§5.1 母体与子样

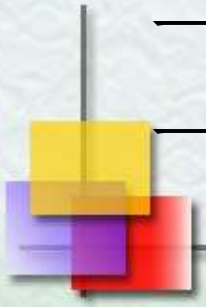
一、母体（总体）和个体

总体（population）—— 研究对象的全体

个体 —— 每一个对象

例如　研究某企业生产的一批电视机显象管的平均使用寿命，那么这一批显象管的全体就组成一个总体，其中每一只显象管就是一个个体。

例如　研究某大学一年级学生的身高情况，这时一年级大学生的全体就是总体；每个大学生就是一个个体。



在实际中我们真正所关心的是总体的某种数量指标，例如显象管的寿命指标X，学生的身高指标Y，它们都是r.v.。称这样的r. v. 为表征总体的随机变量。

为了方便起见，我们就将表征总体的随机变量定义为总体。

总体 $\longleftrightarrow$ r.v.X（Y, ξ,η ）

若X的分布函数为F(x)，则称总体的分布函数为F(x) 。

有限总体和无限总体

（按总体所含个体的数量分）

当有限总体包含的个体的总数很大时, 可近似地将它看成是无限总体.

K维总体： 所研究的数量指标有K个

对总体进行研究时，对总体中每个个体逐一进行考察，这在实际中往往是行不通的，一是试验具有破坏性，二是需花费大量的人力物力。

常用的方法是：从总体中随机地抽取若干个个体，根据对这部分个体的研究结果推断总体某方面的特征。

二、样本（sample）

定义 从总体X中随机地抽取n个个体，称之为总体X的一个**样本容量**为n的**样本（子样）**，样本中的个体称为**样品**。

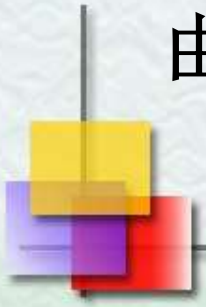
从总体中抽取若干个个体的过程称为**抽样**

假设抽样满足下述两个条件：

（1）**随机性** 为了使样本具有充分的代表性，抽样必须是随机的，应使总体中的每一个个体都有同等的机会被抽取到 。

（2）**独立性** 各次抽样必须是相互独立的，即每次抽样的结果既不影响其它各次抽样的结果，也不受其它各次抽样结果的影响。

这种随机的、独立的抽样方法称为**简单随机抽样**，由此得到的样本称为**简单随机样本**.



讨论：有放回和无放回抽样是否都是简单随机抽样？

- **有放回抽样：是**
- **无放回抽样：否，不满足独立性**

例5.1 P235(157)

第二次抽样的结果依赖于第一次抽样的结果

若第一次抽到不合格品，则第二次抽到不合格品的概率为

$$P(\xi_2 = 1 \mid \xi_1 = 1) = \frac{Np - 1}{N - 1}$$

若第一次抽到合格品，则第二次抽到不合格品的概率为

$$P(\xi_2 = 1 \mid \xi_1 = 0) = \frac{Np}{N - 1}$$

可以看到上述两种情形的概率都近似地等于p. 所以在N很大，n（样本容量）不大(一个经验法则是n/N≤0.1)时，可以把所得的子样近似地看成一个简单随机子样.

本课程中凡是提到抽样与样本，都是指简单随机抽样与简单随机样本。

例如　总体X是一批显象管的使用寿命，现从总体X中抽取n个显象管，X_i表示抽到的第i个显象管的使用寿命，i=1, 2, …,n；由于抽取的随机性，显然，每一个X_i都是随机变量，并且有着和总体X相同的分布。另外，由于抽取的独立性，

$X_1, X_2, \cdots, X_n$ 相互独立。

记$(X_1, X_2, \cdots, X_n)$为总体X的一个样本容量为n的样本。则 $X_1, X_2, \cdots, X_n$ 独立同分布（与总体X同分布）。

样本具有所谓的二重性:

一方面,由于样本是从总体中随机抽取的,抽取前无法预知它们的数值,因此,样本是随机变量,用大写字母$X_1, X_2, \cdots, X_n$表示;另一方面, 样本在抽取以后经观测就有确定的观测值，因此，样本又是一组数值。此时用小写字母$x_1, x_2, \cdots, x_n$表示。

例如　从某厂生产的显象管中随机抽取10个显象管，测得寿命如下（单位千小时）：

4.8，3.4，5.2，4.7，5.5，4.2，4.5，3.9，5.0， 4.9

这十个数据就是样本容量为10的样本 $X_1, X_2, \cdots, X_{10}$ 的一组观测值 $x_1, x_2, \cdots, x_{10}$ 。

若将样本$X_1, X_2, \cdots, X_n$看作是一n维随机变量$(X_1, X_2, \cdots, X_n)$,则

(1)当总体X是离散型随机变量, 则样本$(X_1, X_2, \cdots, X_n)$的联合分布律为:

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$
$$= P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n)$$

(2)当总体X是连续型随机变量,且具有概率密度函数$f(x)$时 ,则样本$(X_1, X_2, \cdots, X_n)$的联合概率密度为

$$f(x_1, x_2, \cdots, x_n) = f(x_1) f(x_2) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

例1 设总体 X 服从参数为 $\lambda(\lambda>0)$ 的指数分布，$(X_1,X_2,\cdots,X_n)$ 是来自总体的样本，求样本 $(X_1,X_2,\cdots,X_n)$ 的概率密度.

解 总体 X 的概率密度为 $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0,\\ 0, & x\le 0,\end{cases}$

所以 $(X_1,X_2,\cdots,X_n)$ 的概率密度为

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)$$

$$=\begin{cases}\prod_{i=1}^{n}\lambda e^{-\lambda x_i}, & x_i>0,\\ 0, & \text{其他.}\end{cases}\quad=\begin{cases}\lambda^n e^{-\lambda\sum_{i=1}^{n}x_i}, & x_i>0,\\ 0, & \text{其他.}\end{cases}$$

例2 设总体 X 服从两点分布 $B(1,p)$, 其中 $0<p<1$, $(X_1, X_2, \cdots, X_n)$ 是来自总体的样本, 求样本 $(X_1, X_2, \cdots, X_n)$ 的分布律.

解 总体 X 的分布律为 $P\{X=x\}=p^x(1-p)^{1-x}$ $(x=0,1)$

所以 $(X_1, X_2, \cdots, X_n)$ 的分布律为

$$P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

其中 $x_1, x_2, \cdots, x_n$ 在集合 $\{0,1\}$ 中取值.

二、经验分布函数（样本分布函数，Empirical cdf）

设 $X_1, X_2, \cdots, X_n$ 是总体 F 的一个样本，

用 $S(x)(-\infty < x < +\infty)$ 表示 $X_1, X_2, \cdots, X_n$ 中不大于 x 的随机变量的个数，

定义经验分布函数 $F_n(x)$ 为

$$F_n(x) = \frac{1}{n} S(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x), \quad -\infty < x < +\infty$$

一般地，

设 $x_1, x_2, \cdots, x_n$ 是总体 F 的一个容量为 n 样本值，

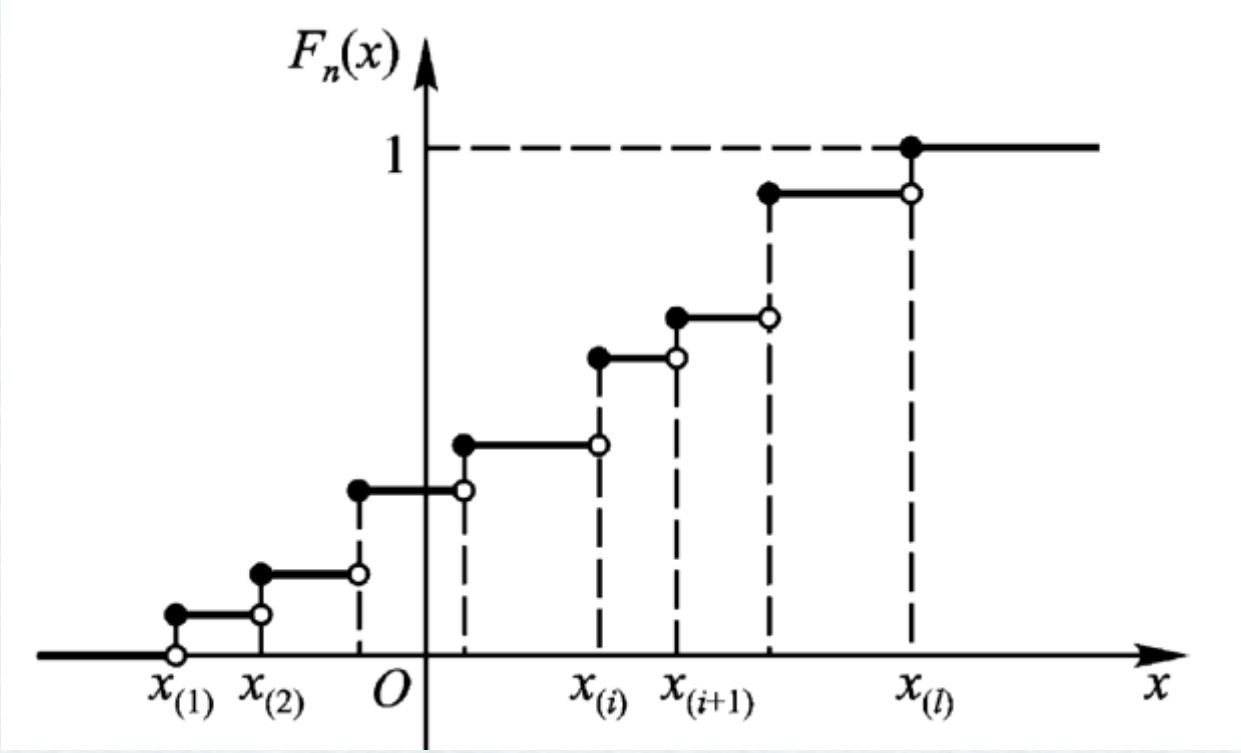
先将 $x_1, x_2, \cdots, x_n$ 按自小到大的次序排列，

并重新编号，　　$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，

则经验分布函数 $F_n(x)$ 的观察值为

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

样本分布函数$F_n(x)$的图形如图所示



对于一个样本值，$F_n(x)$的观察值容易求得.

($F_n(x)$的观察值仍以 $F_n(x)$ 表示.)

例3 设总体 X 具有一个样本值 1, 2, 3,

则经验分布函数 $F_3(x)$的观察值为

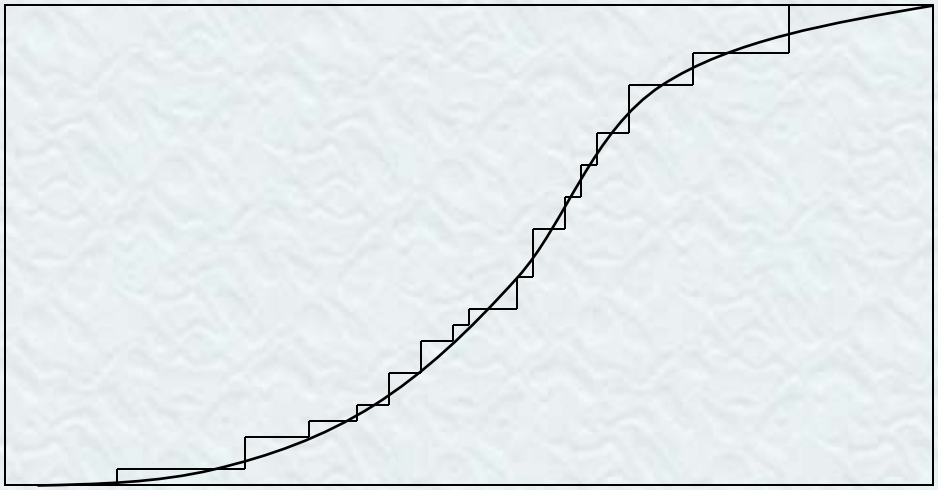
$$F_3(x)=\begin{cases}0, & x<1,\\ \frac{1}{3}, & 1\le x<2,\\ \frac{2}{3}, & 2\le x<3,\\ 1, & x\ge 3.\end{cases}$$

例4 设总体X 具有一个样本值 1, 1, 2,

则经验分布函数 $F_3(x)$的观察值为

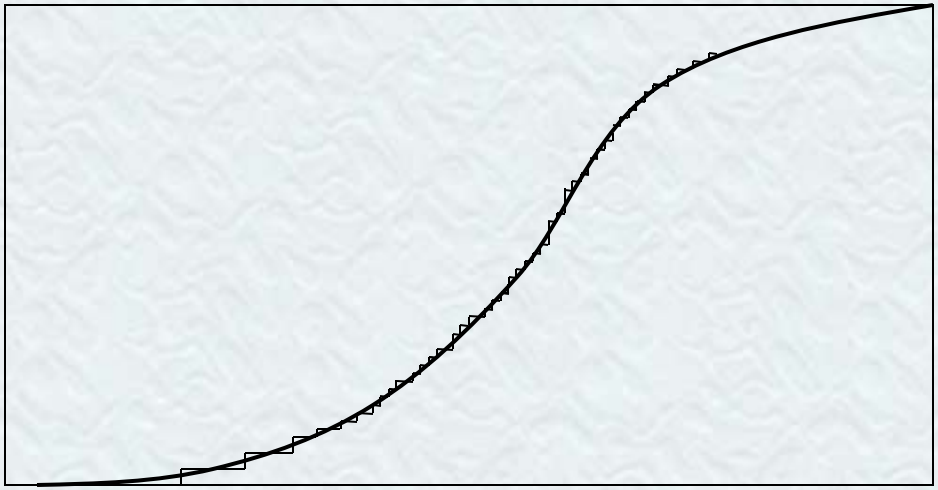
$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{2}{3}, & 1\le x<2,\\ 1, & x\ge 2.\end{cases}$$

(a)



2000

(b)



20000

对于不同的样本，得到的样本分布函数也不尽相同。随着n的增大，样本分布函数所描绘的曲线越来越光滑，而且与总体分布函数越来越接近。

有定理表明，样本分布函数实际上将近似地等于总体的分布函数。

由伯努利大数定律，可得：对任意固定的 x，

$$\lim_{n\to\infty} P\left\{\left|F_n(x)-F(x)\right|>\varepsilon\right\}=0,$$

即 $F_n(x)\xrightarrow{P}F(x),\quad n\to\infty$

对于任一实数 x，当 n充分大时，经验分布函数是总体分布函数 $F(x)$的一个良好的近似。

——样本推断总体的理论依据

格里汶科定理

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

格里汶科资料

Boris Vladimirovich Gnedenko

Born: 1 Jan 1912 in Simbirsk (now Ulyanovskaya), Russia

Died: 27 Dec 1995 in Moscow, Russia

补充：样本数据的整理与显示

频数频率表；直方图；茎叶图

直方图

数理统计中研究连续型随机变量X的样本分布时，通常需要作出样本的频率直方图（简称直方图），作直方图的步骤如下：

1. 找出样本观测值$x_1, x_2, \cdots, x_n$ 中的最小值与最大值,分别记作x_1^*与 x_n^*，即

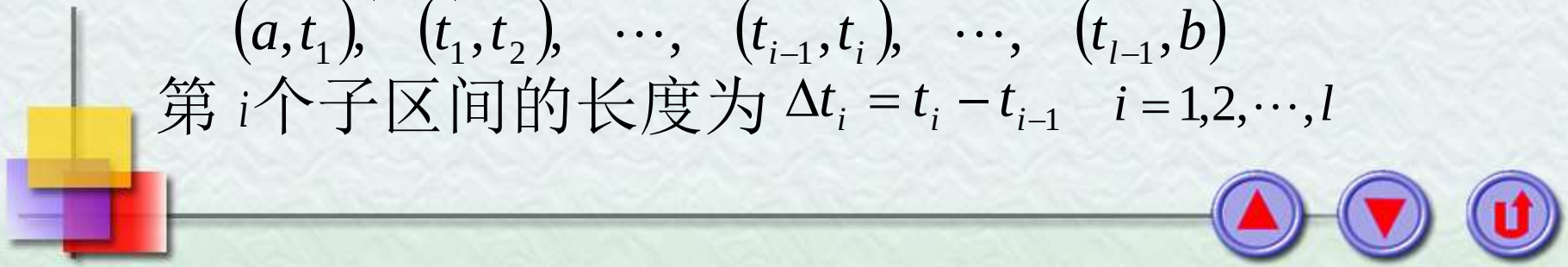
$$x_1^* = \min(x_1, x_2, \cdots, x_n), \qquad x_n^* = \max(x_1, x_2, \cdots, x_n)$$

2. 适当选取略小于x_1^*的数a与略大于x_n^*的数b，并用分点 $a = t_0 < t_1 < t_2 < \cdots < t_{l-1} < t_l = b$

把区间(a,b)分成 l 个子区间

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

第 i个子区间的长度为 $\Delta t_i = t_i - t_{i-1} \quad i = 1,2,\cdots,l$



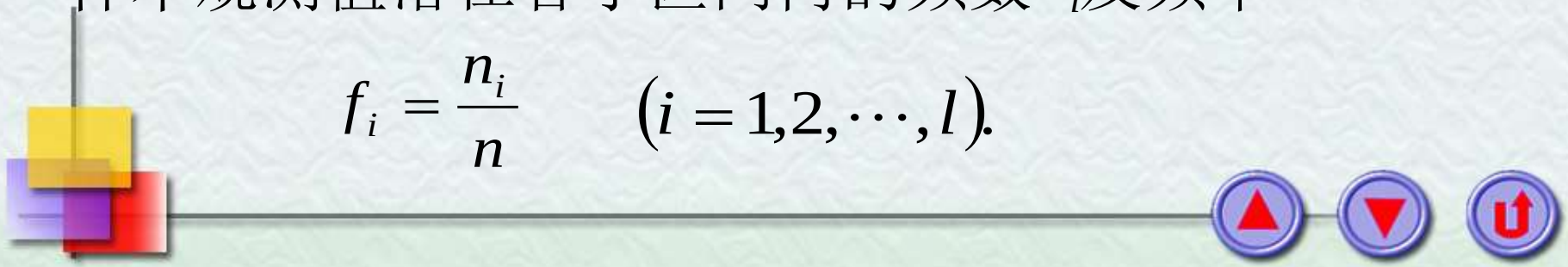
各子区间的长度可以相等，也可以不等；若使各子区间的长度相等，则有 $\Delta t_i = \dfrac{b-a}{l}$

子区间的个数一般取为8至15个，太多则由于频率的随机摆动而使分布显得杂乱，太少则难于显示分布的特征。

此外，为了方便起见，分点t_i 应比样本观测值x_i多取一位小数。

3.把所有样本观测值逐个分到各子区间内，并计算样本观测值落在各子区间内的频数n_i及频率

$$f_i = \frac{n_i}{n} \qquad (i = 1,2,\cdots,l).$$



4.在Ox轴上截取各子区间，并以各子区间为底，

以$\dfrac{f_i}{t_i - t_{i-1}}$为高作小矩形，各个小矩形的面积$\Delta S_i$

就等于样本观测值落在该子区间内的频率，即

$$\Delta S_i = (t_i - t_{i-1})\frac{f_i}{t_i - t_{i-1}} = f_i \quad (i = 1,2,\cdots,l).$$

所有小矩形的面积的和 $\sum_{i=1}^{l} \Delta S_i = \sum_{i=1}^{l} f_i = 1.$

这样作出的所有小矩形就构成了直方图。

因为样本容量n充分大时，随机变量X的取值落在各个子区间(t_{i-1}, t_i)内的频率近似等于其概率

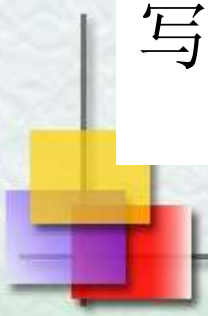
即 $f_i \approx P(t_{i-1} < X < t_i) \quad (i = 1,2,\cdots,l)$

所以直方图大致地描述了总体X的概率分布。

例 测量100个某种机械零件的质量，得到样本观测值如下（单位：g）

246 251 259 254 246 253 237 252 250 251
249 244 249 244 243 246 256 247 252 252
250 247 255 249 247 252 252 242 245 240
260 263 254 240 255 250 256 246 249 253
246 255 244 245 257 252 250 249 255 248
258 242 252 259 249 244 251 250 241 253
250 265 247 249 253 247 248 251 251 249
246 250 252 256 245 254 258 248 255 251
249 252 254 246 250 251 247 253 252 255
254 247 252 257 258 247 252 264 248 244

写出零件质量的频率分布表并作直方图。



解 因为样本观测中最小值为237，最大值为265，

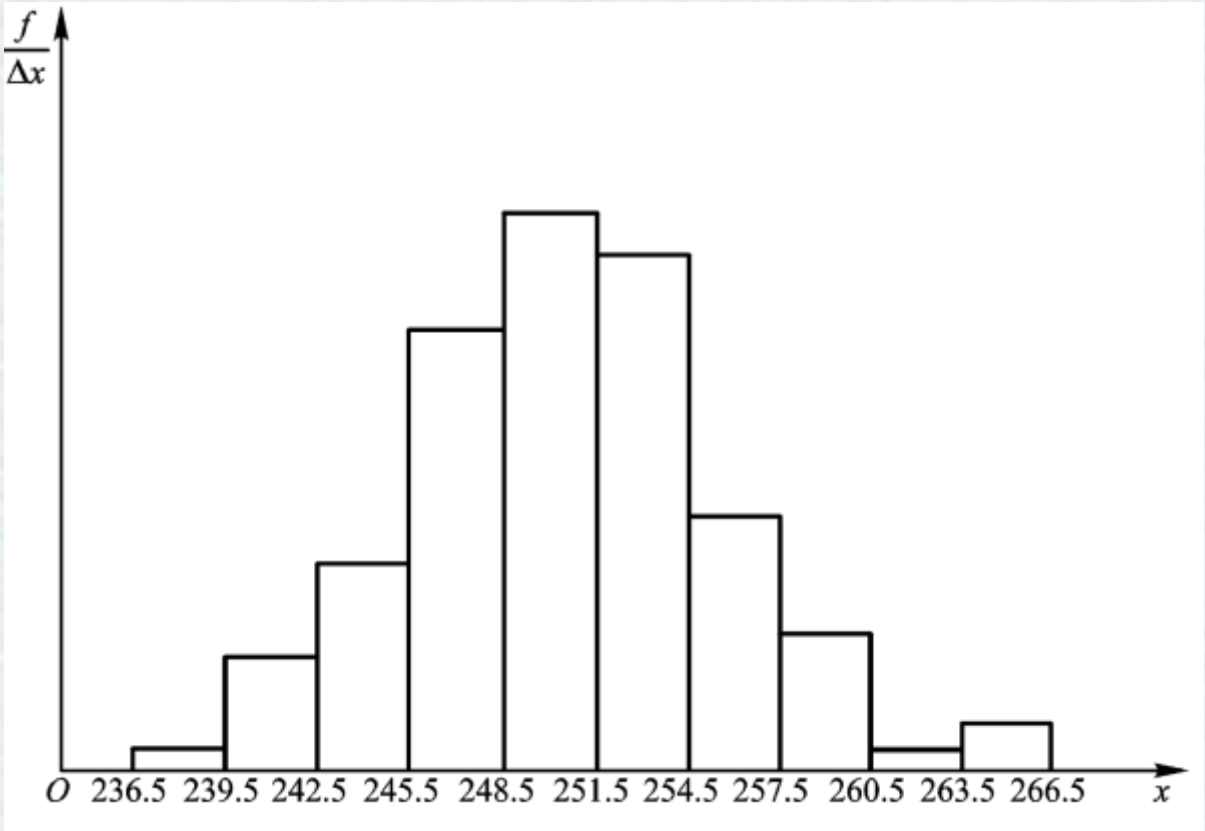
所以我们把数据的分布区间确定为（236.5，266.5）

并把这个区间等分为10个子区间

(236.5,239.5),
(239.5，242.5),
…,
(263.5,266.5)

由此得到零件质量的频率分布表：

零件质量/ g	频数 n_i	频率 f_i
236.5～239.5	1	0.01
239.5～242.5	5	0.05
242.5～245.5	9	0.09
245.5～248.5	19	0.19
248.5～251.5	24	0.24
251.5～254.5	22	0.22
254.5～257.5	11	0.11
257.5～260.5	6	0.06
260.5～263.5	1	0.01
263.5～266.5	2	0.02
总计	100	1.00



直方图

作业：5.14（1），5.15