概率统计

张尧庭 编

工程数学

中央广播电视大学出版社

工程数学

概率统计

张尧庭 编

中央广播电视大学出版社

前　言

一本50学时能教完的书要介绍概率论、数理统计的全貌是困难的，必需有所取舍，才能重点突出，使学的人有所收获。本书的重点是数理统计，介绍一些概率论的概念是使得对一些统计的理论有更好的了解，因此舍去了一般教材中关于古典概型的比较详细的讨论。在数理统计中也是重点选取最有用的回归分析作为一条主线，目的是使学的人至少能掌握一种比较有用的方法。为此，书中对一些例子的计算步骤及结论作了较详细的讨论，希望通过对这些实例的讨论，读者对于应用中遇到的问题会进行分析。本书所引用的数字的实例，基本上都是选自国内的实际应用的问题，并注明了数据的来源，但计算及讨论大都是不一样的，个别的例是完全引用别的书上的，我也在注中说明了。为了使读者能对数理统计应用的广泛性有一点感性的认识，本书的例子和习题有意选用了各个不同行业的材料，并不只限于工业的应用。

很明显，这是一本带有探索性的教材，是想探索一种新的编写体系与方法，使读者在学习时少遇到一些困难，多感到一些兴趣。是否能达到预期的效果，只能由读者来评定了。欢迎提出各种改进、批评的意见，使得如有再版、改写的可能时，来作进一步的修改。

编　者　1984年9月

目　录

第一章　数据的简单分析

本章叙述数据分析的简单方法，计算平均数、方差、极差、中位数、频数等，为今后各章的概念提供直观的想法和背景。

§1　均　值

均值，又称平均数，是一般人都熟悉的概念。但是，为什么要用它？应怎样正确地使用它？这两个问题并不是都能回答清楚的。

给定一组数据 $x_1, \cdots, x_n$，我们称

$$\bar{x}=\frac{1}{n}(x_1+\cdots+x_n)=\frac{1}{n}\sum_{i=1}^{n}x_i$$

是这一组数据的**算术平均数**，简称为**平均数**或**均值**。为什么要求均值呢？因为它能反映这组数据的总的情况，或者说它有代表性。例如广东省 1961—1965 年带鱼的年产量是：

年　份	1961	1962	1963	1964	1965
产量(吨)	4108	2288	2911	4397	5373

于是平均年产量是

$$\frac{1}{5}(4108+2288+2911+4397+5373)=3815.4$$

这个数简单明瞭地反映了这几年带鱼的产量。

下面我们来进一步分析一下均值的代表性。对给定的 n 个数 $x_1, \cdots, x_n$，原则上可以用任何一个常数 c 去“代表”它们，问题只在于“代表”得好不好。$x_i - c$ 反映了 c 偏离 x_i 的程度，也就是 c“代表”x_i 的好坏。为了消除符号的影响，用

$$(x_1-c)^2+(x_2-c)^2+\cdots+(x_n-c)^2=\sum_{i=1}^{n}(x_i-c)^2$$

来衡量 c“代表”$x_1, \cdots, x_n$ 的好坏。显然，代表性最好的 c 应使 $\sum_{i=1}^{n}(x_i-c)^2$ 达到最小。而均值 $\bar{x}$ 恰巧是合于这一要求的唯一的数，下面我们来证明这一点，为此先证几个今后经常要用的公式。

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

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

即

$$n\bar{x}=\sum_{i=1}^{n}x_i=x_1+\cdots+x_n$$

$$=(x_1-\bar{x})+\cdots+(x_n-\bar{x})+n\bar{x}$$

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

公式 1.2　任给一个常数 c，成立等式

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

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

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

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

$$\xlongequal{\text{由}(1.1)}\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-c)^2$$

从(1.2)式可以看出,

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

而且只有在 $c=\bar{x}$ 时,才达到最小值 $\sum_{i=1}^{n}(x_i-\bar{x})^2$。这就说明了均值 $\bar{x}$ 的"代表"性。

计算均值 $\bar{x}$ 有一些常用的方法,我们列成公式,以后会不断地用到它们。

公式 1.3 若 $y_i=x_i-a$, $i=1,2,\cdots,n$, 则

$$\bar{y}=\bar{x}-a,\qquad \sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad (1.3)$$

证 $$\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}x_i-na$$

两边同除 n 即得 $\bar{y}=\bar{x}-a$, 又

$$y_i-\bar{y}=x_i-a-(\bar{x}-a)=x_i-\bar{x}$$

两边平方求和就得(1.3)的第二式。

公式 1.4 若 $y_i=bx_i$, $\quad i=1,2,\cdots,n$, 则

$$\bar{y}=b\bar{x} \qquad \sum_{i=1}^{n}(x_i-\bar{x})^2=b^{-2}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{1.4}$$

证 （留作习题）

从公式(1.3), (1.4)就可得:

公式 1.5 若 $y_i=bx_i-a$, $i=1, 2, \cdots, n$, 则

$$\bar{y}=b\bar{x}-a \qquad \sum_{i=1}^{n}(x_i-\bar{x})^2=b^{-2}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{1.5}$$

在求 $x_1, \cdots, x_n$ 的均值时，如果 x_i 中有相同的，自然可以合并，不妨设不同的只有 k 个值 $a_1, \cdots, a_k$，并且 a_i 出现了 n_i 次, $i=1, 2, \cdots, k$。此时

$$x_1+x_2+\cdots+x_n=n_1a_1+n_2a_2+\cdots+n_ka_k$$

因此

$$\bar{x}=\frac{1}{n}(x_1+\cdots+x_n)=\frac{1}{n}\sum_{i=1}^{K}n_ia_i=\sum_{i=1}^{K}a_i\cdot\frac{n_i}{n} \tag{1.6}$$

从上式右端可以看到，均值 $\bar{x}$ 与 a_i 出现的次数的关系是由比值 n_i/n 决定的，即只与 a_i 在全部数据中所占的比例有关。比值 n_i/n 越大，$\bar{x}$ 受 a_i 的影响大；n_i/n 越小，$\bar{x}$ 受 a_i 的影响就小。

把(1.6)式进一步推广就得到**加权平均数**的概念：给了一组数据 $x_1, \cdots, x_n$，又给了一组正数 $p_1, \cdots, p_n$ 且 $\sum_{i=1}^{n}p_i=1$，则

$$p_1x_1+p_2x_2+\cdots+p_nx_n$$

就称为 $x_1, \cdots, x_n$ 的加权平均数，$p_1, \cdots, p_n$ 称为 $x_1, \cdots, x_n$ 相应的权。很明显，当权 $p_1=\cdots=p_n=1/n$ 时，加权平均数就是算

术平均数。“权”p_i 就是衡量数据 x_i 在平均时的重视程度，因此怎样合理地决定权是重要的问题。

§2 方差、标准差、极差

对给定的一组数据 $x_1, \cdots, x_n$ 来说，$\bar{x}$ 是与这 n 个数最接近的数，即 $\sum_{i=1}^{n}(x_i-\bar{x})^2$ 在 $\sum_{i=1}^{n}(x_i-c)^2$ 中当 c 变化时达到最小值。$(x_i-\bar{x})^2$ 是 x_i 与 $\bar{x}$ 偏差的平方，这 n 个数据与 $\bar{x}$ 的平均的偏差平方就是

$$\frac{1}{n}[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2]$$

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

它就称为数据 $x_1, \cdots, x_n$ 的**方差**(或均方差——平均的平方偏差)，它的算术平方根称为**标准差**，记作 s，即有

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad s=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (2.1)$$

容易看出，s^2 越大，这组数据就越“分散”，或者说，这组数据的变异性(即互相不同的程度)就大；s^2 越小，这组数据的变异性就小，也就更“集中”。当 $s^2=0$，$x_1=x_2=\cdots=x_n=\bar{x}$。

因此对一组数据 $x_1, \cdots, x_n$ 来分析时，$\bar{x}$ 与 s(或 s^2)是最常用的两个量，一个是代表性的值(指 $\bar{x}$)，一个是描述数据的

变异性的值(指 s^2 或 s)。§1中的一些公式就是为了简化一些计算的。下面举一个例来说明这些公式的应用。

例1 已从调查资料获得健康人血清粘蛋白含量与矽肺病人血清粘蛋白含量如下表*:

健康人	二期矽肺病人
含量(mg/100ml)	
42.84	65.45
48.19	69.63
48.19	69.73
52.48	74.97
58.90	80.44
64.26	80.44
69.61	95.20
80.22	96.39

我们用 $x_1, \cdots, x_8$ 表示健康人的八个数据，将 x_i 均减去 50 后乘以 100，就不需要用小数来表示了，即 $y_i=100(x_i-50)$。于是得

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
−716	−181	−181	248	890	1426	1961	3022

$$\sum_{i=1}^{8} y_i = 6469 \qquad \bar{y} = \frac{1}{8}6469 = 808.63$$

于是

$$\bar{x} = \frac{\bar{y}}{100} + 50 = 58.0863$$

* 资料见白求恩医科大学卫生系统计教研室《交叉积差法》一文

将 x_i 逐个减去 $\bar{x}$，然后平方、求和、平均，可算得 s^2 为 140.15，标准差为 11.84。也可用(1.2)，取 $c=0$ 得

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2 \tag{2.2}$$

这是计算方差常用的一个公式。用(2.2)式计算

$$\sum_{i=1}^{8}x_i^2=28113.326$$

$$\sum_{i=1}^{8}x_i^2-n\bar{x}^2=1121.180$$

$$s^2=\frac{1}{8}1121.180=140.1475\doteq 140.15$$

与刚才那种直接计算的结果相同。(2.2) 式是一个很有用的公式。

对矽肺病人的数据进行类似的计算，就得均值是 79.03，方差是 117.82，标准差是 10.85。

从数据上看矽肺病人的均值高，方差小，健康人均值小，方差大，数据更分散些。

例 2 从收集到的资料，已知测得光速的值是下列八个：

299792.3, 299792.5, 299793.1

299794.2, 299792.6, 299793.0

299795.1, 299789.8

将上述各数减去 299790 之后，记为 y_i，于是有

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
2.3	2.5	3.1	4.2	2.6	3.0	5.1	−0.2

如果说例 2.1 中减去一个常数后求均值、方差的计算并没有多大的变化的话，这一例就明显地感到简单得多。计算后得

$$\bar{y}=\frac{1}{8}(2.3+2.5+3.1+4.2+2.6+3.0+5.1-0.2)$$

$$=\frac{1}{8}22.6=2.825$$

因此得光速的均值为

$$299790+2.825=299792.825$$

由于 y_i 的值是原数据各减同一个值得来的，方差不会改变，直接算 y_i 的方差比较方便。于是可以求得

$$s_y^2=\frac{1}{8}[(2.3)^2+(2.5)^2+(3.1)^2+(4.2)^2+(2.6)^2$$

$$+(3.0)^2+(5.1)^2+(0.2)^2]-(2.825)^2$$

$$=\frac{80.6}{8}-(2.825)^2=2.094$$

于是这组光速资料的标准差

$$s=1.447$$

通过计算，我们可以理解到(1.3)—(1.5)式的含意是：

1. 若将各数据加同一常数 a，则均值也加同一常数 a，而方差不变。

2. 若将各数据乘同一常数 b，则均值也乘同一常数 b，而方差则乘以 b^2。

当然描述数据之间差异的程度(变异性)，也可以用别的量，下面列举一些有时会遇到的量，并给以一定的说明：

1. **极差** $x_1,\cdots,x_n$ 中的最大值减去最小值，即 $\max\limits_{1\leqslant i\leqslant n}x_i-\min\limits_{1\leqslant i\leqslant n}x_i$ 称为 $x_1,\cdots,x_n$ 的极差。很明显，它反映了数据之间

最大的差距是多少。

2. **变异系数** $x_1, \cdots, x_n$ 的方差 s^2 或标准差 s 都是有单位的量，单位不同时不好比较。为了清除单位的影响，考虑相对的变异性，这时采用变异系数 $s/|\bar{x}|$，对均值取绝对值是为了使变异系数只取正值。

下面我们算一个具体的例子：

例 3 用例 1 的数据算出相应的极差与变异系数。

解 对于正常人

$$\text{极差} = 80.22 - 42.84 = 37.38$$

$$\text{变异系数} = 11.84/58.0863 = 0.2038$$

对于矽肺病人

$$\text{极差} = 96.39 - 65.45 = 30.94$$

$$\text{变异系数} = 10.85/79.03 = 0.1373$$

从极差和变异系数来看，也是健康人的变化幅度大，矽肺病人的数据比较集中。

§ 3 中位数、众数、直方图

对一组数据 $x_1, \cdots, x_n$，是否除 $\bar{x}$ 以外，还可以用别的数来“代表”它呢？回答是肯定的。从 $\bar{x}$ 的定义可以看到，$x_1, \cdots, x_n$ 中个别的很大的值或很小的值，对 $\bar{x}$ 的影响较大。有时这些极端值并不反映真实的情况，但又不能随便剔除它们，这时往往采用**中位数**。将数据 $x_1, x_2, \cdots, x_n$ 按大小次序排列，当 n 是奇数时，居中的一个就是；当 n 是偶数时，居中的两个数取平均就是中位数。当数据较少时，常常先分组（按大小次序），然

而再求比较方便。例如广东省 1961—1965 年这五年的带鱼产量(见 § 1)按大小次序排列是

2288	2911	4108	4397	5373
(1962)	(1963)	(1961)	(1964)	(1965)

因此中位数是 4108(均值是 3815.4)。例如测试得到上海树脂的伸长数据(%)为

21.4　20.9　22.1　19.7　20.4　19.8

按大小排列为

19.7　19.8　20.4　20.9　21.4　22.1

因此中位数是 $\frac{1}{2}(20.4+20.9)=20.65$(均值是 20.72)。

除了中位数外,还会遇到的就是众数。众数这一概念使用**直方图**较易说明,因此我们先介绍直方图。在数据很多时,往往先将数据分组,标明落入第 i 组数据的个数,记为 n_i,称为第 i 组相应的**频数**。根据分组的间隔,在第 i 组相应的间隔上画一矩形,其高度就是该组的频数,这样的图形就称为直方图。下面用一个数例来说明。

下表是上海市中心气象台九十九年 (1884—1982 年)的年降水量的资料(mm)。

将表中数据分成 11 组,各组的范围及相应的频数(即有多少年的降水量是在这个组内)均列在资料下面另一张表内,表下面就是相应的直方图。

从图 1-1 上可以看出当资料总数逐渐增大,分组更细一些时,上述直方图可以趋向于一条连续曲线(在图 1-1 中用虚线表示)。因此直方图可以给我们一个很好的关于这组数据分布