

# 样本与统计量



# 引言

前面五章我们讲述了概率论的基本内容，随后的四章将讲述数理统计。数理统计是具有广泛应用的一个数学分支，它以概率论为理论基础，根据试验或现象得到的数据，来研究随机现象，对研究对象的客观规律性作出种种合理的估计和判断。

数理统计的内容包括：如何收集、整理数据资料；如何对所得到的数据资料进行分析研究，从而对所研究对象的性质、特点作出推断。后者就是我们所说的统计推断问题。本书只讲述统计推断的基本内容。

# 引言

随机变量及其所伴随的概率分布全面描述了随机现象的统计性规律。

概率论的许多问题中，随机变量的概率分布通常是已知的，或者假设是已知的，而一切计算与推理都是在这已知是基础上得出来的。

但实际中，情况往往并非如此，一个随机现象所服从的分布可能是完全不知道的，或者知道其分布概型，但是其中的某些参数是未知的。

# 引言

例如：

某公路上行驶车辆的速度服从什么分布是未知的；

电视机的使用寿命服从什么分布是未知的；

产品是否合格服从两点分布，但参数——合格率 $p$ 是未知的；

数理统计的任务则是以概率论为基础，根据试验所得到的数据，对研究对象的客观统计规律性做出合理的推断。

# 学习的基本内容

从第六章开始，我们学习数理统计的基础知识。

数理统计的任务是以概率论为基础，根据试验所得到的数据，对研究对象的客观统计规律性作出合理的推断。数理统计所包含的内容十分丰富，本书介绍其中的参数估计、假设检验、方差分析、回归分析等内容。第六章主要介绍数理统计的一些基本术语、基本概念、重要的统计量及其分布，它们是后面各章的基础。

■ **引言：数理统计学**是一门关于数据收集、整理、分析和推断的科学。在概率论中已经知道，由于大量的随机试验中各种结果的出现必然呈现它的规律性，因而从理论上讲只要对随机现象进行足够多次观察，各种结果的规律性一定能清楚地呈现，但是实际上所允许的观察永远是有限的，甚至是少量的。

例如：若规定灯泡寿命低于1000小时者为次品，如何确定次品率？由于灯泡寿命试验是破坏性试验，不可能把整批灯泡逐一检测，只能抽取一部分灯泡作为样本进行检验，以样本的信息来推断总体的信息，这是数理统计学研究的问题之一。

09



## § 6.1 随机样本

### 总体和样本

**总体** —— 研究对象全体元素组成的集合

所研究的对象的某个(或某些)数量指标的全体

**个体** —— 组成总体的每一个元素

**容量** —— 总体中所包含个体的个数

**有限总体** —— 容量为有限的总体

**无限总体** —— 容量为无限的总体



例如考察某大学一年级男生的身高这一试验中，若一年级男生共2000人，每个男生的身高是一个可能观察值，所形成的总体中共有2000个可能的观察值，是一个**有限总体**。又如，考察某一湖泊中某种鱼的含汞量所得的总体也是**有限总体**。

观察并记录某一地点每天（包括以往、现在和将来）的最高气温，或者测量某一湖泊任一地点的深度，所得的总体是**无限总体**。

有些有限总体，它的容量很大，我们可以认为它是一个无限总体。例如考察全国正在使用的某种型号灯泡的寿命所形成的总体，由于可能观察值的个数很多，就可以认为是无限总体。





我们所要研究的个体的某一个数量指标（例如男生的身高），它对总体中不同的个体来说取不同的值，既具有**不确定性**。我们自总体中随机取一个个体，观察它的数量指标的值，这就是一个**随机试验**。而**数量指标 $X$** 作为随机试验中被观察的量，它的取值随试验结果而定，它是一个**随机变量**。

**我们对总体的研究就是对随机变量 $X$ 的研究。** $X$ 的分布函数和数字特征，分别称为总体的分布函数和数字特征。这样，**一个总体对应于一个随机变量 $X$** 。今后不再区分总体与相应的随机变量，笼统称为总体 $X$ 。即如下：



**总体** —— 所研究的对象的某个(或某些)数量指标的全体, 它是一个**随机变量** (或多维随机变量). 记为 $\mathbf{X}$ .

$\mathbf{X}$  的分布函数和数字特征称为总体的分布函数和数字特征.

**个体** —— 即总体的每个数量指标, 可看作随机变量

$\mathbf{X}$  的某个取值. 用  $X_i$  表示.

例如: 我们检验自动生产线出来的零件是次品还是正品, 用1表示产品为次品, 用0表示产品为正品, 设出现次品的概率为 $p$ , 那么总体就是由一些具有数量指标为1和一些具有数量指标为0的个体所组成。这个总体对应于一个参数为 $p$ 的0-1分布, 我们就将它说成是0-1分布的总体。



## 样本 —— 从总体中抽取的部分个体.

用 $(X_1, X_2, \dots, X_n)$ 表示,  $n$ 为样本容量。称 $(x_1, x_2, \dots, x_n)$ 为总体 $X$ 的一个容量为 $n$ 的样本观测值, 或称样本的一个实现.

所谓从总体中抽取一个个体, 就是对 $X$ 进行一次观察并记录其结果, 我们在相同的条件下对总体 $X$ 进行 $n$ 次重复的、独立的观察, 并将 $n$ 次观察结果按试验的次序记为 $X_1, X_2, \dots, X_n$ , 由于 $X_1, X_2, \dots, X_n$ 是对随机变量 $X$ 观察的结果, 各次观察是在相同的条件下独立进行的, 所以有理由认为 $X_1, X_2, \dots, X_n$ 是相互独立的, 且都是与 $X$ 具有相同分布的随机变量。



## 简单随机样本

若总体  $X$  的样本

$(X_1, X_2, \dots, X_n)$  满足:

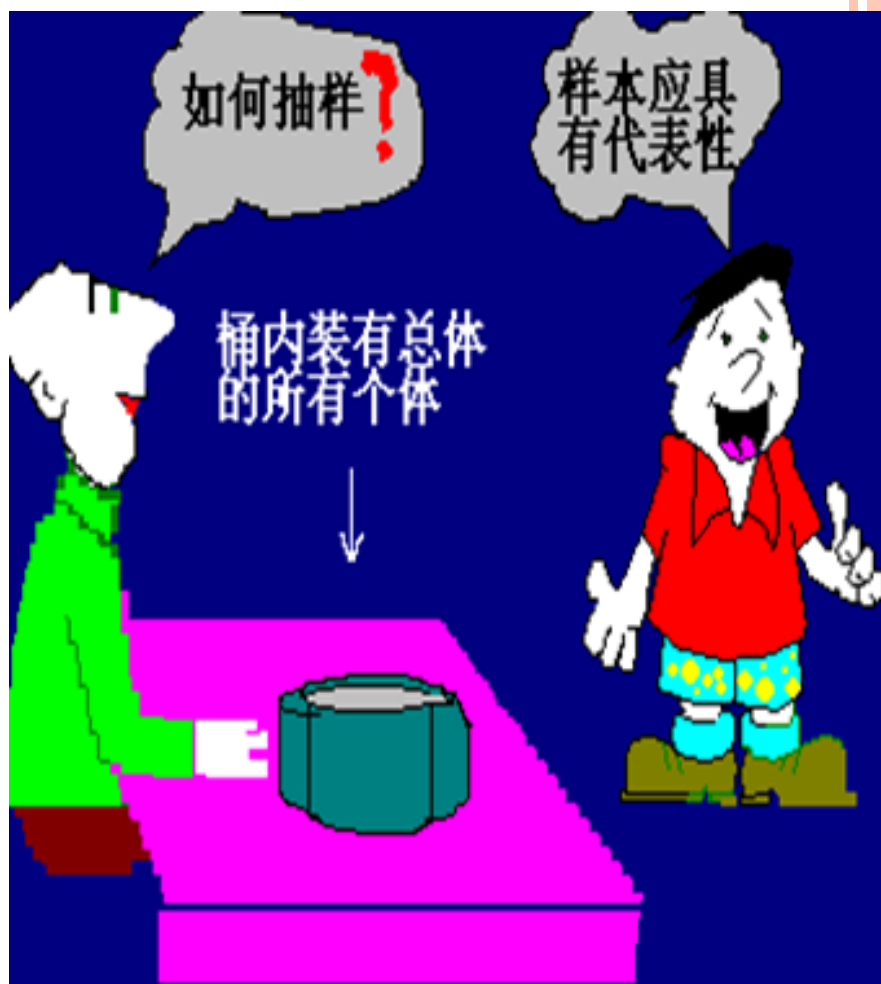
1. 与  $X_1, X_2, \dots, X_n$   
有相同的分布

(2)  $X_1, X_2, \dots, X_n$  相互独立

则称  $(X_1, X_2, \dots, X_n)$  为  
简单随机样本, 简称样本。

样本的观察值  
称为样本值。

$x_1, x_2, \dots, x_n$



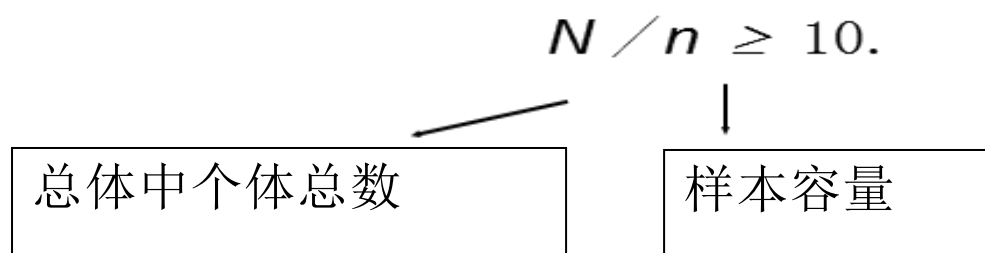
样本的两重性：

- 1、在泛指任一次抽取的结果时，  $X_1, X_2, \dots, X_n$  表示  $n$  个随机变量（样本）
- 2、在具体的依次抽取之后，  $X_1, X_2, \dots, X_n$  表示  $n$  个具体的数值（样本值）。



在数理统计中，有意义的样本容量一般要求 $n \geq 50$ 才有统计意义。

一般, 对有限总体, 放回抽样所得到的样本为简单随机样本, 但使用不方便, 常用不放回抽样代替. 而代替的条件是



设总体  $\mathbf{X}$  的分布函数为  $\mathbf{F}(\mathbf{x})$ , 则样本  $(X_1, X_2 \cdots X_n)$

的联合分布函数为

$$F_{\text{总}}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^n F(x_i)$$

若总体  $\mathbf{X}$  的概率密度为  $\mathbf{f}(\mathbf{x})$ , 则样本的联合概率密度为

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



例如 设某批产品共有 $N$ 个, 其中的次品数为 $M$ , 其  
次品率为

$$p = M / N$$

若  $p$  是未知的, 则可用抽样方法来估计它.

从这批产品中任取一个产品, 用随机变量  
 $X$ 来描述它是否是次品:

$$X = \begin{cases} 1, & \text{所取的产品是次品} \\ 0, & \text{所取的产品不是次品} \end{cases}$$

$X$  服从参数为 $p$  的0-1分布, 可用如下表示  
方法:

$$f(x, p) = p^x (1 - p)^{1-x}, \quad x = 0, 1$$





设有放回地抽取一个容量为  $n$  的样本

$$(X_1, X_2, \dots, X_n)$$

其样本值为

$$(x_1, x_2, \dots, x_n)$$

样本空间为

$$\{(x_1, x_2, \dots, x_n) \mid x_i = 0, 1, i = 1, 2, \dots, n\}$$

$(X_1, X_2, \dots, X_n)$  的联合分布为

$$\begin{aligned} f_{\text{总}}(x_1, x_2, \dots, x_n) &= \prod_{i=1}^n f(x_i) \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$



若抽样是无放回的,则前次抽取的结果会影响后面抽取的结果. 例如

$$P\{X_2 = 1 \mid X_1 = 1\} = \frac{M-1}{N-1} = \frac{p - \frac{1}{N}}{1 - \frac{1}{N}} \xrightarrow{N \rightarrow \infty} p$$

$$P\{X_2 = 1 \mid X_1 = 0\} = \frac{M}{N-1} = \frac{p}{1 - \frac{1}{N}} \xrightarrow{N \rightarrow \infty} p$$

所以, 当样本容量  $n$  与总体中个体数目  $N$  相比很小时, 可将无放回抽样近似地看作放回抽样.



# 统计量

样本是统计推断的依据，但在实际问题中，往往不是直接使用样本本身，而是针对不同的问题构造不同的样本函数，利用这种样本的函数进行统计推断。



# 统计量

设  $(X_1, X_2, \dots, X_n)$  是取自总体  $X$  的一个样本,

$g(X_1, X_2, \dots, X_n)$  是  $X_1, X_2, \dots, X_n$  的实值连续函数  
 且不含未知参数, 则称随机变量  $g(X_1, X_2, \dots, X_n)$  为统计量.

若  $(X_1, X_2, \dots, X_n)$  是一个样本值, 称  $g(x_1, x_2, \dots, x_n)$   
 为统计量  $g(X_1, X_2, \dots, X_n)$  的一个样本值

例如: 设  $(X_1, X_2, X_3)$  是服从正态总体  $N(\mu, \sigma^2)$  中抽取  
 的一个样本, 其中  $\mu$  为已知参数,  $\sigma$  为未知参数,

则  $X_1 + X_2 + 3\mu X_3$        $X_1^2 + 3\mu X_2 X_3$       是统计量

$X_1 + \sigma X_2 + X_3^2$        $X_1 X_2 X_3 + \sigma$       不是统计量



## 几个常用的统计量

设  $(X_1, X_2, \dots, X_n)$  是总体  $X$  的一个样本，

样本均值 (**sample mean**)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

描述数据分布的**中心位置**。

样本方差(**sample variance**)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

描述数据分布的**离散程度**。



## 几个常用的统计量

设  $(X_1, X_2, \dots, X_n)$  是总体  $X$  的一个样本,

样本均方差或标准差

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

它们的观测值用相应的小写字母表示.



## 几个常用的统计量

设  $(X_1, X_2, \dots, X_n)$  是总体  $X$  的一个样本,

样本**K**阶原点矩

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

样本**K**阶中心矩

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$



**例1** 从一批机器零件毛坯中随机地抽取10件，测得其重量为(单位：公斤)：

**210, 243, 185, 240, 215,**

**228, 196, 235, 200, 199**

求这组样本值的均值、方差、二阶原点矩与二阶中心矩.

**解** 令  $(x_1, x_2, \dots, x_{10})$   
 $= (210, 243, 185, 240, 215,$   
 $228, 196, 235, 200, 199)$

则  $\bar{x} = \frac{1}{10}(210 + 243 + 185 + 240 + 215$   
 $+ 228 + 196 + 235 + 200 + 199) = 217.19$



$$s^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2 = 433.43$$

$$A_2 = \frac{1}{10} \sum_{i=1}^{10} x_i^2 = 47522.5$$

$$B_2 = \frac{9}{10} s^2 = \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x})^2 = 390.0$$



## $k$ 阶矩的概念

定义 设  $X$  为随机变量，若  $X$  的  $k$  阶原点矩，记作

存在，则称  $E(X^k)$

$$E(X^k)$$

样本的  $k$  阶原点矩，记作

$$\mu_k = E(X^k)$$
$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

结论：

$$A_k \xrightarrow{P} \mu_k \quad (n \rightarrow \infty), k = 1, 2, \dots$$

原因：辛钦大数定律

作用：矩估计法的理论依据



# 数据的简单处理

为了研究随机现象，首要的工作是收集原始数据。一般通过抽样调查或试验得到的数据往往是杂乱无章的，需要通过整理后才能显示出它们的分布状况。

数据的简单处理是以一种直观明了方式加工数据。

它包括两个方面——数据整理

计算样本特征数

# 数据的简单处理

数据整理：将数据分组

作频率分布表

计算各组频数

作频率直方图

计算样本特征数：

(1) 反映趋势的特征数

样本均值

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

中位数：数据按大小顺序排列后，位置居中的那个数  
或居中的两个数的平均数。

众数：样本中出现最多的那个数。

# 数据的简单处理

(2) 反映分散程度的特征数：极差、四分位差

**极差**——样本数据中最大值与最小值之差，

$$R = M - m$$

**四分位数**——将样本数据依概率分为四等份的**3**个数据，

依次称为第一、第二、第三四分位数。

第一四分位数 $Q_1$ :

$$P\{X < Q_1\} = 0.25$$

第二四分位数 $Q_2$ :

$$P\{X < Q_2\} = 0.5$$

第三四分位数 $Q_3$ :

$$P\{X < Q_3\} = 0.75$$

**例1** 为对某小麦杂交组合 $F_2$ 代的株高 $X$ 进行研究，抽取容量为**100**的样本，测试的原始数据记录如下(单位：厘米)，试根据以上数据，画出它的频率直方图，求随机变量 $X$ 的分布状况。

87	88	111	91	73	70	92	98	105	94
99	91	98	110	98	97	90	83	92	88
86	94	102	99	89	104	94	94	92	96
87	94	92	86	102	88	75	90	90	80
84	91	82	94	99	102	91	96	94	94
85	88	80	83	81	69	95	80	97	92
96	109	91	80	80	94	102	80	86	91
90	83	84	91	87	95	76	90	91	77
103	89	88	85	95	92	104	92	95	83
86	81	86	91	89	83	96	86	75	92



第一. 整理原始数据, 加工为分组资料, 作出频率分布表, 画直方图, 提取样本分布特征的信息. 步骤如下:

1. 找出数据中最小值  $m=69$ , 最大值  $M=111$ ,

现取区间  $[67.5, 112.5]$ , 它能覆盖区间  $[69, 111]$

2. 数据分组, 根据样本容量  $n$  的大小, 决定分**组数**  $k$ 。

一般规律	$30 \leq n \leq 40$	$5 \leq k \leq 6$
	$40 \leq n \leq 60$	$6 \leq k \leq 8$
	$60 \leq n \leq 100$	$8 \leq k \leq 10$
	$100 \leq n \leq 500$	$10 \leq k \leq 20$

数据分组数参考表

数据数	40 ~6 0	10 0	15 0	20 0	40 0	60 0	80 0	10 00	15 00	20 00	50 00	10 00 0
分组数	6~ 8	7~ 9	10 ~1 5	16	20	24	27	30	35	39	56	74





本例取组数**k=9**.

一般采取等距分组（也可以不等距分组），组距为

$$\frac{112.5 - 67.5}{9} = 5$$



### 3. 确定组限

小区间的端点称为组限。

分组如下：

<b>[67.5,72.5)</b>	<b>[72.5,77.5)</b>	<b>[77.5,82.5)</b>
<b>[82.5,87.5)</b>	<b>[87.5,92.5)</b>	<b>[92.5,97.5)</b>
<b>[97.5,102.5)</b>	<b>[102.5,107.5)</b>	<b>[107.5,112.5)</b>

### 4. 将数据分组，计算出各组频数，作频数、频率分布表

数出落在每个小区间内的数据的频率 $f_k$ ，算出频率 $f_k/n$

( $n=100, i=1, 2 \cdots 9$ ). 如下表

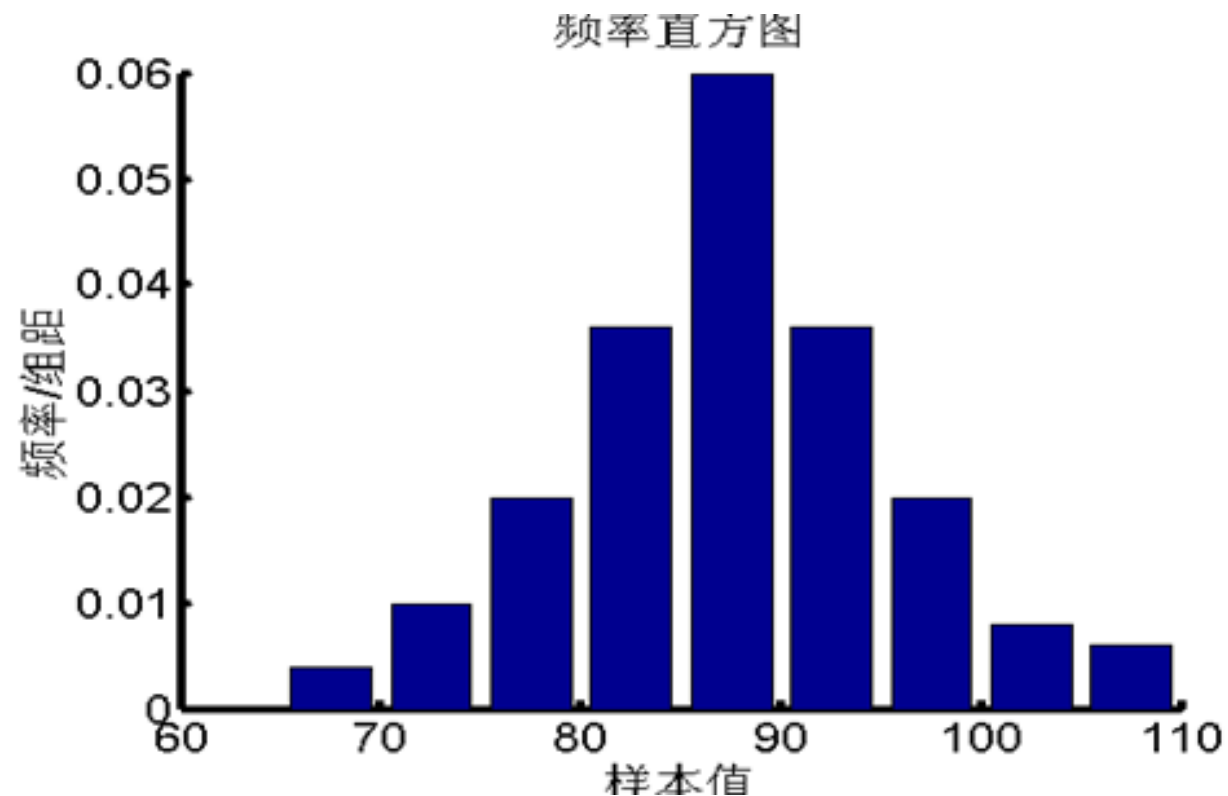


组序	区间范围	频数 $f_j$	频率 $W_j=f_j/n$	累计频率 $F_j$
1	<b>[67.5,72.5)</b>	2	0.02	0.02
2	[72.5, 77.5)	5	0.05	0.07
3	[77.5, 82.5)	10	0.10	0.17
4	[82.5, 87.5)	18	0.18	0.35
5	[87.5, 92.5)	30	0.3	0.65
6	[92.5, 97.5)	18	0.18	0.83
7	[97.5, 102.5)	10	0.1	0.93
8	[102.5, 107.5)	4	0.04	0.97
9	[107.5, 112.5)	3	0.03	1.00

## 5. 作出频率直方图

以样本值为横坐标，频率/组距为纵坐标；

现在自左至右依次做以分组区间为底，以频率/组距为高的小矩形，如图。这样的图形叫频率直方图



从频率直方图可看到：靠近两个极端的数据出现比较少，而中间附近的数据比较多，即中间大两头小的分布趋势，——随机变量分布状况的最粗略的信息。

在频率直方图中， 每个矩形面积恰好等于样本值落在该矩形对应的分组区间内的频率、

频率直方图中的小矩形的面积近似地反映了样本数据落在某个区间内的可能性大小，故它可近似描述 $X$ 的分布状况。



## 第二. 计算样本特征数

1.反映集中趋势的特征数: 样本均值、中位数、众数等

样本均值**MEAN**

中位数**MEDIAN**

众数

$$\bar{X} = 90.3$$

91

91, 94

2.反映分散程度的特征数: 样本方差、样本标准差、  
极差、四分位差等

样本方差    样本标准差    **Q1**    **Q3**    极差    四分位差

68.6909    8.288    85.25    95    42    4.875

上述差异特征统计量的值越小, 表示离散程度越小.

## 例1 DOS状态下的MINITAB操作

MTB > **set c1**

DATA> 87 88 111 91 73 70 92 98 105 94 99 91 98

DATA> 110 98 97 83 90 83 92 88 86 94 102 99 89 104

DATA> 94 94 92 96 87 94 92 86 102 88 75 90 90 80

DATA> 84 91 82 94 99 102 91 96 94 94 85 88 80 83

DATA> 81 69 95 80 97 92 96 109 91 80 80 94 102

DATA> 80 86 91 90 83 84 91 87 95 76 90 91 77 103

DATA> 89 88 85 95 92 104 92 95 83 86 81 86 91 89 83

DATA> 96 86 75 92

MTB > **end**

MTB > **describe c1**

显示:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

中位数

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

	N	MEAN	MEDIAN	TRMEAN	STDEV
C1	100	90.300	91.000	90.322	8.288

	SEMEAN	MIN	MAX	Q1	Q3
C1	0.829	69.000	111.000	85.250	95.000

$$\frac{S}{\sqrt{n}}$$

第一四分位数


第三四分位数



```
MTB>CODE (67.5:72.49)70 (72.5:77.49)75  
      (77.5:82.49)80 (82.5:87.49)85  
      (87.5:92.49)90 (92.5:97.49)95  
      (97.5:102.49)100 (102.5:107.49)105  
      (107.5:112.49)110 C1 C2
```

```
MTB>TALLY C2;
```

```
SUBC>ALL.
```



显示各列数据的频数、  
累计频数、频率、累计频率



将**C1**数据列重新编码，  
并保存到**C2**数据列



## 显示结果

C2	COUNTS	CUMCNTS	PERCENTS	CUMPCENTS
	(频数)	(累计频数)	(频率)	(累计频率)
	1	2	0.02	0.02
	5	7	0.05	0.07
	10	17	0.10	0.17
	18	35	0.18	0.35
	30	65	0.30	0.65
	18	83	0.18	0.83
	10	93	0.10	0.93
	4	97	0.04	0.97
	3	100	0.03	1.00

