概率统计 B
第五章 统计估值

原著：陈家鼎、刘婉如、汪仁官
制作：李东风，邓明华

2025 春季学期

本章目录

本节目录

随机变量模型与数据

- 随机变量是刻画随机结果及随机结果的取值分布情况的数学模型。
- 在实际问题中，假定某个随机结果的模型为随机变量 X，需要求 X 的分布（分布函数、分布密度或概率函数），至少要求其数字特征（期望、方差等）。
- 还可能需要回答 X 的有关问题。
- 这些需要数据的支持。

例 1.1

- 钢铁厂一天生产了 10000 根 16Mn 型钢筋。强度小于 52kg/mm^2 的算次品。
- 如何求这批产品的次品率 p?
- 检验所有 10000 根不可能：时间、费用、或破坏。
- 概率统计模型：设随机取一根，结果用随机变量 X 表示：

$$X = \begin{cases} 1 & \text{是次品} \\ 0 & \text{不是次品} \end{cases}$$

$$P(X = 1) = p$$

- 抽取少量样品来估计 p。

例 1.2

- 灯泡厂生产灯泡，寿命是随机的。求一批灯泡的平均寿命和寿命的分布情况。
- 不能全部检验：破坏性。
- 设随机抽取一只灯泡的寿命为随机变量 X。设 $X \sim \mathrm{Exp}(\lambda)$。求 λ 可以回答平均寿命以及寿命分布问题。

随机抽样法

- 从要研究的对象的全体中抽取一小部分来进行观察和研究，从而对整体进行推断。
- 重要意义：普查方法经常不可行，因为人力、物力、时间限制，或破坏性试验。
- **例 1.1（续）** 从 10000 根中抽取 50 根，对这 50 根进行检验。用 50 根的次品率作为所有全体的次品率的估计。
- 为什么这样是科学的?
- 随机抽样法包括：
  - 如何抽样，抽多少，怎样抽取；
  - 得到抽样结果（一批数据）后如何进行数据分析，进行统计推断。

总体

- 把所研究的对象的全体称为**总体**。
- 把总体中每一个基本单位称为**个体**。
- 主要关心每个个体的某一特性值（即数量指标，如钢筋的强度、灯泡的寿命）机器在总体中的分布情况（如钢筋强度在 50kg/mm^2 到 60kg/mm^2 之间的在 10000 根钢筋中所占的比例，灯泡寿命在 1000 小时到 2000 小时之间的占全天生产的灯泡的比例）。
- 要考察总体中个体特性值的分布规律，可以将个体特性值看成一个随机变量 X，代表从总体中随机抽取一个个体的特性值，其概率分布就可以体现总体中个体特性值的分布规律。
- 因为只关心总体的某个特性值，所以把总体认作是其特性值的随机变量 X。

样本

- 在一个总体（如 10000 根钢筋，或 10000 根钢筋的强度）X 中，抽取 n 个个体 $X_1, X_2, \ldots, X_n$, 这 n 个个体 $X_1, X_2, \ldots, X_n$ 称为总体 X 的一个容量为 n 的 **样本** (或叫**子样**)，也称 n 为**样本量**。
- 由于 $X_1, X_2, \ldots, X_n$ 是从总体 X 随机抽取出来的可能结果，可以看成是 n 个随机变量。
- 在一次抽取之后，又可以看成是 n 个具体的数值，称为**样本值**, 在使用这个意义时记为小写的 $x_1, x_2, \ldots, x_n$。
- 有时大小写也会混用。

样本的代表性

- 样本值应该对总体具有代表性。
- 如果样本 $X_1, X_2, \ldots, X_n$ 是相互独立的而且与总体 X 具有相同的概率分布，则称其为**简单随机样本**。
- 有放回逐次随机抽样法可以得到简单随机样本。当总体个数很大时，无放回地逐次随机抽样也可以认为是得到简单随机样本。
- 对总体 X 进行多次独立的重复观测，可以认为是得到简单随机样本。

总体与样本的数学模型

- 总体就是一个随机变量 X，我们关心其分布。
- 样本就是 n 个相互独立且与 X 有相同概率分布的随机变量 $X_1, X_2, \ldots, X_n$。
- 每一次具体抽样，所得的样本的值就是这 n 个随机变量的值（样本值），用 $x_1, x_2, \ldots, x_n$ 表示。
- **定义 1.1** 称随机变量 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的容量为 n 的 (简单随机) 样本，如果 $X_1, X_2, \ldots, X_n$ 相互独立，而且每个 X_i 与 X 有相同的概率分布。这时，若 X 有分布密度 $p(x)$，则称 $X_1, X_2, \ldots, X_n$ 是来自总体 $p(x)$ 的样本。
- **定理 1.1** 若 $X_1, X_2, \ldots, X_n$ 是来自总体 $p(x)$ 的样本，则 $(X_1, X_2, \ldots, X_n)$ 有联合密度

$$p(x_1)p(x_2)\ldots p(x_n).$$

本节目录

经验分布函数

- 描述随机变量分布，可以用分布函数、密度函数或概率函数。
- 给定样本值 $x_1, x_2, \ldots, x_n$，如何估计分布函数 $F(x)$?
- 注意到

$$F(x) = P(X \leq x)$$

联系概率的频率含义以及简单随机样本可以看成是独立重复试验结果，用 $x_1, x_2, \ldots, x_n$ 中小于等于 x 的比例估计概率 $F(x)$。

- **定义 2.1** 设 $x_1, x_2, \ldots, x_n$ 是 X 的样本，称 x 的函数

$$F_n(x) = \frac{\nu_n}{n}$$

为 X 的经验分布函数，其中 ν_n 表示 $x_1, x_2, \ldots, x_n$ 中小于等于 x 的个数。

次序统计量

- 将样本值 $x_1, x_2, \ldots, x_n$ 从小到大排列后记为

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

$x_{(i)}$ 叫做样本的第 i 个次序统计量 $(i = 1, 2, \ldots, n)$。

经验分布函数的阶梯函数表示

- 经验分布函数是一个阶梯函数：
- 若 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 两两不同，易见

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

这是仅在每个 $x_{(i)}$ 处向上跳跃 $\frac{1}{n}$ 的阶梯函数，每一段在左端点连续，右端点不连续。

- 如果某个 $x_{(j)}$ 有 m 个相同的样本值，则 $F_n(x)$ 在 $x_{(j)}$ 处向上跳跃 $\frac{m}{n}$。

经验分布函数的 (强) 相合性

- 根据大数定律和强大数定律，对固定的 x，只要 n 相当大，$F_n(x)$ 与 $F(x)$ 很接近。
- 给定 x 后可以定义新的随机变量

$$Y_i = \begin{cases} 1 & X_i \le x \\ 0 & X_i > x \end{cases} \quad (i = 1, 2, \ldots, n)$$

- 则

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- $Y_1, Y_2, \ldots, Y_n$ 相互独立同 $\mathrm{b}(1, F(x))$ 二点分布，$E(Y_i) = F(x)$ $(i = 1, 2, \ldots, n)$。有

$$P\left(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} Y_i = F(x) \right) = 1$$

经验分布函数的一致强相合性

- 对所有 x，当 $n \to \infty$ 时，$F_n(x)$ 一致地逼近 $F(x)$:
- **定理 (Glivenko-Cantelli)** 设

$$D_n = \sup_x |F_n(x) - F(x)|$$

则

$$P\left(\lim_{n\to\infty} D_n = 0\right) = 1$$

分位数估计

- 当 $F(x)$ 严格单调上升且连续时，反函数 $F^{-1}(p)$ $(p \in (0,1))$ 为分位数函数。
- 一般地，$F(x)$ 的 p 分位数为满足

$$P(X \leq x_p) \geq p, \quad P(X \geq x_p) \geq 1-p$$

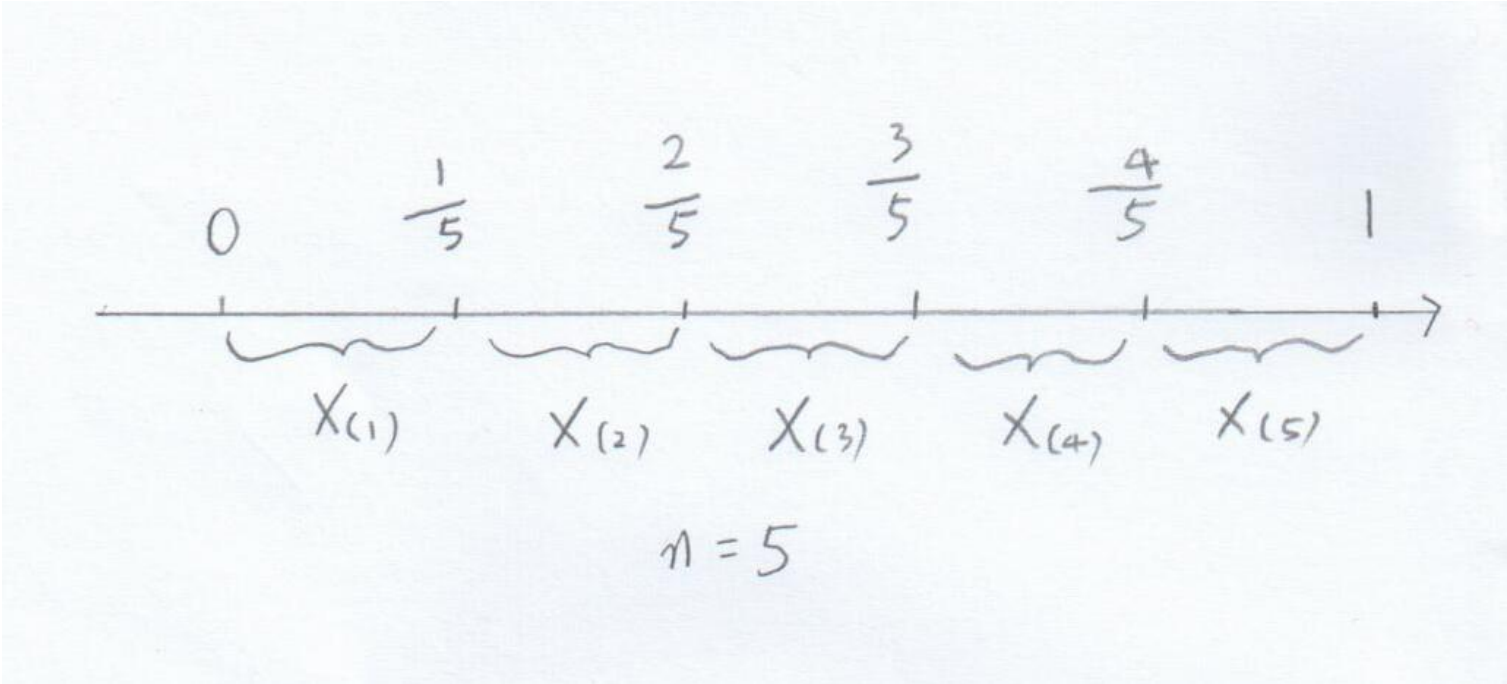
的数 x_p。
- 对样本值 $x_1, x_2, \ldots, x_n$，从小到大排列为次序统计量 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$，令

$$r = [pn] + 1$$

用 $x_{(r)}$ 估计 x_p。
- 当 $F(x) = p$ 至多有一个根时，x_p 存在唯一，

$$P\left(\lim_{n\to\infty} x_{(r)} = x_p\right) = 1$$

- 例如，$n = 5$，则 $x_{(1)}, x_{(2)}, \ldots, x_{(5)}$ 代表的 p 范围如下图：



例 2.1

- **例 2.1** 自动装罐头的净重随机，额定 345 克。
- 从生产线随机抽取 10 个罐头：

$$344, 336, 345, 342, 340, 338, 344, 343, 344, 343$$

- 要求估计分布函数 $F(x)$ 和中位数。
- **解** 可以认为是简单随机样本。用经验分布函数 $F_n(x)$ 估计 $F(x)$。

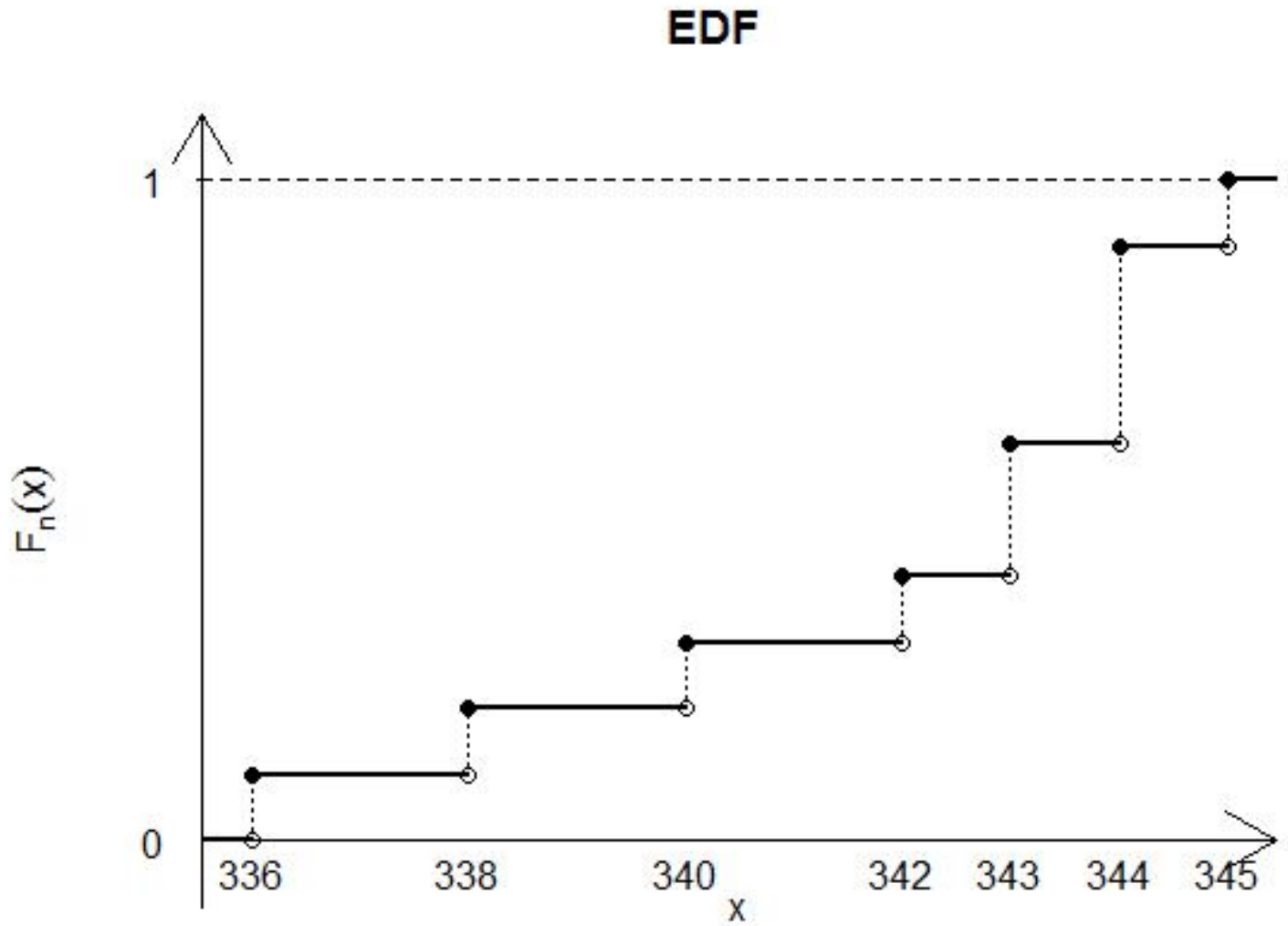
- 样本值从小到大排列得

$$336, 338, 340, 342, 343, 343, 344, 344, 344, 345$$

- 经验分布函数为

$$F_n(x) = \begin{cases} 0 & x < 336 \\ \frac{1}{10} & 336 \le x < 338 \quad (\text{跳跃 } \frac{1}{10}) \\ \frac{2}{10} & 338 \le x < 340 \quad (\text{跳跃 } \frac{1}{10}) \\ \frac{3}{10} & 340 \le x < 342 \quad (\text{跳跃 } \frac{1}{10}) \\ \frac{4}{10} & 342 \le x < 343 \quad (\text{跳跃 } \frac{1}{10}) \\ \frac{6}{10} & 343 \le x < 344 \quad (\text{跳跃 } \frac{2}{10}) \\ \frac{9}{10} & 344 \le x < 345 \quad (\text{跳跃 } \frac{3}{10}) \\ 1 & x \ge 345 \quad (\text{跳跃 } \frac{1}{10}) \end{cases}$$

- 经验分布函数图：



- 中位数估计：用次序统计量中间一个（奇数个时）或中间两个的平均（偶数个时），或者直接用 $x_{[0.5n]+1}$。即 $(x_{(5)}+x_{(6)})/2=(343+343)/2=343$, 或 $x_{([0.5\times 10]+1)}=x_{(6)}=343$。

分布密度估计

- 对于连续型总体（随机变量），分布函数不如分布密度直观。高维时分布函数更不实用。
- 分布密度估计有直方图法、核估计法、最近邻估计法等。

直方图法

- 直方图法用阶梯函数估计密度函数。
- 把样本 $x_1, x_2, \ldots, x_n$ 从小到大排列为次序统计量 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 后，把数轴分成 m 个小区间，在每个小区间中用

$$\frac{\text{落入小区间的样本点个数}}{n} \cdot \frac{1}{\text{小区间长度}}$$

估计该小区间的密度值 $p(x)$。

直方图估计的理论依据

- 设 $x_1, x_2, \ldots, x_n$ 为来自密度为 $p(x)$ 的总体的样本。
- 用 $R_n(a,b)$ 表示落入区间 $(a,b]$ 的样本点个数。
- 若 $(a,b]$ 很短，可认为 $x \in (a,b]$ 时 $p(x)$ 近似为常数，于是

$$P(a < X \leq b) = \int_a^b p(x)dx \approx p(x)(b-a)$$

- 用频率 $R_n(a,b)/n$ 估计概率 $P(a < X \leq b)$，有

$$\begin{aligned} \frac{R_n(a,b)}{n} &\approx p(x)(b-a) \\ p(x) &\approx \frac{R_n(a,b)}{n} \cdot \frac{1}{b-a}, \quad x \in (a,b] \end{aligned}$$

直方图法

- 为了估计密度函数 $p(x)$，设

$$t_0 < t_1 < \cdots < t_m$$

是 $m+1$ 个实数，通常假定 $t_i - t_{i-1} \equiv h > 0\ (i = 1, 2, \ldots, m)$。

- 令

$$p_n(x) = \begin{cases} \frac{R_n(t_{i-1}, t_i)}{n} \frac{1}{h} & \text{当} x \in (t_{i-1}, t_i],\ (i = 1, 2, \ldots, m) \\ 0 & \text{当} x \le t_0 \text{ 或} x > t_m \end{cases}$$

- 用 $p_n(x)$ 作为 $p(x)$ 的估计，这就是**直方图估计法**。

- 在第 i 个小区间 $(t_{i-1}, t_i]$ 用

$$p_n(x) = \frac{R_n(t_{i-1}, t_i)}{n} \frac{1}{h}$$

估计 $p(x)$，其图像是以底边为 h、高为 $p_n(x)$ 的矩形，面积为

$$\begin{aligned} p_n(x) \cdot h =& \frac{R_n(t_{i-1}, t_i)}{n} \\ \approx& P(t_{i-1} < X \leq t_i) = \int_{t_{i-1}}^{t_i} p(x)dx \\ \approx& p(x) \cdot h \end{aligned}$$

作直方图步骤

- 步骤一、对样本值 $x_1, x_2, \ldots, x_n$ 进行分组。找到最小值 $x_{(1)}$ 和最大值 $x_{(n)}$。
- 取 a 比 $x_{(1)}$ 略小，b 比 $x_{(n)}$ 略大，取适当分组数 m 把区间 $(a, b]$ m 等分，分点为

$$t_i = a + i\frac{b-a}{m} \quad (i = 0, 1, \ldots, m)$$

记每组的区间长度为

$$h = \frac{b-a}{m}$$

- m 的取法：n 小的时候 m 也较小，n 很大时 m 可以随之增大也可以不再增大。应尽可能使得多数小区间中包含有样本的值。一般取 a, b, m 使得各分点的小数位数比观测值的小数位数多一位，这样不会有样本值落在小区间端点上。
- m 的建议公式之一：

$$m \approx 1 + 3.322 \lg n$$

- m 的建议取法二：

n	m
< 50	$5 \sim 6$
$50 \sim 100$	$6 \sim 10$
$100 \sim 250$	$7 \sim 12$
> 250	$10 \sim 20$

- 步骤二、决定了分点后，用唱票的方法数出样本值落入第 i 组 $(t_{i-1}, t_i]$ 中的个数，记为 $\nu_i (i = 1, 2, \ldots, m)$。
- 计算样本值落入各组的频率

$$f_i = \frac{\nu_n}{n} \quad (i = 1, 2, \ldots, m)$$

- 步骤三、做直方图。画坐标系，x 轴范围为 (a, b)，y 轴范围为 $[0, \max\limits_i f_i]$。
- 对 $i = 1, 2, \ldots, m$，以 x 轴的区间 $[t_{i-1}, t_i]$ 为底，以 f_i/h 为高做矩形框。这一系列矩形叫做直方图。
- 每个竖着的长方形的面积，近似代表 X 取值落入“底边”的概率。
- 某区间上 $p(x)$ 下的面积为 X 落入该区间的概率，直方图用频率估计概率，从而估计 $p(x)$。（示意图）

例 2.2

- **例 2.2** $n = 120$。
- $x_{(1)} = 0.64, x_{(n)} = 0.95, x_{(n)} - x_{(1)} = 0.31$。
- 把距离 0.31 略增大为 0.32 就容易分解因数。可取 $m = 16, h = 0.02$, $a = x_{(1)} - 0.005 = 0.635, b = x_{(n)} + 0.005 = 0.955$，各分点千分位都有 0.005，没有样本点落在区间端点上。
- 算出各区间端点，用唱票法统计出 $n_i, i = 1, 2, \ldots, m$。
- 作图。

直方图估计的相合性

- 若密度函数 $p(x)$ 在 $(-\infty,\infty)$ 上一致连续，对某个 $\delta>0$，

$$\int_{-\infty}^{\infty}|x|^{\delta}p(x)dx$$

收敛，又小区间长度 h_n 满足

$$\lim_{n\to\infty}h_n=0$$

$$h_n\geq\frac{(\ln n)^2}{n}$$

- 则有

$$P\left(\lim_{n\to\infty}\sup_{-\infty<x<\infty}|p_n(x)-p(x)|\right)=1.$$

(一致强相合)

Rosenblatt 估计

- 为了估计 $p(x)$，若 $p(x)$ 连续，可根据

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

- 其中 $F(x)$ 可以用经验分布函数 $F_n(x)$ 估计。有

$$\hat{p}_n(x) = \frac{1}{2h}\left[F_n(x+h) - F_n(x-h)\right], \quad x \in (-\infty, \infty)$$

- 这叫做 Rosenblatt 密度估计。
- 注意到

$$F_n(x+h) - F_n(x-h) = \frac{R_n(x-h, x+h)}{n}$$

$$\hat{p}_n(x) = \frac{R_n(x-h, x+h)}{n}\frac{1}{2h}$$

- 所以 Rosenblatt 估计和直方图估计类似。

- 但是，直方图估计中小区间 $(t_{i-1}, t_i]$ 是固定的，而这里的小区间 $(x-h, x+h]$ 随自变量 x 而变化。
- 记

$$K_0(x) = \frac{1}{2} I_{[-1,1]}(x)$$

则

$$\begin{aligned} K_0\left(\frac{x-x_i}{h}\right) &= \frac{1}{2} I_{[x-h,x+h]}(x_i) \\ R_n(x-h, x+h) &= 2\sum_{i=1}^{n} K_0\left(\frac{x-x_i}{h}\right) \\ \hat{p}_n(x) &= \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{x-x_i}{h}\right) \end{aligned}$$

- $K_0(\cdot)$ 叫做一个“核函数”，$\hat{p}_n(x) \cdot 2h$ 是 x 邻域内的样本值百分比。

核密度估计

- Rosenblatt 估计采用了 x 邻域 $[x-h, x+h]$ 内的样本点数，x 邻域内样本点越多，密度估计越大。
- 推广：采纳样本点时，不一刀切，而是离 x 越近的样本点加权越大。
- **定义 2.2** 设 $K(x)$ 是非负函数且 $\int_{-\infty}^{\infty} K(x)dx = 1$, 则称 $K(x)$ 是核函数。此时称

$$\tilde{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

  为 $p(x)$ 的核估计。
- 核函数一般选为偶函数，且在正半轴单调下降（类似于正态分布曲线）。

常用核函数

- $$
\begin{aligned}
K_0(x) =& \frac{1}{2} I_{[-1,1]}(x) \\
K_1(x) =& \left(1-|x|^3\right)^3 I_{[-1,1]}(x) \\
K_2(x) =& \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \\
K_3(x) =& \frac{1}{\pi\left(1+x^2\right)} \\
K_4(x) =& \frac{1}{2\pi}\left(\frac{\sin\frac{x}{2}}{\frac{x}{2}}\right)^2
\end{aligned}
$$

- 演示。

核估计的相合性

- 当 n 无限增大且 $h = h_n$ 无限减小时，核估计 $\tilde{p}(x)$ 与密度 $p(x)$ 无限接近。
- 若 $p(x)$ 在 $(-\infty, \infty)$ 上一致连续，且

$$\lim_{n\to\infty} h_n = 0$$

$$\sum_{n=1}^{\infty} \exp\left\{-rnh_n^2\right\} < \infty \quad (\forall r > 0)$$

又核函数 $K(x)$ 为有界变差函数，则

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |\tilde{p}_n(x) - p(x)| = 0\right) = 1.$$

(一致强相合)

- 在核函数和 h 选取合适时核估计比直方图估计精度更高。

最近邻估计

- 核估计是 x 附近的样本点越多则密度估计值越大。
- 最近邻估计是固定 x 附近需要有的样本点数，令邻域区间长度可变。
- 取自然数 $K(n)$（n 为样本量），令

$$a_n(x) = \min\{t : t > 0, R_n(x-t, x+t) \geq K(n)\}$$

$$p_n^*(x) = \frac{K(n)}{n} \frac{1}{2a_n(x)}$$

最近邻估计的相合性

- 适当条件下 $n \to \infty$ 时 $p_n^*(x)$ 与 $p(x)$ 可以任意接近。
- 若 $p(x)$ 在 $(-\infty, \infty)$ 上一致连续，且

$$\lim_{n\to\infty} \frac{K(n)}{n} = 0 \qquad \lim_{n\to\infty} \frac{K(n)}{\ln n} = \infty$$

则

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |p_n^*(x) - p(x)| = 0\right) = 1.$$

(一致强相合)

本节目录

参数方法与非参数方法

- 经验分布函数、直方图估计、核密度估计、最近邻密度估计都不需要假定总体来自那一种分布，称为**非参数**统计方法。
- 但是，直接估计一个函数需要的信息量很大。
- 如果已知总体分布类型，只是分布参数未知，则可以只估计分布参数，然后总体分布就知道了。这种方法叫做**参数**统计方法。
- 例如，设产品指标服从正态分布 $\mathrm{N}(\mu, \sigma^2)$，但 μ, σ^2 未知，就可以由样本估计 μ, σ^2，代入密度函数中作为分布密度估计。
- 又如，产品寿命常服从威布尔分布或对数正态分布，可以从样本中估计分布参数后得到总体分布的估计。

参数估计问题

- 设总体 X 的密度函数或概率函数为 $p(x;\theta_1,\theta_2,\ldots,\theta_m)$，其中 $\theta_1,\theta_2,\ldots,\theta_m$ 是未知参数。
- 若 X 的样本值为

$$x_1,x_2,\ldots,x_n$$

- 问：如何估计参数 $\theta_1,\theta_2,\ldots,\theta_m$?
- 估计方法有很多，常用的有最大似然估计法和矩估计法。

似然函数

- 给定样本值 $x_1, x_2, \ldots, x_n$ 后，令

$$\begin{aligned} & L_n(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m) \\ = & \prod_{i=1}^{n} p(x_i; \theta_1, \theta_2, \ldots, \theta_m) \end{aligned}$$

  把样本值 $x_1, x_2, \ldots, x_n$ 看作常数，这样 $L_n(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)$ 是参数 $\theta_1, \theta_2, \ldots, \theta_m$ 的函数，叫做样本 $x_1, x_2, \ldots, x_n$ 的**似然函数**。
- 似然函数如果看成自变量 $x_1, x_2, \ldots, x_n$ 的函数，实际是样本 $(X_1, X_2, \ldots, X_n)$ 看成随机变量时的联合密度函数。

最大似然估计

- **定义 3.1** 如果 $L_n(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)$ 在 $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ 达到最大值，则称 $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ 为参数 $(\theta_1, \theta_2, \ldots, \theta_m)$ 的**最大似然估计** (MLE)。

最大似然估计直观解释例子

- **例** 假设两个人一同出去打猎，两人只允许开一枪。
- 甲击中概率为 0.9，乙击中概率为 0.2。
- 二人返回时，带回一头猎物。
- 众人会认为是谁的开枪?

- 两人射击结果用随机变量 X 表示：

$$X=\begin{cases}1 & \text{击中}\\ 0 & \text{未击中}\end{cases}$$

- $P(X=1)$ 可能为 $\{0.9,0.2\}$ 两种取值（对应甲开枪和乙开枪）。
- 在甲开枪的情况下，得到已发生的“命中”结果可能性大于乙开枪的情况下，得到已发生的“命中”结果可能性。
- 所以推测是甲开的枪。
- 在结果（样本值）给定时，似然函数代表参数取不同值时，观测到已发生的结果的可能性大小。

最大似然估计直观解释例子 II

- 假定一个盒子里有许多黑球和白球，比例 3:1，但不知道黑球多还是白球多。
- 则随机抽取一个球，得到黑球的概率或者是 $\frac{1}{4}$, 或者是 $\frac{3}{4}$。
- 如果有放回地从盒子里抽 3 个球，则黑球数目服从二项分布：

$$P(X=x) = C_3^x p^x (1-p)^{3-x}, \quad x = 0, 1, 2, 3$$

$$(p = \frac{1}{4}, \frac{3}{4} \text{为抽到黑球的概率})$$

- 根据抽取到的黑球数判断到底是黑球多还是白球多。
- 直观看，如果 $x = 0, 1$，则猜白球多；如果 $x = 2, 3$，则猜黑球多。

- 样本值固定后，取不同参数值的似然函数大小代表该种参数下观测到已发生的情况的可能性大小。
- 我们取使得已发生的情况出现的可能性最大的参数作为估计值。
- 分别计算 $p=\frac{1}{4},\frac{3}{4}$ 的似然函数值：

x	0	1	2	3
$p=\frac{1}{4}$ 时 $P(X=x)$ 的值	$\mathbf{\frac{27}{64}}$	$\mathbf{\frac{27}{64}}$	$\frac{9}{64}$	$\frac{1}{64}$
$p=\frac{3}{4}$ 时 $P(X=x)$ 的值	$\frac{1}{64}$	$\frac{9}{64}$	$\mathbf{\frac{27}{64}}$	$\mathbf{\frac{27}{64}}$

- 于是 $x=0,1$ 时应取 $\hat{p}=\frac{1}{4}$（猜白球多），$x=2,3$ 时应取 $\hat{p}=\frac{3}{4}$（猜黑球多）。

最大似然估计求法

- 把似然函数简记为 L_n。L_n 与 $\ln L_n$ 同时达到最大值，可以求 $\ln L_n$ 的最大值点。
- 当 $\ln L_n$ 的一阶偏导数存在时，最大值点处一阶偏导数都等于零：

$$\begin{cases} \frac{\partial \ln L_n}{\partial \theta_1} = 0 \\ \frac{\partial \ln L_n}{\partial \theta_2} = 0 \\ \cdots\cdots \\ \frac{\partial \ln L_n}{\partial \theta_m} = 0 \end{cases} \tag{3.1}$$

  这个关于参数 $(\theta_1, \theta_2, \ldots, \theta_m)$ 的方程组叫做似然方程组。
- 注意一阶偏导存在条件下似然方程组成立，但似然方程组的解不能保证为最大值点。

最大似然估计的优良性

- 在相当一般的条件下：
- 相合性：n 充分大时最大似然估计结果与参数真值之间可以无限接近。
- 有效性：一定意义下没有比最大似然估计更精确的估计。
- 渐近正态性：n 充分大时最大似然估计近似服从正态分布。

指数分布参数的最大似然估计

- 密度

$$p(x;\lambda)=\lambda e^{-\lambda x},\ x>0,\lambda>0$$

- 样本 $x_1,x_2,\ldots,x_n$ 的似然函数和对数似然函数

$$L_n(x_1,x_2,\ldots,x_n;\lambda)=\lambda^n\exp\left\{-\lambda\sum_1^n x_i\right\}$$

$$\ln L_n=n\ln\lambda-\lambda\sum_1^n x_i$$

- 似然方程

$$\frac{d\ln L_n}{d\lambda}=\frac{n}{\lambda}-\sum_1^n x_i=0$$

$$\hat{\lambda}=\frac{n}{\sum_1^n x_i}=\frac{1}{\bar{x}}$$

此 $\hat{\lambda}$ 确实是 $\ln L_n$ 的最大值点，是 λ 的最大似然估计。

例 3.1

- **例 3.1** 一直某种电子设备的使用寿命服从指数分布 $\mathrm{Exp}(\lambda)$。今随机抽取 18 台，测得寿命数据如下（单位：小时）：

$$16, 29, 50, 68, 100, 130, 140$$
$$270, 280, 240, 410, 450, 520, 620$$
$$190, 210, 800, 1100$$

- 求 λ 的估计。
- **解** 用最大似然估计。$\bar{x} = 318$，

$$\hat{\lambda} = \frac{1}{318} \approx 0.03144$$

是 λ 的估计值。

正态分布参数的最大似然估计

- $\mathrm{N}(\mu, \sigma^2)$ 的密度函数为 (记 $\delta = \sigma^2$)

$$p(x; \mu, \delta) = \frac{1}{\sqrt{2\pi\delta}} \exp\left\{-\frac{1}{2\delta}(x-\mu)^2\right\}$$

- 样本 $x_1, x_2, \ldots, x_n$ 的似然函数与对数似然函数为

$$\begin{aligned} & L_n(x_1, x_2, \ldots, x_n;\ \mu, \delta) \\ =& (2\pi)^{-\frac{n}{2}} \delta^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\delta}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\ & \ln L_n \\ =& -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\delta - \frac{1}{2\delta}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

- 似然方程组为

$$\begin{cases} \dfrac{\partial \ln L_n}{\partial \mu} = \dfrac{1}{\delta} \displaystyle\sum_{i=1}^{n} (x_i - \mu) = 0 \\ \dfrac{\partial \ln L_n}{\partial \delta} = -\dfrac{n}{2\delta} + \dfrac{1}{2\delta^2} \displaystyle\sum_{i=1}^{n} (x_i - \mu)^2 = 0 \end{cases}$$

- 解得

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\delta} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- $(\hat{\mu}, \hat{\delta})$ 确实是 $\ln L_n$ 的最大值点，所以是 (μ, σ^2) 的最大似然估计。

威布尔分布参数的最大似然估计

- Weibull(m,η) 分布密度为

$$p(x;m,\eta)=\frac{m}{\eta^m}x^{m-1}\exp\left\{-\left(\frac{x}{\eta}\right)^m\right\},$$
$$x>0;\ m>0,\eta>0$$

- 似然函数和对数似然函数分别为

$$L_n(x_1,x_2,\ldots,x_n;m,\eta)$$
$$=m^n\eta^{-mn}\left(\prod_{i=1}^n x_i\right)^{m-1}\exp\left\{-\frac{1}{\eta^m}\sum_{i=1}^n x_i^m\right\}$$
$$\ln L_n$$
$$=n\ln m-nm\ln\eta+(m-1)\sum_{i=1}^n\ln x_i-\frac{1}{\eta^m}\sum_{i=1}^n x_i^m$$

- 似然方程组为

$$\frac{\partial \ln L_n}{\partial m} = \frac{n}{m} - n \ln \eta + \sum_{i=1}^{n} \ln x_i$$

$$- \frac{1}{\eta^m} \sum_{i=1}^{n} x_i^m \ln x_i + \frac{\ln \eta}{\eta^m} \sum_{i=1}^{n} x_i^m = 0 \tag{3.2a}$$

$$\frac{\partial \ln L_n}{\partial \eta} = - \frac{nm}{\eta} + \frac{m}{\eta^{m+1}} \sum_{i=1}^{n} x_i^m = 0 \tag{3.2b}$$

- 由 (3.2b) 得

$$\eta = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^m \right)^{\frac{1}{m}} \tag{3.3}$$

- 代入 (3.2a) 可得

$$\frac{1}{m}+\frac{1}{n}\sum_{i=1}^{n}\ln x_i-\frac{\sum_{i=1}^{n}x_i^m\ln x_i}{\sum_{i=1}^{n}x_i^m}=0 \tag{3.4}$$

- 当 $n\geq 2, x_1, x_2, \ldots, x_n$ 不完全相等时，方程 (3.4) 恰有一个根 $\hat{m}$，代回 (3.3) 中得

$$\hat{\eta}=\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\hat{m}}\right)^{\frac{1}{\hat{m}}}$$

- 可以证明 $(\hat{m},\hat{\eta})$ 是似然方程组的解，也是对数似然函数的最大值点，所以是参数 (m,η) 的最大似然估计。
- 方程 (3.4) 一定有唯一解且方程左边是 m 的严格单调减函数，可以用二分法求解 (计算机数值算法求解)。

例 3.2

- **例 3.2** 轴承的寿命一般服从威布尔分布。$n = 20$ 的样本数据如下 (单位：小时)：

$$153, 223, 313, 373, 378, 385, 424,$$
$$232, 452, 452, 547, 561, 634, 699,$$
$$759, 859, 1000, 1132, 1152, 1466$$

估计形状参数 m 和刻度参数 η。

- **解** 用最大似然估计法，解方程 (3.4) 得 m 的最大似然估计 $\hat{m} = 1.9$。再利用 (3.5) 可得 η 的最大似然估计 $\hat{\eta} = 685$。

均匀分布参数的最大似然估计

- 参数似然函数不可导时，可以具体研究似然函数求最大值点。
- 均匀分布 $\mathrm{U}(a,b)$ 的密度函数为

$$p(x;a,b)=\frac{1}{b-a}I_{[a,b]}(x)$$

($a<b$ 是未知参数)

- 似然函数

$$\begin{aligned}L_n&=\frac{1}{(b-a)^n}\prod_{i=1}^{n}I_{[a,b]}(x_i)\\&=\begin{cases}\frac{1}{(b-a)^n} & \text{当}x_{(1)}\geq a\text{且}x_{(n)}\leq b\\0 & \text{其它}\end{cases}\end{aligned}$$

- 似然函数最大，首先要求不为零，即 $a \leq x_{(1)}, b \geq x_{(n)}$。在此条件下 $b-a$ 最小，应取 a 最大, b 最小，所以

$$\hat{a} = x_{(1)} \qquad \hat{b} = x_{(n)}$$

本节目录

数字特征的估计

- 直接估计分布函数、分布密度、概率函数要求数据很多，参数最大似然估计有时比较复杂。
- 如果只是需要估计期望、方差等数字特征，则比较容易。

期望的点估计

- 例 1.1 中钢筋次品率估计可以看成是二点分布总体 X 的期望 $E(X)$ 的估计问题。
- 一般地，对总体 X 的样本 $x_1, x_2, \ldots, x_n$，估计 $E(X)$ 可以用样本平均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 这个估计是好的，而且样本量 n 越大，估计越精确。

统计量和抽样分布

- 把样本 $X_1, X_2, \ldots, X_n$ 看成随机变量，样本平均值

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

  也是随机变量。
- 若函数 $\psi(x_1, x_2, \ldots, x_n)$ 是不依赖于未知参数的函数，样本的函数 $Y = \psi(X_1, X_2, \ldots, X_n)$ 叫做样本**统计量**。
- 统计量是随机变量，其分布叫做**抽样分布**。
- 比如 $\bar{X}$ 是统计量。

样本平均值的无偏性

- 虽然用 $\bar{X}$ 估计 $E(X)$ 有时比 $E(X)$ 大，有时比 $E(X)$ 小，但平均来说是等于 $E(X)$ 的：
- **定理 4.1** 设 $E(X)$ 存在，则

$$E(\bar{X}) = E(X)$$

- **证**

$$\begin{aligned} E(\bar{X}) =& E\left[\frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)\right] \\ =& \frac{1}{n} E\left(X_1 + X_2 + \cdots + X_n\right) \\ =& \frac{1}{n}\left[E(X_1) + E(X_2) + \cdots + E(X_n)\right] \\ =& \frac{1}{n} \cdot nE(X) = E(X) \end{aligned}$$

- 称这样的估计为**无偏**估计。

有效性

- 记

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

- 按常理，$\bar{X}_2$ 应该比 $\bar{X}_1$ 更好。
- 在无偏的条件下，

$$D(\bar{X}_n) = E\left[\bar{X}_n - E(\bar{X}_n)\right]^2 = E\left[\bar{X}_n - E(X)\right]^2$$

  代表了估计误差大小。$D(\bar{X}_n)$ 越小则估计约精确。
- 抽样分布方差越小，称统计量越有效。

- **定理 4.2** 设 X 的期望、方差都存在，则

$$D(\bar{X}_n) = \frac{D(X)}{n}$$

- **证**

$$\begin{aligned} D(\bar{X}_n) =& D\left[\frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)\right] \\ =& \frac{1}{n^2} D\left(X_1 + X_2 + \cdots + X_n\right) \\ =& \frac{1}{n^2}\left[D(X_1) + D(X_2) + \cdots + D(X_n)\right] \\ =& \frac{1}{n^2}\left[nD(X)\right] = \frac{D(X)}{n} \end{aligned}$$

- 定理说明样本量越大，估计越精确。

说明

- 为什么 n 越大，估计越精确?
- 利用切比雪夫不等式:

$$P\{|\bar{X}-E(\bar{X})|<\varepsilon\}\geq 1-\frac{D(\bar{X})}{\varepsilon^2}$$

- 其中 $E(\bar{X})=E(X), D(\bar{X})=\frac{D(X)}{n}$, 有

$$P\{|\bar{X}-E(X)|<\varepsilon\}\geq 1-\frac{D(X)}{n\varepsilon^2} \tag{4.1}$$

- 从而

$$\lim_{n\to\infty} P\{|\bar{X}-E(X)|<\varepsilon\}=1 \tag{4.2}$$

即 n 充分大时可以有充分大把握保证 $|\bar{X}-E(X)|<\varepsilon$，即 $\bar{X}\approx E(X)$。

关于估计的优良性

- 估计的优良性包括无偏性、有效性、相合性等。
- 设 X 的分布密度为 $p(x;\theta)$，其中 $\theta=(\theta_1,\theta_2,\ldots,\theta_m)\in\Theta$, Θ 是 m 维空间 R^m 中的某个集合。（当 X 为离散型时可做类似讨论）。
- 设 $g(\theta)$ 是参数 θ 的函数，$X_1,X_2,\ldots,X_n$ 是 X 的样本。
- **定义 4.1** 称样本的函数 $\varphi(X_1,X_2,\ldots,X_n)$ 为 $g(\theta)$ 的**估计量**。
- 估计量是统计量，只要给定样本值就可以计算出数值，计算不能依赖于参数 θ。
- 如何选择估计量?

无偏性

- 由于 $X_1, X_2, \ldots, X_n$ 的联合密度与 θ 有关，所以 $\varphi(X_1, X_2, \ldots, X_n)$ 的期望与 θ 有关，为此显式地记为

$$E_\theta\left[\varphi(X_1, X_2, \ldots, X_n)\right]$$

- **定义 4.2** 称 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的无偏估计，若

$$E_\theta\left[\varphi(X_1, X_2, \ldots, X_n)\right] = g(\theta) \quad (\forall \theta \in \Theta)$$

有效性

- **定义 4.3** 若 $\varphi_1(X_1, X_2, \ldots, X_n)$ 和 $\varphi_2(X_1, X_2, \ldots, X_n)$ 都是 $g(\theta)$ 的估计量，满足

$$\begin{aligned} & E_\theta \left[\varphi_1(X_1, X_2, \ldots, X_n) - g(\theta)\right]^2 \\ \leq & E_\theta \left[\varphi_2(X_1, X_2, \ldots, X_n) - g(\theta)\right]^2 \quad (\forall \theta \in \Theta) \end{aligned}$$

  且存在 θ_0 使上式中小于号成立，则称 φ_1 比 φ_2 **有效**。
- 比如，$\bar{X}_k$ 和 $\bar{X}_{k-1}$ 都是 $E(X)$ 的无偏估计，但 $\bar{X}_k$ 比 $\bar{X}_{k-1}$ 有效（设 $D(X) > 0$）。

最小方差无偏估计

- **定义 4.4** 如果 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的无偏估计量，而且对于 $g(\theta)$ 的任一无偏估计量 $\psi(X_1, X_2, \ldots, X_n)$ 都有

$$D\left(\varphi(X_1, X_2, \ldots, X_n)\right) \leq D\left(\psi(X_1, X_2, \ldots, X_n)\right) \quad (\forall \theta \in \Theta)$$

则称 $\varphi(X_1, X_2, \ldots, X_n)$ 为 $g(\theta)$ 的最小方差无偏估计量。

相合性

- **定义** 称 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的相合估计，若对任意 $\varepsilon > 0$，

$$\lim_{n\to\infty} P\left(|\varphi(X_1, X_2, \ldots, X_n) - g(\theta)| > \varepsilon\right) = 0.$$

- **定义** 称 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的强相合估计，若

$$P\left(\lim_{n\to\infty} |\varphi(X_1, X_2, \ldots, X_n) - g(\theta)| = 0\right) = 1.$$

方差的点估计

- 设 $X_1, X_2, \ldots, X_n$ 为 X 的样本，为估计 $D(X)$，使用

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

  称为**样本方差**。
- 显然，如果所有样本值都相等，样本值波动最小，$S^2 = 0$。样本值之间差异越大，S^2 越大。
- 为什么除以 $n-1$ 而不是除以 n?

样本方差的无偏性

- **定理 4.3** 设 X 的方差存在，则

$$E(S^2) = D(X)$$

- **证**

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}\left(X_i^2 - 2\bar{X}\cdot X_i + \bar{X}^2\right) \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\cdot\sum_{i=1}^{n} X_i + n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\cdot n\bar{X} + n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \qquad (4.3)
\end{aligned}$$

- 于是

$$
\begin{aligned}
E(S^2) =& E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\
=& \frac{1}{n-1}E\left[\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right] \\
=& \frac{1}{n-1}\sum_{i=1}^{n}E(X_i^2)-\frac{n}{n-1}E\left(\bar{X}^2\right) \\
=& \frac{n}{n-1}\left[E\left(X^2\right)-E\left(\bar{X}^2\right)\right]
\end{aligned}
$$

- 其中

$$
\begin{aligned}
E\left(X^2\right) =& D(X)+[E(X)]^2 \\
E\left(\bar{X}^2\right) =& D\left(\bar{X}\right)+\left[E\left(\bar{X}\right)\right]^2 \\
=& \frac{D(X)}{n}+[E(X)]^2
\end{aligned}
$$

- 所以

$$
E(S^2)=\frac{n}{n-1}\left[D(X)-\frac{1}{n}D(X)\right]=D(X)
$$

- S^2 是 $D(X)$ 的无偏估计。如果用

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

估计 $D(X)$，则

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

- 但是，如果知道了所有总体的值 $x_1,x_2,\ldots,x_N$，则应该使用 $\frac{1}{N}$ 来计算总体方差。
- 注：知道所有总体时，可以认为分布为所有总体值上的离散均匀分布，这时 $E(X)=\bar{X}, D(X)=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2$。

矩估计法

- 如果总体参数可以从各阶矩表示，则估计了各阶矩后可以估计参数。
- 设随机变量 X 的分布密度是 $p(x;\theta_1,\theta_2,\ldots,\theta_m)$，$\nu_k$ 是 X 的 k 阶矩 $(k=1,2,\ldots)$，ν_k 是 $\theta_1,\theta_2,\ldots,\theta_m$ 的函数：

$$\nu_k = E\left(X^k\right) = g_k(\theta_1,\theta_2,\ldots,\theta_m)$$

- 设 $\nu_1, \nu_2, \ldots, \nu_m$ 已知，如果从方程组

$$\begin{cases} g_1(\theta_1, \theta_2, \ldots, \theta_m) = \nu_1 \\ g_2(\theta_1, \theta_2, \ldots, \theta_m) = \nu_2 \\ \qquad\qquad \cdots\cdots\cdots\cdots \\ g_m(\theta_1, \theta_2, \ldots, \theta_m) = \nu_m \end{cases}$$

可以求出

$$\begin{cases} \theta_1 = f_1(\nu_1, \nu_2, \ldots, \nu_m) \\ \theta_2 = f_2(\nu_1, \nu_2, \ldots, \nu_m) \\ \qquad\qquad \cdots\cdots\cdots\cdots \\ \theta_m = f_m(\nu_1, \nu_2, \ldots, \nu_m) \end{cases}$$

- 设 $x_1, x_2, \ldots, x_n$ 是 X 的样本值，用

$$\hat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad (k = 1, 2, \ldots, m)$$

来估计 ν_k；

- 用

$$\hat{\theta}_k = f_k(\hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_m)$$

估计 $\theta_k (k = 1, 2, \ldots, m)$。

- 这种估计未知参数的方法叫做**矩估计法**。

例 4.2

- **例 4.2** 设 $X \sim \mathrm{N}(\mu, \sigma^2)$，$x_1, x_2, \ldots, x_n$ 是样本，求 μ, σ^2 的矩估计。
-

$$\begin{aligned} \nu_1 &= E(X) = \mu \\ \nu_2 &= E(X^2) = \sigma^2 + \mu^2 \end{aligned}$$

- 反解得

$$\begin{cases} \mu &= \nu_1 \\ \sigma^2 &= \nu_2 - \nu_1^2 \end{cases}$$

- 估计 ν_1, ν_2 为

$$\hat{\nu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\nu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

- 估计 μ, σ^2 为

$$\hat{\mu} = \hat{\nu}_1 = \bar{x}$$

$$\hat{\sigma}^2 = \hat{\nu}_2 - \hat{\nu}_1^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 与最大似然估计相同。
- 其它分布不一定。样本量较大时，最大似然估计一般更精确。

例 4.3

- **例 4.3** $X \sim \mathrm{U}(0, \theta)$，$\theta$ 为未知参数。
- 最大似然估计 $\hat{\theta} = x_{(n)}$。
- 因为 $\nu_1 = E(X) = \frac{\theta}{2}$，所以 θ 矩估计为

$$\tilde{\theta} = 2\hat{\nu}_1 = 2\bar{x} = \frac{2}{n}\sum_{i=1}^{n} x_i$$

例 4.4

- **例 4.4** 台风可以引起内陆降雨。以下 $n = 36$ 个观测值为 24 小时降雨量实际观测数据：

$$31.00, 2.82, 3.95, 4.02, 9.50, 4.50, 11.40,$$
$$10.71, 6.31, 4.95, 5.64, 5.51, 13.40, 9.72,$$
$$6.47, 10.16, 4.21, 11.60, 4.75, 6.85, 6.25,$$
$$3.42, 11.80, 0.80, 3.69, 3.10, 22.22, 7.43,$$
$$5.00, 4.58, 4.46, 8.00, 3.73, 3.50, 6.20, 0.67$$

- 降雨量一般服从伽玛分布。密度为

$$p(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

- 用矩法估计 α, β。

$$\nu_1 = \frac{\alpha}{\beta}$$
$$\nu_2 = \frac{\alpha(\alpha+1)}{\beta^2}$$

- $\hat{\nu}_1 = 7.29, \hat{\nu}_2 = 85.59$。
- 解方程组

$$\begin{cases} \dfrac{\alpha}{\beta} = 7.29 \\ \dfrac{\alpha(\alpha+1)}{\beta^2} = 85.59 \end{cases}$$

- 得 $\hat{\alpha} = 1.64, \hat{\beta} = 0.22$。

本节目录

置信区间

- 前面找到了期望 $E(X)$ 和方差 $D(X)$ 的估计量，这种估计量又称为**点估计**，因为它们是用一个数值来估计未知的参数或数字特征的。
- 我们有时还希望了解估计的准确程度，这时应该用一个可能取值的范围（区间）来估计未知参数和数字特征。
- 将讨论正态总体的区间估计：
  (1) 已知方差，对 $E(X)$ 进行区间估计；
  (2) 未知方差，对 $E(X)$ 进行区间估计；
  (3) 方差 $D(X)$ 的区间估计在下一节讨论。

方差已知时期望的置信区间

- 设总体 X 服从 $\mathrm{N}(\mu, \sigma^2)$ 分布。
- 则 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 也是正态分布随机变量，分布为

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) . = \mathrm{N}\left(E(X), \frac{D(X)}{n}\right)$$

- (参见 P144 习题十四第 9 题及 P422 附录二定理 5 的系)。
- 于是

$$\eta = \frac{\bar{X} - E(X)}{\sqrt{\frac{D(X)}{n}}} \sim \mathrm{N}(0, 1)$$

- 根据正态分布的经验规则，

$$P(|\eta| \leq 1.96) = 0.95$$

- 即

$$P\left(|\bar{X} - E(X)| \leq 1.96\sqrt{\frac{D(X)}{n}}\right) = 0.95 \tag{5.1}$$

- 从 (5.1) 式看出，有 95% 的把握保证

$$|E(X) - \bar{X}| \leq 1.96\sqrt{\frac{D(X)}{n}}$$

即

$$\bar{X} - 1.96\sqrt{\frac{D(X)}{n}} \leq E(X) \leq \bar{X} + 1.96\sqrt{\frac{D(X)}{n}}$$

- 即随机区间

$$\left[\bar{X} - 1.96\sqrt{\frac{D(X)}{n}}, \bar{X} + 1.96\sqrt{\frac{D(X)}{n}}\right] \tag{5.2}$$

  以很大的概率包含 $E(X)$。
- 这样的区间叫做**置信区间**。
- 称这样的置信区间的**置信水平**或**置信度**为 0.95。
- 如果做 100 次抽样（每次抽 n 个样品），则从平均的意义讲，算出的 $\bar{x}$ 值有 95 次，使得区间 (5.2) 包含真值 μ。(演示)
- 注意：在计算出置信区间后，我们不能说 $E(X)$ 属于这个区间的概率是 95%。因为一个计算出来的区间或者包含 $E(X)$, 或者不包含 $E(X)$。
- 样本量 n 越大，置信区间越短。
- 置信度越高，计算的置信区间越长。

例 5.1

- **例 5.1** 滚珠直径 X 服从正态分布。随机抽取 $n = 6$ 个，测量值（单位: mm）:

$$14.70,\ 15.21,\ 14.90,\ 14.91,\ 15.32,\ 15.32$$

- 估计直径的平均值。
- 如果知道 X 的方差为 0.05, 求平均直径的置信区间。

- 解

$$\begin{aligned}\bar{x} =& \frac{1}{6}(14.70 + 15.21 + 14.90 + 14.91 + 15.32 + 15.32) \\ =& 15.06(\text{mm})\end{aligned}$$

为 $E(X)$ 的估计。

- 为计算 $E(X) = \mu$ 的置信区间，计算半径

$$1.96\sqrt{\frac{D(X)}{n}} = 1.96 \times \sqrt{\frac{0.05}{6}} = 0.18$$

- $E(X)$ 的置信区间为

$$[15.06 - 0.18, 15.06 + 0.18] = [14.88, 15.24]$$

非正态分布的情形

- 如果 X 不是服从正态分布，根据中心极限定理，当 n 充分大时

$$\eta = \frac{\bar{X} - E(X)}{\sqrt{\frac{D(X)}{n}}}$$

近似服从标准正态分布。

- 所以 $E(X)$ 的置信度为 95% 的置信区间仍可用公式

$$\left[\bar{X} - 1.96\sqrt{\frac{D(X)}{n}}, \bar{X} + 1.96\sqrt{\frac{D(X)}{n}}\right]$$

计算。

方差未知时均值的置信区间

- 如果 $D(X)$ 未知时求 $E(X)$ 的置信区间，想到的是用样本方差 S^2 代替 (5.2) 中的 $D(X)$。
- 但是，这时

$$T = \frac{\bar{X} - E(X)}{\sqrt{\frac{S^2}{n}}}$$

不再服从标准正态分布，不能利用正态分布经验规则。

- 需要推导 T 的分布。

t 分布

- 当总体 $X \sim \mathrm{N}(\mu, \sigma^2)$，$X_1, X_2, \ldots, X_n$ 是 X 的样本的时候，可以证明 T 的分布密度为

$$p_{n-1}(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} \tag{5.4}$$

- T 的分布只依赖于样本量 n 而与总体参数 μ, σ^2 无关。
- **定义 5.1** 如果随机变量 Y 的分布密度为

$$p_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \cdot \sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \tag{5.6}$$

  则称 Y 服从 n 个自由度的 t 分布，记为 $Y \sim \mathrm{t}(n)$。
- 统计量 T 服从 $\mathrm{t}(n-1)$ 分布。

- t(n) 密度为偶函数，形状与标准正态分布类似，但两个尾部比正态分布厚。
- 可以证明

$$\lim_{n\to\infty} p_n(t) = \phi(t), \quad t \in (-\infty, \infty)$$

- 可以找到 $\lambda > 0$ 使得

$$\int_{-\lambda}^{\lambda} p_{n-1}(t)dt = 0.95$$

- 即

$$P(|T| \leq \lambda) = 0.95$$

$$P\left(\bar{X} - \lambda\sqrt{\frac{S^2}{n}} \leq E(X) \leq \bar{X} + \lambda\sqrt{\frac{S^2}{n}}\right) = 0.95$$

- 于是 $E(X)$ 的置信度为 95% 的置信区间为

$$\left[\bar{X} - \lambda\sqrt{\frac{S^2}{n}}, \bar{X} + \lambda\sqrt{\frac{S^2}{n}}\right]$$

- λ 叫做 t 分布双侧 0.05 临界值，在 P432 附表 2 中列有不同自由度和不同置信度的对应值。

例 5.2

- **例 5.2** 用某仪器间接测量温度，重复测量 5 次，得到的结果如下（单位：°C）

$$1250,\ 1265,\ 1245,\ 1260,\ 1275$$

假设仪器没有系统偏差，求真值的范围。

- **解** 用 μ 表示温度真值，X 表示测量值。X 通常服从正态分布。有 $n = 5$ 的样本。

- μ 的置信区间为

$$\left[\bar{x} - \lambda\sqrt{\frac{S^2}{n}}, \bar{x} + \lambda\sqrt{\frac{S^2}{n}}\right]$$

- 计算得 $\bar{x} = 1259, S^2 = \frac{570}{4}$，自由度为 $n - 1 = 4$，查 t 分布临界值表 $(\alpha = 0.05)$ 得 $\lambda = 2.776$，半径为

$$\lambda\sqrt{\frac{S^2}{n}} = 2.776 \times \sqrt{\frac{570}{4 \times 5}} \approx 14.8$$

- 置信区间为

$$[1259 - 14.8, 1259 + 14.8] = [1244.2, 1273.8]$$

例 5.3

- **例 5.3** 对飞机的飞行速度进行 15 次独立试验，测得飞机的最大飞行速度 ($\mathrm{m}\cdot\mathrm{s}^{-1}$) 如下：

$$422.2,\ 418.7,\ 425.6,\ 420.3,\ 425.8$$
$$423.1,\ 431.5,\ 428.2,\ 438.3,\ 434.0$$
$$412.3,\ 417.2,\ 413.5,\ 441.3,\ 423.7$$

根据长期经验，可以认为最大飞行速度服从正态分布。求最大飞行速度期望的置信区间。

- **解** 用 X 表示最大飞行速度。$D(X)$ 未知，求 $E(X)$ 的置信区间。
- 这里 $\bar{x} = 425.047, S^2 = \frac{1006.34}{14}$。
- 自由度 $n - 1 = 14$，查表得 $\lambda = 2.145$。
- 半径

$$\lambda\sqrt{\frac{S^2}{n}} = 2.145\sqrt{\frac{1006.34}{14 \times 15}} = 4.696$$

- 置信区间为

$$[425.047 - 4.696, 425.047 + 4.696] = [420.35, 429.74]$$

方差未知时求期望置信区间的步骤

- 由样本值 $x_1, x_2, \ldots, x_n$ 计算出 $\bar{x}, S^2$。
- 查 t 分布临界值表，自由度为 $n-1$，α 为 $1-$ 置信度，得临界值 λ。
- 计算半径

$$d = \lambda \sqrt{\frac{S^2}{n}}$$

- 得到 $E(X)$ 的置信度为 $1-\alpha$ 的置信区间为

$$[\bar{x} - d, \bar{x} + d]$$

本节目录

方差的置信区间

- 希望对方差 $D(X)$ 给出区间估计。方差本身也是一个重要指标。
- **例 6.1** 某自动车床加工零件，抽查 16 个零件，测得长度如下（单位: mm）:

$$12.15,\ 12.12,\ 12.01,\ 12.08,\ 12.09,\ 12.16$$
$$12.03,\ 12.01,\ 12.06,\ 12.13,\ 12.07,\ 12.11$$
$$12.08,\ 12.01,\ 12.03,\ 12.06$$

- 估计方差。$\bar{x} = 12.075, S^2 = 0.00244$。
- 如何给出方差的区间估计?

正态分布总体方差的置信区间

- 设总体 $X \sim \mathrm{N}(\mu, \sigma^2)$，$X_1, X_2, \ldots, X_n$ 是 X 的样本。由样本值 $x_1, x_2, \ldots, x_n$ 给出 σ^2 的置信区间。
- 已知 S^2 是 σ^2 的无偏估计，但不知道 S^2 与 σ 的具体偏离情况。
- 来求 $\frac{S^2}{\sigma^2}$ 的分布。
- 可以证明 $\eta = \frac{(n-1)S^2}{\sigma^2}$ 的分布密度为

$$p(u) = \begin{cases} \dfrac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} u^{\frac{n-3}{2}} e^{-\frac{u}{2}} & u > 0 \\ 0 & u \leq 0 \end{cases} \tag{6.1}$$

- $n \geq 3$ 时图形手绘示意。

卡方分布

- η 的分布叫做卡方分布，是一种特殊的伽玛分布。
- **定义 6.1** 如果随机变量 Y 的分布密度函数为

$$k_n(u) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} u^{\frac{n}{2}-1} e^{-\frac{u}{2}} & u > 0 \\ 0 & u \leq 0 \end{cases}$$

  则称 Y 服从 n 个自由度的**卡方分布**，记作 $Y \sim \chi^2(n)$。
- 易见 $\chi^2(n)$ 分布是 $\text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$ 分布。
- $\eta = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$。

正态分布总体方差置信区间（续）

- η 分布已知，可以取 $\lambda_1, \lambda_2\ (0 < \lambda_1 < \lambda_2)$，使得

$$P(\lambda_1 \leq \eta \leq \lambda_2) = 0.95 \tag{6.2}$$

- 一般选

$$\int_0^{\lambda_1} p(u)du = 0.025 \tag{6.3}$$

$$\int_{\lambda_2}^{\infty} p(u)du = 0.025 \tag{6.4}$$

- λ_1 和 λ_2 可以从 P433 的附表 3 查到（附表三给出的是卡方分布的右侧分位数）。

- 查出 λ_1, λ_2 后，以 95% 把握保证如下不等式：

$$\lambda_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq \lambda_2$$

- 即

$$\frac{(n-1)S^2}{\lambda_2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\lambda_1}$$

- 即

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\lambda_2} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\lambda_1}$$

这就是 σ^2 的置信度为 0.95 的置信区间。

- 为了得到标准差 σ 的置信区间，只要把 σ^2 的置信区间端点开平方根。

- **例 6.1（续）** 这里

$$\sum_{n=1}^{n}(x_i-\bar{x})^2=0.0366$$

$n=16$, 查 15 个自由度的卡方分布临界值得 $\lambda_1=6.26, \lambda_2=27.5$,

$$\begin{aligned}&\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\lambda_2}, \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\lambda_1}\right]\\ =&\left[\frac{0.0366}{27.5}, \frac{0.0366}{6.26}\right]\\ =&[0.0013, 0.0058]\end{aligned}$$

- σ 的置信区间为 $[0.036, 0.076]$。

本节目录

寻求置信区间和置信限的一般方法

- 概念：置信区间；置信水平（置信度）；置信系数。单侧的置信限。
- 方法：枢轴量方法；统计量方法；假设检验接受域方法。