第八章 统计量及其分布

前面一到六章的研究属于概率论的范畴。我们已经看到,随机变量及其概率分布全面地描述了随机现象的统计规律性。在概率论的许多问题中,概率分布通常被假定为已知的，而一切计算及推理均基于这个已知的分布进行，在实际问题中，情况往往并非如此，看一个例子：

例 8。0.1 某公司要采购一批产品，每件产品不是合格品就是不合格品，但该批产品总有一个不合格品率 p .由此,若从该批产品中随机抽取一件，用 X 表示这一件产品的不合格数,不难看出 X 服从一个二点分布 $b(1,p)$, 但分布中的参数 p 却是不知道的。显然， p 的大小决定了该产品的质量，它直接影响采购行为的经济效益。因此,人们会对 p 提出一些问题，比如，

⑴ p 的大小如何决定;

⑵ p 大概落在什么范围内；

⑶能否认为 p 满足设定要求（如 $p \le 0.05$ ）。

诸如例 8。0.1 研究的问题属于数理统计的范畴。接下来我们从统计中最基本的概念—总体和样本开始介绍统计学内容。

§8。1 总体与样本

8.1.1 总体与个体

在数理统计中,称所研究的对象全体为**总体**（或**母体**）.总体中的个元素称为**个体**。若总体中的个体数目有限,则称之为**有限总体**;否则称之为**无限总体**.

例 8.1。1 有一批产品共 1000 个，每个产品可区分为一等、二等、次品。我们要研究这批产品的质量，1000 个产品的等级构成一个总体，每个产品的等级是个体。

例 8.1。2 为考察在某种工艺条件下织出的一批布匹的疵点数，共取 5000 匹布。那么这 5000 匹布中每匹布茨点数的全体构成一个总体，每匹布的各自疵点数是个体.

例 8。1。3 在检查某军工厂生产的一大批炮弹的质量时,若只考察炮弹的射程，那么,这批炮弹中每一颗的射程的全体构成一个总体，每棵炮弹的各自射程是个体。

从例 2、例 3 可见，总体中的元素常常不是指元素本身，而是指元素的某种数量指标。在例 2 中，总体中的元素指每匹布的疵点数，在例 3 中，总体中的元素指每颗炮弹的射程。在例 1 中，如果一等品用“1”表示，二等品用“2”表示，次品用“0”表示，总体中元素是指每个产品的等级指标，同样，总体可看成数“1”、“2"、“0”的集合。从三个例子可以看出，数量指标取同一值的元素可以有几个，也就是每一个值可以重复.总体是一个可重复的（即允许相同）数的集合。在例 1 的个产品中，值为“1”的有 721 个，值为“2”的有 213 个,值为“0”的有 66 个,因此“1”占 $\frac{721}{1000}$，“2”占 $\frac{213}{1000}$，“0”占 $\frac{66}{1000}$。从

数学角度说，**总体**是指所研究的数量指标可能取的各种不同数值的全体，而各种不同数值含有一定的比率。这样一来，若抛开实际背景，总体就是一堆数，这堆数中有大有小，有的出现机会多，有的出现机会少，因此用一个概率分布去描述和归纳总体是恰当的，从这个意义看，**总体就是一个分布**。以后说“从总体中抽样”与“从某分布中抽样”是同一个意思.

总体的数量指标用$\mathbf{X}$ 表示。从总体中随意地取得的一个个体是随机变量，记为X 。显然，随机变量X 所有可能取得的数值就是$\mathbf{X}$ 可能取得不同值得全体.X 的概率分布与总体的分布有什么关系呢？以例 1 为例，随机变量X 的概率分布列为

X	1	2	0
p	721/1000	213/1000	66/1000

与$\mathbf{X}$ 取各种不同值的比率相同，即X 的概率分布与$\mathbf{X}$ 的总体分布相同。这个结论具有普遍性。以后总体数量指标与相应的随机变量都用X 表示，并不严加区分.**总体分布**指相应的随机变量X 的概率分布,可用分布列、分布密度、分布函数具体表现出来.**总体分布的数字特征**指的是相应随机变量的数字特征。

为方便起见，总体数量指标X 有时简称为总体X ，总体X 的分布和数字特征采用概率论中随机变量的相应量的记号.

上面是从总体得到随机变量。反之,从随机变量亦可得到总体.例如,扔一颗骰子出现的点数是随机变量，它可能取得的不同值的全体“1”、“2”、“3”、“4”、“5"、“6”构成一个总体，它的分布随机变量的概率分布。

在有些问题中,我们对每一研究对象可能要观测两个甚至更多个指标，此时可用多维随机向量及其联合分布来描述总体，这种总体称为多维总体，譬如，我们要了解某校大学生的三个指标:年龄、身高、月生活支出,则我们可用一个三维随机向量描述该总体，这是一个三维总体,它是多元分析所研究的对象。

8.1。2 样本

从总体中取得一部份个体,总体中的这一部分个体称为**样本**。取得样本的过程称为**抽样**。一个样本中每一个个体称为**样品**。样本中个体的个数称为**样本容量**。

在数理统计中,采取抽样的方法是**随机抽样法**，即样本中每一个个体（样品）是从总体中被随意地取出来的。随机抽样分重复抽样和非重复抽样两种。以例 1 为例，从1000个产品中抽取一个容量为10的样本，如果随机抽取地抽取一个产品检查后放回,再随机地抽取一个检查后又放回，直至取到10个个体为止,这种方法称为**重复**（或**返回**）**抽样**。如果每取一个检查后不再放回,直至取得10个个体为止,或者一次抽取10个,这种方法称为**非重复**（或**无返回)抽样**。需要指出，随机抽样得到的样本，所含样品是有一定次序的，通常按它被抽到的先后顺序排列。

从总体X 随机抽样得到的样本可以用n 维随机向量表示$(X_1, X_2, \mathrm{L}\ , X_n)$。现在考察它的概率分布。在重复抽样情形,由于每次取出一个个体检查后放回，总体成分不变（总体分布不变）,所以$X_1, X_2, \mathrm{L}\ , X_n$ 是独立同分布的,并且每一个随机变量的分布于总体分布相同。对于非重复抽样，则分两种情形：在有限总体情形，因取出一个个体后改变了总体的成分，所以随机变量$X_1, X_2, \mathrm{L}\ , X_n$ 不相互独立；在无限总体情形，每取出一个个体后并不改变总体的成分，所以随机变量$X_1, X_2, \mathrm{L}\ , X_n$ 仍然是独立同分布的,并且每一随机变量的概率分布

都是总体分布.

在实际情况中，我们有时遇到的是有限总体，而采用无返回抽样.此时，如果样本容量n相对于总体容量N(总体中个体总数）很小，实际上要求$\frac{n}{N}\leqslant 0.1$，可以把X_1,X_2,K,X_n近似地看成独立同分布,而且每个随机变量的分布都是总体分布。

如果样本(X_1,X_2,K,X_n)中各个个体独立同分布,且每一随机变量的概率分布是总体分布，则称它为**简单随机样本**.这种样本数学上比较容易处理。

样本(X_1,K,X_n)是n维随机向量,这是对具体进行一次抽样而言。在抽样后获得它的一组观察值(x_1,K,x_n),称为样本值。为方便起见,有的时候样本与样本值亦可统称为样本.

设总体X的分布函数是$F(x)$，则样本(X_1,X_2,K,X_n)的概率分布函数为

$$F_n(x_1,x_2,\mathsf{L},x_n)=F(x_1)F(x_2)\mathsf{L}\ F(x_n).$$

在总体离散分布情形，设总体分布列为$p(x^{(i)})=\mathsf{P}\{X=x^{(i)}\},i=1,2,\mathsf{K}$（有限个或可列多个）。则样本的概率分布列为

$$\begin{aligned}p_n(x_1,x_2,\mathsf{K},x_n)&=\mathsf{P}\{X_1=x_1,X_2=x_2,\mathsf{K},X_n=x_n\}\\&=p(x_1)p(x_2)\mathsf{L}\ p(x_n),\end{aligned}$$

其中x_1,x_2,K,x_n每一个值都是在X所有可能取的值$x^{(1)},x^{(2)},\mathsf{K}$之中。

在总体连续分布情形，设总体分布密度为$f(x)$,则

$$f(x_1,x_2,\mathsf{K},x_n)=f(x_1)f(x_2)\mathsf{L}\ f(x_n)$$

对应于随机变量X有一个总体，如何讲抽样与样本呢？如果对于随机变量独立重复地做n次试验，所得观察值为一个简单随机样本，那么，进行n次试验观察相当于进行一次抽样，而且这是重复抽样。例如，对靶射击一次得到的环数是一个随机变量，今独立重复地对靶射击7次，可以看作进行一次重复抽样，所得7个环数构成一个样本。

§8。2 样本数据的整理与显示

8。2。1 经验分布函数

设X_1,X_2,K,X_n是取自总体分布函数为$F(x)$的样本，若将样本观测值由小到大进行排列，为$x_{(1)},x_{(2)},\mathsf{K},x_{(n)}$，则称$X_{(1)},X_{(2)},\mathsf{K},X_{(n)}$为有序样本，用有序样本定义函数如下

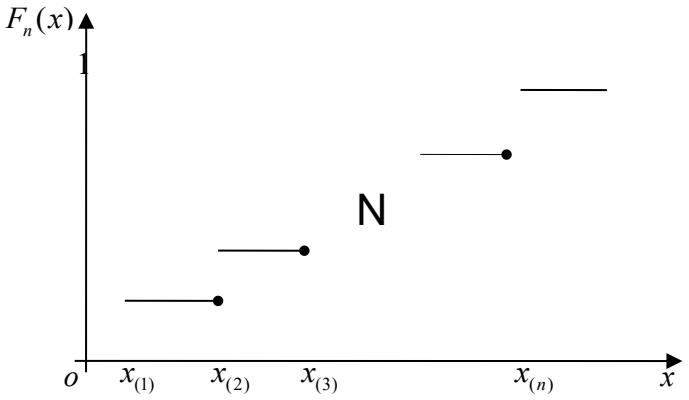
$$F_n(x)=\begin{cases}0, & x<X_{(1)},\\ \dfrac{k}{n}, & X_{(k)}\le x<X_{(k+1)}, k=1,2,\mathrm{K},n-1,\\ 1, & x\ge X_{(n)},\end{cases}$$

则 $F_n(x)$ 是一非减右连续函数，且满足

$$F_n(-\infty)=0, F_n(+\infty)=1.$$

由此可见，$F_n(x)$ 是一个分布函数，并称 $F_n(x)$ 为**经验分布函数**.

经验分布函数的图形如图 8.1



（图 8。1:经验分布函数）

例 8.2。1 某食品厂生产听装饮料，现从生产线上随机抽取 5 听饮料，称得其净重为（单位:克）

351　347　355　344　351

这是一个容量为 5 的样本,经排序可得有序样本：

$$X_{(1)}=344, X_{(2)}=347, X_{(3)}=351, X_{(4)}=351, X_{(5)}=355,$$

其经验分布函数为

$$F_n(x)=\begin{cases}0, & x<344,\\ 0.2, & 344\le x<347,\\ 0.4, & 347\le x<351,\\ 0.8, & 351\le x<355,\\ 1, & x\ge 355,\end{cases}$$

对每一固定的 $x, F_n(x)$ 是样本事件中“ $x_i\le x$ ”发生的频率，当 n 固定时， $F_n(x)$ 是样本的函数，它是一个随机变量，由伯努利大数定律：只要 n 相当大， $F_n(x)$ 依概率收敛于 $F(x)$。更深刻的结果也是存在的，这就是格里纹科定理。

定理 8.2.1（格里纹科定理）当 $n \to +\infty$ 时，经验分布函数 $F_n(x)$ 关于 x 均匀地依概率收敛到 $F(x)$，既对任意 $\varepsilon > 0$

$$\lim_{n\to+\infty} P\{\sup_{-\infty<x<+\infty} |F_n(x) - F(x)| < \varepsilon\} = 1.$$

其中记号“sup”表示上确界.

此定理的证明超出了本课程要求范围,故略去。由定理结论，对任意给定的很小的正数 ε,当 n 充分大时，$\{\sup_{-\infty<x<+\infty} |F_n(x) - F(x)| < \varepsilon\}$ 是大概率事件,即对所有 x,$|F_n(x) - F(x)| < \varepsilon$ 是大概率事件。由实际推断原理，一次抽样后此事件必定发生，因而可用一次抽样获得的 $F_n(x)$ 近似于 $F(x)$，这里的近似对 x 是一致的。需要指出的是，这里要求 n 很大,也就是说抽取的是大样本（容量大的样本)。

8。2。2 频数频率分布表

样本数据的整理是统计研究的基础，整理数据的最常用的方法之一是给出其频数分布表或频率分布表。我们从一个例子开始介绍。

例 8。2.2 为研究某工厂工人生产某种产品的能力，我们随机调查了 20 位工人每天生产的该种产品的数量，数据如下

160	196	164	148	170
175	178	166	181	162
161	168	166	162	172
156	170	157	162	154

对这 20 个数据（样本）进行整理，具体步骤如下：

⑴**对样本进行分组**。首先确定组数 k，作为一般性原则，组数通常在 $5 \sim 20$ 个，对容量较小的样本,通常将其分为 5 组或 6 组，容量为 100 左右的样本可分 7 到 10 组,容量为 200 左右的样本可分 9 到 13 组，容量为 300 左右及以上的样本可分 12 到 20 组，目的是使用足够的组来表示数据的变异。本例中只有 20 个数据,我们将之分为 5 组，即 $k = 5$.

⑵**确定每组组距**。每组区间长度可以相同也可以不同，实用中常选长度相同的区间以便于进行比较，此时各组区间的长度称为组距，其近似公式为

组距 d =（样本最大观测值 - 样本最小观测值）/组数

本例中，数据最大观测值为 196，最小观测值为 148，故组距近似为 $d = 9.6$, 方便起见，取组距为 10.

⑶**确定每组组限**.各组区间端点为 $a_0, a_0 + d = a_1, a_0 + 2d = a_2, \cdots, a_0 + kd = a_k,$，形成如下的分组区间

$$(a_0, a_1], (a_1, a_2], \cdots, (a_{k-1}, a_k],$$

其中 a_0 略小于最小观测值，a_k 略大于最大观测值，本例中可取 $a_0 = 147, a_5 = 197$, 于是本

例的分组区间为：

$$(147,157],(157,167],(167,177],(177,187],(187,197],$$

通常可用每组中的组中值来代表改组的变量取值，组中值＝（组上限＋组下限）/2 。

⑷**统计样本数据落入每个区间的个数—频数,并列出频数频率分布表**。本例的频数频率分布表见下表，从表中可以读出很多信息，如：40% 的工人产量在157 到167 之间；产量少于167 个的有12 人，占60%；产量高于177 的有3 人,占15% .

组序	分组区间	组值	频数	频率	累计频率 / %
1	(147,157]	152	4	0.20	20
2	(157,167]	162	8	0.40	60
3	(167,177]	172	5	0.25	85
4	(177,187]	182	2	0.10	95
5	(187,197]	192	1	0.05	100
合计			20	1	

8.2。3 样本数据的图形显示

前面我们介绍了频数频率分布的表格形式,它也可以用图形表示,这在许多场合更直观。

1.**直方图**

频数分布最常用的图形表示是直方图,它在组距相等场合常用宽度相等的长条矩形表示，矩形的高低表示频数的大小。在图形上,横坐标表示所关心变量的取值区间，纵坐标表示频数，这样就得到频数直方图，若把纵轴改成频率就得到频率直方图。

为使诸长条矩形面积和为1，可将纵轴取为频率组距，如此得到的直方图称为单位频率直方图，或简称频率直方图。此三种直方图的差别仅在于纵轴刻度的选择，直方图本身无变化。

2.**茎叶图**

除直方图外，另一种常用的方法就是**茎叶图**，下面，我们从一个例子谈起。

例 8。2。3 某公司对应聘人员进行能力测试，测试成绩总分为150 分。下面是50 位应聘人员的测试成绩（已经排过顺序)

64	67	70	72	74	76	76	79	80	81
82	82	83	85	86	88	91	91	92	93
93	93	95	95	95	97	97	99	100	100
102	104	106	106	107	108	108	112	112	114
116	118	119	119	122	123	125	126	128	133

我们用这批数据给出一个茎叶图，把每一个数值分为两部分，前面一部分(百位和十位）称为**茎**。后面部分（个位）称为**叶**，如

数值　　分开　　茎和叶

112 → 11|2 → 11和2

然后画一条竖线，**在竖线的左侧写上茎,右侧写上叶，就形成了茎叶图**。应聘人员的茎叶图见图 8.2

6	4 7
7	0 2 4 6 6 9
8	0 1 2 2 3 5 6 8
9	1 1 2 2 3 3 3 5 6 6 7 7 9
10	0 0 2 4 6 6 7 8 8
11	2 2 4 6 8 9 9
12	2 3 5 6 8
13	3

（图 8。2 测试成绩的茎叶图）

茎叶图的外观很象横放的直方图,但茎叶图中叶增加了具体的数值，使我们对数据的具体取值一目了然，从而保留了数据中全部的信息。

在比较两组样本时,可画出它们的**背靠背的茎叶图**,这是一个简单直观而有效的对比方法.

例 8.2。4 下面的数据是某工厂两个车间某天各 40 名员工生产的产品数量（见下表），我们将这些数据放到一个背靠背茎叶图上(图 8.3）

甲车间						乙车间					
50	52	56	61	61	62	56	66	67	67	68	68
64	65	65	65	67	67	72	72	74	75	75	75
67	68	71	72	74	74	75	76	76	76	76	78
76	76	77	77	78	82	78	79	80	81	81	83
83	85	87	88	90	91	83	83	84	84	84	86
86	92	86	93	93	97	86	87	87	88	92	
100	100	103	105			93	95	98	107		

甲车间		乙车间
620	5	6
87775554211	6	67788
877664421	7	2245555666889
8766532	8	011333444667 78
73210	9	22358
5300	10	7

（图 8.3 两车间产量的背靠背茎叶图）

在图 8.3 中，茎在中间，左边表示甲车间的数据，右边表示乙车间的数据。从茎叶图可以看出,甲车间员工的产量偏于上方，乙车间员工的产量大多位于中间，乙车间的平均产量

高于甲车间.乙车间各员工的产量比较集中，而甲车间员工的产量比较分散。

§8。3 统计量及其分布

8.3。1 统计量与抽样分布

样本来自总体，样本的观测值中含有总体各方面的信息，但这些信息较为分散，有时显的杂乱无章。为将这些分散在样本中的有关总体的信息集中起来以反映总体的各种特征，需要对样本进行加工,表和图是一类加工形式，它使人们从中获得对总体的初步认识。当人们需要从样本获得对总体各种参数的认识时，最常用的方法是构造样本的函数，不同的函数反映总体的不同特征。

定义 8.3。1 设 $X_1, X_2, \cdots, X_n$ 为取自总体 X 的样本， 若样本函数

$T = T(X_1, X_2, \cdots, X_n)$ 中不含有任何未知参数，则称 T 为**统计量**，统计量的分布称为**抽样分布**或称**诱导分布**。

按照这一定义，若 $X_1, X_2, \cdots, X_n$ 为样本，则 $\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2$ 以及 8.2。1 节中的 $F_n(x)$ 都是统计量。而当 μ, σ^2 未知时， $X_1 - \mu, X_1/\sigma$ 等均不是统计量。必须指出的是: 尽管统计量不依赖于未知参数，但是它的分布一般都是依赖于未知参数的。下面我们介绍一些常见的统计量及其抽样分布

8。3.2 样本均值及其抽样分布

定义 8.3.2 设 $X_1, X_2, \cdots, X_n$ 为取自总体 X 的样本值，其算术平均值称为样本均值，一般用 $\bar{X}$ 表示，即

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{8.3.1}$$

在分组样本场合,样本均值的近似公式为

$$\bar{X} = \frac{X_1 f_1 + \cdots + X_k f_k}{n} \quad (n = \sum_{k=1}^{k} f_k). \tag{8.3.2}$$

其中 k 为组数， X_i 为第 i 组的组中值， f_i 为第 i 组的频数.

例 8。3。1 某单位收集到 20 名青年人的某月的娱乐支出费用数据：

79	84	84	88	92	93	94	97	98	99
100	101	101	102	102	108	110	113	118	125

则该月这20名青年的平均娱乐支出为

$$\bar{X}=\frac{79+\mathrm{L}\ +125}{20}=99.4$$

将这个数据分组可得到如下频数频率分布：

组序	分组区间	组值	频数	频率/%
1	(77,87]	82	3	15
2	(87,97]	92	5	25
3	(97,107]	102	7	35
4	(107,117]	112	3	15
5	(117,127]	122	2	10
合计			20	100

对上表的分组样本，使用公式(8.3.1)进行计算可得

$$\bar{X}=\frac{1}{20}(82\times3+92\times5+\mathrm{L}\ +122\times2)=100.$$

我们看到两种计算结果的不同,事实上,由于（8.3.2）式未用到真实的样本观测数据，因而给出的是近似结果。

关于样本均值，有如下几个性质。

定理8。3。1若把样本与样本均值之差称为偏差，则样本的所有偏差之和为0，即

$\sum_{i=1}^{n}(X_i-\bar{X})=0$。

证明：

$$\sum_{i=1}^{n}(X_i-\bar{X})=\sum_{i=1}^{n}X_i-n\bar{X}=\sum_{i=1}^{n}X_i-n\cdot\frac{\sum_{i=1}^{n}X_i}{n}=0.$$

证毕。

定理8。3.2样本与样本均值的偏差平方和最小，即在形如$\sum_{i=1}^{n}(X_i-c)$的函数中，

$\sum_{i=1}^{n}(X_i-\bar{X})$最小，其中$c$为任意给定常数。

证明：对任意给定的常数c

$$\sum_{i=1}^{n}(X_i-c)^2=\sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-c)^2$$
$$=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-c)^2+2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-c)$$
$$=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-c)^2\geq\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

证毕。

下面考虑样本均值的分布。现在我们给出关于样本均值抽样分布的一个重要结论.

定理 8.3。3 设 X_1,X_2,K,X_n 是来自总体 X 的样本，$\bar{X}$ 为样本均值。

⑴如总体分布为 $N(\mu,\sigma^2)$，则 $\bar{X}$ 的**精确分布**为 $N(\mu,\sigma^2/n)$;

⑵若总体分布未知或不是正态分布，但 $E(X)=\mu$, 则 n 较大时 $\bar{X}$ 的**渐进分布**为 $N(\mu,\sigma^2/n)$, 常记为 $\bar{X}\ \square\ N(\mu,\sigma^2/n)$.这里渐进分布是指 n 较大时的近似分布.

证明：⑴利用卷积公式，可得知 $\sum_{i=1}^{n}X_i\ \square\ N(n\mu,n\sigma^2)$, 由此可知 $\bar{X}\ \square\ N(\mu,\sigma^2/n)$.

⑵由中心极限定理，$\sqrt{n}(\bar{X}-\mu)/\sigma\xrightarrow{L}N(0,1)$, 这表明 n 较大时的 $\bar{X}$ 渐进分布为 $N(\mu,\sigma^2/n)$.

证毕。

8.3。3 样本方差与样本标准差

定义 8。3.3 设 $X_1,X_2\mathsf{K},X_n$ 为取自总体 X 的样本，则它关于样本均值的平均偏差平方和

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{8.3.3}$$

称为**样本方差**， 其算术根 $S=\sqrt{S^2}$ 称为**样本标准差**。相对样本方差而言，样本标准差通常更有实际意义，因为它与样本均值具有相同的度量单位. 在 n 不大时，常用

$$S^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{8.3.4}$$

作为样本方差（也称**无偏方差**），其算术根 $S^*=\sqrt{S^{*2}}$ 也称为样本标准差，在实际中，S^{*2} 比 S^2 更常用.

在这个定义中，n 为样本量，$\sum_{i=1}^{n}(X_i-\bar{X})^2$ 称为偏差平方和，$n-1$ 称为偏差平方和的自由度. 其含义是：在 $\bar{X}$ 确定后，n 个偏差中 $X_1-\bar{X}, X_2-\bar{X}, \cdots, X_n-\bar{X}$ 只有 $n-1$ 个值可以自由变动，而第 n 个则不能自由取值, 因为 $\sum_{i=1}^{n}(X_i-\bar{X})=0.$

样本偏差平方和有三个不同的表达式：

$$\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}X_i^2-\frac{(\sum_{i=1}^{n}X_i)^2}{n}=\sum_{i=1}^{n}X_i^2-n\bar{X}^2. \tag{8.3.5}$$

它们都可用来计算样本方差。

在分组样本场合，样本方差的近似计算公式为

$$S^{*2}=\frac{1}{n-1}\sum_{i=1}^{k}f_i(X_i-\bar{X})^2=\frac{1}{n-1}[\sum_{i=1}^{k}f_iX_i^2-n\bar{X}^2] \tag{8.3.6}$$

其中 X_i, f_i 为为第 i 个区间的组中值和频数。

例 8。3.2 考察例 8.3。1 的样本, 我们已经计算得 $\bar{X}=99.4$，其样本方差与样本标准差分别为

$$\begin{aligned}S^{*2}&=\frac{1}{20-1}[(79-99.4)^2+(84-99.4)^2+\cdots+(125-99.4)^2]\\&=133.9368\\S^{*}&=\sqrt{133.9368}=11.5731.\end{aligned}$$

下面的定理给出样本均值的数学期望和方差以及样本方差的数学期望，它不依赖于总体的分布形式。

定理 8.3。4 设总体 X 具有二阶矩，即 $E(X)=\mu, D(X)=\sigma^2<+\infty, X_1, \cdots, X_n$ 为从该总体得到的样本，$\bar{X}$ 和 S^{*2} 分别是样本均值和样本方差，则

$$E(\bar{X})=\mu, D(\bar{X})=\sigma^2/n, \tag{8.3.7}$$

$$E(S^{*2})=\sigma^2. \tag{8.3.8}$$

此定理表明, 样本均值的均值与总体均值相同，而样本均值的方差是总体方差的 $n/1.$

证明：由于

$$E(\bar{X})=\frac{1}{n}E(\sum_{i=1}^{n}X_i)=\frac{n\mu}{n}=\mu,$$
$$D(\bar{X})=\frac{1}{n^2}D(\sum_{i=1}^{n}X_i)=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n},$$

故（8。3。7）成立，下证（8。3。8)，注意到

$$\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}X_i^2-n\bar{X}^2,$$

而 $E(X_i^2)=(E(X_i))^2+D(X_i)=\mu^2+\sigma^2, E(\bar{X})^2=(E(\bar{X}))^2+D(\bar{X})=\mu^2+\sigma^2/n,$

于是

$$E(\sum_{i=1}^{n}(X_i-\bar{X})^2)=n(\mu^2+\sigma^2)-n(\mu^2+\sigma^2/n)=(n-1)\sigma^2,$$

两边各除 $n-1$，即得（8.3。8）式。

证毕。

8.3。4 样本矩及其函数

样本均值和样本方差的更一般的推广是样本矩,这是一类常见的统计量.

定义 8。3。4 设 $X_1,X_2,\cdots,X_n$ 是样本，则统计量

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \tag{8.3.9}$$

称为**样本 k 阶原点矩**，特别，样本一阶原点矩就是样本均值。统计量

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k \tag{8.3.10}$$

称为**样本 k 阶中心矩**. 特别，样本二阶中心矩就是样本方差。

当总体分布关于中心对称时，我们用 $\bar{X}$ 和 S^* 刻画样本特征很有代表性。而当其不对称时，只用 $\bar{X}$ 和 S^* 就显得很不够。为此,需要一些刻画分布形状的统计量。这里我们介绍样本偏度和样本峰度，它们都是样本中心矩的函数。

定义 8.3。5 设 $X_1,\cdots,X_n$ 是样本,则称统计量

$$\gamma_1=\frac{B_3}{B_2^{3/2}} \tag{8.3.11}$$

为**样本偏度**。

样本偏度 γ_1 反映了总体分布密度曲线的对称性信息。如果数据完全对称,则不难看出 $B_3=0,$ 对不对称的数据则 $B_3\neq 0.$ 这里用 B_3 除以 $B_2^{3/2}$ 是为了消除量纲的影响, γ_1 是个相对

数，它很好地刻画了数据分布的偏斜方向和程度。

定义 8。3.6 设 X_1,K ,X_n 是样本，则称统计量

$$\gamma_2 = \frac{B_4}{B_2^2} - 3 \tag{8.3.12}$$

为**样本峰度**。

样本峰度 γ_2 反映了总体分布密度曲线在其峰值附近的陡峭程度。当 $\gamma_2 > 0$ 时, 分布密度曲线在其峰值附近比正态分布来得陡，称为尖顶型; 当 $\gamma_2 < 0$ 时，分布密度曲线在其峰值附近比正态分布来得平坦，称为平顶型。

例 8.3.3 下表是两个班（每班 50 名同学）的英语课程的考试成绩。

成绩	组中值	甲班人数 $f_{甲}$	乙班人数 $f_{乙}$
90－100	95	5	4
80－89	85	10	14
70－79	75	22	16
60－69	65	11	14
50－59	55	1	2
40－49	45	1	0

下面我们分别计算两个班级的平均成绩、标准差、样本偏度及样本峰度. 以下两个表分别给出甲班和乙班的计算过程。

甲班成绩的计算过程：

X	$f_{甲}$	$X \cdot f_{甲}$	$(X-\bar{X}_{甲})^2 \cdot f_{甲}$	$(X-\bar{X}_{甲})^3 \cdot f_{甲}$	$(X-\bar{X}_{甲})^4 \cdot f_{甲}$
95	5	475	1843.20	35389.440	679477.2480
85	10	850	846.40	7786.880	71639.2960
75	22	1650	14.08	-11.364	9.0112
65	11	715	1283.04	-13856.832	149653.7856
55	1	55	432.64	-8998.912	187177.3696
45	1	45	948.64	-29218.112	899917.8496
和	50	3790	5368	-8908.8	1987874.56

乙班成绩的计算过程：

X	$f_{乙}$	$X \cdot f_{乙}$	$(X-\bar{X}_{乙})^2 \cdot f_{乙}$	$(X-\bar{X}_{乙})^3 \cdot f_{乙}$	$(X-\bar{X}_{乙})^4 \cdot f_{乙}$
95	4	380	1474.56	28311.552	543581.7984
85	14	1190	1184.96	10901.632	100295.0144
75	16	1200	10.24	-8.192	6.5536
65	14	910	1632.96	-17635.968	190468.4544
55	2	110	865.28	-17997.824	374354.7392
和	50	3790	5168	3571.2	1208706.56

可算得两个班的平均成绩、标准差、样本偏度、样本峰度分别为：

$$\bar{X}_{甲}=\frac{3790}{50}=75.8,\bar{X}_{乙}=\frac{3790}{50}=75.8,$$

$$S^*_{甲}=\sqrt{\frac{5368}{49}}=10.47,S^*_{乙}=\sqrt{\frac{5168}{49}}=10.27,$$

$$\gamma_{1甲}=\frac{-8908.8/50}{(5368/50)^{3/2}}=-0.16,\gamma_{1乙}=\frac{3571.2/50}{(5168/50)^{3/2}}=0.068,$$

$$\gamma_{2甲}=\frac{1987874.56/50}{(5368/50)^2}-3=0.45,\gamma_{2乙}=\frac{1208706.56/50}{(5168/50)^2}-3=-0.74.$$

由此可见，两个班的平均成绩相同，标准差也几乎相同，样本偏度分别为 -0.16 和 0.068，显示两个班的成绩基本对称的。但两个班的样本峰度明显不同，乙班的成绩分布比较平坦，而甲班则稍显尖顶。

8。3.5 次序统计量及其分布

除了样本矩以外，另一类常见的统计量是次序统计量，它在实际和理论中都有广泛的应用。

㈠定义

定义 8。3。7 设 X_1,K,X_n 是取自总体 X 的样本，$X_{(i)}$ 称为该样本的**第 i 个次序统计量**，它的取值是将样本观测值 x_1,K,x_n 由小到大排列后得到的第 i 个观测值。$(X_{(1)},\mathsf{K},X_{(n)})$ 称为该样本的**次序统计量**，$X_{(1)}$ 称为该样本的**最小次序统计量**，$X_{(n)}$ 称为该样本的**最大次序统计量**。

我们知道，在一个（简单随机）样本中，X_1,K,X_n 是独立同分布的，而次序统计量 $(X_{(1)},\mathsf{K},X_{(n)})$ 则即不独立，分布也不相同。如下例

例 8.3。4 设总体 X 的分布为仅取 $0,1,2$ 得离散均匀分布，分布列为

X	0	1	2
p	1/3	1/3	1/3

现从中抽取容量为 3 的样本，其一切可能取值有 $3^3=27$ 种，现将它们列在下表，其右侧是相应的次序统计量观测值。

样本取值及其次序统计量取值

$X_1\ X_2\ X_3$	$X_{(1)}\ X_{(2)}\ X_{(3)}$	$X_1\ X_2\ X_3$	$X_{(1)}\ X_{(2)}\ X_{(3)}$

0	0	0	0	0	0	1	2	0	0	1	2
0	0	1	0	0	1	2	1	0	0	1	2
0	1	0	0	0	1	0	2	2	0	2	2
1	0	0	0	0	1	2	0	2	0	2	2
0	0	2	0	0	2	2	2	0	0	2	2
0	2	0	0	0	2	1	1	2	1	1	2
2	0	0	0	0	2	1	2	1	1	1	2
0	1	1	0	1	1	2	1	1	1	1	2
1	0	1	0	1	1	1	2	2	1	2	2
1	1	0	0	1	1	2	1	2	1	2	2
0	1	2	0	1	2	2	2	1	1	2	2
0	2	1	0	1	2	1	1	1	1	1	1
1	0	2	0	1	2	2	2	2	2	2	2
2	0	1	0	1	2						

由于样本取上述每一组观测值的概率相同，都为$1/27$，由此可给出的分布列如下

$X_{(1)}$	0	1	2
p	$\frac{19}{27}$	$\frac{7}{27}$	$\frac{1}{27}$

$X_{(2)}$	0	1	2
p	$\frac{7}{27}$	$\frac{13}{27}$	$\frac{7}{27}$

$X_{(3)}$	0	1	2
p	$\frac{1}{27}$	$\frac{7}{27}$	$\frac{19}{27}$

我们可以清楚地看到这三个次序统计量的分布是不相同的。

进一步,我们给出两个次序统计量的联和分布，如$X_{(1)}, X_{(2)}$的联合分布列为

$X_{(1)}$ \ $X_{(2)}$	0	1	2
0	$\frac{7}{27}$	$\frac{9}{27}$	$\frac{3}{27}$
1	0	$\frac{4}{27}$	$\frac{3}{27}$
2	0	0	$\frac{1}{27}$

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/288115101062006052