第 2 章　　描述统计

第 35 页　倒数第 1 行

几何平均数通常应用在无名数动态相对指标，例如第 14 章的发展速度。

第 41 页

四分位数公式 Excel 有 Quartile.INC(间距)和 Quartile.EXC(个数)。

以数据间距计算

Quartile.INC(array,0)= Percentile.INC(array,0)= min，

Quartile.INC(array,1)= Percentile.INC(array,0.25)= Q1 ，

Quartile.INC(array,2)= Percentile.INC(array,0.5)= Q2 = Me ，

Quartile.INC(array,3)= Percentile.INC(array,0.75)= Q3 ，

Quartile.INC(array,4)= Percentile.INC(array,1)= max。

R> quantile(x, 0.25, type=7) ;　quantile(array, 0.75, type=7)

以数据个数计算

Quartile.EXC(array,0)= Percentile.EXC(array,0)= min，

Quartile.EXC(array,1)= Percentile.EXC(array,0.25)= Q1 ，

Quartile.EXC(array,2)= Percentile.EXC(array,0.5)= Q2 = Me，

Quartile.EXC(array,3)= Percentile.EXC(array,0.75)= Q3 ，

Quartile.EXC(array,4)= Percentile.EXC(array,1)= max

R> quantile(x, 0.25, type=6) ;　quantile(array, 0.75, type=6)

第 43 页

请注意，为什么样本数据的方差公式的分母要用 $n-1$ ？因为样本数据方差的分母若用 n，则会低估总体方差 σ^2 以及与自由度不符(请见第 7 章)

第 45 页

如果两(或 m)组独立抽样的样本数据，来自两(m)个“不同”总体(平均数可能不同)，但两(m)个总体“方差相同”(9.3 节)，则两(m)组样本数据，可合并计算其共同的方差。 每组样本容量 n_i ，每组样本方差 s_i^2 ，合并方差 s^2_{pool} (或记作 s_p^2)。请见 9.4 及 10.3 节 MSE 公式：

$$s^2_{\text{pool}} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)} \text{ 或 } s_p^2 = \frac{\sum_{i=1}^{m}(n_i-1)s_i^2}{n-m}$$

以下有：例题 2.1x 或 例题 2.11a 等，表示在书本《大话统计学》没有编号，是增加的。

補充教材在最后，有分组数据的描述统计。

例题 2.1： 学生成绩的频数分布表。 下列 30 个数据是学生的成绩：频数分配表

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：本章都用这 30 个数据为例，计算描述统计。

组界	频数	累积频数	累积频率
[20,30)	1	1	3.33%
[30,40)	5	6	20%
[40,50)	8	14	46.67%
[50,60)	4	18	60%
[60,70)	1	19	63.33%
[70,80)	2	21	70%
[80,90)	5	26	86.67%
[90,100)	4	30	100%
	30		

例题 2.2： 学生成绩的直方图。

请见 2.9.2 节

例题 2.3： 不同组距频数分布表画直方图。以最小的组宽为基准。

组界	相对次数	组宽	修正相对次数
[20,40)	0.05	20	0.05
[40,60)	0.10	20	0.10
[60,120)	0.45	60	0.15
[120,180)	0.30	60	0.10
[180,500)	0.10	320	0.00625
	1.00		

解答：

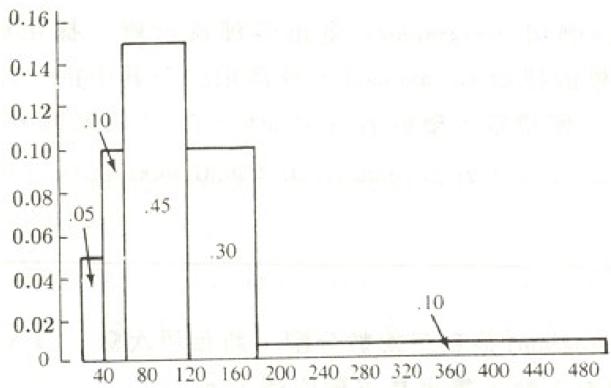


图 2.a 修正的直方图

2.3.2 多边形图

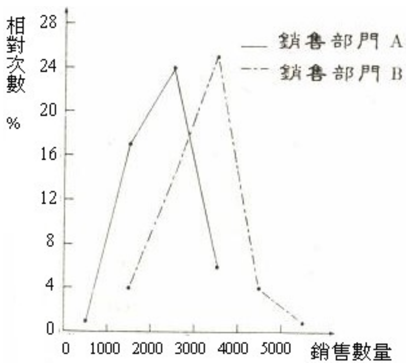


图 2.b　两组数据的多边形图

例题 2.x： 下列 25 个数据

20,21,22,22,23,24,25,26,27,29,30,30,32,33,33,36,37,40,41,42,48,55,56,61,64

解答：以茎叶图表示，结果如下。左边的茎叶图，茎(枝)的长度为 10。中间的茎叶图，茎(枝)的长度为 5。最右边茎叶图左边的数字是，由上而下与由下而上的累积频数，括号是中位数所在。

```
2 | 0122345679
3 | 0023367
4 | 0128
5 | 56
6 | 14
```

```
2* | 012234
2. | 5679
3* | 00233
3. | 67
4* | 012
4. | 8
5* |
5. | 56
6* | 14
```

```
 6    2* | 012234
10    2o | 5679
(5)   3* | 00233
10    3o | 67
 8    4* | 012
 5    4o | 8
 4    5* |
 4    5o | 56
 2    6* | 14
```

图 2.c　茎叶图

茎叶图逆时针旋转 90 度，可以看作直方图。

例题 2.5： 学生成绩的箱线图

请见 2.9.5 节

例题 2.6： 柏拉图

请见 2.9.7 节

2.3.8　箱线图

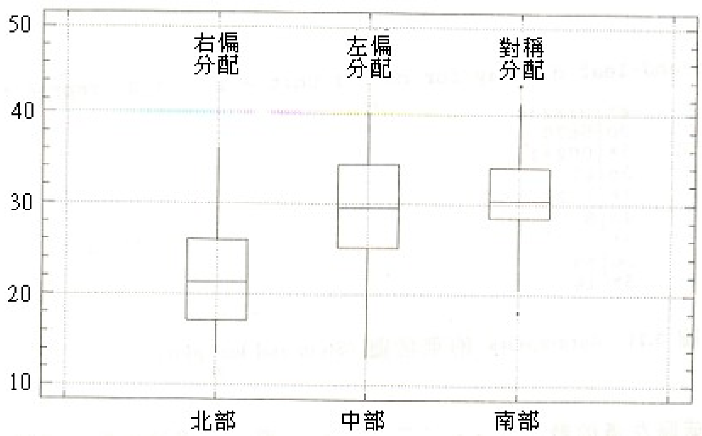


图 2.d　多组箱线图

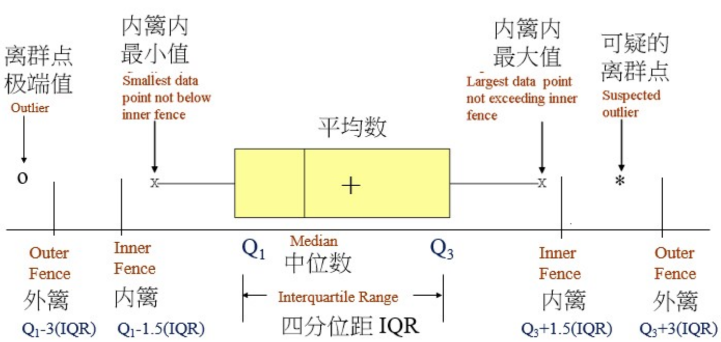


图 2.e 箱线图另一种表达方式

2.3.9 柏拉图

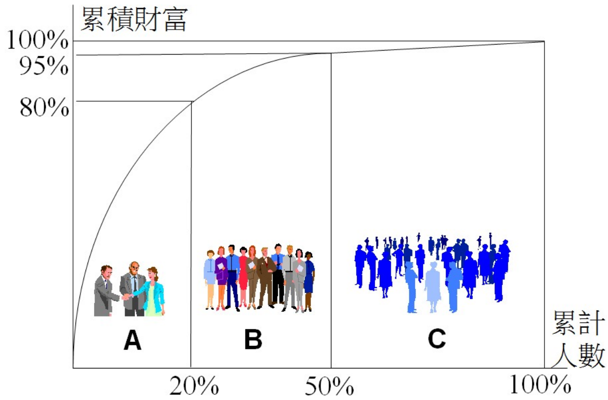


图 2.f 柏拉图 (Perato chart)

在信息系统设计有句名言：「一图值千言」(A picture is worth a thousand words)。一个良好的统计图，能让数据作出更有意义的表达。例如图 2.g 是 1812 年拿破仑攻打俄罗斯，从 40 万大军出发，到最后只剩下 1 万人的人数与地图的统计图，图形中除了人数统计数值外，还有时间、温度、地形等变量数据，这个图形可以称为「最佳统计图」。

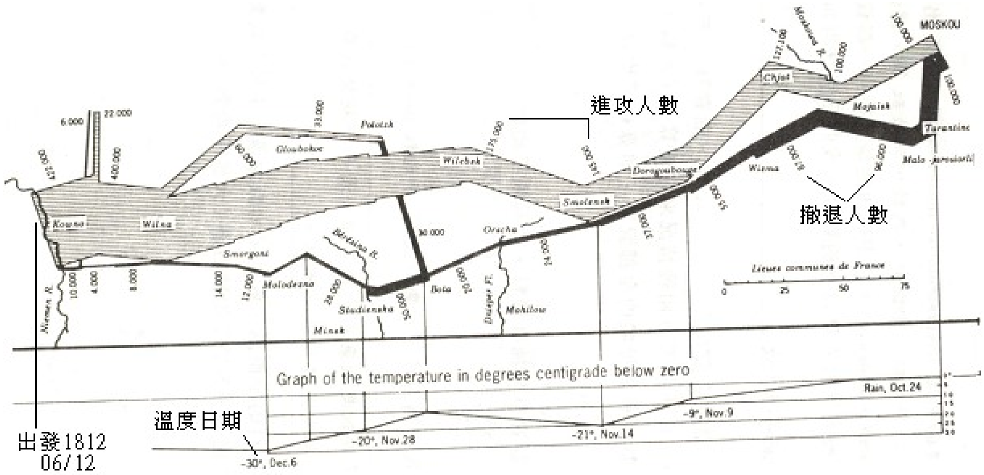
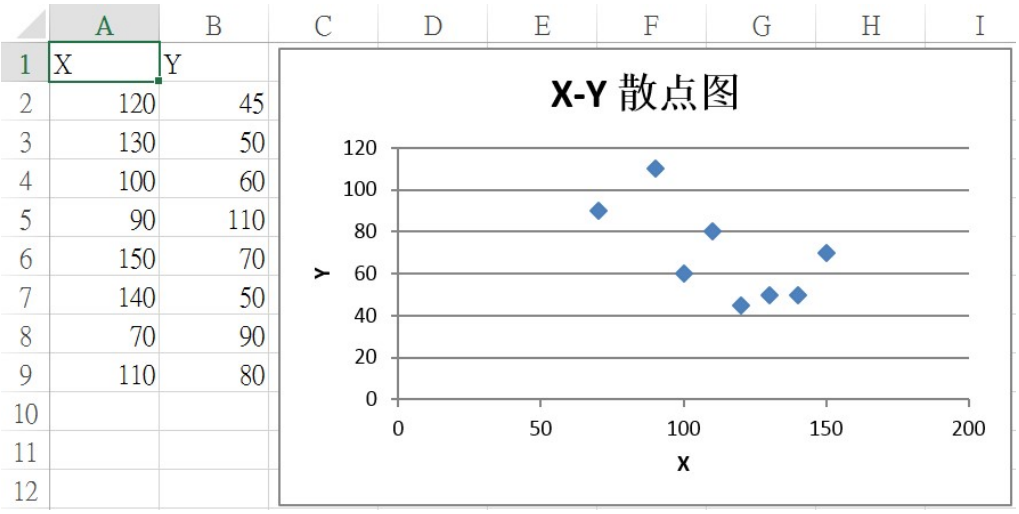


图 2.g 拿破仑进攻苏俄兵力统计图

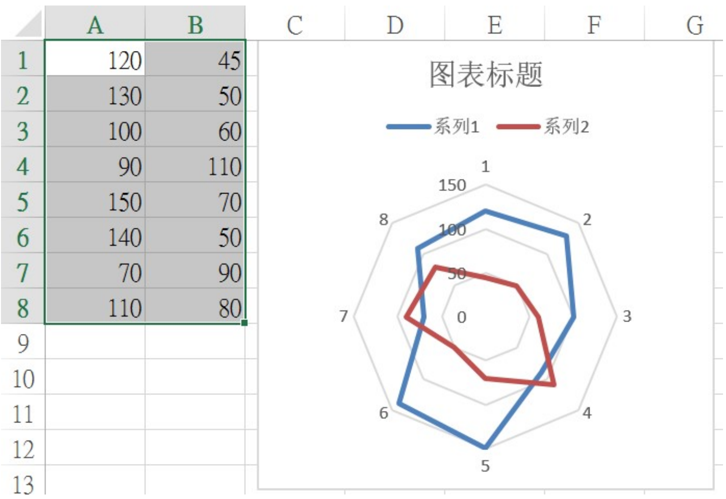
数据来源 ： Tufte``The Visual Display of Quantitative Information"(1983) , p.41

例题 2.7： 散点图



例题 2.8： 雷达图

不是用中文统计，先将 A1:B8 标示起来，再 Excel 菜单(选单) → 插入 → 图表 → 所有图表 → 雷达图



2.4.2 相对指标

例题 2.9 分为下列六个子题

例题 2.9a： 比重相对指标

表 2.1 的 GDP 国内生产总值

$$\text{第一产业比重相对指标} = \frac{56957}{568845} = 10.0\%$$

$$\text{第二产业比重相对指标} = \frac{249684}{568845} = 43.9\%$$

$$\text{第三产业比重相对指标} = \frac{262204}{568845} = 46.1\%$$

例题 2.9b： 比率相对指标

2016 年我国人口有男性 706652506 人，女性 672362042 人

$$\text{男性人口和女性人口的比率相对指标} = \frac{706652506}{672362042} = 105\%$$

男女性别比率 $= 105 : 100$

某校有学生 3000 人，教师 150 人

生师比率相对指标 $= 20 : 1$

例题 2.9c： 比较相对指标

2015 年北京的 GDP 国内生产总值 $= 23000$ 亿元，人均 GDP $= 106751$ 元

2015 年上海的 GDP 国内生产总值 $= 25300$ 亿元，人均 GDP $= 102920$ 元

$$\text{2015 年北京和上海的 GDP 比较相对指标} = \frac{23000}{25300} = 91\%$$

$$\text{2015 年北京和上海的人均 GDP 比较相对指标} = \frac{106751}{102920} = 104\%$$

例题 2.9d： 动态相对指标

A 企业 2014 年营业收入= 2000 万元，2015 年营业收入= 2500 万元，

$$\text{A 企业营业收入动态相对指标} = \frac{25,000,000}{20,000,000} = 125\%$$

动态相对指标即指数或发展速度，请见第 3 章及第 4 章。

例题 2.9e： 强度相对指标

某市面积 1000 平方公里，人口 400 万人，有 6000 间便利商店，下列强度相对指标

$$人口密度 = \frac{4000,000}{1000} = 4000\ 人/平方公里$$

$$便利商店密度 = \frac{6000}{1000} = 6\ 间/平方公里$$

$$便利商店平均服务人数 = \frac{4000,000}{6000} = 6667\ 人/间$$

有的强度相对指标的分子和分母可以互换，因此有正指标和逆指标，如果上述是正指标，则逆指标如下：

$$每万人拥有的土地面积 = \frac{1000}{400} = 2.5\ 平方公里/万人$$

$$便利商店平均服务面积 = \frac{1000}{6000} = 0.167\ 平方公里/间$$

$$每万人平均便利商店数目 = \frac{6000}{400} = 15\ 间/万人$$

强度相对指标还有下列应用：人均国内生产总值、人均生产量(营业额)、成本利润率、资产利润率、投资收益(报酬)率、存货周转率、资产负债率、速动比率等。

例题 2.9f: 计划完成相对指标

A 企业 2016 年计划生产 5 万件产品，2016 年 6 月实际生产 4 万件产品。

$$计划完成相对指标 = \frac{40000}{50000} = 80\%$$

2.5 平均指标

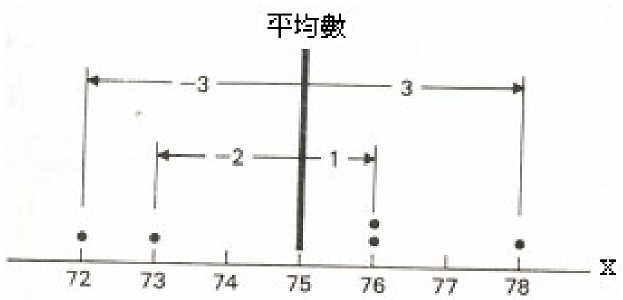


图 2.h 算术平均数是数据的平衡点(重心)

例题 2.10：学生成绩 30 个数据

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：算术平均数 $\mu = \frac{\sum_{i=1}^{30} x_i}{30} = \frac{1800}{30} = 60$

例题 2.10a：大学计算平均总成绩是，每科成绩乘以每科学分数(权数)，除以总学分数。

小明上学期修三门课，成绩如下：统计学 3 学分成绩 86 分，策略管理 3 学分成绩 78 分，通识课程 2 学分成绩 90 分，请问小明的学期总平均？

解答： $\bar{x}=\dfrac{w_1x_1+w_2x_2+w_3x_3}{w_1+w_2+w_3}=\dfrac{3\times 86+3\times 78+2\times 90}{3+3+2}=84.375$

例题 2.11a： 下列 4 个数据，是连续四年的成长率，求平均成长率：

+20%，+33.33%，-6.25%，+10%

解答：几何平均数 $G=\sqrt[4]{(1.2)(1.3333)(0.9375)(1.1)}=\sqrt[4]{1.659}=1.1334$

每年平均成长率 13.34%

例题 2.11： 学生成绩 30 个数据

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：几何平均数 G = 55.83

例题 2.12： 学生成绩 30 个数据

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答： 调和平均数 $H=51.86$

例题 2.12a： 甲、乙两地距离 120 公里，去程开车 1 小时(速度每小时 120 公里)，回程开车 2 小里(速度每小时 60 公里)，计算平均每小时平均速度。

解答： $H=\dfrac{2}{\dfrac{1}{120}+\dfrac{1}{60}}=80$ 等于 $\dfrac{120+120}{3}=80$

例题 2.12b： 第 1 小时开车 120 公里(速度每小时 120 公里)，接着 2 小时开车 180 公里(速度每小时 90 公里)，计算平均每小时平均速度。

解答： $H=\dfrac{300}{\dfrac{120}{120}+\dfrac{180}{90}}=100$ 等于 $\dfrac{120+180}{3}=100$

例题 2.13： C 先生做定期定额投资，每个月固定投资 A 股票 12500 元，第一个月买 A 股票每股 10 元；第二个月买 A 股票每股 12.5 元。请问 C 先生平均每股买多少钱？

解答： $H=\dfrac{2}{\dfrac{1}{10}+\dfrac{1}{12.5}}=11.1111$ $\dfrac{1250\times 10+1000\times 12.5}{1250+1000}=11.1111$

上面右式是用加权平均数：第一个月买 1250 股，第二个月买 1000 股

例题 2.15： 学生成绩 30 个数据

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答： 中位数 $M_e=\dfrac{x_{15}+x_{16}}{2}=\dfrac{50+50}{2}=50$

例题 2.16： 学生成绩 30 个数据

解答： 众数 $M_O=35$

例题 2.17： 学生成绩 30 个数据

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答： 25％ 截尾平均数是 $k=\left\lfloor \dfrac{30\times 25}{100} \right\rfloor = \lfloor 7.5 \rfloor = 7$

$$\bar{x}_T = \frac{\sum_{i=k+1}^{n-k} x_i}{n-2k} = \frac{\sum_{i=8}^{23} x_i}{30-14} = \frac{926}{16} = 57.875$$

5％ 截尾平均数是 $k=\left\lfloor \dfrac{5\times 30}{100} \right\rfloor = \lfloor 1.5 \rfloor = 1$

$$\bar{x}_T = \frac{\sum_{i=k+1}^{n-k} x_i}{n-2k} = \frac{\sum_{i=2}^{29} x_i}{30-2} = \frac{1680}{28} = 60$$

温瑟平均数是 $k=\left\lfloor \dfrac{30}{4} \right\rfloor = \lfloor 7.5 \rfloor = 7$

$$W = \frac{k(Q_1+Q_3)+\sum_{i=k+1}^{n-k} x_i}{n} = \frac{7\times(42.5+85.75)+\sum_{i=8}^{23} x_i}{30} = \frac{897.75+926}{30} = 60.79$$

例题 2.19： 学生成绩 30 个数据

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

(1). 求第 20 个百分位数、 (2). 求第 85 个百分位数、

(3). 求 $x=60$ 的百分位秩、 (4). 求 $x=80$ 的百分位秩。

解答：1. 以数据间距计算百分位数 P_k

(1) 计算第 20 个百分位数(第 2 个十分位数)

$k^* = \lfloor 30\times(0.2)+1-0.2 \rfloor = \lfloor 6.8 \rfloor = 6,$

$P_{20} = x_6 + 0.8(x_7 - x_6) = 36 + 0.8(40-23) = 39.2$

(2) 计算第 85 个百分位数

$k^* = \lfloor 30\times(0.85)+1-0.85 \rfloor = \lfloor 25.65 \rfloor = 25,$

$P_{20} = x_{25} + 0.65(x_{26} - x_{25}) = 88 + 0.65(89-88) = 88.65$

(3) 计算 $x=60$ 的百分位秩

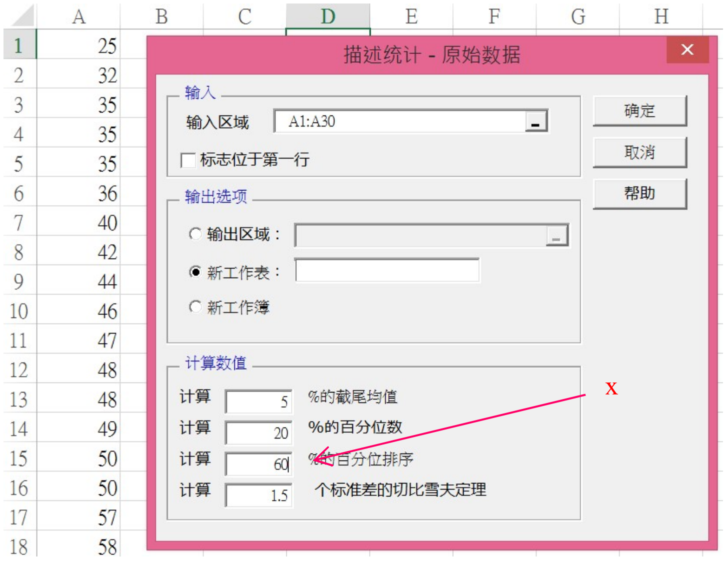
$x_{18} = 58 \le 60 \le 66 = x_{19}$

$m = 18 + \dfrac{60-58}{66-58} - 1 = 17.25$ $\qquad p = 17.25\times\dfrac{100}{29} = 59.48 \Rightarrow P_{59} = 60$

(4) 计算 $x=80$ 的百分位秩

$x_{21} = 78 \le 80 \le 85 = x_{22}$

$m = 21 + \dfrac{80-78}{85-78} - 1 = 20.286$ $\qquad p = 20.286\times\dfrac{100}{29} = 69.95 \Rightarrow P_{70} = 80$



相对位置量数 - 百分位数和百分位排序(百分位阶)

p (%)	第 p 百分位数	
	计算间距	计算个数
20	39.2	36.8

百分位数 x	百分位排序(百分位阶 %)	
	计算间距	计算个数
60	59.48	58.87096774

相对位置量数 - 百分位数和百分位排序(百分位阶)

p (%)	第 p 百分位数	
	计算间距	计算个数
85	88.65	89.7

百分位数 x	百分位排序(百分位阶 %)	
	计算间距	计算个数
80	69.95	68.66359447

2. 以数据个数计算(中文统计 3.0)百分位数 P_k

(1) 计算第 20 个百分位数(第 2 个十分位数)

$$k^* = \left\lfloor \frac{nk+k}{100} \right\rfloor = \left\lfloor \frac{30\times 20+20}{100} \right\rfloor = \lfloor 6.2 \rfloor = 6,$$

$$P_{20} = x_6 + 0.2(x_7 - x_6) = 36 + 0.2(40-36) = 36.8$$

(2) 计算第 85 个百分位数

$$k^* = \left\lfloor \frac{nk+k}{100} \right\rfloor = \left\lfloor \frac{30\times 85+85}{100} \right\rfloor = \lfloor 26.35 \rfloor = 26,$$

$$P_{85} = x_{26} + 0.35(x_{27} - x_{26}) = 89 + 0.35(91-89) = 89.7$$

(3) 计算 $x = 60$ 的百分位秩

$$x_{18} = 58 \le 60 \le 66 = x_{19} \Rightarrow m = 18 + \frac{60-58}{66-58} = 18.25$$

$$p = 18.25 \times \frac{100}{31} = 58.87 \Rightarrow P_{59} = 60$$

(4) 计算 $x=80$ 的百分位秩

$$x_{18}=58\le 60\le 66=x_{19}\Rightarrow m=18+\frac{60-58}{66-58}=18.25$$

$$p=18.25\times\frac{100}{31}=58.87\Rightarrow P_{59}=60$$

3. 以数据个数计算(近似)：

第 20 个百分位数(第 2 个十分位数)

$$\frac{nk}{100}=\frac{30\times 20}{100}=6,\qquad P_{20}=\frac{x_6+x_7}{2}=\frac{36+40}{2}=38$$

第 85 个百分位数

$$\frac{nk}{100}=\frac{30\times 85}{100}=25.5,\qquad P_{85}=x_{26}=89$$

近似法没有百分位秩的公式。

例题 2.20：学生成绩 30 个数据，请计算四分位数。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：

	n=30	中位数 四分位数
资料间距 .INC	$1+\frac{n-1}{4}=8.25$	$Q_1=x_8+0.25(x_9-x_8)$
	$1+\frac{n-1}{2}=15.5$	$Q_2=x_{15}+0.5(x_{16}-x_{15})$
	$1+\frac{3(n-1)}{4}=22.75$	$Q_3=x_{22}+0.75(x_{23}-x_{22})$
资料个数 .EXC	$\frac{n+1}{4}=7.75$	$Q_1=x_7+0.75(x_8-x_7)$
	$\frac{n+1}{2}=15.5$	$Q_2=x_{15}+0.5(x_{16}-x_{15})$
	$\frac{3(n+1)}{4}=23.25$	$Q_3=x_{23}+0.25(x_{24}-x_{23})$

1. 以数据间距计算：

$$Q_1=P_{25}=x_8+0.25(x_9-x_8)=42+0.25(44-42)=42.5,$$
$$Q_2=P_{50}=M_e=x_{15}+0.5(x_{16}-x_{15})=50,$$
$$Q_3=P_{75}=x_{22}+0.75(x_{23}-x_{22})=85+0.75(86-85)=85.75$$

2. 以数据个数计算：

$$Q_1=P_{25}=x_7+0.75(x_8-x_7)=40+0.75(42-40)=41.5,$$
$$Q_2=P_{50}=M_e=x_{15}+0.5(x_{16}-x_{15})=50,$$
$$Q_3=P_{75}=x_{23}+0.25(x_{24}-x_{23})=86+0.25(86-85)=86.25$$

2.6 离差量数

例题 2.21： 1978 年到 1982 年汽车销售量

国家＼年	1978	1979	1980	1981	1982	总和
美国	22	23	7	13	20	85
欧洲	6	4	9	5	2	26
日本	8	2	13	12	9	44
总和	36	29	29	30	31	155

解答：国家变量的异众比率 $V_r = \dfrac{\sum f_i - f_{M_o}}{\sum f_i} = 1 - \dfrac{f_{M_o}}{\sum f_i} = 1 - \dfrac{85}{155} = 0.45$

年度变量的异众比率 $V_r = \dfrac{\sum f_i - f_{M_o}}{\sum f_i} = 1 - \dfrac{f_{M_o}}{\sum f_i} = 1 - \dfrac{36}{155} = 0.77$

例题 2.21a： 表 1.2 泰坦尼克号旅客和组员人数

头等舱	二等舱	三等舱	组员	总和
329	285	710	899	2223

解答：异众比率 $V_r = \dfrac{\sum f_i - f_{M_o}}{\sum f_i} = 1 - \dfrac{f_{M_o}}{\sum f_i} = 1 - \dfrac{899}{2223} = 0.60$

例题 2.22： 学生成绩 30 个数据。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答： 极差(全距) $R = x_{30} - x_1 = 95 - 25 = 70$

四分位差(数据个数) $Q = Q_3 - Q_1 = 86.25 - 41.5 = 44.75$

四分位差(数据间距) $Q = Q_3 - Q_1 = 85.75 - 42.5 = 43.25$

全距因为只计算最大值与最小值，没有考虑中间数据的变化；四分位差也只计算中间 50%的范围，没有两端或内部的变化。全距和四分位差与中位数，合起来就是 3.3.8 节的箱线图。

例题 2.23： 某生英文成绩 80 分，全校平均 70 分，标准差 10 分；他的数学成绩 65 分， 全校平均 55 分，标准差 5 分。请问该生英文和数学，那一科考得比较好？

解答：英文成绩的 Z 分数 $z_i = \dfrac{x_i - \mu}{\sigma} = \dfrac{80-70}{10} = 1$

数学成绩的 Z 分数 $z_i = \dfrac{x_i - \mu}{\sigma} = \dfrac{65-55}{5} = 2$

数学考得比较好。

例题 2.24： 学生成绩 30 个数据。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：(1) 若这群数据是总体的全部数据，则方差 $\sigma^2 = 498.4$，标准差 σ =22.32

(2) 这群数据是抽样数据，则方差 $s^2 = 515.59$，标准差 s = 22.70，

例题 2.25： 学生成绩 30 个数据。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答： 以平均数为中心，平均差为

$$MD_{\mu}=\frac{\sum_{i=1}^{n}|x_i-\mu|}{n}=\frac{\sum_{i=1}^{25}|x_i-60|}{30}=20.2$$

以中位数为中心，平均差为

$$MD_{M_e}=\frac{\sum_{i=1}^{n}|x_i-M_e|}{n}=\frac{\sum_{i=1}^{25}|x_i-50|}{30}=19.2$$

如果是抽样数据，则平均绝对差为

$$MAD=\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n}=\frac{\sum_{i=1}^{25}|x_i-60|}{30}=20.2$$

例题 2.26： 学生成绩 30 个数据。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：　1.　变异系数 VC

$$VC=\frac{\sigma}{\mu}=\frac{22.32}{60}=0.372 \quad 或 \quad VC=\frac{s}{\bar{x}}=\frac{22.7}{60}=0.378$$

2. 平均差系数 MC

$$MC=\frac{MD_{\mu}}{\mu}=\frac{20.2}{60}=0.337 \quad 或 \quad MC=\frac{MD_M}{M_e}=\frac{19.2}{50}=0.384$$

3. 全距系数 RC

$$RC=\frac{x_n-x_1}{x_n+x_1}=\frac{95-25}{95+25}=0.583$$

4. 四分距系数 QC

$$QC=\frac{Q_3-Q_1}{Q_3+Q_1}=\frac{86.25-41.5}{86.25+41.5}=0.35$$

2.7　形态量数

例题 2.27： 学生成绩 30 个数据。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：　(1) 若这群数据是总体的全部数据，则

三级中心距 $M_3=3217.4$　　四级中心距 $M_4=390989.2$

(2) 若这群数据是抽样数据，则

三级中心距 $m_3=3328.3$　　四级中心距 $m_4=404471.6$

例题 2.28： 学生成绩 30 个数据。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：　1. 若这群数据是总体的全部数据，则

(1) 皮尔生偏度指数(Pearson's index of skewness)

$$SK=\frac{3(\mu-M_d)}{\sigma}=\frac{3\times(60-50)}{22.32}=1.344$$ 所以是右偏型 。

(2) 三级距偏度系数

$$SK=\frac{M_3}{\sigma^3}=\frac{3217.4}{(22.32)^3}=0.2889>0$$ 所以是右偏型。

(3) SPSS 和 Statgrahpics 偏度系数

$$SK=\frac{nM_3}{(n-2)\sigma^3}=\frac{30\times(3217.4)}{(30-2)(22.32)^3}=0.31>0$$ 所以是右偏型 。

2. 若这群数据是抽样数据，则

(1) 皮尔生偏度指数(Pearson's index of skewness)

$$SK=\frac{3(\bar{x}-M_d)}{s}=\frac{3\times(60-50)}{22.7}=1.322$$

(2) 三级距偏度系数

$$SK=\frac{m_3}{s^3}=\frac{3328.3}{(22.7)^3}=0.285>0$$

(3) SPSS、Statgrahpics 和 Excel 偏度系数

$$SK=\frac{nm_3}{(n-2)s^3}=\frac{30\times3328.3}{(30-2)(22.7)^3}=0.305>0$$

2.7.3　峰度

例题 2.29： 学生成绩 30 个数据。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：　1. 若这群数据是总体的全部数据，则

(1) 四级距峰度系数

$$K=\frac{M_4}{\sigma^4}=\frac{390989.2}{(22.32)^4}=1.575<3$$ 所以是平峰型。

(2) Excel, SPSS 和 Statgrahpics 峰度系数

$$K==\frac{30(30+1)(390989.2)}{(30-2)(30-3)(22.32)^4}-\frac{3(30-1)^2}{(30-2)(30-3)}=-1.4<0$$ 所以是平峰型。

2. 若这群数据是抽样数据，则

(1) 四级距峰度系数

$$K=\frac{m_4}{s^4}=\frac{404471.6}{(22.7)^4}=1.52<3$$ 所以是平峰型

(2) SPSS 和 Statgrahpics 峰度系数

$$K==\frac{30(30+1)(404471.6)}{(30-2)(30-3)(22.7)^4}-\frac{3(30-1)^2}{(30-2)(30-3)}=-1.47<0$$ 所以是平峰型

查比希夫定理

例题 2.30： 学生成绩 30 个数据。

25,32,35,35,35,36,40,42,44,46,47,48,48,49,50,50,57,58,66,72,78,85,86,87,88,89,91,92,94,95

解答：　平均数和标准差分别为 $\mu = 60$ 和 $\sigma = 22.32$，

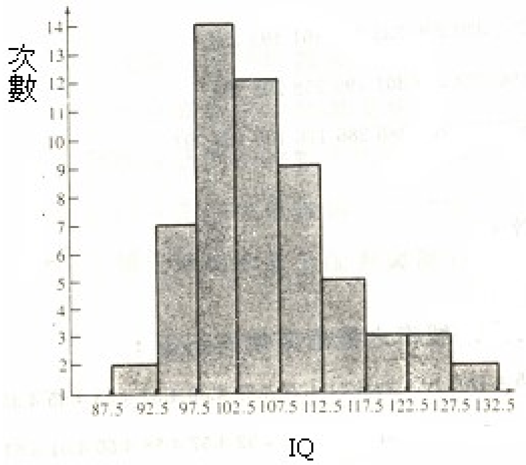
k = 1.5 则 $(\mu - 1.5\sigma, \mu + 1.5\sigma) = (26.52, 93.48)$ 包含 $\frac{27}{30} = 90\%$ 数据，大于 $1 - \frac{1}{k^2} = 56\%$

k = 2 则 $(\mu - 2\sigma, \mu + 2\sigma) = (15.4, 104.6)$ 包含 $\frac{30}{30} = 100\%$ 数据，大于 $1 - \frac{1}{k^2} = 75\%$

习题

1. 就下列各组数据，计算平均数与中位数：
   (1) 4, 7, 3, 6, 5
   (2) 24, 28, 36, 30, 24, 29
   (3) -2, 1, -1, 0, 3, -2, 1
2. 某一保险公司 7 位职员，月薪(以元为单位)分别为
   29500, 27750, 29250, 45000, 31500, 28500, 29750
   (1) 试计算薪水平均数与中位数。　(2) 何者较适合做为集中趋势的代表值，为什么？
3. 就下列数据，计算：
   27, 43, 52, 53, 53, 53, 61, 63, 63, 65, 68, 70, 72, 75, 83, 95, 96, 97, 101,105
   110, 115, 115, 115, 115, 126, 127, 134, 145, 152, 153, 182, 190, 197, 197, 282, 322, 322, 342, 521
   (1) 中位数与四分位数。　(2) 第 20 个百分位数与第 70 个百分位数。
4. 就下列数据，计算：
   239, 212, 249, 227, 218, 310, 281, 330, 226, 233, 223, 161, 195, 233, 249,
   284, 245, 174, 154, 256, 196, 299, 210, 301, 199, 258, 205, 195, 227, 244,
   355, 234, 495, 179, 357, 282, 265, 286, 286, 176, 195, 163, 297
   (1) 平均数与标准差。　(2) 中位数与四分位数。　(3) 全距与四分位差。
5. 下面 50 个观察值为测量某城市之酸雨浓度的记录：
   3.58, 3.80, 4.01, 4.01, 4.05, 4.05, 4.12, 4.18, 4.20, 4.21, 4.27, 4.28, 4.30,
   4.32, 4.33, 4.35, 4.35, 4.41, 4.42, 4.45, 4.45, 4.50, 4.50, 4.50, 4.50, 4.51, 4.52,
   4.52, 4.52, 4.57, 4.58, 4.60, 4.61, 4.61, 4.62, 4.62, 4.65, 4.70, 4.70, 4.70, 4.70,
   4.72, 4.78, 4.78, 4.80, 5.07, 5.20, 5.26, 5.41, 5.48
   (1) 求中位数与四分位数。　(2) 求出第 90 百分位数。
   (3) 求平均数与标准差。　(4) 绘出此组数据的方块图。

6. 转换下列直方图为频数分布表。



7. 利用下列数据：

162, 151, 167, 167, 167, 170, 158, 163, 175, 169, 169, 158, 150, 156, 157, 174, 162, 150, 151, 165, 170, 156, 170, 153, 154, 157, 155, 171, 150, 163, 150, 172, 161, 154, 174, 163, 148, 152, 163, 149, 158, 176, 164, 157, 153, 169, 161, 160, 164, 155

(1) 建立频数分布表。　　(2) 画出各种可能的统计图。

(3) 计算各种代表值与离差值衡量。　　(4) 验证查比希夫定理。

8. 下列 25 个报摊的每日利润额：

55.31, 81.47, 64.90, 70.88, 86.02, 77.25, 76.73, 84.21, 56.02, 84.92, 90.23, 78.01, 88.05, 73.37, 87.09, 57.41, 85.43, 74.76, 86.51, 86.37, 76.15, 88.64, 84.71, 66.05, 83.91

(1) 估计其平均数及众数。

(2) 将数据分为 8 组，组中点分别为 55,60,65,...,90。作频数分布表，估计其分组平均数及众数。

(3) 将数据分为 4 组，组中点分别为 60,70,80,90，作出频数分布表，估计其分组平均数及众数。

(4) 何者的估计与未分组的真实平均数较接近？

9. 下列数据，计算：

学生编号	性别	身高	主修	学生编号	性别	身高	主修
1	F	167	S	26	M	167	B
2	M	172	A	27	M	168	A
3	M	170	S	28	M	172	B
4	M	170	B	29	F	168	A
5	F	161	A	30	F	165	B
6	F	166	B	31	F	165	B
7	M	171	F	32	M	164	B
8	M	167	B	33	M	172	F
9	M	165	S	34	M	167	B
10	F	167	B	35	M	173	S
11	M	174	F	36	F	171	B
12	M	168	S	37	M	171	B
13	M	174	A	38	M	169	S
14	F	164	A	39	F	169	A
15	M	169	S	40	M	147	S
16	M	164	B	41	M	173	B
17	M	172	A	42	M	168	B
18	M	171	B	43	F	166	S
19	F	167	S	44	M	173	A
20	M	170	S	45	M	173	S
21	M	166	S	46	M	167	S
22	F	167	B	47	F	162	S
23	M	168	S	48	M	168	B
24	M	171	F	49	M	171	S
25	M	175	S				

(1) 男性身高的平均数与标准差。 (2) 女性身高的平均数与标准差。

(3) 男性身高的中位数与四分位数。 (4) 女性身高的中位数与四分位数。

(5) 各类主修的频数分布表与饼状图。

10. 试根据下表 65 名职员薪俸数据：（假设此数据为样本数据）

薪资金额(元)	25000	35000	45000	65000	75000	85000	95000
人数	8	10	16	14	10	5	2

(1) 计算薪资平均数。 (2) 计算薪资标准差。 (3) 计算薪资中位数。

(4) 计算薪俸众数。

11. 求得 21 个数值之平均数为 55，样本标准差为 3。后来发觉其中「60」一数必须剔除，如不改变其它的原始数据，试设法计算出剔除「60」一数后，所余 20 个数值之平均数及标准差。

12. 某校抽取学生 50 人，分成甲、乙两组，甲组学生 20 人，其平均分数为 78 分，标准差为 8 分；乙组学生 30 人，其平均分数为 72 分，标准差为 10 分，试求全部样本 50 人之平均成绩及标准差。

13. 某轮胎制造公司每天生产的轮胎数量不一定，最近 30 天中 ，每天生产的数量分别

如下： 93, 86, 100, 92, 88, 80, 85, 93, 88, 78, 95101, 99, 86, 87, 79, 84, 76, 71, 109, 85, 79, 89, 110, 97, 93, 79, 86, 88

(1) 问该公司最近 30 天，平均每天生产多少个轮胎？

(2) 求中位数，与第十个百分位

(3) 求方差，标准差与变异系数

(4) 求全距与四分位差

(5) 求平均绝对差

14. 人事经理由该公司的人事数据中，获悉该公司员工的平均年龄为 32.4 岁，标准差为 19.6 岁；而由薪资档案中，获悉该公司员工的薪资平均为每人每个月 9506 元，标准差为 8613 元。请问此公司员工年龄的离散程度与薪资的离散程度何者较大？为什么？

15. 有一个橡皮筛工厂，生产了一堆圆柱形橡皮筛，半径平均是 1.02 公分，半径的标准差是 0.02 公分。

(1) 判断出这堆橡皮筛中半径介于 0.98 公分与 1.06 公分之间所占的比例？

(2) 有顾客要买一堆此形的橡皮筛，但要求的规格必须是半径介于 0.90 公分与 1.10 公分之间的比例不得低于 93%，该工厂生产的这堆橡皮筛是否符合这位顾客的规格要求？

16. 利用下列样本数据：

1, 2, 12, 3, 15, 5, 12, 11, 3, 4, 3, 5, 0, 7, 17, 6, 17, 13, 2, 5, 5, 7, 1, 11, 3, 9, 9, 8, 18, 8, 10, 9, 4, 12, 1, 8, 8, 7, 11, 9, 15, 11, 8, 4, 5, 11, 3, 14, 12, 10

(1) 建立频数分布表。 (2) 画出各种可能的统计图。

(3) 计算各种代表值与离差值衡量。 (4) 验证查比希夫定理。

17. 利用下列数据：

67, 71, 90, 46, 51, 71, 66, 54, 46, 22, 74, 34, 65, 55, 63, 69, 61, 57, 46, 84

(1) 建立频数分布表。 (2) 画出各种可能的统计图。

(3) 计算各种代表值与离差值衡量。 (4) 验证查比希夫定理。

18. 如何将直方图转换为箱线图?

19. 如何将箱线图转换为直方图?

18. 甲县内共有 4 家汽车经销商: 运输汽车(C)、马可汽车(M)、三角汽车(T)与全球汽车(U) ，抽样 40 位买车主被抽样询问甲县内那家汽车经销商提供的服务最好，以下为抽样数据:

T	C	C	C	G	C	M	T	C	G
G	M	C	M	T	C	M	M	C	M
T	C	C	T	G	M	M	C	C	T
T	G	C	G	T	M	M	C	G	T

(1) 请建构出频数分布表与频率分布表。

(2) 请建构条形图，请问那个汽车经销商服务最差?

(3) 请建构圆饼图，请问那个汽车经销商服务最好?

19. 下表为 40 位销售员的年龄:

47	21	37	53	28	30	45	31	41	56
40	30	32	34	26	39	31	33	35	25
34	24	24	35	45	33	23	25	36	46
38	35	28	43	45	39	34	27	42	44

(1) 请分成 4 组，建构出直方图，请问本直方图的形态?

(2) 请分成 6 组，建构出直方图，请问本直方图的形态?

(3) 请以 10 位数为茎、个位数为叶画出茎叶图，请问本茎叶图的形态?

(4) 请画出累积频率图(ogive)，并估计 30 岁以下的销售员比率?40 岁以上的销售员比率? 40 岁以上与 50 岁以下的销售员比率?

20. 下表为 50 位学生微积分考试的成绩:

63	74	42	65	51	54	36	56	68	57
62	64	76	67	79	61	81	77	59	38
84	68	71	94	71	86	69	75	91	55
48	82	83	54	79	62	68	58	41	47
83	67	72	95	70	85	70	76	90	46

(1) 请分成 4 组，建构出直方图，请问本直方图的形态?

(2) 请分成 6 组，建构出直方图，请问本直方图的形态?

(3) 请以 10 位数为茎、个位数为叶画出茎叶图，请问本茎叶图的形态?

(4) 请画出累积频率图(ogive)，并估计 60 分以下的学生比例？80 分以上的学生比例？介于 70 分与 80 分的学生比例？

21. 下表为过去一年 12 个月爱乐超级市场的销售状况(以百万计):

月份	1	2	3	4	5	6	7	8	9	10	11	12
销售	78	74	83	87	85	93	100	105	103	89	78	94

(1) 请建构出条形图，请问本条形图的趋势?

(2) 请建构出折线图，请问本折线图的趋势?

(3) 请从条形图与折线图找出爱乐超级市场销售的特性。

22. 假设你投资一个三年的连续投资项目，总金额为 NT$1,000,000，第一年的收益为 100%总金额成长为 NT$2,000,000，第二年则损失 40%总金额降至 NT$1,200,000，第三年再损失 10%总金额再降至 NT$1,080,000。

(1) 请计算三年收益或损失的算术平均数。

(2) 请计算三年收益或损失的几何平均数。

(3) 请比较以上的两个平均数，那个较能代表实际的状况。

23. 以下为两项基金过去 10 年的回收状况:

基金 A:	7.1	7.4	19.7	3.9	32.4	41.7	23.2	4.0	1.9	29.3
基金 B:	10.8	4.1	5.1	10.9	26.5	24.0	16.9	9.4	2.6	10.1

(1) 请计算两项基金的平均收益或损失，请问若依据此平均收益或损失，你会建议投资那项基金。
(2) 请计算两项基金的收益或损失标准差，请问若依据此收益或损失标准差，你会建议投资那项基金。
(3) 请计算两项基金的收益或损失变异系数，请问若依据此收益或损失变异系数，你会建议投资那项基金。
(4) 若两项基金的相关系数为 0.25，请问若两项基金都各投资一半，此组合收益或损失变异系数为?
(5) 若两项基金的相关系数为 0.25，请问若两项基金都投资，你会建议投资比率为。

24. 以下为从政府部门公务机关抽样出来的 50 位公务员的年龄：31, 43, 56, 23, 49, 42, 33, 61, 44, 28, 48, 38, 44, 35, 40, 64, 52, 42, 47, 39, 53, 27, 36, 35, 20, 30, 44, 55, 22, 50, 41, 34, 60, 43, 27, 49, 37, 43, 36, 41, 63, 51, 43, 48, 40, 52, 28, 35, 36, and 21。
(1) 请计算这 50 位公务员的年龄平均数、中位数。
(2) 请计算这 50 位公务员的年龄全距、方差与标准差。
(3) 请分成 5 组，建构这 50 位公务员的年龄直方图。
(4) 依据此直方图，决定分布状况与偏度，依此分布状况与偏度决定经验法则或薛比雪夫法则较能适用于描述这 50 位公务员的年龄? 请依此法则计算 95%的最高年龄与最低年龄。
(5) 请将 50 位公务员年龄做成箱线图，依此图解释这 50 位公务员年龄的分布状况与偏度。并计算是否有过低或过高的年龄(偏离值，outliers)。请比较此方块图与直方图的结果。
(6) 找出这 50 位公务员年龄 95%百分位数、与公务员年龄为 55 的百分比级数。

25. 以下为抽样 40 位销售员的销售状况(以千计)：164, 148, 137, 157, 173, 156, 177, 172, 169, 165, 145, 168, 163, 162, 174, 152, 156, 168, 154, 151, 174, 146, 134, 140, 171, 146, 167, 164, 161, 175, 151, 157, 169, 153, 152, 175, 145, 135, 139, 172。
(1) 请计算这 40 位销售员的销售平均数、中位数。
(2) 请计算这 40 位销售员的销售全距、方差与标准差。
(3) 请分成 5 组，建构这 40 位销售员的销售直方图。
(4) 依据此直方图，决定分布状况与偏度，依此分布状况与偏度决定经验法则或薛比雪夫法则较能适用于描述这 40 位销售员的销售? 请依此法则计算 95%的最高年龄与最低年龄。
(5) 请将 40 位销售员销售做成方块图，依此图解释这 40 位销售员销售的分布状况与偏度。并计算是否有过低或过高的年龄(偏离值，outliers)。请比较此方块图与直方图的结果。
(6) 找出这 40 位销售员销售 95%百分位数、与公务员年龄为 150 的百分比级数。

26. 给定以下为 16 位公司雇员的每月薪资(以千计)与年资(以年计)：

年资	5	3	7	9	2	4	6	8	6	4	8	10	3	5	7	9	6
薪资	40	28	30	52	24	42	34	58	45	22	36	67	32	48	56	58	53

(1) 请画出这 16 位公司雇员的每月薪资与年资的散布图，其中年资为横轴(X)而每月薪资为纵轴(Y)。从散点图解释这 16 位公司雇员的每月薪资与年资间的关系。

(2) 请计算 16 位公司雇员的每月薪资与年资的共变量，从共变量解释这 16 位公司雇员的每月薪资与年资间的关系。

(3) 请计算 16 位公司雇员的每月薪资与年资的相关系数，从相关系数解释这 16 位公司雇员的每月薪资与年资间的关系。

描述统计 补充教材

2.12　分组资料的描述统计

所谓分组数据是，数据已经变成「次数分配表」的型式。这种数据通常是次级数据，换言之，这是已经由别人整理过的资料。我们知道次数分配表有三种型式：间断数据次数分配表，连续数据次数分配表，以及单值分组次数分配表。要计算分组资料的叙述统计，前两种型式(有上组界及下组界)，取每组的组中点。单值分组，则取其单值作组中点，但是等于课本的不分组资料叙述统计。

如果分组数据有开放组界(open-ended class)，则不能计算各种平均数、全距、方差标准差等量数(集中趋势与离差)，但是可以计算中位数、众数、四分位数等量数，除非这些量数刚好在开放组界。

2.12.1　算术平均数

定义：假设数据有 k 组，$x_1, x_2, \cdots, x_k$ 为各组的组中点(middle point)，$f_1, f_2, \cdots, f_k$ 为各组的次数(frequency)，$N = \sum f_i$ 是全部资料的数目。算术平均数(arithmetic mean)，记作 μ 或 $\bar{x}$。以下我们分两种情况来定义算术平均数：

(1) 这群资料是母体的全部资料，则算术平均数

$$\mu = \frac{\sum_{i=1}^{k} f_i x_i}{N}$$

(2) 这群数据是抽样数据，则算术平均数

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

通常我们将算术平均数，简称平均数。

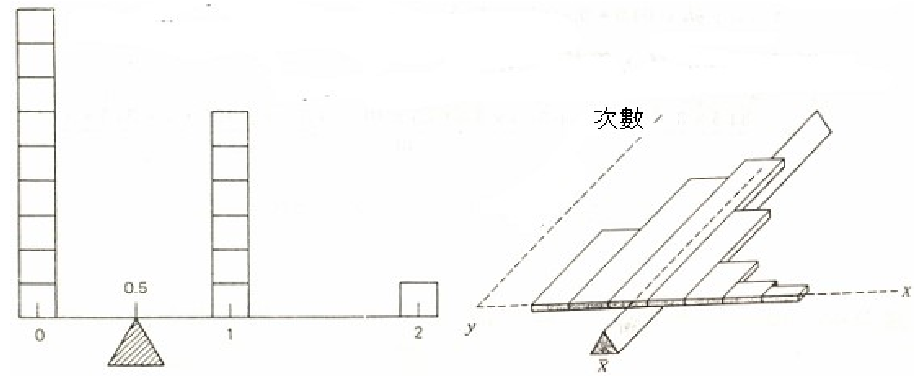


图 2.22 分组数据算术平均数也是数据平衡点

上述公式可能计算起来很繁，以下我们介绍简捷计算法。

假设组宽均相等为 h，数据总数是 N。先找一个假定平均数 a，通常是中间那个组的组中点。将 x_i 转换到 y_i： $y_i = \frac{x_i - a}{h}$ （使 y_i 为整数）。

计算 $\bar{y} = \frac{\sum_{i=1}^{k} f_i y_i}{N}$

原来分组数据的平均数： $\bar{x} = a + h\bar{y}$

例题 2.31：下列分组数据，计算平均数：

表 2.7 分组资料

组界	组中点 x_i	频数 f_i	$y_i = \frac{x_i - 64.5}{10}$	$f_i y_i$
30 ~ 39	34.5	3	-3	-9
40 ~ 49	44.5	1	-2	-2
50 ~ 59	54.5	7	-1	-7
60 ~ 69	64.5	10	0	0
70 ~ 79	74.5	8	1	8
80 ~ 89	84.5	7	2	14
90 ~ 99	94.5	4	3	12
		40		16

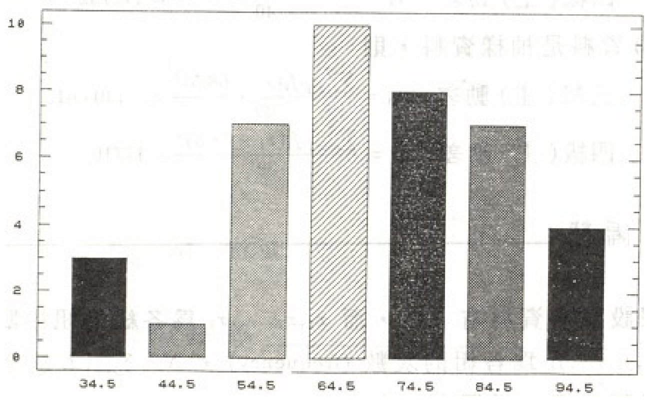


图 2.23 分组资料直方图

解答：计算 $\bar{y}=\frac{\sum_{i=1}^{k} f_i y_i}{N}=\frac{16}{40}=0.4$

原来分组数据的平均数： $\bar{x}=a+h\bar{y}=64.5+10\times(0.4)=68.5$

依照原公式计算，算术平均数

$$\bar{x}=\frac{3\times34.5+1\times44.5+7\times54.5+10\times64.5+8\times74.5+7\times84.5+4\times94.5}{40}=68.5$$

2.12.2 几何平均数

定义：假设数字数据有 k 组，$x_1,x_2,\cdots,x_k$ 为各组的组中点，$f_1,f_2,\cdots,f_k$ 为各组的次数， $N=\sum_{i=1}^{k} f_i$ 是全部资料的数目。几何平均数(geometric mean)，记作 G，定义如下：

$$G=\sqrt[N]{x_1^{f_1}\cdot x_2^{f_2}\cdot\cdots\cdot x_k^{f_k}}=\sqrt[N]{\prod_{i=1}^{k} x_i^{f_i}}$$

例题 2.32：表 2.7 的分组数据，计算几何平均数。

解答：

$$G=\sqrt[N]{x_1^{f_1}\cdot x_2^{f_2}\cdot\cdots\cdot x_k^{f_k}}=\sqrt[40]{(34.5)^3(44.5)^1(54.5)^7(64.5)^{10}(74.5)^8(84.5)^7(94.5)^4}=66.371$$

2.12.3 调和平均数

定义：假设数字数据有 k 组，$x_1,x_2,\cdots,x_k$ 为各组的组中点，$f_1,f_2,\cdots,f_k$ 为各组的次数，$N=\sum_{i=1}^{k} f_i$ 是全部资料的数目。调和平均数(harmonic mean)，记作 H，定义如下：

$$H=\frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_k}{x_k}}=\frac{N}{\sum_{i=1}^{k}\frac{f_i}{x_i}}$$

例题 2.33：表 2.7 的分组数据，计算调和平均数。

解答：$H=\frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_k}{x_k}}=\frac{40}{\frac{3}{34.5}+\frac{1}{44.5}+\frac{7}{54.5}+\frac{10}{64.5}+\frac{8}{74.5}+\frac{7}{84.5}+\frac{4}{94.5}}=63.953$

我们注意到：$\mu\geq G\geq H$

2.12.4 中位数

定义：假设数字数据有 k 组，$f_1,f_2,\cdots,f_k$ 为各组的次数，$N=\sum_{i=1}^{k} f_i$ 是全部资料的

数目。中位数(median)，记作 M，计算如下：

(1) 从第一组累计次数，找出中位数所在的组，假设为第 m 组。

$$\sum_{i=1}^{m} f_i \ge \frac{N}{2} \quad \text{且} \quad \sum_{i=m}^{k} f_i \ge \frac{N}{2}$$

(2) 令 L_m 为第 m 组的下界， U_m 为第 m 组的上界。

(3) $$M = L_m + \left(\frac{N}{2} - \sum_{i=1}^{m-1} f_i\right)\left(\frac{U_m - L_m}{f_m}\right)$$

例题 2.34：表 2.7 的分组数据，计算中位数。

解答：中位数在第 4 组，m= 4, $L_m = 60$, $U_m = 69$，

$$M = L_m + \left(\frac{N}{2} - \sum_{i=1}^{m-1} f_i\right)\left(\frac{U_m - L_m}{f_m}\right) = 60 + \left(\frac{40}{2} - [3+1+7]\right)\left(\frac{69-60}{10}\right) = 68.1$$

2.12.5 众数

定义：假设数字数据有 k 组，$f_1, f_2, \cdots, f_k$ 为各组的次数，$N = \sum_{i=1}^{k} f_i$ 是全部资料的数目。找出次数最多的组，假设为第 m 组。 $f_m = \max_i f_i$ 。令 L_m 为第 m 组的下界，U_m 为第 m 组的上界。分组资料的众数(mode)，记作 M_o，有下列几种计算方法：

(1) 金氏众数(King's mode)

令 $$\frac{M_o - L_m}{U_m - L_m} = \frac{f_{m+1}}{f_{m-1} + f_{m+1}}$$

$$M_o = L_m + \left(\frac{f_{m+1}}{f_{m-1} + f_{m+1}}\right)(U_m - L_m)$$

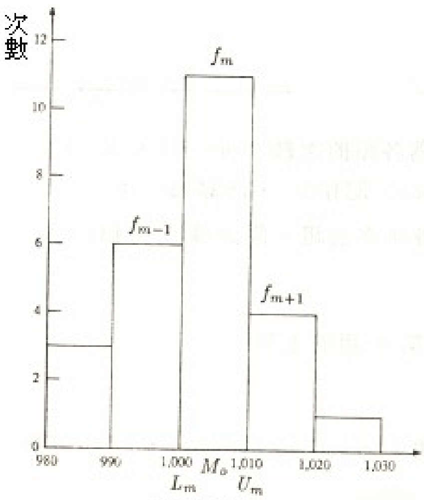
(2) 苏伯众数(Czuber's mode)

令 $$\frac{M_o - L_m}{U_m - L_m} = \frac{f_m - f_{m-1}}{f_m - f_{m-1} + f_m - f_{m-1}}$$

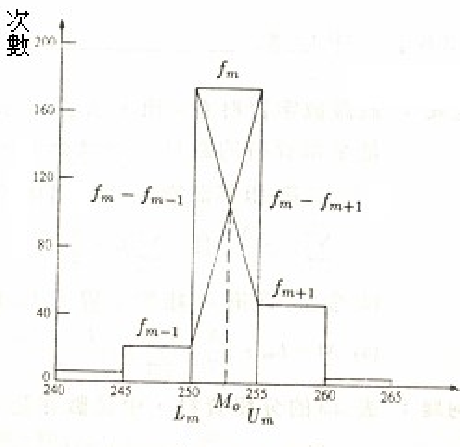
$$M_o = L_m + \left(\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}}\right)(U_m - L_m)$$

(3) 皮尔生众数(Pearson's mode)

$$M_o = \mu - 3(\mu - M)$$



金氏衆數



蘇伯衆數

图 2.24 分组资料众数

例题 2.35： 表 2.7 的分组数据，计算众数。

解答：$m=4$，$L_m=60$，$U_m=69$

(1) 金氏众数

$$M_o=L_m+\left(\frac{f_{m+1}}{f_{m-1}+f_{m+1}}\right)(U_m-L_m)=60+\left(\frac{8}{7+8}\right)(69-60)=64.8$$

(2) 苏伯众数

$$M_o=L_m+\left(\frac{f_m-f_{m-1}}{2f_m-f_{m-1}-f_{m+1}}\right)(U_m-L_m)=60+\left(\frac{10-7}{2\times10-7-8}\right)(69-60)=65.4$$

(3) 皮尔生众数 $M_o=68.5-3(68.5-68.1)=67.3$

2.12.6 四分位数

定义： 假设数字数据有 k 组，$f_1,f_2,\cdots,f_k$ 为各组的次数，$N=\sum_{i=1}^{k}f_i$ 是全部资料的数目。第 i 个四分位数(quartiles)，记作 Q_i，$i=1,2,3$，计算如下：

(1) 从第一组累计次数，找出第 i 个四分位数所在的组，假设为第 q 组。

$$\sum_{j=1}^{q}f_j\ge\frac{N\times i}{4} \quad 且 \quad \sum_{j=q}^{k}f_j\ge1-\frac{N\times i}{4}$$

(2) 令 L_q 为第 q 组的下界，U_q 为第 q 组的上界。

(3) $$Q_i=L_q+\left(\frac{N\times i}{4}-\sum_{j=1}^{q-1}f_j\right)\left(\frac{U_q-L_q}{f_q}\right)$$

例题 2.36： 表 2.7 的分组数据，计算四分位数。

解答：$$Q_1=L_q+\left(\frac{N\times i}{4}-\sum_{j=1}^{q-1}f_j\right)\left(\frac{U_q-L_q}{f_q}\right)=50+\left(\frac{40\times1}{4}-[3+1]\right)\left(\frac{59-50}{7}\right)=57.714$$

$$Q_2=M=60+\left(\frac{40\times2}{4}-[3+1+7]\right)\left(\frac{69-60}{10}\right)=68.1$$

$$Q_3 = 80 + \left(\frac{40\times 3}{4} - [3+1+7+10+8]\right)(\frac{89-80}{7}) = 81.286$$

2.12.7 百分位数

定义：假设数字数据有 k 组，$f_1, f_2, \cdots, f_k$ 为各组的次数，$N = \sum_{i=1}^{k} f_i$ 是全部资料的数目。第 i 个百分位数(percentiles)，记作 P_i，计算如下：

(1) 从第一组累计次数，找出第 i 个百分位数所在的组，假设为第 p 组。

$$\sum_{j=1}^{p} f_j \ge \frac{N\times i}{100} \quad 且 \quad \sum_{j=p}^{k} f_j \ge 1 - \frac{N\times i}{100}$$

(2) 令 L_p 为第 p 组的下界，U_p 为第 p 组的上界。

(3) $$P_i = L_p + \left(\frac{N\times i}{100} - \sum_{j=1}^{p-1} f_j\right)(\frac{U_p - L_p}{f_p})$$

例题 2.37：表 2.7 的分组数据，计算百分位数。

解答：第 10 个百分位数(第 1 个十分位数) $P_{10} = 40 + \left(\frac{40\times 10}{100} - 3\right)(\frac{49-40}{1}) = 49$

第 20 个百分位数(第 2 个十分位数) $P_{20} = 50 + \left(\frac{40\times 20}{100} - [3+1]\right)(\frac{59-50}{7}) = 55.143$

第 85 个百分位数 $P_{85} = 80 + \left(\frac{40\times 850}{100} - [3+1+7+10+8]\right)(\frac{89-80}{7}) = 86.429$

2.12.8 全距与四分位距

定义：假设数字数据有 k 组，$f_1, f_2, \cdots, f_k$ 为各组的次数，$N = \sum_{i=1}^{k} f_i$ 是全部资料的数目。令 L_1 为第 1 组的下界，U_k 为第 k 组的上界。则全距(range) R 为 $R = U_k - L_1$

定义：若第一个四分位数和第三个四分位数，分别为 Q_1 和 Q_3，则四分位距 Q (interquartile range) 为 $Q = Q_3 - Q_1$

例题 2.38：表 2.7 的分组数据，计算全距与四分位距。

解答：全距 R 为 $R = U_k - L_1 = 99 - 30 = 69$

四分位距 Q 为 $Q = Q_3 - Q_1 = 81.286 - 57.714 = 23.572$

2.12.9 方差与标准差

定义：假设数字数据有 k 组，$x_1, x_2, \cdots, x_k$ 为各组的组中点，$f_1, f_2, \cdots, f_k$ 为各组的次数，$N = \sum_{i=1}^{k} f_i$ 是全部资料的数目。以下我们分两种情况来定义方差(variance)：

(1) 这群资料是母体的全部资料，则方差

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{k} f_i x_i^2}{N} - \mu^2$$

(2) 这群数据是抽样数据 ， 则方差

$$s^2=\frac{\sum_{i=1}^{k}f_i(x_i-\bar{x})^2}{N-1}=\frac{\sum_{i=1}^{k}f_ix_i^2-N\bar{x}^2}{N-1}=\frac{1}{N-1}\left[\sum_{i=1}^{k}f_ix_i^2-\frac{1}{N}(\sum_{i=1}^{k}f_ix_i)^2\right]$$

定义：方差的正平方根，称为标准差(standard deviation)。

(1) 这群数据是母体的全部资料，则标准差

$$\sigma=\sqrt{\frac{\sum_{i=1}^{k}f_i(x_i-\mu)^2}{N}}=\sqrt{\frac{\sum_{i=1}^{k}f_ix_i^2}{N}-\mu^2}$$

(2) 这群数据是抽样数据，则标准差

$$s=\sqrt{\frac{\sum_{i=1}^{k}f_i(x_i-\bar{x})^2}{N-1}}=\sqrt{\frac{\sum_{i=1}^{k}f_ix_i^2-N\bar{x}^2}{N-1}}$$

以下我们介绍简捷计算法。假设每组组宽均相等为 h，数据总数是 N。

1. 先找一个假定平均数 a，通常是中间那个组的组中点。
2. 将各组组中点将 x_i 转换到 y_i： $y_i=\dfrac{x_i-a}{h}$
3. 计算 $\mu=\dfrac{\sum_{i=1}^{k}f_iy_i}{N}$ 或 $\bar{y}=\dfrac{\sum_{i=1}^{k}f_iy_i}{N}$
4. 计算 $\sigma_y^2=\dfrac{\sum_{i=1}^{k}f_i(y_i-\mu)^2}{N}=\dfrac{\sum_{i=1}^{k}f_iy_i^2}{N}-\mu^2$

$$s_y^2=\frac{\sum_{i=1}^{k}f_iy_i^2-N\bar{y}^2}{N-1}=\frac{N\sigma_y^2}{N-1}$$

5. 原来分组数据的平均数：$\bar{x}=a+h\bar{y}$
6. 原来分组数据的方差：母体资料 $\sigma_x^2=h^2\sigma_y^2$ ，样本数据 $s_x^2=h^2s_y^2$

例题 2.39：表 2.7 的分组数据，计算方差与标准差。

解答：(1) 这群资料是母体的全部资料，则方差 $\sigma^2=\dfrac{\sum_{i=1}^{k}f_i(x_i-\mu)^2}{N}=259$

标准差 $\sigma=\sqrt{\dfrac{\sum_{i=1}^{k}f_i(x_i-\mu)^2}{N}}=16.093$

(2) 这群数据是抽样数据，则方差 $s^2=\dfrac{\sum_{i=1}^{k}f_i(x_i-\bar{x})^2}{N-1}=265.641$

标准差　$s = \sqrt{\dfrac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{N-1}} = 16.298$

组界	组中点 x_i	频数 f_i	$y_i = \dfrac{x_i - 64.5}{10}$	$f_i y_i$	y_i^2	$f_i y_i^2$
30 ~ 39	34.5	3	-3	-9	9	27
40 ~ 49	44.5	1	-2	-2	4	4
50 ~ 59	54.5	7	-1	-7	1	7
60 ~ 69	64.5	10	0	0	0	0
70 ~ 79	74.5	8	1	8	1	8
80 ~ 89	84.5	7	2	14	4	28
90 ~ 99	94.5	4	3	12	9	36
$\sum$		$N = 40$		16		110
$\sum \div N$				0.4		2.75

计算　$\sigma_y^2 = \dfrac{\sum_{i=1}^{k} f_i y_i^2}{N} - \bar{y}^2 = \dfrac{110}{40} - (0.4)^2 = 2.59$

$$s_y^2 = \frac{N\sigma_y^2}{N-1} = \frac{40(2.59)}{39} = 2.6564$$

原来分组数据的方差：

$$\sigma_x^2 = h^2\sigma_y^2 = 100 \times 2.59 = 259$$

$$s_x^2 = \frac{N}{N-1}\sigma_x^2 = \frac{40}{39} \times 259 = 265.64$$

2.12.10　三阶距与四阶距

定义：假设数字数据有 k 组，$x_1, x_2, \cdots, x_k$ 为各组的组中点，$f_1, f_2, \cdots, f_k$ 为各组的次数，$N = \sum_{i}^{k} f_i$ 是全部资料的数目。以下我们分两种情况来定义三阶距(third moment)：

(1)这群资料是母体的全部资料，则三阶中心距

$$M_3 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^3}{N}$$

(2) 这群数据是抽样数据，则三阶中心距

$$m_3 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^3}{N-1}$$

定义：假设数字数据有 k 组，$x_1, x_2, \cdots, x_k$ 为各组的组中点，$f_1, f_2, \cdots, f_k$ 为各组的次数，$N = \sum_{i=1}^{k} f_i$ 是全部资料的数目。以下我们分两种情况来定义四阶距(fourth moment)：

(1) 这群资料是母体的全部资料，则四阶中心距

$$M_4 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^4}{N}$$

(2) 这群数据是抽样数据，则四阶中心距

$$m_4 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^4}{N-1}$$

例题 2.40： 表 2.7 的分组数据，计算三阶及四阶中心距。

解答：(1) 数据是母体的全部资料，则三阶中心距

$$M_3 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^3}{N} = \frac{\sum_{i=1}^{7} f_i (x_i - 68.5)^3}{40} = -1272$$

四阶中心距

$$M_4 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^4}{N} = \frac{\sum_{i=1}^{7} f_i (x_i - 68.5)^4}{40} = 172732$$

(2) 数据是抽样数据，则三阶中心距

$$m_3 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^3}{N} = \frac{\sum_{i=1}^{7} f_i (x_i - 68.5)^3}{39} = -1304.615$$

四阶中心距

$$m_4 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^4}{N} = \frac{\sum_{i=1}^{7} f_i (x_i - 68.5)^4}{39} = 177161$$

2.12.11　偏度

定义：假设数字数据有 k 组，$x_1, x_2, \cdots, x_k$ 为各组的组中点，$f_1, f_2, \cdots, f_k$ 为各组的次数，$N = \sum_{i=1}^{k} f_i$ 是全部资料的数目。以下我们有几个方法来定义偏度(skewness) SK：数据是母体数据或抽样数据，则

(1) 皮尔生偏度指数(Pearson's index of skewness)

$$SK = \frac{3(\mu - M)}{\sigma} \quad \text{或样本数据} \quad SK = \frac{3(\bar{x} - M)}{s}$$

(2) 利用三阶距

$$SK = \frac{M_3}{\sigma^3} \quad \text{或样本数据} \quad SK = \frac{m_3}{s^3}$$

(3) SPSS 和 Statgrahpics 的公式

$$SK=\frac{NM_3}{(N-2)\sigma^3},\qquad \sigma\neq 0,\ \ N>3$$

$$SK=\frac{Nm_3}{(N-2)s^3},\qquad \sigma\neq 0,\ \ N>3$$

例题 2.41：表 2.7 的分组数据，计算偏度。

解答：(1) 皮尔生偏度指数(Pearson's index of skewness)

$$SK=\frac{3(\mu-M)}{\sigma}=\frac{3(68.5-68.1)}{16.093}=0.075$$

样本数据 $SK=\dfrac{3(\bar{x}-M)}{\sigma}=\dfrac{3(68.5-68.1)}{16.298}=0.074$

(2) 利用三阶距

$$SK=\frac{-1272}{(16.093)^3}=-0.305$$

样本数据 $SK=\dfrac{-1304.615}{(16.298)^3}=-0.301$

(3) SPSS 和 Statgrahpics 利用 Bliss(1967) 的公式

$$SK=\frac{40(-1272)}{38(16.093)^3}=-0.321$$

样本数据 $SK=\dfrac{40(-1304.615)}{38(16.298)^3}=-0.317$

注意：以上偏度系数，皮尔生指数和另两者不同。

皮尔生偏度指数大于 0，所以是右偏；三阶距偏度系数小于 0，所以是左偏。

但是我们数据表或直方图，可以看出：右边的概率较高，所以应该是左偏。

2.12.12 峰度

定义：假设数字数据有 k 组，$x_1,x_2,\cdots,x_k$ 为各组的组中点，$f_1,f_2,\cdots,f_k$ 为各组的次数，$N=\sum^k f_i$ 是全部资料的数目。以下我们有几个方法来定义峰度(kurtosis) K：数据是母体数据或抽样数据，则

(1) 利用四阶距

母体资料 $K=\dfrac{M_4}{\sigma^4}$ 或 样本数据 $K=\dfrac{m_4}{s^4}$

若 K < 3，则为平峰型。若 K = 3，则为正态峰。若 K > 3，则为尖峰型。

(2) SPSS 和 Statgrahpics 的公式

$$K=\frac{N(N+1)M_4}{(N-2)(N-3)\sigma^4}-\frac{3(N-1)^2}{(N-2)(N-3)},\qquad \sigma\neq 0,\ \ N>4$$

$$K=\frac{N(N+1)m_4}{(N-2)(N-3)s^4}-\frac{3(N-1)^2}{(N-2)(N-3)},\qquad s\neq 0,\ \ N>4$$

若 K < 0，则为平峰型。若 K = 0，则为正态峰。若 K > 0，则为尖峰型。

例题 2.42：表 2.7.的分组数据，计算峰度。

解答：(1) 利用四阶距

$$K = \frac{M_4}{\sigma^4} = \frac{172732}{(16.093)^4} = 2.575$$

$$K = \frac{m_4}{s^4} = \frac{177161}{(16.298)^4} = 2.510$$

峰度系数小于 3，所以是平峰型。

(2) SPSS 和 Statgrahpics 的公式

$$K = \frac{N(N+1)M_4}{(N-2)(N-3)\sigma^4} - \frac{3(N-1)^2}{(N-2)(N-3)} = \frac{40(41)(172732)}{(38)(37)(16.093)^4} - \frac{3(39)^2}{(38)(37)} = -0.241$$

$$K = \frac{N(N+1)m_4}{(N-2)(N-3)s^4} - \frac{3(N-1)^2}{(N-2)(N-3)} = \frac{40(41)(177161)}{(38)(37)(16.298)^4} - \frac{3(39)^2}{(38)(37)} = -0.317$$

峰度系数小于 0，所以是平峰型。