(\frac{n+1}{n})^{-1}}{\frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}=\frac{78\cdot\frac{79}{80}(16.16-5)^2}{1695}$$

$$=\frac{78\times 125}{1695}=\frac{650}{113}>5 \qquad (3.3)$$

可见1880年的数据确实是异常的。

然而，我们再仔细推敲一下，这样来分析处理也还有它不合理的地方，统计方法的发展历史也是这样，不断地发现以前方法中的优点和缺点，保持优点、改进缺点，就会发展出新的方法。后面我们还要讨论例3.1提出的问题，逐步地深化，以引出更好的更合理的处理方法。

例3.2　以表1.1的数据为基本资料，现在又知道1940年以后有两年三月份的降水日数分别是18和23，问这两年是否是异常的年份。

解：从例3.1的讨论中知表1.1中1880年的数据是异常的，除去后共有79年的资料，其均值和方差是：$\bar{x}=16.16, s^2=\frac{1}{79}1695$。现在对18和23用(3.3)式的F来计算，n＝79，对18就得到

$$F=\frac{(16.16-18)^2\cdot\frac{79}{80}}{\frac{1}{78}\cdot 1695}\approx\frac{(16.16-18)^2\times 78}{1695}$$

$$\approx\frac{(1.84)^2\times 78}{1695}\approx\frac{3.4\times 78}{1695}<1$$

因此，可以说18天不是异常的情况。对23得到

$$F=\frac{(16.16-23)^2\cdot\frac{79}{80}}{\frac{1}{78}\cdot 1695}\approx\frac{(16.16-23)^2\times 78}{1695}$$

$$\approx\frac{(6.84)^2\times 78}{1695}\approx\frac{46.8\times 78}{1695}\approx\frac{3650}{1695}<3$$

因此，可以说23天也不是异常的情况。

气象上的异常与所选取标准有关，例如这里取F＝5为标准，所得到的是罕见的特大异常，如用气象上常用的F＝3，或F＝1，则上述异常情况就不同了。

如果我们要比较的不是一组数据和一个数据，而是比较两组数据，那么相应的计算公式和分析方法又该如何进行呢？

我们用第二讲中的表2.1的数据为例，想分析一下1900年以前的60年和1900年以后的50年，北京地区的降水情况是否有显著的差异。以下用$x_1,\cdots,x_{n_1}$表示前60年的数据$(n_1=42)$，$y_1,\cdots,y_{n_2}$表示后50年的数据$(n_2=37)$。

把这两组数据总起来考虑，n_1+n_2个数据的总和除以n_1+n_2，得全部数据的均值，记为$\overline{GX}$，n_1个数据$x_1,\cdots,x_{n_1}$的均值和n_2个数据$y_1,\cdots,y_{n_2}$的均值分别用$\bar{x}$和$\bar{y}$来表示，自然有

$$\overline{GX}=\frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}x_i+\sum_{j=1}^{n_2}y_j\right)$$

$$=\frac{1}{n_1+n_2}(n_1\bar{x}+n_2\bar{y}), \qquad (3.4)$$

全部数据对$\overline{GX}$的偏差平方和是

$$\sum_{i=1}^{n_1}(x_i-\overline{GX})^2+\sum_{j=1}^{n_2}(y_j-\overline{GX})^2,$$

它反映了全部数据的差异程度，称为全部数据总的偏差平方和，记为$SS_{总}$。把$SS_{总}$的每一项运用第一讲的平方和分解公式（把$\overline{GX}$看成c），就得到

$$\sum_{i=1}^{n_1}(x_i-\overline{GX})^2=\sum_{i=1}^{n_1}(x_i-\bar{x})^2+n_1(\bar{x}-\overline{GX})^2,$$

$$\sum_{j=1}^{n_2}(y_j-\overline{GX})^2=\sum_{j=1}^{n_2}(y_j-\bar{y})^2+n_2(\bar{y}-\overline{GX})^2,$$

于是就有

$$SS_{总}=\sum_{i=1}^{n_1}(x_i-\bar{x})^2+n_1(\bar{x}-\overline{GX})^2$$

$$+\sum_{j=1}^{n_2}(y_j-\bar{y})^2+n_2(\bar{y}-\overline{GX})^2, \qquad (3.5)$$

注意到$\overline{GX}$与$\bar{x},\bar{y}$有关系式(3.4)，因此，

$$\bar{x}-\overline{GX}=\bar{x}-\frac{1}{n_1+n_2}(n_1\bar{x}+n_2\bar{y})=\frac{n_2}{n_1+n_2}(\bar{x}-\bar{y})$$

$$\bar{y}-\overline{GX}=\bar{y}-\frac{1}{n_1+n_2}(n_1\bar{x}+n_2\bar{y})=\frac{n_1}{n_1+n_2}(\bar{y}-\bar{x})$$

代入(3.5)式，便得到

$$SS_{总}=\sum_{i=1}^{n_1}(x_i-\bar{x})^2+\frac{n_1n_2^2}{(n_1+n_2)^2}(\bar{x}-\bar{y})^2$$

$$+\sum_{j=1}^{n_2}(y_j-\bar{y})^2+\frac{n_1^2n_2}{(n_1+n_2)^2}(\bar{x}-\bar{y})^2$$

$$=\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2+\sum_{j=1}^{n_2}(y_j-\bar{y})^2\right]+\frac{n_1n_2}{n_1+n_2}(\bar{x}-\bar{y})^2, \qquad (3.6)$$

很明显，(3.6)式右端［ ］内两个平方和反映了各组数据内部本身的差异程度，称为组内的偏差平方和，记为$SS_{组内}$；(3.6)式右端第二部分反映两组数据之间差异程度，称为组间差的平方和，记为$SS_{组间}$。因此(3.6)式就可以写成：

$$SS_{总}=SS_{组内}+SS_{组间} \qquad (3.6)'$$

这是比较复杂一些的平方和分解公式，和前面讨论过的情况一样，要判断组间的差异是否是显著的，就要考虑比值$\frac{SS_{组间}}{SS_{组内}}$，完全相同的理由，分母反映了n_1+n_2个数据的内部差异，应该用n_1+n_2去除一下，反映单个数据的波动程度，更准确一些，应该用n_1+n_2-2去除，于是就得到F值的公式

$$F=\frac{SS_{组间}}{\frac{1}{n_1+n_2-2}SS_{组内}}=\frac{\frac{n_1n_2}{n_1+n_2}(\bar{x}-\bar{y})^2}{\frac{\sum\limits_{i=1}^{n1}(x_i-\bar{x})^2+\sum\limits_{j=1}^{n2}(y_j-\bar{y})^2}{n_1+n_2-2}} \tag{3.7}$$

我们把(3.7)式的一个特殊情况$n_2=1$写出，它就是：(这是y_j只有一个数据y_1)

$$F=\frac{\frac{n_1}{n_1+1}(\bar{x}-y_1)^2}{\frac{\sum\limits_{i=1}^{n1}(x_i-\bar{x})^2}{n_1-1}}$$

它就是前面在例3.1中出现过的(3.3)，只是$n\to n_1$，$x_{n+1}\to y_1$而已。所以，我们只要弄清楚(3.7)，(3.3)自然是很特殊的情形了。

现在，用(3.7)的F值来检验一下表2.1中的数据，看看北京地区的降水量在1900年前后这几十年内是否有显著的差异。经过计算，算得：

$$\bar{x}=652.5,$$

$$SS_{x组内}=\sum_{i=1}^{42}(x_i-\bar{x})^2=1939589,\ \bar{y}=586.5,$$

$$SS_{y组内}=\sum_{j=1}^{37}(y_j-\bar{y})^2=1281765,$$

用(3.7)式，得：($n_1=42, n_2=37$)

$$F=\frac{\frac{42\times37}{42+37}(652.5-586.5)^2}{\frac{1939589+1281765}{42+37-2}}$$

$$=\frac{\frac{42\times37\times(66)^2}{3221354\times79}}{77}=\frac{42\times37\times42\times11^3\times6}{3221354\times79}$$

$$=\frac{7897428\times11\times6}{3221354\times79}\approx2.02$$

这个数比4小，因此没有理由说北京地区的降水在1900年前60年和后50年之间有显著的差异。如果不是用F值来检验，单看均值的差别，一个是652.5，一个是586.5，相差66mm，这似乎也是一个不小的数字了，经过F值的检验，知道这个差别还属于正常的波动范围之内，并不是降水的数量发生了重要的变化。

对于过去历史资料的分析，很重要的是能比较准确地将过去的资料划分出几个时期，每个时期有它自己的特点，不同的时期，各个指标之间应该有显著的差异，而F值检验正好可以帮助我们来鉴别是否真正存在显著性的差异，以后就要介绍怎样利用F比值来较好地划分历史资料的时期。

读者可以利用(3.7)来鉴别一下本地区的某一个指标在某一年前后是否有显著性的差异，以此作为练习，熟悉本讲的内容。

这样的短训班好

今年2月27日到3月6日，我们地区气象台开办了一期预报短训班。参加学习的同志一致反映，这样的短训班办得好，时间短，内容新，收获大，见效快。我们的做法是：

一、深入揭批“四人帮”。训练班一开始，就紧紧抓住揭批“四人帮”这个纲，紧密联系气象工作实际，批“四人帮”的罪行，揭“四人帮”的危害，肃清“四人帮”的流毒。大家说：粉碎“四人帮”，思想大解放，革命加拼命，预报要大上。决心把预报工作做得更好，为农业学大寨、普及大寨县作贡献。

二、面向县站，抓住重点。这次参加学习的主要是县站预报员。他们在使用单站资料和群众看天经验方面是有许多长处的，但对如何结合天气形势进行综合分析，普遍感到“棘手”，对大台广播的简易天气形势，有的点绘不合理，有的会点不会用，作预报往往是“听听广播看看天，天气形势搁一边”。针对这种情况，我们一方面深入县站，和县站同志实行“三同”，既向县站学习，又传授业务技术。另一方面分期开办以简易天气形势图分析为主的图、资、群结合的预报短训班，使全区十来个台站普遍来个大学习、大交流，这样做，深受县站欢迎。

三、互教互学，共同实践。这期短训班，采取能者为师，互教互学的办法，地区台预报员人人讲课，县站的同志人人介绍当地经验。讲课内容既有基本理论，又有实践经验。每讲一种灾害性天气系统，都能紧密结合本地区的特点，讲清各类天气系统在什么条件下对本地区会有什么样的影响。有的同志还把天气形势和本地天气的关系编成歌谣形式讲课，如“南高北低，久晴无疑”，“东北高压伸黄海，晴天一日有雨来”等等。大家听后，认为这样做，在形势分析中抓住了主要矛盾，既形象化，又好记，好用。

短训班发扬理论联系实际的革命学风，注重实践。每天除了讲课外，还抄收天气广播，分析简易天气图集体观云，参加会商，制作天气预报，使短训班办得团结、紧张、严肃、活泼。大家一致认为，办这样的短训班好。

江苏扬州地区气象台报道组