§1.5 样本数据的整理

一、经验分布函数

分布函数是研究随机变量的重要工具，总体可以用随机变量来表示，而样本反映了总体的信息. 因此，可利用样本$(X_1, X_2, \cdots, X_n)$去估计总体X的分布函数$F(x)$.

总体X的分布函数$F(x)$为

$$F(x) = P(X \le x).$$

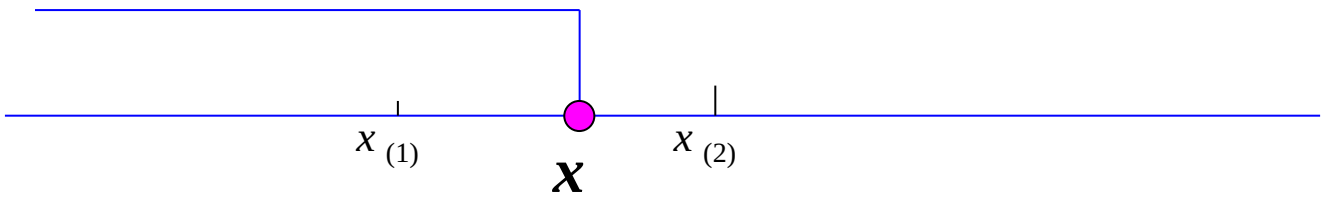
用事件$\{X \le x)$发生的频率作为其估计即可. 这就引出了下面所谓经验分布函数的概念.

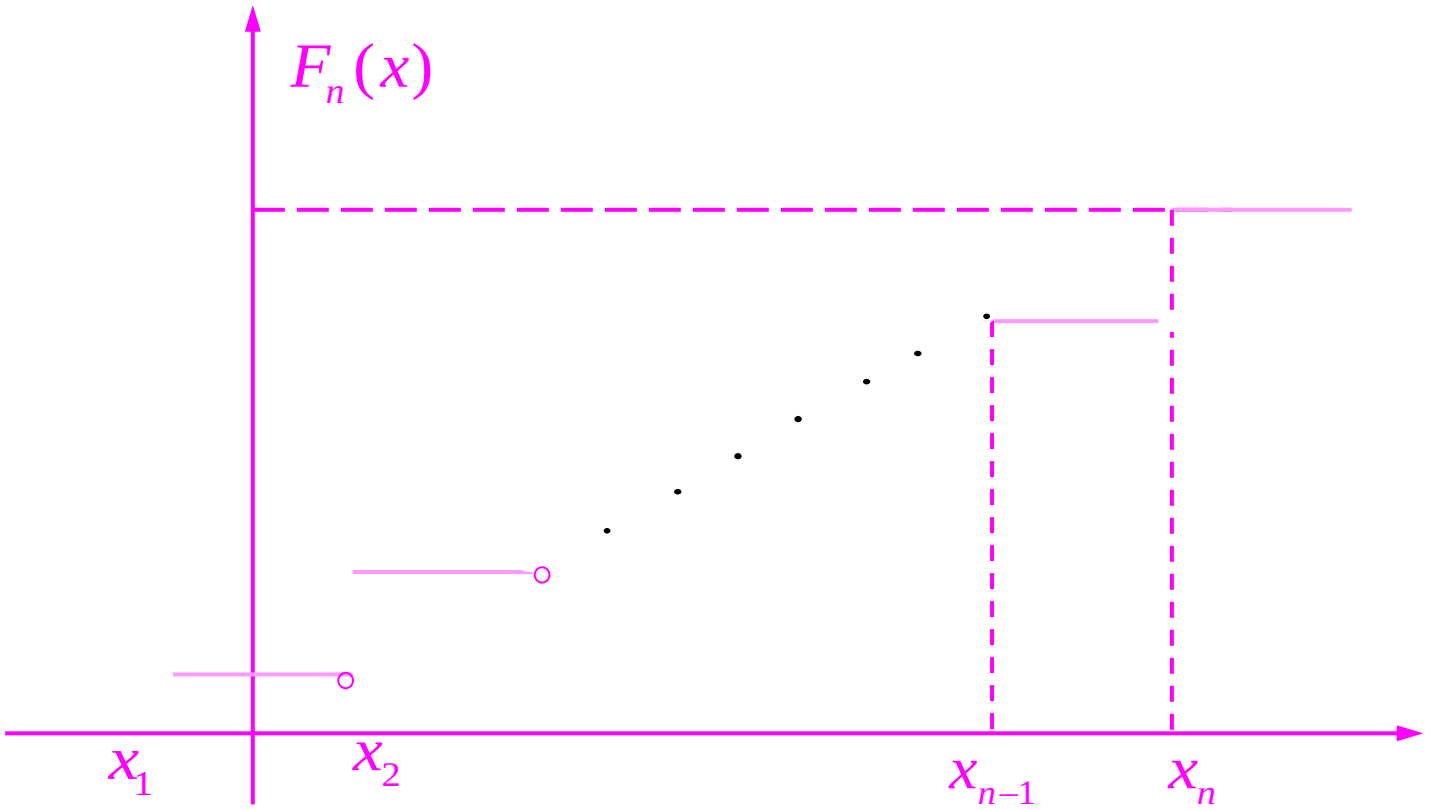
定义5.5.1 设$X_1, X_2, \cdots, X_n$是来自总体X的一个样本，观察值为$x_1, \cdots, x_n$，大小顺序排列为$x_{(1)} \le \cdots \le x_{(n)}$，则总体$X$的经验分布函数定义为

$$F_n(x)=\frac{\text{样本观测值中不超过 } x \text{ 的个数}}{n}, \quad \forall x \in R.$$

即

$$F_n(x)=\begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, & x_{(k)} \le x < x_{(k+1)}, \quad k=1,2,\cdots,n-1. \\ 1, & x \ge x_{(n)} \end{cases}$$





经验分布函数的示意图

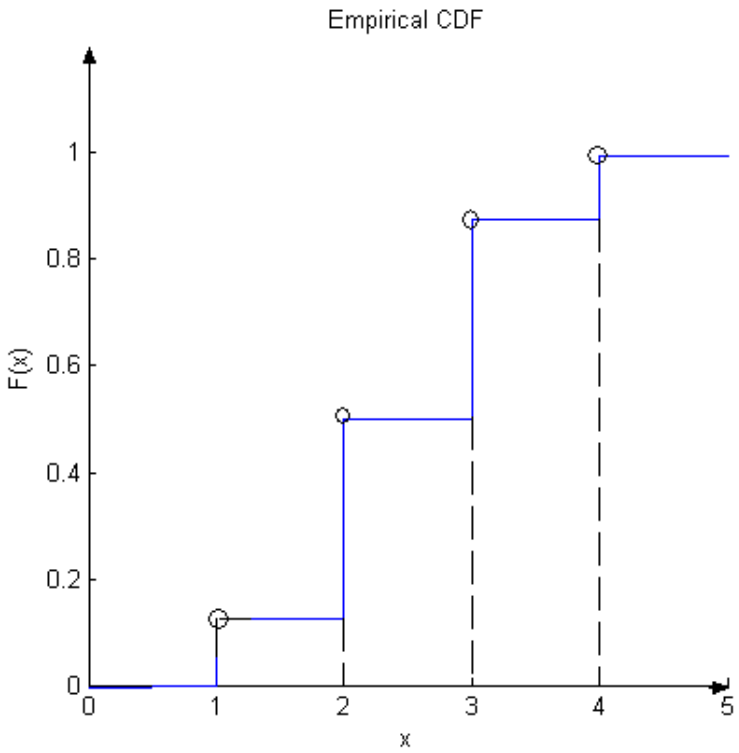
例1 总体X的8个样本观察值为

1, 2, 2, 2, 3, 3, 3, 4.

则经验分布函数为

$$F_8(x)=\begin{cases}0, & x<1;\\ 1/8, & 1\le x<2;\\ 4/8, & 2\le x<3;\\ 7/8, & 3\le x<4;\\ 1, & x\ge 4.\end{cases}$$

经验分布函数如右图所示



经验分布函数是样本的函数，是统计量(也是随机变量). 根据定义：$F_n(x)$是事件$\{X \le x\}$发生的频率. 因此对任意给定的实数 x，有

$$nF_n(x) \sim B(n, F(x)).$$

由伯努里大数定律知：当$n \longrightarrow \infty$时

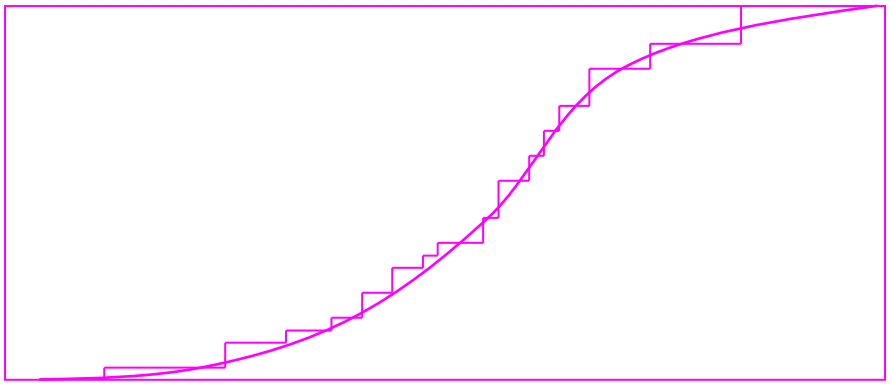
$$F_n(x) \xrightarrow{P} F(x).$$

更深刻的结果是下面的格里纹科定理.

定理5.5.1 （格里纹科定理）设$x_1, x_2, \cdots, x_n$是总体分布函数为$F(x)$的样本，$F_n(x)$为其经验分布函数，当$n \to +\infty$时,有

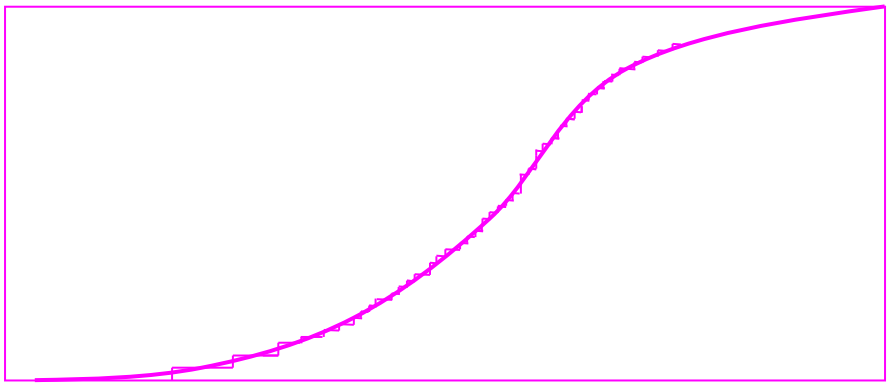
$$P\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |F_n(x) - F(x)| = 0\} = 1.$$

格里纹科定理表明：当n相当大时，经验分布函数是总体分布函数$F(x)$的一个良好的近似. 经典的统计学中一切统计推断都以样本为依据，其理由就在于此.



(a)

20



(b)

200

二、频数—频率分布表

样本数据的整理是统计研究的基础，整理数据的最常用方法之一是给出其频数分布表或频率分布表.

例2 为研究某厂工人生产某种产品的能力，我们随机调查了20位工人某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：

（1）对样本进行分组：作为一般性的原则，组数通常在5 ~ 20个，

数据个数n	分组数k
≤ 50	$5 \sim 7$
$50 \sim 100$	$6 \sim 10$
$100 \sim 250$	$7 \sim 12$
250以上	$10 \sim 20$

（2）确定每组组距：近似公式为

组距d = (最大观测值 – 最小观测值) / 组数.

（3）确定每组组限：各组区间端点为

$$a_0, a_1 = a_0 + d, a_2 = a_0 + 2d, \cdots, a_k = a_0 + kd.$$

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.

（4）统计样本数据落入每个区间的个数—频数，并列出其频数频率分布表.

例2的频数—频率分布表

组序	分组区间	组中值	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	

三、直方图

直方图是频数分布的图形表示，它的横坐标表示所关心变量的取值区间，纵坐标有三种表示方法：频数，频率，最准确的是频率 / 组距，它可使得诸长条矩形面积和为1. 凡此三种直方图的差别仅在于纵轴刻度的选择，直方图本身并无变化..

例3 某钢铁加工厂生产内径为25.40*mm*的钢管，为了掌握产品的生产状况，需定期对产品进行检测，下面的数据是一次抽样中的100件钢管的内径尺寸：试画出钢管内径的直方图.

25.39	25.36	25.34	25.42	25.45	25.38	25.39	25.42	25.47	25.35
25.40	25.43	25.44	25.48	25.45	25.43	25.46	25.40	25.51	25.45
25.41	25.39	25.41	25.36	25.38	25.31	25.56	25.43	25.40	25.38
25.37	25.44	25.33	25.46	25.40	25.49	25.34	25.42	25.50	25.37
25.35	25.32	25.45	25.40	25.27	25.43	25.54	25.39	25.45	25.43
25.40	25.43	25.44	25.41	25.53	25.37	25.38	25.24	25.44	25.40
25.36	25.42	25.39	25.46	25.38	25.35	25.31	25.34	25.40	25.36
25.41	25.32	25.38	25.42	25.40	25.33	25.37	25.41	25.49	25.35
25.47	25.34	25.30	25.39	25.36	25.46	25.29	25.40	25.37	25.33
25.40	25.35	25.41	25.37	25.47	25.39	25.42	25.47	25.38	25.39

最大值

最小值

（1）求极差： $R_n = x_{(n)} - x_{(1)} = 25.56 - 25.24 = 0.32.$

（2）决定组数与组距：$k = 11$，$d = 0.03.$

（3）决定分点，将数据分组.

（4）登记频数，计算频率，列出频率分布表.

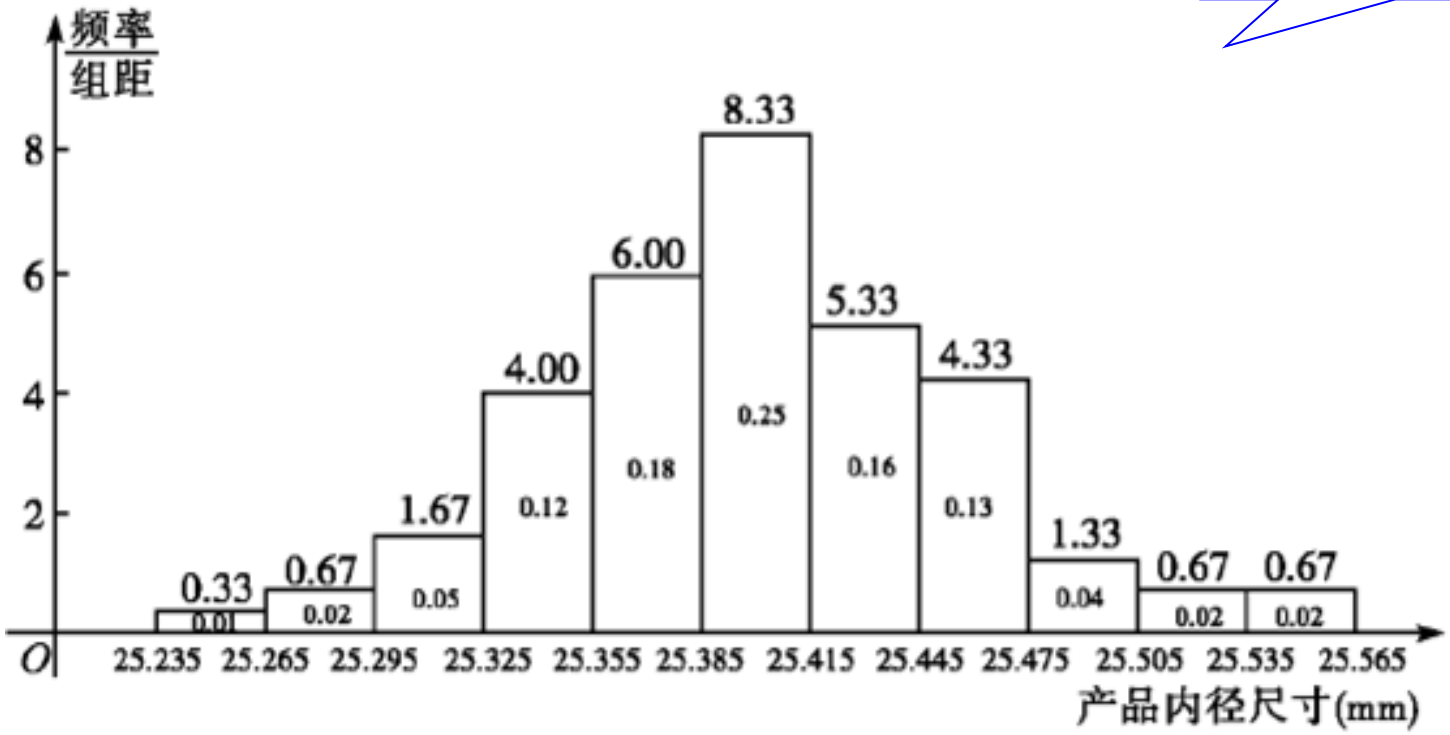
（5）绘制频率分布直方图.

频率分布表

分组	个数累计	频数	频率
25.235～25.265	一	1	0.01
25.265～25.295	下	2	0.02
25.295～25.325	正	5	0.05
25.325～25.355	正正下	12	0.12
25.355～25.385	正正正下	18	0.18
25.385～25.415	正正正正正	25	0.25
25.415～25.445	正正正一	16	0.16
25.445～25.475	正正下	13	0.13
25.475～25.505	下	4	0.04
25.505～25.535	下	2	0.02
25.535～25.565	下	2	0.02
合计	100	100	1.00

频率分布直方图

注意 纵坐标是
频率/组距



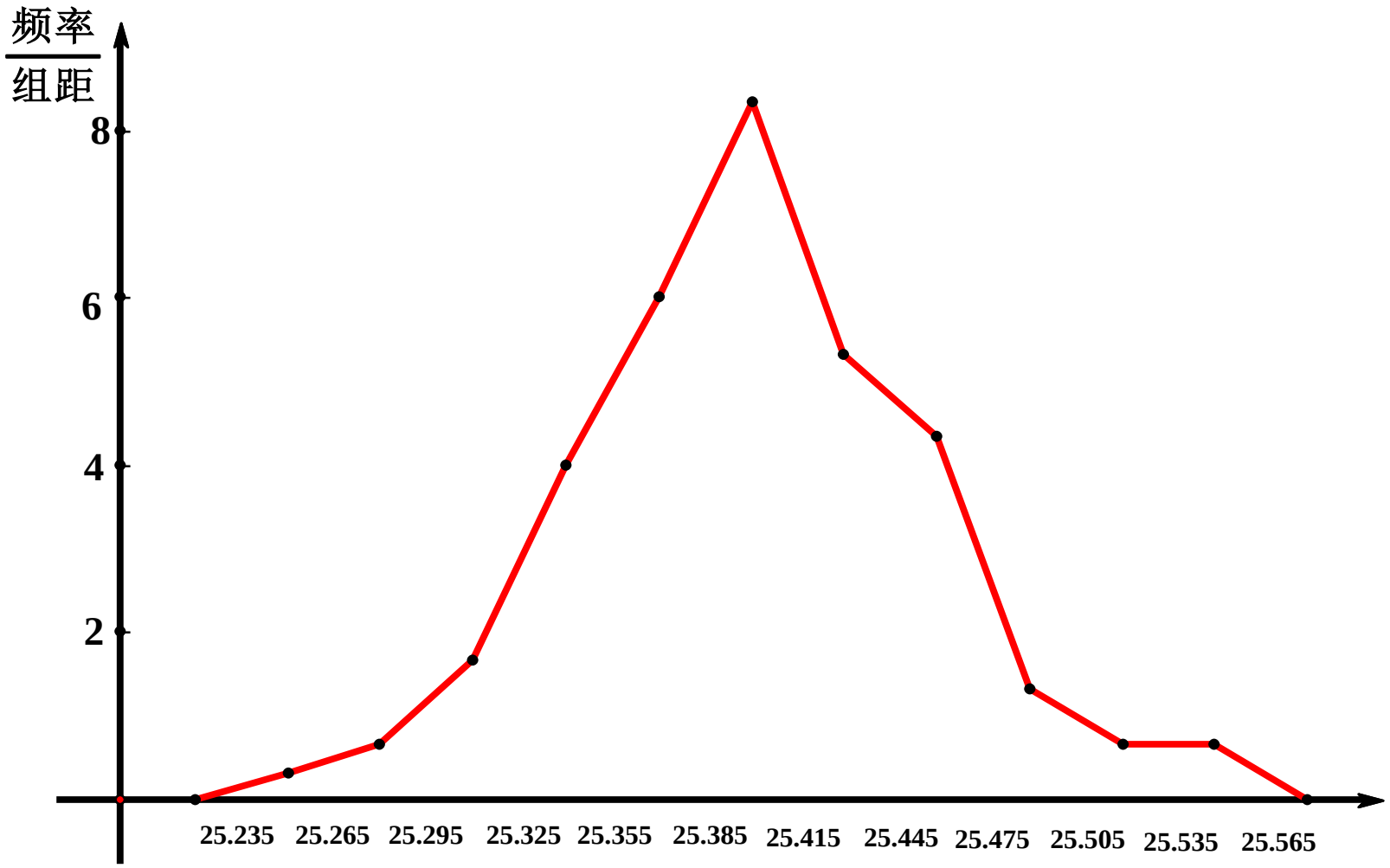
$$\text{小长方形面积}=\text{组距}\times\frac{\text{频率}}{\text{组距}}=\text{频率}$$

从频率分布直方图可以清楚的看出数据分布的总体态势，但是数据的实际值体现不出来了.

频率分布折线图：　把频率分布直方图各个长方形上边的中点用线段连接起来，就得到频率分布折线图.

为了方便看图，一般习惯于把频率分布折线图画成与横轴相连，所以横轴上的左右两端点没有实际的意义. 例如，前面的钢管内径的频率分布折线图，如下图所示.

频率分布折线图


频率
组距
8
6
4
2
25.235
25.265
25.295
25.325
25.355
25.385
25.415
25.445
25.475
25.505
25.535
25.565

频率分布折线图的性质：

如果样本容量无限增大，分组的组距无限减小，那么频率分布折线图就会无限接近于总体密度曲线.

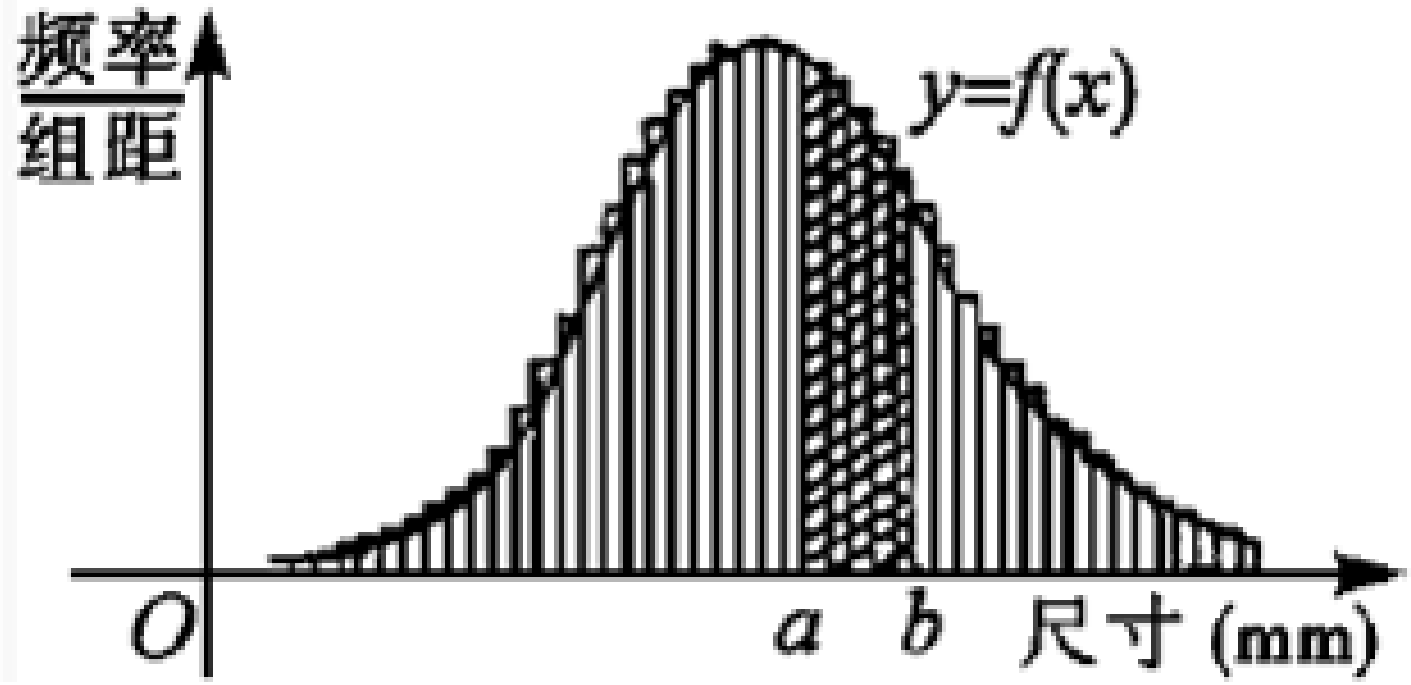


图 2

四、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，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. 如

数值	分开	茎	和	叶
112	⟶ 11 | 2	⟶ 11	和	2

例4 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(已经过排序)：

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，见下页。

测试成绩的茎叶图

6	4 7
7	0 2 4 6 6 9
8	0 1 2 2 3 5 6 8
9	1 1 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

注意： 茎叶图保留数据中全部信息. 当样本量较大，数据很分散，横跨二、三个数量级时，茎叶图并不适用.

在比较两组样本时，可画出它们的背靠背的茎叶图.

甲车间		乙车间
	5	6
8 7 7 7 5 5 5 4 2 1 1	6	6 7 7 8 8
8 7 7 6 6 4 4 2 1	7	2 2 4 5 5 5 5 6 6 6 8 8 9
8 7 6 6 5 3 2	8	0 1 1 3 3 3 4 4 4 6 6 7 7 8
7 3 2 1 0	9	0 2 3 5 8
5 3 0 0	10	7

叶　　茎　　叶

五、五数概括与箱线图

顺序统计量的应用之一就是五数概括与箱线图. 在得到有序样本后，容易计算如下五个值：

最小观测值$x_{min} = x_{(1)}$；　最大观测值$x_{max} = x_{(n)}$；

中位数$m_{0.5}$；　第一4分位数$Q_1 = m_{0.25}$；

第三4分位数$Q_3 = m_{0.75}$.

所谓五数概括就是指用这五个数

$$x_{min},\quad Q_1,\quad m_{0.5},\quad Q_3,\quad x_{max}$$

来大致描述一批数据的轮廓.

下面就通过一个具体的实例说明之.

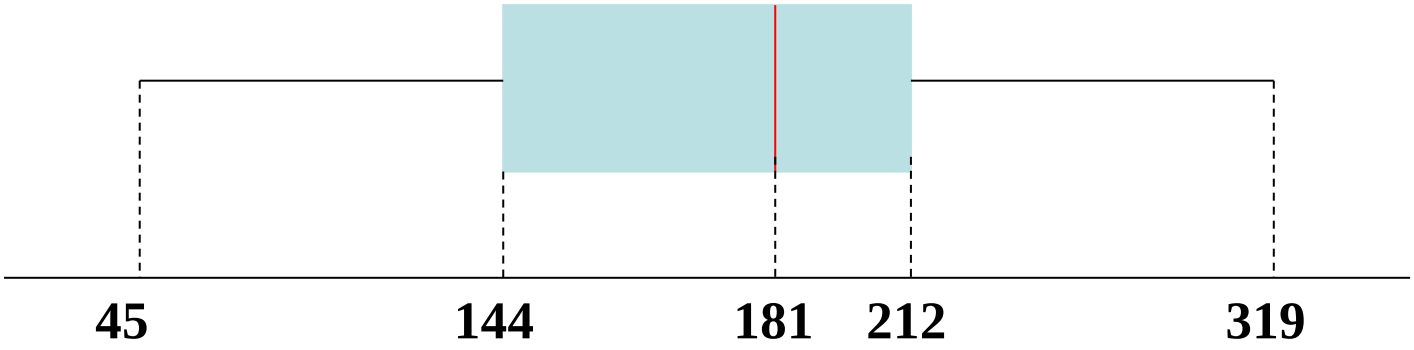
例5 下表是某厂160名销售人员某月的销售量数据的有序样本，由该批数据可计算得到：

$$x_{\min}=45,\quad x_{\max}=319,\quad m_{0.5}=181,$$
$$Q_1=144,\quad Q_3=212.$$

五数概括的图形表示称为箱线图，由箱子和线段组成. 其作法如下

(1) 画一个箱子，其两侧恰为第一4分位数和第三4分位数，在中位数位置上画一条竖线，它在箱子内，这个箱子包含了样本中50%的数据；

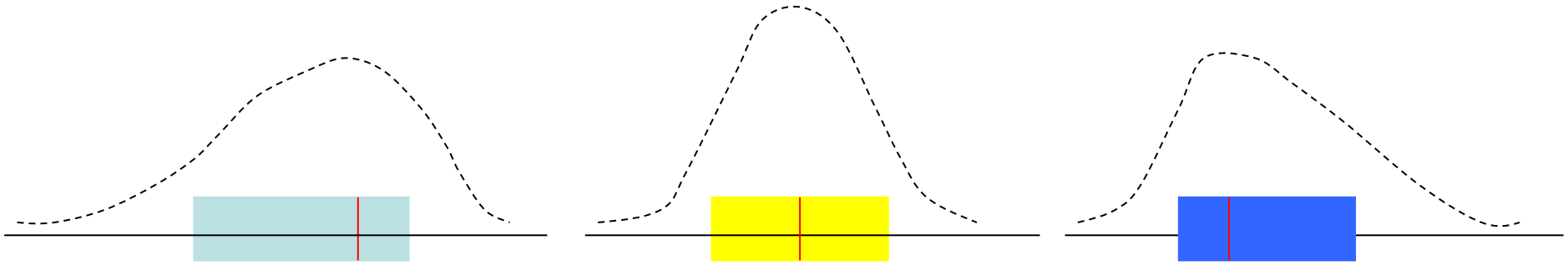
45	74	76	80	87	91	92	93	95	96
98	99	104	106	111	113	117	120	122	122
124	126	127	127	129	129	130	131	131	133
134	134	135	136	137	137	139	141	141	143
145	148	149	149	149	150	150	153	153	153
153	154	157	160	160	162	163	163	165	165
167	167	168	170	171	172	173	174	175	175
176	178	178	178	179	179	179	180	181	181
188	189	189	191	191	191	192	192	194	194
194	194	195	196	197	197	198	198	198	199
200	201	202	204	204	205	205	206	207	210
214	214	215	215	216	217	218	219	219	221
221	221	221	221	222	223	223	224	227	227
228	229	232	234	234	238	240	242	242	242
244	246	253	253	255	258	282	290	314	319



月销售量数据的箱线图

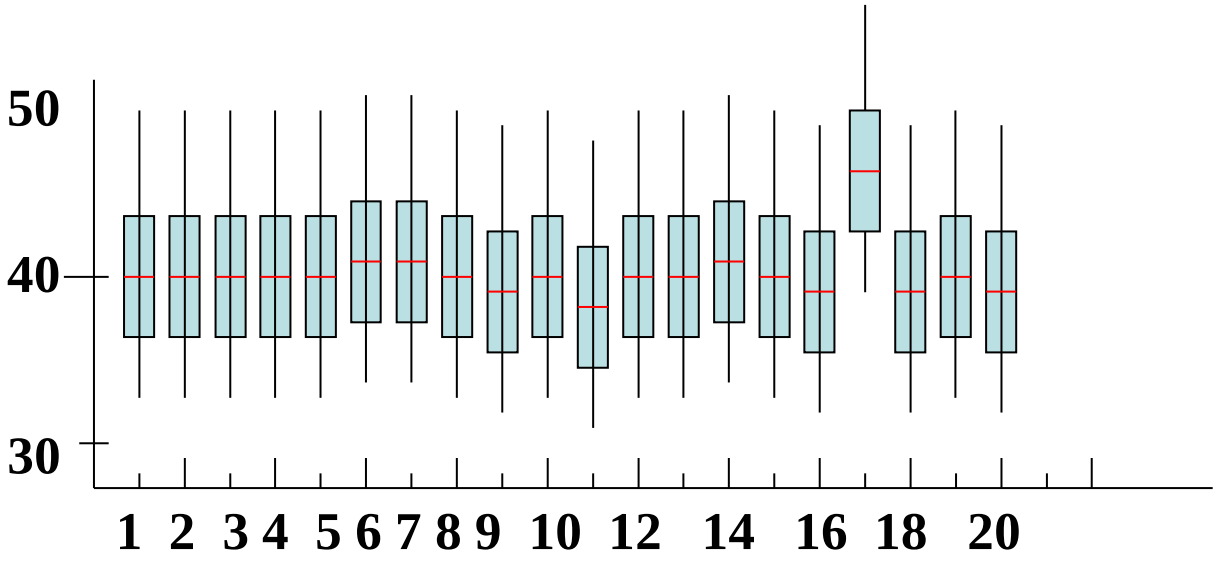
（2）在箱子左右两侧各引出一条水平线，分别至最小值和最大值为止，每条线段包含了样本中25%的数据.

箱线图可用来对数据分布的形状进行大致的判断. 下图给出三种常见的箱线图，分别对应对称分布、左偏分布和右偏分布.



三种常见的箱线图及其对应的分布轮廓

如果我们要对几批数据进行比较，则可以在一张纸上同时画出这几批数据的箱线图. 下图是某厂20天生产的某种产品的直径数据画成的箱线图，从图中可以清楚地看出，第17天的产品出现了异常.


50
40
30
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20