Ch 8 中心极限定理

回顾前一次课

Chernoff方法：利用Markov不等式有

$$P[X \geq \epsilon] = P[e^{tX} \geq e^{t\epsilon}] \leq e^{-t\epsilon} E[e^{tX}]$$

特别地, 有$P[X \geq \epsilon] \leq \min_{t>0}\{e^{-t\epsilon} E[e^{tX}]\}$

- $X_i \in \{0,1\}$
- Rademacher随机变量：$P(X = +1) = P(X = -1) = 1/2$
- $X_i \in [a, b]$, Chernoff引理
- $X_i \sim \mathrm{N}(\mu, \sigma^2)$

基于方差的concentration: Bennet, Bernstein

随机投影Random projection

大数定律总结

依概率收敛$\boldsymbol{X_n \xrightarrow{P} a}$: $\lim\limits_{n\to\infty} P[|X_n - a| < \epsilon] = 1 \quad \lim\limits_{n\to\infty} P[|X_n - a| > \epsilon] = 0$

大数定律： 随机变量序列$X_1, X_2, \cdots, X_n, \cdots$满足 $\sum_{i=1}^{n} \frac{X_i}{n} \xrightarrow{P} \sum_{i=1}^{n} \frac{E[X_i]}{n}$

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

Khintchine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$X_n \sim \mathrm{B}(n, p)$, 有$X_n/n \xrightarrow{P} p$

依分布收敛

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

$$\lim_{n\to\infty}P[Y_n\le y]=P[Y\le y]$$

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

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

中心极限定理的问题

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

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

的极限分布是否为服从正态分布

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$X_1, X_2, \cdots, X_n, \cdots$的期望$E(X_1) = \mu$和方差$\mathrm{Var}(X_1) = \sigma^2$, 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

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

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：$n \to \infty$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： $n \to \infty$时随机变量平均值$\frac{1}{n}\sum_{i=1}^{n} X_i$的的具体分布

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

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

- 已知n和$P[X_n \le y]$, 求y
- 已知n和y, 求$P[X_n \le y]$
- 已知y和$P[X_n \le y]$, 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$\left|\frac{k}{n}-p\right| \le 0.05$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$X_1, X_2, \ldots, X_n, \ldots$的期望$E[X_n] = \mu_n$和方差$\mathrm{Var}(X_n) = \sigma_n^2 > 0$. 记$B_n^2 = \sum_{k=1}^{n} \sigma_k^2$, 若存在$\delta > 0$, 当$n \to \infty$时有

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

成立, 则有

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

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$\mathrm{Var}(X_k)=\sigma^2$, 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 $X_n \xrightarrow{d} N\big(np, np(1-p)\big)$
- 李雅普诺夫定理: 独立不同分布中心极限定理

Ch 9 统计的基本概念

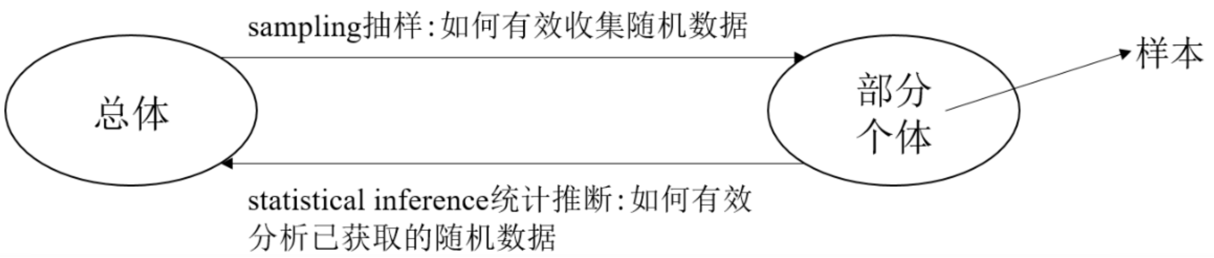
统计学

19世纪末20世纪初, 随着近代数学和概率论发展, 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$X_1, X_2, \cdots, X_n$，称$X_1, X_2, \cdots, X_n$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: $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$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

样本$X_1, X_2, \cdots, X_n$是总体X的**简单随机样本**, 简称**样本**, 指样本满足

1) 代表性, 即X_i与X同分布;

2) 独立性, 即$X_1, X_2, \cdots, X_n$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$X_1,\cdots,X_n$的联合分布函数为

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

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

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

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

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

统计量

设$X_1,X_2,\cdots,X_n$来自总体X的一个样本，$g(X_1,X_2,\cdots,X_n)$ 是关于$X_1,X_2,\cdots,X_n$的一个连续、且不含任意参数的函数，称$g(X_1,X_2,\cdots,X_n)$是一个**统计量**

- $g(X_1,X_2,\cdots,X_n)$是随机变量: $X_1,X_2,\cdots,X_n$是随机变量
- $g(x_1,x_2,\cdots,x_n)$ 为$g(X_1,X_2,\cdots,X_n)$的一次观察值

样本均值

设$X_1, X_2, \cdots, X_n$是总体X的一个样本, 定义**样本均值**为

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

引理：总体X的期望为$E[X] = \mu$, 方差$\mathrm{Var}(X) = \sigma^2$, 有

$$E[\bar{X}] = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

样本方差

设$X_1, X_2, \cdots, X_n$是总体X的一个样本, 定义**样本方差**为

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}{X_i}^2 - \bar{X}^2$$

定义**样本标准差**为 $S_0 = \sqrt{S_0^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

引理： 总体X的期望为$E[X] = \mu$, 方差$\mathrm{Var}(X) = \sigma^2$, 有

$$E[S_0^2] = \frac{n-1}{n}\sigma^2$$

修正后的样本方差

设 $X_1, X_2, \cdots, X_n$ 是总体 X 的一个样本，定义**修正后的样本方差**为

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

引理： 总体X的期望为$E[X] = \mu$, 方差$\mathrm{Var}(X) = \sigma^2$, 有

$$E[S^2] = \sigma^2$$

原点矩和中心矩

假设$X_1, X_2, \cdots, X_n$是来自总体X的一个样本, 样本***k*****阶原点矩**为:

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad k = 1,2,\cdots.$$

样本***k*****阶中心矩**为:

$$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k \qquad k = 1,2,\cdots$$

例题

设总体$X \sim N(20,3)$，从总体中抽取两独立样本，容量分别为10和15. 求这两个样本均值之差的绝对值大于0.3的概率