第三章　统计资料的综合

第二章主要介绍了统计资料的整理，得到分布表和统计图等。这当然已能使我们对一份统计资料（数据）有概括地了解。但为了更进一步综合地说明统计资料的特征，以及为了与类似问题进行比较等，有必要用一些数值将资料的特征表示出来，这样的数值称为特征数。我们在本章只考虑单变量的问题，将介绍三类特征数：表示集中位置的特征数，表示变异（分散）程度的特征数和表示偏倚程度的特征数。

3.1　表示集中位置的特征数

3.1.1 平均数

1、 算术平均数（Arithmetic average）

（1）定义

一组 n 个观测值 $x_1, x_2, \cdots, x_n$ 的算术平均数，定义为 $\overline{x}$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3\text{-}1)$$

如果资料已经分组，组数为 k，用 $x_1, x_2, \cdots, x_k$ 表示各组中点，$f_1, f_2 \cdots, f_k$ 表示相应的频数，那么

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \qquad (3\text{-}2)$$

当然，各组中数值都用中点值代替了，所得结果只能是近似的。因此，在求算术平均数时应尽可能用分组前的原始数据。

[例 3.1]某学院一年级学生有 200 人，二年级学生 150 人，三年级学生 100 人，四年级学生有 200 人。某日学院召开全院大会，一年级学生缺席 4%，二年级学生缺席 6%，三年级学生缺席 5%，四年级学生缺席 8%，试问全院学生缺席百分之几？

表 3-1 某学院学生缺席状况表

X	f
4	200
6	150
5	100
8	200

$$\bar{x}=\frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

$$\bar{x}=\frac{4\times200+6\times150+8\times200+5\times100}{200+150+100+200}$$

$$=\frac{3800}{650}=5.85$$

∴该日全院学生缺席 5.85%

[例 3.2]

表 3-2 某校 125 位大学一年级新生体重表

体重（公斤）	组中值(x)	人数(f)
46—48	47	4
49—51	50	20
52—54	53	25
55—57	56	38
58—60	59	21
61—63	62	12
64—66	65	5

平均体重：$\bar{x}=\frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$

$$=\frac{6949}{125}=55.592$$

（2）性质

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

② $\sum_{i=1}^{n}\left(x_i-A\right)^2$ 当 $A=\bar{x}$ 时最小

$$\because \sum_{i=1}^{n}\left(x_i-A\right)^2=\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)+\left(\bar{x}-A\right)\right]^2$$

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

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

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

$\therefore$ 当 $\bar{x}-A=0$ 时， $\sum_{i=1}^{n}\left(x_i-A\right)^2$ 为最小

2、 几何平均数（Geometric Mean）

在数据为环比类型的问题中，算术平均数是不适用的。例如表 3-3 是天津市工农业总产值在“六五”期间的逐年增长率，如求该期间平均增长率，算术平均数是不恰当的。几何平均数可以解决这个问题。

表 3-3 天津市工业总产值

年份	比上年增长%
2000	
2001	14.0
2002	19.6
2003	24.1
2004	31.0
2005	20.8

（天津市统计年鉴 2005 年）

（1） 定义

一组 n 个数据 $r_1, r_2, ..., r_n$ 的几何平均数 G 定义为：

$$G = \sqrt[n]{r_1 r_2 \cdots r_n} \qquad (3\text{-}3)$$

在上例中，令 $r_1, r_2, ..., r_5$ 依次为 105.9，106.9，108.2，111.6，115.1 于是几何平均数

$$\sqrt[5]{105.9 \times 106.9 \times 108.2 \times 111.6 \times 115.1} = 109.5$$

（2）几何平均数性质

设观测的数据是 $s_0, s_1, s_2, ..., s_n$（在上例中为各年工农业总产值），令

$$r_i = \frac{s_i}{s_{i-1}} \qquad (i = 1,2,...,n)$$

在上例中为各年产值与上年的比率，那么

$$s_n = s_0 \cdot \frac{s_1}{s_0} \cdot \frac{s_2}{s_1} \cdots \frac{s_n}{s_{n-1}} = s_0 \cdot r_1 \cdot r_2 \cdots r_n$$

由 $G = \sqrt[n]{r_1 r_2 \cdots r_n}$ 得

$$s_n = s_0 G^n$$

与 $n\bar{x} = \sum x_i$ 相对照，可以看出，当 $r_1, r_2, ..., r_n$ 是环比的情形，几何平均数作为它们的平均数是恰当的。

3、 调和平均数

如果数据是相对变化率时，对其求平均数，用算术平均数也是不恰当的。

例如甲乙两地相距 120 公里，某人乘车往返甲乙两地之间，去时速度为每小时 20 公里，回来时速度为每小时 30 公里，若求平均速度，这时用算术平均数 $\frac{1}{2}(20+30)$ 是不对的，但调和平均数可解决此类问题。

（1）定义

一组 n 个数据 $R_1, R_2, ..., R_n$ 的调和平均数 H，由下式定义：

$$\frac{1}{H}=\frac{1}{n}\left(\frac{1}{R_1}+\frac{1}{R_2}+\cdots+\frac{1}{R_n}\right) \qquad (3\text{-}4)$$

在上例中，$\frac{1}{H}=\frac{1}{2}\left(\frac{1}{20}+\frac{1}{30}\right)=\frac{1}{24}$，$H=24$（公里/小时）

（2）性质

设 R 表示两个变量 M, N 的相对变化率

$$R_i=\frac{M_i}{N_i} \qquad (i=1,2,\cdots,n)$$

n 组观测值的总平均相对变化率为 $\sum M_i/\sum N_i$

$$\sum M_i=\text{（平均变化率）}\times\sum N_i=\text{（平均变化率）}\times\sum\frac{M_i}{R_i}$$

当 $M_1=M_2=\cdots=M_n$ 时，

即得 $\frac{1}{H}=\frac{1}{n}\left(\frac{1}{R_1}+\frac{1}{R_2}+\cdots+\frac{1}{Rn}\right)$

就是说，一组表示两个变量相对变化率的数据，如果分子一定，那么总的平均变化率就应由调和平均数来表示。

几何平均数和调和平均数都只适用于特定问题，选用时要注意。当数据中出现零和负数时不能求几何平均数和调和平均数。

3.1.2 众数（Mode）

算术平均数表示了集中位置特征，它照顾到每一个值，但它不见得是出现次数最多的值（甚至也可能不是观测值中的一个）。所以有必要研究表示集中位置的其它的特征数。

定义：对于有频数分布的变量，它的众数指频数最大的变量的值。

表 3-4 频数分布表

x	f
3	15
5	2
7	3

对于已分组且等组距的频数分布，根据最大频数，可求得众数所在组。根据众数定义，可知众数不唯一。

3.1.3 中位数（Median）

算术平均数作为集中位置的特征还有一缺点，就是受观测值中极端值的影响很大，而一组观测值中的极端值常常没有代表性。中位数将避免这种影响。

（1）定义

一组 n 个观测值按数值大小排列如 $x_1, x_2, \ldots, x_n$，处于中央位置的值称为中位数，以 Me 表示，即

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{，当 n 为奇数} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) & \text{，当 n 为偶数} \end{cases} \qquad (3\text{-}5)$$

（2）性质

①一组观测值其中小于 Me 的个数和大于 Me 的个数相等（可与算术平均数性质（2）对比）。Me 是累积频率为 0.50 所对应的 X 的值，如图 3-1 所示。

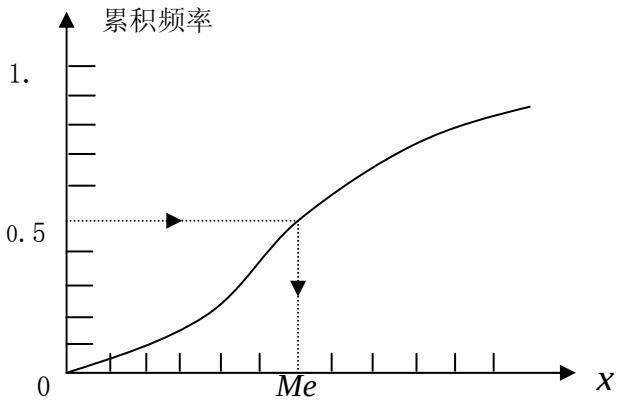

图 3-1

这是在观测值没有重复的情形下中位数的性质。如果观测值中重复数值很多，这一性质就不一定成立。例如 9 名学生的成绩是：

95 90 90 85 85 85 85 80 75

那么 Me=85。大于 Me 的有 3 个，小于 Me 的有 2 个，个数不等。

②我们还常用到各 x_i 与某一定值 A 的离差的绝对值（也称绝对离差）之和

$$\sum_{i=1}^{n}|x_i - A|$$

中位数 Me 有这样的性质：上述绝对值离差和以 $A = Me$ 时为最小。中位数具有的这种最小性说明它同样表示了集中位置的特征。

3.1.4 百分位数（ Percentile）

（1）定义

一组 n 个观测值按数值大小排列，如 $x_1, x_2, \cdots, x_n$，处于 $p\%$ 位置的值称第 p 百分位数。

中位数是第 50 百分位数。

第 25 百分位数又称第一个四分位数（First Quartile），用 Q_1 表示；第 50 百分位数又称第二个四分位数（Second Quartile），用 Q_2 表示；第 75 百分位数又称第三个四分位数（Third Quartile），用 Q_3 表示。若求得第 P 百分位数为小数，可向上取整为整数。

分位数是用于衡量数据位置的量度，但它所衡量的，不一定是中心位置。百分位数提供了有关各数据项如何在最小值与最大值之间分布的信息。对于无大量重复的数据，第 p 百分位数将数据分为两个部分：大约 P%的数据项的值比第 p 百分位数小；而大约（100－p）%的数据项的值比第 p 百分位数大。对第 p 百分位数，严格的定义如下：

第 P 百分位数是这样一个值，它使得至少有 p%的数据项小于或等于这个值，且至少有（100－p）%的数据项大于或等于这个值。

高等院校的入学考试成绩经常以百分位数的形式报告。比如，假设某个考生在入学考试中语文部分的原始分数为 54 分，相对于参加同一考试的其

他学生来说，他的成绩如何并不容易知道。但是如果原始分数 54 分恰好对 30%的学生考分比他高。

（2）计算第 p 百分位数的步骤

第 1 步：以递增顺序排列原数据（即从小到大排列）。

第 2 步：计算指数 $i = np\%$

其中，p 是所求的百分位数的位置，n 是项数。

第 3 步：①若 i 不是整数，将 i 向上取整。大于 i 的毗邻整数即为第 p 百分位数的位置。

②若 i 是整数，则第 P 百分位数是第 i 项与第（i＋1）项数据的平均值。

3.1.5 四分位数

人们经常会将数据划分为 4 个部分，每一个部分大约包含有 1/4 即 25%的数据项。这种划分的临界点即为四分位数。它们定义如下：

Q_1=第 1 四分位数，即第 25 百分位数

Q_2=第 2 四分位数，即第 50 百分位数

Q_3=第 3 四分位数，即第 75 百分位数

下面是按递增顺序排列的起始起薪数据。Q_2 即第 2 四分位数（中位数），已被确知为 2405。

2210 2255 2350 2380 2380 2390 2420 2440 2450 2550 2630 2825

计算 Q_1 和 Q_3 需要用到计算第 25 百分位数和第 75 百分位数的方法。它们的计算如下：

对 Q_1：

i=(p/100)n=(25/100)12=3

由于 i 为整数，由第 3 步的（2）可知，第 1 四分位数即第 25 百分位数即为第 3 项与第 4 项的平均值。所以 Q_1=（2350＋2380）/ 2=2365。

对于 Q_3：

i=(p/100)n=(75/100)12=9

同样 i 为整数，由第 3 步的（2）可知，第 3 四分位数（即第 75 百分位数）为第 9 项与第 10 项的平均值。所以 Q_3=（2455+2550）/2 =2500。

如下所示，四分位数将 12 个数据分为了 4 个部分，每个部分含有 25%的数据项。

2210 2255 2350 ｜ 2380 2380 2390 ｜ 2420 2440 2450 ｜ 2550 2630 2825

Q_1＝2365　　Q_2＝2405（中位数）　　Q_3＝2500

我们已将四分位数分别定义为第 25，50，75 百分位数。因此，四分位数的计算方法与其他百分位数的计算方法是相同的。但是在计算四分位数时有些方法的约定是不同的；而计算出来的值也会因这些约定的不同而稍有差异。尽管如此，无论采用何种计算过程，计算四分位数的目的都是将数据划分为大致相等的 4 个部分。

3.2 表示变异（分散）程度的特征数

一组数据，即对变量的一组观测值，除用算术平均数等表示它的集中位置的特征外，各观测值的相互之间变异情况，或说分散情况，也是一组数据的一个重要特征。如果这组数据是产品质量检查的结果，那么数据的变异情况说明生产是否稳定；如果数据是测量的结果，那么变异的情况说明测量方法或仪器是精密还是粗糙；如果数据是学生的成绩，那么变异的情况说明成绩是否整齐（而不是高低）。

3.2.1 极差（或称全距 Range）R

定义 极差

$$R = x_{\max} - x_{\min} \tag{3-6}$$

其中 $x_{\max}$ 和 $x_{\min}$ 分别为数据中的极大值和极小值。

可以看出，极差的计算非常简便，所以在现场检查时经常用到。但是极差没有考虑各中间值。

3.2.2 四分位数间距

能够克服极端值影响的一种衡量变异程度的量度是四分位数间距（IQR）。这种衡量变异程度的量度即为第 3 四分位数 Q_3 与第 1 四分位数 Q_1 的差。也就是说，四分位数间距是在中间的 50％的数据的全距。

四分位数间距：

$$IQR = Q_3 - Q_1$$

3.2.3 平均差（Mean Absolute Deviation）

定义 平均差 M.D.是离差的绝对值的平均数，即

$$M.D.=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\bar{x}\right| \qquad (3\text{-}7a)$$

对于已分组的频数分布（组数为 k）

$$M.D.=\frac{1}{n}\sum_{i=1}^{k}f_i\left|x_i-\bar{x}\right| \qquad (3\text{-}7b)$$

平均差虽然能较好地区别出不同组数据的分散情况或程度，但它的缺点是绝对值不适于作进一步的数学分析。

3.2.4 方差（Variance），标准差（Standard Deviation）

定义 方差

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

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

对于已分组的频数分布（组数为 k）

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

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

方差可以起到与平均差相同的作用，且避免了绝对值的运算。但方差也有缺点，那就是所取的单位为 X 的单位的平方（如：当 X 为身高，单位为厘米时，方差的单位为平方厘米；当 X 为成绩，单位为分时，方差单位为分的平方，这都是不好解释的），所以用它的正平方根，作为标准差。

定义 标准差

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

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

对于已分组的频数分布（组数为 k）

总体 $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}$

样本 $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}$

标准差的单位与 X 的单位相同。

3.2.5 变异系数（Coefficient of Variation）

定义　变异系数 C

$$C = \frac{S}{\bar{x}} \times 100(\%) \qquad (3\text{-}8)$$

是一个无量纲的量。它适于用在比较有不同算术平均数或有不同量纲的两组数据的情况。例如比较大学生身高与小学生身高，或比较 130 名大学生身高和体重哪个变化波动范围比较大时，都可用变异系数。

3.3 表示偏倚情况或程度的特征数

偏倚性用以表示各观测值分布的不对称情况或程度。

3.3.1 比较众数、中位数和算术平均数的相对位置

图 3-2 举出了对称的、具有左偏态（负偏态）和右偏态（正偏态）的频数分布的例子。注意到它们的特点是：①对称的分布的众数、中位数和算术平均数相同；②具有偏倚性的分布，算术平均数突出在外，偏向分布的尾端，而中位数则介于众数与算术平均数之间。

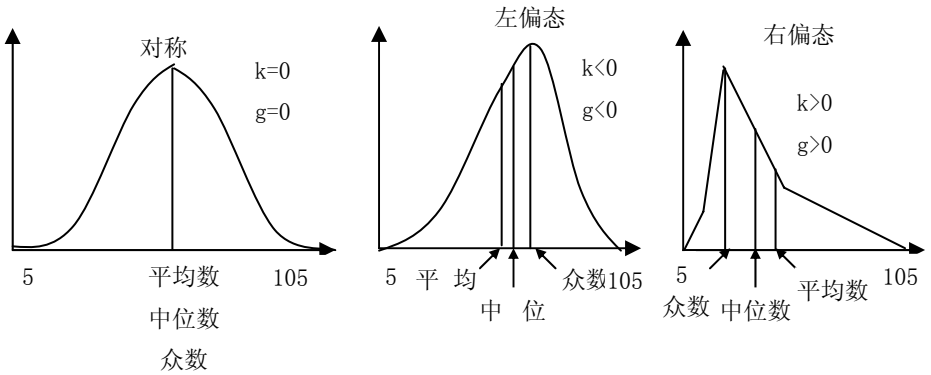

图　3-2

可以看出，对于单峰的分布，

对称态： $\bar{x}$ =Me=Mo

左偏态： $\bar{x}$ <Me<Mo

右偏态： $\bar{x}$ >Me>Mo

3.3.2 定量地描述偏倚性，常用的两个公式

（1）Pearson 偏倚系数

$$k = \frac{3\left(\bar{x} - Me\right)}{s} \qquad (3\text{-}9a)$$

如分布是对称的，则 k=0；若具有左偏态，平均数小于中位数，所以 k<0；若具有右偏态，平均数大于中位数，k>0。

（2）用标准化的三阶矩阵 g 表示

$$g = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3 / n}{s^3} \qquad (3\text{-}9b)$$

k 和 g 都不受 x 的单位的影响。

注意总体和样本的区别。如果所研究的统计资料是一个样本，那么本章的各特征数的名称都应当冠以“样本“二字，如“样本平均数”、“样本方差”等等，以便与总体的相应特征数相区别。

3.4 五数概括法

五数概括法即用下面的五个数来概括数据：

（1）最小值。

（2）第 1 四分位数（Q_1）。

（3）中位数（Q_2）。

（4）第 3 四分位数（Q_3）。

（5）最大值。

运用五数概括法的最简单方式是首先将数据按递增顺序排列，然后很容易就能确定最小值、3 个四分位数和最大值了。对 12 个起薪数据的样本，按照递增顺序排列如下：

2210 2255 2350 | 2380 2380 2390 | 2420 2440 2450 | 2550 2630 2825

Q_1＝2365　　Q_2＝2405　　Q_3＝2500

（中位数）

中位数 2405 以及四分位数 Q_1＝2365 和 Q_3＝2500 前面已经计算出来了。对上述数据的观察可以知道最小值为 2210，最大值为 2825。因此，上述起薪数据以五数概括为：2210，2365，2405，2500，2825。在相邻的每两个数之间，大约有 1 / 4 或 25%的数据项。

3.5 盒形图

盒形图实际上是以图形来概括数据。我们将盒形图延至这一章才介绍是因为它的关键是计算中位数和四分位数 Q_1 和 Q_3。此外还将用到四分位数间距 IQR＝Q_3－Q_1。

盒形图的画法步骤如下：

（1）画一个方盒，其边界恰好是第 1 和第 3 四分位数。对于上述的起薪数据，　Q_1＝2365，　Q_3＝2500。

这个方盒包含了中间的 50%的数据。

（2)在方盒上中位数的位置画一条垂线(对起薪数据,中位数为 2405)。因此中位数将数据分为相等的两个部分。

（3）利用四分位数间距 IQR=Q_3－Q_1，来设定界限。盒形图的界限定于低于 Q_1 以下 1．5 个 IQR 和高于 Q_3 以上 1．5 个 IQR 的位置。上述的起薪数据中，IQR＝Q_3－Q_1＝2500－2365=135。因此，上、下限分别为：2365－1.5x135=2162．5 和 2500＋1.5x135＝2702.5，上、下限以外的数值作为异常值。

（4）在图 3－4 中的横线叫做须线（whisker)，须线从方盒的边线出发，直至在上、下限之内的最大值和最小值。对于起薪数据，其须线止于 2210 和 2630。

（5）最后，任一异常值的位置以符号“＊”标出。在图 3－4 中可以看到一个异常值——2825。

图 3-3 是起薪数据的盒形图，在图中，我们用一些竖线显示上、下限的位置。这些竖线用来表明对于起薪数据，上、下限是如何计算出来的，但是一般情况下它们并不在盒形图上画出。在图 3-4 中显示了正常情况下的起薪数据盒形图的外观。

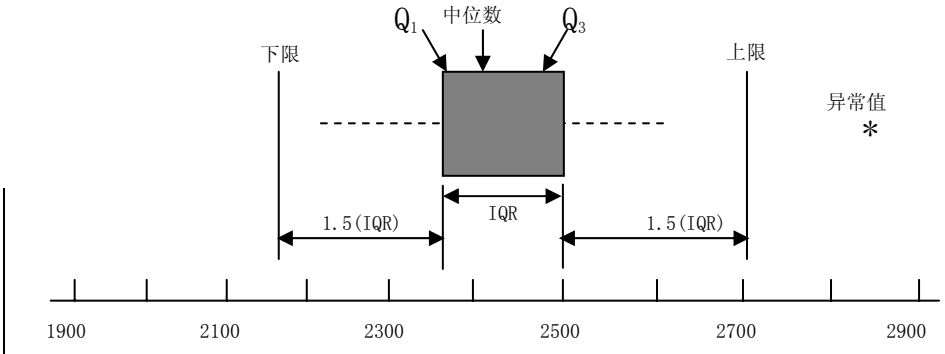

图 3-3

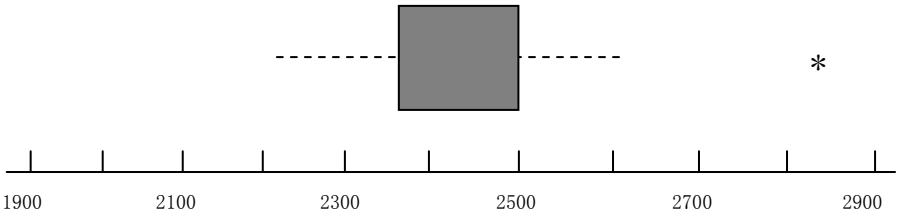

图 3-4

习题

1．随着电脑的广泛应用，越来越多的人在工作时需要使用电脑。下面是一个样本，它显示的是工作中需要使用电脑的人的年龄数据。

22	58	24	50	29	52	57	31	30	41
44	40	46	29	31	37	32	44	49	29

a. 计算平均数和众数。

b．计算第 1 和第 3 四分位数。

d．计算第 32 百分位数并解释其含义。

2. 北京作为为世界旅游胜地，吸引着成千上万的国内外游客。下面的数据来自 2002 年 10 月中的若干天来北京观光的旅游者的数据，是一个具有代表性的数据样本。其中旅游人数以千人计。

来自国内的：

108.70	112.25	94.01	144.03	162.44	161.61	76.20
102.11	110.87	79.36	129.04	95.16	114.16	121.88

来自国外的：

29.89	41.13	40.67	40.41	43.07	24.86
31.61	21.60	27.34	64.57	32.98	41.31

a．计算来自国内和国外的旅游者的每日人数的平均数和中位数。

b．计算来自这国内和国外的旅游者的每日人数的全距、标准差和标准差系数。

c．对于来自这国内和国外的旅游人数进行比较。

3．假设下面的数据是A和B两个公司供货所需天数。

A：	11	10	9	10	11	11	10	11	10	10
B：	8	10	13	7	10	11	10	7	15	12

利用全距和标准差来证明前面的结论，即Dawson Supply公司在供货时间上更具有一致性和可靠性。

4. 一名板球投球手在6局比赛中的得分为：182，168，184，190，170和174。以上述数据作为一个样本，计算下列的描述统计量。

a．全距

b．方差

c．标准差

d．变异系数

5．某生产部门利用一种抽样程序来检验新生产出来的产品的质量，该部门使用下面的法则来决定检验结果：如果一个样本中的14个数据项的方差大于0.005，则生产线必须关闭整修。假设搜集的数据如下：

3.43	3.45	3.43	3.48	3.52	3.50	3.39
3.48	3.41	3.38	3.49	3.45	3.51	3.50

问此时的生产线是否必须关闭？为什么？

6．下面是20个长途电话通话时间的频数分布，计算该数据的平均数、方差

和标准差。

通话时间/分钟	频数	通话时间/分钟	频数
4-7	4	20-23	1
8-11	5	24-27	1
12-15	7	合计	20
16-19	2		