

Ch 7.2 中心极限定理



回顾前一次课

随机投影: $f(x) = xP/c$ 其中 P 是 $d \times k$ 的随机矩阵, 其每个元素之间相互独立, c 为常数

JL-引理: 设 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ 为 d 维空间的 n 个点, 随机矩阵 $P = (p_{ij})_{d \times k}$, 每个元素相互独立且 $p_{ij} \sim N(0,1)$, 令

$$\mathbf{y}_i = f(\mathbf{x}_i) = \mathbf{x}_i P / \sqrt{k}$$

对任意 $\epsilon \in (0, 1/2)$, 当 $k \geq 8 \ln 2n / (\epsilon^2 - \epsilon^3)$ 时, 对任意 $i \neq j$, 以至少 $1/2$ 的概率有

$$(1 - \epsilon) |\mathbf{x}_i - \mathbf{x}_j|_2^2 \leq |\mathbf{y}_i - \mathbf{y}_j|_2^2 \leq (1 + \epsilon) |\mathbf{x}_i - \mathbf{x}_j|_2^2$$

回顾前一次课

Markov 大数定律：若随机变量序列 $\{X_i\}$ 满足 $\text{Var}(\sum_{i=1}^n X_n)/n^2 \rightarrow 0$, 则满足大数定律

Chebyshev 大数定律：若独立随机变量序列 $\{X_i\}$ 满足 $\text{Var}(X_i) \leq c$, 则满足大数定律

Khinchine 大数定律：若独立同分布随机变量序列 $\{X_i\}$ 期望存在, 则满足大数定律;

Bernoulli 大数定律：对二项分布 $X_n \sim B(n, p)$, 有 $X_n/n \xrightarrow{P} p$

中心极限定理的问题

对独立的随机变量序列 $X_1, X_2, \dots, X_n, \dots$, 考虑标准化后随机变量

$$Y_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n E(X_i)}{\sqrt{\text{Var}(\sum_{i=1}^n X_i)}}$$

的极限分布是否为服从正态分布

依分布收敛

设随机变量 Y 的分布函数为 $F_Y(y) = P(Y \leq y)$, 以及随机变量序列 $Y_1, Y_2, \dots, Y_n, \dots$ 的分布函数分别为 $F_{Y_n}(y) = P(Y_n \leq y)$, 如果

$$\lim_{n \rightarrow \infty} P[Y_n \leq y] = P[Y \leq y]$$

$$\text{即 } \lim_{n \rightarrow \infty} F_{Y_n}(y) = F_Y(y)$$

则称随机变量序列 $Y_1, Y_2, \dots, Y_n, \dots$ 依分布收敛于 Y , 记 $Y_n \xrightarrow{d} Y$.

林德贝格-勒维(Lindeberg-Lévy)中心极限定理

设独立同分布的随机变量 $X_1, X_2, \dots, X_n, \dots$ 的期望 $E(X_1) = \mu$ 和方差 $\text{Var}(X_1) = \sigma^2$, 则

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$$

前面介绍标准正态分布的分布函数为 $\Phi(x)$, 则上述中心极限定理等价于

$$\lim_{n \rightarrow \infty} P[Y_n \leq y] = \Phi(y)$$

变形公式: $\sum_{i=1}^n X_i \xrightarrow{d} N(n\mu, n\sigma^2)$ 等

大数定理与中心极限定理的区别

大数定律： $n \rightarrow \infty$ 时随机变量平均值 $\frac{1}{n} \sum_{i=1}^n X_i$ 的趋势

中心极限定理： $n \rightarrow \infty$ 时随机变量平均值 $\frac{1}{n} \sum_{i=1}^n X_i$ 的具体分布

例题：设一电压接收器同时接收到20个独立同分布的信号电压 V_k ($k \in [20]$)，且 $V_k \sim U(0,10)$ ，粗略估计电压和大于105的概率.

例题

某产品装箱, 每箱重量是随机的, 假设其期望是50公斤, 标准差为5公斤. 若最大载重量为5吨, 问每车最多可装多少箱能以0.997以上的概率保证不超载?

棣莫弗-拉普拉斯(De Moivre-Laplace)中心极限定理

设随机变量 $X_n \sim B(n, p)$, 则

$$Y_n = \frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1)$$

根据上述中心极限定理有

$$P[X_n \leq y] = P\left[\frac{X_n - np}{\sqrt{np(1-p)}} \leq \frac{y - np}{\sqrt{np(1-p)}}\right] \approx \Phi\left(\frac{y - np}{\sqrt{np(1-p)}}\right)$$

- 已知 n 和 $P[X_n \leq y]$, 求 y ;
- 已知 n 和 y , 求 $P[X_n \leq y]$;
- 已知 y 和 $P[X_n \leq y]$, 求 n .

例题

车间有200台独立工作的车床，每台工作的概率为0.6，工作时每台耗电1千瓦，至少供电多少千瓦才能以99.9%的概率保证正常生产.

例题

系统由100个相互独立的部件组成, 每部件损坏率为0.1, 至少85个部件正常工作系统才能运行, 求系统运行概率

依分布收敛

一次电视节目调查中调查 n 人, 其中 k 人观看了电视节目, 因此收看比例 k/n 作为电视节目收视率 p 的估计, 要以90%的概率有 $\left| \frac{k}{n} - p \right| \leq 0.05$ 成立, 需要调查多少对象?

李雅普诺夫(Lyapunov)中心极限

设独立随机变量 $X_1, X_2, \dots, X_n, \dots$ 的期望 $E[X_n] = \mu_n$ 和方差 $\text{Var}(X_n) = \sigma_n^2 > 0$. 记 $B_n^2 = \sum_{k=1}^n \sigma_k^2$, 若存在 $\delta > 0$, 当 $n \rightarrow \infty$ 时有

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n E[|X_k - \mu_k|^{2+\delta}] \rightarrow 0$$

成立, 则有

$$Y_n = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n E[X_k]}{\sqrt{\text{Var}(\sum_{k=1}^n X_k)}} \xrightarrow{d} N(0,1).$$

中心极限定理小结

- 林德贝格-勒维中心极限定理: 独立同分布随机变量, 若 $E[X_k] = \mu$ 和 $\text{Var}(X_k) = \sigma^2$, 则 $\sum_{k=1}^n X_k \xrightarrow{d} N(n\mu, n\sigma^2)$
- 棣莫弗-拉普拉斯中心极限定理: 若 $X_n \sim B(n, p)$, 则 $X_n \xrightarrow{d} N(np, np(1-p))$
- 李雅普诺夫定理: 独立不同分布中心极限定理

Ch 8 统计的基本概念



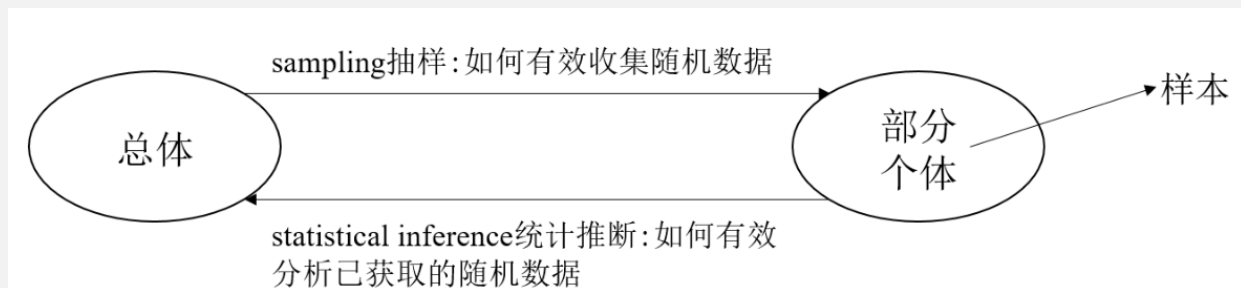
统计学

到19世纪末20世纪初，随着近代数学和概率论的发展，诞生了统计学。

统计学：以概率论为基础，研究如何有效收集研究对象的随机数据，以及如何运用所获得的数据揭示统计规律的一门学科。

统计学的研究内容包括：

- 抽样
- 参数估计
- 假设检验



总体与个体

总体：研究问题所涉及的对象全体，分为有限或无限总体，例如：全国人民的收入是总体

个体：总体中每个元素，例如：一个人的收入是个体

通常关心：总体的某项或某些数量指标

每个个体是随机试验的一个观察值, 即随机变量 X 的值
对总体的研究即对随机变量 X 的分布或数字特征的研究
后面总体与随机变量 X 的分布不再区分, 简称总体 X

基本概念

总体: 研究对象的全体, 用随机变量 X 表示(分布未知)

样本: 从总体中随机抽取一些个体, 表示为 X_1, X_2, \dots, X_n , 称 X_1, X_2, \dots, X_n 为取自总体 X 的随机样本, 其样本容量为 n

抽样: 抽取样本的过程

样本值: 观察样本得到的数值, 例如: $X_1 = x_1$, $X_2 = x_2, \dots, X_n = x_n$ 为样本观察值或样本值.

样本的二重性: i) 就一次观察而言, 样本值是确定的数;
ii) 不同的抽样下, 样本值会发生变化, 可看作随机变量.

简单随机样本

称样本 X_1, X_2, \dots, X_n 是总体 X 的简单随机样本, 简称样本, 是指样本满足:

- 1) 代表性, 即 X_i 与 X 同分布;
- 2) 独立性, 即 X_1, X_2, \dots, X_n 之间相互独立.

本书后面所考虑的样本均为简单随机样本.

样本的分布

设总体 X 的联合分布函数为 $F(x)$, 则 X_1, X_2, \dots, X_n 的联合分布函数为

$$F(x_1, x_2, \dots, x_n) = F(x_1)F(x_2) \cdots F(x_n)$$

若总体 X 的概率密度为 $f(x)$, 则样本 X_1, X_2, \dots, X_n 的联合概率密度为

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \cdots f(x_n)$$

若总体 X 的分布列 $P(X = x_i)$, 则样本 X_1, X_2, \dots, X_n 的联合分布列为

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$$