数理统计：对数据进行收集、整理、分析与推断

收集数据:

从总体X中抽取样本：$X_1, X_2, \ldots, X_n$

$X_1, X_2, \ldots, X_n$

"杂乱无章"的数据

包含了有用的"信息"

Q: 如何提炼出有用的信息？

第二节　直方图与样本分布函数

一、　直方图

二、　样本分布函数

一、 直方图

设总体X中抽取到样本观测值$x_1, x_2, \cdots, x_n$,则做直方图的一般步骤如下：

（1）找出$x_1, x_2, \cdots, x_n$中的最小值$x_{(1)}$和最大值$x_{(n)}$。选取略小于$x_{(1)}$的数a和略大于$x_{(n)}$的数b。

（2）根据样本容量确定组数k，如果样本容量小，则组数少些。如果样本容量大，则组数多些。一般来说，组数k取为8～16.

表　　样本含量与组数

样本含量（n）	组数
60-100	7-10
100-200	9-12
200-500	12-17
500以上	17-30

（3）选取分点

$$a=t_0<t_1<\cdots<t_{i-1}<t_i<\cdots<t_k=b.$$

把区间(a,b)分为k个子区间

$$(a,t_1],(t_1,t_2],\cdots,(t_{i-1},t_i],\cdots,(t_{k-1},b)$$

第 i 个子区间$(t_{i-1},t_i]$的长度为

$$\Delta t_i=t_i-t_{i-1},\quad i=1,2,\cdots,k$$

若取各子区间长度相等，则有

$$\Delta t_i=\frac{b-a}{k},\quad i=1,2\cdots,k.$$

记$\Delta t=\dfrac{b-a}{k}$.称 Δt 为组距。此时分点

$$t_i=a+i\Delta t,\quad i=1,2,\cdots,k$$

注：分点 t_i 应比样本观测值x_i多取一位有效数字。

（4）数出$x_1, x_2, \cdots, x_n$落在每个子区间$(t_{i-1}, t_i]$内的频数n_i，再算出频率

$$f_i = \frac{n_i}{n}, \quad i = 1, 2, \cdots, k.$$

（5）在Ox轴上画出各个分点$t_i\left(i = 0, 1, 2, \cdots, k\right)$，并以各子区间$(t_{i-1}, t_i]$为底，以$y_i = \dfrac{f_i}{\Delta t_i}$为高做小矩形，这样做出的所有小矩形构成了直方图。

直方图作用

第 $i\left(i=1,2,\cdots,k\right)$ 个小矩形的面积等于样本观测值落在该子区间内的频率。

所有小矩形的面积之和等于1.

当样本容量 n 充分大时，随机变量 X 落在第 i 个小区间 $(t_{i-1},t_i]$ 内的频率近似等于其概率，即

$$f_i \approx P\left\{t_{i-1} < X \leq t_i\right\}, i=1,2,\cdots,k,$$

所以直方图可以大致反映随机变量的概率分布。

例6.2.1 某门课程有120人参加考试，考试成绩 X 如下：

86	83	77	81	81	80	79	82	82	81
83	65	64	78	75	82	80	80	77	81
81	87	82	78	80	81	87	81	77	78
77	78	77	77	77	71	95	78	81	79
80	77	76	82	80	82	84	79	90	82
79	82	79	86	76	78	83	75	82	78
73	83	81	81	83	89	81	86	82	82
78	84	84	84	81	81	74	78	78	80
74	78	75	79	85	75	74	71	88	82
76	85	73	78	81	79	77	78	81	87
75	83	90	80	85	81	77	78	82	84
85	84	82	85	84	82	85	84	78	78

试根据这些数据作出直方图，并根据直方图估计X 的分布.

解 从n=120个数据中找出

最小值$x_{(1)}=64$及最大值$x_{(120)}=95$.

取 $a=63.5$, $b=95.5$, 分 $k=16$ 组，组距

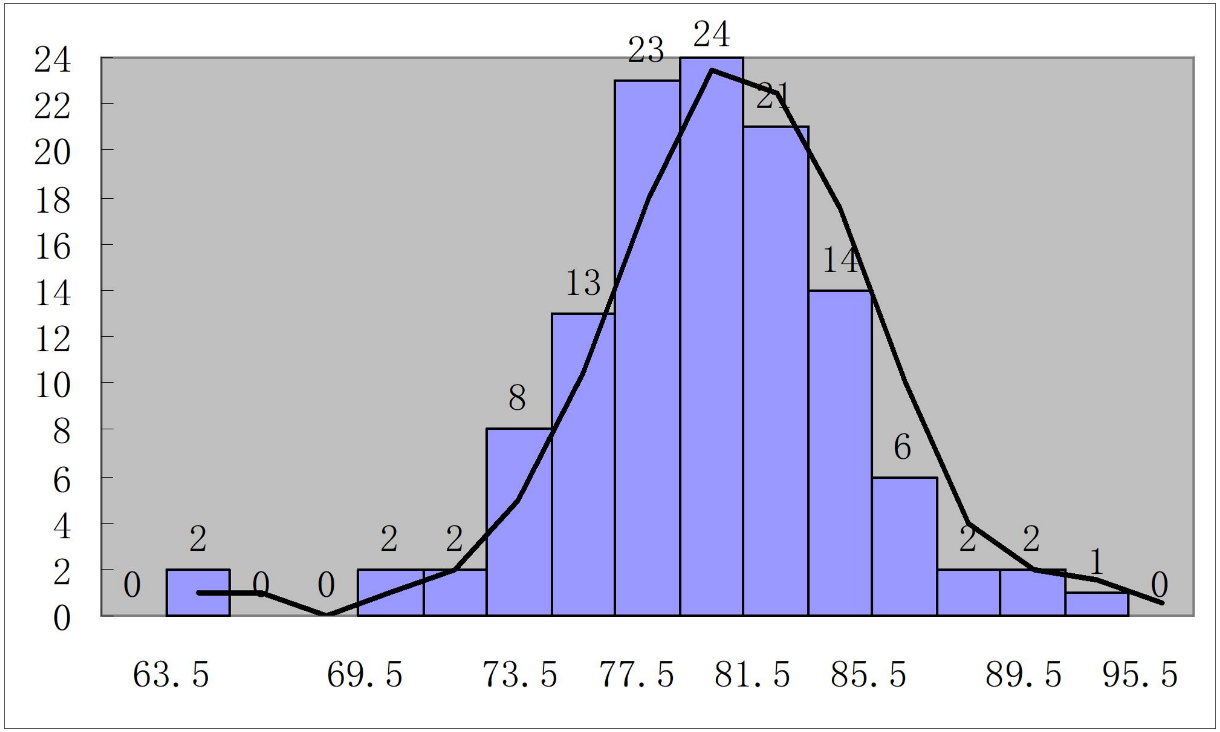
$$\Delta t=\frac{95.5-63.5}{16}=2$$

分组$(t_{i-1},t_i]$	频数
63.5~65.5	2
65.5~67.5	0
67.5~69.5	0
69.5~71.5	2
71.5~73.5	2
73.5~75.5	8
75.5~77.5	13
77.5~79.5	23

分组$(t_{i-1},t_i]$	频数
79.5~81.5	24
81.5~83.5	21
83.5~85.5	14
85.5~87.5	6
87.5~89.5	2
89.5~91.5	2
91.5~93.5	0
93.5~95.5	1

以横轴 x 轴表示成绩，$a = t_0 = 63.5$， $t_1 = 65.5, \ldots, t_{15} = 93.5$，$b = t_{16} = 95.5$, $\Delta t = 2$， 在$(t_{i-1}, t_i]$上，做高为

$$y_i = \frac{f_i}{\Delta t} = \frac{n_i}{n} \cdot \frac{1}{\Delta t} = \frac{n_i}{240}$$ 的矩形。



二、 样本分布函数

样本能够反映总体X的信息, 总体X的分布函数$F(x)$是否能由样本来“表示”？ 回答是 肯定的.

定义 设总体X的分布函数为$F(x)$, 从总体X中抽取容量为n的样本, 样本观测值为$x_1, x_2, \cdots, x_n$, 如果样本容量n较大, 则相同的观测值可能重复出现若干次. 假如在n个样本观测值$x_1, x_2, \cdots, x_n$中有k个不同的值, 按由小到大的顺序依次记为$x_{(1)} < x_{(1)} < \cdots < x_{(k)}$, $k \le n$, 并假设各个$x_{(i)}$出现的频数为n_i, 则各个$x_{(i)}$出现的频率为

$\mathrm{f}_i = \dfrac{n_i}{n}$, $i=1,2,\cdots,k$, $k \le n$, 显然有$\sum\limits_{i=1}^{k} n_i = n$, $\sum\limits_{i=1}^{k} f_i = 1$.

称$F_n(x)=\begin{cases} 0, & x < x_{(1)}; \\ \sum_{j=1}^{i} f_j, & x_{(i)} \le x < x_{(i+1)}, \quad i=1,2,\cdots,k-1; \\ 1, & x \ge x_{(k)}. \end{cases}$

为总体X的样本分布函数.

注：对于样本观察值 x_1，x_2，…，x_n, 为了求其对应的样本分布函数$F_n(x)$之值，只须将这 n 个值中小于或等于 x 的个数除以样本容量 n 即可.

样本分布函数$F_n(x)$具有以下性质：

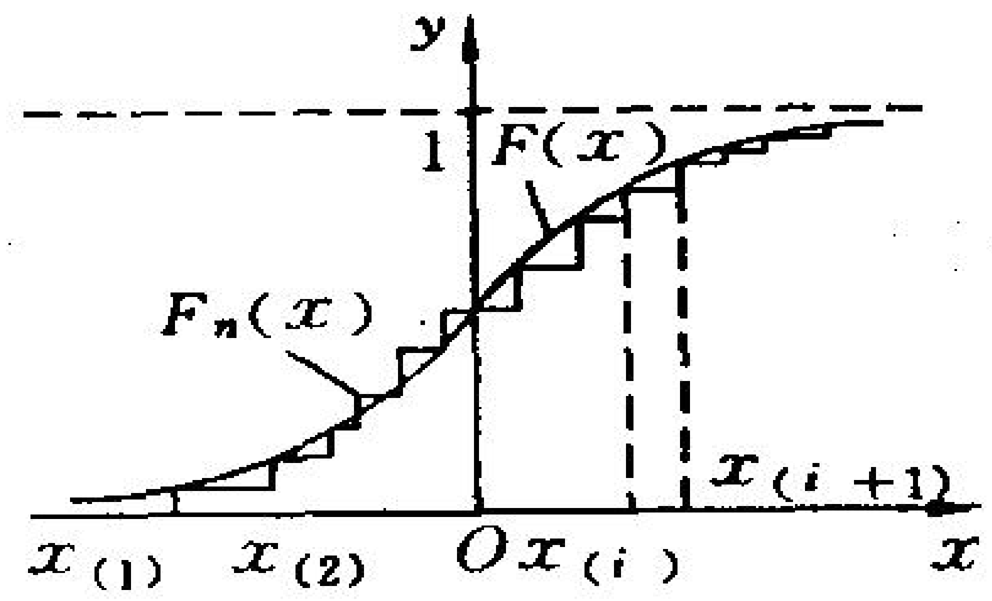
(1) $0 \le F_n(x) \le 1$；

(2) $F_n(x)$是单调不减函数；

(3) $F_n(-\infty)=0$, $F_n(+\infty)=1$,

(4) $F_n(x)$是处处右连续的.

样本分布函数$F_n(x)$不仅与样本容量 n 有关，还与所得到的样本观察值有关，$F_n(x)$的图形呈跳跃上升的阶梯状，图中的曲线是总体 X 的理论分布函数$F(x)$的图形.



例6.2.2 试根据总体X的下列两组样本容量为10的样本

观测值:

Ⅰ组	观测值	1	2	3	4	5
	频 数	2	3	1	3	1
Ⅱ组	观测值	1	2	3	4	5
	频 数	1	2	2	3	2

分别求出样本分布函数 $F_{10}(x)$, 并求出 $F_{10}(3.5)$.

解 (1) $n=10$, 计算得频率和样本分布函数分别为

$$f_1=\frac{2}{10},f_2=\frac{3}{10},f_3=\frac{1}{10},f_4=\frac{3}{10},f_5=\frac{1}{10},$$

$$F_{10}(x)=\begin{cases}0, & x<1,\\ 0.2, & 1\le x<2,\\ 0.5, & 2\le x<3,\\ 0.6, & 3\le x<4,\\ 0.9, & 4\le x<5,\\ 1, & x\ge 5.\end{cases}$$

从而有$F_{10}(3.5)=0.6.$

(2) n**=10,** 计算得频率和样本分布函数分别为

$$f_1=\frac{1}{10},f_2=\frac{2}{10},f_3=\frac{2}{10},f_4=\frac{3}{10},f_5=\frac{2}{10},$$

$$F_{10}(x)=\begin{cases}0, & x<1,\\ 0.1, & 1\le x<2,\\ 0.3, & 2\le x<3,\\ 0.5, & 3\le x<4,\\ 0.8, & 4\le x<5,\\ 1, & x\ge 5.\end{cases}$$

从而有F_{10}**(3.5)=0.5.**

对于给定的x，$F_n(x)$是 n 次重复独立试验中事件$\{X\leq x\}$出现的频率，而理论分布函数$F(x)$是事件$\{X\leq x\}$发生的概率，

由伯努利定理（大数定律）知，对任意给定的正数ε，有

$$\lim_{n\to\infty} P\{|F_n(x)-F(x)|<\varepsilon\}=1$$

即$F_n(x)$依概率收敛于$F(x)$.

定理*　格利文科(W.Glivenko)定理

设总体X的分布函数为$F(x)$，样本分布函数为$F_n(x)$，则对于任何实数x, 当n$\to\infty$时,有$F_n(x)$依概率1关于x一致收敛于$F(x)$, 即

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

这一结论是数理统计中依据样本来推断总体特征的理论基础.