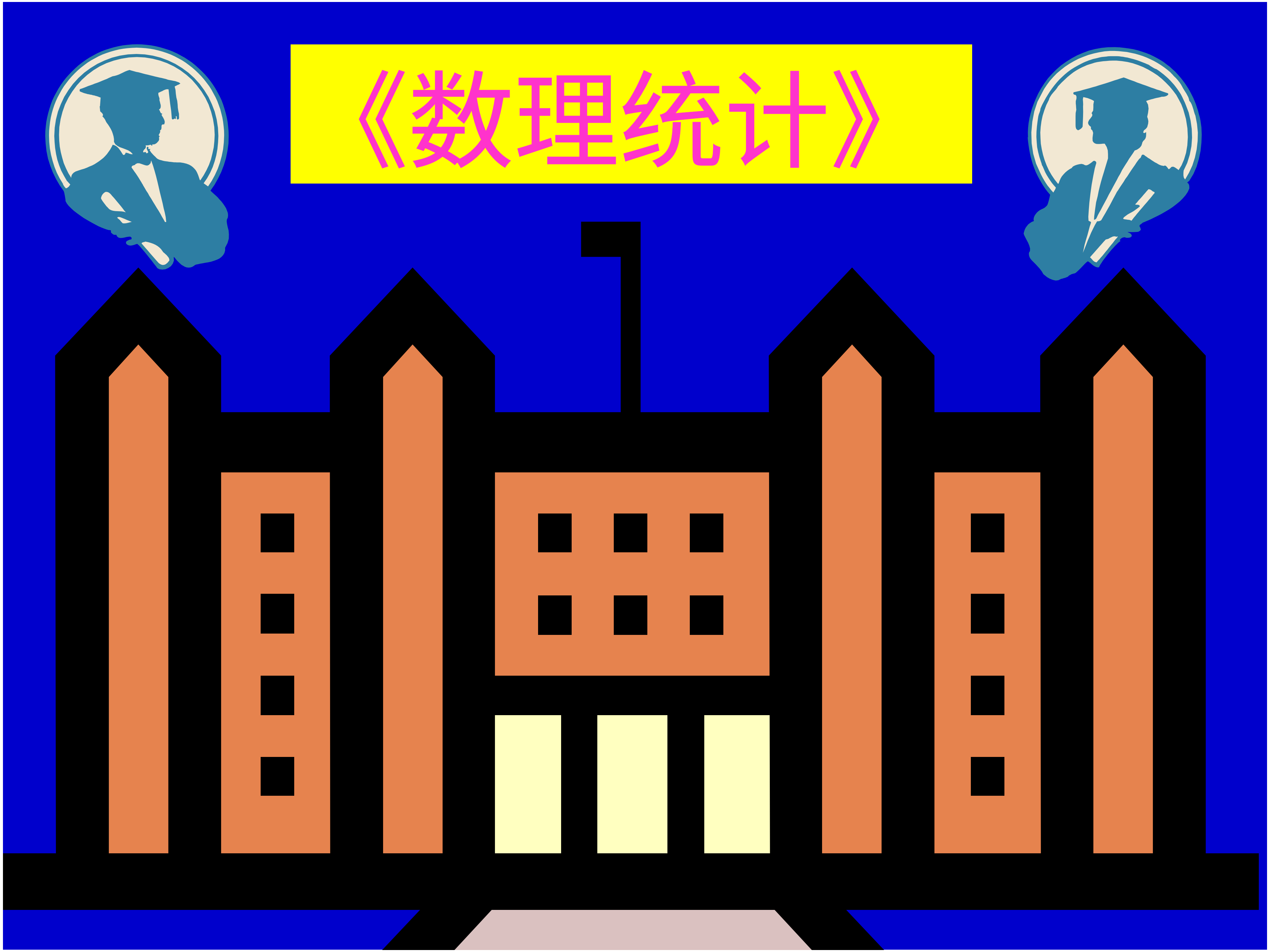
《数理统计》



第六章　数理统计的基本概念

数理统计的分类

描述统计学——

对随机现象进行观测、试验，以取得有代表性的观测值

推断统计学——

对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性

推断统计学

- 参数估计 (第七章)
- 假设检验 (第八章)
- 方差分析 (第九章)
- 回归分析 (第九章)

§ 6.1 基本概念

一、总体、个体和随机样本

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,它是一个随机变量(或多维随机变量).记为X .

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 —— 组成总体的每一个元素

即总体的每个数量指标,可以看作随机变量 X 的某个取值.用 X_i 表示.

样本 —— 从总体中抽取的部分个体.

用 $(X_1, X_2, \cdots, X_n)$ 表示样本, n 为样本容量.

称 $(x_1, x_2, \cdots, x_n)$ 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.

样本空间 ——样本所有可能取值的集合.

简单随机样本

若总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 满足：

(1) $X_1, X_2, \cdots, X_n$ 相互独立

(2) $X_1, X_2, \cdots, X_n$ 与 X 有相同的分布

则称 $(X_1, X_2, \cdots, X_n)$ 为简单随机样本.

一般,对有限总体,采用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

$$N / n \geq 10.$$

其中N为总体中个体的数目，n 为样本容量.

设$(X_1, X_2, \cdots, X_n)$为总体X的简单随机样本，X 的分布函数为$F(x)$，则$(X_1, X_2, \cdots, X_n)$的联合分布函数为

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

若总体X的概率密度函数为$f(x)$，则 $(X_1, X_2, \cdots, X_n)$ 的联合概率密度函数为

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

例如　某批产品共有N个,其次品数为M，其次品率为

$$p=\frac{M}{N}$$

若 p未知,则可用抽样的方法来估计它.

从这批产品中任取一个产品，记

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

X 服从参数为 p 的0-1分布,

$$P\{X=x\}=p^{x}(1-p)^{1-x}, \qquad x=0,1$$

设有放回地抽取了一个容量为 n 的样本

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

其样本值为 $(x_1, x_2, \cdots, x_n)$

样本空间为

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

$(X_1, X_2, \cdots, X_n)$ 的联合分布律为

$$P\{X_1 = x_1, \cdots, X_n = x_n\} = \prod_{i=1}^{n} P\{X_i = x_i\}$$

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

若抽样是无放回地,则前次抽取的结果会影响后面抽取的结果.例如

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

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

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.

二、频率分布和直方图

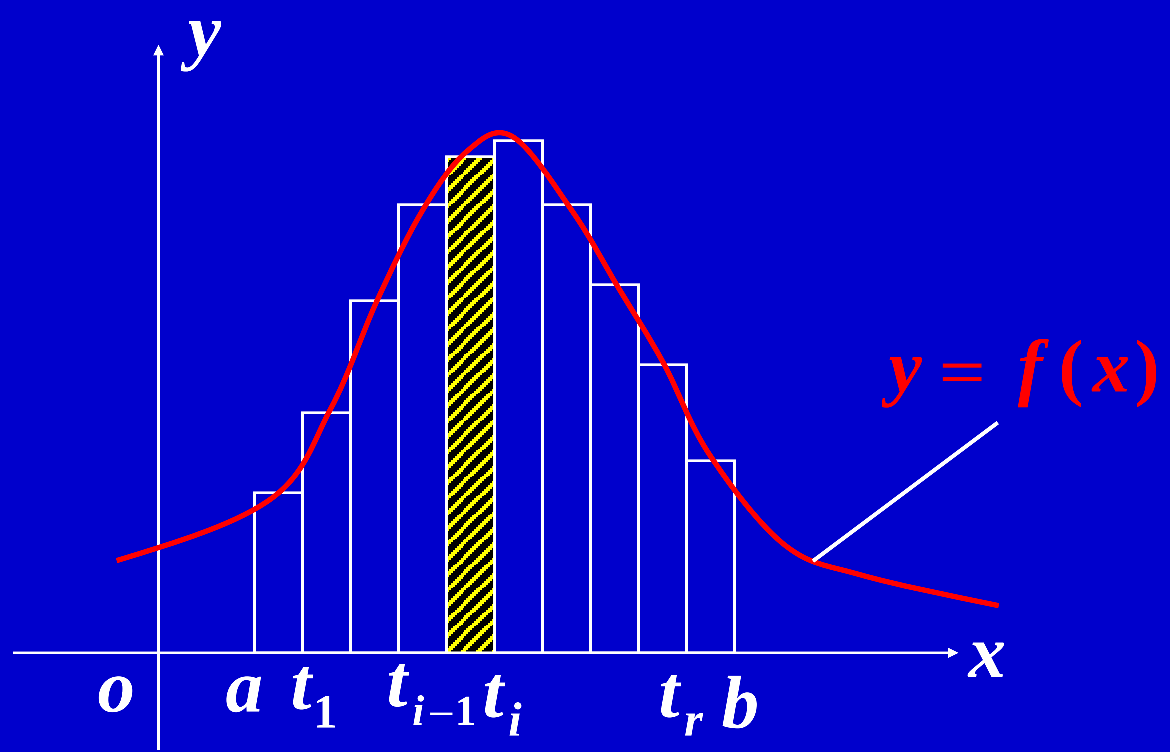
设总体X是一个连续型$r.v.$，x_1^* 和 x_n^* 分别是 $x_1, x_2, \cdots, x_n$ 的最小值和最大值，选取略小于 x_1^* 的a和略大于 x_n^* 的b， 并在区间$[a,b]$中插入若干个分点 $a = t_0 < t_1 < \cdots < t_r < t_{r+1} = b$ 把区间分成 $r+1$ 个小区间，且每个区间的区间长度均相等.

算出样本观察值落入第 i个小区间 $(t_{i-1}, t_i]$ 的个数（频数），记为 m_i ，称 $v_i = \dfrac{m_i}{n}$为样本观察值落入区间 $(t_{i-1}, t_i]$的频率，$i = 1,2,\cdots,r+1$。

列频率分布表如下：

分组	频数m_i	频率v_i	累计频率
$(t_0,t_1]$	m_1	$\frac{m_1}{n}$	$\frac{m_1}{n}$
$(t_1,t_2]$	m_2	$\frac{m_2}{n}$	$\frac{m_1+m_2}{n}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$(t_r,t_{r+1}]$	m_{r+1}	$\frac{m_{r+1}}{n}$	$\sum_{i=1}^{r+1}\frac{m_i}{n}=1$
合计	$\sum_{i=1}^{r+1}m_i=n$	$\sum_{i=1}^{r+1}\frac{m_i}{n}=1$	

在xoy平面上画一排矩形，即对于每个$i\ (1 \le i \le r+1)$ 作出以x轴上的区间为底，以$y_i = \dfrac{v_i}{t_i - t_{i-1}}$ 为高的矩形。这样的图叫做频率直方图。



三、经验分布函数

设$x_1, x_2, \cdots, x_n$是总体X的样本观察值，将它们 按从小到大的顺序排列成 ：$x_1^* \le x_2^* \le \cdots \le x_n^*$,对$\forall x \in R$,

定义函数$F_n^*(x)$:

$$F_n^*(x) = \begin{cases} 0\ , & x < x_1^* \\ \dfrac{k}{n}\ , & x_k^* \le x < x_{k+1}^* \quad (k = 1,2,\cdots,n-1) \\ 1\ , & x \ge x_n^* \end{cases}$$

称$F_n^*(x)$为总体X的经验分布函数。

例1 已知总体X的一组样本观察值为：0，0.2，0.25，-0.3，-0.1，2，0.15，1，-0.7, -1。求样本经验分布函数 $F_{10}^{*}(x)$ 。

解 $$F_{10}^{*}(x)=\begin{cases} 0 & x<-1 \\ 0.1 & -1\le x<-0.7 \\ 0.2 & -0.7\le x<-0.3 \\ 0.3 & -0.3\le x<-0.1 \\ 0.4 & -0.1\le x<0 \\ 0.5 & 0\le x<0.15 \\ 0.6 & 0.15\le x<0.2 \\ 0.7 & 0.2\le x<0.25 \\ 0.8 & 0.25\le x<1 \\ 0.9 & 1\le x<2 \\ 1 & x\ge 2 \end{cases}$$

四、统计量

1、定义

设$(X_1, X_2, \cdots, X_n)$是总体X的一个样本，

$$g(r_1, r_2, \cdots, r_n)$$

为一实值连续函数，且不含有未知参数，则称随机变量$g(X_1, X_2, \cdots, X_n)$为**统计量**.

若$(x_1, x_2, \cdots, x_n)$是一个样本观察值，

称 $$g(x_1, x_2, \cdots, x_n)$$

为统计量$g(X_1, X_2, \cdots, X_n)$的**一个观察值**

例2 $X \sim N(\mu,\sigma^2)$，μ,σ^2是未知参数，

$(X_1, X_2, \cdots, X_n)$ 是一样本,则

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

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

是统计量，其中 $X_i \sim N(\mu,\sigma^2)$

但 $$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$$

不是统计量.若 μ ,σ 已知,则为统计量

2、常用的统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本，称统计量

(1) $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ 为样本均值

(2) $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ 为样本方差

$S = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$ 为样本标准差

(3) $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ 为样本的**k**阶原点矩

(4) $B_k = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^k$ 为样本的**k**阶中心矩

例如

$$A_1 = \overline{X}$$

$$B_2 = \frac{n-1}{n} S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \triangleq S_n^2$$

(5) 顺序统计量与极差

设 $(x_1, x_2, \cdots, x_n)$ 为样本 $(X_1, X_2, \cdots, X_n)$

的一个实现，且 $x_1^* \le x_2^* \le \cdots \le x_n^*$

当 $(X_1, X_2, \cdots, X_n)$ 取值为 $(x_1, x_2, \cdots, x_n)$ 时，

定义随机变量 $X_{(k)} = x_k^*, k = 1,2,\cdots,n$

则称统计量 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为顺序统计量.

其中 $X_{(1)} = \min\limits_{1 \le k \le n} \{X_k\}, X_{(n)} = \max\limits_{1 \le k \le n} \{X_k\}$

称 $D_n = X_{(n)} - X_{(1)}$ 为极差

注 样本方差 $\boxed{S^2}$ 与样本二阶中心矩 $\boxed{S_n^2}$ 的不同

1） 关系式 $S^2 = \dfrac{n}{n-1} S_n^2$

推导 $$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)$$

$$= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2$$

$$= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = n(A_2 - \overline{X}^2)$$

故 $B_2 = A_2 - \overline{X}^2$ $\quad S^2 = \dfrac{n}{n-1}(A_2 - \overline{X}^2) = \dfrac{n}{n-1} S_n^2$

2) $E(S_n^2)=\dfrac{n-1}{n}\sigma^2$ $E(S^2)=\sigma^2$

推导 设$E(X)=\mu$, $D(X)=\sigma^2$ 则

$$E\left(\overline{X}\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\mu \qquad D\left(\overline{X}\right)=\frac{1}{n}\sigma^2$$

$$E(S_n^2)=EA_2-E\left(\overline{X}^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-\left[D\left(\overline{X}\right)+E^2\left(\overline{X}\right)\right]$$

$$=\sigma^2+\mu^2-\left(\frac{1}{n}\sigma^2+\mu^2\right)=\frac{n-1}{n}\sigma^2$$

$$E(S^2)=E\left[\frac{n}{n-1}S_n^2\right]=\frac{n}{n-1}ES_n^2=\sigma^2$$

例3　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 $(x_1, x_2, \cdots, x_{10})$

$$=(210,\ 243,\ 185,\ 240,\ 215,\ 228,\ 196,\ 235,\ 200,\ 199)$$

则 $\bar{x}=\frac{1}{10}(230+243+185+240+215$

$+228+196+235+200+199)$

$=217.19$

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

$$a_2=\frac{1}{10}\sum_{i=1}^{10}x_i^2=47522.5$$

$$b_2=\frac{9}{10}s^2=\frac{1}{10}\sum_{i=1}^{10}(x_i-\bar{x})^2=390.0$$

例4 在总体 $N(52, 6.3^2)$中，随机地抽取一个容量为36的样本，求样本均值 $\overline{X}$ 落在50.8到53.8之间的概率.

解 $$\overline{X} \sim N(52, \frac{6.3^2}{36})$$

故 $$P\{50.8 < \overline{X} < 53.8\} = F_{\overline{X}}(53.8) - F_{\overline{X}}(50.8)$$

$$= \Phi\left(\frac{53.8-52}{\frac{6.3}{6}}\right) - \Phi\left(\frac{50.8-52}{\frac{6.3}{6}}\right)$$

$$= \Phi(1.7143) - \Phi(-1.1429)$$

$$= 0.8239$$

例5　设总体X的概率密度函数为

$$f(x)=\begin{cases}|x| & |x|<1\\ 0 & |x|\ge 1\end{cases}$$

$(X_1, X_2, \cdots, X_{50})$ 为总体的样本，求

(1) $\overline{X}$ 的数学期望与方差　**(2)** $E(S^2)$

(3) $P\{|\overline{X}|>0.02\}$

解 (1) $E(\overline{X})=E(X)=\int_{-1}^{1} x|x|\mathrm{d}x=0$

$$D(\overline{X})=\frac{1}{50}D(X)=\frac{1}{50}E(X^2)$$

$$=\frac{1}{50}2\int_0^1 x^2|x|\mathrm{d}x=\frac{1}{100}$$

$$(2)\ E(S^2) = D(X) = E(X^2) = \frac{1}{2}$$

(3) $\overline{X} \sim N(0, \frac{1}{100})$ （近似）,由中心极限定理

$$P\{|\overline{X}| > 0.02\} = 1 - P\{|\overline{X}| \le 0.02\}$$

$$= 2\left(1 - \Phi\left(\frac{0.02 - 0}{0.1}\right)\right)$$

$$= 2(1 - \Phi(0.2))$$

$$= 0.8414$$