

## 第二节 直方图和箱线图

一、直方图

二、箱线图

三、小结

# 一、直方图

例1 下面给出了84个伊特拉斯坎 (Etruscan) 人男子的头颅的最大宽度 (mm) , 现在来画这些数据的“频率直方图”。

141	148	132	138	154	142	150	146	155	158	150	140
147	148	144	150	149	145	149	158	143	141	144	144
126	140	144	142	141	140	145	135	147	146	141	136
140	146	142	137	148	154	137	139	143	140	131	143
141	149	148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138	142	149
142	137	134	144	146	147	140	142	140	137	152	145





**步骤:**

1. 找出最小值126, 最大值158, 现取区间  $[124.5, 159.5]$ ;
2. 将区间  $[124.5, 159.5]$  等分为7个小区间, 小区间的长度记成  $\Delta$ ,  $\Delta = (159.5 - 124.5) / 7 = 5$ ,  $\Delta$  称为组距;
3. 小区间的端点称为组限, 数出落在每个小区间的数据的频数  $f_i$ , 算出频率  $f_i / n$ .



列表如下：

组 限	频 数	频 率	累计频率
124.5~129.5	1	0.0119	0.0119
129.5~134.5	4	0.0476	0.0595
134.5~139.5	10	0.1191	0.1786
139.5~144.5	33	0.3929	0.5715
144.5~149.5	24	0.2857	0.8572
149.5~154.5	9	0.1071	0.9643
154.5~159.5	3	0.0357	1.0000

现在自左向右依次在各个小区间上作以  $\frac{f_i}{n} / \Delta$  为高的小矩形，这样的图形叫**频率直方图**。



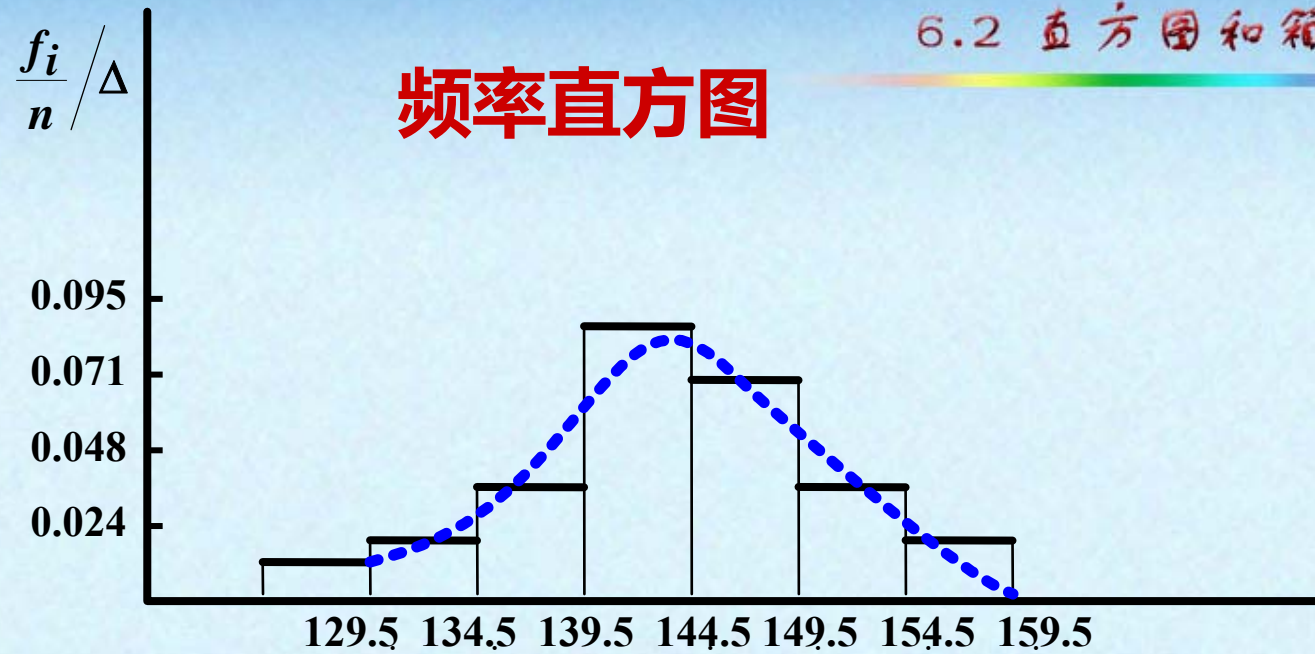


图 6-1

每个小矩形的面积正是频率。若 $n$ 很大，频率趋近于概率，而概率密度曲线所覆盖的面积正是概率，因此，直方图的轮廓近似为概率密度曲线。





## 二、箱线图

**定义** 设有容量为 $n$ 的样本观察值  $x_1, x_2, \dots, x_n$ , 样本  $p$  分位数 ( $0 < p < 1$ ) 记为  $x_p$ , 它具有以下的性质:

- (1) 至少有  $np$  个观察值小于或等于  $x_p$ ;
- (2) 至少有  $n(1-p)$  个观察值大于或等于  $x_p$ .

样本  $p$  分位数可按以下法则求得. 将  $x_1, x_2, \dots, x_n$  按从小到大的顺序排列成  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ .



例:  $n=12, p=0.9, np=10.8, n-np=1.2$

1° 若 $np$ 不是整数，则只有一个数据满足定义中的两点要求，这一数据位于大于 $np$ 的最小整数处，即为位于 $[np]+1$ 处的数。

2° 若 $np$ 是整数，就取位于 $[np]$ 和 $[np]+1$ 处的中位数。

**综上，**

$$x_p = \begin{cases} x_{([np]+1)}, & \text{当 } np \text{ 不是整数,} \\ \frac{1}{2}[x_{(np)} + x_{(np+1)}], & \text{当 } np \text{ 是整数.} \end{cases}$$





特别, 当  $p = 0.5$  时, 0.5分位数  $x_{0.5}$  也记为  $Q_2$  或  $M$  称为样本中位数, 即有

$$x_{0.5} = \begin{cases} x_{(\lfloor \frac{n}{2} \rfloor + 1)}, & \text{当 } np \text{ 不是整数,} \\ \frac{1}{2} [x_{(\frac{n}{2})} + x_{(\frac{n}{2} + 1)}], & \text{当 } np \text{ 是整数.} \end{cases}$$

0.25分位数  $x_{0.25}$  称为第一四分位数, 又记为  $Q_1$ ;

0.75分位数  $x_{0.75}$  称为第三四分位数, 又记为  $Q_3$ .





**例2 设有一组容量为18的样本如下（已经排过序）**

**122 126 133 140 145 145 149 150 157**

**162 166 175 177 177 183 188 199 212**

**求样本分位数： $x_{0.2}$ ,  $x_{0.25}$ ,  $x_{0.5}$ .**

**解 (1) 因为  $np = 18 \times 0.2 = 3.6$ ,**

**$x_{0.2}$  位于第  $[3.6] + 1 = 4$  处, 即有  $x_{0.2} = x_{(4)} = 140$ .**

**(2) 因为  $np = 18 \times 0.25 = 4.5$ ,**

**$x_{0.25}$  位于第  $[4.5] + 1 = 5$  处, 即有  $x_{0.25} = 145$ .**

**(3) 因为  $np = 18 \times 0.5 = 9$ ,  $x_{0.5}$  是这组数中间两**



个数的平均值, 即有  $x_{0.5} = \frac{1}{2}(157 + 162) = 159.5$ .

**数据集的箱线图是由箱子和直线组成的图形, 它是基于以下五个数的图形概括: 最小值 Min, 第一四分位数  $Q_1$ , 中位数  $M$ , 第三四分位数  $Q_3$  和最大值 Max. 它的作法如下:**

(1) 画一水平数轴, 在轴上标上 Min,  $Q_1$ ,  $M$ ,  $Q_3$ , Max. 在数轴上方画一个上、下侧平行于数轴的矩形箱子, 箱子的左右两侧分别位于  $Q_1$ ,  $Q_3$  的上方.





在 $M$ 点的上方画一条垂直线段. 线段位于箱子内部.

(2) 自箱子左侧引一条水平线  $Min$ ; 在同一水平高度自箱子右侧引一条水平线直至最大值. 如图所示.

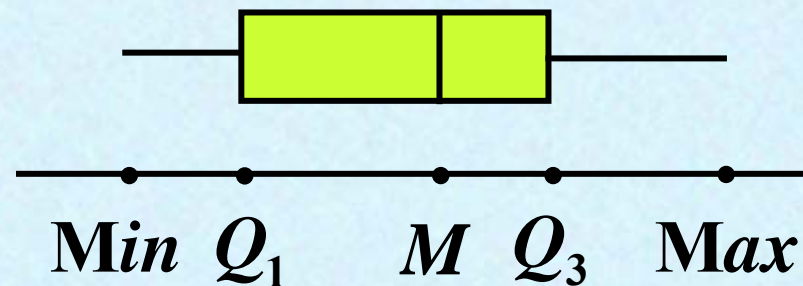


图 6-2



例3 以下是8个病人的血压（收缩压，mmHg）数据（已经过排序），试作出箱线图.

102 110 117 118 122 123 132 150

解 因为 $np = 8 \times 0.25 = 2$ ，故

$$Q_1 = \frac{1}{2}(110 + 117) = 113.5.$$

因为 $np = 8 \times 0.5 = 4$ ，故

$$x_{0.5} = Q_2 = \frac{1}{2}(118 + 122) = 120.$$

因为 $np = 8 \times 0.75 = 6$ ，故





$$x_{0.75} = Q_3 = \frac{1}{2}(123 + 132) = 127.5.$$

**Min = 102,    Max = 150,**

**作出箱线图如图所示.**

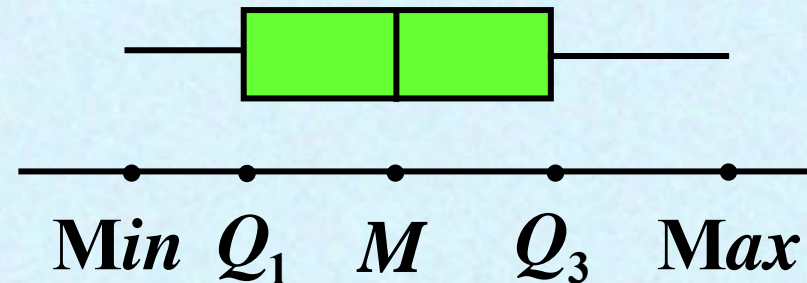


图 6-3



**例4** 下面分别给出了25个男子和25个女子的肺活量（以升计.数据应经过排序）

**女子组** 2.7 2.8 2.9 3.1 3.1 3.1 3.2 3.4 3.4  
3.4 3.4 3.4 3.5 3.5 3.5 3.6 3.7 3.7  
3.7 3.8 3.8 4.0 4.1 4.2 4.2

**男子组** 4.1 4.1 4.3 4.3 4.5 4.6 4.7 4.8 4.8  
5.1 5.3 5.3 5.3 5.4 5.4 5.5 5.6 5.7  
5.8 5.8 6.0 6.1 6.3 6.7 6.7

**试分别画出这两组数据的箱线图.**





解 **女子组**  $\text{Min} = 2.7$ ,  $\text{Max} = 4.2$ ,  $M = 3.5$ ,

因  $np = 25 \times 0.25 = 6.25$ ,  $Q_1 = 3.2$ .

因  $np = 25 \times 0.75 = 18.75$ ,  $Q_3 = 3.7$ .

**男子组**  $\text{Min} = 4.1$ ,  $\text{Max} = 6.7$ ,  $M = 5.3$ ,

因  $np = 25 \times 0.25 = 6.25$ ,  $Q_1 = 4.7$ .

因  $np = 25 \times 0.75 = 18.75$ ,  $Q_3 = 5.8$ .

作出箱线图如图所示.



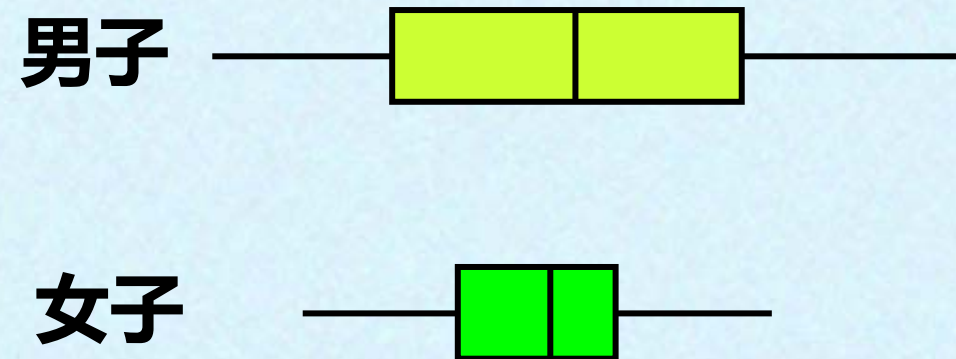


图 6-4



## 疑似异常值

在数据集中, 某一个观察值不寻常地大于或小于该数据集中的其他数据, 称为疑似异常值.

第一四分位数  $Q_1$  与第三四分位数  $Q_3$  之间的距离:

$$Q_3 - Q_1 = IQR$$

称为四分位数间距.

若数据小于  $Q_1 - 1.5IQR$  或大于  $Q_3 + 1.5IQR$ , 则认为它是疑似异常值 .



## 修正箱线图

(1') 同(1);

(2') 计算 $IQR = Q_3 - Q_1$ , 若一个数据小于 $Q_1 - 1.5IQR$  或大于 $Q_3 + 1.5IQR$ , 则认为它是一个疑似异常值. 画出疑似异常值, 并以\*表示;

(3') 自箱子左侧引一水平线段直至数据集中除去疑似异常值后的最小值, 又自箱子右侧引一水平线直至数据集中除去疑似异常值后的最大值.





例5 下面给出了某医院21个病人的住院时间（以天计），试画出修正箱线图（数据已经过排序）。

1 2 3 3 4 4 5 6 6 7 7 9 9

10 12 12 13 15 18 23 55

解  $\text{Min} = 1, \text{Max} = 55, M = 7,$

因  $21 \times 0.25 = 5.25$ , 得  $Q_1 = 4$ ,

又  $21 \times 0.75 = 15.75$ , 得  $Q_3 = 12$ ,

$IQR = Q_3 - Q_1 = 8, Q_3 + 1.5IQR = 12 + 1.5 \times 8 = 24,$

$Q_1 - 1.5IQR = 4 - 12 = -8.$



观察值  $55 > 24$ , 故 55 是疑似异常值, 且仅此一个疑似异常值.

作出修正箱线图如图所示.

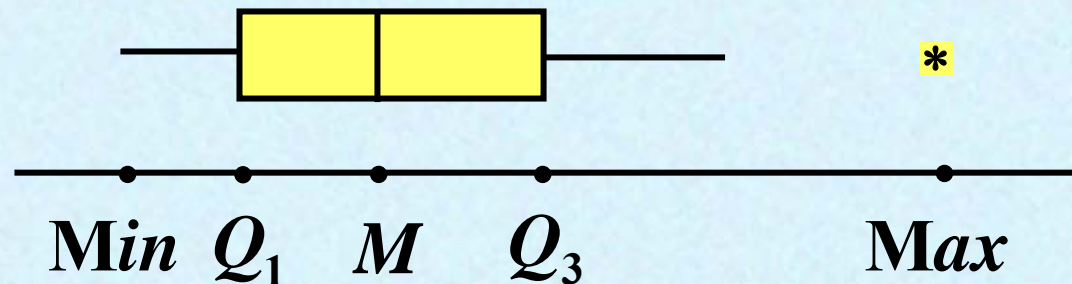


图 6-5





## 三、小结

### 1. 频率直方图作图步骤

- (1) 找出最小值和最大值;
- (2) 将选定区间分为 $k$ 个小区间;
- (3) 算出频率 $f_i / n$ . 在各个小区间上作以  $\frac{f_i}{n} / \Delta$  为高的小矩形.



## 2.箱线图作图步骤

(1) 画一水平数轴，在轴上标上  $\text{Min}$ ,  $Q_1$ ,  $M$ ,  $Q_3$ ,  $\text{Max}$ . 在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于  $Q_1$ ,  $Q_3$  的上方.

在  $M$  点的上方画一条垂直线段. 线段位于箱子内部.

(2) 自箱子左侧引一条水平线  $\text{Min}$ ; 在同一水平高度自箱子右侧引一条水平线直至最大值.



作业：10

