第六章 非参数概率密度函数的估计

参数估计：假设数据服从某种分布

非参数估计：直接利用样本估计分布

参数估计的局限：

- **参数估计方法**：要求**已知**分布的具体形式
- **实际情况可能是**：不知道分布的形式或者不是典型的分布形式

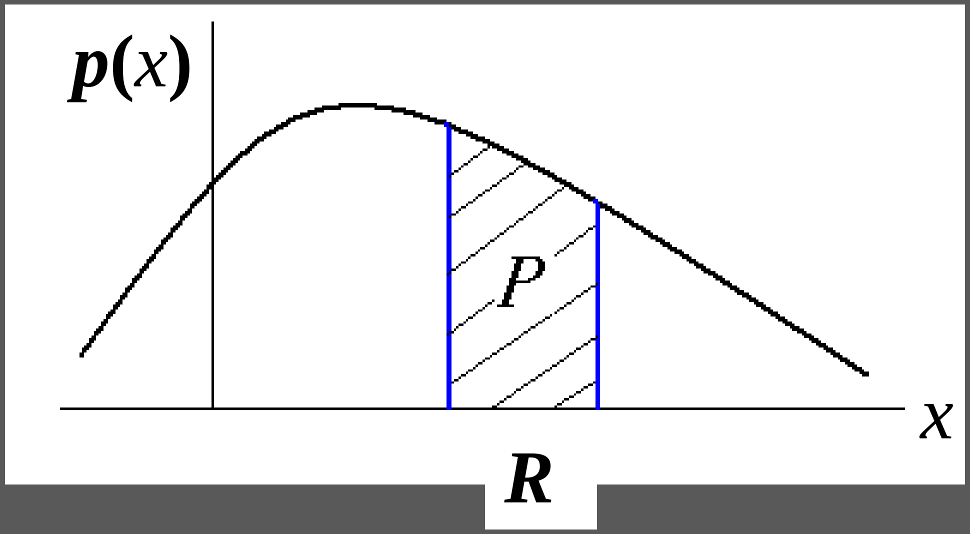
基本原理

Parzen窗法

k_N 近邻法

基本原理

设 $p(x)$是 x 的总体概率**密度函数**



随机变量 x 落入区域 R 的概率:

$$P = \int_R p(x)dx$$

k 的期望： $\boxed{E[k] = NP}$

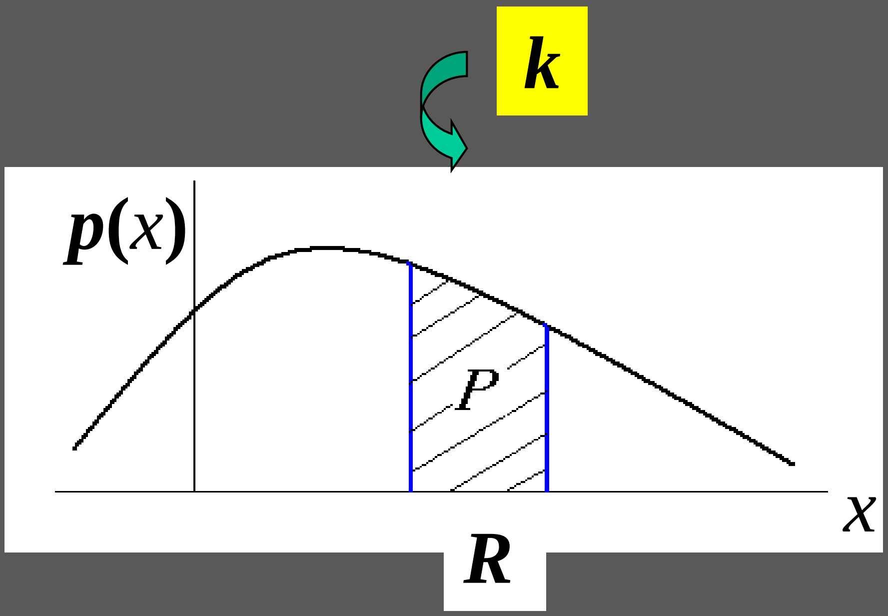
$$E[k] = NP$$

得出

$$k \approx N\widehat{P}$$

则

$$\widehat{P} \approx \frac{k}{N}$$



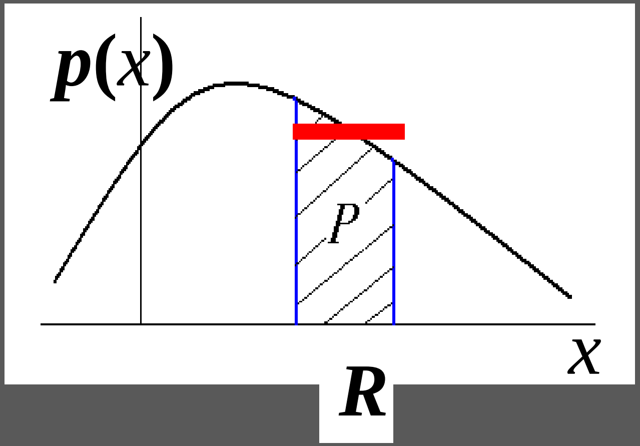
意义：k / N 是 P 的一个好的估计

我们的目标是估计概率密度函数 $p(x)$ 本身

假设 $p(x)$ 是连续函数，并且取区域 R 足够小，这样可以认为 $p(x)$ 在区域 R 上保持不变，则

概率

$$P = \int_R p(x)dx = p(x)V$$



其中，V 是区域 R 的体积，x 是 R 中的某一个点

$$\hat{P} \approx \frac{k}{N}$$

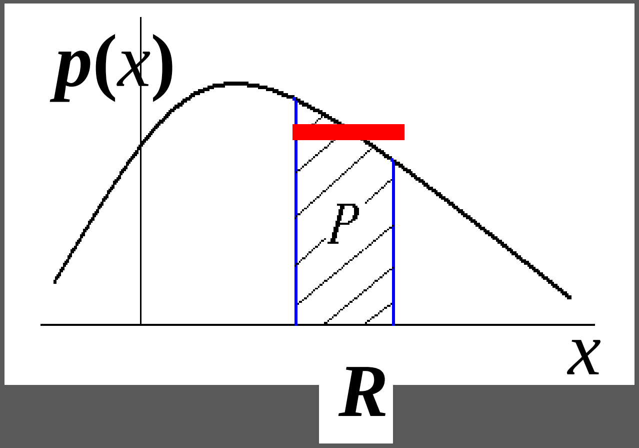
$$\hat{P} = \int_R p(x)dx = \hat{p}(x)V$$

点 x 的概率密度 p(x) 的估计值

$$\hat{p}(x) = \frac{k/N}{V}$$

它依赖于：

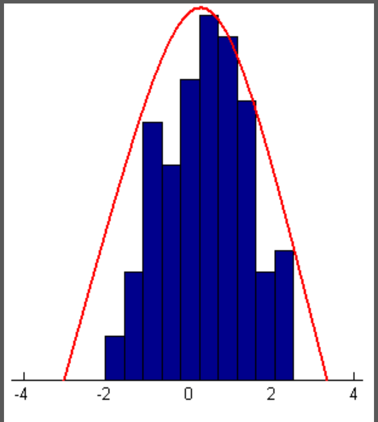
- ✓ 样本数目*N*
- ✓ 包含x的区域 *R* 的体积*V*
- ✓ 落入 区域 *R* 的样本数*k*



分析公式的性质

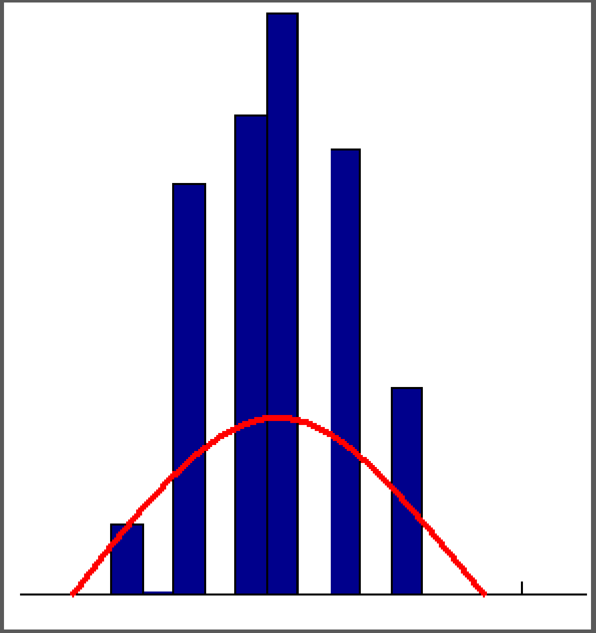
$$\hat{p}(x) = \frac{k/N}{V}$$

情形1：体积 V 固定。**理论上**，N 越来越多，则比值 $k/(NV)$ 按概率意义收敛于概率。**实际上**，只能得到 $p(x)$ 的空间平均估计：



$$p(\mathrm{x}) \approx \frac{k}{NV} = \frac{\overset{\frown}{P}}{V} = \frac{\int_R p(\mathrm{x})d\mathrm{x}}{\int_R d\mathrm{x}}$$

情形2：样本数固定。体积 V 趋近于零，随着区域 R 的不断缩小，一种可能性是出现不包含任何样本，从而使 $p(x)=0$；另一种可能性是包含几个样本，但是 $p(x)=\infty$



非常不光滑

<u>情形3</u>：(实际情况），样本数是有限的，体积不可能足够小。此时，k/N 和 $p(x)$ 存在随机性，即它们有一定的方差

假设可以使用无限多样本，从理论上分析公式的性质

估计的方法：为了估计点 **x** 的概率密度，构造一列包含 **x** 的区域序列 $R_1, R_2, \ldots, R_N, \ldots$

其中，对于 R_1 用一个样本来估计，R_2 用两个样本来估计，依次类推

记：

V_N：R_N的体积

k_N：落入区域 R_N 的样本数

$\hat{p}_N(\mathbf{x})$：$p(\mathrm{x})$的第 N 次估计

则

$$\hat{p}_N(\mathbf{x}) = \frac{k_N / N}{V_N} \rightarrow p(x)$$

如果它要收敛于 $p(\mathrm{x})$，则需要满足三个条件

需满足的三个条件：

(1) $$\boxed{\lim_{N\to\infty} V_N = 0}$$

<u>**含义**</u>：

如果区域平滑地缩小，$p(x)$ 在点 x 连续，这一条件使空间平均 P/V 收敛于概率 $p(x)$

(2) $$\boxed{\lim_{N \to \infty} k_N = \infty}$$

含义:

对于 $p(x) \neq 0$ 的点有意义，可以使频度 k_N/N 在

概率意义上收敛于概率 P

(3) $$\lim_{N\to\infty}\frac{k_N}{N}=0$$

尽管在一个小区域 R_N 内落入了大量样本，但是与总样本数相比仍然**非常少**

现在，有**两种**方法，可以选取满足上述**三个**条件的区域序列，从而导出**两种**非参数估计方法：

选定

$$\hat{p}_N(\mathbf{x}) = \frac{k_N/N}{V_N}$$

选定

k_N近邻估计：体积可变，落入其中的样本数不变

Parzen窗法：体积不变，落入其中的样本数可变

Parzen窗法：使区域序列V_N 为 N 的某个函数, 对于N*较*大的数据集, V_N *的取值应该较小*

k_N近邻估计：使 k_N 为 N 的某个函数，V_N 定义为刚好包含 **x** 的 k_N 个近邻样本的 区域 R_N 的体积

Parzen窗法

1 Parzen窗估计的基本概念

2 估计量 $\hat{p}_N(\mathbf{x})$ 为概率密度的条件

3 窗函数的选择

4 窗宽对估计量的影响

5 估计量的性质

1 Parzen窗估计的概念

估计概率密度的基本公式

$$\hat{p}_N(\mathbf{x}) = \frac{k_N / N}{V_N}$$

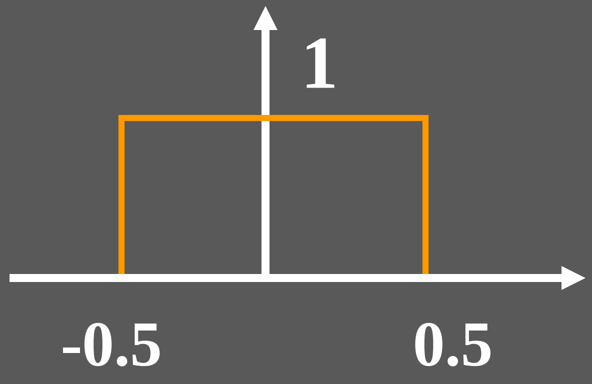
设区域 R_N 是以 h_N 为棱长的 d 维超立方体，则

立方体的体积为

$$V_N = h_N^d$$

定义一个窗函数：

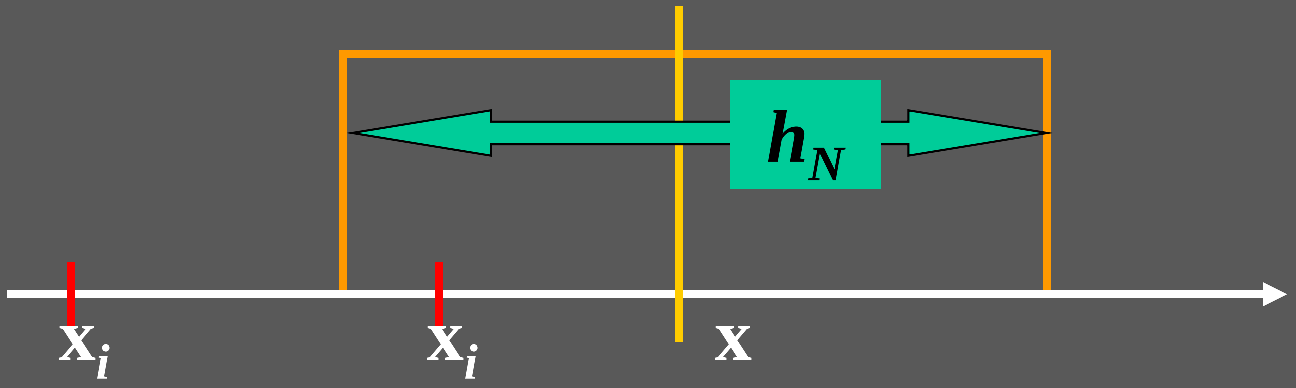
$$\varphi(\mathbf{u})=\begin{cases}1, & |u_j|\le \frac{1}{2},\ j=1,\ldots,d\\ 0, & \text{其他}\end{cases}$$



上述函数也称为方窗函数

求出落入超立方体内的样本个数？

立方体：以 $\mathbf{x}$ 为中心、体积为 V_N 的立方体



如果某一个样本 x_i 落入该立方体，则 $\varphi\left(\frac{x-x_i}{h_N}\right)=1$

否则 $\varphi\left(\frac{x-x_i}{h_N}\right)=0$

落入该立方体的样本数为

$$k_N = \sum_{i=1}^{N} \varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)$$

点 **x** 的概率密度

$$\hat{p}_N(\mathbf{x}) = \frac{k_N / N}{V_N} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)$$

Parzen窗法估计的基本公式

说明：

1 可以取其他形式的**窗函数**

2 窗函数的可选定义：每一个样本对窗函数的贡献依赖于它到 x 的距离 （ x 为立方体的中心）

3 这种估计可以看成 x 与 x_i 的函数的**平均**

2 估计量是密度函数的条件

概率密度函数应该满足的条件：

$$(1)\quad p(\mathbf{x}) \geq 0, \mathbf{x} \in \Re^d$$

$$(2)\quad \int_{\Re^d} p(\mathbf{x})d\mathbf{x} = 1$$

$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)$$

这一估计量是概率密度吗？

只要窗函数满足下列条件

$$(1)\quad \varphi(\mathrm{u}) \geq 0$$

$$(2)\quad \int \varphi(\mathrm{u}) d\mathrm{u} = 1$$

（意义：窗函数本身具有概率密度函数的形式）

$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)$$ 就是**概率密度函数**

证明：满足条件(1)

因为 $\boxed{\varphi(\mathbf{u}) \geq 0}$

则 $$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right) \geq 0$$

满足条件(2)

$$\int \hat{p}_N(\mathbf{x})d\mathbf{x} = \int \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)d\mathbf{x}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)d\mathbf{x}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \varphi(\mathbf{u})d\mathbf{u}$$

$$= \frac{1}{N}N$$

$$= 1$$

$$V_N = h_N^d$$

3 窗函数的选择

基本原则：窗函数必须满足两个基本条件

(1) $\varphi(u) \geq 0$

(2) $\int \varphi(u)du = 1$

常用的窗函数有：

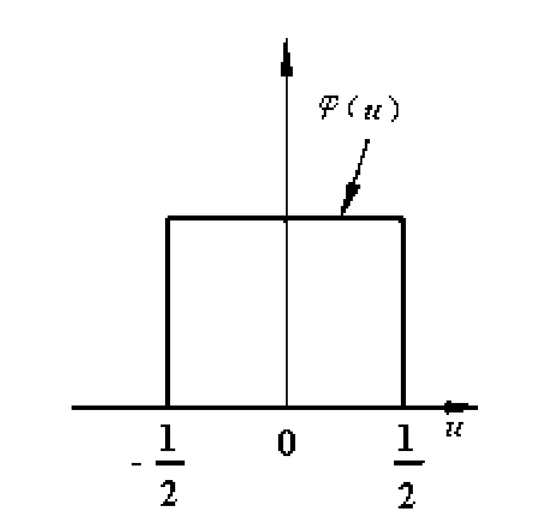
(1) 方窗函数

(2) 正态窗函数

(3) 指数窗函数

方窗函数（前面已介绍）

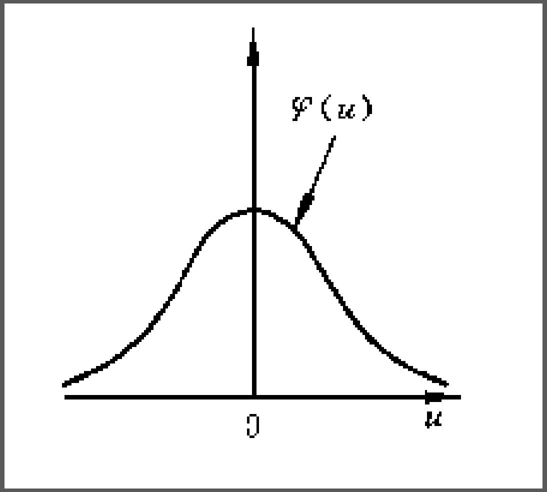
$$\varphi(\mathbf{u}) = \begin{cases} 1, & |u_j| \le \frac{1}{2},\ j=1,\ldots,d \\ 0, & \text{其他} \end{cases}$$



一维的图形

正态窗函数

$$\varphi(\mathbf{u}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{u}\right)$$

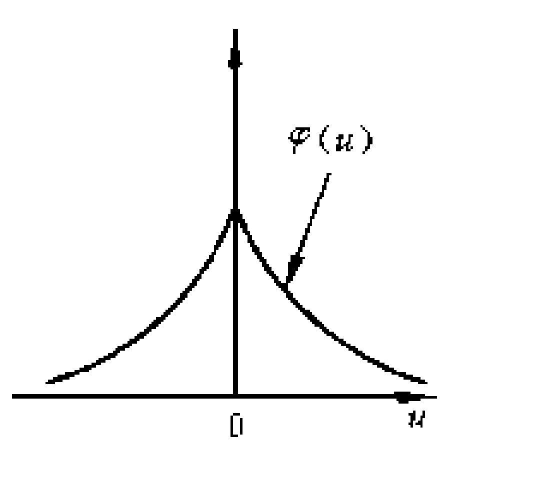


一维图形

指数窗函数

向量的长度或者模

$$\varphi(u)=\exp\left(-\frac{1}{2}|u|\right)$$



一维图形

$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)$$

$$V_N = h_N^d$$

说明：

对于一个实际问题，估计结果的**好坏**取决于窗函数的类型及**参数**（**窗宽**）

4 窗宽对估计量的影响

问题：

只有有限个样本数的条件下，讨论窗宽对估计量的影响

定义一个新函数 $$\delta_N(\mathbf{x}) = \frac{1}{V_N}\varphi\left(\frac{\mathbf{x}}{h_N}\right)$$

则

$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\delta_N(\mathbf{x}-\mathbf{x}_i)$$

取平均

$$V_N = h_N^d$$

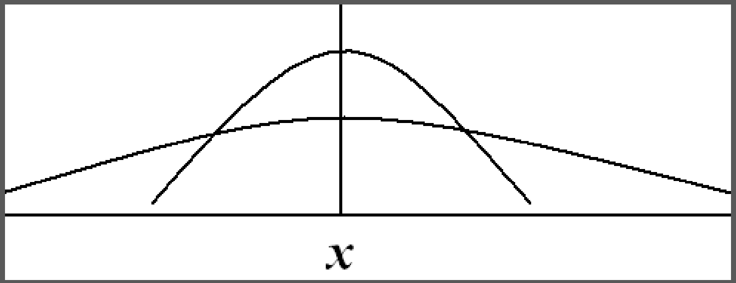
$$\delta_N(x) = \frac{1}{V_N}\varphi\left(\frac{x}{h_N}\right)$$

窗宽 h_N 的大小会影响 $\delta_N(x)$ 的幅度大小，两者之间成**反比**关系

$$V_N = h_N^d$$

$$\delta_N(\mathbf{x}) = \frac{1}{V_N}\varphi\left(\frac{\mathbf{x}}{h_N}\right)$$

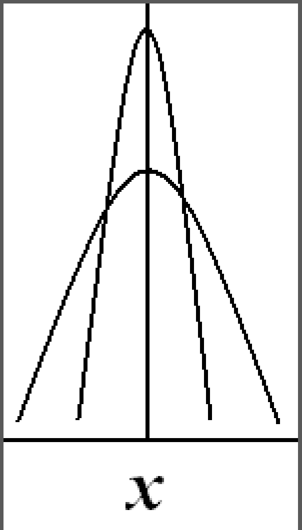
$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\delta_N(\mathbf{x}-\mathbf{x}_i)$$



情形 1：窗宽 h_N 很大，此时$\delta_N(x\text{-}x_i)$ 的幅度很小，并且只有离 x 很远的样本对概率密度才无贡献。因此，密度估计值是 N 个宽度较大、变化缓慢的函数的叠加，这是一个概率密度平均的估计，其分辨率很低----平均效应较大

$$V_N = h_N^d$$

$$\delta_N(\mathbf{x}) = \frac{1}{V_N}\varphi\left(\frac{\mathbf{x}}{h_N}\right)$$



$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\delta_N(\mathbf{x}-\mathbf{x}_i)$$

<u>情形 2</u>：窗宽 h_N 很小，此时$\delta_N(\mathbf{x}-\mathbf{x}_i)$ 的幅度很大。此时，密度估计值是 N 个以样本为中心的尖峰函数的叠加，变化剧烈。当 $h_N \rightarrow 0$ 时，$\delta_N(\mathbf{x}-\mathbf{x}_i)$ 趋于以 $\mathbf{x}_i$ 为中心的 δ 函数，估计量变成样本为中心的 δ 函数的叠加----不平滑效应明显

- 因为样本数是有限的，对于实际问题，我们只能折中考虑
- 一般而言，样本数目较多时，窗宽可以取小一些

5 估计量的统计性质

$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)$$

当满足下列三个基本条件时，**估计量**是渐近无偏和平方误差一致的：

(1) 总体密度 $p(\mathbf{x})$ 在点 $\mathbf{x}$ 处连续

(2) 窗函数满足下列关系式

$$\varphi(\mathbf{u}) \geq 0$$

$$\int \varphi(\mathbf{u})d\mathbf{u} = 1$$

保证估计量是密度函数

$$\sup_{u} \varphi(\mathbf{u}) < \infty$$

窗函数有界

$$\lim_{\|\mathbf{u}\| \to \infty} \varphi(\mathbf{u}) \prod_{i=1}^{d} u_i = 0$$

窗函数随着 u 的增加而迅速趋于零

(3)窗宽满足下列条件

$$\lim_{N\to\infty} V_N = 0$$

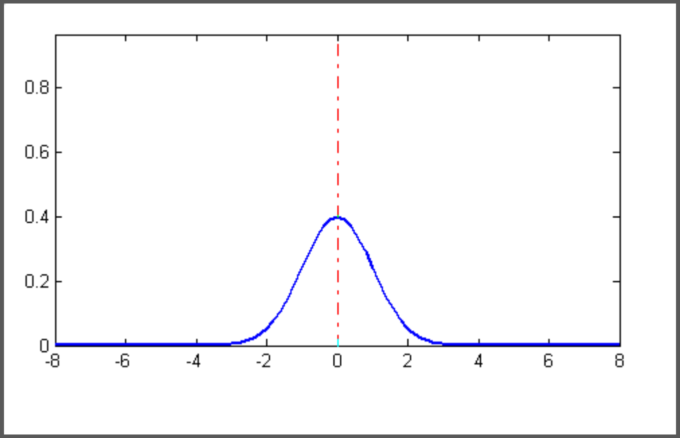
$$\lim_{N\to\infty} NV_N = \infty$$

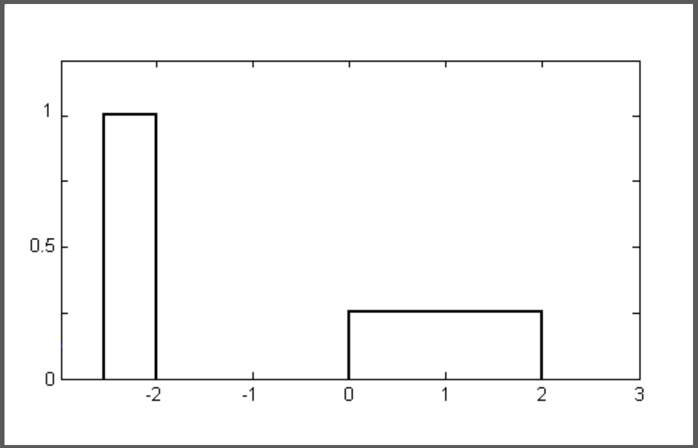
意义：体积随着 N 的增大而趋于零，缩减的速度不能太快，其速率低于 $1/N$

6 实例

(1) 标准正态分布的密度函数 $N(0, 1)$

(2) 两个均匀分布的混合密度 $$p(x)=\begin{cases}1, -2.5 < x < -2\\ 0.25, 0 < x < 2\\ 0, 其他\end{cases}$$





窗函数为正态窗函数

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^{2}\right)$$

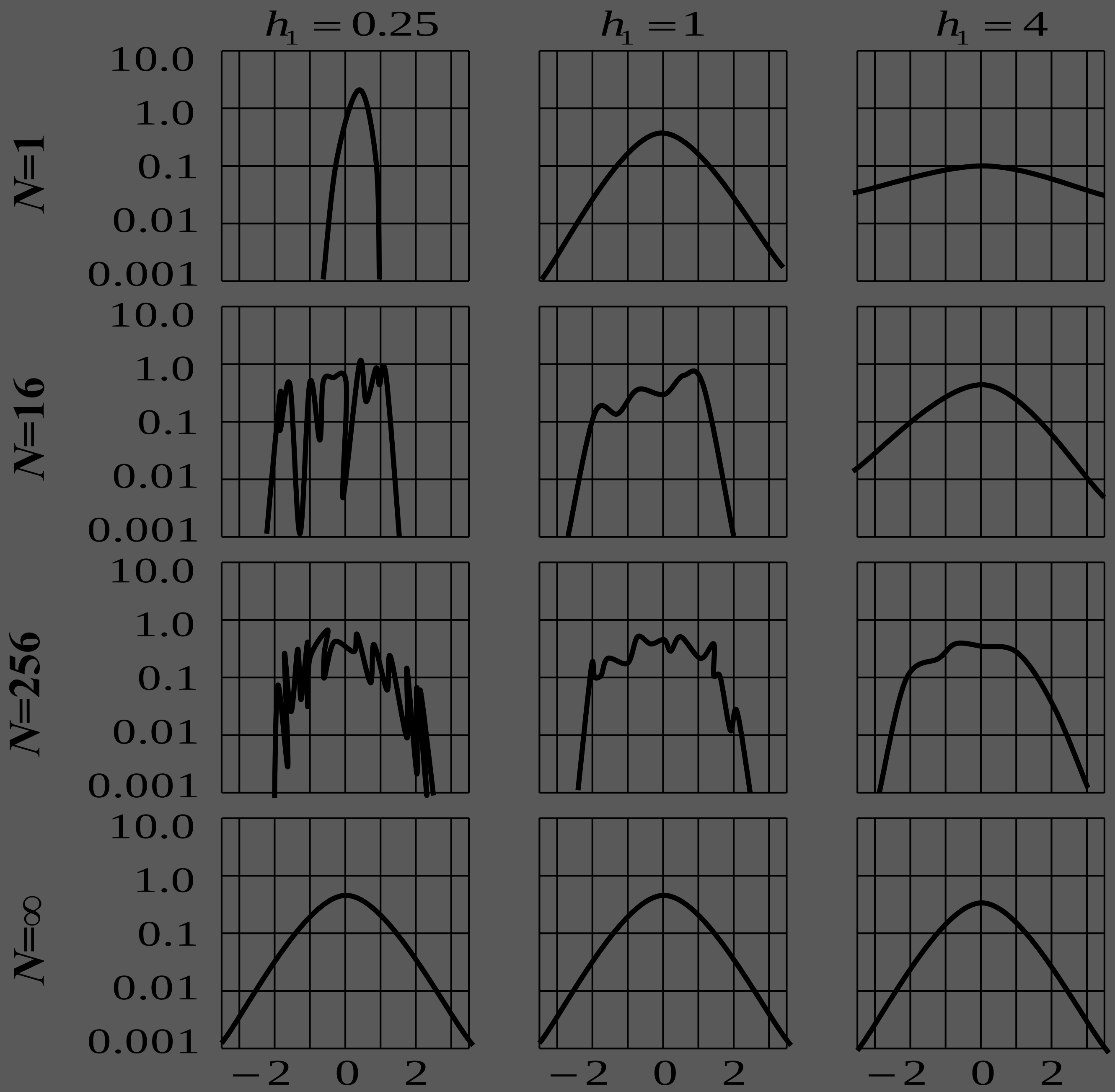
窗宽为

$$h_N = h_1 \big/ \sqrt{N}$$

其中 h_1 是可以调节的参数（使用者可以选取）

$$\hat{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_N}\varphi\left(\frac{x-x_i}{h_N}\right)$$

考察估计量与样本数 N、可调参数 h_1 的关系



$h_1 = 0.25$
$h_1 = 1$
$h_1 = 4$
N=1
N=16
N=256
N=∞
10.0
1.0
0.1
0.01
0.001
−2
0
2

讨论：$p_N(x)$ 随N, h_1 的变化情况

①当$N=1$时，$p_N(x)$是一个以第一个样本为中心的正态形状的小丘，就是窗函数本身

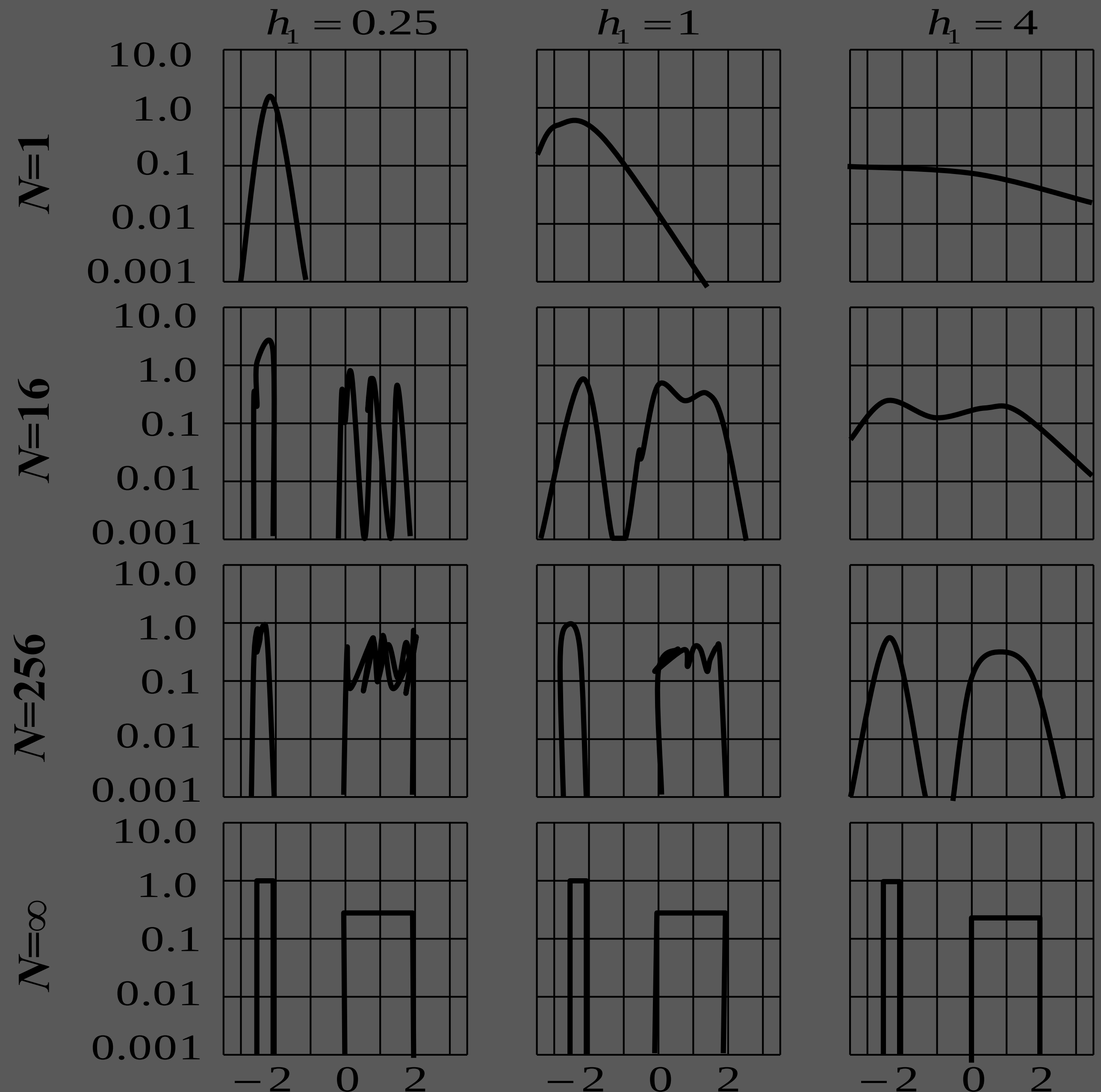
②当 $N=16$ 及 $N=256$ 时

$h_1=0.25$　曲线起伏很大，噪声大

$h_1=1$　起伏减小

$h_1=4$　曲线平坦

③当$N\to\infty$时，$p_N(x)$ 收敛于一平滑的正态曲线，得到了较精确的估计



$h_1 = 0.25$
$h_1 = 1$
$h_1 = 4$
N=1
N=16
N=256
N=∞
10.0
1.0
0.1
0.01
0.001
−2
0
2

①当 $N=1$ 时， $p_N(x)$ 实际是窗函数

②当 $N=16$ 及 $N=256$ 时

$h_1=0.25$ 曲线起伏大

$h_1=1$ 曲线起伏减小

$h_1=4$ 曲线平坦

③当 $N\to\infty$ 时，曲线较好

Parzen窗法的优缺点：

①优点：普遍适应。对规则分布，非规则分布，单锋或多峰分布都可用此法进行密度估计

②缺点：要求样本足够多，才能有较好的估计，计算量大，存储量大

使用Parzen窗法时存在的主要困难：

窗宽的选择非常重要，也非常困难。窗宽太大，得到一个较为平坦的分布，反映不出真实分布的变化。如果选得太小，估计量变化太大，很不稳定

3.5.3 k_N近邻估计

Parzen窗法的基本思想：固定体积，每一体积内包含不同数量的样本

k_N近邻估计基本思想：固定每一个体积内所包含的样本个数，而体积本身的大小可以变化

具体做法:

对于给定的 N 个样本，我们事先确定 N 的某一个函数 k_N，表示**样本个数**

在点 $\mathbf{x}$ 的周围，选择一个体积，不断地扩大，直到捕捉到 k_N 个样本为止，这些样本称为 $\mathbf{x}$ 的k_N个近邻

估计点 $\mathbf{x}$ 的概率密度为

$$\hat{p}_N(\mathbf{x}) = \frac{k_N / N}{V_N}$$

优点:

（1）如果点 x 附近的密度比较高，则包含k_N个样本的体积就比较小，可以提高分辨率

（2）反之，如果点 x 附近的密度比较低，则体积就比较大，分辨率较低

基本估计公式：

$$\hat{p}_N(\mathbf{x}) = \frac{k_N / N}{V_N}$$

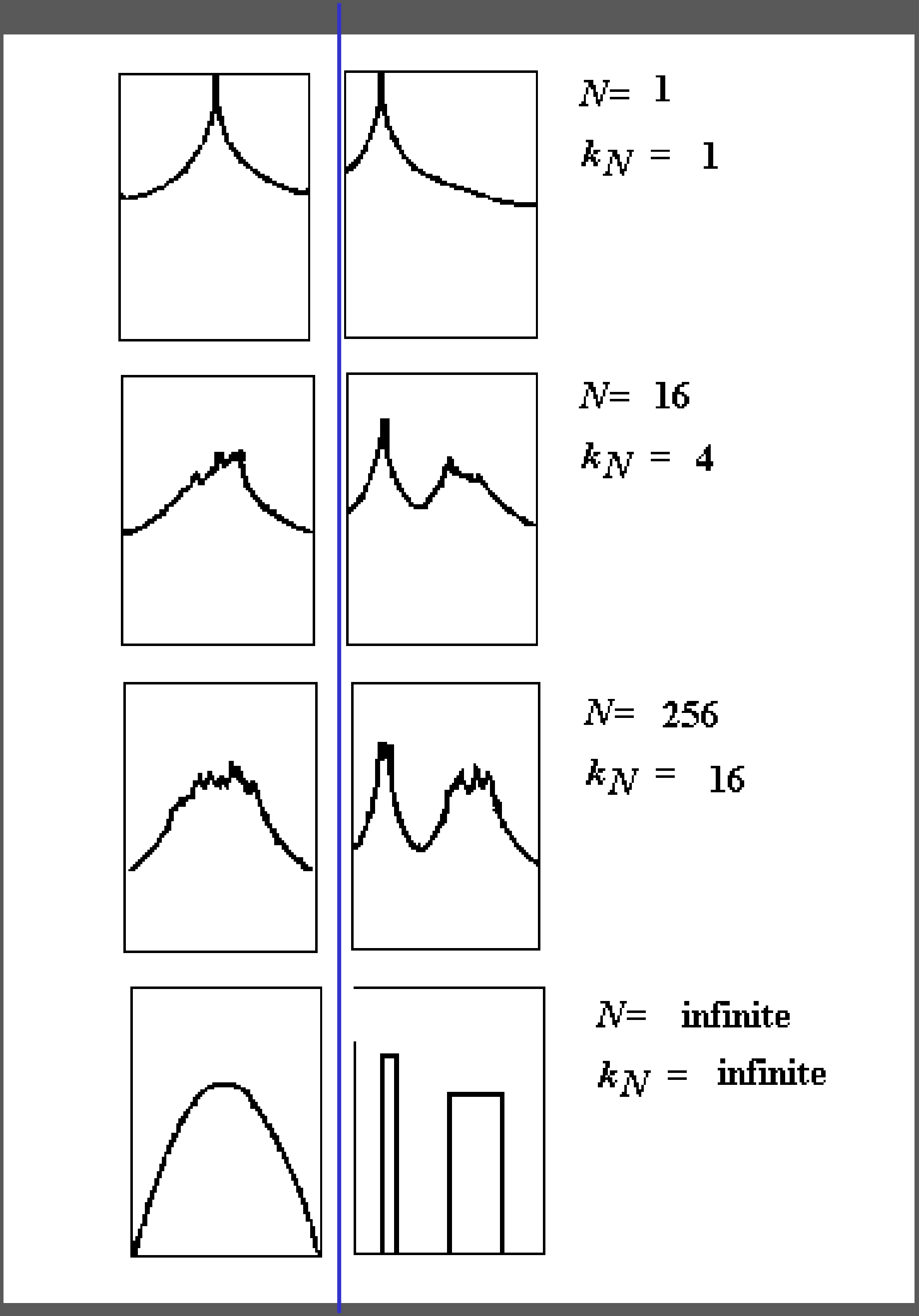
满足的条件是

$$\lim_{N\to\infty} V_N = 0$$

$$\lim_{N\to\infty} k_N = \infty$$

$$\lim_{N\to\infty} \frac{k_N}{N} = 0$$

标准
正态
分布



混合
分布

k_N近邻法的缺点：

- 需要样本多
- 计算量大
- 存储量大