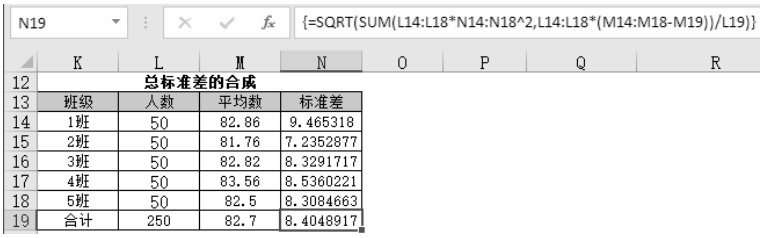


	K	L	M	N	O	P	Q	R
12		总标准差的合成						
13	班级	人数	平均数	标准差				
14	1班	50	82.86	9.465318				
15	2班	50	81.76	7.2352877				
16	3班	50	82.82	8.3291717				
17	4班	50	83.56	8.5360221				
18	5班	50	82.5	8.3084663				
19	合计	250	82.7	8.4048917				

图 4-44　两类数据标准差的计算结果

【函数公式解析】

在 N19 单元格的数组公式中，“SUM(L14:L18*N14:N18^2)”一段相当于标准差合成公式的 $\sum N_i S_i^2$ 部分，“SUM(L14:L18*(M14:M18−M19))”一段相当于“$\sum N_i \left(\overline{X}_i - \overline{X}_T\right)^2$”部分。这 2 段可以当成 SUM 的 2 个参数，因而可以合并成为“SUM(L14:L18*N14:N18^2, L14:L18*(M14:M18−M19))”，就是备选公式中的一段。

4.2.6　变异系数

变异系数（Coefficient of Variation）也称差异系数，是标准差与平均数的比值，是衡量资料中各观测值变异程度的另一个统计量，用 CV 来表示。当进行两个或多个资料变异程度的比较时，如果度量单位与平均数相同，可以直接利用标准差来比较。如果两个或多个资料的测量尺度相差太大，或者数据量纲不同，直接使用标准差来进行比较不合适，此时就应当消除测量尺度和量纲的影响，而变异系数可以做到这一点。变异系数没有量纲，这样就可以进行客观比较了。变异系数的大小，同时受平均数和标准差两个统计量的影响，因而在利用变异系数表示资料的变异程度时，最好将平均数和标准差也列出。公式为

$$\mathrm{CV} = \frac{S}{\overline{X}}$$

例 4-21　在例 4-15 的基础上，分别按原始成绩和频数分布表计算变异系数。

解题思路：计算变异系数，必须要已知标准差与平均数。在 Excel 中，对于未分组数据和分组数据，都要运用前面介绍的方法计算出标准差与平均数，再计算变异比值。

解题过程：由于例 4-15 中已建立统计表，这里直接在表中输入公式。

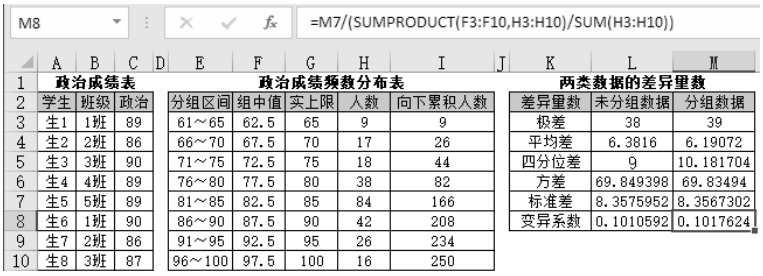
在 L8 单元格输入公式“=STDEV.S(C3:C252)/AVERAGE(C3:C252)”。

在 M8 单元格输入公式“=M7/(SUMPRODUCT(F3:F10,H3:H10)/SUM(H3:H10))”。

计算结果如图 4-45 所示。

【函数公式解析】

在 M8 单元格的公式中，“(SUMPRODUCT(F3:F10,H3:H10)/SUM(H3:H10))”一段为根据频数分布表计算加权算术平均数。



M8 =M7/(SUMPRODUCT(F3:F10,H3:H10)/SUM(H3:H10))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	政治成绩表				政治成绩频数分布表						两类数据的差异量数		
2	学生	班级	政治		分组区间	组中值	实上限	人数	向下累积人数		差异量数	未分组数据	分组数据
3	生1	1班	89		61～65	62.5	65	9	9		极差	38	39
4	生2	2班	86		66～70	67.5	70	17	26		平均差	6.3816	6.19072
5	生3	3班	90		71～75	72.5	75	18	44		四分位差	9	10.181704
6	生4	4班	89		76～80	77.5	80	38	82		方差	69.849398	69.83494
7	生5	5班	89		81～85	82.5	85	84	166		标准差	8.3575952	8.3567302
8	生6	1班	90		86～90	87.5	90	42	208		变异系数	0.1010592	0.1017624
9	生7	2班	86		91～95	92.5	95	26	234				
10	生8	3班	87		96～100	97.5	100	16	250				

图 4-45　两类数据标准差的计算结果

例 4-22　在文件“第 4 章统计量.xlsx”的“差异量数”工作表中，有两个企业 1—6 月的产值数据，如图 4-46 所示。如何比较离散程度？

解题思路：两个企业的月产值差异悬殊，不能直接用标准差比较离散程度，而要用变异系数来比较。要先计算出标准差与平均数，再计算比值。

解题过程：建立表格，输入公式。

（1）建立表格。在原始数据表下扩建表格，如图 4-47 所示。

	K	L	M
21	两个企业的月产值（万元）		
22	月份	砖厂	鞋厂
23	1月	650	100
24	2月	640	120
25	3月	660	110
26	4月	620	115
27	5月	680	130
28	6月	640	90

图 4-46　两个企业 1—6 月的产值

	K	L	M
21	两个企业的月产值（万元）		
22	月份	砖厂	鞋厂
23	1月	650	100
24	2月	640	120
25	3月	660	110
26	4月	620	115
27	5月	680	130
28	6月	640	90
29	标准差		
30	平均值		
31	变异系数		

图 4-47　两类数据的差异量数统计表

（2）输入公式。

在 L29 单元格输入公式“=STDEV.S(L23:L28)”。

在 L30 单元格输入公式“=AVERAGE(L23:L28)”。

在 L31 单元格输入公式“=L29/L30”。

将 L29:L31 区域的公式向右填充到 M31 单元格。

计算结果如图 4-48 所示。

从表中可以看出，砖厂的标准差为 20.41，而鞋厂的标准差为 14.29，砖厂显著高于鞋厂。砖厂的平均值为 648.33，而鞋厂的平均值为 110.33，砖厂远远地高于鞋厂。砖厂的变异系数为 0.03，而鞋厂的变异系数为 0.13，砖厂低于鞋厂。这表明，砖厂的月产值相对稳定。

M31 =M29/M30

	K	L	M	N
21	两个企业的月产值（万元）			
22	月份	砖厂	鞋厂	
23	1月	650	100	
24	2月	640	120	
25	3月	660	110	
26	4月	620	115	
27	5月	680	130	
28	6月	640	90	
29	标准差	20.412415	14.28869	
30	平均值	648.33333	110.83333	
31	变异系数	0.0314844	0.1289205	

图 4-48　两类数据标准差的计算结果

4.2.7　分类的差异量数

前面在介绍极差、平均差、四分位差、方差、标准差、变异系数 6 个差异量数时，所举

例子均没有分类计算差异量数。实际上，这批原始数据可能是有类别的。下面介绍分类计算差异量数的技巧。

例 4-23 利用例 4-15 的原始数据，如何按班计算极差、平均差、四分位差、方差、标准差、变异系数？

解题思路 1：在 Excel 中，使用 IF 函数的条件判断功能结合其他函数，可以分类计算极差、平均差、四分位差、方差、标准差、变异系数。

解题过程：建立统计表，输入公式。

（1）建立统计表。建立一个“未分组数据分类的差异量数”统计表，如图 4-49 所示。

	P	Q	R	S	T	U	V	W
1	未分组数据分类的差异里数							
2	班级	MAX-MIN求极差	MAXIFS-MINIFS求极差	平均差	四分位差	方差	标准差	变异系数
3	1班							
4	2班							
5	3班							
6	4班							
7	5班							

图 4-49 差异量数统计表

（2）输入公式。

在 Q3 单元格公式输入数组公式“{=MAX(IF($B$3:$B$252=P3,$C$3:$C$252))-MIN(IF($B$3:$B$252=P3,$C$3:$C$252))}”。

在 R3 单元格输入公式“=MAXIFS($C$3:$C$252,$B$3:$B$252,P3)-MINIFS($C$3:$C$252,$B$3:$B$252,P3)”。

在 S3 单元格公式输入数组公式“{=SUM(ABS(IFERROR(IF($B$3:$B$252=P3,$C$3:$C$252,"")-AVERAGEIF($B$3:$B$252,P3,$C$3:$C$252),)))/L14}”。

在 T3 单元格公式输入数组公式“{=QUARTILE.INC(IF($B$3:$B$252=P3,$C$3:$C$252),3)-QUARTILE.INC(IF($B$3:$B$252=P3,$C$3:$C$252),1)}”。

在 U3 单元格公式输入数组公式“{=VAR.S(IF($B$3:$B$252=P3,$C$3:$C$252))}”。

在 V3 单元格公式输入数组公式“{=STDEV.S(IF($B$3:$B$252=P3,$C$3:$C$252))}”。

在 W3 单元格公式输入数组公式“{=V3/AVERAGEIF($B$3:$B$252,P3,$C$3:$C$252)}”。

将 Q3:W3 区域的公式向下填充到 W7 单元格。

结果如图 4-50 所示。

W7 =V7/AVERAGEIF($B$3:$B$252,P7,$C$3:$C$252)

	P	Q	R	S	T	U	V	W
1	未分组数据分类的差异里数							
2	班级	MAX-MIN求极差	MAXIFS-MINIFS求极差	平均差	四分位差	方差	标准差	变异系数
3	1班	37	37	6.9112	8.75	89.5922	9.46532	0.11423
4	2班	33	33	5.5984	8	52.3494	7.23529	0.08849
5	3班	31	31	6.4456	8.75	69.3751	8.32917	0.10057
6	4班	37	37	6.6	10	72.8637	8.53602	0.10215
7	5班	35	35	6.4	8.75	69.0306	8.30847	0.10071

图 4-50 计算集中量数的结果

解题思路 2：在 Excel 中，可利用数据透视表强大的统计功能分类统计方差和标准差。

解题过程：插入数据透视表，更改值汇总依据。

（1）插入数据透视表。操作过程为：

❶ 单击“插入”选项卡。

❷ 单击“表格”组中的“数据透视表”按钮。

❸ 在弹出的“创建数据透视表”对话框中，将鼠标放置于“选择一个表或区域”单选按钮右侧的“表/区域”框，使用鼠标拖动选择 A2:C252 区域。

❹ 在“选择放置数据透视表的位置”组中选择“现有工作表”单选按钮。

❺ 将鼠标放置于“位置”框中，使用鼠标单击 P12 单元格。

❻ 单击“确定”按钮，随即生成空白数据透视表并弹出“数据透视表字段”任务窗格。

❼ 在“数据透视表字段”任务窗格，将字段节区域中的“班级”字段拖放到“行”区域节，将“政治”字段分 2 次拖放到“值”区域节。

操作过程及结果如图 4-51 所示。

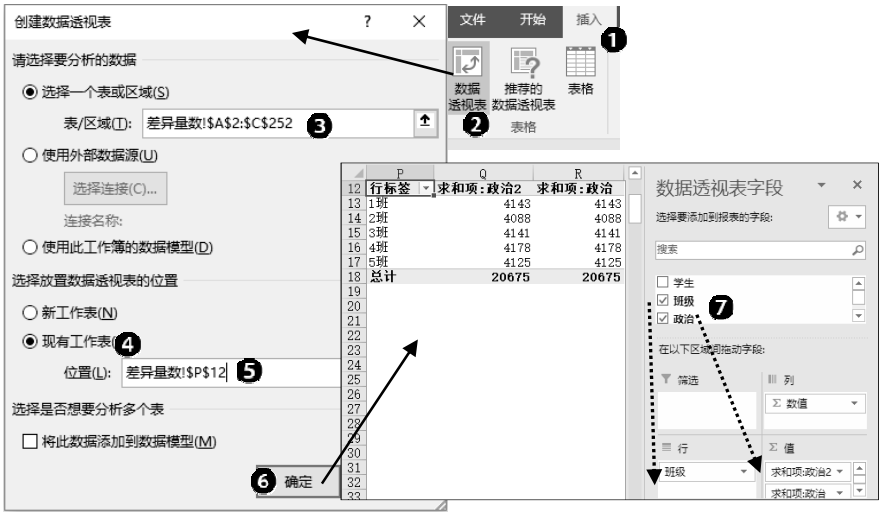


图 4-51　插入数据透视表

（2）更改值汇总依据。操作过程为：

❶ 右击数据透视表“求和项：政治 2”字段的“值”区域中的任意单元格，例如 R13 单元格。

❷ 在快捷菜单中选择“值汇总依据”级联菜单的“其他项”命令。

❸ 在弹出的“值字段设置”对话框中，已默认选择“值汇总方式”选项卡。

❹ 在“计算类型”列表框中选择“方差”选项。

❺ 单击“确定”按钮。

❻ 重复第 1～5 步，区别在于，右击的单元格为“求和项：政治”字段的“值”区域中的任意单元格，例如 S13 单元格；在“值字段设置”对话框的“计算类型”列表框中选择“标准偏差”选项。

操作过程及结果如图 4-52 所示。

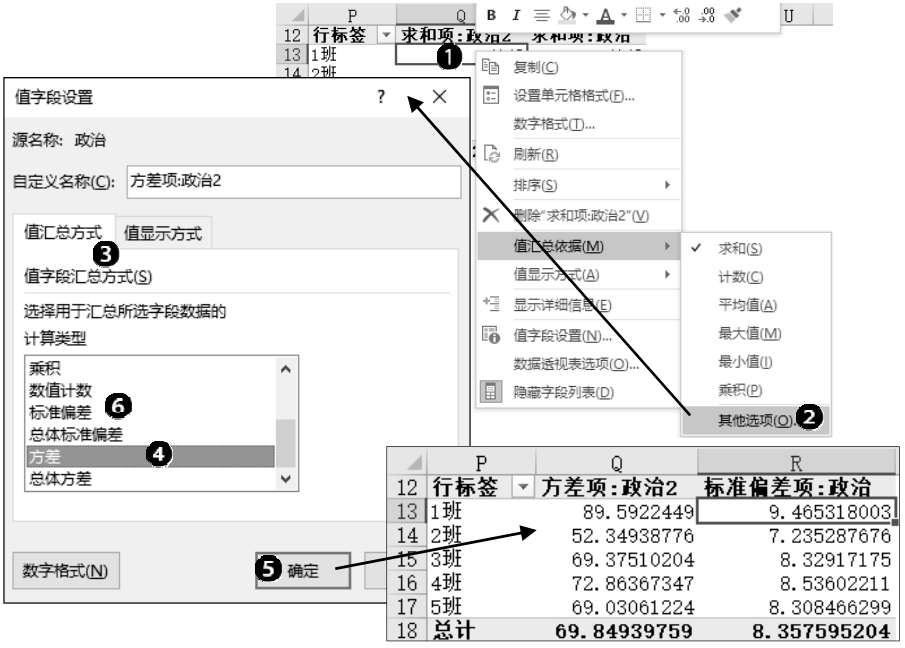


图 4-52 更改值汇总依据

4.3 地位量数

在一组数据或一个次数分布中，每一个数据都有相应的地位，都可用一定的地位量数来说明它所处的位置。反映数据所处地位的量就叫作地位量数。百分位数、百分等级分数、标准分数等都是地位量数。

4.3.1 百分位数

百分位数（Percentile）又称百分位分数，是一种相对地位量数。把一个次数分布排序后，分为 100 个单位，百分位数就是次数分布中相对于某个特定百分点的原始分数，表明在次数分布中特定个案百分比低于该分数。百分位数用 P_m 表示。例如，若 P_{24} 等于 60，则表明在该次数分布中有 24%的个案低于 60 分。

对于未分组数据，Excel 提供了 PERCENTILE.INC 函数计算百分位数。

对于分组数据，可根据累积频数分布表，运用插值法，按比例计算。公式为

$$P_m = L + \frac{\frac{m}{100}n - F_{m-1}}{f_m} \times d$$

式中，L 为 P_m 所在组的下限；n 为总频数；F_{m-1} 为 P_m 所在组以下的累积频数；f_m 为 P_m 所在组的频数；d 为组距。

例 4-24 在文件“第 4 章统计量.xlsx”中的“地位量数”工作表中，有某级 5 个班 250 人的政治成绩表及频数分布表，如图 4-53 所示（隐藏了部分行），如何分别按原始成绩和频数分布表计算前 10%的百分位数？

解题思路：对于未分组数据，可以直接使用 PERCENTILE.INC 函数计算百分位数。对于分组数据，则按百分位数公式计算。

解题过程：建立统计表，输入公式。

（1）建立统计表。建立一个“两类数据的地位量数”统计表（包含后面将要介绍到的百分等级分数），如图 4-54 所示。

	A	B	C	D	E	F	G	H	I	J
1	政治成绩表					政治成绩频数分布表				
2	学生	班级	政治			分组区间	组中值	实上限	人数	向上累积人数
3	生1	1班	89			61～65	62.5	65	9	9
4	生2	2班	86			66～70	67.5	70	17	26
5	生3	3班	90			71～75	72.5	75	18	44
6	生4	4班	89			76～80	77.5	80	38	82
7	生5	5班	89			81～85	82.5	85	84	166
8	生6	1班	90			86～90	87.5	90	42	208
9	生7	2班	86			91～95	92.5	95	26	234
10	生8	3班	87			96～100	97.5	100	16	250
252	生250	5班	68							

图 4-53 政治成绩表及频数分布表

	L	M	N
1	两类数据的地位量数		
2	项目	未分组数据	分组数据
3	前10%的百分位分数		
4	94分的百分等级分数		

图 4-54 两类数据的地位量数表

（2）输入公式。

在 M3 输入公式“=PERCENTILE.INC(C3:C252,1−0.1)”。

在 N3 输入公式“=91+(0.9*J10−J8)/I9*5”。

计算结果如图 4-55 所示（隐藏了 D 列）。

N3 =91+(0.9*J10-J8)/I9*5

	A	B	C	E	F	G	H	I	J	K	L	M	N
1	政治成绩表				政治成绩频数分布表						两类数据的地位量数		
2	学生	班级	政治		分组区间	组中值	实上限	人数	向上累积人数		项目	未分组数据	分组数据
3	生1	1班	89		61～65	62.5	65	9	9		前10%的百分位分数	94	94.2692
4	生2	2班	86		66～70	67.5	70	17	26		94分的百分等级分数		
5	生3	3班	90		71～75	72.5	75	18	44				
6	生4	4班	89		76～80	77.5	80	38	82				
7	生5	5班	89		81～85	82.5	85	84	166				
8	生6	1班	90		86～90	87.5	90	42	208				
9	生7	2班	86		91～95	92.5	95	26	234				
10	生8	3班	87		96～100	97.5	100	16	250				

图 4-55 两类数据百分位数的计算结果

【函数公式解析】

PERCENTILE.INC 函数返回区域中数值的第 k 个百分点的值。具体语法为：

```
PERCENTILE.INC(array,k)
```

array：必需，定义相对位置的数组或数据区域。

k：必需，范围内 0～1（包含 0 和 1）的百分点值。

4.3.2 百分等级分数

百分等级分数是心理测量学中的术语，是应用最广泛的表示测验分数的方法。一个测验分数的百分等级是指在常模样本中低于这个分数的人数百分比。换句话说，百分等级指出的是个体在常模团体中所处的位置，百分等级越低，个体所处的位置越低。百分等级分数用 PR 表示。

对于未分组数据，Excel 提供了 PERCENTRANK.INC 函数计算百分等级分数。

对于分组数据，可根据累积频数分布表，运用插值法，按比例。公式为

$$\mathrm{PR}=\frac{F_{m-1}+\dfrac{(x-L)f_m}{d}}{n}\times 100$$

式中，L 为 P_m 所在组的下限；F_{m-1} 为 P_m 所在组以下的累积频数；f_m 为 P_m 所在组的频数；d 为组距；n 为总频数。

例 4-25 在例 4-24 的基础上，分别按原始成绩和频数分布表计算 94 分的百分等级分数。

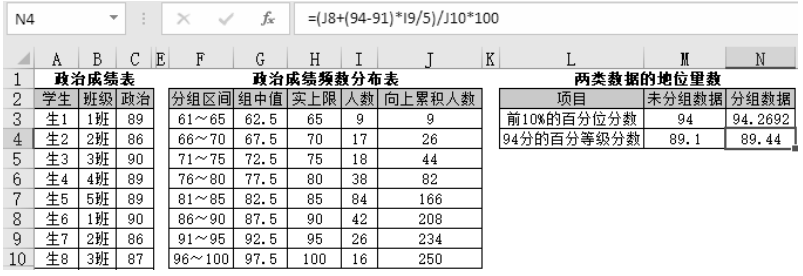
解题思路：在 Excel 中，对于未分组数据，可以使用 PERCENTRANK.INC 函数计算百分等级分数。对于分组数据，则按公式计算。

解题过程：由于在例 4-24 已建立统计表，这里直接在表中输入公式。

在 M4 输入公式“=PERCENTRANK.INC(C3:C252,94)*100”。

在 N4 输入公式“=(J8+(94−91)*I9/5)/J10*100”。

计算结果如图 4-56 所示（隐藏了 D 列）。



	A	B	C	F	G	H	I	J	L	M	N
1	政治成绩表			政治成绩频数分布表					两类数据的地位里数		
2	学生	班级	政治	分组区间	组中值	实上限	人数	向上累积人数	项目	未分组数据	分组数据
3	生1	1班	89	61～65	62.5	65	9	9	前10%的百分位分数	94	94.2692
4	生2	2班	86	66～70	67.5	70	17	26	94分的百分等级分数	89.1	89.44
5	生3	3班	90	71～75	72.5	75	18	44			
6	生4	4班	89	76～80	77.5	80	38	82			
7	生5	5班	89	81～85	82.5	85	84	166			
8	生6	1班	90	86～90	87.5	90	42	208			
9	生7	2班	86	91～95	92.5	95	26	234			
10	生8	3班	87	96～100	97.5	100	16	250			

图 4-56 两类数据百分位数的计算结果

【函数公式解析】

PERCENTRANK.INC 函数将某个数值在数据集中的排位作为数据集的百分比值返回，此处的百分比值的范围为 0～1（含 0 和 1）。具体语法为：

```
PERCENTRANK.INC(array,x,[significance])
```

array：必需，定义相对位置的数值数组或数值数据区域。

x：必需，需要得到其排位的值。

[significance]：可选，用于标识返回的百分比值的有效位数的值。如果省略，则使用 3 位小数(0.xxx)。

4.3.3 标准分数

标准分数（Standard Score）也叫 Z 分数（Z-score），是一个分数与一组数据平均数的差再除以标准差的商。用公式表示为

$$Z=\frac{X-\overline{X}}{S}$$

Z值的量代表着原始分数和母体平均值之间的距离，是以标准差为单位的一个相对量，没有实际单位。在原始分数低于平均值时，则Z为负数，反之，则为正数。一组原始分数的Z分数的标准差为 1。若原始分数呈正态分布，转换成Z分数后，会得到一个均值为 0，标准差为 1 的标准正态分布。Z分数被教育统计学家誉为“多学科表示量数”，有着广泛的用途。根据标准分数公式，Excel 提供了 STANDARDIZE 函数计算标准分数。

例 4-26 在例 4-24 的基础上，如何将原始成绩转换为Z分数？

解题思路：在 Excel 中，可以使用 STANDARDIZE 函数将原始成绩转换为 Z 分数。

解题过程：建立表格，输入公式。

（1）建立表格。在原数据“政治成绩表”表后增加 1 列，如图 4-57 所示。

（2）输入公式。

在 D3 单元格输入公式“=STANDARDIZE(C3,AVERAGE($C$3:$C$252),STDEV.S($C$3:$C$252))”。将 D3 单元格的公式向下填充到 D252 单元格。

计算结果如图 4-58 所示（隐藏了部分行）。

	A	B	C	D
1	政治成绩表			
2	学生	班级	政治	标准分数
3	生1	1班	89	
252	生250	5班	68	

图 4-57 原数据表后增加 1 列

D3 =STANDARDIZE(C3,AVERAGE($C$3:$C$252),STDEV.S($C$3:$C$252))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	政治成绩表					政治成绩频数分布表						两类数据的地位量数	
2	学生	班级	政治	标准分数		分组区间	组中值	实上限	人数	向上累积人数		项目	未分组数据
3	生1	1班	89	0.75381		61～65	62.5	65	9	9		前10%的百分位分数	94
252	生250	5班	68	-1.7589									

图 4-58 两类数据百分位数的计算结果

【函数公式解析】

STANDARDIZE 函数返回由平均数和标准差表示的分布的规范化值。具体语法为：

```
STANDARDIZE(x,mean,standard_dev)
```

x：必需，需要进行正态化的数值。

mean：必需，分布的算术平均值。

standard_dev：必需，分布的标准偏差。

例 4-27 在文件“第 4 章统计量.xlsx”的“地位量数”工作表中，有两名学生的各科成绩和全体学生的平均分和标准差，如图 4-59 所示，试比较两名学生的成绩。

解题思路：可以从原始成绩和Z分数两个角度进行比较。

解题过程：建立表格，输入公式。

（1）建立表格。在原数据表后增加 2 列和 1 行，如图 4-60 所示。

	P	Q	R	S	T
1	两名学生成绩的比较				
2	科目	原始成绩		全体学生	
3		生A	生B	平均分	标准差
4	语文	85	89	70	10
5	政治	70	62	65	5
6	外语	68	72	69	8
7	数学	53	40	50	6
8	物理	72	87	75	8

图 4-59　两名学生的各科成绩和全体学生的平均分和标准差

	P	Q	R	S	T	U	V
1	两名学生成绩的比较						
2	科目	原始成绩		全体学生		Z分数	
3		生A	生B	平均分	标准差	学生A	学生B
4	语文	85	89	70	10		
5	政治	70	62	65	5		
6	外语	68	72	69	8		
7	数学	53	40	50	6		
8	物理	72	87	75	8		
9	合计						

图 4-60　在原数据表后增加 2 列和 1 行

（2）输入公式。

在 U4 单元格输入公式“=(Q4−$S4)/$T4”。

将 U4 单元格的公式向右向下填充到 V8 单元格。

在 Q9 单元格输入公式“=SUM(Q4:Q8)”。

将 Q9 单元格的公式向右填充到 V9 单元格，再清除 S9:T9 区域的公式。

结果如图 4-61 所示。

从表中可以看出，虽然生 A 的原始成绩总分低于生 B，但生 A 的 *Z* 分数总分高于生 B，所以生 A 的成绩更好。对原始成绩直接计算总分是不科学的，因为各科的难易程度、评分的宽严程度不同，各科的分数是不同质的。而 *Z* 分数是相对地位量数，消除了量纲，各科就可以等量齐观了。

U4　=(Q4-$S4)/$T4

	P	Q	R	S	T	U	V
1	两名学生成绩的比较						
2	科目	原始成绩		全体学生		Z分数	
3		生A	生B	平均分	标准差	学生A	学生B
4	语文	85	89	70	10	1.5	1.9
5	政治	70	62	65	5	1	-0.6
6	外语	68	72	69	8	-0.125	0.375
7	数学	53	40	50	6	0.5	-1.667
8	物理	72	87	75	8	-0.375	1.5
9	合计	348	350			2.5	1.5083

图 4-61　两名学生成绩比较的结果

4.4 分布形态

只用集中趋势和离散程度来分析数据还不够准确和全面，还要分析数据的分布形态。数据的分布形态可以从分布的对称程度和分布的高低来描述。对前者的描述叫偏度，对后者的描述叫峰度。

4.4.1 偏度

偏度指数据分布不对称的方向和程度。偏态分布是与正态分布相对而言的。偏态分布又可分为正偏态分布和负偏态分布两种类型。如果频数分布的集中位置偏向数值小的一侧，分布图高峰向左偏移，长尾向右侧延伸，就称为正偏态分布，也称右偏态分布；同样的，如果频数分布的集中位置偏向数值大的一侧，分布图高峰向右偏移，长尾向左延伸，则成为负偏态分布，也称左偏态分布。偏态分布的两种类型如图 4-62 所示。

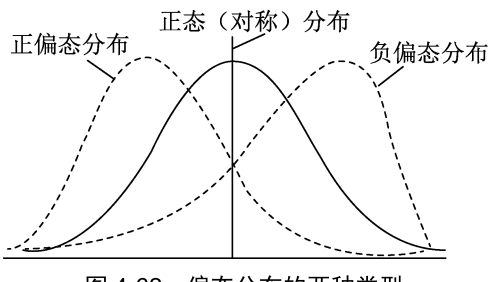


图 4-62　偏态分布的两种类型

在一个正态分布中，平均数、中数、众数三者相等，在数轴中完全重合。在描述正态分布时，只需报告平均数即可。在正偏态分布中，在数轴上，众数＜中数＜平均数，平均数在最右边。在负偏态分布中，在数轴上，平均数＜中数＜众数，平均数在最左边。在偏态分布中，中数把分布下的面积分成两等份的点值上，平均数永远位于尾端，如图 4-63 所示。

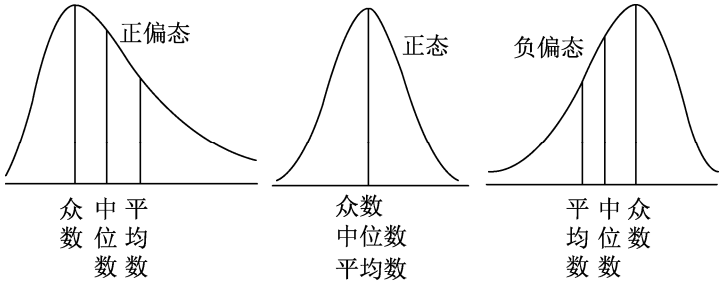


图 4-63　偏态分布中 3 个集中量的关系

显然，从平均数、中数、众数三者的关系只能大致地推断一个分布的对称与否及不对称的方向，不对称的程度如何度量呢？在统计学中，常用偏斜度、皮尔逊偏度系数、矩偏度系数等来衡量偏度。

1. 偏斜度

偏斜度（Skewness）是对统计数据分布偏斜方向及程度的度量。在偏态分布中，当偏斜度为正值时，分布正偏，即众数位于算术平均数的左侧；当偏斜度为负值时，分布负偏，即众数位于算术平均数的右侧。可以利用众数、中数和平均数之间的关系判断分布是左偏态还是右偏态，但要度量分布偏斜的程度，就需要计算偏斜度了。

设一组数据为 X_1，X_2，…，X_n，样本量为 n，偏斜度的计算公式为

$$\mathrm{SKEW}=\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{S}\right)^3$$

如果原始数据被分为 k 组，各组的组中值分别用 X_1，X_2，…，X_k 表示，各组变量值出现的频数分别用 f_1，f_2，…，f_k 表示，样本量为 n，则偏斜度的计算公式为

$$\mathrm{SKEW}=\frac{n}{(n-1)(n-2)}\sum_{i=1}^{k}\left(\frac{X_i-\overline{X}}{S}\right)^3 f_i$$

对于未分组数据，Excel 提供了 SKEW 函数计算偏斜度。

例 4-28　在文件“第 4 章统计量.xlsx”的“分布形态”工作表中，有某年级 5 个班 250 人的政治成绩表及频数分布表，如图 4-64 所示（隐藏了部分行），如何分别按原始成绩和频数分布表计算偏斜度？

解题思路：在 Excel 中，对于未分组数据，可以直接使用 SKEW 函数计算偏斜度或使用定义的公式计算偏斜度。对于分组数据，则只能使用推演公式计算偏斜度。

解题过程：建立统计表，输入公式。

（1）建立统计表。建立一个“两类数据的偏斜度、皮尔逊偏度系数”统计表（包括后面将要介绍到的皮尔逊偏度系数），如图 4-65 所示。

	A	B	C	D	E	F	G	H	I
1	政治成绩表				政治成绩频数分布表				
2	学生	班级	政治		分组区间	组中值	实上限	人数	向下累积人数
3	生1	1班	89		61～65	62.5	65	9	9
4	生2	2班	86		66～70	67.5	70	17	26
5	生3	3班	90		71～75	72.5	75	18	44
6	生4	4班	89		76～80	77.5	80	38	82
7	生5	5班	89		81～85	82.5	85	84	166
8	生6	1班	90		86～90	87.5	90	42	208
9	生7	2班	86		91～95	92.5	95	26	234
10	生8	3班	87		96～100	97.5	100	16	250
252	生250	5班	68						

图 4-64　政治成绩表及频数分布表

	K	L	M
1	两类数据的偏斜度、皮尔逊偏度系数		
2	项目	未分组数据	分组数据
3	平均数		
4	标准差		
5	中位数		
6	偏斜度		
7	皮尔逊偏度系数		

图 4-65　两类数据的偏斜度、皮尔逊偏度系数统计表

（2）输入公式。

在 L3 单元格输入公式“=AVERAGE(C3:C252)”。

在 M3 单元格输入公式“=SUMPRODUCT(F3:F10,H3:H10)/SUM(H3:H10)”。

在 L4 单元格输入公式“=STDEV.S(C3:C252)”。

在 M4 单元格输入公式“=SQRT(SUMPRODUCT(POWER(F3:F10−SUMPRODUCT(F3:F10,H3:H10)/SUM(H3:H10),2),H3:H10)/(I10−1))”。

在 L6 单元格输入数组公式“=SKEW(C3:C252)”或“{=SUM(((C3:C252−L3)/L4)^3)*(I10/(I10−1)/(I10−2))}”。

在 M6 单元格输入公式“=SUMPRODUCT(((F3:F10−M3)/M4)^3,H3:H10)*(I10/(I10−1)/(I10−2))”。

计算结果如图 4-66 所示。

M6　=SUMPRODUCT(((F3:F10-M3)/M4)^3,H3:H10)*(I10/(I10-1)/(I10-2))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	政治成绩表				政治成绩频数分布表						两类数据的偏斜度、皮尔逊偏度系数		
2	学生	班级	政治		分组区间	组中值	实上限	人数	向下累积人数		项目	未分组数据	分组数据
3	生1	1班	89		61～65	62.5	65	9	9		平均数	82.7	82.12
4	生2	2班	86		66～70	67.5	70	17	26		标准差	8.3575952	8.35673
5	生3	3班	90		71～75	72.5	75	18	44		中位数		
6	生4	4班	89		76～80	77.5	80	38	82		偏斜度	-0.381125	-0.3263
7	生5	5班	89		81～85	82.5	85	84	166		皮尔逊偏度系数		
8	生6	1班	90		86～90	87.5	90	42	208				
9	生7	2班	86		91～95	92.5	95	26	234				
10	生8	3班	87		96～100	97.5	100	16	250				

图 4-66　两类数据偏斜度的计算结果

从表中可以看出，偏斜度为负值，超过−0.3，这组原始数据呈现一定负偏态，高分段学生较多。以 H3:H10 区域的各组人数为源数据，插入“带平滑线和数据标记的散点图”，可以比较明显地看出数据呈现负偏态，如图 4-67 所示。

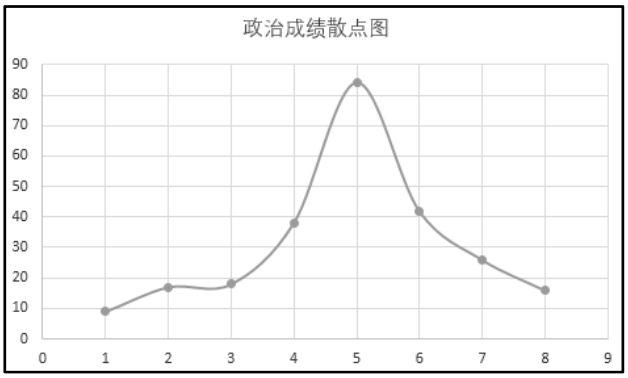


图 4-67　人数的负偏态

【函数公式解析】

在 L6 单元格公式中，SKEW 函数返回分布的偏斜度。具体语法为：

```
SKEW(number1,[number2],…)
```

number1，[number2]，…：number1 是必需的，后续数字是可选的。用于计算偏斜度的 1～255 个参数。也可以用单一数组或对某个数组的引用来代替用逗号分隔的参数。参数可以是数字或者是包含数字的名称、数组或引用。如果数组或引用参数包含文本、逻辑值或空白单元格，则这些值将被忽略；但包含零值的单元格将计算在内。

备选公式是按偏斜度公式计算的。

2. 皮尔逊偏度系数

皮尔逊（Pearson）偏度系数以平均值与中位数或众数之差再与标准差之比来衡量偏斜的程度，用 SK 表示。这是根据众数、中位数与均值各自的性质，通过比较众数或中位数与均值占比来衡量偏斜度的。其计算公式为

$$\mathrm{SK}=\frac{\overline{X}-M_{\mathrm{d}}}{S}\ \text{或}\ \mathrm{SK}=\frac{\overline{X}-M_{\mathrm{o}}}{S}$$

式中，$\overline{X}$ 为均值；M_{d} 为中位数；M_{o} 为众数。偏度系数小于 0，平均数在众数之左，是一种左偏的分布，又称为负偏态。偏度系数大于 0，均值在众数之右，是一种右偏的分布，又称为正偏态。

例 4-29　在例 4-28 的基础上，分别按原始成绩和频数分布表计算皮尔逊偏度系数。

解题思路：由于分组数据的众数为估计值，因此本例利用中位数来计算皮尔逊偏度系数。所需平均值、中位数、标准差均可以用前面介绍的方法获得。

解题过程：由于在例 4-24 已建立统计表，这里直接在表中输入公式。

在 L5 单元格输入公式“=MEDIAN(C3:C252)”。

在 M5 单元格输入公式“=81+(123−I6)/(I7−I6)*5”。

在 L7 单元格输入公式“=(L3−L5)/L4”。

将 L7 单元格的公式向右填充到 M7 单元格。

计算结果如图 4-68 所示。

L7 =(L3-L5)/L4

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	政治成绩表				政治成绩频数分布表						两类数据的偏斜度、皮尔逊偏度系数		
2	学生	班级	政治		分组区间	组中值	实上限	人数	向下累积人数		项目	未分组数据	分组数据
3	生1	1班	89		61～65	62.5	65	9	9		平均数	82.7	82.12
4	生2	2班	86		66～70	67.5	70	17	26		标准差	8.3575952	8.35673
5	生3	3班	90		71～75	72.5	75	18	44		中位数	83	83.4405
6	生4	4班	89		76～80	77.5	80	38	82		偏斜度	-0.381125	-0.3263
7	生5	5班	89		81～85	82.5	85	84	166		皮尔逊偏度系数	-0.035895	-0.158
8	生6	1班	90		86～90	87.5	90	42	208				
9	生7	2班	86		91～95	92.5	95	26	234				
10	生8	3班	87		96～100	97.5	100	16	250				

图 4-68　两类数据皮尔逊偏度系数的计算结果

3. 矩偏度系数

矩也称为动差。在统计学中，未分组变量和分组变量的 k 阶样本中心矩的公式分别为：

$$m_k=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^k}{n-1},\quad m_k=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^k f_i}{\sum_{i=1}^{n}f_i-1}$$

当 k=0 时，为零阶中心矩。

当 k=1 时，为一阶中心矩，用于表示数据分布的差异度。因为 $\sum_{i=1}^{n}(X_i-\overline{X})=0$，所以在实际应用中，一般不取其代数和，而取绝对值和 $\sum_{i=1}^{n}\left|X_i-\overline{X}\right|$，这就是平均差。

当 k=2 时，为二阶中心矩，用于表示数据的离中趋势，也就是方差。方差的平方根就是标准差，是应用最为广泛的一种差异量数。

当 k=3 时，为三阶中心矩，用于表示一个分布的偏斜度或偏态性。

当 k=4 时，为四阶中心矩，用于表示一个分布的峰度或峰态性。

矩偏度系数是以变量的三阶中心动差除以标准差三次方，来衡量一个分布的不对称程度或偏斜程度的指标。也就是说，三阶中心动差是以标准差为单位的系数。一个样本未分组变量和分组变量的矩偏度系数公式为：

$$\alpha=\frac{m_3}{S^3}=\frac{\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})^3}{n-1}}{\left(\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}\right)^3},\quad \alpha=\frac{m_3}{S^3}=\frac{\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})^3 f_i}{\sum_{i=1}^{n}f_i-1}}{\left(\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})^2 f_i}{\sum_{i=1}^{n}f_i-1}\right)^3}$$

当 $\alpha>0$ 时，为正偏态；当 $\alpha<0$ 时，为负偏态；当 $\alpha=0$ 时，为正态分布，如图 4-69 所示。

例 4-30 在例 4-28 的基础上，分别按原始成绩和频数分布表计算矩偏度系数。

解题思路：在 Excel 中，对于未分组数据和分组数据，都可以按公式计算矩偏度系数。

解题过程：建立统计表，输入公式。

（1）建立统计表。建立一个“两类数据的矩偏度系数、矩峰度系数”统计表（包含 4.4.2 节将要介绍到的矩峰度系数），如图 4-70 所示。

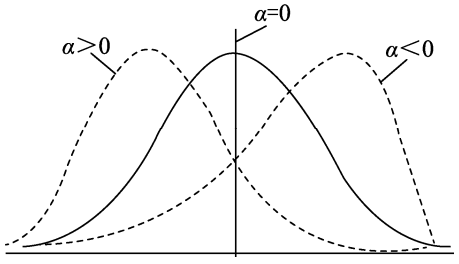


图 4-69 α 值决定分布形态

	K	L	M
9	两类数据的矩偏度、矩峰度系数		
10	项目	未分组数据	分组数据
11	$n-1$		
12	平均数		
13	矩偏度系数		
14	矩峰度系数		

图 4-70 两类数据的矩偏度系数、矩峰度系数表

（2）输入公式。

在 L11 单元格输入公式“=SUM($H$3:$H$10)-1”。

在 M11 单元格输入公式“=L11”。

在 L12 单元格输入公式“=L3”。

在 M12 单元格输入公式“=M3”。

在 L13 单元格输入公式“{=SUM((C3:C252-L12)^3)/L11/(SUM((C3:C252-L12)^2)/L11)^3}”。

在 M13 单元格输入公式“{=SUM(H3:H10*(F3:F10-M12)^3)/M11/(SUM((F3:F10-M12)^2*H3:H10)/M11)^3}”。

计算结果如图 4-71 所示。

L13 {=SUM((C3:C252-L12)^3)/L11/(SUM((C3:C252-L12)^2)/L11)^3}

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	政治成绩表				政治成绩频数分布表						两类数据的偏斜度、皮尔逊偏度系数		
2	学生	班级	政治		分组区间	组中值	实上限	人数	向下累积人数		项目	未分组数据	分组数据
3	生1	1班	89		61～65	62.5	65	9	9		平均数	82.7	82.12
4	生2	2班	86		66～70	67.5	70	17	26		标准差	8.3575952	8.35673
5	生3	3班	90		71～75	72.5	75	18	44		中位数	83	83.4405
6	生4	4班	89		76～80	77.5	80	38	82		偏斜度	-0.381125	-0.3263
7	生5	5班	89		81～85	82.5	85	84	166		皮尔逊偏度系数	-0.035895	-0.158
8	生6	1班	90		86～90	87.5	90	42	208				
9	生7	2班	86		91～95	92.5	95	26	234				
10	生8	3班	87		96～100	97.5	100	16	250		两类数据的矩偏度、矩峰度系数		
11	生9	4班	90								项目	未分组数据	分组数据
12	生10	5班	87								$n-1$	249	249
13	生11	1班	89								平均数	82.7	82.12
14	生12	2班	90								矩偏度系数	-0.000648	-0.0006
											矩峰度系数		

图 4-71 两类数据的矩偏度系数的计算结果

4.4.2 峰度

峰度是指数据分布图形的尖峭程度或峰凸程度。以正态分布为标准，如果一个分布比正态分布更高更瘦，则称为高峰态；如果一个分布比正态分布更矮更胖，则称为低峰态。

如图 4-72 所示。

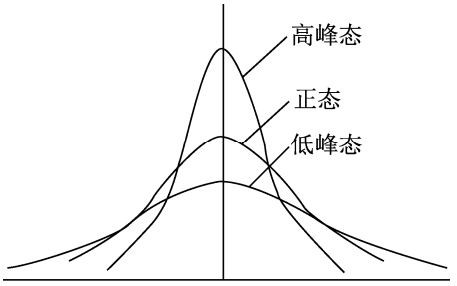


图 4-72　三类峰态

1. 峰值

峰值是以变量的四阶中心动差除以标准差的四次方，并将结果再减去 3，用来衡量频数分布的集中程度，也是衡量分布曲线相对尖锐度或平坦度的指标。在统计学中，未分组变量和分组变量的 k 阶样本中心矩的公式分别为：

$$\mathrm{KUPT}=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{S}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

$$\mathrm{KUPT}=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{S}\right)^4 f_i-\frac{3(n-1)^2}{(n-2)(n-3)}$$

峰值指标是以正态分布为比较标准，正峰值表示一个频数分布比正态分布更集中，分布呈尖峰状态，平均数代表性更高；负峰值表示一个频数分布比正态分布更分散，分布呈平坦峰，平均数代表性较低。

对于未分组数据，Excel 提供了 KURT 函数计算偏斜度。

例 4-31　在例 4-28 的基础上，分别按原始成绩和频数分布表峰值。

解题思路：在 Excel 中，对于未分组数据，可以直接使用 KURT 函数计算峰值或使用公式进行计算峰值。对于分组数据，则只能使用公式计算峰值。

解题过程：建立统计表，输入公式。

（1）建立统计表。建立一个“两类数据的峰值”统计表，如图 4-73 所示。

	O	P	Q
1	两类数据的峰值		
2	项目	未分组数据	分组数据
3	平均数		
4	标准差		
5	峰值		

图 4-73　两类数据的峰值表

（2）输入公式。

所需平均值、标准差可以直接使用前面例子的计算结果。

在 P3 单元格输入公式“=L3”。将 P3 单元格的公式向右向下填充到 Q4 单元格。

在 P5 单元格输入公式“=KURT(C3:C252)”或“{=I10*(I10+1)/(I10−1)/(I10−2)/(I10−3)*(SUM(((C3:C252−P3)/P4)^4))−(3*(I10−1)^2/(I10−2)/(I10−3))}”。

在 Q5 单元格输入数组公式“{=I10*(I10+1)/(I10−1)/(I10−2)/(I10−3)*(SUM(((F3:F10−Q3)/P4)^4*H3:H10))−(3*(I10−1)^2/(I10−2)/(I10−3))}”。

计算结果如图 4-74 所示（隐藏了部分行列）。

P5 =KURT(C3:C252)

	A	B	C	D	E	F	G	H	I	J	O	P	Q
1	政治成绩表				政治成绩频数分布表						两类数据的峰值		
2	学生	班级	政治		分组区间	组中值	实上限	人数	向下累积人数		项目	未分组数据	分组数据
3	生1	1班	89		61～65	62.5	65	9	9		平均数	82.7	82.12
4	生2	2班	86		66～70	67.5	70	17	26		标准差	8.3575952	8.35673
5	生3	3班	90		71～75	72.5	75	18	44		峰值	-0.0282	-0.0734
6	生4	4班	89		76～80	77.5	80	38	82				
7	生5	5班	89		81～85	82.5	85	84	166				
8	生6	1班	90		86～90	87.5	90	42	208				
9	生7	2班	86		91～95	92.5	95	26	234				
10	生8	3班	87		96～100	97.5	100	16	250				

图 4-74　两类数据峰值的计算结果

从图 4-74 中可以看出，峰值虽然为负值，但接近 0，表明这个分布基本呈正态分布。

【函数公式解析】

在 P5 单元格的公式中，KURT 函数返回一组数据的峰值。具体语法为：

```
KURT(number1,[number2],…)
```

number1，[number2]，…：number1 是必需的，后续数字是可选的。用于计算峰值的 1～255 个数。也可以用单一数组或对某个数组的引用来代替用逗号分隔的参数。参数可以是数字或者是包含数字的名称、数组或引用。如果数组或引用参数包含文本、逻辑值或空白单元格，则这些值将被忽略；但包含零值的单元格将计算在内。

备选公式是按峰值公式计算的。

2. 矩峰度系数

矩峰度系数表示四阶中心动差与标准差四次方的比值。未分组变量和分组变量的矩偏度系数公式为：

$$\beta=\frac{m_4}{S^4}=\frac{\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})^4}{n-1}}{\left(\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}\right)^4},\quad \beta=\frac{m_4}{S^4}=\frac{\dfrac{\sum_{i=1}^{n}f_i(X_i-\overline{X})^4 f_i}{\sum_{i=1}^{n}f_i-1}}{\left(\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})^2 f_i}{\sum_{i=1}^{n}f_i-1}\right)^4}$$

当 $\beta>0$ 时，为高峰态；当 $\beta<0$ 时，为低峰态；当 $\beta=0$ 时，为正态分布，如图 4-75 所示。

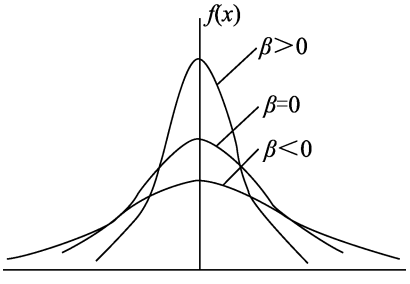


图 4-75　β 值决定分布形态

例 4-32　在例 4-28 的基础上，分别按原始成绩和频数分布表计算矩峰度系数。

解题思路：在 Excel 中，对于未分组数据和分组数据，都可以按公式计算。

解题过程：由于例 4-28 中已建立统计表，这里直接

在表中输入公式。

在 L14 单元格公式输入数组公式“{=SUM((C3:C252−L12)^4)/L11/(SUM((C3:C252−L12)^2)/L11)^4}”。

在 M14 单元格公式输入数组公式“{=SUM((F3:F10−M12)^4*H3:H10)/M11/(SUM((F3:F10−M12)^2*H3:H10)/M11)^4}”。

计算结果如图 4-76 所示。

L14 {=SUM((C3:C252-L12)^4)/L11/(SUM((C3:C252-L12)^2)/L11)^4}

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	政治成绩表				政治成绩频数分布表						两类数据的偏斜度、皮尔逊偏度系数		
2	学生	班级	政治		分组区间	组中值	实上限	人数	向下累积人数		项目	未分组数据	分组数据
3	生1	1班	89		61～65	62.5	65	9	9		平均数	82.7	82.12
4	生2	2班	86		66～70	67.5	70	17	26		标准差	8.3575952	8.35673
5	生3	3班	90		71～75	72.5	75	18	44		中位数	83	83.4405
6	生4	4班	89		76～80	77.5	80	38	82		偏斜度	-0.381125	-0.3263
7	生5	5班	89		81～85	82.5	85	84	166		皮尔逊偏度系数	-0.035895	-0.158
8	生6	1班	90		86～90	87.5	90	42	208				
9	生7	2班	86		91～95	92.5	95	26	234		两类数据的矩偏度、矩峰度系数		
10	生8	3班	87		96～100	97.5	100	16	250		项目	未分组数据	分组数据
11	生9	4班	90								n-1	249	249
12	生10	5班	87								平均数	82.7	82.12
13	生11	1班	89								矩偏度系数	-0.000648	-0.0006
14	生12	2班	90								矩峰度系数	0.0006019	0.00059

图 4-76　两类数据的矩峰度系数的计算结果

从图 4-76 中可以看出，矩峰度系数虽然为正值，但极其接近 0，表明这个分布基本呈正态分布。

4.4.3　分类的偏斜度和峰值

前面在介绍偏度和峰度时，所举例子均没有分类计算偏度和峰度。实际上，这批原始数据可能是有类别的。下面介绍分类计算偏度和峰度的技巧。

例 4-33　在例 4-28 的基础上，如何按班计算偏斜度和峰值？

解题思路：在 Excel 中，使用 IF 函数的条件判断功能结合其他函数，可以分类计算偏斜度和峰值。

解题过程：建立统计表，输入公式。

（1）建立统计表。建立一个“未分组数据分类的偏斜度和峰值”统计表，如图 4-77 所示。

（2）输入公式。

在 T3 单元格输入数组公式“{=SKEW(IF($B$3:$B$252=S3,$C$3:$C$252))}”。

在 U3 单元格输入数组公式“{=KURT(IF($B$3:$B$252=S3,$C$3:$C$252))}”。

将 T3:U3 区域的公式向下填充到 U7 单元格。

	S	T	U
1	未分组数据分类的偏斜度和峰值		
2	班级	偏斜度	峰值
3	1班		
4	2班		
5	3班		
6	4班		
7	5班		

图 4-77　未分组数据分类的偏斜度和峰值统计表

计算结果如图 4-78 所示。

T3 {=SKEW(IF($B$3:$B$252=S3,$C$3:$C$252))}

	A	B	C	D	S	T	U	V	W
1	政治成绩表				未分组数据分类的偏斜度和峰值				
2	学生	班级	政治		班级	偏斜度	峰值		
3	生1	1班	89		1班	-0.502798775	0.036534568		
4	生2	2班	86		2班	-0.693406203	0.448858646		
5	生3	3班	90		3班	0.104738925	-0.500948698		
6	生4	4班	89		4班	-0.607128211	0.408346987		
7	生5	5班	89		5班	-0.403977104	-0.254713526		

图 4-78 未分组数据分类的偏斜度和峰值的计算结果

4.5 描述统计工具

在 Excel 中，对于一组数据，可以使用描述统计工具快速获得一些统计量，如平均数、中位数、众数、标准差、方差、峰度、偏度等，这比使用公式更为快速、方便。

例 4-34 在文件“第 4 章统计量.xlsx”中的“描述统计工具”工作表中，有某级 5 个班 250 人的政治成绩，如图 4-79 所示（隐藏了部分行），如何运用“描述统计”工具进行分析？

	A	B	C
1	政治成绩表		
2	学生	班级	政治
3	生1	1班	89
4	生2	2班	86
252	生250	5班	68

图 4-79 政治成绩表

解题思路： 在 Excel 中，在“数据”选项卡中启用“数学分析”功能，再启用“描述统计”工具。

解题过程： 操作过程如下：

❶ 单击“数据”选项卡。

❷ 在“分析”组中单击“数据分析”按钮。

❸ 在弹出的“数据分析”对话框的“分析工具”列表框中选择“描述统计”选项。

❹ 单击“确定”按钮，弹出“描述统计”对话框。

❺ 将鼠标放置于“输入”组的“输入区域”框中选择 C2:C252 区域。本区域是按列排列的，“分组方式”组中已默认为“逐列”，所以不再选择“分组方式”。

❻ 勾选“标志位于第一行”复选框。本例标志“政治”在 C2 单元格，已包含在“输入区域”。如果标志未包含在“输入区域”，也未勾选“标志位于第一行”复选框，Excel 会自动以列 1，列 2，列 3，…作为数据的列标志。

❼ 在“输出选项”组中选择“输出区域”单选按钮。如果需要，也可以选择其他地方存放统计结果。

❽ 将鼠标放置于“输出区域”框中，选择 E1 单元格。E1 单元格将作为输出的统计表的左上角单元格引用。

❾ 勾选“汇总统计”复选框。“汇总统计”是指描述统计结果，包括平均数、标准误差（相对于平均数）、中位数、众数、标准差、方差、峰度、偏度、区域、最小值、最大值、求

和、观测数等。如果不勾选，则不会显示这些结果。

⑩ 勾选“平均数置信度”复选框，还可在文本框中设置具体的置信度。常用置信度为90%、95%、97%、99%，分别代表显著性水平为10%、5%、3%、1%时的平均数置信度。

⑪ 勾选“第K大值”复选框。如果不修改文本框中的默认数值“1”，则为数据的最大值。

⑫ 勾选“第K小值”复选框。如果不修改文本框中的默认数值“1”，则为数据的最小值。

⑬ 单击“确定”按钮。

操作过程及结果如图4-80所示。

图4-80 进行“描述统计”的过程和结果

总之，统计量代表了样本的集中趋势或离散程度等特征和分布形态，并为进一步的数据分析和推断奠定基础。

第 5 章 概率及其分布

前面各章介绍了 Excel 在描述性统计方面的应用。显然，仅对数据进行描述性统计是不够的。科学研究的目的是通过对样本数据的统计分析推测总体的基本特征，并对推断的正确性大小进行概率检验。这种从样本出发来推断总体分布的过程就叫统计推断。统计推断的数学基础是概率论。

本章将介绍概率的基本知识，并从 Excel 应用的角度，重点介绍排列组合和包括二项分布、泊松分布、正态分布在内的三大概率分布。

5.1 概率原理

5.1.1 什么是概率

概率（Probability）又称或然率、机会率、机率（几率）或可能性，是概率论的基本概念，是对随机事件发生的可能性的度量。在概率论中，把不可能发生的事件的概率称为最小概率，定为 0，而把必然发生的事件的概率称为最大概率，定为 1。那么，事件 A 出现的概率 $P(A)$ 为 0～1，即随机事件发生的可能性为 0～1，是一个非负数。

概率与事件有关。在一定的条件下可能发生也可能不发生的事件，叫作随机事件。在一个特定的随机试验中，每一可能出现的结果被称为一个基本事件，随机事件（简称事件）是由某些基本事件组成的。如果一次试验中可能出现的结果有 n 个，即此试验由 n 个基本事件组成。例如，在掷骰子的随机试验中，每一点（1、2、3、4、5、6）表示一个基本事件。

在试验中不包含任何基本事件的事件，称为不可能事件。在试验中包含所有基本事件，一定发生的事件，称为必然事件。例如，在连续掷两次骰子的随机试验中，“点数之和为 1”这个事件是一个不可能事件，P=0。“点数之和小于 40”是一个必然事件，P=1。

一次试验中所有结果出现的可能性都相等，那么这种事件就叫作等可能事件。例如，在掷骰子的随机试验中，每一点（1、2、3、4、5、6）出现的可能性都相等。