数据汇总

主讲人：刘宏志
liuhz@ss.pku.edu.cn

经验累积分布函数

$$(X_1, X_2, \cdots X_n) \text{———} (x_1, x_2, \cdots, x_n)$$

$$\downarrow$$

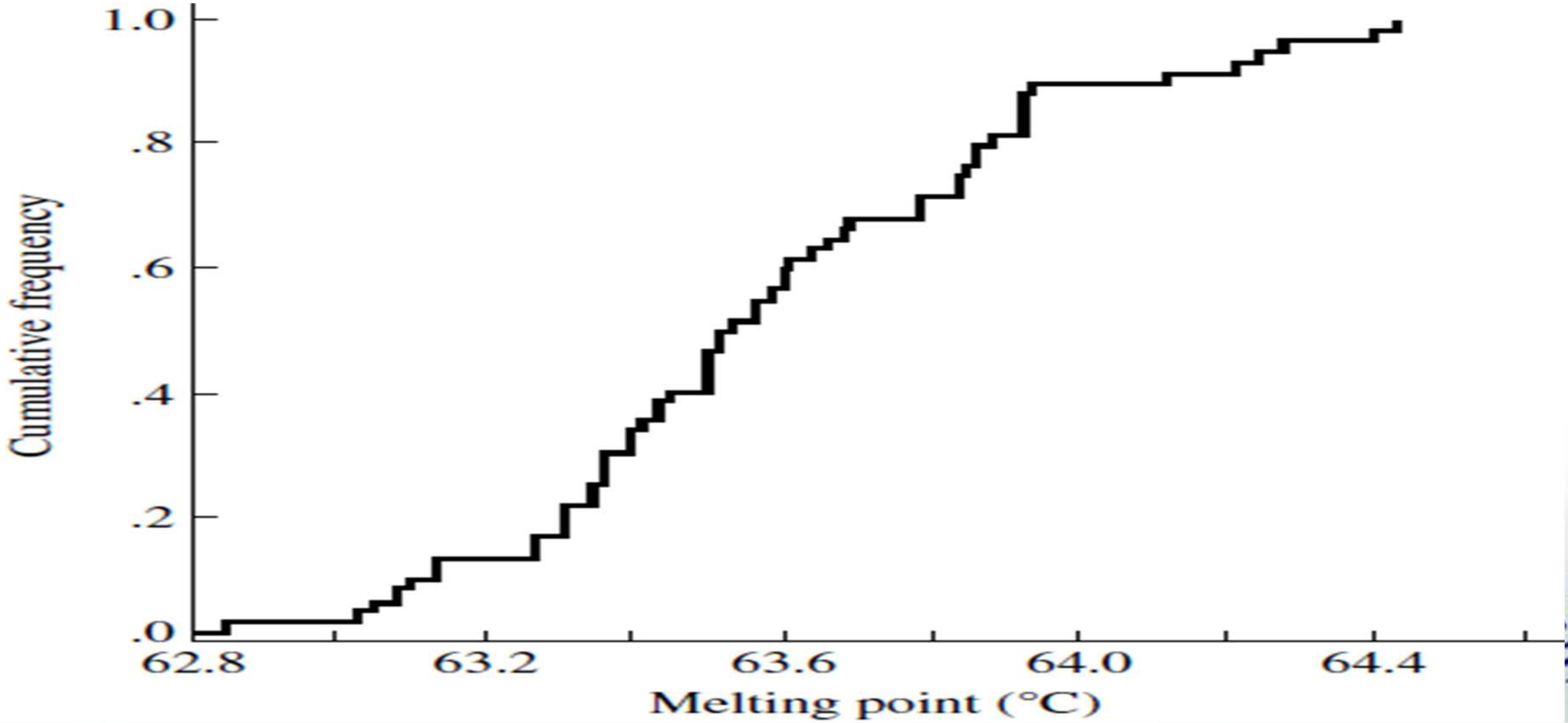
$$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$$

$$\text{定义函数} F_n(x) \overset{\Delta}{=} \begin{cases} 0, & \text{当} x < x_{(1)} \\ \vdots & \vdots \\ k/n, & \text{当} x_{(k)} \le x < x_{(k+1)} \\ \vdots & \vdots \\ 1, & \text{当} x \ge x_{(n)} \end{cases}$$

北京大学

经验分布函数$F_n(x)$的性质

- $0 \le F_n(x) \le 1$，$F_n(-\infty) = 0$，$F_n(+\infty) = 1$
- $F_n(x)$为非减函数
- $F_n(x)$在每个$x_{(i)}$处右连续，点$x_{(i)}$是跳跃点，跳跃度为该点的频率w_i
- $\forall \varepsilon > 0, \lim\limits_{n\to\infty} P(|F_n(x) - F(x)| \le \varepsilon) = 1$
- 格里汶科定理：$P\left\{\lim\limits_{n\to\infty} \sup\limits_{-\infty<x<+\infty} |F_n(x) - F(x)| = 0\right\} = 1$



生存时间函数

- 动机：
  - 医疗上，需要考虑各种药物或疗法的效果
  - 保险公司需要评估各种人群的寿命，以制定投保方案
  - 工程上，需要考虑材料（原件、设备等）的寿命
  - ……
- 生存分析：研究生存时间的分布规律以及生存时间和相关因素之间关系的一种统计分析方法
- 生存时间：从某起始事件到某终止事件经历的时间跨度
- 生存时间函数：描述生存时间分布规律的函数
  - 例如，生存函数、死亡函数、死亡密度函数、风险函数

生存时间数据类型

- 完全数据
  - 提供的关于生存时间的信息是完整确切的
  - 准确地度量了观察对象实际生存的时间
- 截尾数据
  - 提供的关于生存时间的信息是不完整不确切的
  - 没有准确地度量观察对象实际生存的时间
    - 在随访过程中某些观察对象失访
    - 死于其它原因
    - 在规定的研究过程结束时观察对象的终止事件还未发生

生存时间函数的估计

- 生存函数：$S(t) = P(\ T > t\ )$
  - 观察对象的生存时间T大于某时刻t的概率
  - 性质：$S(0)=1$， $S(\infty)=0$， 且$0 \leq S(t) \leq 1$

$$\hat{S}(t) = \frac{t\text{时刻尚生存的观察对象数}}{\text{观察对象总数}}$$

- 死亡函数： $F(t) = P(\ T \leq t\ )$
  - 观察对象的生存时间T不大于某时刻t的概率
  - 性质：$F(0)=0$， $F(\infty)=1$， 且$0 \leq F(t) \leq 1$

$$\hat{F}(t) = 1 - \hat{S}(t)$$

生存时间函数的估计

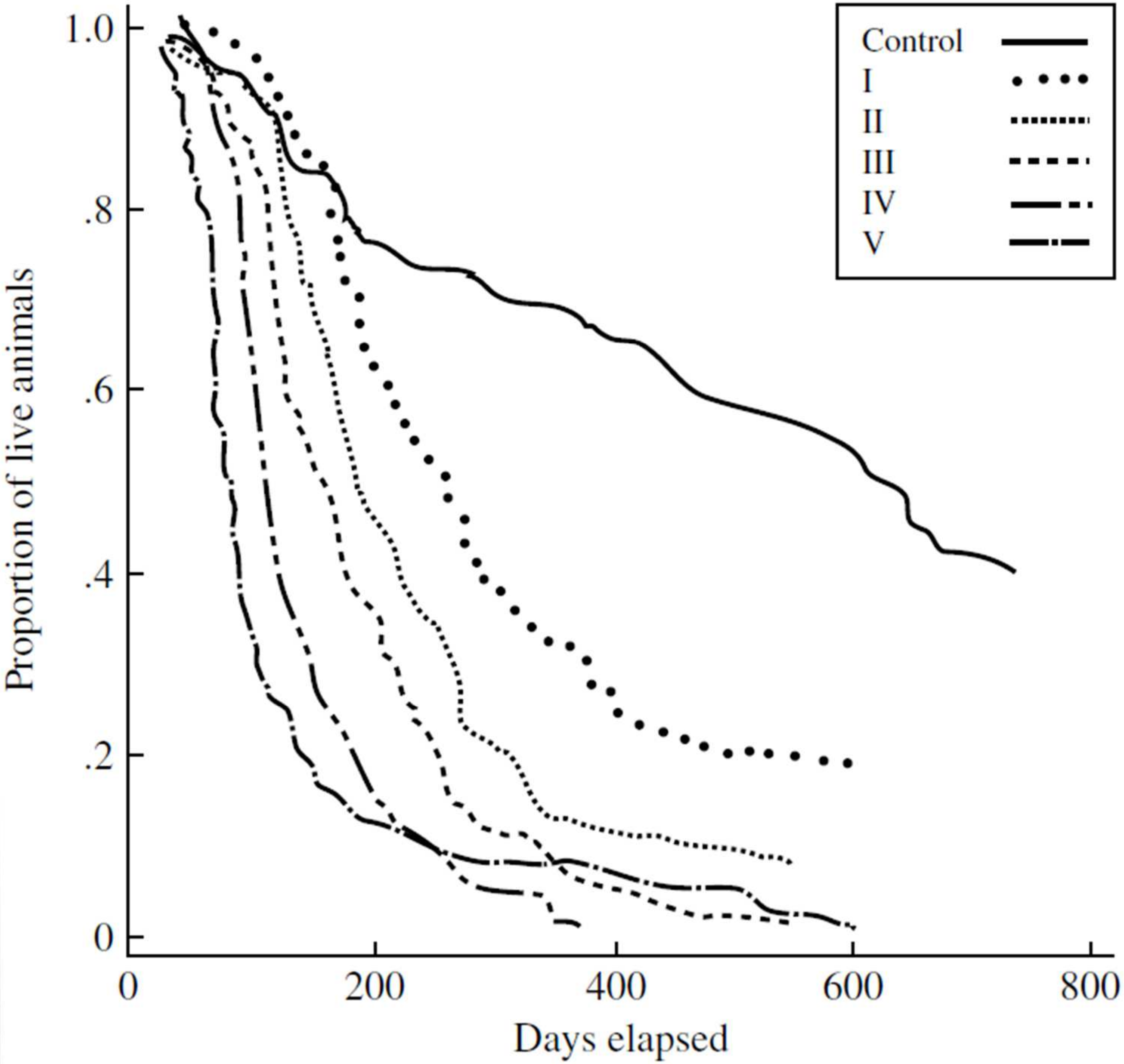
- 死亡密度函数：$f(t)=\lim_{\delta\to 0}\frac{F(t+\delta)-F(t)}{\delta}$
  - 观察对象在某时刻t 的瞬时死亡率

$$\hat{f}(t)=\frac{\text{观察对象在时间区间}[t,t+\Delta t]\text{内死亡数}}{\text{观察对象总数}\times\text{区间}[t,t+\Delta t]\text{所含单位时间数}}$$

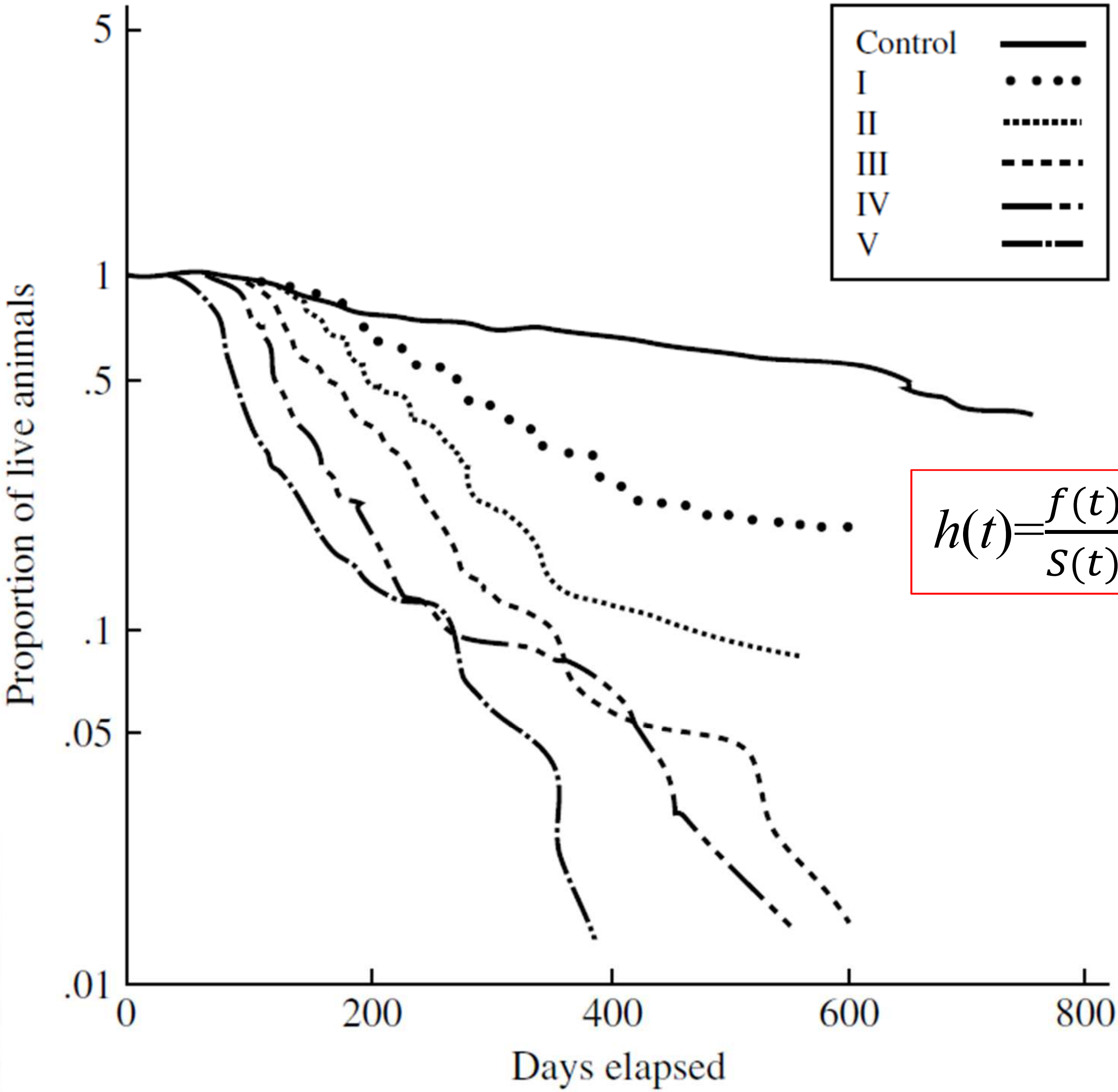
- 风险函数：　$h(t)=\frac{f(t)}{S(t)}$
  - 生存到时刻 t 的观察对象在时刻t 的瞬时死亡率

$$\hat{h}(t)=\frac{\hat{f}(t)}{\hat{S}(t)}=\frac{\text{观察对象在时间区间}[t,t+\Delta t]\text{内死亡数}}{t\text{时刻生存者数量}\times\text{区间}[t,t+\Delta t]\text{所含单位时间数}}$$

示例：豚鼠寿命的生存函数



示例：豚鼠寿命的对数生存函数



$$h(t)=\frac{f(t)}{S(t)}=\frac{d}{dt}[-\ln S(t)]$$

分位数-分位数图（Q-Q图）

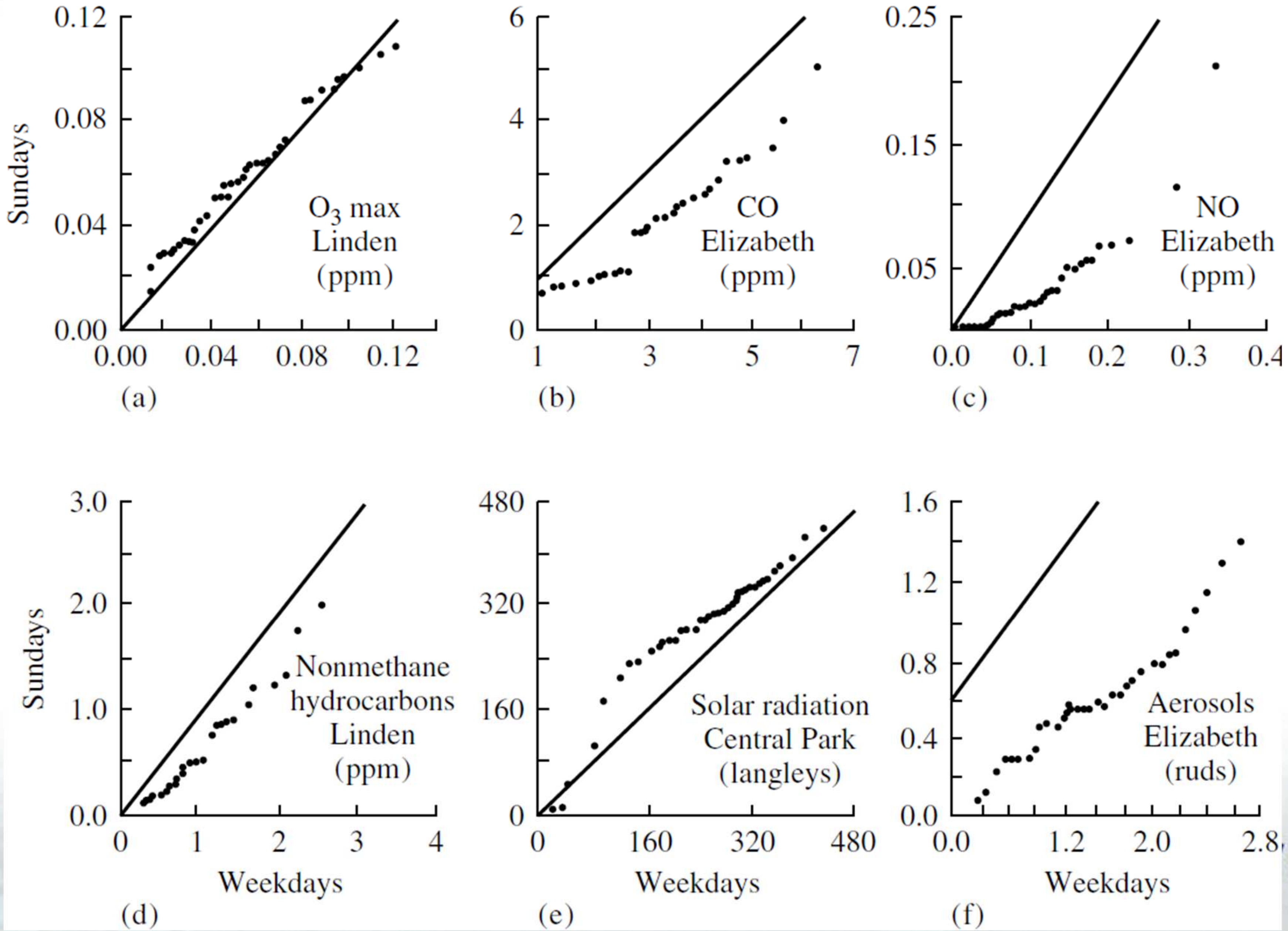
- 如果X是具有严格单增分布函数F的连续型随机变量，该分布的第p分位数为满足下式的x值：

$$F(x)=p$$

- $x_p=F^{-1}(p)$：连续分布函数中的一点，该点的一侧对应概率p
- Q-Q图：比较两个分布函数
  - 两组容量为n的数据，顺序统计量分别为$X_{(1)},\ldots,X_{(n)}$和$Y_{(1)},\ldots,Y_{(n)}$
  - 利用点对（$X_{(i)}, Y_{(i)}$）简单构造Q-Q图
- 性质：
  - 如果两个分布相似，则对应的Q-Q图趋近直线$y=x$
  - 如果两个分布线性相关，则对应的Q-Q图趋近一条直线

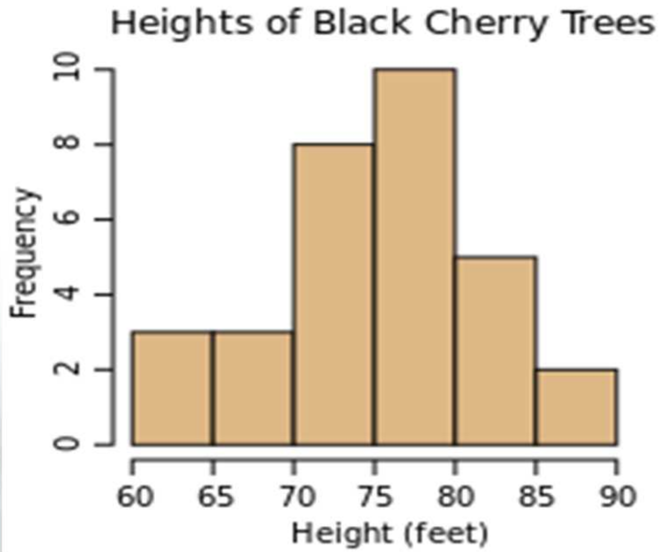
给定n个观测，顺序统计量为$X_{(1)},\ldots,X_{(n)}$，数据的$(k\text{-}0.5)/n$分位数分配给$X_{(k)}$

案例：周日和平日空气污染对比

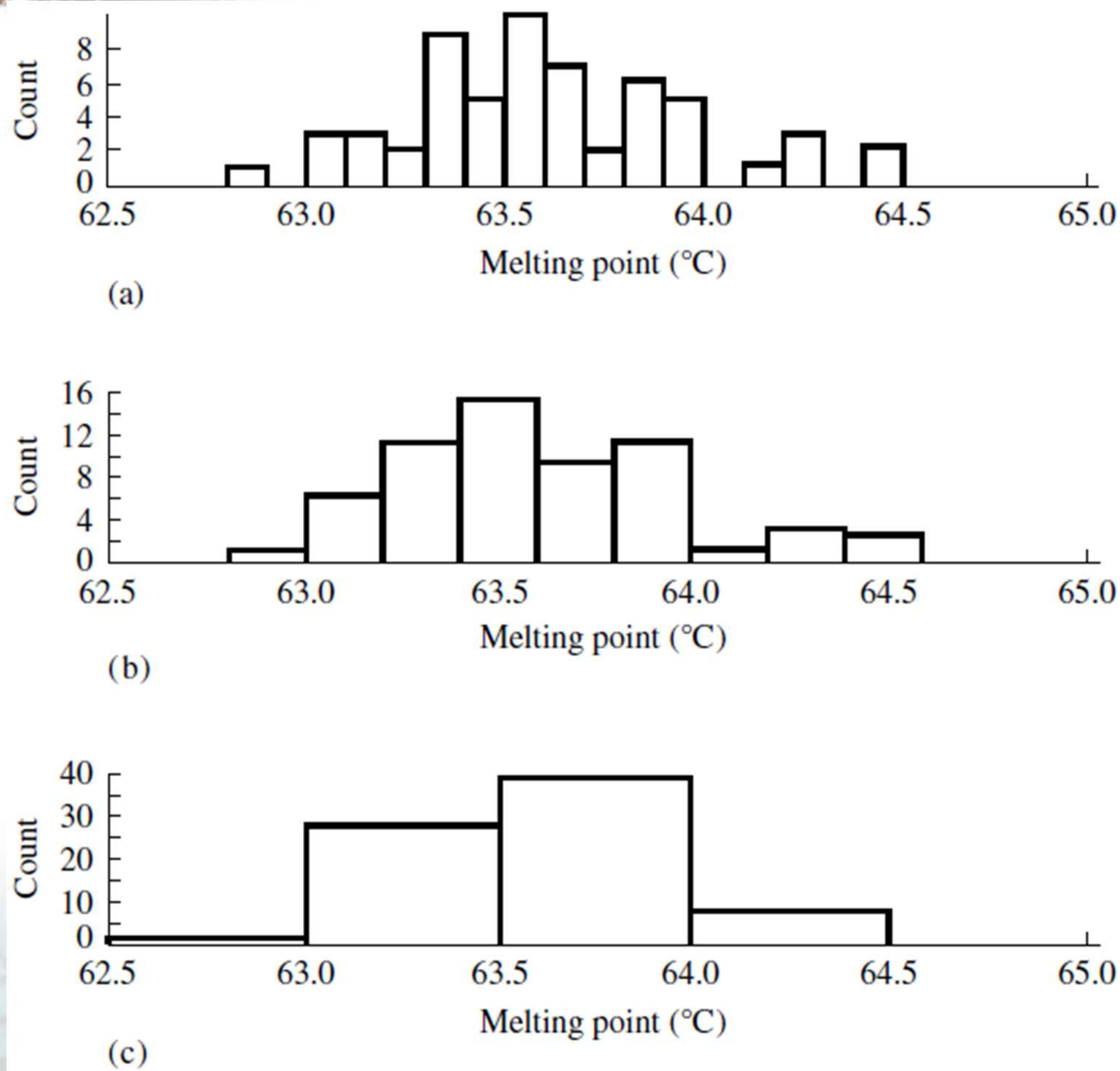


直方图

- 直方图：
  - 将数据区域划分成几个区间或频带，画出落入每个频带的观测数或比例
  - 常用于显示没有任何随机模型假设的数据图形
  - 展示数据分布形状的方式类似于密度函数显示概率
- 基本步骤：
  - 计算最大值与最小值的差(确定变动范围)
  - 决定频带宽度与频带数(将数据分组)
  - 决定分点
  - 列出频率分布表
  - 画出频率分布直方图

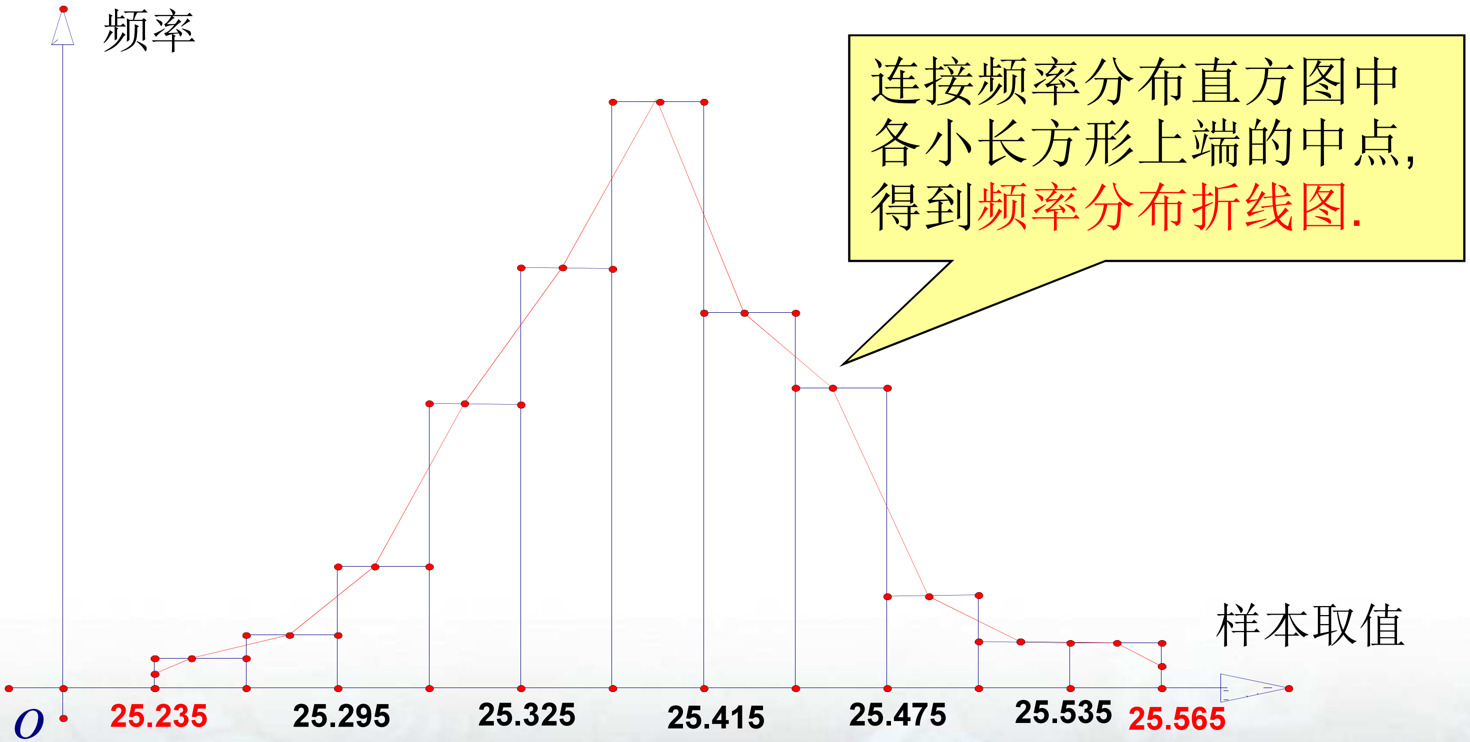


示例：蜂蜡熔点的直方图


Count
8
6
4
2
0
62.5
63.0
63.5
64.0
64.5
65.0
Melting point (°C)
(a)
Count
16
12
8
4
0
62.5
63.0
63.5
64.0
64.5
65.0
Melting point (°C)
(b)
Count
40
30
20
10
0
62.5
63.0
63.5
64.0
64.5
65.0
Melting point (°C)
(c)

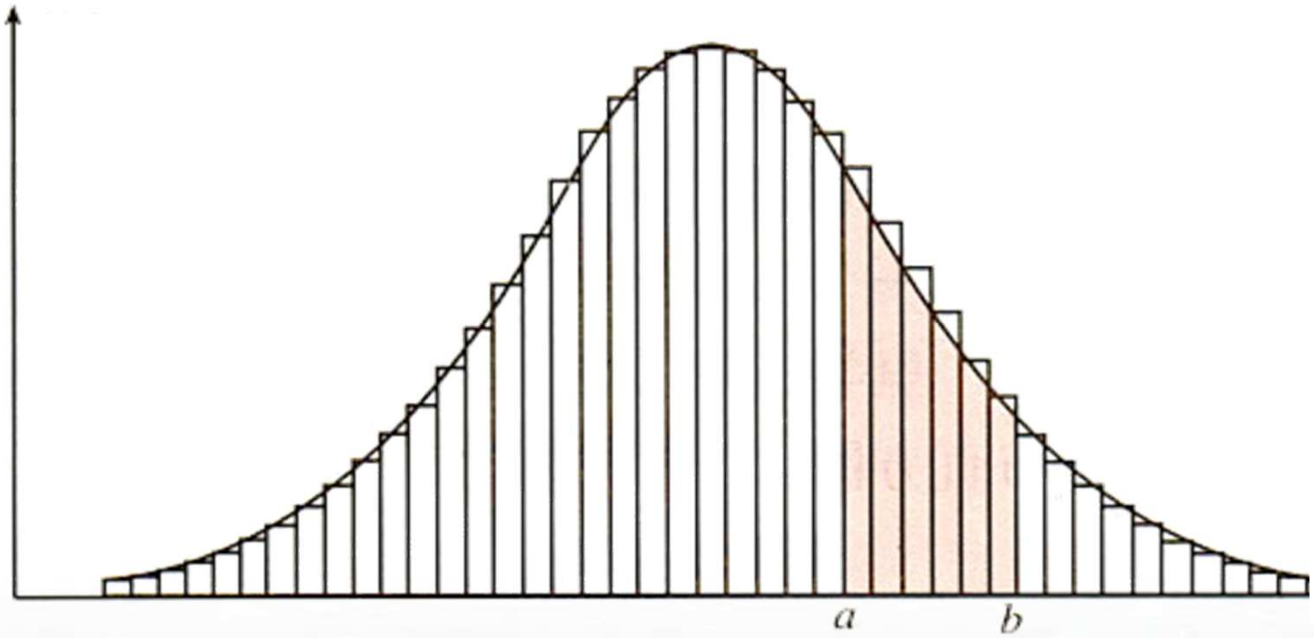
总体密度曲线

- 样本容量越大，所分组数越多，各组的频率就越接近于总体在相应各组取值的概率
- 如果样本容量无限增大，分组的组距无限缩小，那么频率分布直方图折线就会无限接近于一条光滑曲线——总体密度曲线



总体密度曲线

- 样本容量越大，所分组数越多，各组的频率就越接近于总体在相应各组取值的概率
- 如果样本容量无限增大，分组的组距无限缩小，那么频率分布直方图折线就会无限接近于一条光滑曲线——总体密度曲线



缺点：要求样本容量足够大，但实际中样本收集困难，样本容量有限

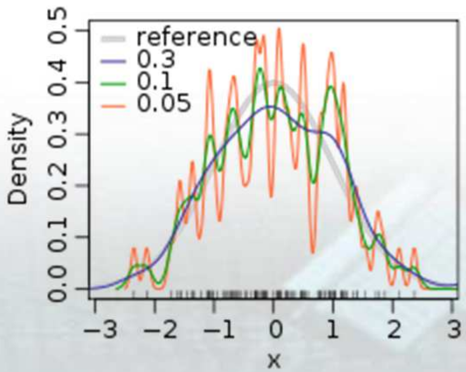
核概率密度估计

- 基本思想：
  - 如果某一个数在观察中出现了，可认为这个数的概率密度较大
  - 和这个数较近的数的概率密度也会较大，而那些远离这个数的数的概率密度会较小

- 估计方法：
  - 针对观察中的每个数，以$K_h(x-x_i)$拟合想象中的那个远小近大概率密度
  - 针对每个观察中出现的数拟合出多个概率密度分布函数，取平均

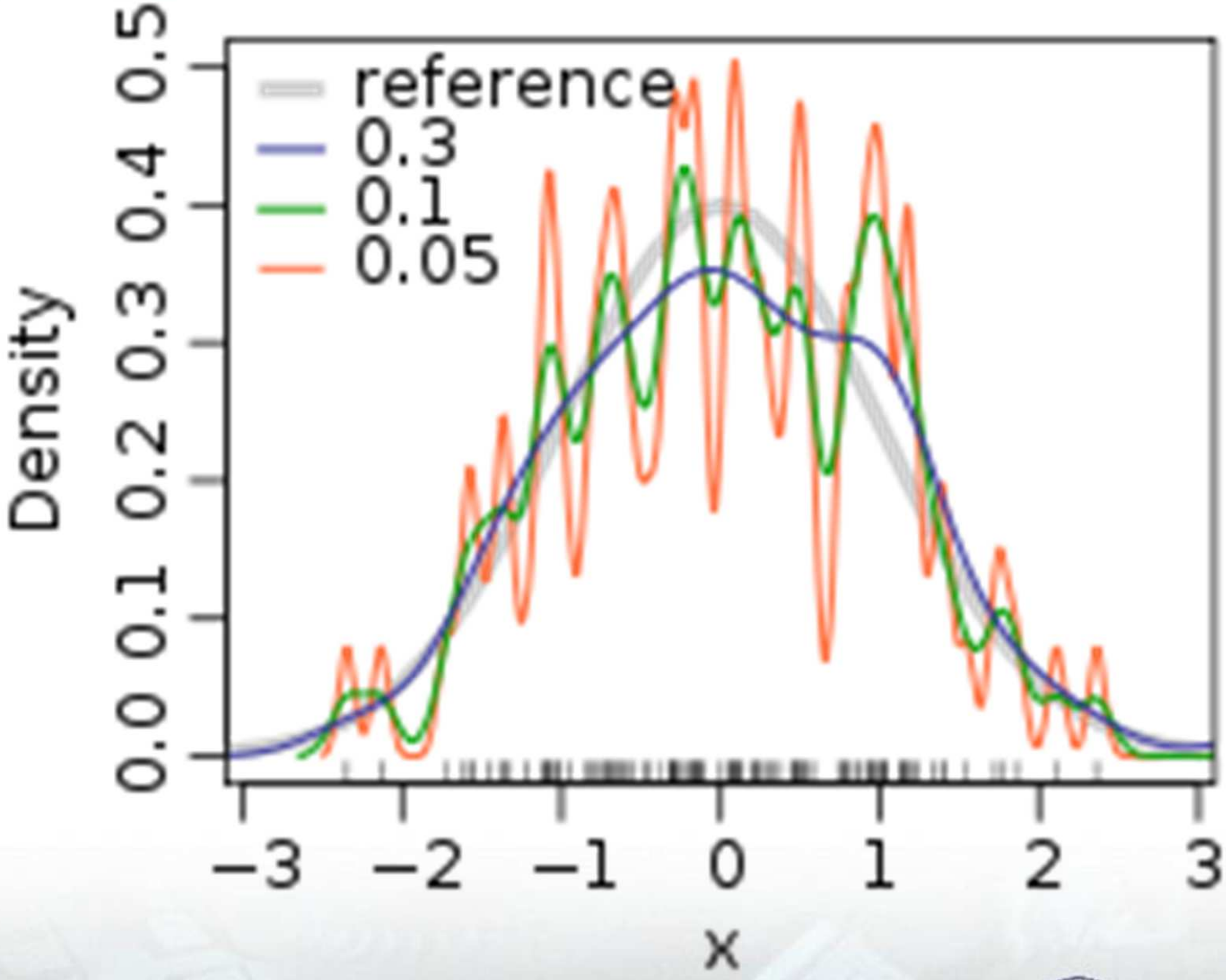
$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \quad = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big),$$

  - $K(.)$为核函数，通常取标准正态分布密度函数
  - 参数h为估计函数的带宽，控制着函数的光滑性



北京大学

示例：核概率密度估计



茎叶图

- 基本思想：
  - 将数组中的数按位数进行比较
  - 将数的大小基本不变或变化不大的位作为一个主干（茎）
  - 将变化大的位的数作为分枝（叶），列在主干的后面
  - 这样可清楚看到每个主干后面的几个数，每个数具体是多少
- 茎叶图vs.直方图
  - 茎叶图保留原始资料的信息，直方图则会丢失原始资料的信息
  - 将茎叶图茎和叶逆时针旋转90度，实际上就是一个直方图，可从中统计出次数，计算出各数据段的频率

		STEM	LEAF
1	1	628	:5
1	0	629	:
4	3	630	:358
7	3	631	:033
9	2	632	:77
18	9	633	:001446669
23	5	634	:01335
	10	635	:0000113668
26	7	636	:0013689
19	2	637	:88
17	6	638	:334668
11	5	639	:22223
6	0	640	:
6	1	641	:2
5	3	642	:147
2	0	643	:
2	2	644	:02

位置度量

- 位置度量是一组数据中心的测量值
- 基本思想：
  - 如果数据是同一个量不同的测量结果，利用位置度量来代替单个观测值，更精确地表示测量尺寸
- 常用的位置度量：
  - 算术平均、中位数、截尾均值、M估计

位置度量

- 算术平均：$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- 中位数：
  - 如果样本容量是奇数，中位数为顺序观测的中间值
  - 如果样本容量为偶数，中位数为两个中间值的平均
- 截尾均值：
  - 丢掉最小的100α%和最大的100α%观测数据
  - 计算剩余数据的算术平均
- M估计：
  - 当标的分布为正态时，样本均值是位置参数μ的最大似然估计

示例：铂的升华温度

Heats of Sublimation of Platinum (kcal/mol)

136.3	136.6	135.8	135.4	134.7	135.0	134.1	143.3
147.8	148.8	134.8	135.2	134.9	146.5	141.2	135.4
134.8	135.8	135.0	133.7	134.4	134.9	134.8	134.5
134.3	135.2						

- 算术平均：137.5
- 中位数：135.1
- 截尾均值(α =0.2)：135.29
- M估计：135.28

离群点(outliers): 偏离主体太远的观测

没有任何一个估计对所有分布都最好