

第五章、统计估值

例1.1. 钢铁厂一天生产了10000根16Mn型钢筋. 强度小于52kg/mm²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: 随机取一根, 结果用随机变量 X 表示:

$$X = \begin{cases} 1, & \text{是次品,} \\ 0, & \text{不是次品.} \end{cases} \quad \begin{cases} P(X = 1) = p, \\ P(X = 0) = 1 - p. \end{cases}$$

- 重点: p 未知, 此乃目标.
- 抽取少量(如, $n = 20, 50$)样品得到数据 X_1, \dots, X_n .
- 可认为它们独立, 且与 X 同分布.
- 重点: X_1, \dots, X_n 视为已知.
- 直观: $p \approx \bar{X}$.

例1.2. 灯泡厂生产了一批灯泡, 如何估计它们的平均寿命及其长短差异?

- 概率统计模型: 随机抽取一只灯泡, 其寿命视为随机变量 X .
- 设 $X \sim \text{Exp}(\lambda)$. λ 未知.
- 抽取 n 个样品得到 X_1, \dots, X_n .
- 可以认为它们独立同分布.
- 直观: 平均寿命 $EX \approx \bar{X}$, 长短差异 $\sigma^2 \approx ?$.

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.
- 抽样方法: 如何抽样, 抽多少, 怎样抽取;
- 统计推断: 得到抽样结果(一批数据)后, 如何分析、处理.

总体/总体分布.

- **总体**: 所研究的对象的全体. 如, 10000根钢筋, 一批灯泡.
- **个体**: 总体中的每一个.
- 关心每个个体的某一**特性值**(如, 钢筋的强度, 灯泡的寿命), 及其在总体中的**分布**情况(如, 强度 < 52 的(次品)在10000根钢筋中所占的比例, 寿命在1000~2000小时之间灯泡占这一批中的比例).
- 建模: 将个体特性值视为**随机变量 X** : 从总体中**随机抽取一个个体的特性值**. 目标: 研究 X 的**分布**.
- 简化: **总体**指 X , **总体分布**指 X 的分布. 如, 两点分布 $B(1, p)$, 指数分布 $\text{Exp}(\lambda)$.
- **重点**: 总体分布, 或, 其参数(如, p, λ), 是**未知的**. 此乃目标.

样本.

- 在一个**总体** (如, 10000根钢筋, 或10000根钢筋的强度) X 中, 随机抽取 n 个个体 X_1, \dots, X_n (如, n 个样品钢筋, 或其强度), 它们称为总体 X 的一个容量为 n 的**样本**(或叫**子样**). 称 n 为**样本(容)量**.
- 简化: 可将 X_1, \dots, X_n 视为 n 个**随机变量**.
- 抽取后, 得到 n 个具体的数值, 称为**样本值**, 记为 x_1, \dots, x_n .
- 注: 大小写可混用. 大写强调理论, 小写强调实用.
- 若样本 X_1, \dots, X_n 相互独立, 且均与总体 X 具有相同的分布, 则称其为**简单随机样本**.
- 例, 对总体 X 进行独立重复观测, 得到简单随机样本.
- 注: 总体数目很大时, 可认为无放回抽样得到简单随机样本.

数学模型(定义1.1).

- 总体: 随机变量 X . **重点:** 其分布未知, 是我们关心的.
- 样本: 独立且与 X 同分布的随机变量 X_1, \dots, X_n .
- 样本值: x_1, \dots, x_n , 抽样/试验完成后, 样本的取值. 即, 数据.
重点: x_1, \dots, x_n 视为已知.
- 样本容量: n .
- 若 X 有分布密度 $p(x)$, 则称 X_1, X_2, \dots, X_n 是来自总体 $p(x)$ 的样本.
- 定理1.1. 若 X_1, \dots, X_n 是来自总体 $p(x)$ 的样本, 则 (X_1, X_2, \dots, X_n) 有联合密度

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

§5.2 分布函数与分布密度的估计

- 描述随机变量的分布: 分布函数、密度函数、分布列.
- 给定样本值 x_1, x_2, \dots, x_n , 如何估计分布函数 $F(x)$?
- 固定 x , 令 $A = \{X \leq x\}$. 则 $F(x) = P(A)$.
- 大数定律: $P(A) \approx A$ 的频率.
- 定义2.1. 给定 X 的样本(值) x_1, x_2, \dots, x_n . 称 x 的函数

$$F_n(x) = \frac{\nu_n}{n}, \quad \nu_n = |\{i : 1 \leq i \leq n \text{ 且 } x_i \leq x\}|$$

为 X 的经验分布函数.

- 注: 固定 x , $F_n(x)$ 即为 A 的频率. 故 $F_n(x) \approx F(x)$, (n 很大).

- 将样本值 x_1, x_2, \dots, x_n 从小到大排列后记为

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

$x_{(i)}$ 叫做样本的第 i 个次序统计量.

- 若 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ 两两不同, 易见

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ \frac{k}{n}, & x_{(k)} \leq x < x_{(k+1)}, \quad (k = 1, 2, \dots, n-1) \\ 1, & x \geq x_{(n)}. \end{cases}$$

- 若某个 $x_{(j)}$ 有 m 个相同的样本值, 则 $F_n(x)$ 在 $x_{(j)}$ 处向上跳跃 $\frac{m}{n}$ 即可.
- 分位数估计: 若 $x_p = F^{-1}(p)$ 存在唯一. 取 $r_n = [pn] + 1$, 则 $x_{(r_n)} \rightarrow x_p, \text{ a.s.}$

例2.1. 罐头净重是随机的, 额定345克. 随机抽取10个, 得到数据:

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

试估计分布函数 $F(x)$ 及其中位数.

● 解: 用经验分布函数 $F_n(x)$ 估计 $F(x)$.

● 样本值从小到大排列:

336, 338, 340, 342,

343, 343, 344, 344, 344, 345

● 经验分布函数:

● 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 $\frac{1}{2}(x_{(5)} + x_{(6)}) = 343$, 或 $x_{([0.5 \times 10]+1)} = x_{(6)} = 343$.

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

分布密度估计

直方图法.

- 总体密度: $p(x)$. 样本: x_1, x_2, \dots, x_n .
- 步骤一、取 a 比 $x_{(1)}$ 略小, b 比 $x_{(n)}$ 略大. 把区间 $(a, b]$ 等分为 m 个小区间 I_1, \dots, I_m . 区间长: $h = \frac{b-a}{m}$.
- 步骤二、 $\forall i$, 统计出 $\nu_i = \{k : x_k \in I_i\}$, 计算出 $f_i = \frac{\nu_i}{n}$.
- 步骤三、做直方图. 在 $x \in I_i$ 上, 作高为 $\hat{p}(x) = \frac{f_i}{h}$ 的矩形框.
- 结论: 用 $\hat{p}(x)$ 来估计 $p(x)$.
- 理由: 第 i 个框的面积 $f_i \approx P(X \in I_i)$ (频率 \approx 概率), 从而

$$\hat{p}(x) \approx \frac{1}{|I_i|} P(X \in I_i) \approx p(x), \quad \forall x \in I_i.$$

- 注: 不必等分区间, 近似即可.

- a, b, m 的取法: 使多数小区间中包含样本值; 小数位数比观测值的小数位数多一位, 避免样本值落在小区间端点.

- m 的建议取法一:

$$m \approx 1 + 3.322 \ln n.$$

- 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	$10 \sim 20$

- 若 $p(x)$ 一致连续; $\exists \delta > 0$, 使得

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

且小区间长度 h_n 满足:

$$\lim_{n \rightarrow \infty} h_n = 0, \quad h_n \geq \frac{(\ln n)^2}{n}.$$

则一致强相合:

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

例2.2. $n = 120$. 数据: $0.86, 0.83, \dots$

- $x_{(1)} = 0.64, x_{(n)} = 0.95, x_{(n)} - x_{(1)} = 0.31$.
- 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, 统计出 $\nu_i, i = 1, 2, \dots, m$.
- 作图: 高为 $\frac{f_i}{h}$.

核估计.

- 理由1: 若 $p(x)$ 连续, 则 h 很小时,

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

- 理由2: 可用经验分布函数 $F_n(x)$ 估计 $F(x)$.
- **Rosenblatt估计**: 用 $\hat{p}_n(x)$ 来估计 $p(x)$.

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

- 注: $F_n(x+h) - F_n(x-h) = \{k : x_k \in (x-h, x+h]\}$, 思想与直方图法相似, 区别在于小区间.
- 记 $K_0(x) = \frac{1}{2}$, 若 $-1 \leq x < 1$; $K_0(x) = 0$, 其他. 则

$$\hat{p}_n(x) = \frac{1}{2nh} |\{i : x-h < x_i \leq x+h\}| = \frac{1}{nh} \sum_{i=1}^n K_0\left(\frac{x-x_i}{h}\right).$$

- **核函数**(定义2.2): $K(x)$ 为非负函数, $\int_{-\infty}^{\infty} K(x)dx = 1$.
- 此时, 称 $\tilde{p}_n(x)$ 为 $p(x)$ 的核估计, 其中

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

- 注: 一般选为偶函数, 且在正半轴单调下降(类似正态曲线).
- 其他常用核函数: 如,

$$K_2(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \quad K_3(x) = \frac{1}{\pi(1+x^2)}.$$

- 若 $p(x)$ 一致连续; $\lim_{n \rightarrow \infty} h_n = 0$, $\sum_{n=1}^{\infty} e^{-rnh_n^2} < \infty$, $\forall r > 0$;
且核函数 $K(x)$ 为有界变差函数, 则**一致强相合**:

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

- 注: 在 $K(x)$ 和 h 选取合适时, 比直方图的估计精度更高.

最近邻估计.

- 核估计: 固定区间长 $2h$, x 附近的样本点多则密度大.
- 最近邻估计: 固定 x 附近所需的样本点数, 所需的邻域区间越短则密度越大.
- 数据: x_1, \dots, x_n .
- 取正整数 $K(n)$, 令

$$a_n(x) = \min \{t : t > 0, |\{i : x_i \in (x - t, x + t)\}| \geq K(n)\}$$
$$p_n^*(x) = \frac{K(n)}{n} \cdot \frac{1}{2a_n(x)}.$$

- 若 $p(x)$ 一致连续, $\lim_{n \rightarrow \infty} \frac{K(n)}{n} = 0$ 且 $\lim_{n \rightarrow \infty} \frac{K(n)}{\ln n} = \infty$, 则一致强相合:

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

§5.3 最大似然估计法

- 经验分布函数、直方图估计、核密度估计、最近邻密度估计等, 都不需要假定总体分布的类型, 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, 则只需估计参数. 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 $N(\mu, \sigma^2)$, 但 μ, σ^2 未知.
- 又如, 设产品寿命服从威布尔分布/对数正态分布, 参数未知.
- 设总体 X 的密度函数或概率函数(分布列)为 $p(x; \theta_1, \theta_2, \dots, \theta_m)$, 其中 $\theta_1, \theta_2, \dots, \theta_m$ 是未知参数.
- 问: 根据样本值: x_1, x_2, \dots, x_n . 如何估计参数?

- 参数: $\theta = (\theta_1, \dots, \theta_m)$. 范围: Θ . 样本: $\vec{x} = (x_1, \dots, x_n)$.
- 似然函数:

$$L(\theta) = L_n(\theta) = L(\vec{x}; \theta) = L_n(\vec{x}; \theta) := \prod_{i=1}^n p(x_i; \theta).$$

- 注1: θ 是参数. $L(\vec{x}; \theta)$ 是实向量 \vec{x} 的函数, 它即为 $\vec{X} = (X_1, \dots, X_n)$ 的联合分布列/密度函数.
- 注2: \vec{x} 是 \vec{X} 的取值. $Y = L(\vec{X}; \theta)$ 是随机变量. 如, $X \sim N(\mu, \sigma^2)$, 则 $Y = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$.
- 注3: 样本值 \vec{x} 为已知, 视为常数. 似然函数是参数 θ 的函数.
- 定义3.1. 如果 $L_n(\vec{x}; \theta)$ 在 $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$ 达到最大值, 则称 $\hat{\theta}$ (或 $\hat{\theta}_i$) 为参数 θ (或 θ_i) 的最大似然估计(MLE).
- 注: $\hat{\theta}$ 依赖于 \vec{x} , 故为 $\theta(\vec{x})$.
- 注: 称 $g(\hat{\theta})$ 为 $g(\theta)$ 的最大似然估计.

例. 盒中有许多黑球和白球, 比例为3 : 1, 猜哪种多.

- 直观: 随机取3个, 见到哪种多就猜哪种多.
- 数据: $n = 1, X \sim B(3, p), p \in \{\frac{1}{4}, \frac{3}{4}\}$,

似然函数:

$$L(p) = C_3^x p^x (1-p)^{3-x}, \quad 0 \leq x \leq 3.$$

$p \backslash x$	0	1	2	3
$\frac{1}{4}$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$
$\frac{3}{4}$	$\frac{1}{64}$	$\frac{9}{64}$	$\frac{27}{64}$	$\frac{27}{64}$
$\hat{\theta}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{3}{4}$

- 理论分析: $\hat{\theta}(\vec{X})$ 是随机变量.
- 在相当一般的条件下, 有如下性质:
 - 相合性: $n \rightarrow \infty$, 估计结果与参数真值无限接近.
 - 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, $\hat{\theta}(\vec{X})$ 近似服从正态分布.
- $L_n(\theta)$ 与对数似然函数 $\ln L_n(\theta)$ 的最大值点相同.
- $\theta = (\theta_1, \dots, \theta_m) \in \Theta$ 满足如下似然方程组:

$$\frac{\partial \ln L_n}{\partial \theta_1} = \dots = \frac{\partial \ln L_n}{\partial \theta_m} = 0.$$

- 注: 似然方程组的解不能保证为最大值点.

指数分布

$X \sim \text{Exp}(\lambda)$. 参数: λ , 范围: $(0, \infty)$. 求 λ 的最大似然估计 $\hat{\lambda}$.

- 密度: $p(x; \lambda) = \lambda e^{-\lambda x}$, $x > 0$.
- 似然函数、对数似然函数: $x_i \geq 0$, $\forall x_i$.

$$L_n(\vec{x}; \lambda) = \lambda^n \exp\{-\lambda \sum_{i=1}^n x_i\}, \quad \ln L_n(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^n x_i.$$

- 似然方程: 记 $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$,
 $(\ln L_n(\lambda))' = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \Rightarrow \hat{\lambda} = n / \sum_{i=1}^n x_i = 1/\bar{x}$.
- **例3.1.** $n = 18$, 数据: 16, 29, 50, 68, 100, 130, 140, 270, 280, 240, 410, 450, 520, 620, 190, 210, 800, 1100.
代入数据: $\bar{x} = 318$, $\hat{\lambda} = \frac{1}{318} \approx 0.03144$.

- 进一步, $\mu = EX = \frac{1}{\lambda}$ 的最大似然估计为 $\hat{\mu} = \frac{1}{\hat{\lambda}} = \bar{X}$.

正态分布

$X \sim N(\mu, \sigma^2)$. 参数 $\mu, \delta = \sigma^2$, 范围: $\mu \in \mathbb{R}, \delta > 0$.

求 μ, δ 的最大似然估计 $\hat{\mu}$ 与 $\hat{\delta} = \hat{\sigma}^2$.

- 密度: $p(x; \mu, \delta) = \frac{1}{\sqrt{2\pi\delta}} \exp\left\{-\frac{1}{2\delta}(x - \mu)^2\right\}$.
- 似然函数、对数似然函数:

$$L(\vec{x}; \mu, \delta) = \frac{1}{\sqrt{2\pi\delta}^n} \exp\left\{-\frac{1}{2\delta} \sum_{i=1}^n (x_i - \mu)^2\right\},$$

$$\ln L_n(\mu, \delta) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \delta - \frac{1}{2\delta} \sum_{i=1}^n (x_i - \mu)^2.$$

- 似然方程组:

$$\frac{\partial \ln L_n}{\partial \mu} = \frac{1}{\delta} \sum_{i=1}^n (x_i - \mu) = 0,$$

$$\frac{\partial \ln L_n}{\partial \delta} = -\frac{n}{2\delta} + \frac{1}{2\delta^2} \sum_{i=1}^n (x_i - \mu)^2 = 0.$$

- 解得: $\hat{\mu} = \bar{x}, \hat{\delta} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. (确实是最大值点.)

威布尔分布

参数: m, η , 范围: $m > 0, \eta > 0$. 密度:

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

求 m, η 的最大似然估计 $\hat{m}, \hat{\eta}$.

- 对数似然函数:

$$\ln L_n(m, \eta) = n \ln m - nm \ln \eta + (m-1) \sum_{i=1}^n \ln x_i - \sum_{i=1}^n \left(\frac{x_i}{\eta} \right)^m.$$

- 似然方程组: ($a^m = e^{m \ln a}$).

$$(3.2a) \quad \frac{\partial \ln L_n}{\partial m} = \frac{n}{m} - n \ln \eta + \sum_{i=1}^n \ln x_i - \sum_{i=1}^n \left(\frac{x_i}{\eta} \right)^m \ln \frac{x_i}{\eta} = 0,$$

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

- 由(3.2b)得 $\eta = \left(\frac{1}{n} \sum_{i=1}^n x_i^m \right)^{\frac{1}{m}}$.
- 再代入(3.2a).

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

$$(3.2a) \quad \frac{\partial \ln L_n}{\partial m} = \frac{n}{m} - n \ln \eta + \sum_{i=1}^n \ln x_i - \sum_{i=1}^n \left(\frac{x_i}{\eta}\right)^m \ln \frac{x_i}{\eta} = 0.$$

- 整理:

$$\begin{aligned} \star &= \frac{1}{\eta^m} \sum_{i=1}^n x_i^m (\ln x_i - \ln \eta) = \frac{1}{\eta^m} \sum_{i=1}^n x_i^m \ln x_i - \frac{1}{\eta^m} \sum_{i=1}^n x_i^m \ln \eta \\ &= \frac{1}{\eta^m} \cdot \star - \frac{1}{\eta^m} n \eta^m \ln \eta = \frac{1}{\eta^m} \cdot \star - \star. \end{aligned}$$

- 故, (3.2a) 化为

$$\varphi(m) := \frac{1}{m} + \frac{1}{n} \sum_{i=1}^n \ln x_i - \frac{\sum_{i=1}^n x_i^m \ln x_i}{\sum_{i=1}^n x_i^m} = 0. \quad (3.4)$$

- 当 $n \geq 2$, x_1, \dots, x_n 不完全相等时, 上述方程恰有一个根 \hat{m} .
- 注: $\hat{m}(\vec{x})$ 没有显示表达. $\varphi'(m) < 0$, 可用二分法找到 \hat{m} .
- 进一步, $\hat{\eta} = \left(\frac{1}{n} \sum_{i=1}^n x_i^{\hat{m}}\right)^{\frac{1}{\hat{m}}}$.
- 注: 可以证明 $(\hat{m}, \hat{\eta})$ 是最大似然估计.

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

求形状参数 m 和刻度参数 η 的最大似然估计.

- 解: 解方程(3.4) 得 m 的最大似然估计 $\hat{m} = 1.9$.

进一步, $\hat{\eta} = \left(\frac{1}{n} \sum_{i=1}^n x_i^{\hat{m}}\right)^{\frac{1}{\hat{m}}} = 685$.

均匀分布

设 $X \sim U[a, b]$, 参数: a, b , 范围: $a < b$.

求: a, b 的最大似然估计 \hat{a}, \hat{b} .

- 密度: $p(x; a, b) = \frac{1}{b-a}$, 若 $a \leq x \leq b$; $p(x; a, b) = 0$, 否则.
- 似然函数:

$$L_n(\vec{x}; a, b) = \begin{cases} \frac{1}{(b-a)^n}, & a \leq x_{(1)} \text{ 且 } x_{(n)} \leq b, \\ 0, & \text{否则.} \end{cases}$$

- 故, $\hat{a} = x_{(1)}, \hat{b} = x_{(n)}$.

§5.4 期望与方差的点估计

- 直接估计分布函数、分布密度、概率函数要求数据很多.
- 参数的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.
- 例1.1. 钢筋次品率估计问题.
等价地, 总体 $X \sim B(1, p)$ 的期望 EX 估计问题.

- 用**样本均值** $\bar{X}_n = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 来估计 EX .
- \bar{X} 是随机变量.
- 定理4.1. 设 EX 存在, 则 $E\bar{X} = EX$.
- 注: 称这样的估计为**无偏估计**.
- 定理4.2. 设 X 的期望、方差都存在, 则 $D(\bar{X}_n) = \frac{1}{n}D(X)$.
- 注: n 越大, $D(\bar{X}_n)$ 越小, 估计越精确, 称为**越有效**.
- 若 $\psi(x_1, \dots, x_n)$ 是不依赖于未知参数的函数, 则称 $Y = \psi(X_1, \dots, X_n)$ 为**统计量**.
- 如, \bar{X} 是统计量.
- 统计量是随机变量, 其分布称为**抽样分布**.

- 设 $g(\theta)$ 是参数 θ 的函数, X_1, \dots, X_n 是 X 的样本.
- 定义4.1. 称统计量 $\varphi(X_1, \dots, X_n)$ 为 $g(\theta)$ 的估计量.
- 注: 一种估计方法产生 $\varphi_n(X_1, \dots, X_n), n = 1, 2, \dots$
- $\varphi(X_1, \dots, X_n)$ 的分布/期望与 θ 有关, 为此显式地记为

$$E_{\theta} [\varphi(X_1, \dots, X_n)].$$

- 定义4.2. 若 $E_{\theta} [\varphi(X_1, \dots, X_n)] = g(\theta), \forall \theta \in \Theta$, 则称 $\varphi(X_1, \dots, X_n)$ 为 $g(\theta)$ 的无偏估计,
- 定义4.3. 若 $g(\theta)$ 的两个估计量满足

$$E_{\theta} [\varphi_1(X_1, \dots, X_n) - g(\theta)]^2 \leq E_{\theta} [\varphi_2(X_1, \dots, X_n) - g(\theta)]^2, \forall \theta \in \Theta,$$

且存在 θ_0 使LHS < RHS, 则称 φ_1 比 φ_2 有效.

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

$$D(\varphi(X_1, \dots, X_n)) \leq D(\psi(X_1, \dots, X_n)), \quad \forall \theta \in \Theta$$

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

- 定义. 若对任意 $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\varphi(X_1, \dots, X_n) - g(\theta)| > \varepsilon) = 0.$$

则称 $\varphi(X_1, \dots, X_n)$ 是 $g(\theta)$ 的相合估计.

- 定义. 若

$$P\left(\lim_{n \rightarrow \infty} \varphi(X_1, X_2, \dots, X_n) = g(\theta)\right) = 1,$$

则称 $\varphi(X_1, \dots, X_n)$ 是 $g(\theta)$ 的强相合估计.

方差的点估计

- 用如下定义的**样本方差估计** $D(X)$.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

- 定理4.3. 设 X 的方差存在, 则 $E(S^2) = D(X)$.
- 证:
$$\begin{aligned} \frac{n-1}{n} S^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - 2\bar{X}\bar{X} + \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2, \end{aligned}$$
故,
$$E \frac{n-1}{n} S^2 = \underbrace{E X^2}_{D(\bar{X})} - \underbrace{(E \bar{X})^2}_{(E \bar{X})^2} = \underbrace{D(X)}_{D(X)} - \underbrace{\frac{1}{n} D(X)}_{\frac{1}{n} D(X)}.$$
即,
$$E \frac{n-1}{n} S^2 = \frac{n-1}{n} D(X).$$
- 注: 若 x_i 's 罗列了总体, 则 $D(X) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$.
- 用 $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ 估计标准差 σ .
一般情况下, $(ES)^2 < ES^2 = \sigma^2$, 即 $ES < \sigma$.

- 设 X 的分布密度是 $p(x; \theta_1, \dots, \theta_m)$.
- ν_k 是 X 的 k 阶矩: $\nu_k = E_{\theta} X^k = g_k(\theta_1, \dots, \theta_m)$.
- 设 $\nu_1, \nu_2, \dots, \nu_m$ 已知, 则可从方程组

$$\begin{cases} g_1(\theta_1, \dots, \theta_m) = \nu_1 \\ \dots\dots\dots \\ g_m(\theta_1, \dots, \theta_m) = \nu_m \end{cases} \quad \text{解得} \quad \begin{cases} \theta_1 = f_1(\nu_1, \dots, \nu_m) \\ \dots\dots\dots \\ \theta_m = f_m(\nu_1, \dots, \nu_m) \end{cases}$$

- 样本: x_1, \dots, x_n .
用样本矩 $\hat{\nu}_k = \frac{1}{n} \sum_{i=1}^n x_i^k$ 来估计 $\nu_k, k = 1, 2, \dots, m$.
- 用 $\hat{\theta}_k = f_k(\hat{\nu}_1, \hat{\nu}_2, \dots, \hat{\nu}_m)$ 估计 $\theta_k, k = 1, 2, \dots, m$.

例4.2. $X \sim N(\mu, \sigma^2)$, 求 μ, σ^2 的矩估计.

- 列方程组: $\nu_1 = \mu, \nu_2 = EX^2 = D(X) + (EX)^2 = \sigma^2 + \mu^2$.
- 反解: $\mu = \nu_1, \sigma^2 = \nu_2 - \nu_1^2$.
- 估计: $\hat{\nu}_1 = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \hat{\nu}_2 = \frac{1}{n} \sum_{i=1}^n x_i^2$.
- 代入解: $\hat{\mu} = \hat{\nu}_1 = \bar{x}, \hat{\sigma}^2 = \hat{\nu}_2 - \hat{\nu}_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.
- 与最大似然估计相同.

例4.3. $X \sim U(0, \theta)$. 求 θ 的矩估计.

- 方程: $\nu_1 = EX = \frac{\theta}{2}$.
- 反解: $\theta = 2\nu_1$.
- 估计: $\hat{\nu}_1 = \bar{X}$, $\theta = 2\bar{X}$.
- 矩估计 $2\bar{X}$ vs 最大似然估计 $X_{(n)}$.

(1) 合理性: 有可能 $2\bar{X} < X_{(n)}$.

(2) 无偏性: $E(2\bar{X}) = 2EX = \theta$,

$$EX_{(n)} = \frac{n}{n+1}\theta, \text{ 调整: } \tilde{\theta} = \frac{n+1}{n}X_{(n)}.$$

(3) 有效性: $D(2\bar{X}) = 4 \cdot \frac{1}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}$; $D(\tilde{\theta}) = ??$

(4) 相合性: 都有.

例4.4. 降雨量 $X \sim \Gamma(\alpha, \beta)$. $n = 36$, 数据: \dots . 求 α, β 的矩估计.

- 密度: $p(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x > 0.$

- 列方程: $\nu_1 = \frac{\alpha}{\beta}, \nu_2 = \frac{\alpha(\alpha+1)}{\beta^2}.$

- 代数据: $\hat{\nu}_1 = 7.29, \hat{\nu}_2 = 85.59$. 求解如下方程组:

$$\frac{\hat{\alpha}}{\hat{\beta}} = 7.29, \quad \frac{\hat{\alpha}(\hat{\alpha} + 1)}{\hat{\beta}^2} = 85.59.$$

- 解得: $\hat{\alpha} = 1.64, \hat{\beta} = 0.22.$

§5.5 期望的置信区间

- 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 $N(\mu, \sigma^2)$** 的区间估计:
 - (1) 已知方差 σ^2 , 对 $\mu = EX$ 进行区间估计;
 - (2) 未知 σ^2 , 对 μ 进行区间估计;

σ^2 已知, 估计 μ .

- 枢轴量: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$,

$$Z = Z(\bar{X}, \mu) = \bar{X}^* = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

- 查表: $P(|Z| \leq 1.96) = 0.95$.

$$P\left(|\bar{X} - \mu| \leq 1.96 \cdot \sigma/\sqrt{n}\right) = 0.95.$$

- 概率角度: $\bar{X} \in [\mu - \varepsilon, \mu + \varepsilon]$, $\varepsilon = 1.96 \cdot \sigma/\sqrt{n}$.
- 统计角度: $\mu \in [\bar{X} - \varepsilon, \bar{X} + \varepsilon]$.
- 置信区间: 随机区间; 置信度/水平: $1 - \alpha = 0.95$.
- ** 对应的置信区间最短; n 越大, 置信区间越短.
- 非正态情形. Z 近似服从 $N(0, 1)$. 仍可用 $[\bar{X} - \varepsilon, \bar{X} + \varepsilon]$.

例5.1. 滚珠直径 $X \sim N(\mu, 0.05)$ (单位: mm). $n = 6$, 数据: 14.70, 15.21, 14.90, 14.91, 15.32, 15.32. 求: μ 的区间估计.

- 代数据:

$$\bar{x} = \frac{1}{6}(14.70 + 15.21 + 14.90 + 14.91 + 15.32 + 15.32) = 15.06.$$

- 半径: $\varepsilon = 1.96\sqrt{\sigma^2/n} = 1.96\sqrt{0.05/6} = 0.18.$

- 置信度为0.95 的置信区间为

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

σ^2 未知(讨厌参数), 估计 μ .

- 用样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 代替 σ^2 .

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

- 需要推导 T 的分布. 以下推导过程可见附录二定理5 ~ 7.

- $X_i = \mu + \sigma Z_i$, $\bar{X} = \mu + \sigma \bar{Z}$; $X_i - \bar{X} = \sigma(Z_i - \bar{Z})$,

- 分子: $\sqrt{n}(\bar{X} - \mu) = \sigma\sqrt{n}\bar{Z}$,

分母: $\sqrt{S^2} = \sigma\sqrt{\tilde{S}^2}$, 其中 $\tilde{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$.

- 取正交矩阵 $\mathbf{B}_{n \times n}$, 使得 $b_{1i} = \frac{1}{\sqrt{n}}$, $\forall i$.

$$\vec{Y} = \mathbf{B}\vec{Z} \sim N(\vec{0}, \mathbf{I}), \quad Y_1 = \sqrt{n}\bar{Z}.$$

- 结论1:** 分子与分母独立.

$$\sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2.$$

- **结论1:** \bar{X} 与 S^2 独立. $T = \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{S^2}} = \frac{Y_1}{\sqrt{\tilde{S}^2}}$,

其中, $\tilde{S}^2 = \frac{1}{n-1} \sum_{i=2}^n Y_i^2$, Y_1, \dots, Y_n i.i.d., $\sim N(0, 1)$.

- $Y_i^2 = \Gamma(\frac{1}{2}, \frac{1}{2})$. (§2.4, 例).
- 引理: 若 X, Y 独立, $X \sim \Gamma(\alpha, \beta)$, $Y \sim \Gamma(\alpha', \beta)$, 则

$$X + Y \sim \Gamma(\alpha + \alpha', \beta).$$

- 证明: 用§4.2, (2.2) 计算得到.
- 推论: $Y_2^2 + \dots + Y_n^2 \sim \Gamma(\frac{n-1}{2}, \frac{1}{2})$.
- 称 $\Gamma(\frac{n}{2}, \frac{1}{2})$ 为自由度为 n 的卡方分布, 记为 $\chi^2(n)$.

$\chi^2(n)$ 的分布密度如下: (定义6.1):

$$q_n(u) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} u^{\frac{n}{2}-1} e^{-\frac{u}{2}}, \quad u > 0. \quad (6.1)$$

- **结论2:** $\sum_{i=2}^n Y_i^2 \sim \chi^2(n-1)$.

- 结论1: \bar{X} 与 S^2 独立. $T = \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{S^2}} = \frac{Y_1}{\sqrt{\tilde{S}^2}}$,

其中, $\tilde{S}^2 = \frac{1}{n-1} \sum_{i=2}^n Y_i^2$, Y_1, \dots, Y_n i.i.d., $\sim N(0, 1)$.

- 结论2: $\sum_{i=2}^n Y_i^2 \sim \chi^2(n-1)$.

- 设 Z 与 K_n 独立, $Z \sim N(0, 1)$, $K_n \sim \chi^2(n)$. 记 $W = \frac{Z}{\sqrt{K_n/n}}$.

- 自由度为 n 的 t 分布, 记为 $t(n)$, 密度如下. 往证 $W \sim t(n)$.

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

- $F_W(t) = P(Z \leq t\sqrt{K_n/n}) = \int_{-\infty}^{\infty} \int_{-\infty}^{t\sqrt{x/n}} \phi(z)q_n(x)dzdx$
 $= \int_{-\infty}^t \int_{-\infty}^{\infty} \phi(w\sqrt{x/n})q_n(x)\sqrt{x/n}dx dw, \quad (z = w\sqrt{x/n}).$

- $p_n(t) = C \int_0^{\infty} e^{-\frac{t^2 x}{2n}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} x^{\frac{1}{2}} dx$
 $= C \int_0^{\infty} e^{-\frac{1}{2}\left(1+\frac{t^2}{n}\right)x} x^{\frac{n-1}{2}} dx = \hat{C} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$

- 结论3: $T \sim t(n-1)$.

关于 T 的分布的总结:

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

- **结论1:** \bar{X} 与 S^2 独立, (分子与分母独立).

分子: $\sqrt{n}(\bar{X} - \mu) = \sigma Y_1$, 其中 $Y_1 \sim N(0, 1)$.

- **结论2:** 分母: $\sqrt{S^2} = \sigma \sqrt{\tilde{S}^2}$, 其中 $\tilde{S}^2 = \frac{1}{n-1} \sum_{i=2}^n Y_i^2$,
 $\sum_{i=2}^n Y_i^2 \sim \chi^2(n-1)$.

- **结论3:** $T \sim t(n-1)$.

- 注1: $t(n)$ 的密度 $p_n(t)$ 是偶函数.

- 注2: $p_n(t)$ 形如 $\phi(t)$; 尾更厚, 即 $P(|T_n| > x) \geq P(|Z| > x)$.

因此, 若 $P(Z \in [-z, z]) = P(T_n \in [-t, t]) = 95\%$, 则 $t > z$.

- 注3: 可以证明 $\lim_{n \rightarrow \infty} p_n(t) = \phi(t)$, $\forall t$.

- 枢轴量:

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S^2}} \sim t(n-1).$$

- 查表: $\lambda = t_{1-\alpha/2}(n-1)$, 即 $P(|T| \leq \lambda) = 1 - \alpha = 95\%$.

注: λ 叫做t分布双侧 $\alpha = 0.05$ 临界值.

- μ 的置信度为95% 的置信区间:

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

例5.2. 测量温度(单位:度). $n = 5$ 次, 数据: 1250, 1265, 1245, 1260, 1275. 假设仪器没有系统偏差, 求: 真值的范围, ($\alpha = 0.05$).

- μ : 真值. (合理地)假设 $X \sim N(\mu, \sigma^2)$.
- μ 的置信区间为

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

- 代入数据: 计算得 $\bar{x} = 1259$, $s^2 = S^2(\vec{x}) = \frac{570}{4}$.
- 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. 最大飞行速度的 $n = 15$ 个测量数据(单位: 米/秒): 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. 求: 最大飞行速度的置信区间. ($\alpha = 0.05$)

- 注: 根据长期经验, 可以认为最大飞行速度 $X \sim N(\mu, \sigma^2)$.
- 代入数据: $\bar{x} = 425.047$, $s^2 = \frac{1006.34}{14}$.
- 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

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

μ 的区间估计总结.

σ^2 已知:

- (1) 由样本值 x_1, \dots, x_n 计算出 \bar{x} .
- (2) 查 $N(0, 1)$ 分布表, 得临界值 $\lambda = z_{1-\alpha/2}$.
- (3) 算出半径 $\varepsilon = \lambda\sqrt{\sigma^2/n}$.
- (4) 置信区间为 $[\bar{x} - \varepsilon, \bar{x} + \varepsilon]$.

σ^2 未知:

- (1) 由样本值 x_1, \dots, x_n 计算出 \bar{x}, s^2 .
- (2) 查 $t(n-1)$ 分布表, 得临界值 $\lambda = t_{1-\alpha/2}(n-1)$.
- (3) 算出半径 $\varepsilon = \lambda\sqrt{s^2/n}$.
- (4) 置信区间为 $[\bar{x} - \varepsilon, \bar{x} + \varepsilon]$.

§5.6 方差的置信区间

总体 $X \sim N(\mu, \sigma^2)$. 求 σ^2 的区间估计. (置信度 $1 - \alpha = 0.95$).

- 样本方差: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.
- 枢轴量: $K_{n-1} = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.
- 取 $\lambda_1 = \chi_{\alpha/2}^2(n-1)$, $\lambda_2 = \chi_{1-\alpha/2}^2(n-1)$ 使得
$$P(\lambda_1 \leq K_{n-1} \leq \lambda_2) = 1 - \alpha = 0.95.$$

- σ^2 的置信区间:

$$\left[\sum_{i=1}^n (X_i - \bar{X})^2 / \lambda_2, \sum_{i=1}^n (X_i - \bar{X})^2 / \lambda_1 \right].$$

- 取 $\lambda = \chi_{1-\alpha}^2(n-1)$, 则 $P(\sigma^2 \leq \bar{\delta}) = 1 - \alpha = 0.95$.

置信上限: $\bar{\delta} = \sum_{i=1}^n (X_i - \bar{X})^2 / \lambda$.

- 注: σ 的置信区间/上限, 端点开平方根即可.
- 注: 若 μ 已知, 则枢轴量 $K_n = \sum_{i=1}^n (X_i - \mu)^2 / \sigma^2 \sim \chi^2(n)$.

例6.1. 某自动车床加工零件, 设零件的长度 $X \sim N(\mu, \sigma^2)$, (单位: mm). $n = 16$ 个测量数据: 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. 求 σ^2 的区间估计.

- 代数据: $\bar{x} = 12.075$, $(n - 1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = 0.0366$.
- 查 $\chi^2(15)$ 的表: 得 $\lambda_1 = 6.26$, $\lambda_2 = 27.5$.
- σ^2 的置信区间: $[\sum_{i=1}^n (x_i - \bar{x})^2 / \lambda_2, \sum_{i=1}^n (x_i - \bar{x})^2 / \lambda_1]$.

$$\left[\frac{0.0366}{27.5}, \frac{0.0366}{6.26} \right] = [0.0013, 0.0058].$$

- 注: σ 的置信区间为 $[0.036, 0.076]$.

§5.7 寻求置信区间和置信限的一般方法

- 概念: 置信区间/置信限、置信水平(置信度)、置信系数.
- 方法: 枢轴量方法、统计量方法、假设检验接受域方法.