概率统计B

Probability and Statistics

张思容
zhangsirong@buaa.edu.cn

数学与系统科学学院，北京航空航天大学
School of Mathematics and System Sciences, BUAA

December 1, 2014

1. 不等式与大数定律
   - 矩不等式
   - 弱大数定律
2. 中心极限定理
   - 强大数定律***
   - 中心极限定理
3. 统计思想
   - 抽样调查
   - 统计量与估计
   - 统计学三大分布

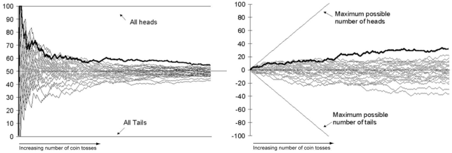
Review：回顾

Last Chapter:

- 多元随机向量
- 协方差
- 复合函数

This Chapter:

- 不等式和大数定律
- 中心极限定理
- 统计思想与三大分布



大量实验的投硬币结果逼近期望(均值)! (但是误差可能越来越大.) 大数定律不是平均律!

概率不等式

利用随机变量的特征估计概率:

Proposition (Markov马尔可夫不等式)

设随机变量X取非负值，则任意$a > 0$,
$P(X > a) \leq \frac{EX}{a}$.

证明：定义随机变量$I = a, X \geq a; I = 0, X < a$,
易得$I \leq X$; 利用期望的保号性，$EI \leq EX$,可得。

Corollary (Chebyshev切比雪夫不等式)

设随机变量X的均值为μ,方差为σ^2，则任意$a > 0$, 有
$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$

证明：令$Y = (X - \mu)^2$,利用马尔可夫不等式， $P(Y > a) \leq EY/a$,
令$a = k^2$, 有$P((X - \mu)^2 \geq a) = P(|X - \mu| \geq k)$, $E(X - \mu)^2 = \sigma^2$,命题得证。

应用不等式估计概率

EXAMPLE (估计销售量)

假设一个商店平均每天卖*50*件商品，

1. 问某一天卖掉*75*件以上的大概概率是多少？
2. 如果已知销售量的方差是*25*，问销售量在40到60之间的概率是多少?

解答 设X是销售量，$EX=50$,

1. 由马尔可夫不等式 $P(X>75)\leq 50/75=2/3$;
2. 如果$VarX=\sigma^2=25$,则$P(|X-50|<10)=1-P(|X-50|\geq 10)=1-25/100=3/4$.

以上不等式仅仅给出上界，不是很精确。(一般用于理论研究).
如果已知分布是正态分布，可以得到2的答案为0.955。

弱大数定律

- 从一个随机变量X到一个随机变量系列X_n:
- 假设: X_i都是相同的概率分布且互相独立(称为IID),期望为μ,方差为σ^2. 记均值随机变量$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{i=n} X_i$.

Theorem (弱大数定律)

给定独立同分布的随机变量列X_n,有有限期望μ(和有限方差σ^2), 则有 对任意$\epsilon > 0$,有当$n \to \infty$时 $P(|\bar{X}_n - \mu| \geq \epsilon) \to 0$.

证明: 易得 $E\bar{X}_n = \mu$, $Var(\bar{X}_n) = \sigma^2/n$;
由切比雪夫不等式 $P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0$.

Definition (依概率收敛)

设X_n, X是随机变量,如果对任意$\epsilon > 0$, 有$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$, 称$X_n$依概率收敛到$X$. 记为$X_n \xrightarrow{p} X$.

弱大数定律即 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{i=n} X_i \xrightarrow{p} \mu$.

实例: 社会调查(或选举)

EXAMPLE (早餐)

设我们希望知道大学生吃早餐的比例p。通过在图书馆门口调查n个人, 如果m个人吃早餐。 我们就估计$p = m/n$.

1. 给定误差ϵ,估计错误的概率是多少?
2. 如果希望误差不超过0.01, 错误的概率不超过0.05, 即5%,需要调查多少人?

解答:.

记X_i为两点分布$B(1,p)$,则均值$\bar{X}_n$的期望p,方差为$p(1-p)/n$.(分布?)
1)利用切比雪夫不等式, 设误差为$\epsilon = 0.1$,
则$P(|\bar{X}_n - p| \geq 0.1) < p(1-p)/(n * 0.01) < 1/4(n * 0.01)$.
如果$n = 100$,误差大于0.1的概率小于0.25
2)同上, $P(|\bar{X}_n - p| \geq 0.01) < 1/(4n * 0.01^2) < 0.05$, 保守估计要求$n > 50000$. □

常数随机变量***

EXAMPLE (常数随机变量)

设$X = a$是常数，则$EX = \mu, Var = 0$.
如果已知$EX = \mu, Var = 0$,X是什么?
仅仅有$P(X = \mu) = 1$.

证明: 任意k, 有$P(|X - \mu| > k) \leq 0/k = 0$;
即$P(X \neq \mu) = 0$。

Proposition (推论)

特殊特征与概率分布.

- $E(|X|) = 0 \to P(X = 0) = 1$(马尔可夫不等式)
- $DX = 0 \to P(X = EX) = 1$
- $\rho(X, Y) = \pm 1 \to P(Y = aX + b) = 1$.
  (构造$Z = X/\sigma_x \pm Y/\sigma_y$,利用$DZ = 0$结论。)

随机变量的极限***

数学分析中的极限

1. 数列极限$a_n \to a$
2. 函数列的极限:$f_n(x) \to f(x)$.
   点点收敛, 一致收敛;
3. 级数的收敛:

随机变量的极限(参见实变函数,泛函分析***)

1. 依概率收敛: 记为$X_n \xrightarrow{p} X$, 如果$P(|X_n - X| \geq \epsilon) \to 0$
2. 以概率1收敛(几乎处处): $X_n \xrightarrow{as} X$,或者 $X_n \to X, wp1.$,如果$P(\lim X_n = X) = 1$;
3. 依分布收敛: 记为$X_n \xrightarrow{d} X$,如果$\lim P(X_n \leq x) = F_X(x), \lim F_{X_n}(x) = F_X(x)$;

强大数定律***

Definition (以概率1收敛)

设X_n, X是随机变量,如果$P(\lim X_n = X) = 1$ 称X_n以概率*1*收敛*(*几乎处处*)*到X, 记为$X_n \xrightarrow{as} X$,或者 $X_n \to X, wp1.$

Theorem (强大数定律)

给定独立同分布的随机变量列X_n,有有限期望μ,*(*有限四阶矩$EX^4 < \infty$*)*,则 $\bar{X}_n \xrightarrow{as} \mu$, 即$P(\lim \bar{X}_n = \mu) = 1.$

证明提示: 设期望为0,构造级数$E(\sum \bar{X}_n{}^4) < \infty$, 得到$\lim \bar{X}_n{}^4 = 0$,
即$\bar{X}_n \xrightarrow{as} 0$.
注记:强大数定律说明了频率学派的概率解释。 给定两点分布$B(1, p)$, 频率$\bar{X}_n \xrightarrow{as} p$。
(比课本贝努利定理强)。

中心极限定理

Theorem (中心极限定理)

给定独立同分布的随机变量列X_n,有有限期望μ和有限方差σ^2, 定义部分和随机变量$S_n = \sum_{i=1}^{i=n} X_i$,其标准化为$\xi_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$, 则 $\xi_n \xrightarrow{d} N(0,1)$,
即$\lim P(\xi_n \leq x) = \Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt$.其中$\Phi(x)$为标准正态分布的分布函数。

证明提示: 利用矩母函数可以证明。
说明: 中心极限定理有很多推广, 独立最重要;
均值, 方差是最重要的分布特征, 利用统计估计均值和方差即可。

Corollary (分布逼近)

对充分大n,$S_n \sim N(n\mu, n\sigma^2)$,逼近公式: $P(S_n \leq x) \approx F_{N(n\mu,n\sigma^2)}(x)$。
特别二项分布$B(n,p)$可用正态分布$N(np, np(1-p))$逼近。*(DeMoivre-Laplace定理)*.

(续) 实例: 社会调查(或选举)

EXAMPLE (早餐)

设我们希望知道大学生吃早餐的比例p。通过在图书馆门口调查n个人,如果m个人吃早餐。 我们就估计$p = m/n$.

1. 给定误差ϵ,估计错误的概率是多少?
2. 如果希望误差不超过0.01, 错误的概率不超过0.05, 即5%,需要调查多少人?

解答: 记X_i为两点分布$B(1,p)$,则利用中心极限定理的推论:
$\bar{X}_n \sim N(p, p(1-p)/n)$, 可近似计算$\bar{X}_n \approx N(p, 1/(4n))$.
1)$P(|\bar{X}_n - p| \geq \epsilon) \sim 2P(\bar{X}_n - p \geq \epsilon)$. 令$Z = (\bar{X}_n - p)/\sqrt{1/(4n)}$. 如果$n = 100$,$\epsilon = 0.1$,则概率约$2(1-\Phi(Z)) = 2 - 2\Phi(2) = 0.046$;
2)同上, $P(|\bar{X}_n - p| \geq \epsilon) = 2 - 2\Phi(2\epsilon\sqrt{n}) \leq 0.05$, 注意:
$\Phi(1.96) = 0.975$!!!, 计算有$n > (1.96)^2/(4 * 0.01^2) = 9604$.
一般的报刊上社会抽样调查误差约为0.03, $n > 1067$人即可。

作业

北航教材:
P125 习题六. 1， 2， 3， 4， 5.

从概率论到统计

概率论的理论:

- 概率模型(样本空间+概率律);
- 随机变量(分布律)X ,数字特征(矩)EX^i;
- 大数定律: 大量分布的平均值是稳定的;
- 中心极限定理：很多随机变量的分布是Bell 钟型曲线;
  Galton：混乱世界中的最高规律:

经典统计学的推断过程:

- 得到概率模型(随机变量)的大量实验数据x_i(样本)
- 计算数据(样本)的特征(矩)或其他统计量;
- 参数估计: (大数定律)可以用数据的统计量来估计未知模型的参数;得到概率模型;
- 统计决策: 给出估计的误差或假设检验；利用统计量的分布，一般假定X的分布是正态分布(中心极限定理);

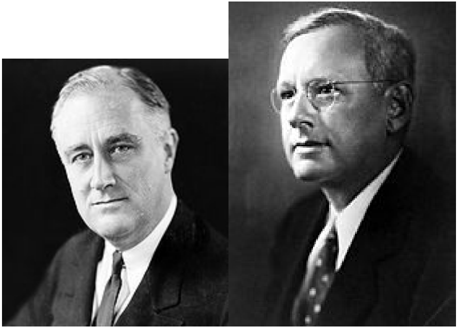
Review： 回顾

Last Week:

- (矩)不等式: 特殊特征与随机变量
- 大数定律： 数的逼近
- 中心极限定理: 和分布的正态逼近或误差的逼近

This Week:

- 统计与抽样
- 统计学三大分布



美国1936年总统选举预测：
文学文摘(1千万问卷,预测Landon赢)
盖洛普George Gallup(5000问卷,预测Roosevelt 赢)

什么是统计?

A Joke:

- 一种疾病平均每一千人死亡3.2人。
- 健康大臣问秘书3.2人是如何死法?
- 秘书说：一个统计学家说死了3.2人时，意味着3个人已经死了，两个人正要死。

统计学：大不列颠百科全书定义：
Statistics is the art and science of collecting and analyzing data.

- 美国统计局："美国是运行在数据上的国家。"
- H.G.Wells：统计思维总有一天会向读与写一样 成为一个有效率公民的必备能力。
- C.R.Rao: 不确定知识+不确定性度量的知识=可用的知识

总体与样本 population vs sample

- 总体(N)：统计研究对象。比如人口普查。
  具体指标: 身高，寿命，教育:
- 总体的分布: 或一个指标的分布: 是一个确定的分布!!! 但无法得到!
  数学: (抽象化)一个"典型"个体的(指标)是一个随机变量,设为X。
- 抽样调查: 随机选取n个个体的过程。每一次选取是一个抽样(sampling); n是样本容量。
- 样本: "随机" 得到的n个个体是随机向量$(X_1, X_2, \ldots, X_n)$
  具体指标：称为样本值，记为$x_1, x_2, \ldots, x_n$.
- 经典统计: 由样本值推断总体分布X的参数 (即参数估计) 或分布函数(非参数估计).

抽样过程

常见抽样方法:

- 简单随机抽样: 每个个体有等概率被抽取;
- 分层抽样: 先根据总体分类,再简单抽样分析。比如男女。主要提高结果的代表性。
- 分阶抽样: 先抽样一个子总体, 再从子总体中抽样,... 比如北航学生调查:
  先抽取若干学院, 再抽取若干班级。 主要为了提高抽样效率(省钱).

数学描述:总体分布X

- ***简单抽样过程: 对每个个体$S_i(\omega) = 1$ 或 0, 选中概率 $\sim B(1, n/N)$.
- 假定: 得到$X_1, X_2, \ldots, X_n$是独立同分布(IID)与X.
  设$F(x)$为X的分布函数, 则联合分布$F(x_1, x_2, \ldots, x_n) = \prod F(x_i)$.
  怎样得到F的信息?

从数据到统计量

简单数据统计

- 图表：直方图，圆饼图，
- 样本统计量: 均值，中位数，分位数(1/4),范围(箱线图)
- 样本的矩: 样本均值，样本方差，样本矩$\frac{1}{n}\sum x_i^k$. 样本中心矩;
- 次序统计量: 样本排序$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$,
  经验分布:$F_n(x) = k/n, x_{(k)} \leq x < x_{(k+1)}$.

Definition (统计量)

$X_1, X_2, \ldots, X_n$是独立同分布*(IID)*与X，g为任一n元函数，
称$g(X_1, X_2, \ldots, X_n)$为 一个统计量*(*是一个随机变量*)*。

大数定律: 样本矩逼近总体矩；统计量的取值逼近某个参数;
中心极限定理:统计量的分布服从与正态分布相关的一些分布；

三大统计(抽样)分布

假定总体是正态分布!!!

1. $\chi^2(n)$分布(卡方分布)：自由度为n. $\chi^2 = \sum_i X_i^2, X_i \sim N(0,1)$.
   $\to \chi^2$分析
   性质：$E\chi^2 = n, D\chi^2 = 2n$.
   $\chi^2(n) + \chi^2(m) = \chi^2(n+m)$
2. $t(n)$ 分布(学生分布)：$T = X/\sqrt{Y/n}, X \sim N(0,1), Y \sim \chi^2(n)$。
   性质: 对称，小样本分布$n \le 30$，$n > 30$与正态分布接近；
3. $F(n,m)$分布(纪念Fisher)：
   $F(m,n) = \frac{X/m}{Y/n}, X \sim \chi^2(m), Y \sim \chi^2(n)$，且$X, Y$相互独立。
   $\to$ ANOVA 方差分析
   性质: $F_{1-\alpha}(m,n) = \frac{1}{F_\alpha(n,m)}$.

分布查表: 下α分位数$z_\alpha, \chi^2_\alpha(n), t_\alpha(n), F_\alpha(m,n)$ 满足$P(S \le S_\alpha) = \alpha$.
分布计算: 常见分布计算器。

常见统计量的分布 IMPORTANT!

- 样本常见特征的分布是统计推断的关键!
  一般通过样本计算(构造)出一些统计量: 样本均值, 样本方差, 样本相关系数, 希望统计量的分布是已知的!
- 中心极限定理给出样本均值的逼近分布正态分布, 其他特征呢?
- 基本假设:如果总体是正态分布, $X \sim N(\mu, \sigma^2)$,
  记$\bar{X}_n = \frac{1}{n}\sum X_i, S^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2$. 特别$\bar{X}_n$和$S^2$独立。
- 样本均值与正态分布 $(\bar{X}_n - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$
- 样本方差与χ^2分布: $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$
- (未知总体方差) 样本均值与t分布: $\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t(n-1)$
- *(两个总体的均值差) 假设方差相同未知, P135定理五.
- *(两个总体的方差比) 两个独立总体X, Y,
  $\frac{S_X^2/\sigma_x^2}{S_Y^2/\sigma_Y^2} \sim F(n-1, m-1), n, m \geq 2$.

作业

北航教材:

P138 习题七. 1, 2, 3, 4, 5, 6

QUIZ 小测验 三

1. 设随机变量$X \sim N(\mu, \sigma^2)$,则$E(X-\mu)^4 =$. ()
(A) σ^4; (B) $2\sigma^4$; (C) $6\sigma^4$; (D) $3\sigma^4$ 。
2. 设随机变量X存在数学期望EX 和方差DX ,则对任意正数ϵ有(),
（A） $P(|X-EX| \geq \epsilon) > \frac{DX}{\epsilon^2}$, （B） $P(|X-EX| < \epsilon) > 1-\frac{DX}{\epsilon^2}$;
（C） $P(|X-EX| \geq \epsilon\sqrt{DX}) \leq \frac{1}{\epsilon^2}$;
（D） $P(|X| \geq \epsilon) \leq \frac{E(X-EX)^k}{\epsilon^k}(k \geq 1)$,
3. 设二维随机变量的概率密度
为$f(x,y) = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}$,则$z = \sqrt{x^2+y^2}$的概率密度是:
4. 设X_n是相互独立的随机变量序列,且其分布律为
$P(X_n = -\sqrt{n}) = \frac{1}{2^{n+1}}, P(X_n = \sqrt{n}) = \frac{1}{2^{n+1}}, P(X_n = 0) = 1-\frac{1}{2^n}$. 记
$Y_n = \frac{1}{n}\sum_i X_i$。
试求: （1） EX_n, DX_n; （2） EY_n, DY_n;
（3） 证明: 对任给$\epsilon > 0$,成立$\lim_{n\to\infty} P(|Y_n| \geq \epsilon) = 0$。

QUIZ 小测验 三：答案

1. 设随机变量$X \sim N(\mu, \sigma^2)$,则$E(X-\mu)^4 =$. (D)
   (A) σ^4； (B) $2\sigma^4$; (C) $6\sigma^4$; (D) $3\sigma^4$ 。
2. 设随机变量X存在数学期望EX 和方差DX ,则对任意正数ϵ有((c)),
   （A） $P(|X-EX| \geq \epsilon) > \frac{DX}{\epsilon^2}$, （B） $P(|X-EX| < \epsilon) > 1-\frac{DX}{\epsilon^2}$;
   （C） $P(|X-EX| \geq \epsilon\sqrt{DX}) \leq \frac{1}{\epsilon^2}$;
   （D） $P(|X| \geq \epsilon) \leq \frac{E(X-EX)^k}{\epsilon^k}(k \geq 1)$,
3. 设二维随机变量的概率密度
   为$f(x,y) = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}$,则$z = \sqrt{x^2+y^2}$的概率密度是：
   $(f(z) = \frac{z}{\sigma^2}e^{-\frac{z^2}{2\sigma^2}}, z > 0)$
4. 设X_n是相互独立的随机变量序列,且其分布律为
   $P(X_n = -\sqrt{n}) = \frac{1}{2^{n+1}}, P(X_n = \sqrt{n}) = \frac{1}{2^{n+1}}$, $P(X_n = 0) = 1-\frac{1}{2^n}$. 记
   $Y_n = \frac{1}{n}\sum_i X_i$。
   试求： （1） $EX_n, DX_n; (0, n/2^n)$ （2） $EY_n, DY_n; (0, \frac{1}{n^2}\sum_i i/2^i)$
   （3）证明：对任给$\epsilon > 0$,成立$\lim_{n\to\infty} P(|Y_n| \geq \epsilon) = 0$。