概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

September 17, 2010

Week 1: 概率模型与公理化

相互介绍

教师：

- 张思容: Ph.D. 几何分析，医学图像分析；
- 办公时间：周五(2pm-4pm)或预约。图书馆西配楼501。
- 概率统计课程答疑：主216 每周星期三(6pm-8pm)
- 联系方式：134-3920-1025. zhangsirong@buaa.edu.cn
- 课程网站：sirongzhang.wordpress.com/2010fallPSS/
- 欢迎大家学期中提建议和问题，不要考试后!

学生介绍：3821,22,3831,32
探测制导与控制技术,飞行器动力工程(航天)

课程介绍: 参考书目

预备要求：微积分.
参考书目：

- (教材) 概率统计及随机过程。张福渊等，北家航空航天大学出版社。ISBN: 7810770047
- (推荐) 概率导论. (MIT教材) 人民邮电出版社ISBN 9787115215444。
- (推荐）应用随机过程概率模型导论。Sheldon M.Ross，龚光鲁译，人民邮电出版社。ISBN：9787115167330
- (考研) 概率论与数理统计。盛骤等(浙江大学教材). 高等教育出版社。ISBN 9787040238969
- (习题）北航概率统计习题集；浙江大学概率统计习题解答；

课程介绍：内容学习

概率统计：最有用的数学课程。

- 例子: 赌博, 股票，市场调查。
- 内容: 随机现象建模+统计决策。
- 工具：理论(数学分析...), 应用：软件SAS,...。

课程学习

- 目标：学会概率建模的思想和方法；掌握应用及技巧(解题).
- 作业：每周交一次。TBA
- 成绩： 平时成绩10+期末考试90=100
  小测验若干(quizs).

问题？Q&A(Questions and Answers)

样本空间

- 样本空间: 表示一个试验的所有可能结果的集合。记为S或Ω。
  确定现象：数→ 函数
  随机现象：集合→ 随机过程
- 样本空间的例子：
  投硬币：(Head, Tail)
  掷骰子：(1,2,3,4,5,6)
  射击打靶：？
- 事件：样本空间中关心的结果(子集合). 用A, B等表示。
  $A = \{$ 骰子是偶数$\}$
- 事件与子集；集合运算。空集即不可能事件；全集是必然事件；
  注意：不是所有子集都是事件!

概率律(law)

概率律P:给每个事件指定一个概率大小。

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. $A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$

EXAMPLE

投硬币：$(Head, \ Tail)$ $A = \{Head\}, B = \{Tail\}$, 指定$P(A) = P(B) = \frac{1}{2}$.
类似对掷骰子可以指定$P(\{1\}) = \frac{1}{6}\ldots$
(有限空间)古典概率模型：$P(A) = \frac{\#A}{\#\Omega}$

问题： 空间无穷怎么办？古怪子集怎么办？

概率与真实世界

什么是概率？《概率沉思录》E.T.Jaynes

- 赌博与彩票等：概率是等可能的→ 古典概率模型Laplace
- 红楼梦是否是曹雪芹写的？明天会下雨吗？概率是主观判断(经验) → 贝叶斯学派Bayes
- 人口出生率：男孩vs 女孩105 : 100 概率是数据发生的频率→ 频率学派Fisher

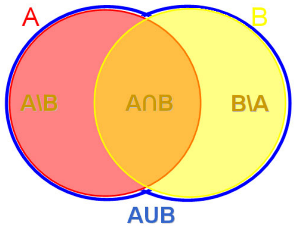
概率模型：在不同的应用中选择不同的模型。"抓住老鼠就是好猫!"

思考题

将52张扑克扣在桌上一张张翻开；一直到出现一张"A"为止。再翻一张牌，问下一张牌是黑桃A的概率和方块2的概率谁大？
一样大！$p = \frac{1}{52}$.

事件与集合

- 集合Ω: 元素$\omega \in \Omega$
  记子集$A = \{\omega | \omega \in A\}$.
- 集合运算: $\bar{A} = \Omega - A$, 乘法$A \cap B = AB$,"加法" $A \cup B$, 减法$A - B = A\bar{B}$
  对应的事件意义？
- 不相容事件: $A \cap B = \emptyset$.
  对立事件：$\bar{A}$
- 集合运算的规律：交换律，结合律；
  Venn韦恩图
- 有用公式：
  $A \cup B = A + \bar{A}B, A = AB + A\bar{B}$
  De-Morgan公式：
  $\overline{\bigcup_{j=1}^{j=n} A_i} = \bigcap_{j=1}^{j=n} \bar{A}_i$.



配对问题

EXAMPLE (配对问题)

任意n个同学交了n本作业，随机每人发回一本作业，试求有$k = 0, 1, 2, ..., n$个同学得到自己作业的概率？简单情形：$n = 2$, $n = 3$,

解答.

设$n = 3$, 记三个人A, B, C,作业a, b, c, Ab表示A拿到b的作业。
样本空间：$\Omega = \{AaBbCc, AaBcCb, AbBaCc, AbBcCa, AcBbCa, AcBaCb\}$
事件: $X = k$个同学配对成功；$k = 0$: $X = \{AbBcCa, AcBaCb\}$
$k = 1$: $X = \{AaBcCb, AbBaCc, AcBbCa\}$
$k = 2$: $X = \emptyset$; $k = 3$: $X = \{AaBbCc\}$
古典概率计算有$P(\{X = 0\}) = \frac{1}{3}, P(\{X = 1\}) = \frac{1}{2}$,
$P(\{X = 2\}) = 0, P(\{X = 3\}) = \frac{1}{6}$, □

n变大时，情况迅速变复杂！见后续讲解。

计数方法counting

Theorem (乘法原理)

完成一件事要k步，每一步分别有$n_1, n_2, n_3, \ldots, n_k$方法，则完成这件事的方法共有$n = n_1 \cdot n_2 \cdots n_k$.

基本结果：

- (m次试验): 从n个元素中有放回的每次取一个；取出m个元素,排成一列；共有n^m种可能；不同排列是等可能的；
- (m元排列) 从n个元素中无放回的每次取一个；取出m个元素,排成一列；共有$A_n^m = \frac{n!}{(n-m)!}$种可能；不同排列是等可能的；
- (m元组合) 从n个元素中无放回的每次取一个；取出m个元素,放在一组；共有$C_n^m = \frac{n!}{m!(n-m)!}$种可能；不同组合是等可能的；

例子：(集合子集的个数) 集合有n个元素，则所有子集的个数为2^n。利用：组合数即二项式系数, 由二项式展开$(x + y)^n = \sum C_i^n x^i y^{n-i}$可得。

例子

EXAMPLE (生日问题)

任意n个人中有[至少两个人]有相同生日的概率。

解答.

假设每个人等可能出生于365天中任一天。
n个人生日的样本空间Ω大小：365^n
$\bar{A}$ =[没有两个人生日相同], $\bar{A}$的集合大小为$n!C_{365}^n$,
$P(A) = 1 - P(\bar{A}) = 1 - \frac{n!C_{365}^n}{365^n}$

n	20	30	40	50	60	70	80
p_n	0.411	0.706	0.891	0.970	0.994	0.999	0.9999

□

作业

北航教材:
P31 习题一.
1(2),(3),(5);
4
5
6

无穷样本空间: 可数

- 古典模型: 集合有n个元素, 每个元素可看成一个子集,称为基本事件(样本点);
- 可数模型(离散模型)：集合有可数个元素。每个元素可看成一个子集,可看成事件;

EXAMPLE (射击命中)

设一个人打靶命中率为p, 重复射击直到打中靶停止。我们关心的要射击多少次才能打中？

解答.

样本空间：$\Omega = \{1, 2, 3, \ldots, n, \ldots, \infty\}$
概率律：$P(x = 1) = p, P(X = 2) = p(1-p), \ldots, P(X = n) = p(1-p)^{n-1}$.
验证$P(X = 1) + P(X = 2) + \cdots = 1$ □

一般的：可数模型的概率律可对应一个收敛的正项级数(和为1)。

无穷样本空间: 连续空间(不可数)

连续模型：集合有不可数个元素。每个元素可看成一个子集,但一点作为事件无意义! 因为一点的概率必须为零!
古典概率推广：几何概率定义$P(A) = \frac{L(A)}{L(\Omega)}$,其中$L(A)$表示其区域的面积或长度或体积。

EXAMPLE (约会问题)

两个人约定在8点钟见面，都可能迟到至多一个小时。如果先到的等候15分钟，问两个人能见面的概率。

解答.

设到达时间分别为x, y, 样本空间：$\Omega = \{(x, y) | 8 \le x \le 9, 8 \le y \le 9\}$
事件：两个人见面$A = \{(x, y) || x - y| \le 15/60\}$
概率律：$P(A) = \frac{L(A)}{L(\Omega)}$, 计算有$P(A) = 1 - P(\bar{A}) = 1 - \frac{3}{4} \cdot \frac{3}{4} = \frac{7}{16}$ □

一般的：连续模型的概率律对应一个可积函数。

概率律的公理化

Definition (概率: 1930 Kolmogorov)

P是定义在样本空间Ω上所有事件$\mathfrak{F}$的一个函数；$P:\mathfrak{F}\to[0,1]$, 满足

1. 非负性：$P(A)\geq 0$;
2. 规范性：$P(\Omega)=1$;
3. 可数(列)可加性：如果A_i是互不相容事件，则$P(\bigcup_{i=1}^{\infty}A_i)=\sum_{i=1}^{\infty}P(A_i)$;

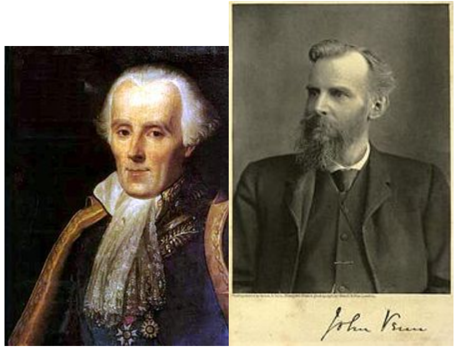
称P为Ω的一个概率(律)，记$P(A)$为事件A的概率。

基本属性

- $P(\emptyset)=0$；有限可加性；
- 单调性：$A\subset B$，则$P(A)\leq P(B)$;
- ***连续性：有单调增序列A_i,单调减序列B_i, 则 $P(\bigcup_{n=1}^{\infty}A_n)=\lim_{n\to\infty}P(A_n)$,$P(\bigcap_{n=1}^{\infty}B_n)=\lim_{n\to\infty}P(B_n)$
- ***数学上称概率为测度；见实变函数论。

从古典到频率学派

P.Laplace VS J.Venn

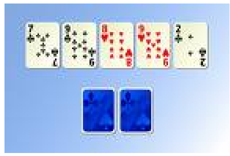


概率的真实大小？

- 投硬币：coin flipping 正面的概率？
- 拉普拉斯：1/2
- De Morgan 德摩根：1061/2048
- 蒲丰Buffon：2048/4040
- John Kerrich: 5067/10000
- K Pearson: 12012/24000, Feller: 4979/10000;

频率学派：概率由大量数据的频率决定。

Texas Hold'em



德州扑克Texas Hold'em,是世界上最流行的扑克游戏。每个玩家最后用五张扑克牌比大小。得克萨斯扑克的大小规则如下：

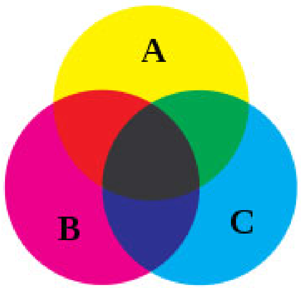
1. 皇家同花顺Royal straight flush (比如黑桃10,J,Q,K,A)
2. 同花顺Straight flush (比如黑桃3,4,5,6,7)
3. 四条,炸弹Four of a kind (比如四条9和一个其它任何牌)
4. 满堂彩Full house (三条加一对) 5.清一色Flush (比如梅花2,5,6,8,J)
5. 一条龙Straight (比如4,5,6,7,8不同花色混杂)
6. 三条Three of a Kind (比如三个8和其它任意两张单牌)
7. 两对Two Pair
8. 一对Pair

思考：这个规则合理吗？各种情形的概率是多少？

加法公式

Proposition (加法公式)

1. $P(A\cup B)=P(A)+P(B)-P(A\cap B)$;
2. $P(A_1\cup A_2\cup A_3)=\sum P(A_i)-\sum P(A_iA_j)+P(A_1A_2A_3)$;
3. *Jordan* 约旦公式：$P(\bigcup_{i=1}^{n}A_i)=\sum_{k=1}^{n}(-1)^{k-1}p_k$; 其中$p_k=\sum P(A_{i_1}A_{i_2}\dots A_{i_k})$.



证明(1).

$P(A\cup B)=P(A+\bar{A}\cap B)=P(A)+P(\bar{A}B)$
$=P(A)+P(\bar{A}B)+P(AB)-P(AB)=$
$P(A)+P(B)-P(AB)$ □

再谈：配对问题

EXAMPLE (配对问题)

任意n个同学交了n本作业，随机每人发回一本作业，问没有一个同学得到自己作业的概率？

解答.

(对立事件$\bar{A}$)计算至少有一个人得到自己的作业的概率.设E_i是第i个人得到自己作业的事件；则$\bar{A}=\bigcup_{i=1}^{i=n}E_i$.
Jordan公式：$P(\bar{A})=P(\bigcup_{i=1}^{i=n}E_i)=\sum_{k=1}^{n}(-1)^{k-1}p_k$,
$p_k=\sum P(E_{i_1}E_{i_2}\ldots E_{i_k})$.
记结果为$(1,2,3,\ldots,n)$,$(E_{i_1}E_{i_2}\ldots E_{i_k})$为其中$k$个人拿到自己作业；剩下$n-k$个任意排列；样本空间是$n$个任意排列，断定:$P(E_{i_1}E_{i_2}\ldots E_{i_k})=\frac{(n-k)!}{n!}$
另外: n中选k个人共有C_n^k可能; Jordan公式中每一项求和有$p_k=\frac{1}{k!}$
$P(\bar{A})=1-\frac{1}{2!}+\frac{1}{3!}-\cdots+(-1)^{n-1}\frac{1}{n!}$, $P(A)=1-P(\bar{A})$.
特别：n充分大时，$P(A)\approx e^{-1}\approx 0.3678$. □

产品检验

EXAMPLE (超几何分布)

假设N个产品中有M个次品,抽取n件产品检验，其中恰有m件次品的概率。

Proof.

A_m是恰有m件产品的事件$(m\le n)$.
$P(A_m)=\frac{C_M^m C_{N-M}^{n-m}}{C_N^n}$ □

排列组合续：分组问题

- (m次试验): 从n个元素中有放回的每次取一个；取出m个元素,排成一列；共有n^m种可能；不同排列是等可能的；
- (m元排列) 从n个元素中无放回的每次取一个；取出m个元素,排成一列；共有$A_n^m=\frac{n!}{(n-m)!}$种可能；不同排列是等可能的；
- (m元组合) 从n个元素中无放回的每次取一个；取出m个元素,放在一组；共有$C_n^m=\frac{n!}{m!(n-m)!}$种可能；不同组合是等可能的；
- (k重分组) 将n个元素分成不同的k组，不考虑每组中的元素次序；第i个组恰有n_i个元素的可能分组为$\frac{n!}{n_1!n_2!\ldots n_k!}$种可能；不同分组是等可能的；

例子：导师有4个研究生12个本科生，分配到4个项目中，随机分配，问每个项目正好有一个研究生的概率？

Proof.

样本空间大小：$\frac{16!}{4!4!4!4!}$.
事件大小: $\frac{12!}{3!3!3!3!}\cdot 4!$. 概率$P(A)=\frac{64}{455}$ □

作业

北航教材:
P31 习题一. 8,11,16,18

概率统计与随机过程
Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

October 8, 2010

Week 3,4：条件概率与独立性

Review：回顾

RECALL:

- 概率模型=样本空间+概率律
- 公理化：有限到无穷(数学分析工具!)
- 计数，加法公式，例子。

TODAY

- 计算公式：乘法，全概率，贝叶斯公式；
- 条件概率→ 统计推断。

EXAMPLE (赌牌游戏)

有三张牌：一张两面都是红的，一张两面都是黑的，一张两面是一红一黑。随机取出一张，如果正面是红的，反面是黑还是红？
怎样赢钱？

赢钱策略：永远赌反面和正面一样!

条件概率的定义

EXAMPLE (概率的变化)

掷骰子一次，结果为6的概率为$\frac{1}{6}$. 如果已知结果为偶数，结果为6的概率为$\frac{1}{3}$. 如果已知结果比4小呢？

解释：概率空间发生了变化。

Definition (条件概率)

设A, B是事件，已知A发生，B发生的概率记为$P(B|A)$. 其公式为$P(B|A) = \frac{P(AB)}{P(A)}$, 其中$P(A) > 0$。

- 古典模型解释: $P(A\cap B) = \frac{\sharp(A\cap B)}{\sharp\Omega} = \frac{\sharp(A\cap B)}{\sharp A}\cdot\frac{\sharp(A)}{\sharp\Omega} = P(A)P(B|A)$
- 公理化模型→ 定理：记$P_A(B) = P(B|A)$,P_A是一个概率(满足概率公理);

条件概率是一个概率律! ***

Theorem (条件概率)

条件概率满足概率公理化的三个条件。

- 非负：$P_A(B) \geq 0$
- 归一化：$P_A(\Omega) = 1$
- (可列) 可加性$P_A(B \cup C) = P_A(B) + P_A(C)$.

Corollary (条件概率下的条件概率)

设有事件A, B, C, 已知A发生，有条件概率P_A, 又已知B发生，有条件概率P_{AB}, 则$P_A(C|B) = P_{AB}(C)$.

乘法公式

Proposition (乘法公式)

$P(A \cap B) = P(A)P(B|A)$,可以推广到n个事件。$P(A_1A_2\dots A_n) = P(A_1)P(A_2|A_1)\dots P(A_n|A_1A_2\dots A_{n-1})$

EXAMPLE (扑克)

一副牌(52张)连续抽三张都不是红桃的概率？

Proof.

设A_1, A_2, A_3表示第i张牌不是红桃。
$P(A_1) = \frac{39}{52}$, $P(A_2) = ?$
$P(A_2|A_1) = \frac{38}{51}$, $P(A_3|A_1A_2) = \frac{37}{50}$
则$P(A_1A_2A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)$. □

全概率公式

Proposition (全概率公式)

$P(B) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A})$,可以推广到$n$个互不相容事件(满足$\bigcup A_i = \Omega$)。

证明：$P(B) = P(B \cap \Omega) = P(B \cap (A \cup \bar{A}))$
$P(B) = P(BA) + P(B\bar{A}) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A})$

EXAMPLE (风险评估)

保险公司认为开车者分为两类。一类容易出事故，一类为安全者。统计发现容易出事故在一年内发生事故的概率是0.4；安全者为0.2. 设第一类人口比例为30%. 现有一个新司机买保险，问他在一年内出事故的概率是多少？如果他一年内出了事故，他是容易出事故者的概率是多大？

解答：记A =容易出事故者；B =一年内发生事故；
(1): $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 0.4 \times 0.3 + 0.2 \times 0.7 = 0.26$
(2): $P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B|A)}{P(B)} = 6/13$.

Bayes概率与因果推理***

Thomas Bayes(1702-1761)

Theorem (贝叶斯定理)

$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = P(A)\frac{P(B|A)}{P(B)}$。

Proof: $P(AB) = P(A|B)P(B) = P(B)P(B|A)$

- 设A是原因，B是结果。已知B发生，推断A是否发生？
- 称$P(A)$是先验概率(prior)，$P(A|B)$是后验概率(posterior). $P(B|A)$是似然性(likeliwood), $P(B)$是边缘概率(marginal).
- 贝叶斯推断：新的事件或信息B改变了A的概率(贝叶斯定理)。→ 通过更多信息得到更确切的原因判断.

贝叶斯公式

Proposition (贝叶斯公式(逆概率公式))

$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(\bar{A})P(B|\bar{A})}$,可以推广到$n$个互不相容事件。

[证明]: 全概率公式有$P(B) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A})$，代入贝叶斯定理可得。

EXAMPLE

赌牌游戏：有三张牌。一张两面都是红的，一张两面都是黑的，一张两面是一红一黑。随机取出一张，如果是红的，反面是黑的概率多大？

记牌为RR, RB, BB,取出的牌朝上为红事件为A,

$P(RB|A) = \frac{P(RB\cap A)}{P(A|RR)P(RR)+P(A|RB)P(RB)+P(A|BB)P(BB)} = 1/3.$

抽签原理

EXAMPLE (抽签原理)

n个签中有m个为中奖。有放回的抽签：任一次中奖的概率为m/n。无放回的随机抽签，任一次的中奖概率也是m/n. 即抽奖与次序无关!

Proof.

有放回情形：显然任一次概率m/n.

无放回情形：记A_k表示第k次中奖的事件。则$P(A_1) = \frac{m}{n}$.

数学归纳法:设$P(A_{k-1}) = m/n$, 对所有m, n都成立，则用全概率公式

$P(A_k) = P(A_k|A_1)P(A_1) + P(A_k|\bar{A}_1)P(\bar{A}_1)$

$P(A_k) = \frac{m-1}{n-1}\frac{m}{n} + \frac{m}{n-1}\frac{n-m}{n} = m/n.$

直接证明：设第k次中奖，样本空间为n个取k个的排列共A_n^k. 事件A_k,第k次抽奖中奖的可能性为m种;其余$k-1$个任意排列共A_{n-1}^{k-1}.

$P(A_k) = m * A_{n-1}^{k-1}/A_n^k = m/n.$ □

假阳性之谜

医学里诊断疾病时对某种化验结果或试验结果，也有阴阳性的区分。阳性表示体内有某种病原体存在或者对某种药物有过敏反应.

EXAMPLE (疾病判断)

已知人群中某种疾病的发病率是0.1%. 有个抽血试验可以诊断该疾病，但准确率是90%(有病为阳性或无病为阴性的概率).有个人(甲)抽血试验为阳性，问医生有多大把握(概率)判断这个人有该疾病？

解答.

记A =甲患病，B =甲的试验结果阳性;

$P(A) = 0.001, P(B|A) = 0.9, P(B|\bar{A}) = 0.1$

$P(A|B) = \frac{0.001\times0.9}{0.001\times0.9+0.999\times0.1} \approx 0.0089$ □

罕见病的检验结果更有可能是假阳性!

Monty Hall 三门问题

Monty Hall: 美国电视节目主持人.

问题：有三扇门，其中一个后面有大奖(汽车). 幸运观众选好一个门后，主持人打开一扇门，显示其中没有奖，问该观众应该坚持原来的选择还是改变选择？

- 答案：永远应该改变选择!
- 概率分析：坚持(一种可能) → 赢奖概率1/3, 改变(两种可能) →赢奖概率2/3
- 理解性解释：假设有一万扇门，选定一个门后，主持人打开所有9998扇门，只剩一扇门，是否改变选择？
- 条件概率分析：假设$C = 1, 2, 3$表示汽车在那个门,$S = 1, 2, 3$表示观众的选择，$H = 1, 2, 3$表示主持人打开的门。设$S = 1, H = 3$, 计算概率$P(C = 2|H = 3, S = 1) = 2/3$。

作业

北航教材:

P33 习题一. 24,26,28,31

Review：回顾

Remark (作业事宜)

每周四上午十点交到各班概率统计信箱！
下周一上午在信箱取回作业。
助教：李翠英*cuiyinglee@163.com*
答疑：每周三晚上主*216(*概率统计课程答疑*)*

RECALL:

- 概率模型=样本空间+概率律
- 计算：加法公式，乘法公式，全概率，贝叶斯公式；
- 例子: 配对问题，抽样原理，保险评估，假阳性之谜；

TODAY

- 条件概率与独立性。
- 真实世界的概率模型？影响因素极多(独立性) → 简化模型
- 例子：系统与重复试验

独立性的定义

Definition (两个事件独立)

如果A, B满足$P(AB) = P(A)P(B)$, 称事件A, B互相独立。

- 条件概率解释：如果A, B独立($P(A) > 0$)，则$P(B|A) = P(B)$。
- 理解: 不相容事件($A \cap B = \emptyset$) 不相互独立!
- 必然事件Ω和不可能事件$\emptyset$与任何事件相互独立。
- 如果A, B独立,则$\bar{A}, B$相互独立.
- 例子：掷骰子：A =结果为偶数；B =结果为三的倍数。问A, B是否独立？
  正四面体，正八面体骰子呢？

独立性的判断

EXAMPLE

掷骰子两次，事件A =第一次为1, 事件B =第二次为6, A与B是相互独立。记事件C =两次的和为7,A与C是否相互独立？

Proof.

$$P(A = 1, B = 6) = \frac{1}{36} = P(A = 1)P(B = 6)$$
$$P(A = 1, C = 7) = \frac{1}{36} = P(A = 1)P(C = 7) \qquad \square$$

问题：已知C发生，A, B独立吗？

多次事件的独立性

EXAMPLE (条件影响独立)

掷硬币两次。事件A =第一次为正面H, 事件B =第二次为正面H, 事件C =两次的结果不同,如果已知C发生，问A, B是否相互独立?

解答 $P(A|C) = P(A) = 1/2, P(B|C) = P(B) = 1/2$,
$P(AB|C) = 0 \neq P(A|C)P(B|C)$.
事实上A, B, C两两独立但三个不独立。

Definition (n个事件独立)

如果$A_1, A_2, \ldots, A_n$满足,任取其中$k(1 \leq k \leq n)$个事件,有$P(A_{i_1}A_{i_2}\cdots A_{i_k}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k})$. 称$A_1, A_2, \ldots, A_n$相互独立。

- n次独立试验; n次无放回抽样?
- 所有事件相互独立→ 任意两组事件相互独立。
- 理解: 两两相互独立不一定是所有事件相互独立。

系统的可靠性

EXAMPLE (系统)

系统有若干元件组成，系统的可靠性由元件的可靠性和系统的结构(网络)共同决定。

- 基本假定：所有元件是相互独立的。
- 连接方式：并联或串联。
  设每个元件正常工作的概率为$P(A_i)$.
  概率公式：串联系统：$P(A_1A_2\cdots A_n) = P(A_1)P(A_2)\cdots P(A_n)$.
  并联$P(A_1 \cup A_2 \cdots \cup A_n) = 1 - (1 - P(A_1))(1 - P(A_2)\cdots(1 - P(A_n))$.

EXAMPLE (教材P29,例4)

$2n$个元件可以组成两个系统，比较两个系统的可靠性。

独立试验与二项概率模型

EXAMPLE (教务网站堵塞)

设教务网站有c个服务器，每个可以处理100个网页访问请求，用户帐号为n个。如果固定时间段内每个帐号访问的概率是p,问网站堵塞的概率？（即服务器不够用)。

- 基本假定：所有帐号访问是独立的。
- 固定时间段内访问帐号的个数记为X,则$P(X = k) = C_n^k p^k(1 - p)^{n-k}$
- 网站堵塞的概率$P = \sum_{i=100c}^{\infty} P(X = i)$
- 若$c = 15, n = 20000, p = 0.1$ 呢？

一般称n次重复试验结果出现k次的概率为二项概率。

QUIZ 小测验一

作业：北航教材: P35 习题一. 35,38,39,40

1. 设A, B 为任意两事件,则下列关系成立的有( )
   (A) $(A + B) - B = A$;(B) $(A + B) - B = A - B$;
   (C) $(A - B) + B = A$;(D) $(A - B) + B = AB$.
2. 从$0 \to 9$ 这十个数码中任意取出4个排成一串数码,则数码恰成四位偶数的概率为: （A）$\frac{41}{90}$;（B）$\frac{1}{2}$; （C）$\frac{40}{90}$; （D）$\frac{32}{90}$
3. 一盒子内装有5个红球,15个白球,从中不放回取10次,每次取一个球, 则第5次取到的是红球的概率为多少？
4. 袋中装有编号的八个球,从中任取3个,则最小号码为偶数的概率为多少？

答案：1. B, 2.A , 3 1/5, 4 11/28

概率统计与随机过程

Probability, Statistics and Stochastic Process

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 Systems sciences, BUAA

October 15, 2010

Week 5：随机变量与离散分布

Review：回顾

RECALL:

- 概率模型=样本空间+概率律
- 古典模型：一个因素影响下的概率模型。
- 复杂系统(或重复试验)：很多个(独立)因素影响下的概率模型？

TODAY

- 随机变量与分布函数；
- 二项分布。

Galton box: (bean machine)

随机变量的定义

什么是随机变量？random variable

- 直接解释：对每一个样本结果给出一个(实)数值。
- 数学解释；定义在一个样本空间上的函数。
  对比：函数的自变量与因变量。

EXAMPLE (投硬币)

投硬币三次，样本空间有8个元素。记每个结果中正面(H)的个数为X. X是个随机变量。

X取值$0, 1, 2, 3 \in \mathcal{R}$. 同理可定义随机变量$Y$为每个结果中反面的个数。

Definition (定义:随机变量)

随机变量X是定义在样本空间Ω上的实值函数:即每一个样本点ω,$X(\omega)$是个实数。通常用X, Y, Z或ξ, η等表示。

例子：人体正常体温T =多少摄氏度。华氏度$= 18/10c + 32$?

概率律的转移

Remark (push-back)

随机变量引导概率律转移到实数上。
设$x \in \mathcal{R}$, $X^{-1}(x) = \{\omega_1, \omega_2, \dots\}$.
则$P(X = x) = \sum_{\omega_i} P(\omega_i), \omega_i \in \Omega$.

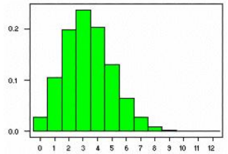
- X取值为离散的,称为离散随机变量, 概率律变成一个概率序列，称为分布列;
- X取值包含连续区间, 概率律呢？
  注意:$P(X = x_0) = 0!$，概率律变成一个(密度)函数。

EXAMPLE (人体温度)

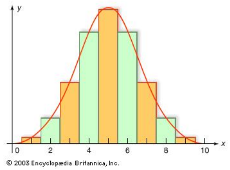
正常体温在$\mathcal{R}$上的有意义表示$36.8 \le X \le 37.2$(事件), 概率计算：$P(36.8 \le X \le 37.2) = P(\{\omega : 36.8 \le X(\omega) \le 37.2\})$。
区间足够小？得到一个函数。

分布函数的定义

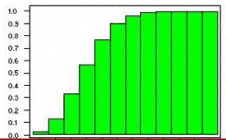
离散分布



连续分布



累积分布



Definition (概率分布函数)

给定随机变量X，定义函数

$$F(x) = P(X \le x), -\infty \le x \le +\infty$$

称F为X的概率分布函数，简称分布函数或累积分布函数CDF。

分布函数的充分必要条件：

- $0 \le F(x) \le 1$.
- $F(x)$单调非减
- $F(x)$右连续；***

通常随机变量对应一些常见分布函数，直接称随机变量为某个分布。

随机变量的注解***

- 随机变量不仅仅是一个函数$X : \Omega \to \mathcal{R}$;
  更重要的是得到分布函数。可以记为一个变换$(\Omega, P) \to (\mathcal{R}, F)$。
  事实上分布函数更重要。
- 随机变量的函数还是随机变量(复合函数)。$\Omega \to \mathcal{R} \to \mathcal{R}$
- 区分分布函数可以用函数的数字特征：期望，方差等等。
- 随机变量分类：离散随机变量，非离散随机变量(我们一般仅仅考虑其中的连续随机变量);

分布函数的判断

分布函数的性质

- $F(-\infty) = 0, F(+\infty) = 1$
- $P(a \le X \le b) = F(b) - F(a)$
- $P(X = x) = F(x) - F(x^-)$ ***

EXAMPLE

设$F(x) = a + be^{-x}, x \ge 0$, $F(x)$是否可以是某个随机变量的分布函数？特别求概率$P(X \le \ln 2)$.

$a = 1, b = -1, P = 1/2$
说明：有很多分布函数，但常见只有几种。

二项分布

EXAMPLE (两点分布，Bernouli贝努利分布)

X取值为0或1时的概率分布是
$P(x=1)=p, P(X=0)=1-p$, 记为$B(1,p)$.
模型：最简单随机变量=一次试验是否成功。

Definition (二项分布或二项随机变量)

随机变量的取值为$0,1,2,\ldots,N$,且满足$P(X=k)=C_N^k p^k q^{N-k}, k=1,2,\ldots,N, p+q=1$.
称其满足二项分布，记为$B(N,p)$.

- 模型：N次独立试验中的成功次数。
- 称为二项分布因为其值为满足二项式定理的各项。
  $(p+q)^N=\sum C_N^k p^k q^{N-k}$,可递推计算,$N$充分大时,需近似计算;
- 应用：投票权大小(委员会组成); 比赛规则(三局两胜,五局三胜?)...

应用例子

EXAMPLE (基因遗传)

设某个生物特征(眼睛颜色，或左撇子)由一对基因决定。D为显性基因，r为隐性基因。后代从父母中各得到一个基因。只要有一个显性基因，就必然出现该生物特征。问一对混合型父母(基因为Dr)的四个后代中有三个有该生物特征的概率是多少?

解答： 服从二项$B(4,3/4)$，$P(X=3)=C_4^3(\frac{3}{4})^3\frac{1}{4}=\frac{27}{64}$.

EXAMPLE (安全)

设飞机上每个引擎的失效概率相同为p且互相独立.如果一个飞机上的一半引擎正常，则飞机可以正常运行。问p为多少时，用四个引擎的飞机比用两个引擎安全?

解答： 四个引擎正常个数服从$B(4,p)$，飞机正常飞行即$X\geq 2$,
$P(X\geq 2)=1-(1-p)^4-4p(1-p)^3$;类似两个引擎飞机正常飞行$Y\geq 1, P(Y\geq 1)=1-(1-p)^2$。计算可得$p\geq 2/3$时，四个引擎的飞机更安全。

随机变量与决策

EXAMPLE (轮盘赌Roulette)

美式轮盘赌有38个数值，最简单的玩法押一个数值一元，可赢得35元。问一次押注的期望赢钱多少?

解答： $EX=35\frac{1}{38}+(-1)\frac{37}{38}=-1/19\approx -0.053$

作业

北航教材:
P60 习题二. 1,4,6,8

定义

Definition (离散随机变量)

随机变量取值为有限或可数个时，称为离散随机变量。
记其值为$x_1.x_2,\dots,x_k,\dots$。
特别记$P(X=x_k)=p_k, k=1,2,\dots$，$P$是$X$对应的概率分布，称$\{p_k\}$为概率分布列。

- 常记为

X	x_1	x_2	x_3	$\dots$
P	p_1	p_2	p_3	$\dots$

- $p_k \geq 0, \sum p_k = 1.$
  注：$\Omega = \bigcup_i \{\omega | X(\omega) = x_i\}$
- 分布函数$F(x)$是阶跃函数。

N次独立试验中的分布

- 模型：N次独立试验或N次有放回抽签。
- 二项分布: $X=k$,试验中的成功次数, $B(N,p)$
  $X=X_1+X_2+\cdots+X_N$,其中X_i是第i次试验是否成功的两点分布。
- 几何分布：$Y=k$,首次成功时的试验次数。
  [几何分布]: 随机变量取值为$1,2,\dots,\infty$,且满足$P(Y=k)=p(1-p)^{k-1}, k=1,2,\dots,\infty.$
- ***负二项分布：$Y=k$,第r次成功时的试验次数。$NB(r,p)$
  $Y=Y_1+Y_2+\cdots+Y_r$, Y_i一次成功时的试验次数。
  [负二项分布] 随机变量取值为$1,2,\dots,\infty$,且满足$P(Y=k)=C_{k-1}^{r-1}p^r(1-p)^{k-r}, k=1,2,\dots,\infty.$

Banach火柴盒问题：有两个火柴盒(各有N个)，随机选取一个,直到发现一个盒子空了为止。问另外一个盒子有k个火柴的概率。
答案：固定一个盒子：$Y\sim NB(N+1,1/2)$,试验次数$Y=2N+1-k$.
$P(Y=2N+1-k)=C_{2N-k}^{N}(1/2)^{2N-k+1}$.
两个盒子的概率为上面两倍。

无放回抽签的分布

Definition (超几何分布)

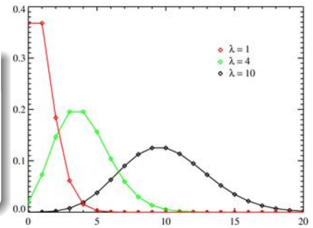
随机变量取值为$0,1,2,\dots,n$,且满足$P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, k=1,2,\dots,n;$
称其满足超几何分布，记为$H(n,M,N)$.

- 模型：从(N,M)大样本中抽取小样本n,M是N中特别一类(如次品),则n中次品个数k服从超几何分布。特别$n=1,k=1$有抽签原理!.
- ***较大N,M时,可用二项分布逼近。$B(n,p), p=M/N$,即有放回与无放回差别不大。
- 实用例子：估计某地区的某动物总数N。
  捕捉一批动物M，做标记放回，过一段时间，再捕捉一批动物n, 其中有标记的为k, 利用超几何分布，估计N。
  概率为$P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$是关于$N$的一个序列，
  我们求最大值对应的N,计算有$N=nM/k$. 称为最大似然估计。

定义

Definition (泊松分布)

随机变量取值为$0,1,2,\dots,k,\infty$,满足$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,2,\dots;\lambda$为正常数. 称其满足泊松分布，记为$\mathcal{P}(\lambda)$.



- 模型：大量试验中的小概率事件发生次数。
  每年发生战争(或地震)的次数,某一小时进入某邮局的顾客数；寿命超过100岁的人数；
- 逼近二项分布：当n大,p小，np合适大小时，可用$\lambda=np$的泊松分布近似计算。
- *** 泊松范例：n个小概率事件发生的概率p_i,所有事件互相独立或弱相依，则事件发生次数服从$\lambda=\sum p_i$的泊松分布。
  配对问题：N个人得到自己作业的次数近似服从$\lambda=N*1/N=1$的泊松分布.

应用例子

- (事故发生) 设某高速公路段每天发生事故的次数服从$\lambda = 3$的泊松分布，求今天不发生事故的概率。
  解答： $P(X = 0) = e^{-3} \approx 0.05$.
- (印刷错误)设某本书任一页有印刷错误的次数服从$\lambda = 1$的泊松分布，求某一页出现至少一个错误的概率
  解答： $P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-1} \approx 0.633$

EXAMPLE (生日问题***)

任意两个人生日相同的事件是小概率，弱相依的；设发生事件的次数服从$\lambda = C_n^2 \frac{1}{365}$的泊松分布，求$n$个人里至少两个人同生日的概率。

解答： 一次试验为任选两个人为同一天生日；$p = 1/365$, 试验次数(任选两个人)C_n^2. 出现两个人生日相同次数X服从泊松分布。
有$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda}$
如果要求$P(X \geq 1) > 0.5$,即$e^{-n(n-1)/730} \geq 0.5$; 计算有$n \geq 23$.
多少人中有三个人生日相同？$\lambda = C_n^3 (\frac{1}{365})^2$.

复合函数的分布

构造新的分布：$g(X)$

- 设X是随机变量,g是$R \to R$的函数，$Y = g(X) : \Omega \to R \to R$，$Y$是随机变量；
- 离散分布: $p_0, p_1, p_2 \ldots$
  例子: $Y = X^2$, $Y = 4 \to X = \pm 2$,
  $P(Y = 4) = P(g^{-1}(4)) = P(X = 2) + P(X = -2)$;
  参见第四章例1.

分布的特征：期望expectation

Definition (离散随机变量期望)

设离散随机变量X的概率分布列为p_i,且$\sum_{i=0}^{+\infty} |x_i| p_i < +\infty$,
称$EX = \sum_{i=0}^{+\infty} x_i p_i$为$X$的期望(值).

- 随机变量的平均特征：又称均值(概率加权平均);
- 期望可能无穷，有限取值随机变量有有限期望；
- 掷骰子的期望值是$EX = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$

常见离散分布的期望；

- 两点分布(贝努利分布) $EX = p$
- 二项分布$B(N, p)$，$EX = np$
- 泊松分布$\mathcal{P}(\lambda)$,$EX = \lambda$
- 几何分布$EX = 1/p$;
- 超几何分布$H(n, M, N)$, $EX = nM/N$.

作业

北航教材:
P60 习题二. 12，13，14，18

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

October 22, 2010

Week 6：连续随机变量及分布特征

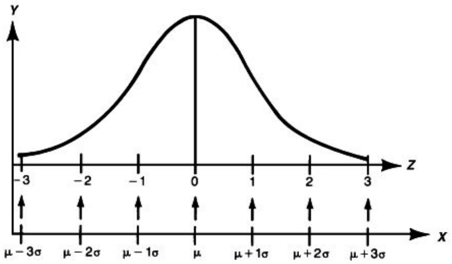
Review：回顾

RECALL:

- 随机变量与分布函数
- 离散分布：两点分布，二项分布，几何分布，泊松分布；
- 问题：
  超几何分布与抽取次序无关。(抽签原理)
  泊松分布逼近：生日问题的试验次数？

TODAY

- 连续随机变量：
- 常见连续分布：



The Translation of X to Z by the Transformation $Z = (X - \mu)/\sigma$

Figure 3

正态分布，高斯分布，Bell曲线(钟型曲线)，

连续随机变量的定义

从离散到连续

- 样本空间：有限(或可数) → 不可数(一般是区间)
- 概率律：概率分布列p_i → 概率密度函数$f(x)$
  $p_i = P(X = x_i)$; $P(x \le X \le x + \delta) \sim f(x)\delta$

例子：打靶的环数:$X = 10, 9, 8, ..., [0, 10]$.

Definition (连续随机变量)

随机变量X取值为实数中区间时，如果存在非负函数$f(x)$,使得任意$-\infty \le a < b \le \infty$, 有$P(a < X \le b) = \int_a^b f(x)dx$, 称$X$为(绝对)连续随机变量。特别称$f(x)$是$X$的概率密度函数*probability density*。

性质：

- 密度函数的充分必要条件：$0 \le f(x), \int_{-\infty}^{\infty} f(x)dx = 1$.
- 分布函数与密度函数$F'(x) = f(x)$(除有限点).

均匀分布

Definition (uniform 分布)

如果连续随机变量X的密度函数满足，$f(x)=\frac{1}{b-a}, x\in(a,b)$, $f(x)=0$, 其他情形。记为$X\sim\mathcal{U}(a,,b)$.
特别称$I_{(a,b)}=1, x\in(a,b)$为示性函数，$f(x)=\frac{1}{b-a}I_{(a,b)}$

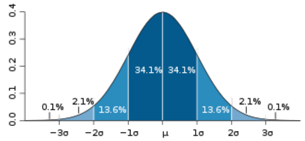
模型：等可能事件的推广．特别有分段均匀函数(离散简化：两点分布)

EXAMPLE (随机到达)

设某路公共车从5点开始每十五分钟发车；假设一个人在8点到8:30之间随机到达车站，问(a) 他等候时间小于5分钟的概率(b)他等候时间大于10分钟的概率。

解答 设到达时间X服从均匀分布$\mathcal{U}(0,30)$；(a)
$P=P(10<X<15)+P(25<X<30)=\int_{10}^{15}\frac{1}{30}dx+\int_{25}^{30}\frac{1}{30}dx=1/3$
(b)类似$P=P(0<X<5)+P(15<X<20)=1/3$

正态分布

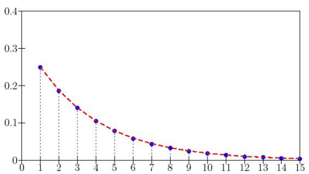


Definition (正态分布,高斯分布，Bell曲线Normal)

如果连续随机变量X的密度函数满足，
$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x\in R$,称$X$服从参数为$(\mu,\sigma)$的正态分布。记为$X\sim N(\mu,\sigma^2)$.
特别称$N(0,1)$为标准正态分布。记密度函数为$\Phi(x)$.

- 二项分布的逼近：n充分大当可用正态分布近似计算。
Demoivre-Laplace 定理：$B(n,p)\to N(np,np(1-p))$
- 模型：大量物理或生物数据的特征指标
高斯分布:研究测量误差的分布；IQ分数，成绩；
- 特别μ表示平均值，σ表示差异大小；
$f(x)$关于$x=\mu$对称，$f(\mu)=1/(\sqrt{2\pi\sigma^2})$最大；

指数分布



Definition (指数分布)

如果连续随机变量X的密度函数满足，$f(x)=\lambda e^{-\lambda x}, x\geq 0$, $f(x)=0$, 其他情形。记为$X\sim\varepsilon(\lambda)$.

- $F(\alpha)=1-e^{-\lambda\alpha}, \alpha\geq 0$;
几何分布与指数分布：$F_{geom}(N)=1-(1-p)^N\sim F(N\delta)$
- 模型：小概率事件之间的间隔时间(等待时间);
下一次类似汶川地震发生的时间；人或机器的寿命；排队时间；
- (定理) 指数分布的无后效性：$P(X>s+t|X>s)=P(X>t)$
- 排队等待概率:设有两个窗口，两个人正在办理，每个人的办理时间服从指数分布$\varepsilon(\lambda)$, 设你是下一个，则你最后办完的概率是1/2.

其他连续分布

- $\Gamma(\alpha,\beta)$分布：α次成功的等待时间。
$\Gamma(n,\beta)=\epsilon_1(\beta)+\epsilon_2(\beta)+\ldots$.
离散模型：负二项分布。
- Weibull分布：最弱环节的寿命(可靠性) $F(x)=1-e^{-(x/\eta)^\beta}$.
- Beta分布：$f(x)=\frac{1}{\beta(a,b)}x^{a-1}(1-x)^{b-1}, 0<x<1$
$\beta(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma a+b}$.
- 其他复合随机变量的分布：对数正态分布，...

密度函数的判断

密度函数的性质

- $f(x) \geq 0, \int_{-\infty}^{\infty} f(x)dx = 1$
- $F(x) = \int_{-\infty}^{x} f(t)dt$

EXAMPLE

设$f(x) = a\sin x, \pi \geq x \geq 0$, $f(x)$是否可以是某个随机变量的密度函数？特别求其分布函数.

$a = 1/2$
$F(x) = 1/2(1 - \cos x), 0 \leq x < \pi$.

指数分布与失效函数

Definition (失效函数:失效率，死亡率)

定义$\lambda(t) = \frac{f(t)}{1-F(t)}$为分布$F$的失效函数。

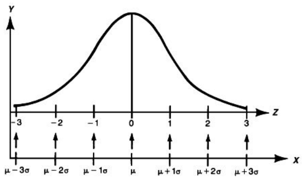
注记： 解释：寿命到达t时刻后在dt内失效的概率.
指数分布的失效函数为λ,称为死亡率。一般的失效函数决定一个分布。

EXAMPLE (吸烟与寿命)

一般可设人的寿命满足死亡率为λ的指数分布。研究说明吸烟者的死亡率是非吸烟者的两倍。设一个不吸烟的50岁人的正常死亡率为$1/30$,则同龄的吸烟者死亡率为$1/15$. 计算两个人活到60岁的概率。

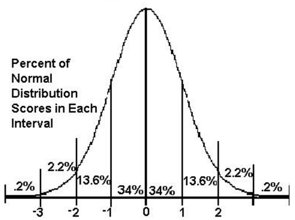
解答： $P(X > 10) = 1 - F_X(10) = e^{-1/3} \approx 0.71$
$P(Y > 10) = 1 - F_Y(10) = e^{-2/3} \approx 0.51$.

标准正态分布及分位数



The Translation of X to Z by the Transformation $Z = (X-\mu)/\sigma$

Figure 3



- 正态分布的线性变换：$X \sim N(0,1), Y = \sigma X + \mu \sim N(\mu, \sigma^2)$
- 正态分布的取值分布:$\mu \pm \sigma, 68.27\%$, $\mu \pm 2\sigma, 95.45\%$, $\mu \pm 3\sigma, 99.73\%$, $P(|X - \mu| \leq 6\sigma) = 99.99999980\%$; 六西格玛管理!
- 正态分布表：$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. 特别有$z_\alpha$称为$\alpha$分位数:满足$P(Z \leq z_\alpha) = \alpha, 0 \leq \alpha \leq 1$.

应用(考试成绩)： 希望一个好的考试成绩是正态分布,设成绩分类为A, B, C, D, F, 可以给成绩如下: $A = \{X > \mu + \sigma\}, 15.67\%$,
$B = \{\mu < X < \mu + \sigma\}, 34.13\%$, $C = \{\mu - \sigma < X < \mu\}, 34.13\%$
$D = \{\mu - 2\sigma < X < \mu - \sigma\}, 13.59\%$, $F = \{X < \mu - 2\sigma\}, 2.28\%$.

作业

北航教材:
P60 习题二. 20,23,27,33

标准正态分布

Remark (标准正态分布)

$N(0,1)$为标准正态分布,密度函数满足，$f(x) = \frac{1}{\sqrt{2\pi}} exp^{-\frac{x^2}{2}}, x \in R$,
特别记分布函数为$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$.

- 结论：$\int_{-\infty}^{\infty} e^{-t^2/2} dt = \sqrt{2\pi}$.
- 密度函数性质: 关于0点对称，0点最大，两边单调。
- 分布函数性质: $\Phi(0) = 1/2, \Phi(-x) = 1 - \Phi(x)$.
- 分位数点: $\Phi(-1.96) = 0.025, \Phi(-1.645) = 0.05$
  $P(X > z_{1-\alpha}) = \alpha$, $P(|X| > z_{1-\alpha/2}) = \alpha$

3σ,6σ管理的概率：$X \sim N(\mu, \sigma^2)$.
则$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 0.9974$.

正态分布与信号噪音

Remark

一般假设噪音服从正态分布$N(0, \sigma^2)$(与其他随机变量独立)。

EXAMPLE (信号传输)

假设传输二进制信号0, 1, 发送信号为s,接收信号为R, 由于有随机噪音$N \sim N(0, 1)$影响, 可设$R = s + N$。如果信号是1,发送$s = 2$，信号是0，发送$s = -2$。指定接收信号的判断法则：$R \ge 0.5$,$s^* = 1$,
$R < 0.5, s^* = 0$. 求该法则的错误概率。

解答:.

两类错误: 信号为0,判断为1, 信号为1,判断为0. 即$P(s^* = 1|s = -2) = P(R \ge 0.5) = P(-2 + N > 0.5) = \Phi(-2.5) \approx 0.062$,
类似有$P(s^* = 0|s = 2) = P(R < -1.5) = \Phi(-1.5) \approx 0.0668$. □

复合函数的分布

已知X的分布，问$Y = g(X)$的分布；

- g是$R \to R$的函数；$Y : \Omega \to R \to R$，Y是随机变量；
- 离散分布: 求原像。例子$Y = X^2$,$Y = 4 \to X = \pm 2$,
  $P(Y = 4) = P(g^{-1}(4)) = P(X = 2) + P(X = -2)$;
- 连续分布：分布函数求原像$F(y) = F(g(x))$;
  $P(g(X) \le y) \sim ?P(X \le g^{-1}(y))$

EXAMPLE (正态分布的线性变换)

设$X \sim N(0, 1)$,则$Y = \sigma X + \mu$服从$N(\mu, \sigma^2)$。

EXAMPLE (χ^2分布)

设$X \sim N(0, 1)$,则$Y = X^2$服从χ^2分布。$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}$.

复合函数的密度函数

Theorem (复合函数的密度函数)

设X的密度函数$f(x)$,$Y = g(X)$;Y的取值范围为D, 对任意$y \in D$, 设其原像可以写成有限点(即$\sharp(g^{-1}(y))$有限).每一个可以写成一个逆函数$h_i(y) = x_i$; 则Y的密度函数为$f_Y(y) = \sum_{i=1}^{n} f(h_i(y))|h_i'(y)|$。

注记：参见教材4.2节定理1.推广到分段单调函数即可。

特例证明:

假设$g(x)$是严格单调增的(必然有唯一原像),设逆函数$h(y)$.
$F_y(y) = P(Y \le y) = P(g(X) \le y) = P(x \le h(y)) = F_x(h(y))$,
两边求导数有$f_y(y) = f_x(h(y))h'(y)$. □

例子: $X \sim N(0, 1)$, $Y = X^2$.
每一个y有两个原像，可以定义两个逆函数$h_1 = \sqrt{y}, h_2 = -\sqrt{y}$; 计算有$f_Y(y) = \Phi(h_1(y))|h_1'(y)| + \Phi(h_2(y))|h_2'(y)| = \frac{e^{-y/2}}{\sqrt{2\pi y}}$.

随机变量的特征

正态分布的特征$N(\mu,\sigma^2)$: μ表示平均值，σ表示差异大小。

Definition (连续随机变量期望)

设连续随机变量X的密度函数为$f(x)$,且$\int_{i=-\infty}^{+\infty}|x|f(x)dx<+\infty$,
称$EX=\int_{-\infty}^{+\infty}xf(x)dx$为$X$的期望(值).

注：离散随机变量：$EX=\sum_{i=0}^{+\infty}x_ip_i$；
离散到连续：$EX=\sum xP(x\le X\le x+dx)\approx\sum x(f(x)dx)$

Definition (方差)

设随机变量X的期望$EX=\mu$有限, 则称$Var(X)=E(X-\mu)^2$为X的方差.又记为DX.

可以验证: $EX=\mu, Var(X)=\sigma^2$. 注：还有其他特征,参见第五章。

常见随机变量的期望与方差

常见离散分布；

- 两点分布，Bernouli贝努利分布$EX=p, VarX=pq$
- 二项分布$B(N,p)$，$EX=np, VarX=npq$
- 泊松分布$\mathcal{P}(\lambda), EX=\lambda, VarX=\lambda$
- 几何分布$EX=1/p, VarX=q/p^2$;

常见连续分布：

- 均匀分布$\mathcal{U}(a,,b), EX=\frac{a+b}{2}, VarX=\frac{(b-a)^2}{12}$;
- 指数分布$\varepsilon(\lambda)$, $EX=1/\lambda, VarX=1/\lambda^2$;

作业

北航教材:
P60 习题二. 35, 37
P120 习题四. 7,13

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

October 29, 2010

Week 7：随机向量及其分布

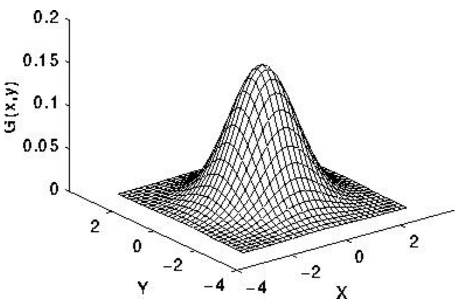
Review：回顾

RECALL:

- 随机变量与分布函数
- 常见离散分布：两点分布，二项分布，几何分布，泊松分布；
- 常见连续分布：均匀分布，正态分布，指数分布；
- 分布的特征与复合：

TODAY

- 随机向量：
- 常见二维分布：



二维正态分布(高斯分布)

随机向量的定义

Definition (随机向量)

随机向量$\vec{X}$是定义在样本空间Ω上的实值多元函数:即每一个样本点ω,$\vec{X}(\omega)$是个向量。通常用$\vec{X},\vec{Y},\vec{Z}$或$\vec{\xi},\vec{\eta}$等表示。

记n是向量的维数：一般考察二维向量，记为$\vec{X}=(X,Y)$.

EXAMPLE (身体指标)

$\vec{I}=(H,W)$。H是身高(m)，W是体重(kg)；假设H,W为正态分布；身体状况是怎么样的？身高与体重的关系？BMI指标$=W/H^2$。

EXAMPLE (多项分布)

掷骰子n次，结果可以为$(1-6)$,记其中为$X_i=n_i$为结果为i的次数；则$\vec{X}=(X_1,X_2,\ldots,X_6)$是六维随机向量,其分布为
$P(X_1=n_1,X_2=n_2,\ldots,X_6=n_6)=\frac{n!}{n_1!n_2!\ldots n_6!}p_1^{n_1}\ldots p_6^{n_6}$. 其中$\sum n_i=n$.
一般的称为多项分布。(***边缘分布是二项分布)

联合分布函数

Definition (联合累积分布)

给定二维随机向量(X,Y),定义$F(x,y)=P(X\le x)\cap P(Y\le y)=P(X\le x, Y\le y)$。称为随机向量$(X,Y)$的分布函数，或随机变量$X,Y$的联合分布。

- ***分布函数可以决定任意事件(集合)的概率$P(\{(x,y)\in D\})$。(测度论)
- 分布函数的充分必要条件:
  1. $0\le F(x,y)\le 1. F(x,-\infty)=F(-\infty,y)=F(-\infty,-\infty)=0, F(+\infty,+\infty)=1.$
  2. $F(x,y)$对x或y单调非减
  3. $F(x,y)$对x或y右连续; ***
  4. 任意$x_1<x_2, y_1<y_2$，有$F(x_2,y_2)+F(x_1,y_1)-F(x_1,y_2)-F(x_2,y_1)\ge 0.$
- 分类：二维离散随机向量，二维连续随机向量，混合型随机向量；

二维离散分布

Definition (二维分布矩阵)

给定(X,Y)二维随机变量，取值至多可数；记$p_{i,j}=P(X=x_i,Y=y_j)$，称$p_{i,j}$为二维离散分布(矩阵)；$\sum_i\sum_j p_{i,j}=1$.

EXAMPLE (人口调查)

设一个家庭至多三个小孩，$X=i, Y=j$对应男孩与女孩的个数。分布如下：

$i:j$	0	1	2	3	p_i
0	0.15	0.10	0.0875	0.0375	0.375
1	0.10	0.175	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2
3	0.0375	0	0	0	0.0375
q_j	0.375	0.3875	0.2	0.0375	1

注意：令$p_i=\sum_j p_{i,j}, q_j=\sum_i p_{i,j}$为矩阵的行和与列和；
则$\sum p_i=\sum q_i=1$ 称为边缘分布!

二维连续分布

Definition (二维连续随机向量)

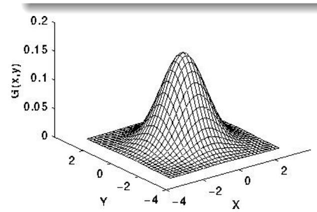
给定(X,Y)二维随机变量，如果存在二元非负可积函数$f(x,y)$使得，任意区域D，有$P((X,Y)\in D)=\iint_D f(x,y)dxdy$,称$(X,Y)$是连续随机向量。$f(x,y)$称为联合密度函数。特别: $\iint_{R^2} f(x,y)dxdy=1$.

- 分布函数$F(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x} f(x,y)dxdy$.
- 定理: 连续可微的联合分布函数有联合密度函数，所以是连续随机向量。即$\frac{\partial^2 F}{\partial x\partial y}=f(x,y)$.
- 均匀分布: $f(x,y)=\frac{1}{Area(D)}, (x,y)\in D$.
  约会问题: 两个人约定见面。假设在一小时内随机到达，求第一个人等十分钟以上的概率。
  解答：
  $P(X+10<Y)=\int_D f(x,y)dxdy=\int_0^{50}dx\int_{x+10}^{y}(1/60)^2dy=25/36$

二维正态分布bivariate normal distribution

EXAMPLE (二维正态分布)

给定(X,Y)二维随机变量，称为方程$N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2,\rho)$的二维正态分布，
如果其概率密度函数是:
$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}]\}$



- 参数μ_1,μ_2是均值，σ_1,σ_2是差异，
- ρ是相关系数。$|\rho|\le 1$.

边缘分布

Definition (边缘分布函数)

给定二维随机向量及其分布函数$F(x,y)$, 可以得到X,Y是随机变量，其分布函数$F_X(x)=F(x,\infty), F_Y(y)=F(\infty,y)$. 称为边缘分布函数。

- 随机向量的投射得到坐标随机变量: 对应的分布函数即边缘分布。
- 离散情形: 边缘分布列即矩阵的行和或列和。
- 连续情形: 存在对应的边缘密度函数$f_X(x)=\int_{-\infty}^{\infty}f(x,y)dy$, $f_Y(y)=\int_{-\infty}^{\infty}f(x,y)dx$.
- 联合分布决定边缘分布，反之不能。

分布函数的计算

EXAMPLE (P70 例2)

$f(x,y)=ae^{-2y}, 0\le x\le 2, y>0$.

- 确定常数a.
- 求分布函数$F(x,y)$.
- 求概率$P(|Y|\le X)$.
- 求X,Y的边缘分布函数;
- 求X,Y的边缘密度函数;

边缘分布与联合分布

Proposition

(二维)正态分布的边缘分布还是正态分布。
设$(X,Y)\sim N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2,\rho)$,则边缘分布$X\sim N(\mu_1,\sigma_1^2)$, $Y\sim N(\mu_2,\sigma_2^2)$。

注：配方求积分可得。

- 边缘分布不能得到联合分布($\rho=?$).
- 一般的联合分布与边缘分布的类型不一定一样。
圆盘上的均匀分布的边缘分布不是均匀分布。

如何利用边缘分布得到联合分布？条件分布或独立性!

作业

北航教材:
P90 习题三. 2,4,13,15

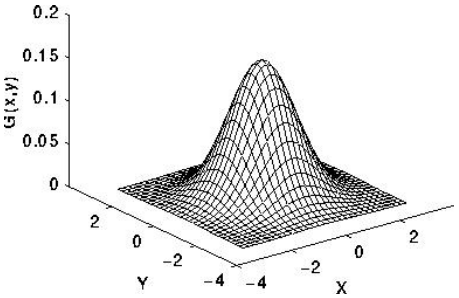
Review：回顾

RECALL:

- 随机向量与联合分布函数
- 离散分布矩阵，连续分布密度；
- 常见随机向量分布：多项分布，均匀分布，正态分布；
- 边缘分布及边缘密度函数；

TODAY

- 条件分布及计算；
- 随机变量的独立性；



二维正态分布(高斯分布)

条件概率与随机变量

- 条件概率公式：$P(B|A) = \frac{P(AB)}{P(A)}$
- *** 关于事件A的随机变量条件概率公式：$P(X \in B|A) = \frac{P(X \in B \cap A)}{P(A)}$.
  (条件)随机变量: $P(X \in B|A) = \frac{\int_{A\cap B} f(x)dx}{P(A)}$. 特别密度函数$f_{X|A} = f(x)/P(X \in A)$.
  ***有类似全概率公式的随机变量密度分解公式。
- 关于随机变量Y的条件概率公式：$P(X = x|Y = y) = ?$
  $P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$.
  说明：希望上面对于任意x, y都成立；
  问题: 连续情形$P(X = x) = P(Y = y) = 0$.

条件分布：离散情形：

Definition (离散条件分布)

给定(X, Y)二维离散随机变量，记$p_{i,j} = P(X = x_i, Y = y_j)$, 记$p_{.j} = \sum_i p_{i,j}, p_{i.} = \sum_j p_{i,j}$, 则$P(X = x_i|Y = y_j) = p_{i,j}/p_{.j}$,称为条件$Y = y_j$下的条件分布律。类似有条件$X = x_i$下的条件分布律。

EXAMPLE (人口调查)

设一个家庭至多三个小孩，$X = i, Y = j$对应男孩与女孩的个数。分布如下：

$i:j$	0	1	2	3	$p_{i.}$
0	0.15	0.10	0.0875	0.0375	0.375
1	0.10	0.175	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2
3	0.0375	0	0	0	0.0375
$p_{.j}$	0.375	0.3875	0.2	0.0375	1

有一个女孩的家庭的男孩分布律：$P(X = i|Y = 1) = p_{i1}/p_{.1}$

条件分布：连续情形

设(X, Y)是连续的二维随机向量，密度函数为$f(x, y)$.

- 局部逼近：$P(x \le X \le x + \delta, y \le Y \le y + \delta) \approx f(x, y)\delta^2$
- 边缘分布: $P(y \le Y \le y + \delta) \approx f_Y(y)\delta$
- 条件分布：$P(x \le X \le x + \delta|y \le Y \le y + \delta) \approx \frac{f(x,y)}{f_Y(y)}\delta$

Definition (连续条件分布)

给定(X, Y)连续二维随机向量，密度函数$f(x, y)$.
则$f_{X|Y}(x) = f(x, y)/f_Y(y)$,称为条件$Y = y$下的条件概率密度。
对应有条件分布函数$F_{X|Y}(x) = \int_{-\infty}^{x} f(u, y)du/f_Y(y)$.

有乘法公式$f(x, y) = f_{X|Y}(x)f_Y(y)$，可以推广到多个随机变量。

连续二维分布的例子

EXAMPLE (圆盘上均匀分布)

已知圆盘上均匀分布的概率密度为$f(x,y)=1/(\pi R^2), x^2+y2\le R^2$.
则边缘密度$f_Y(y)=\frac{2}{\pi R^2}\sqrt{R^2-y^2}, |y|\le R$.
条件密度函数$f_{X|Y}(x)=\frac{1}{2\sqrt{R^2-y^2}}, |x|\le\sqrt{R^2-y^2}$.
结论：边缘分布非均匀分布但条件分布是均匀分布(小区间)。

Proposition (正态分布的边缘与条件分布)

(二维)正态分布的边缘分布还是正态分布。
设$(X,Y)\sim N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2,\rho)$,则边缘分布$X\sim N(\mu_1,\sigma_1^2)$, $Y\sim N(\mu_2,\sigma_2^2)$。特别条件分布还是正态分布。$X|Y\sim N(\mu_1+\rho\sigma_1/\sigma_2(y-\mu_2),(\sigma_1\sqrt{1-\rho^2})^2)$.

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}]\}$$

独立性

Definition (独立性)

随机变量X,Y称为独立的，如果任意x,y,事件$\{X\le x\}$与$\{Y\le y\}$是独立的。即$P(X\le x,Y\le y)=P(X\le x)P(Y\le y)$. 可以推广到$n$个随机变量，称为独立序列。

X,Y相互独立等价于$F(x,y)=F_X(x)F_Y(y)$等价于$f(x,y)=f_X(x)f_Y(y)$等价于$p_{ij}=P_{i.}p_{.j}$. 一般的，随机向量的各个坐标分量独立可以得到简化的联合分布函数。

Theorem (独立的性质)

- 常数与任一随机变量独立；
- X,Y独立，则任意事件$X\in A$与$Y\in B$都独立；
- X,Y独立，任意函数$U=g(X),V=f(Y)$,则U,V是互相独立的；

二维正态分布的坐标随机变量相互独立的充要条件$\rho=0$.

***连续贝叶斯推断

Remark (观测量推测)

随机变量X是未知变量，Y是观测到的随机变量，推断X的信息？贝叶斯公式: $f_{X|Y}=\frac{f_Xf_{Y|X}}{f_Y}$.

EXAMPLE (信号检测)

假设原始信号$S=0,1$, $P(S=1)=p, P(S=0)=1-p=q$. 检测到信号为R, 受随机噪音$N\sim N(0,1)$影响, 可设$R=S+N$。如果观测到$R=y$,问$S=1$的概率？

注意:本题为混合型随机向量。其中边缘分布可用离散项求和得到。
$P(S=1|R=y)=\frac{P(R=y|S=1)P(S=1)}{P(R=y)}=\frac{pe^y}{pe^y+(1-p)e^{-y}}$
显然当y越大时，$S=1$的概率越大。

作业

北航教材:
P90 习题三. 12，17，20,25

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

November 5, 2010

Week 8：随机向量的函数及其分布

Review：回顾

RECALL:

- 随机向量及其分布
- 边缘分布与条件分布;
- 常见分布：均匀分布与二维正态分布；
- 独立性；

TODAY

- 随机向量的函数；
- 常见二维随机向量函数的分布；

二维正态分布(高斯分布)

随机向量的函数

Remark (随机向量的函数)

随机向量$\vec{X}$是定义在样本空间Ω上的实值N元向量函数，$g: R^n \to R$的实值函数，记$Y = g(\vec{X})$,则Y是定义在Ω上的一个随机变量。一般考察二维向量，记为$Z = g(x, y)$.

例子(身体指标):H是身高(m)，W是体重(kg). 假设(H, W)为二维正态分布，则BMI指标$= W/H^2$ 是一个随机变量(分布是什么？)

- 一般复杂g得到的分布函数比较复杂；我们仅关心简单情形或分布函数的特征；
- 应用: 到达过程(随机变量求和); 失效模型(随机变量的极小值)；延迟模型(随机变量极大值);
- 计算: $F_Z(z) = P(g(x, y \le z)) = \int\int_{g(x,y)\le z} f(x, y)dxdy$
  类似二重积分的累次积分方法: $F_Z(z) = \int_{-\infty}^{z}\int_{g(x,y)=u} f(x, y)dl$ (后面是曲线积分). ***来源于二维变换$U = g_1(X, Y), V = g_2(X, Y)$,对应于二重积分的坐标变换公式。

离散随机向量的变换

- 一维情形: $Z = g(X)$, $p_Z(z) = \sum_{g(x)=z} p(x)$.
- 二维情形: $p_Z(z) = \sum_{g(x,y)=z} p(x,y)$. 沿曲线$g(x,y) =$常数，求和。(等高线!)

EXAMPLE (泊松分布求和:例3)

设$X \sim \mathcal{P}(\lambda_1), Y \sim \mathcal{P}(\lambda_2)$, 则$Z = X + Y$服从$\mathcal{P}(\lambda_1 + \lambda_2)$.

注解: 两个到达过程的和即对应强度的相加的一个到达过程。

一般连续情形

- 一维情形:

Theorem (复合函数的密度函数)

设X的密度函数$f(x), Y = g(X)$;Y的取值范围为D, 对任意$y \in D$, 设其原像可以写成有限点(即$\sharp(g^{-1}(y))$有限).每一个可以写成一个逆函数$h_i(y) = x_i$; 则Y的密度函数为$f_Y(y) = \sum_{i=1}^{n} f(h_i(y))|h_i'(y)|$。

注记：一般地在每个单调区间内有一个逆函数。
解答作业:P121. 13

- 二维情形：$F_Z(z) = \int\int_{g(x,y)\leq z} f(x,y)dxdy$. 沿曲线$g(x,y) =$常数，求积分。
  特别地：边缘分布密度$Z_1 = X$, $f_{Z_1}(z) = \int_{-\infty}^{\infty} f(z,y)dy$.
  $Z_2 = Y$, $f_{Z_2}(z) = \int_{-\infty}^{\infty} f(x,z)dx$.
  一般的有***$F_Z(z) = \int_{-\infty}^{z} \int_{g(x,y)=u} f(x,y)dl$

随机变量的和与差

随机变量的和: $Z = X + Y$, 设(X,Y)联合分布密度为$f(x,y)$,
则$P(Z \leq z) = \int_{X+Y\leq z} f(x,y)dxdy$

- 公式1：$f_Z(z) = \int_{-\infty}^{\infty} f(x, z-x)dx$;
- 公式2：$f_Z(z) = \int_{x+y=z} f(x,y)dy$ 沿直线积分。
- 特别X, Y独立有$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$
  通常称等式右边为卷积$f_X(x) * f_Y(y)$.
- 例子: X, Y为独立的均匀分布，则$Z = X + Y$是三角分布。(参见信号处理).

EXAMPLE (约会问题: 随机变量的差)

设两个人约定八点见面，每人迟到的时间服从指数分布$\epsilon(\lambda)$. 设迟到时间相互独立，求$Z = X - Y$的分布函数。

答案: 分布密度函数$f_Z(z) = \frac{\lambda}{2}e^{-\lambda|z|}$，称为双边指数分布。

**随机变量的乘与除

设(X,Y)联合分布密度为$f(x,y)$,

- $Z = XY$, 则$f_{XY} = \int_{-\infty}^{\infty} \frac{1}{|x|} f(x, z/x)dx$
- $Z = Y/X$, 则$f_{Y/X} = \int_{-\infty}^{\infty} |x| f(x, xz)dx$
- 特别如果$Y > 0$,则$\log Y$是随机变量。常见有对数正态分布$\log Y = X \sim N(\mu, \sigma^2)$.
  以上乘除公式可以用加减法代替。

Remark (概念辨析)

一般的随机向量的有很多复合函数(从而导出很多随机变量)。上面随机变量的加减乘除是：我们把复合函数得到的随机变量看成(坐标)边缘随机变量的运算的结果。但知道边缘分布并不能知道复合函数得到的随机变量的分布函数!!!
应用中我们考虑很多随机变量的关系(及运算)，则假定不同随机变量的独立性是有用的，甚至必须的。

极大极小随机变量

Definition (随机变量排序)

已知$X_1, X_2, \dots, X_n$, 可以排序为$X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$,称$X_{(i)}, 1 \le i \le n$为次序统计量。

- 设X_i是独立同分布的；分布函数为$F(x)$;
- 最大(极大)值分布$F_{max}(x) = P(X_1 \le x, X_2 \le x, \dots, X_n \le x) = P(X_1 \le x)P(X_2 \le x)\dots P(X_n \le x) = F(x)^n$
- 最小(极小)值分布$F_{min}(x) = 1 - (1 - F(x))^n$.

EXAMPLE (灯泡现象)

教室的灯泡越多，则越容易坏。

解释： 假设每个灯泡服从指数分布$\mathcal{E}(\lambda)$. 共有n个灯泡，第一个坏掉灯泡的概率分布即极小分布. $F_{min}(x) = 1 - (1 - F(x))^n = 1 - e^{-n\lambda x}$. 显然$n$越大，$F_{min}(x)$越大，寿命越短。

作业

北航教材:
P120 习题四. 2,16,18,26

简单随机变量的和

$S = X_1 + X_2 + \dots + X_n$. 假设随机变量互相独立，
则$F(S) = F(X_1)F(X_2)\cdots F(X_n)$.
离散情形:

- X_i是两点分布$B(1,p)$，则S是二项分布$B(n,p)$.
- X_i是几何分布, 则S是负二项分布$NB(n,p)$.
- X_i是泊松分布$\mathcal{P}(\lambda_i)$, 则S是泊松分布$\mathcal{P}(\sum_i \lambda_i)$.

连续情形:

- X_i是均匀分布，则S是？.
- X_i是指数分布, 则S是伽马分布$\Gamma(n,\lambda)$.
- X_i是正态分布, 则S是正态分布$N(\sum \mu_i, \sum \sigma_i^2)$.

可以推广到随机变量的线性组合。特别正态分布的线性组合还是正态分布!
***可以推广到可数个随机变量的和。

例子1

EXAMPLE (债券价格)

假设$S(n)$是1万元债券在第n周的价格。一个流行的价格演化模型假定$S(n)/S(n-1)$是互相独立的对数正态分布$\log N(\mu, \sigma^2)$。假定$\mu = 0.0165, \sigma = 0.0730$, *(a)* 计算连续两周每周价格都上涨的概率。*(b)* 计算两周后价格上涨的概率。

解答:

- 设$Z \sim N(0,1), \log(S(n)/S(n-1)) = X \sim N(\mu, \sigma^2)$,则任一周内上涨的概率$P(S(n)/S(n-1) > 1) = P(X > 0) = P((X-\mu)/\sigma > -\mu/\sigma)$ 即$P(Z > -0.0165/0.0730) = P(Z < 0.2260) = 0.5894$. 由独立性，连续两周都上涨的概率$P = (0.5894^2) = 0.3474$
- 两周后上涨即$P(S(2)/S(0) > 1) = P(S(2)/s(1)s(1)/s(0) > 1) = P(X_1 + X_2 > 0)$. 注意$X_1, X_2 \sim X = N(\mu, \sigma^2)$, $X_1 + X_2 \sim N(2\mu, 2\sigma^2)$, 计算有$P(X_1 + X_2 > 0) = P(Z > -0.0330/(0.0730\sqrt{2})) = 0.6254$.

例子2

EXAMPLE (系统寿命)

假设一个系统由若干个独立的元件组成，包含并联，串联和备用元件等连接方式；则系统的寿命由元件的寿命和连接方式决定。设各个元件的寿命服从指数分布$\epsilon(\lambda)$.

基本连接方式:

- 串联: $S = \min(X, Y)$, $F_S = 1 - (1 - F(x))^2 = 1 - e^{-2\lambda s}$;
- 并联: $S = \max(X, Y)$, $F_S = F(x)^2 = (1 - e^{-\lambda s})^2$;
- 备用系统: $S = X + Y$, $f_S = \lambda e^{-\lambda s}\lambda s$. $S \sim \Gamma(2, \lambda)$.

一般可以一步步得到系统的寿命。

作业

北航教材:

P90 习题三. 29，

P120 习题四. 19,28,31.

QUIZ 小测验二

1. 设随机变量X在$(-\pi/2, \pi/2)$上服从均匀分布,则$Y = \tan X$的概率密度$f_Y(y) =$
2. 将红、白、黑三只球随机地逐个放入编号为1，2，3的三个盒内（每盒容纳球的个数不限），以X 表示有球盒子的最小号码，求：（1）随机变量X的分布律；（2）X的分布函数。
3. 某仪器上装有4只独立工作的同类元件。已知每只元件的寿命(以小时计) $X \sim N(5000, \sigma^2)$,当工作的元件不少于2只时,该仪器能正常工作。则该仪器能正常工作5000小时以上的概率为_。
4. ***P63. 23题:设随机变量X的概率密度为$f(x) = \frac{a}{e^x + e^{-x}}$, (1)确定常数$a$; (2)求X的分布函数; (3)求$P(0 \le X \le \ln\sqrt{3})$.

QUIZ 小测验二: 答案

1. $f_Y(y) = \frac{1}{\pi}\frac{1}{1+y^2}, -\infty < y < +\infty$.
2. （1）随机变量X的分布律;
   $P(X = 3) = 1/27; P(X = 2) = 7/27; P(X = 1) = 19/27$;
   （2）X的分布函数。
   $F(x) = 0, x < 1; 19/27, 1 \le x < 2; 26/27, 2 \le x < 3; 1, x \ge 3$
3. 一个元件正常工作概率$p = P(X > 5000) = 1/2$;
   元件工作个数Y服从$B(4, 0.5)$,
   仪器正常工作概率:
   $P(Y \ge 2) = 1 - P(Y = 0) - P(Y = 1) = 1 - 1/16 - 4/16 = 11/16$.
4. $a = 2/\pi, F(x) = 2/\pi \arctan e^x, P = 1/6$.

概率统计与随机过程

Probability, Statistics and Stochastic Process

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 Systems sciences, BUAA

November 12, 2010

Week 9：随机变量的数字特征

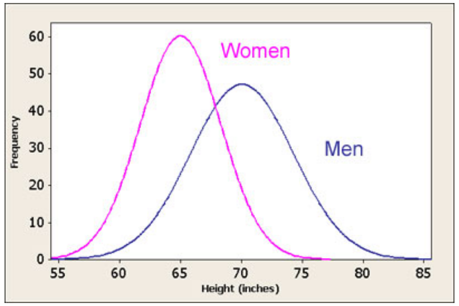
Review：回顾

RECALL:

- 随机向量及其分布
- 随机向量的函数及其分布
- 复合分布函数的计算: 特殊区域上的积分。

TODAY

- 期望及其计算;
- 复合函数的期望;



男女的身高分布如上：→ 简单的数字描述？
从函数到数字特征$EX, EX^2, EX^3, \ldots$

期望与方差

EXAMPLE (寿命预期)

已知一个中国人(50岁)继续生存的时间服从指数分布$\epsilon(1/25)$. 问一个50岁人的预期寿命是多少？大多数人的寿命范围？$\mu \pm s, 75 \pm 25$.

- 随机变量的期望: $EX = \sum x_i p_i$
  即平均值,盈利: 常记为μ;
  例子：Roulette美式轮盘赌有38个数值，最简单的玩法押一个数值一元，可赢得35元。问一次押注的期望赢钱多少？
- 随机变量的方差: $Var(X) = DX = \sum (x_i - EX)^2 p_i$
  即变异, 风险: 通常定义标准方差$s = \sqrt{DX}$,记数据为$\mu \pm s$.
  例子续：如果同上的豪华轮盘赌的最低押金是十元(可赢350元)，你希望参加吗？
- 还有其他特征吗？偏态skewness，峰态：kurtosis;

数学期望和矩

Definition (连续随机变量期望)

设连续随机变量X的密度函数为$f(x)$,且$\int_{i=-\infty}^{+\infty}|x|f(x)dx<+\infty$,称$EX=\int_{-\infty}^{+\infty}xf(x)dx$为$X$的期望(值).

- 从离散到连续: 级数求和到积分。
- 积分可能不存在或无穷;
- 推广: $EX^n=\int_{-\infty}^{+\infty}x^nf(x)dx$,称为$X$的$n$阶矩(moment)。
- 特别$E(X-EX)^n$称为X的n阶中心矩, 二阶中心矩即方差$DX=E(X-EX)^2=EX^2-(EX)^2$.
- 基本性质: (积分性质决定)单调性,线性复合。

***E又称为期望算子(线性):给定一个随机变量(分布函数)得到一个数。参见泛函分析。

一般期望的计算公式

Proposition (复合函数)

设$Y=g(X)$,如果X是离散随机变量, $EY=\sum_{i=0}^{\infty}g(x_i)p_i$; 如果$X$是连续随机变量, $EY=\int_{-\infty}^{+\infty}g(x)f(x)dx$.

- 离散情形证明: $\sum_j g(x_j)p(x_j)=\sum_i\sum_{g(x_j)=y_i}y_ip(x_j)=EY$.
- 连续情形证明: ***

Lemma (非负随机变量的期望)

设$Y\geq 0$,则$EY=\int_0^{\infty}P(Y>y)dy$.

Proof.

$Eg(X)=\int_0^{\infty}P(g(x)>y)dy=\int_0^{\infty}\int_{g(x)>y}f(x)dxdy$

交换积分$Eg(x)=\int_{x:g(x)>0}\int_0^{g(x)}dyf(x)dx=\int_{x:g(x)>0}g(x)f(x)dx$. □

随机向量的期望

Definition (二维随机向量的期望)

设随机向量为$\vec{X}=(X,Y)$,则$\vec{X}$的期望是(EX,EY), 记为$E\vec{X}$.类似有n维向量的期望。

Theorem (随机向量函数的期望)

设$\vec{X}$是随机向量, 如果为连续随机向量, 分布密度为$f(x_1,x_2,\ldots,x_n)$,则$Y=g(\vec{X})$的期望是$EY=\int\int g(\vec{X})f(x_1,\ldots,x_n)dx_1\ldots dx_n$. 对于离散随机向量类似。

推论:

- +++线性:$E(\sum a_iX_i)=\sum a_iE(X_i)$.
- +++(独立)乘法: 设X,Y独立, 则$EXY=EXEY$. 可推广到n维。

实例: 期望

EXAMPLE (淘宝店盈利估计)

设淘宝店销售一种商品。假设卖一件盈利a元, 没卖掉损失b元。设商品的需求量服从某个分布F, 问库存(订购)多少商品可以得到最大平均利润? 特别如果F是指数分布, 应该库存多少商品?

解答: 设Y是商品的需求, $P(Y\leq k)=F(k)$.
设库存量为k,则盈利随机变量Z满足;
$Z=aY-b(k-Y),Y<k,Z=ak,Y\geq k$
设$f(y)$为密度函数; 平均利润
即$m(k)=EZ=\int_0^k(ay-b(k-y))f(y)dy+\int_k^{\infty}akf(y)dy$
计算有$m(k)=(a+b)\int_0^k yf(y)dy+ak-(a+b)k\int_0^k f(y)dy$;
即求$m(k)$的极值点。计算有$m'(k)=a-(a+b)\int_0^k f(y)dy, m''(k)<0$,
极大值点$m'(k)=0$,有$F(k)=\frac{a}{a+b}$; 特别当F为指数分布, $1-e^{-\lambda k}=\frac{a}{a+b}$; $k=(\lambda)^{-1}\ln\frac{a+b}{b}$.

例子：复合函数的期望

Corollary (复合函数的方差)

- $VarX = DX = EX^2 - (EX)^2$;
- 线性变换: $Var(aX + b) = a^2 Var(X)$.
- 设X, Y独立,则$Var(X + Y) = VarX + VarY$.

EXAMPLE (配对问题)

n个信放在n个信封，求平均有几封正确放好。

解答： 设$X_i =$第i个信封放对。则$EX_i = 1/n$,
设Y为所有放对的信封
数，$Y = X_1 + \cdots + X_n, EY = E(X_1 + \cdots + X_n) = 1$.即平均一封信放对。

常见分布的期望与方差

常见离散分布：

- 两点分布，Bernoulli贝努利分布$EX = p, VarX = pq$
- 二项分布$B(N, p)$，$EX = np, VarX = npq$, 超几何分布类似；
- 泊松分布$\mathcal{P}(\lambda)$,$EX = \lambda, VarX = \lambda$
- 几何分布$EX = 1/p, VarX = q/p^2$;

常见连续分布：

- 均匀分布$\mathcal{U}(a, , b)$,$EX = \frac{a+b}{2}, VarX = \frac{(b-a)^2}{12}$;
- 指数分布$\varepsilon(\lambda)$, $EX = 1/\lambda, VarX = 1/\lambda^2$;
  Γ分布可推出:
- 正态分布: $EX = \mu, VarX = \sigma^2$.
  独立正态分布的线性组合还是正态分布(对应有期望,方差的线性组合);

作业

北航教材:
P157 习题五. 6,9,13,15

协方差

Definition (协方差)

设随机变量X, Y的期望存在,定义X, Y的协方差
为$cov(X, Y) = E(X - EX)(Y - EY)$. 可记为$\sigma_{XY}$.
计算公式: $cov(X, Y) = E(XY) - EXEY$

- 对称性：$cov(X, Y) = cov(Y, X)$,特别$Cov(X, X) = Var(X)$
- (双)线性：$cov(aX, Y) = acov(X, Y)$
- 求和：$cov(\sum X_i, \sum Y_j) = \sum_i \sum_j cov(X_i, Y_j)$ 特别$Var(\sum X_i) = Var(X_i) + \sum\sum_{i \neq j} cov(X_i, X_j)$

推广：n维随机向量期望$E(\vec{X}) = \vec{\mu}$,
协方差矩阵: $\Sigma = (\sigma_{X_i X_j})_{n \times n}$, $\Sigma = E(\vec{X} - \vec{\mu})^{\tau}(\vec{X} - \vec{\mu})$.
高维正态分布$f(\vec{X}) = Cexp^{-\frac{1}{2}(\vec{X}-\vec{\mu})^{\tau}\Sigma^{-1}(\vec{X}-\vec{\mu})}$

相关系数

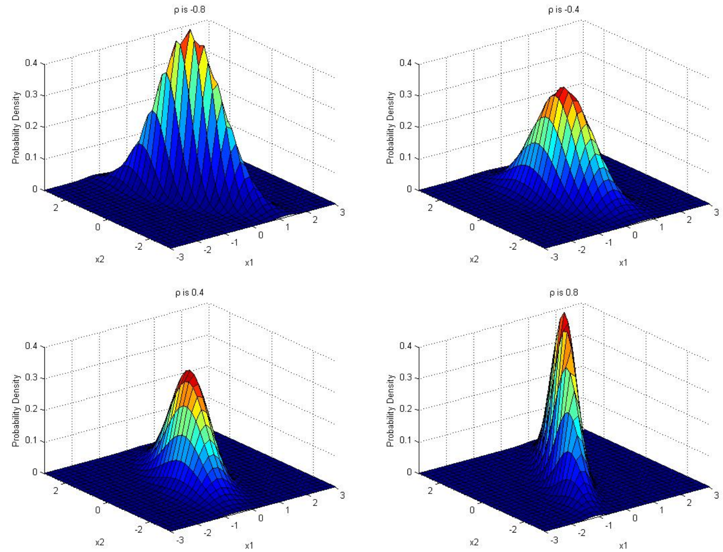
Definition (相关系数)

设随机变量X,Y的期望和方差存在,定义X,Y的相关系数为$\rho(X,Y)=\frac{cov(X,Y)}{\sqrt{VarXVarY}}$. 可记为$\rho_{XY}$.

计算公式: 引入标准化随机变量
$U_X=(X-\mu_X)/\sigma_X, U_Y=(Y-\mu_Y)/\sigma_Y$, 则$\rho(X,Y)=E(U_XU_Y)$

- 相关系数: $-1\le\rho(X,Y)\le 1$,即柯西许瓦茨不等式。
- 如果$\rho(X,Y)=0$,称X,Y互不相关; 一般地如果$cov(X,Y)=0$,称X,Y互不相关。
  独立必然不相关,反之不一定. P152 例5.
- $|\rho|=1\iff P(Y=aX+b)=1$. 注: $VarX=0\to X\equiv Const$.
  相关指线性相关: 不相关不一定没关系,更不一定独立。
  例子: $X\sim N(0,1), Y=X^2$; $cov(X,Y)=0$.
- 二维正态分布$N(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$,$\rho$即相关系数,
  特别边缘随机变量: 不相关等价于独立。

例子: 二维正态分布



矩母函数***

Definition (矩母函数)

设随机变量X的期望存在,定义X的矩母函数为$M_X(s)=E(exp^{sX})$. 可记为$M(s)$.

- 离散情形: $M(s)=\sum e^{sx_i}p(x_i)$.
- 连续情形: $M(s)=\int_{-\infty}^{\infty}e^{sx}f(x)dx$
- 对应有: Z变换: $M(z)=\sum z^ip(i)$.
  拉普拉斯变换$L(f)=\int_0^{\infty}e^{-sx}f(x)dx$.
- 结论: 随机变量的矩即矩母函数在零点的导数。
  $EX^n=\frac{d^nM(s)}{ds^n}|_{s=0}$.
  注: 存在逆变换, 唯一决定随机变量的分布函数。

例子: 指数分布$M(s)=\frac{\lambda}{\lambda-s}, s<\lambda$.
$EX=\frac{1}{\lambda}, DX=\frac{1}{\lambda^2}$.

作业

北航教材:
P157 习题五. 17,18,22,23

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

November 19, 2010

Week 10：极限定理与统计

1. 不等式与大数定律
   - 矩不等式
   - 弱大数定律
   - 例子与应用
2. 中心极限定理与统计思想
   - 强大数定律*
   - 中心极限定理
   - 统计思想

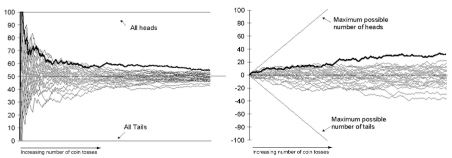
Review：回顾

Last Week:

- 随机变量→ 数字特征(矩)
- 常见特征: 期望，方差，协方差，相关系数；
- ***矩母函数:

This Week:

- (矩)不等式
- 弱大数定律
- 中心极限定理
- 统计思想:



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$P(X > 75) \le 50/75 = 2/3$;
2. 如果$VarX = \sigma^2 = 25$,则$P(|X - 50| < 10) = 1 - P(|X - 50| \ge 10) = 1 - 25/100 = 3/4$.

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$P(|\bar{X}_n - \mu| \ge \epsilon) \to 0$.

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

Definition (依概率收敛)

设X_n, X是随机变量,如果对任意$\epsilon > 0$, 有$\lim_{n\to\infty} P(|X_n - X| \ge \epsilon) = 0$, 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$P(|\bar{X}_n - p| \ge 0.1) < p(1-p)/(n*0.01) < 1/4(n*0.01)$.
如果$n = 100$,误差大于0.1的概率小于0.25
2)同上, $P(|\bar{X}_n - p| \ge 0.01) < 1/(4n*0.01^2) < 0.05$, 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$P(|X - \mu| > k) \le 0/k = 0$;
即$P(X \ne \mu) = 0$。

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$\lim P(X_n \leq x) = F_X(x), \lim F_{X_n}(x) = F_X(x)$;

作业

北航教材:
P157 习题五. 7, 24;
P166 习题六. 1, 2

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。

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:

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);

作业

北航教材:
P166 习题六. 3，4，5.
P184 习题七1.

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

November 26, 2010

Week 11：抽样与点估计

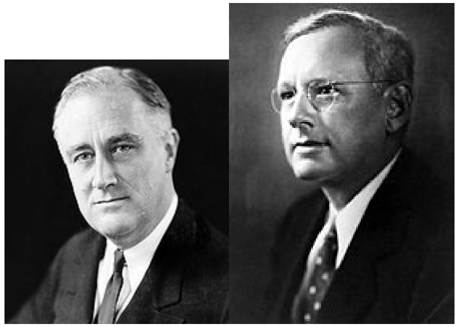
Review：回顾

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
- 参数估计: 矩估计与最大似然估计；



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:$F_n(x) = k/n, x_{(k)} \le x < x_{(k+1)}$.

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。

作业

北航教材:
P184 习题七. 3，4，5，6

QUIZ 小测验三

1. 设随机变量$X \sim N(\mu, \sigma^2)$,则$E(X-\mu)^4 =$. (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. 设随机变量X, Y的二阶矩EX^2, EY^2存在，证明：成立不等式$|EXY| \leq \sqrt{EX^2EY^2}$. (构造$E(X+tY)^2 \geq 0$)
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$\lim_{n\to\infty} P(|Y_n| \geq \epsilon) = 0$。

常见统计量的分布

- 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$\bar{X}_n = \frac{1}{n}\sum X_i, S^2 = \frac{1}{n-1}\sum(X_i-\bar{X}_n)^2$.
- 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，P181定理五.
- *(两个总体的方差比) 两个独立总体X, Y,
$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n-1, m-1), n, m \geq 2$.

点估计

假定总体分布X的模型已知但参数未知。

Definition (参数统计量statistics)

给定简单随机样本$X_1, X_2, \ldots, X_n$, 设θ是总体分布的未知参数，构造函数$g(x_1, x_2, \ldots, x_n)$, 称$\hat{\theta} = g$为θ的估计量或统计量;

常见估计:
- 样本均值$\bar{X}_n = \sum X_i \to EX$ 期望
- 样本方差$S^2 = \sum(X_i-\bar{X})^2/(n-1) \to DX = EX^2-(EX)^2$ 方差, 样本标准差$s \to \sqrt{DX}$ 标准差
- ***样本相关系数$\hat{\rho_{xy}} = \frac{S_{xy}}{s_x s_y} \to \rho$.

Definition (矩估计moment statistics)

给定简单随机样本$X_1, X_2, \ldots, X_n$, 如果$v_k = EX^k$存在，则定义$\hat{V}_k = \frac{1}{n}\sum_i X_i^k$是$v_k$的矩估计统计量。更广义的: 构造函数$g(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_n)$,去估计未知参数$\theta$，称为$\theta$的矩估计量。

矩估计的例子

EXAMPLE (正态分布)

$X \sim N(\mu, \sigma^2)$,参数$\mu = EX, \sigma^2 = DX = EX^2 - (EX)^2$,
则矩估计$\hat{\mu} = \hat{v}_1 = \bar{X}_n$, $\hat{\sigma^2} = \hat{v}_2 - \hat{v}_1{}^2 = \frac{1}{n}\sum(X_j - \hat{\mu})^2$.
任意分布的均值与方差的矩估计的表达式是一样的。(注：与教材说法不一样，关键在矩的定义中通常系数为$1/n$)).

EXAMPLE (均匀分布)

$X \sim U(0, b)$, 则$b = 2EX$, 矩估计$\hat{b} = 2\bar{X}_n$.
一般的$U(a, b)$有两个参数，利用$EX = (a + b)/2, DX = (b - a)^2/12$,解方程可得$a, b = v_1 \pm \sqrt{3(v_2 - v_1^2)}$.

- 指数分布$\epsilon(\lambda)$， $EX = 1/\lambda \to \hat{\lambda} = 1/\bar{X}_n$;
- 泊松分布,$EX = \lambda \to \hat{\lambda} = \bar{X}_n$, 注：$DX = \lambda, \hat{\lambda}_2 = \frac{1}{n}\sum(X_j - \hat{\mu})^2$???
- 二项分布$B(N, p)$, $EX = Np \to \hat{p} = \bar{X}_n/N$; 特别社会调查问题：比例的估计$B(1, p)$, $\hat{p} = \bar{X}_n$

最大似然估计Maximum Likelihood Estimation

EXAMPLE (选择判断)

设某一天下雨的概率p为0.1或0.9，今天没有下雨，问$p = ?$
则应该判断$p = 0.1$；由结果判断概率的思想称为最大似然思想。如果连续几天没下雨，判断更准确!

一般存在一个似然函数(likelihood function)，已知样本值，求使得似然函数取极大值所对应解来估计参数。

Definition (MLE)

给定关于未知参数θ和样本值的似然函数$L(x_1, x_2, \ldots, x_n; \theta)$, 定义$\hat{\theta}$为使得$L$取最大值的解，称$\hat{\theta}$是$\theta$的最大似然估计.

- 离散情形：$L(x_1, x_2, \ldots, x_n; \theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$
- 连续情形：$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1, x_2, \ldots, x_n; \theta)$;
- 求解最大值：$\frac{dL}{d\theta} = 0$, 或者$\frac{d\ln L}{d\theta} = 0$. (对数似然函数).
- 多维情形$(\theta_1, \theta_2, \ldots)$: $\frac{\partial L}{\partial \theta_i} = 0, \forall i$. 现实中可能要用数值解法求解。

MLE的例子

回顾:(估计某地区的某动物总数N)
捕捉一批动物M,做标记放回,过一段时间,再捕捉一批动物n, 其中有标记的为k,利用超几何分布估计N。概率为$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$是关于$N$的一个序列，求最大值对应的$N$,有$N = nM/k$. 称为最大似然估计。

EXAMPLE (正态分布)

$X \sim N(\mu, \sigma^2)$, 计算有$\mu = \bar{X}_n$, $\hat{\sigma^2} = \frac{1}{n}\sum(X_j - \hat{\mu})^2$,与矩估计一致。

EXAMPLE (均匀分布)

$X \sim U(0, b)$, 则MLE估计$\hat{b} = x_{(n)}$,即最大样本值。对于$U(a, b)$,a对应最小样本值。$L(a, b) = \frac{1}{(b-a)^n} \le \frac{1}{(x_{(n)} - x_{(1)})^n}$.

- 指数分布$\epsilon(\lambda)$，MLE $\hat{\lambda} = 1/\bar{X}_n$ 与矩估计一致
- 泊松分布,$\hat{\lambda} = \bar{X}_n$,与矩估计一致
- 两点分布$B(1, p)$, $\hat{p} = \bar{X}_n$，与矩估计一致。

作业

北航教材:
P207 习题八1.2.3.4

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

December 3, 2010

Week 12：点估计的评估与区间估计

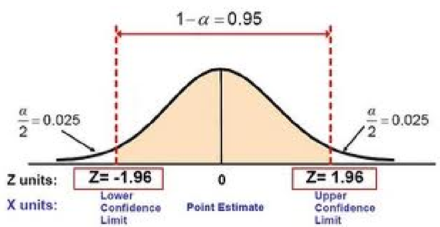
Review：回顾

Last Week:

- 统计与抽样
- 统计学三大分布
- 参数估计: 矩估计与最大似然估计；

This Week:

- 估计(统计量)的评价标准：无偏, 有效，一致。
- 区间估计与置信度；
- 均值与方差的区间估计；



置信区间：给点估计一个误差范围。
置信水平: 0.05,0.01.

无偏估计

Definition (参数统计量statistics)

给定简单随机样本$X_1, X_2, \ldots, X_n$, 设θ是总体分布的未知参数，构造函数$g(X_1, X_2, \ldots, X_n)$, 称$\hat{\theta} = g$为θ的估计量或统计量；

Definition (无偏统计量Unbiased)

如果以上统计量满足$E(\hat{\theta}) = \theta$, 称$\hat{\theta}$是无偏估计。

Proposition

样本均值$\bar{X}_n$和样本方差$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2$是总体分布的均值和方差的无偏估计。

直接验证: $E\bar{X}_n = \mu, E(S^2) = \sigma^2$。
注：方差的矩估计(或MLE估计)$\frac{1}{n}\sum(X_i - \bar{X}_n)^2$不是无偏估计。

无偏估计的例子

Theorem (总体矩的估计)

样本的k次矩是总体的k次矩的无偏估计。
即$\hat{V}_k = \frac{1}{n}\sum_i X_i^k$是$v_k = EX^k$的无偏估计。

注: $E(X_i^k) = v_k$.

- 方差的矩估计(或MLE估计)是有偏估计(n充分大时，相差无几.)
  *** 渐进无偏估计$\lim_n E(\hat{\theta_n}) = \theta$.
  *** 样本标准差s是有偏估计。
- 均匀分布的矩估计: $\hat{b} = 2\bar{X}_n$是无偏估计, MLE估计$Ex_{(n)} = n/(n+1)b$不是。
- 无偏估计也有很多. 比如任意$\hat{\theta} = \sum a_i X_i, \sum a_i = 1$都是期望的无偏估计。

最小方差无偏估计MVU

Definition (有效性)

如果同一个参数θ的两个统计量满足$D(\hat{\theta}_1) \le D(\hat{\theta}_2)$, 称$\hat{\theta}_1$比$\hat{\theta}_2$更有效。

Definition (最小方差无偏估计MVU)

给定无偏估计$\hat{\theta}_0$, 如果任一无偏估计$\hat{\theta}$满足，$D(\hat{\theta}_0) \le D(\hat{\theta})$, 称$\hat{\theta}_0$是最小方差无偏估计。

- 样本均值是最小方差无偏估计MVU。(验证：线性估计中的方差最小, 柯西许瓦茨不等式).
- 均匀分布的MLE估计$D\hat{b} = \frac{n}{(n+2)(n+1)^2}b^2$比矩估计$D\hat{b} = b^2/(3n)$更有效。
- *最小方差无偏估计MVU是统计量的最理想估计，有相关的存在性定理。
- ***信号处理: 维纳滤波器是线性最小方差估计器。

一致性或相合性

估计的最基本性质：收敛性。

Definition (相合性)

给定统计量$\hat{\theta} = g(X_1, X_2, \ldots, X_n)$与样本容量$n$有关，可记为$\hat{\theta}_n$, 如果$\hat{\theta}_n$依概率收敛到$\theta$(即$\hat{\theta}_n \xrightarrow{P} \theta$), 称$\hat{\theta}_n$是$\theta$的相合估计量(或一致估计)。

Theorem (矩估计相合性)

样本矩依概率收敛到随机变量的矩。推广的
有$g(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k) \xrightarrow{P} g(v_1, v_2, \ldots, v_k)$.
所以所有矩估计都是相合估计。

- ***事实上矩估计以概率1收敛. 通常称为强相合估计。
- 样本方差是相合估计。$S^2 \xrightarrow{P} \sigma^2$. 特别情形(正态分布) P196例4.
- ***一般的MLE估计在一定条件下也是相合估计。

例子

EXAMPLE

设总体分布是指数分布$\epsilon(\lambda)$. 记参数$\theta = 1/\lambda$. 则
(1) $\bar{X}_n$是θ的无偏估计; *(2)* $Z = \min(X_1, X_2, \ldots, X_n)$,则$nZ$是$\theta$的无偏估计; *(3)* $\bar{X}_n$比nZ更有效(方差更小).

解答:.

(1),(2)直接验证
(3) $D(\bar{X}_n) = \theta^2/n$, $D(nZ) = \theta^2$. □

作业

北航教材:
P207 习题八6.7.8.9

置信区间

Definition (置信区间)

假定给出两个估计$\hat{\theta}_1, \hat{\theta}_2$, 使得给定$\alpha$,满足$P(\hat{\theta}_1 \le \theta \le \hat{\theta}_2) \ge 1-\alpha$,
称$[\hat{\theta}_1, \hat{\theta}_2]$为$\theta$的置信水平为$1-\alpha$的置信区间。

- 注意$\hat{\theta}_1, \hat{\theta}_2$是随机变量(区间是随机区间!),$\theta$是固定未知参数。
  $P(\hat{\theta}_1 \le \theta \le \hat{\theta}_2) = P(\hat{\theta}_1 \le \theta \bigcap \theta \le \hat{\theta}_2)$, 涉及两个随机变量的分布。
- 通常我们给出点估计的上下误差范围来代替两个估计,
  比如差值区间$[\hat{\theta}-a, \hat{\theta}+b]$ 或比例区间$[\hat{\theta}/a, \hat{\theta}*b]$, $a, b \ge 1$.
  特别如果$a=b$,记为$\hat{\theta} \pm b$, b又称为抽样误差。
- 选取区间满足概率等式: $P(-b \le \hat{\theta}-\theta \le a) = 1-\alpha$,
  已知$\hat{\theta}$的分布,则置信水平$1-\alpha$决定a, b, 从而置信区间的大小。
- 置信区间的频率解释: θ不一定在区间内! 但重复抽样得到的置信区间中有概率$(1-\alpha)$个区间包含θ.

置信水平与下分位数

Remark (由估计$\hat{\theta}$, 怎样计算抽样误差a, b?)

方法: 即构造概率区间满足: $P(-b \le \hat{\theta}-\theta \le a) = 1-\alpha$.
设已知统计量$\hat{\theta}-\theta$的分布$F(x)$(其他统计量参见均值区间估计);
则记下分位数$F(x_{\alpha/2}) = \alpha/2, F(x_{1-\alpha/2}) = 1-\alpha/2$,
取$-b = x_{\alpha/2}, a = x_{1-\alpha/2}$, 即可。

我们称$[\hat{\theta}-x_{1-\alpha/2}, \hat{\theta}-x_{\alpha/2}]$是置信水平为$1-\alpha$关于$\theta$的置信区间的.
注意: 我们取的区间使得$\hat{\theta}$在$[\theta-b, \theta+a]$两边的概率一样$(\alpha/2)$。也可以不一样。

- 应用中: 如果分布是关于0点对称的,则$-x_{\alpha/2} = x_{1-\alpha/2}$,
  即区间为$[\hat{\theta}-x_{1-\alpha/2}, \hat{\theta}+x_{1-\alpha/2}]$。称$x_{1-\alpha/2}$是临界值。
- 一般的分布取标准正态分布或t分布。置信水平为0.95,0.99.
  设$\alpha=0.05$,则$Z_{1-0.025}=1.96$, $\alpha=0.01$,则$Z_{1-0.005}=2.58$.
- 类似可定义: 单侧置信区间估计: $P(\hat{\theta} \ge \theta) = 1-\alpha$
  或$P(\hat{\theta} \le \theta) = 1-\alpha$. 对应区间为$[-\infty, \hat{\theta}]$或$[\hat{\theta}, +\infty]$.

案例分析一

EXAMPLE (测量误差)

我们想得到鲜牛奶的冰点μ。假设温度计的测量值服从正态分布$N(\mu, \sigma^2)$. 用一个测量误差为$e=0.0048$的温度计测量21次(数据略). 计算有样本均值$\bar{X}_n = -0.546$,样本标准差$S=0.005$.

- 可以给出估计$\hat{\mu} = \bar{X}_n = -0.546$,
  $\hat{\sigma^2} = S^2 = 0.005^2$;
- 假设已知$\sigma = e = 0.0048$,则μ的置信水平为95%的置信区间为
  $[\bar{X}_n - 1.96\sigma/\sqrt{n}, \bar{X}_n + 1.96\sigma/\sqrt{n}]$;
- 假设σ未知,则μ的置信水平为95%的置信区间为$[\bar{X}_n - 2.086S/\sqrt{n}, \bar{X}_n + 2.086S/\sqrt{n}]$;
- 方差σ^2的置信水平为95%的置信区间为
  $[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}] \approx [0.0038, 0.0072]$;

对应的统计问题:

- 点估计
- (已知方差)大样本下均值的区间估计;
- (未知方差)小样本下均值的区间估计;
- 方差的区间估计;

均值估计:z,t分位数

设总体分布是正态分布$N(\mu,\sigma^2)$, 怎样给出μ的一个区间估计?
方差未知，®假定$\sigma=0.0048$ ®

1. 点估计$\hat{\mu}=\bar{X}_n=-0.546,\hat{\sigma}^2=S^2=0.005^2$;
2. 统计量分布：$\bar{X}_n\sim N(\mu,\sigma^2/n)$; (一般情形由中心极限定理逼近)
3. 设定95%的置信度,选对称概率区间,查分位数表$Z_{1-0.025}=1.96$。
   概率公式$P(|\bar{X}_n-\mu|\le b)=P(\frac{|\bar{X}_n-\mu|}{\sigma\sqrt{n}}\le z_{1-0.025})=0.95$
4. $b=1.96\sigma/\sqrt{n}$,求解出置信区间上下限$\pm b=0.002$。
   $\mu\in[-0.546-0.002,-0.546+0.002]$.

一般情况方差未知，怎么办? 用样本方差。

- 统计量分布：®假设总体是正态分布®,则$\frac{\bar{X}_n-\mu}{s/\sqrt{n}}\sim t(n-1)$.
- 置信水平$1-\alpha=0.95$,取对称区间$P(\frac{|\bar{X}_n-\mu|}{s\sqrt{n}}\le t_{1-\alpha/2}(n-1))=0.95$
  查表$t_{1-0.025}(20)=2.086$.
- 求解b得到置信区间为$[-0.5483,-0.5437]$.

说明：以上t分位数与z分位数的估计差不多(稍大),因为$n\ge 33$时与正态分布相近；一般用于小样本估计。

方差估计：χ^2分位数

温度计给出的测量误差$e=0.0048$是否准确? 用样本方差去估计。

1. 点估计: $S^2=0.005^2$;
2. 统计量分布：总体是正态分布，样本方差服从χ^2分布，$(n-1)S^2/\sigma^2\sim\chi^2(n-1)$
3. 设置信水平$1-\alpha=0.95$,选择概率区间
   $P(\chi^2_{\alpha/2}(n-1)\le\frac{(n-1)s^2}{\sigma^2}\le\chi^2_{1-\alpha/2}(n-1))=0.95$
   查表$\chi^2_{0.025}(20)=9.591,\chi^2_{1-0.025}(20)=34.170$,
4. 求解区间上下限, 得到σ^2的区间估计$[0.0038^2,0.0072^2]$.

相关说明：

- 可以得到标准差σ的区间估计$[0.0038,0.0072]$. 注意其关于$s=0.005$的非对称性。
- 由于χ^2非对称性，区间的选择有很多种，我们的选择仅仅比较容易计算，不一定是最短的区间。

其他区间估计问题***

EXAMPLE (总体比例估计问题：早餐调查。)

设大学生吃早餐的比例p。抽样n个人, 其中m个人吃早餐。
点估计$p=\bar{X}_n$, 总体分布为二项分布，n充分大，用正态分布逼近，
得区间估计: $\bar{X}_n\pm z_{1-\alpha/2}\sqrt{\bar{X}_n(1-\bar{X}_n)/n}$.
区间大小与n：设抽样误差d, $d=0.03,0.02,0.01$; $n=1068,2401,9604$.

非正态总体的区间估计：用正态总体逼近计算:

- 均值估计: 方差已知，有中心极限定理，
- 均值估计: 方差未知: 小样本情形t分布, 可用正态分布逼近($n\ge 30$).
- 方差估计: 非正态总体没法估计，有其他统计方法；χ^2检验，ANOVA方差分析.

两个(多个)正态总体的比较估计：

- 两个总体的均值差：服从正态分布或某种t分布，依赖与两个总体的方差信息。
- 两个总体的方差比：服从$F(m,n)$分布。

作业

北航教材:
P207 习题八10.11.13.14

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

December 10, 2010

Week 13：假设检验

1. 假设检验的原理
   - 实际推断原理
   - 检验错误:显著性水平
   - 假设检验的问题
2. 正态总体的参数检验
   - 均值检验
   - 方差检验
   - 其他统计方法

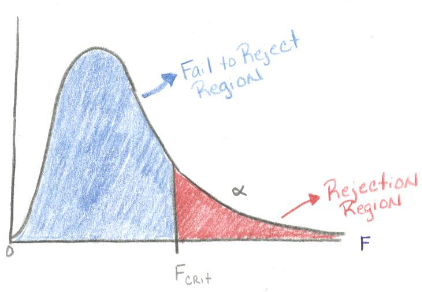
Review：回顾

Last Week:

- 估计(统计量)的评价标准：无偏，有效，一致。
- 区间估计与置信度；
- 均值与方差的区间估计；

This Week:

- 假设检验的原理
- 一个正态总体的参数检验
- 其他统计方法；



区间估计→ 假设检验
计算题→ 判断题(或选择题)

简单统计决策的过程:实际推断原理

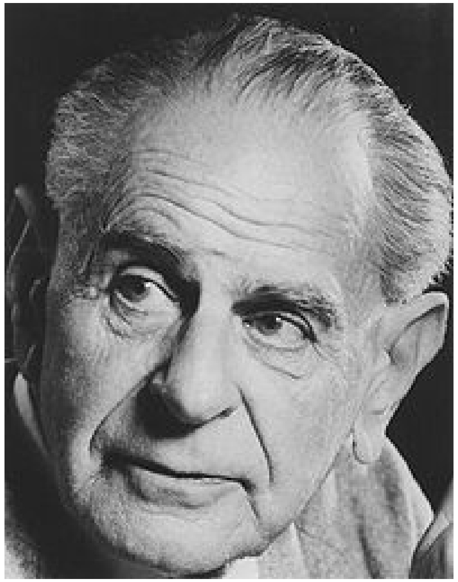
用一个样本统计量的值: "选择判断" 关于未知信息的可能.

EXAMPLE (简单假设检验的过程)

给出H_0, H_1两个假设$\xrightarrow{sampling} X_1, X_2, \ldots, X_n \xrightarrow{T} T$统计量的分布$P_{H_0}(T), P_{H_1}(T) \xrightarrow{decision}$ 利用$\hat{T}$ 选择H_0或H_1.

- 实际推断原理: 如果一个很小概率事件(显著的意外)发生了，说明假设存在不合理的。
- 假设检验与反证法:
- 例子：抽样调查: p是吃早餐的比例可能为$0.4, 0.5, 0.6$. 调查100人有60人吃早餐。问p更可能是什么？"肯定"不是0.4, 但不一定是0.6, 0.5.

Karl Popper: 批判理性主义



卡尔·雷蒙德·波普尔爵士
(1902年7月28日－1994年9月17日)
二十世纪最伟大科学哲学家；
科学研究→ 个人选择→ 社会
证伪主义→ 试错→ 开放社会。
In search of a better world.
"通过知识获得解放".

决策错误与显著性水平

给定两个假设H_0, H_1, 简单随机样本和检验统计量$T = g(X_1, X_2, \ldots, X_n)$.

Definition (拒绝域)

指定区域W, 假定H_0成立，如果统计量的值$t = g_{H_0}(x_0, x_1, \ldots, x_n) \in W$, 则推断$H_0$不成立($H_1$成立)。称$W$为拒绝域。

- 一般称H_0原假设或零假设，H_1为备择假设；希望证伪原假设!
- 通常区域写成$W = \{t : |g(X_1, X_2, \ldots X_n)| > t_0\}$. t_0称为临界值。
- 临界值由事先确定的事件的概率大小决定。即$P_{H_0}(t > t_0) = \alpha$, 称为(统计)显著性水平。通常取为$0.05, 001, 0.1$.

Definition (简单决策的两类错误)

如果真实情况下，原假设是正确的，我们抽样得到的结论否定原假设，称为第一类错误；如果真实情况下，原假设是错误的，我们抽样得到的结论不拒绝原假设，称为第二类错误；记第一类错误的概率为α,第一类错误的概率为β; 显著性水平即第一类错误的概率。

例子: 硬币是否是均匀的?

EXAMPLE (硬币的公平性)

给定一个硬币，问其是否是公平的？即投硬币为正面的概率为0.5.
设投硬币100次，其中31次为正面.

解答:

- 假设$H_0 : p = 0.5$ vs $H_1 : p \neq 0.5$
- 定义$T = \frac{1}{n}\sum X_i$, $X_i \sim B(1, p)$. 则$nT \sim B(n, p)$.
- 设显著水平$\alpha = 0.05$,则拒绝域: $P(|T - p| > t_0) = \alpha = 0.05$.
  利用$Z = \frac{T-p}{\sigma/\sqrt{n}} \sim N(0, 1)$, $z_{0.975} = 1.96$. 和$p = 0.5$!!!
  $|\frac{T-p}{\sigma/\sqrt{n}}| > 1.96$, 计算有$|T - p| > 1.96\sigma/\sqrt{n}$, $t_0 = 0.098$.
- 判断: $|T - p| = 0.19 > t_0$，所以原假设否定，(统计显著意义下) 硬币不是公平的。

说明：如果$T = 0.41$呢？不知道。显著性水平$0.1, 0.01$? P值为0.07.

假设检验的类型

常见假设的类型：

- 参数检验: $H_0 : \theta \in D$ vs. $H_1 : \theta \notin D$,D是一个集合。
- 非参数检验: $H_0 : F = F(x)$ vs. $H_1 : F \neq F(x)$,$F(x)$是一个分布函数。

参数检验的类型

- 双边参数检验: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.
- 单边检验；
  左边检验$H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$.
  右边检验$H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$.

***假设检验的若干问题

1. 错误是不可避免的。给出样本量大小，缩小第一类错误α，会加大第二类错误β. 我们通常控制第一类错误的大小。
2. 假设的选择：H_0选为历史数据给定或保守选择。仅仅当有充分证据否定原假设，才采用备择假设,这样错误的代价较小。
3. 结论：如果小概率事件发生,可以否定原假设. 如果小概率事件没有发生,不能肯定原假设.只能说明没有充分证据，需要进一步考察!
4. 显著性水平：$\alpha = 0.05$,或0.06有区别吗？必须考察实际应用的背景。

应用的谬误:

- 统计显著性与实际显著性:
  样本量充分大后，一般会得到统计显著性，拒绝原假设。
  统计显著性必须与实际问题相结合考虑问题。比如差异是否重要(经济考虑)。统计显著性仅仅是第一步，我们需要进一步发现为什么有差异？
- 不能对一个数据做多次假设检验。比如做二十次检验，很可能有一次的检验结果是显著的。

作业

北航教材:
P236 习题九1.
补充：根据以下历史数据(正面数/总次数)，判断其硬币是否公平$p = 0.5$。$\alpha = 0.1, 0.05, 0.01$.

- De Morgan 德摩根：1061/2048
- 蒲丰Buffon：2048/4040
- John Kerrich: 5067/10000
- 费勒Feller: 4979/10000;
- 罗曼洛夫斯基40173/80664;

惊人的预测：统计思考

网络骗局:
一封电子邮件: 存在高级统计方法以0.95概率预测(足球或篮球)比赛结果。
第一封预测：成功!
第二封预测：成功!
第三封预测：成功!
第四封预测：成功!
第五封预测：成功!
第六封信：请汇款200元买我们的方法。

- 概率思考：五次成功的概率$0.5^5 \sim 0.0313$;
  可能有内幕？值得买!!!
- 统计思考：骗子发8000封邮件，一半猜A赢，一半猜B赢；
  第二次：仅发新预测给4000个得到正确结果的人；
  第三次：发给2000人；第四次：发给1000人；
  第五次：发给500人。最后要钱的电邮发给250人即可!

实际推断原理：仅仅是一次的判断。
但大量试验则小概率事件必然发生。

案例：质量检验

EXAMPLE (case study:白糖的质量检验)

厂家：希望生产的每包白糖的质量服从正态分布$N(500, 0.8)$.
商家：希望得到的白糖不要短斤少两。每包白糖约为500克。
抽样9包，计算有$\bar{X}_n = 499.412, s = 0.676$。

- 厂家：假设检验机器正常工作。
- 均值检验：$H_0 : \mu = 500 vs H_1 : \mu \neq 500$
  正态检验结果：机器不正常。$\alpha = 0.05$
- 方差检验：$H_0 : \sigma^2 = 0.8 vs H_1 : \sigma^2 \neq 0.8$
  或单边方差检验：$H_0 : \sigma^2 \leq 0.8 vs H_1 : \sigma^2 > 0.8$
  χ^2检验结果：机器是否正常不确定。
- 商家：均值检验：$H_0 : \mu = 500 vs H_1 : \mu \neq 500$
  t检验结果：质量不是500克。$\alpha = 0.05$
- 单边均值检验：$H_0 : \mu \geq 500 vs H_1 : \mu < 500$
  t检验结果：质量不足500克。$\alpha = 0.05$

均值检验：正态检验

厂家: 假设检验机器正常工作。

- 假设$H_0: \mu = 500 vs H_1: \mu \neq 500$
- ®假设总体是正态分布$N(500, 0.8)$ ®, 检验统计量$Z = \frac{\bar{X}_n - 500}{\sqrt{0.8/n}} \sim N(0,1)$
- 如果H_0成立，$|Z|$应该比较小，反之应该比较大；特别取$\alpha = 0.05$，使得$P(|Z| > z_{\alpha/2}) = \alpha = 0.05$, 即$|Z| > 1.96$时为小概率事件.
- 代入数据，$|Z| = |(499.412 - 500)/\sqrt{0.8/9}| = 1.97 > 1.96$, 小概率事件发生了，应该否定原假设，即$\mu \neq 500$.

说明:

- $\{|Z| > z_{\alpha/2}\}$称为拒绝域，检验统计量位于其中则拒绝原假设。如果该事件发生，称检验为显著的。使用正态分布决定拒绝域，称为正态检验。
- 现代统计软件可以计算$|Z| \geq 1.97$对应的概率$P = 0.0488$。由于$P < 0.05$否定原假设，称为P值检验。

小样本均值检验：t检验

商家不可能知道方差，要用样本方差，对应有t检验

- 假设$H_0: \mu = 500 vs H_1: \mu \neq 500$
- ®假设总体是正态分布®，则样本方差服从χ^2分布；则检验统计量$T = \frac{\bar{X}_n - \mu}{s\sqrt{n}} \sim t(n-1)$.
- 如果H_0成立，$|T|$应该比较小，反之应该比较大；
- 类似查表有$t_{0.025}(8) = 2.306$, 计算$|T| = 2.609$, 处于拒绝域，否定原假设。
- 用P检验，$P(|T| > 2.609) = 2P(T > 2.609) = 2 * 0.01559 == 0.03118 < 0.05$,否定原假设。

商家可以有理由拒绝该批白糖吗？
其实$\mu > 500$是可以接受的。
我们更关心$\mu < 500$,称为单边检验。

小样本均值检验：单边t检验

$\bar{X}_n = 499.412$,前面t检验显著，我们怀疑短斤少两!

- 假设$H_0: \mu \geq 500 vs H_1: \mu < 500$
- ®假设总体是正态分布®，同样$T_0 = \frac{\bar{X}_n - \mu}{s\sqrt{n}} \sim t(n-1)$.
- 检验统计量$T = \frac{\bar{X}_n - 500}{\sqrt{0.676 * 9}}$的分布未知!
  但是如果假设H_0成立，$T \geq T_0 = \frac{\bar{X}_n - \mu}{\sqrt{0.676 * 9}}$.
- 要控制T的概率，如果$T > T_0$, 选左边区间!!!(见后面解释)
  有$P(T < -t_{0.05}(8)) \leq P(T_0 < -t_{0.05}(8)) = 0.05$
- 查表$-t_{0.05}(8) = -1.86$, $T = -2.609 < -1.86$,所以否定原假设，即该批产品质量不足。

相关说明：

- 一般的假设可以是$H_0: \mu \leq \mu_0, H_0: \mu = \mu_0, H_0: \mu \geq \mu_0$
  第一，三情形都用单边(t)检验。
- 单边与双边检验查表时分位数不同!
- 第一种情形(案例一）：牛奶兑水会改变冰点-0.545 (变高)。

方差检验：χ^2检验

厂家关心质量控制的重点是方差的大小。机器不正常工作的另一个标志是质量不稳定，实际方差大。

- 假设$H_0: \sigma^2 \leq 0.8 vs H_1: \sigma^2 > 0.8$
- 总体是正态分布，样本方差服从χ^2分布，$\xi_0 = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$
- 如果假设H_0成立，检验统计量$\xi = (n-1)S^2/0.8 \leq \xi_0$.
  要控制ξ的概率，如果$\xi \leq \xi_0$, 选右边区间!!! (见后面解释)
  $P(\xi > \chi^2_{0.05}) < P(\xi_0 > \chi^2_{0.05}) = 0.05$
  取单边拒绝域$\xi \geq \chi^2_{0.05}$, 则其发生概率小于0.05.
- 查表$\chi^2_{0.05}(8) = 15.507$,计算$\xi = 8 * (0.676)^2/0.8 = 4.57 < 15.507$ 不能拒绝原假设。
- 说明该机器的质量比较稳定，主要问题是均值!(度量问题？)

相关说明：

- 对应的也可以用P值检验；
- 也可以先做χ^2双边检验.结果同样不显著。

单边检验的拒绝域：解释

设单边检验$H_0: \mu \le \mu_0$或$H_0: \mu \ge \mu_0$

- 单边检验的统计量T是未知分布(μ有不同取值)。但一般与最大或最小值对应的分布$T_0: \mu = \mu_0$有大小关系。具体情况见前面两页：
- 如果随机变量$T > T_0$,则有分布函数$F_T(x) \le F_{T_0}(x)$,
  即$P(T < x) \le P(T_0 < x)$；(证明：事件$\{T < x\} \subset \{T_0 < x\}$.)
  分布密度图像看：T的分布密度函数在T_0的右边。
- 如果随机变量$T > T_0$, 应该选左边区域为拒绝域，因为拒绝域的概率$P(T < x)$有$P(T_0 < x) = \alpha$控制. 若选右边区域，则无法控制。
- (类似)随机变量$T < T_0$,有$F_T(x) \ge F_{T_0}(x)$即$P(T > x) \le P(T_0 > x)$.
  说明：事件$\{T > x\} \subset \{T_0 > x\}$. 图像上$T$的分布密度函数在$T_0$的左边。
- 如果随机变量$T < T_0$, 应该选右边区域为拒绝域，因为拒绝域的概率$P(T > x)$有$P(T_0 > x) = \alpha$控制. 若选左边区域，则无法控制。

其他检验与统计问题***

两个或多个正态总体的假设检验(非正态总体用正态总体逼近计算):

- 两个成对总体的均值差：服从t分布;
- 两个总体的均值差：服从正态分布或某种t分布，依赖与两个总体的方差信息。
- 两个总体的方差比：服从$F(m, n)$分布。

其他检验: 分布检验，秩检验(Wilcox), 偏度，峰度检验；

- 假设检验与区间估计的关系: α是一个；
  如果假设成立$\theta = \theta_0$,则拒绝域可定义为置信区间的外面!
- 相关分析: 用s_{xy} 估计σ_{xy};
  线性回归分析:$Y = aX + b$ 假设检验: $b = 0$;
- 方差分析: 非正态总体没法估计，有其他统计方法；χ^2检验, ANOVA方差分析.
- 抽样方法改进：bootstrap 从一个样本中得到若干个自助样本(每一个自助样本通过在样本中有放回抽取n次得到);

作业

北航教材:
P236 习题九2.3.4.5

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

December 17, 2010

Week 14：随机过程

Review：回顾

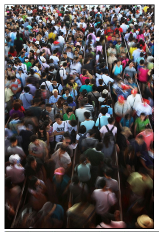
By now:

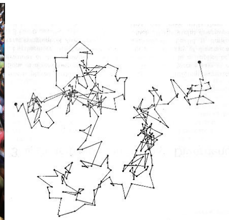
- 概率模型→ 随机变量
- 随机变量的特征与极限定理；
- 统计特征与统计推断；

This Week:

- 随机过程的概念
- 贝努利过程与泊松过程:
- 随机过程的特征：

生活中的随机现象





- 股票价格：
  → 时间序列分析，
  Black-Scholes 公式
- 邮局队列，地震预测；
  → 计数过程，泊松过程，排队论；
- 随机游走,布朗运动：→ 正态过程(维纳过程)，马尔可夫过程；
- 信号处理:
  → 平稳过程

随机过程的定义

Definition (随机过程定义二)

设参数集$T \in (\infty, \infty)$，给定$t \in T$,$X(t)$是一个随机变量，称所有的(一簇)随机变量为随机过程。可以看成一个二元函数$X : \Omega \times T \to R$.

Definition (随机过程定义一)

设随机试验一次得到的是一个函数$x(t)$，所有可能的试验结果是一簇函数$x(t), t \in T$. 称为一个随机过程$X(t)$(随机函数). 参数集$T \in (\infty, \infty)$. 一个函数$x(t)$称为样本函数。

- (理论): 随机过程即一簇相关的随机变量；(可看成无穷维随机向量)
- (应用): 随机过程可看成所有试验结果(样本函数)的一个总体分布。
- 参数集T一般是时间，也可以是空间(遥感图像)。
  状态空间: 二元函数$X : \Omega \times T \to R$的值域称为状态空间；
- 简单信号：$X(t) = A\cos(\omega t + \theta) + N(t)$.

随机过程的概率分布

Definition (随机过程的n维概率分布)

指定参数集T中n个参数$t_1, t_2, \ldots, t_n$，则随机过程在不同参数的随机变量$X(t_i)$有n维联合分布函数$F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots t_n) = P(X(t_1) \leq x_1, X(t_2) \leq x_2, \ldots X(t_n) \leq x_n)$，称其为随机过程的$n$维分布函数。特别如果有联合密度函数，称其为随机过程的n维密度函数。

- 随机过程的所有信息在其概率分布中(维数$n = 1, 2, \ldots$). 但一次随机试验得到的一个样本函数很难推导出分布信息。
- 独立随机过程：任意n个时刻的随机变量是相互独立的。则概率分布由一维分布完全决定。
  参见：贝努利过程(序列). 即不断抛硬币试验。
- 简单随机过程: $X(t) = X + f(t)$.
- 随机过程的运算与复合: $X(t) + Y(t)$.
  一般的必须考察两个随机过程的联合分布函数(很复杂);
  如果两个随机过程互相独立，则分布函数简化为分布的乘积。

随机过程的分类

按参数和状态空间分类

- 离散参数+离散状态：贝努利过程, 马尔可夫链(离散)
- 连续参数+离散状态：泊松过程, 生灭过程，马尔可夫链(连续)
- 离散参数+连续状态：离散时间信号(声音采样)；时间序列；
- 连续参数+连续状态：连续时间信号(声音等)；维纳过程；布朗运动；马尔可夫过程；

按概率分布的结构特征分类；

- 二阶矩过程: 正态过程，宽平稳过程；
- 马尔可夫过程: 马尔可夫链，生灭过程，排队论；
- 独立增量过程: 贝努利过程，泊松过程，维纳过程；
- 更新过程，鞅...

贝努利过程

EXAMPLE (贝努利序列)

$X(n), n = 1, 2, \ldots$，任一n,$X(n) \sim B(1, p)$,且是独立随机过程。

- 常见掷硬币，掷骰子(结果为6或非6),计算机状态(忙或不忙)
- n维联合分布: 是2^n个点上的离散分布(独立分布的乘积)。
  一个样本函数是一个序列$HTTTHHTT$...
- 独立性与无记忆性: 任意两组不同时刻的随机变量复合$g(X(t1), X(t2), \ldots), h(X(s1), X(s2), \ldots)$互相独立。
  特别未来试验与现在和过去独立。任一时刻开始的新贝努利过程与原来贝努利过程是一样的(分布)。(称为过程的重新开始).
- 相关的计数过程: $N(n)$记到时刻n时试验成功的次数；$N(n) \sim B(n, p)$. 特别首次成功次数$T(1)$服从几何分布。第k次成功次数$Y = T_1 + T_2 \cdots + T_k$服从负二项分布(帕斯卡分布);
- 贝努利过程的等价定义: 离散到达过程:每次相邻到达的次数服从独立的几何分布T_i；到达时间$Y_k = T_1 + T_2 + \cdots + T_k$.

贝努利过程例子***

EXAMPLE (下雨)

设下雨后再次下雨的天数服从参数为p的几何分布，则问明天和下周一同时下雨的概率。又问连续下雨七天，问随后的一天不下雨的概率？

解答: 可以用几何分布直接计算(?).
设下雨可以看出参数为p贝努利过程(到达),则在任一天下雨概率为p,互相独立，任两天下雨概率为p^2。
贝努利过程是可以重新开始的，七天后的那天不下雨概率还是$1-p$.

EXAMPLE (姚明的比赛时间)

设姚明每分钟犯规的概率为p,不同分钟内犯规是独立的。姚明犯规六次就会罚下。假设姚明不休息，则求其比赛时间的分布。*(最多48分钟)*.

解答: 犯规次数是个(到达)贝努利过程; 比赛时间$Z=min(Y_6,48)$.
特别$Z=48$概率$P(Y_6\geq 48)=1-\sum_{i=6}^{47}P(Y_6=i)$,
$P(Y_6=m)=C_{m-1}^5p^6(1-p)^{m-6}$(帕斯卡分布).

作业

北航教材:
P282 习题十一1.2.3.4

Quiz 小测验四

1. 设$X_1,X_2,\dots,X_n$是正态总体$X\sim N(\mu,\sigma^2)$的样本,当$C=()$时,$\hat{\mu}^2=\bar{X}^2+cQ^2$是$\mu^2$的无偏估计，其中$\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$, $Q^2=\sum_{i=1}^n(X_i-\bar{X})^2$.
（A）$-\frac{1}{n(n-1)}$,（B）$-\frac{1}{n-1}$（C）$-\frac{1}{n^2}$（D）$-\frac{1}{(n-1)^2}$。
2. 令$X_1,X_2,\dots,X_n$; $Y_1,Y_2,\dots,Y_m$分别为$N(\mu_1,\sigma^2)$, $N(\mu_2,\sigma^2)$的简单样本，记$\bar{X}=\frac{1}{n}\sum_i X_i$, $\bar{Y}=\frac{1}{m}\sum_j Y_j$, $S_1^2=\frac{1}{n-1}\sum_i(X_i-\bar{X})^2$. 试求：（1）$\bar{X}$服从的分布，$\bar{Y}$服从的分布；（2）$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n+1/m}}$服从的分布；（3）$T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_1\sqrt{1/n+1/m}}$服从的分布。
3. 根据大量调查得知，我国健康成年男子的脉搏平均为72次/分，标准差为6.75 次/分，现从某体院男生中，随机抽出25人，测得平均脉搏为69.3次/分。根据经验脉搏服从正态分布.如果标准差不变，试问该体院男生的脉搏与一般健康成年男子的脉搏有无差异？
检验水平$\alpha=0.05$, (已知$Z_{0.95}=1.645$, $Z_{0.975}=1.96$; $t_{0.95}(24)=1.7109$, $t_{0.95}(25)=1.7081$, $t_{0.975}(24)=2.0639$; $t_{0.975}(25)=2.0595$)

Quiz 小测验四: 答案

1. (A). 利用$\bar{X}\sim N(\mu,\sigma^2/n)$, $E\bar{X}^2=D\bar{X}+(E\bar{X})^2=\sigma^2/n+\mu^2$
利用$ES^2=\sigma^2$, $EQ^2=\sigma^2*(n-1)^2$,计算可得。
2. $\bar{X}\sim N(\mu_1,\sigma^2/n)$, $\bar{Y}\sim N(\mu_2,\sigma^2/m)$.
则$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\sigma^2(1/n+1/m))$,标准化得到正态分布；
又$(n-1)S^2/\sigma^2\sim\chi^2(n-1)$, 正态分布除以卡方分布得到$t(n-1)$分布。
3. $H_0:\mu=72.vs.H_1:\mu\neq 72$; 检验统计量$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$;
检验水平对应分位数$z_{0.975}=1.96$,计算
有$|Z|=2.7/6.75*5=2>1.96$, 检验显著，否定原假设，有差异。

时代周刊2010年度人物

- Mark Zuckerberg(1984,5,14) 26岁的马克・扎克伯格是哈佛大学计算机和心理学专业辍学生。
- 2004年创办美国社交网站Facebook, Facebook的注册人数已突破5亿，市值达1800亿美元.
- 2010 马克正在学汉语，即将访问中国。

泊松过程

EXAMPLE (泊松流)

称到达时刻为$t_1, t_2, \ldots$,的过程为强度为λ的泊松流, 如果相邻两次到达时间的分布服从参数为λ的指数分布且相互独立。$N(t)$为到达的个数，称为泊松过程。

- 常见: 事故的发生,顾客的到来；
- n维联合分布？$N(t)$服从参数为λt的泊松分布。
  首次到达服从指数分布，k次到达$Y_k = T_1 + \cdots + T_k$服从Γ分布。
- 独立增量与无记忆性: 因为指数分布的无记忆性, 任一时刻开始的新泊松过程与原来泊松过程是一样的(分布)。(过程的重新开始).
  独立增量：$N(t_2) - N(t_1)$服从$\mathcal{P}(\lambda(t_2 - t_1))$与$N(t_1) - N(t_0)$是互相独立。
- 泊松过程的等价定义:: 贝努利过程的连续化
  (1)在相同时间长度τ内k次到达概率一样$P(k, \tau)$
  (2)不同时间段内到达数目互相独立；
  (3)小区间概率$P(0, \tau) = 1 - \lambda\tau + o(\tau), P(1, \tau) = \lambda\tau + o(\tau)$.

泊松过程例子***

EXAMPLE (排队)

排队等待概率:设有三个窗口，三个人正在办理，每个人的办理时间服从指数分布$\varepsilon(\lambda)$, 设你是下一个，则你最后办完的概率是1/3.

解答: 泊松过程的无记忆性, 你开始办理时，相当于三个人同时开始一个新的泊松过程；三人服务时间的分布一样，你最后离开的概率为1/3。

EXAMPLE (银行的等待时间)

设你进入银行取号,知道有50个人在等待。设平均每分钟可以服务完两个人，问你期待等多长时间？等待时间超过30分钟的概率是多少？

解答: 服务时间服从$\lambda = 2$的指数分布，离开的过程是泊松过程。
$N(t) = 50$服从Γ分布，期望为指数分布的期望和$50 * 1/\lambda = 25$分钟。特别等待时间超过30分钟的概率$P(Y_k \geq 30) = \int_{30}^{\infty} f_{\Gamma(50)}(x)dx$.

均值与方差函数

Definition (均值与方差函数)

给定随机过程$X(t)$, 定义均值函数$\mu(t) = E(X(t)) = \int_{-\infty}^{\infty} x f_1(x; t)dx$; 方差函数$DX(t) = E((X(t) - \mu(t))^2)$, 特别记二阶矩$EX^2(t)$为均方值函数$\Psi_X^2(t)$。

- 可以定义随机过程的n阶矩. 一般仅考察二阶矩，特别如果随机过程的所有二阶矩存在，称其为二阶矩过程。
- 正态过程: 如果随机过程的任意n维分布都是正态分布，称为正态过程。正态过程完全由其所有二阶矩决定。(包含后面的混合矩：自相关函数)

EXAMPLE (随机相位正弦波)

$X(t) = A\cos(\omega t + \theta)$,其中$\theta$是$(0, 2\pi)$上均匀分布。求其均值与方差函数。如果存在噪音$WN(t)$在任一时刻服从正态分布$N(0, \sigma^2)$,不同时刻互相独立。求$Y(t) = X(t) + WN(t)$的均值与方差函数。

自相关与互相关函数

Definition (自相关函数)

给定随机过程$X(t)$, 定义二阶混合矩为自相关函数$R_X(t_1,t_2)=E(X(t_1)X(t_2))$, 特别二阶中心混合矩为自协方差函数$Cov(t_1,t_2)=E((X(t_1)-\mu(t_1))(X(t_2)-\mu(t_2)))$.

Proposition (矩函数的关系)

$\Psi^2(t)=R_X(t,t)$;$Cov(t_1,t_2)=R_X(t_1,t_2)-\mu(t_1)\mu(t_2)$
$DX(t)=Cov(t,t)=\Psi^2(t)-\mu^2(t)$;

Definition (互相关函数)

给定两个随机过程$X(t),Y(t)$,称$(X(t),Y(t))$是二维随机过程(存在联合分布函数); 定义二阶混合矩为互相关函数$R_{XY}(t_1,t_2)=E(X(t_1)Y(t_2))$. 类似互协方差函数$Cov_{XY}(t_1,t_2)=E((X(t_1)-\mu_X(t_1))(Y(t_2)-\mu_Y(t_2)))$. 特别如果互协方差函数恒为零，称两个随机过程不相关。

计算例子

EXAMPLE (随机相位正弦波)

$X(t)=A\cos(\omega t+\theta)$,其中$\theta$是$(0,2\pi)$上均匀分布。求相关函数和协方差函数。
如果存在噪音$WN(t)$在任一时刻服从正态分布$N(0,\sigma^2)$,不同时刻互相独立。求$Y(t)=X(t)+WN(t)$的相关和协方差函数。

Proof.

$\mu_X(t)=0, D_X(t)=A^2/2$;
$R_X(t_1,t_2)=A^2/2\cos\omega(t_2-t_1)=Cov_X(t_1,t_2)$
$\mu_{WN}(t)=0, D_{WN}(t)=\sigma^2, R_{WN}(t_1,t_2)=0$;
$\mu_Y(t)=0, D_Y(t)=A^2/2+\sigma^2$;
一般噪音与信号独立，即互协方差函数为0，这里互相关函数为0.
$R_Y(t_1,t_2)=R_X(t_1,t_2)+0=Cov_Y(t_1,t_2)$.

□

作业

北航教材:
P282 习题十一5.6.11.12

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

December 24, 2010

Week 15：平稳随机过程

1. 平稳随机过程与信号处理
   - 定义
   - 相关函数与谱密度***
   - 随机信号处理***
2. 维纳过程与金融分析***
   - 维纳过程
   - 股票期权定价：Black-Scholes 公式
3. 平稳过程的各态遍历性

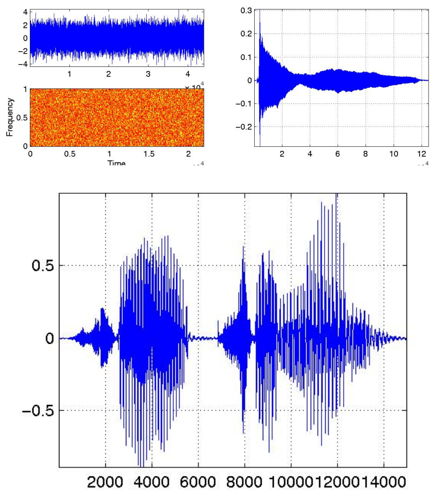
Review：回顾

Last Week:

- 随机过程的概念
- 贝努利过程与泊松过程:
- 随机过程的特征：

This Week:

- 平稳过程
- 随机信号处理
- 正态(维纳)过程与金融分析;
- 平稳过程的统计特性：各态遍历性



严格平稳随机过程

Definition (严格平稳随机过程)

如果对任一h, 随机过程的任意n维分布函数满足$F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots t_n) = F(x_1, x_2, \ldots, x_n; t_1 + h, t_2 + h, \ldots t_n + h)$,称$X(t)$是严格平稳过程。

- 说明：随机过程的有限维分布函数与时间的起点选择无关。
  性质:严格平稳过程的一维分布相同，二维分布仅有赖于时间差。
  独立同分布的随机过程必然是平稳的。
- 例子：飞机正常飞行的高度$H(t)$是平稳的, 起飞下降是不平稳的。
- 贝努利过程$X(t)$,白噪音过程$WN(t)$是严格平稳的；泊松过程$N(t)$不是。随机相位正弦波不是。

Proposition (严格(二阶)平稳过程的数字特征)

$E(X(t)) = \mu(X(t_0))$, $E(X^2(t)) = \Psi^2(X(t_0))$, $D(X(t)) = D(X(t_0))$;
$E(X(t)X(t+\tau)) = R_X(\tau)$,$Cov(t, t+\tau) = Cov(\tau)$.

宽平稳随机过程

Definition (宽平稳随机过程WSS,二阶平稳过程)

如果随机过程是二阶矩过程，且满足称$X(t)$是$E(X(t))$是常数，$E(X(t)X(t+\tau)) = R_X(\tau)$是$\tau$的函数，称$X(t)$是(宽)平稳过程.

- 根据定义有$DX(t)$是常数，$Cov(t, t+\tau)$是τ的函数。
- 结论：如果严格平稳过程有二阶矩，必然是宽平稳过程。
  结论: 正态的宽平稳过程必然是严格平稳过程。
  结论: 同分布,不同时刻互不相关的随机过程是宽平稳的。
- 例子：随机相位正弦波是宽平稳过程。实际应用中可以假定在一段时间内都是平稳的。

Definition (随机过程的平稳相关)

给定平稳随机过程$X(t), Y(t)$,如果互相关函数满足$E(X(t)Y(t+\tau)) = R_{XY}(\tau)$. 称$X, Y$是平稳相关的或联合平稳的。类似可定义相关系数函数$\rho(\tau) = Cov(\tau)/\sigma_X\sigma_Y$.

宽平稳过程的例子

EXAMPLE (随机电报信号)

设有两点分布$P(X_0 = 1) = P(X_0 = -1) = 1/2$和泊松过程$N(t), t \geq 0$, 定义$X(t) = X_0(-1)^{N(t)}, t \geq 0$为在$t$时刻发的电报信号(依赖于泊松过程发生次数的奇偶性). 说明其是(宽)平稳过程.

解答.

设两点分布与泊松过程独立，$E(X(t)) = E(X_0)E((-1)^{N(t)}) = 0$
设$h > 0$, $Cov(X(t), X(t+h)) = E(X(t)X(t+h)) =$
$E(X_0^2(-1)^{N(t)+N(t+h)}) = E((-1)^{N(t)+N(t+h)})$
$Cov(X(t), X(t+h)) = E((-1)^{2N(t)+N(t+h)-N(t)}) =$
$E((-1)^{N(t+h)-N(t)}) = E((-1)^{N(h)})$
$Cov(X(t), X(t+h)) = \sum_{i=0}^{\infty} e^{-\lambda h}\frac{(\lambda h)^i}{i!} = e^{-2\lambda h}$.
$h < 0$, 则结果是$e^{-2\lambda|h|}$. $R(h)$是偶函数。
说明: $X(t)$可看成一个粒子沿直线的速度函数，$N(t)$就是粒子可能发生碰撞反弹的次数. □

平稳随机序列的相关函数与谱密度***

平稳随机过程的主要特征即一维相关函数$R(\tau)$.考察平稳随机序列。

Proposition (相关序列的性质)

自相关序列$R(n)$和自协方差矩阵$Cov(n, m) = (R(m-n))$,满足对称性和正定性：即$R(n) = R(-n)$, $Cov(n, m)$是对称正定(无穷)矩阵(所有特征根非负). 一般情况n充分大，$R(n)$逼近零。

Definition (谱密度)

给定平稳随机过程$X(t)$及其相关函数$R(t)$,则称$S(\omega) = \mathcal{F}(R(t)) = \int R(t)e^{-i\omega t}dt$为其功率谱密度。特别平稳随机序列有周期的功率谱密度。
功率$\int X^2(t)dt \to \Psi^2 = \frac{1}{2\pi}\int S(\omega)d\omega$.

- 白噪声：$R(n) = \sigma^2\delta(n)$, $S(\omega) = \sigma^2$.
- 随机相位正弦波: $R(n) = A^2/2\cos(\omega_0 n)$, $S(\omega) = A^2/2\pi(\delta(\omega-\omega_0) + \delta(\omega+\omega_0))$.

随机信号处理***

随机信号的模型: 宽平稳的二阶随机序列(过程) $X(n)$.

- Wold分解定理: 任一平稳过程$x(n)$可以写成$x(n) = x_r(n) + x_p(n)$,其中$x_r(n)$是有连续功率谱(正则过程)，$x_p(n)$有离散功率谱(可预测过程)。且$E(x_r x_p) = 0$(正交). 有$r_x(k) = r_r(k) + r_p(k)$。
  一般有ARMA模型: 如果平稳随机信号满足$x(n) = -\sum_{k=1}^{p} a(k)x(n-k) + \sum_{k=1}^{q} b(k)w(n-k)$, 其中$w(n)$是白噪音输入;称为ARMA(p,q)过程。自回归移动平均模型。
- MA模型即ARMA(0,q)模型,$x(n) = \sum b(k)w(n-k)$,移动平均模型。正则平稳随机过程：清辅音(不用声带振动的语音).
- AR模型：即ARMA(p,0)模型，$x(n) = \sum a(k)x(n-k) + w(n)$,可预测过程，元音及含声带振动的辅音，拟周期性；离散的功率谱

随机信号处理

- 信号分析: 谱分析，ARMA模型的系数求解；
- 信号处理: 滤波器构造(线性时不变系统); 维纳滤波器，Kalman滤波器；

作业

北航教材:
P305 习题十二1.2.4.5

维纳过程

Definition (维纳过程,布朗运动)

给定$X(t), t \geq 0, X(0) = 0$是独立增量过程，即任意$t_1 < t_2 < \cdots < t_n$,满足$X(t_n) - X(t_{n-1}), X(t_{n-1}) - X(t_{n-2}), \ldots, X(t_2) - X(t_1), X(t_1)$是独立的，且满足增量的分布服从$N(0, \sigma^2(t_{i+1} - t_i))$. 称其为维纳过程或布朗运动。

- 布朗运动是正态过程. n维联合分布已知,任何条件分布都是正态分布。$\sigma = 1$称为标准布朗运动。
- 布朗运动$X(t) \sim N(0, \sigma^2 t)$不是平稳过程。
  $Cov(t_1, t_2) = min(t_1, t_2)\sigma^2$.
- 最有用的概率模型之一。应用于股票价格，量子力学等。
- 常用模型: 布朗运动(带漂移) $X(t) = \mu t + B(t) * \sigma$,其中$B(t)$是标准布朗运动.

股票价格与期权

EXAMPLE (股票的价格模型)

假设股票的价格变化(百分比)是独立同分布的，则股票价格的通用模型为几何布朗运动：$Y(t) = e^{\mu t + B(t) * \sigma}$。

解释：设$Y(0) = 1, Y(n)$是n时刻的股票价格，
则$\frac{Y(n)}{Y(n-1)}, \frac{Y(n-1)}{Y(n-2)}, \cdots, \frac{Y(2)}{Y(1)}, \frac{Y(1)}{Y(0)}$是独立同分布的。
记$X(n) = \frac{Y(n)}{Y(n-1)}$,则$Y(n) = \prod_{i=1}^{n} X(i), \log Y(n) = \sum_i \log X(i)$.
一般$\log X(i)$是独立同分布的(正态分布), 则$\log Y(n)$可逼近一个布朗运动(带漂移)。

EXAMPLE (股票期权)

设时刻*0*股票价格*100*元一股，时刻*1*的股票价格可能是200或50元。若预测股票价格是*200*元，希望现在以每股*150*元的价格预订该股票，则在时刻*1*每股盈利*50*元。
股票期权*option*：在未来时刻t以固定价格K购买某股票的权利。

简单模型的Black-Scholes 布莱克-肖尔斯公式***

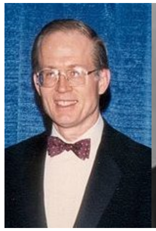
Definition (股票期权的组合投资策略arbitrage(套利))

设在时刻*0*购买x股股票，y股期权，如果存在投资策略在时刻*1*必然赢钱, 称为*arbitrage*。期望给出合理的期权价格使得*arbitrage*套利不存在*!*

- Arbitrage 套利定理: 或者存在arbitrage, 或者存在一个特殊的概率分布(关于时刻1的价格)使得任何人的期望盈利为零。
- 例子: 设每股(购买价$K = 150$)期权的价格为c, 若$c = 50/3$,则无人赚钱。设$c = 20, x = 1, y = -3$,则$Value = 50$,盈利为10.
- 一般情形：股票价格$Y(t) = Y_0 e^{\mu t + B(t) * \sigma}$，考虑利息为$\alpha$,则$e^{-\alpha t} Y(t)$是时刻$t$的真实股票价值。购买价为$K$的一股期权的盈利为$Y(t) - K$或0, 其定价$c$应该满足$E(e^{-\alpha t}(Y(t) - K)_+) = c$. 又要求购买一股股票的期望盈利为零,得到$E(e^{-\alpha t} Y(t)) = Y(0)$,
  有$\mu + \sigma^2/2 = \alpha$.
  计算有$c = Y_0 \phi(\sigma\sqrt{t} + b) - K e^{-\alpha t} \phi(b)$, 其中$\phi$为标准正态分布函数,
  $b = \frac{\alpha t - \sigma^2 t/2 - \log(K/Y_0)}{\sigma\sqrt{t}}$. 此即简单模型的Black-Scholes 公式。

真实世界的金融分析

Fischer Black(1938),
Myron Scholes(1941)




Robert Merton(1944)

- 1973年Fischer Black, Myron Scholes 发表Black-Scholes公式；Robert Merton 给出数学推导(Merton模型).
- 1997年诺贝尔经济学奖授予Scholes和Merton。F.Black1995年去世。
- 1994年Scholes和Merton参与建立长期资本管理基金(Long-Term Capital Management), 1995-1998年年盈利率40%, 1998四个月内损失46亿美元(俄国经济危机),2000年关闭。
- 长期资本管理基金是金融危机的一个前兆和经典案例。

平稳过程的时间平均

Remark (样本函数与数字特征)

随机过程的一个样本函数不能决定随机过程的均值函数或相关函数。简单的统计抽样要求重复得到大量样本函数,这是不现实的。
平稳过程的分布与时间无关,可用不同时间数据估计均值或相关函数。

Definition (样本函数的时间平均)

给定样本函数$x(t)$,定义时间均值$\overline{x(t)} = \lim_{t\to+\infty} \frac{1}{2t}\int_{-t}^{t} x(t)dt$; 时间相关函数$\overline{R(\tau)} = \overline{x(t)x(t+\tau)} = \lim_{t\to+\infty} \frac{1}{2t}\int_{-t}^{t} x(t)x(t+\tau)dt$;

- **随机过程的积分：$Y = \int X(t)dt$是一个随机变量！可看成不同样本函数得到的积分值的总体分布。
  以上时间均值，时间相关函数可定义在随机过程上得到随机过程的时间均值(随机变量)，随机过程的时间相关函数(随机过程)。
- 随机相位正弦波:$\overline{X(t)} = 0, \overline{X(t)X(t+\tau)} = A^2/2\cos\omega\tau$. 与随机过程的均值和相关函数一样!

平稳过程的遍历性ergodicity

Definition (平稳过程的遍历性)

给定平稳随机过程$X(t)$,如果时间均值$\overline{X(t)} \xrightarrow{as} E(X(t)) = \mu_X$, 称随机过程的均值有各态遍历性;
如果时间相关函数$\overline{X(t)X(t+\tau)} \xrightarrow{as} R(\tau)$,称随机过程的相关函数有各态遍历性;
如果以上两个遍历性都满足，称平稳随机过程是遍历的。

- 引理:$E(\overline{X(t)}) = E(X(t))$,
  $D(\overline{X(t)}) = \lim_{t\to+\infty} \frac{1}{t}\int_0^{2t}(1-\frac{s}{2t})(R_X(s)-\mu_X^2)ds$
- 定理：均值遍历$\Leftrightarrow \lim_{t\to+\infty} \frac{1}{t}\int_0^{2t}(1-\frac{s}{2t})(R_X(s)-\mu_X^2)ds = 0$.
  相关函数遍历$\Leftrightarrow \lim_{t\to+\infty} \frac{1}{t}\int_0^{2t}(1-\frac{s}{2t})(B_X(s)-R_X^2(\tau))ds = 0$,其中$B_X(s) = E[X(t)X(t+\tau)X(t+s)X(t+s+\tau)]$.
- 平稳过程的点估计：(一般是随机序列的求和).
  $\hat{\mu_X} = \frac{1}{T}\int_0^T x(t)dt$; $\hat{R}(\tau) = \frac{1}{T-\tau}\int_0^{T-\tau} x(t)x(t+\tau)dt$.

作业

北航教材:
P305 习题十二6.8.9.11

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

December 31, 2010

Week 16：马尔可夫链

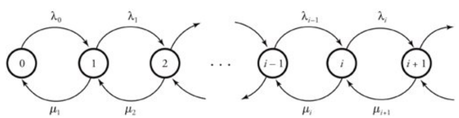
Review：回顾

Last Week:

- 平稳过程
- 随机信号处理
- 正态(维纳)过程与金融分析；
- 平稳过程的统计特性：各态遍历性

This Week:

- 马尔可夫过程
- 马尔可夫链的概率分布
- 离散马尔可夫链的状态与平稳分布
- 应用



赌徒的破产模型：

马尔可夫过程

Definition (无后效性或马尔可夫性)

给定随机过程$X(t)$, 时刻$t_1 < t_2 < \cdots < t_n$, 如果给定条件$X(t_i) = x_i, 1 \le i \le n-1$, $X(t_n)$的条件分布函数$P\{X(t_n) \le x_n | X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_{n-1}) = x_{n-1}\} = P(X(t_n) \le x_n | X(t_{n-1}) = x_{n-1})$,称$X(t)$有无后效性或马尔可夫性，称其为马尔可夫过程。如果状态离散称为马尔可夫链。

- 来源：动力系统(微分方程初值问题)；非平稳过程的简化模型；
- 例子：独立随机过程，独立增量过程($X(0) = 0$)都是马尔可夫过程。包含贝努利过程，白噪音过程，泊松过程,维纳过程。平稳过程一般不是。
- 分类：离散时间马尔可夫链(贝努利过程，随机游走)；连续时间马尔可夫链(泊松过程,生灭过程)；离散时间马尔可夫过程(白噪音序列),连续时间马尔可夫过程(维纳过程,白噪音过程).

齐次离散时间马尔可夫链

Definition (齐次马尔可夫过程)

如果马尔可夫过程的条件概率分布与时间起点无关，称为齐次或时齐的。类似平稳过程与时间起点无关，仅与时间距离有关。(也称为平稳性).

Definition (离散时间马尔可夫链的转移概率)

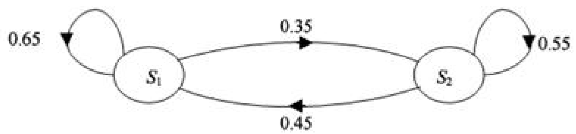
设过程为$X(n)$, 状态空间为$I = \{a_0, a_1, \dots,\}$,则记条件概率分布$P(X(1+n) = a_j | X(n) = a_i) = p_{ij}(n, n+1)$,称为马氏链在时刻$n$下处于状态$a_i$条件下，在时刻$1+n$转移到状态$a_j$的一步转移概率。

- 条件概率分布完全由所有状态的转移概率决定(状态为至多可数的). 一般的可以定义时刻m到时刻$m+n$的转移概率$p_{ij}(t_m, t_m+n)$。
- 齐次马氏链：一步转移概率与起始时刻m无关，是固定常数。可以写成无穷矩阵P_{ij}. 可记时刻m到时刻$m+n$的转移概率为$P_{ij}(n)$或$P_{ij}^{(n)}(t_m)$.

齐次马氏链的表示

矩阵表示：$\begin{bmatrix} P_{00} & P_{01} & \cdots & P_{0j} & \cdots \\ P_{10} & P_{11} & \cdots & P_{1j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i0} & P_{i1} & \cdots & P_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$

状态图表示：(稀疏矩阵用图表示更简单).



一般齐次马氏链可用状态及其一步转移概率图表示。
类似动力系统: 任意n维概率分布由初值和转移矩阵得到。

齐次马氏链的例子

EXAMPLE (下雨预测:两个状态的马氏链)

假设任一天下雨概率仅仅依赖于前一天是否下雨。设有两个状态0, 1分别表示不下雨或下雨. 设已知今天下雨，明天下雨概率a;已知今天不下雨，明天不下雨概率为b, 则其为两个状态马氏链，完全由矩阵$P = [a, 1-a; 1-b, b]$决定。

***如果今天下雨依赖于前几天的天气呢？不是马氏过程，但可以增加状态得到马氏链。

EXAMPLE (一维随机游走: 课本例二)

随机游走: random walk 一般满足$p_{i,i+1} + p_{i,i-1} = 1$.可能存在吸收状态$p_{i,i} = 1$,反射状态$p_{i,i-1} = 1$或$p_{i,i+1} = 1$.

转移概率与CK方程

Proposition (一步转移概率矩阵的性质)

$p_{ij} \geq 0$, $\sum_{j=0}^{\infty} p_{ij} = 1$. 一般称满足以上性质的矩阵为随机矩阵。记为$\mathbb{P}$.

Theorem (Chapman-Kolmogorov方程)

多步转移概率$p_{ij}^{(n+l)} = \sum_{k=0}^{+\infty} p_{ik}^{(n)} p_{kj}^{(l)}$.

证明: 利用条件概率定义和全概率公式。(非齐次有类似结果)

- 定义多步转移概率矩阵$P_{ij}^{(n)}$, 则$P_{ij}^{(n)} = \mathbb{P}^n$, CK方程对应与矩阵乘法。$\mathbb{P}^{n+l} = \mathbb{P}^n\mathbb{P}^l$.
- 多步转移概率矩阵也是随机矩阵。
- 计算今天下雨,两天后下雨的概率。$p_{00}^{(2)} = (P)_{00}^2$.
  ***注: 矩阵乘法的简化: 设有约旦分解$P = AQA^{-1}$,则$P^{(n)} = AQ^nA^{-1}$.

齐次马氏链的概率分布

齐次马氏链的一维分布:

- 初始分布: 记$P(X(0)=j)=p_j(0)$;
- 任一时刻的分布:$P(X(n)=j)=p_j(n)$称为瞬时概率或绝对概率;
  定理: $p_j(n)=\sum_{i=0}^{+\infty} p_i(0)p_{ij}^{(n)}$.
- 例子：未来任一天的下雨分布$X(n)$.
- 例子：随机游走(例二);

Theorem (齐次马氏链的n维分布)

给定齐次马氏链的初始分布$X(0)$,则其n维联合分布$(X(k_1),X(k_2),\ldots,X(k_n))$满足
$P(X(k_1)=i_1,X(k_2)=i_3,\ldots,X(k_n)=i_n)=\sum_{j=0}^{+\infty} p_j(0)p_{ji_1}^{k_1}p_{i_1,i_2}^{k_2-k_1}\ldots p_{i_{n-1}i_n}^{k_n-k_{n-1}}$.

作业

北航教材:
P334 习题十三1.2.3.4

齐次马氏链的状态空间***

Remark (状态图与样本函数)

马氏链的概率分布信息都在状态图中, 任一样本函数(随机过程的一次实现)就是在状态图中游走，特别对齐次链, 一个无穷长度样本函数包含充分的概率分布信息。

状态空间的分类：

- 状态图的分解：连通性: 存在道路$i\to j, j\to i$, 即存在m,n,使得概率$p_{i,j}^{(n)}, p_{j,i}^{(m)}$大于零。
  一般仅仅考虑只有一个连通分支的状态图(称为不可约马氏链);
- 齐次链的常返与非常返(过渡)：无穷次返回或有限次返回;
  常返态：给定i状态, 以概率1返回状态i, 等价于$\sum_{n=0}^{\infty} p_{ii}^{(n)}=\infty$;
  过渡态：给定i状态, 以概率$P<1$返回状态i, 等价于$\sum_{n=0}^{\infty} p_{ii}^{(n)}<\infty$;
- 齐次链的周期与非周期：定期返回或不定；(图中存在唯一环路)
  周期性：$p_{ii}^{(n)}=0$, 除非$n=d,2d,3d,\ldots\ldots$ 称最大的d为周期。否则无周期

齐次马氏链的极限分布

Definition (极限分布)

如果齐次马氏链满足任意i, $\lim_{n\to+\infty} p_{i,j}^{(n)}=p_j$, 且$\sum p_j=1$, 称$p_j$为马氏链的极限分布或平稳分布. 特别其满足方程$p_j=\sum_{i=0}^{\infty} p_i p_{i,j}$.

- 极限分布意义: 即无论初始状态，到达状态j的概率为p_j.
- 平稳分布意义：假设初始分布取为p_i,则马氏链是平稳的.
- 例子: 天气预报下雨：两个状态马氏链有极限分布。$p_0=b/(1+b-a), p_1=(1-a)/(1+b-a)$.
- 应用: 社会科学与生物科学中的极限分布: 社会的阶级分层；一个基因的分布律;

Theorem (极限分布的存在*)**

一个不可约遍历(非周期正常返)的齐次马氏链必然存在极限分布。
特别对有限状态马氏链，如果存在m,使得$\mathbb{P}^{(>)}$的每一个元素大于零, 则存在极限分布。

赌徒破产问题

EXAMPLE (赌徒的随机游走)

设赌徒的资本为i元,每次赌一元(互相独立, 其模型为一个随机游走的马氏链：满足$p_{i,i+1} = p = 1 - p_{i,i-1}$。问赌徒资金达到$N$元而不先破产的概率是多少？

- 资本为N元或0则停止, 有$p_{00} = 1, p_{NN} = 1$.
- 马氏链有两个吸收态，其他都是过渡态，显然最后必然是破产或资本为N元。
- 定义吸收概率a_i为初始状态为i, 最后状态为N的概率。
  有方程组$a_N = 0, a_0 = 0, a_i = (1 - p)a_{i-1} + pa_{i+1}$
- 解方程有$a_i = \frac{1-\rho^i}{1-\rho^N}, \rho \neq 1$或$a_i = i/N, \rho = 1, \rho = (1 - p)/p$.
- 结论:N充分大, $P = 1/2$, 概率与原始资本成正比。
  $P > 0.5$,可能得到无穷的财富；$P < 0.5$,必然破产。

人类消亡问题:分支过程

设$X(n)$是第n代的人口数，假设$X(0) = 1$.假设每一个人的后代数服从独立的分布Z,则$X(n + 1) = \sum_{k=0}^{X(n)} Z_k$. 问$P(X(n) = 0)$即人类消亡的概率!

- 分支过程是齐次马氏链。如果$Z = 0$的概率大于零，则状态0是唯一吸收态。人口或者消亡或者趋于无穷大。
- 记$\mu = EZ = \sum_j jP(Z = j)$为每个人的平均后代数。则第$n$代人口期望$EX(n) = E(E(X(n)|X(n - 1))) = \mu^n$.
- 记人口消失概率$\pi_0 = \lim P(X(n) = 0|X(0) = 1)$. 如果$\mu \leq 1$, $\pi_0 = 1$. (注：$\mu^n = EX(n) \geq P(X(n) \geq 1) = 0$.)
  如果$\mu > 1$,有$P(X(n) = 0|X(0) = 1) = \sum_j P(X(n) = 0|X(1) = j, X(0) = 1)P(X(1) = j)$, 记$P(X(1) = j) = P(Z = j) = b_j$, 取极限有$\pi_0 = \sum_j \pi_0^j b_j$. 解方程可得最小的正数解。
- 例子：设$b_0 = 1/2, b_1 = 1/4, b_2 = 1/4$,则$\mu = 3/4 < 1 \to \pi_0 = 1$
- 例子：设$b_0 = 1/4, b_1 = 1/4, b_2 = 1/2$,则方程为$\pi_0 = 0.25 + 0.25\pi_0 + 0.5\pi_0^2, \pi_0 = 0.5$;
- 不必悲观! 如果$X(0) = n$,则人口消亡的概率为π_0^n.

连续时间马尔可夫链***

Definition (连续时间马氏链的转移概率和速率)

设$X(t)$是连续时间马氏链,定义可转移概率函数$P(X(t + \tau) = j|X(t) = i) = p_{i,j}^{(\tau)}(t)$. 特别对齐次链，函数与$t$无关。
局部有$p_{i,j}^{(\tau)} = q_{i,j}\tau + o(\tau)$, $q_{i,j}$称为转移速率。

类似有CK方程和极限分布。
应用：

- 分支过程: 如果考察每个人的寿命(连续分布),则在不同时刻的人口数是连续时间过程;
- 排队论(一般有生灭过程), 到达时间是泊松过程，离开(服务完成)也是一个泊松过程。

作业

北航教材:
P335 习题十三5.6.7.8

概率统计与随机过程

Probability，Statistics and Stochastic Process

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 Systems sciences, BUAA

January 7, 2011

Week 17：复习和讲解

考试事宜

- 考试时间：一月十八号8点-10点，
  主M203(21,22班)，M204(31,32班).
- 答疑时间：
  课程统一答疑：一月十七号8点30-11点30 主216(张思容等)
  下午2：30-5：30, 晚上：6-9 主216(其他老师,邢家省老师晚上)
  班级答疑：预约或17号下午1-5pm, 图书馆西配楼501室。
- 联系方式：134-3920-1025. zhangsirong@buaa.edu.cn
- 期末考试不考内容：
  1. 第四章：第三节中$Z = max(XY)$与$Z = min(X,Y)$，其中(X,Y)为连续型r.v. 求$F_Z(z)$，当X, Y不独立时不要求。
  2. 第五章：第五节
  3. 第七章：χ^2, t, F分布的概率密度函数表达式不要求记
  4. 第八章：第五节
  5. 第九章：第三、四节
  6. 第十二章：第五节
  7. 第十三章：第三节

概率论(1)：概率模型

- 概念：样本空间+概率律；事件；独立与条件概率；
- 结论：事件复合；加法公式，乘法公式；全概率和贝叶斯公式；排列组合公式；
- 例子：抽签原理，生日问题，配对问题，假阳性，重复射击；...

概率论(2): 随机变量与随机向量

- 概念：随机变量X,(累积)分布函数$F(x)$,概率密度函数$f(x)$, p_i. 事件$(-\infty, x)$
  随机向量,联合分布$F(x, y)$或表格, 事件$(-\infty, x) \times (-\infty, y)$;
- 结论: 边缘分布概率$F_X(x)$公式；条件分布概率公式$F(X|Y)$；边缘密度公式, 条件密度公式及独立；
- 随机变量例子: 贝努利分布,几何分布,泊松分布; 均匀分布，指数分布，正态分布；
  随机向量例子: 二维离散分布；二维正态分布；二维均匀分布(圆盘); n维独立分布乘积；

概率论(3): 复合随机变量和计算

- 概念：随机变量的复合$g(X_1, X_2, \dots)$
- 结论:
  独立随机变量的复合：X, Y独立则$g(X), h(Y)$独立。
  一维情形: 离散$P(g(X) = y) = \sum_{g^{-1}(y)} P(X = g^{-1}(y))$
  连续: $f_Y(y) = \sum_i f(h_i(y))|h_i'(y)|$, $h_i(y) = x$
  二维情形: $F_{Z=g(X,Y)}(z) = \int\int_{g(x,y)\le z} f(x, y)dxdy$
- 例子: 离散随机变量的简单复合；比如$X^2, X + Y, max(X, Y)$
  连续情形: 独立正态分布的线性组合$N(\sum \mu_i, \sum \sigma_i^2)$,$Z = X^2 + Y^2$;
  两个独立分布的加法(密度的卷积),极大值分布(分布函数乘积),极小值分布;
  常见分布复合：二项分布是贝努利分布的和；独立泊松分布的和是泊松分布；指数分布的极大极小值分布；

概率论(4): 数字特征

- 概念：期望算子EX, 矩DX, $Cov(X, Y)$, $\rho(X, Y)$,不相关；
- 结论: $E(g(X)) = \int g(x)f(x)dx$;
  线性$E(X + Y) = EX + EY$,$EXY = EXEY$(独立);
  $D(X + Y) = DX + DY - 2Cov(X, Y)$;
- 例子: 所有常见分布的期望与方差；常见分布的复合的期望；
  正态分布的高阶矩(特别密度函数为偶函数的期望为零);
  二维正态分布的相关系数ρ;
  特殊随机变量特征$Ea = a, Da = 0$, $DX = 0 \iff P(X = EX) = 1$,
  $Cov(X, Y) = 1 \iff P(aX + b = Y) = 1$***

概率论(5): 不等式与极限定理

- 概念：依概率收敛，独立同分布；
- 结论: 切比雪夫不等式的不同形式，(马尔可夫不等式**)
  切比雪夫大数定理:$\lim P(\bar{X}_n = \mu) = 1, \bar{X}_n \xrightarrow{p} \mu$,
  中心极限定理：$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$
- 例子: 概率的频率解释；二项分布的正态分布逼近计算；

数理统计(1): 样本与三大分布

- 概念：总体与样本；简单抽样；样本矩$\bar{X}_n, s_n^2$, 统计量，分布的下分位数；
- 结论: 独立分布的复合：Z_i是标准正态分布；
  $\chi^2(n) = Z_1^2 + \cdots + Z_n^2; EX = 0, DX = 2n$
  $t(n) = Z/(\chi^2(n)/n)$, 对称性；
  $F(m, n) = \frac{\chi^2(m)/m}{\chi^2(n)/n}$. 非负性*,互逆性*；
- 例子：正态总体的统计量分布:
  样本均值与样本方差独立；
  $\bar{X}_n \sim N(\mu, \sigma^2/n), (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$；
  $\to (\bar{X}_n - \mu)/(S/\sqrt{n}) \sim t(n-1)$.
  两个正态总体的均值差的分布,样本方差比的分布***;

数理统计(2): 点估计与评估

- 概念：参数统计量$\theta = g(X_1, \ldots, X_n)$;
  矩估计，极大似然估计；
  无偏性，有效性(方差小)，一致性(相合性);
- 结论: 矩估计都是一致的**；
  样本均值是期望的最小方差无偏估计MVU。
  样本方差是方差的无偏估计。
- 例子：常见分布的矩估计和极大似然估计(均匀分布*);
  正态总体的极大似然估计是有偏估计；
  给定函数的极大似然估计；

数理统计(2): 区间估计与假设检验

- 概念：置信区间与置信水平$1-\alpha$,
  零假设,备择假设,显著性水平α;
  置信区间的意义；假设检验的两类错误；
- 结论: 一个正态总体的统计推断分布：
  样本均值：已知方差，正态分布,正态检验；
  样本均值：未知方差，$t(n-1)$分布,t检验；
  样本方差：$\chi^2(n-1)$分布，卡方检验；
- 例子：以上三类的区间估计。
  置信水平与单侧检验*：$\alpha = 0.05, z_{0.975} = 1.96$

随机过程(1): 定义

- 概念：随机过程；样本函数；n维分布，数字特征；
- 要求: 随机过程的分类；
  简单随机过程的一，二维分布；
  随机过程的数字特征$EX(t), R(t_1, t_2)$ 和相互关系
- 例子：贝努利过程，泊松过程*;
  随机相位正弦波；

随机过程(2): 平稳过程

- 概念：严格平稳过程，平稳过程，正态过程，时间平均和遍历性，
- 要求: (广义)平稳过程验证；遍历性验证；
- 例子：随机正弦波及遍历性

随机过程(3): 马尔可夫链

- 概念：马尔可夫性,
  条件分布与一步转移矩阵,多步转移矩阵, 极限分布
- 结论: C-K方程；
  平稳分布方程；
- 例子：简单随机游走；

QUIZ 小测验一

1. 设A, B 为任意两事件,则下列关系成立的有( )
   (A) $(A+B)-B=A$;(B) $(A+B)-B=A-B$;
   (C) $(A-B)+B=A$;(D) $(A-B)+B=AB$.
2. 从$0 \to 9$ 这十个数码中任意取出4个排成一串数码,则数码恰成四位偶数的概率为：（A）$\frac{41}{90}$;（B）$\frac{1}{2}$ ；（C）$\frac{40}{90}$ ；（D）$\frac{32}{90}$
3. 一盒子内装有5个红球,15个白球,从中不放回取10次,每次取一个球,则第5次取到的是红球的概率为多少？
4. 袋中装有编号$1-8$的八个球,从中任取3个,则最小号码为偶数的概率为多少？

答案：1. B, 2.A , 3 $5/(5+15)=1/4$, 4 $11/28$

QUIZ 小测验二

1. 设随机变量X在$(-\pi/2, \pi/2)$上服从均匀分布,则$Y=\tan X$的概率密度$f_Y(y)=$
2. 将红、白、黑三只球随机地逐个放入编号为1，2，3的三个盒内（每盒容纳球的个数不限），以X 表示有球盒子的最小号码，求：（1）随机变量X的分布律；（2）X的分布函数。
3. 某仪器上装有4只独立工作的同类元件。已知每只元件的寿命(以小时计) $X \sim N(5000, \sigma^2)$,当工作的元件不少于2只时,该仪器能正常工作。则该仪器能正常工作5000小时以上的概率为_。
4. ***P63. 23题:设随机变量X的概率密度为$f(x)=\frac{a}{e^x+e^{-x}}$, (1)确定常数$a$; (2)求X的分布函数; (3)求$P(0 \le X \le \ln\sqrt{3})$.

QUIZ 小测验二：答案

1. $f_Y(y)=\frac{1}{\pi}\frac{1}{1+y^2}, -\infty<y<+\infty.$
2. （1）随机变量X的分布律；
$P(X=3)=1/27; P(X=2)=7/27; P(X=1)=19/27;$
（2）X的分布函数。
$F(x)=0, x<1; 19/27, 1\le x<2; 26/27, 2\le x<3; 1, x\ge 3$
3. 一个元件正常工作概率$p=P(X>5000)=1/2;$
元件工作个数Y服从$B(4,0.5)$,
仪器正常工作概率:
$P(Y\ge 2)=1-P(Y=0)-P(Y=1)=1-1/16-4/16=11/16.$
4. $a=2/\pi, F(x)=2/\pi\arctan e^x, P=1/6.$

QUIZ 小测验三

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）$P(|X-EX|\ge\epsilon)>\frac{DX}{\epsilon^2}$,（B）$P(|X-EX|<\epsilon)>1-\frac{DX}{\epsilon^2}$;
（C）$P(|X-EX|\ge\epsilon\sqrt{DX})\le\frac{1}{\epsilon^2}$;
（D）$P(|X|\ge\epsilon)\le\frac{E(X-EX)^k}{\epsilon^k}(k\ge 1)$,
3. 设随机变量X,Y的二阶矩EX^2,EY^2存在，证明：成立不等式$|EXY|\le\sqrt{EX^2EY^2}$. (构造$E(X+tY)^2\ge 0$)
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$\lim_{n\to\infty}P(|Y_n|\ge\epsilon)=0$。

Quiz 小测验四

1. 设$X_1,X_2,\ldots,X_n$是正态总体$X\sim N(\mu,\sigma^2)$的样本,
当$C=()$时,$\hat{\mu}^2=\bar{X}^2+cQ^2$是$\mu^2$的无偏估计，其中$\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i, Q^2=\sum_{i=1}^n(X_i-\bar{X})^2$.
（A）$-\frac{1}{n(n-1)}$,（B）$-\frac{1}{n-1}$（C）$-\frac{1}{n^2}$（D）$-\frac{1}{(n-1)^2}$ 。
2. 令$X_1,X_2,\ldots,X_n; Y_1,Y_2,\ldots,Y_m$分别为$N(\mu_1,\sigma^2),N(\mu_2,\sigma^2)$的简单样本，记$\bar{X}=\frac{1}{n}\sum_i X_i$, $\bar{Y}=\frac{1}{m}\sum_j Y_j$, $S_1^2=\frac{1}{n-1}\sum_i(X_i-\bar{X})^2$. 试求：
（1）$\bar{X}$服从的分布，$\bar{Y}$服从的分布；（2）$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n+1/m}}$ 服从的分布；（3）$T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_1\sqrt{1/n+1/m}}$ 服从的分布。
3. 根据大量调查得知，我国健康成年男子的脉搏平均为72次/分，标准差为6.75 次/分，现从某体院男生中，随机抽出25人，测得平均脉搏为69.3次/分。根据经验脉搏服从正态分布.如果标准差不变，试问该体院男生的脉搏与一般健康成年男子的脉搏有无差异?
检验水平$\alpha=0.05$, (已知$Z_{0.95}=1.645$, $Z_{0.975}=1.96$; $t_{0.95}(24)=1.7109$, $t_{0.95}(25)=1.7081$, $t_{0.975}(24)=2.0639$; $t_{0.975}(25)=2.0595$)

Quiz 小测验四：答案

1. (A). 利用$\bar{X}\sim N(\mu,\sigma^2/n)$, $E\bar{X}^2=D\bar{X}+(E\bar{X})^2=\sigma^2/n+\mu^2$
利用$ES^2=\sigma^2, EQ^2=\sigma^2*(n-1)^2$,计算可得。
2. $\bar{X}\sim N(\mu_1,\sigma^2/n)$, $\bar{Y}\sim N(\mu_2,\sigma^2/m)$.
则$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\sigma^2(1/n+1/m))$,标准化得到正态分布；
又$(n-1)S^2/\sigma^2\sim\chi^2(n-1)$, 正态分布除以卡方分布得到$t(n-1)$分布。
3. $H_0:\mu=72.vs.H_1:\mu\neq 72$; 检验统计量$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$;
检验水平对应分位数$z_{0.975}=1.96$,计算
有$|Z|=2.7/6.75*5=2>1.96$, 检验显著，否定原假设，有差异。

Quiz 小测验五

1. 设随机过程$Y(t) = e^{-tX}, t \in (-\infty, \infty)$，其中$X$是在$(0,1)$ 上服从均匀分布的随机变量，试求:（1）X的概率密度；
（2）求$EY(t), E(Y(t_1)Y(t_2)), E(Y(t)^2)$,
（3）问$Y(t)$是否为广义平稳过程？
2. 四个位置:1,2,3,4在圆周上逆时针排列.粒子在这四个位置上随机游动.粒子从任何一个位置,以概率$2/3$ 逆时针游动到相邻位置; 以概率$1/3$ 顺时针游动到相邻位置;以$X(n) = j$表示时刻n 粒子处在位置,
（1）求齐次马尔可夫链$X(n), n = 1, 2, \ldots$ 的状态空间；
（2）求一步转移概率矩阵;
（3）求平稳分布(p_1, p_2, p_3, p_4).

Quiz 小测验五：答案

1. （1）X的概率密度$f(x) = 1, 0 < x < 1, f(x) = 0$,其他情形。
（2）$EY(t) = (1 - e^{-t})/t; R(t_1, t_2) = (1 - e^{-(t_1+t_2)})/(t_1 + t_2)$, 特别$EY^2(t) = (1 - e^{-2t})/(2t)$.
（3）不是。因为$EY(t)$不是常数，且$R(t_1, t_2)$不是$t_1 - t_2$的函数。
2. （1）状态空间$= \{1, 2, 3, 4\}$
（2）一步转移概率矩阵$P = [0, 2/3, 0, 1/3; 1/3, 0, 2/3, 0; 0, 1/3, 0, 2/3; 2/3, 0, 1/3, 0]$;
（3）解四个方程加上$p_1 + p_2 + p_3 + p_4 = 1$有平稳分布$(p_1, p_2, p_3, p_4) = [1/4, 1/4, 1/4, 1/4]$.

随机游戏：Taking chances

1. 第一关：随机序列的奥秘；
2. 第二关：统计数字的分布；
3. 第三关：概率的大小；

beat the randomness! 打败随机性

- 第一关：随机序列的奥秘；
潘尼游戏：（1969年发明)
抛掷硬币多次直到出现一个特殊的(三位连续)组合为止。第一个玩家先从8个组合中任选一个，第二个玩家(庄家)再选一个不同组合。游戏中哪个组合先出现，那个人赢。
- 第二关：统计数字的分布；给定一组数据，请猜测其第一个有效数字(非零和小数点)最有可能是什么的两个选择。
2010年联合国人口署给出237个国家人口的估计数。
自然数$1/n, n = 1, 2, \ldots, 200$
以6出发，每次乘以2，得到100个数字

第三关：概率的大小

请猜测有关的概率或平均值大小：

两副牌放在桌子上，54个人从每副牌中各选一张，平均多少人拿到一对完全相同的牌？ $EX = 1$

已知一个家庭有两个小孩，已知其中一个是男孩，问他有一个兄弟的概率是多少？假定生男或生女的概率一样。

$P = 1/3$

设四个人打牌(52张)，每人得到13张扑克牌。如果已知你对家有一个A,则他还有一个A的概率为p_1; 如果已知你对家有一个黑桃A，他还有一个A的概率为p_2，问p_1, p_2谁大？不考虑你的和其他人的牌。

$p_1 \sim 0.37, p2 \sim 0.56$

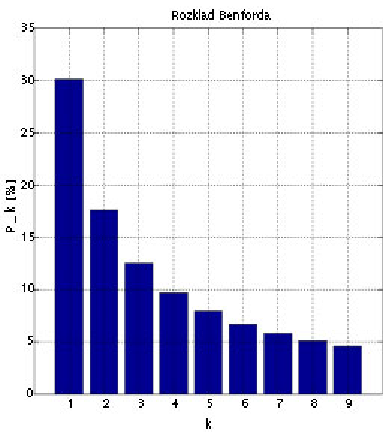
潘尼游戏：Penney's game

潘尼游戏：（1969年发明) 抛掷硬币多次直到出现一个特殊的(三位连续)组合为止。第一个玩家先从8个组合中任选一个，第二个玩家(庄家)再选一个不同组合。游戏中哪个组合先出现，那个人赢。

第一个选择	第二个选择	赔率(庄家)
HHH	THH	7 to 1
HHT	THH	3 to 1
HTH	HHT	2 to 1
HTT	HHT	2 to 1
THH	TTH	2 to 1
THT	TTH	2 to 1
TTH	HTT	3 to 1
TTT	HTT	7 to 1

诀窍：$ABC \to \bar{B}AB$.

本福德规律Benford's law



本福德规律(1938).

大多数真实数据的第一个有效数字不是平均分布的! (服从对数法则)。

世界60最高建筑的高度，人口，河流长度，股票价格，。。。

应用：审计财务报表!税务局查税!

Taking chances

1. "Two roads diverged in a wood, and I... I took the one less traveled by, and that has made all the difference." Robert Frost
2. "Twenty years from now you will be more disappointed by the things you didn't do than by the ones you did. So throw off the bowlines, sail away from the safe harbor, catch the trade winds in your sails. Explore. Dream. Discover." Mark Twain
3. 电影《死亡诗社》Dead poet society:
"Carpe diem, seize the day boys, make your lives extraordinary."