清華大學

1.2 条件概率与独立性

乘法公式：$P(AB)=P(B)P(A|B)$　$P(A_1\cdots A_n)=P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1\cdots A_{n-1})$

拿球问题：有放回/无放回.

全概率公式：$P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i)$

买彩票问题：结论 $\frac{1}{n}$ 机会均等 证明：首计k分析法（设 A_k 以为n个人中的 k个人中奖，求递推公式）

Bayes公式：
$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum P(B_j)P(A|B_j)}$$

独立性 $P(AB)=P(A)P(B)$

相关系数（一种特例）$r(A,B)=\dfrac{P(AB)-P(A)P(B)}{\sqrt{P(A)(1-P(A))P(B)(1-P(B))}}$

1.4 随机变量

离散 vs 连续：
- 取值有限 or 可列多个 ⇒ 离散
- 存在密度函数 ⇒ 连续
- 既不连续也不离散. 广泛存在.

常见随机变量.　　　负二项和泊松都满足可加性.

分布	分布列 or 密度函数	期望	方差	实际意义 && 和其它分布关系 && 性质
二项 $b(n,p)$	$P_k=C_n^k p^k(1-p)^{n-k}$　$k=0,1\cdots n$	np	$np(1-p)$	
负二项 $Nb(r,p)$	$P_k=C_{k-1}^{r-1}(1-p)^{k-r}p^r$　$k=r,r+1\cdots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	第r次成功时总次数，可以和$\boxed{\text{几何分布}}$关联（都是成功次数）
泊松分布 $P(\lambda)$	$P_k=\frac{\lambda^k}{k!}e^{-\lambda}$　$k=0,1,2\cdots$	λ	λ	λ实际也定义为"均分成很多小段后，平均在每段出现事件次数 n重伯努利试验中 $np_n\to\lambda$ 则 $b(n,p)\to P(np)=P(\lambda)$
几何分布 $Ge(p)$	$P_k=(1-p)^{k-1}p$　$k=1,2\cdots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	无记忆性（重要）
正态分布 $N(\mu,\sigma^2)$	$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$　$x\in(-\infty,+\infty)$	μ	σ	标准化 $\frac{x-\mu}{\sigma}\sim N(0,1)$ 可加性
均匀 $U(a,b)$	$p(x)=\frac{1}{b-a}$　$a<x<b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	

指数 $Exp(\lambda)$	$p(x)=\lambda e^{-\lambda x}$ $(x\geq 0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	无记忆性(充要) 几何分布是

L5 期望与方差.

$$E(X)=\begin{cases}\sum x_i P(X=x_i)\\ \int_{-\infty}^{+\infty} x p(x)dx\end{cases} \qquad E(aX+b)=aE(X)+b \qquad E(X_1+\cdots X_n)=E(X_1)+\cdots+E(X_n)$$

$$Var(X)=E\left((X-E(X))^2\right) \qquad Var(aX+b)^2=a^2 Var X \qquad Var(X)=E(X^2)-E^2(X)$$

期望和方差可能不存在 — Cauchy分布.

性质: ① $E(X)$ 是 $E(X-C)^2$ 取min时的C (中位数是 $E(|X-C|)$ min时的C)

② 切比雪夫不等式 $\forall \varepsilon>0$ $P(|X-EX|\geq\varepsilon)\leq\frac{Var(X)}{\varepsilon^2}$ 切比雪夫估计离散变量时一般不依 0.5倍付 而且估计都是大于等的概率.

③ X、Y独立, 则 $E(XY)=E(X)E(Y)$ $Var(X+Y)=Var X+Var Y$.

L6 多元随机变量

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} p(u,v)dudv \qquad p(x,y)=\frac{\partial^2}{\partial x\partial y}F(x,y)$$

边际分布 $F_X(x)=\int_{-\infty}^{x}\left(\int_{-\infty}^{+\infty}p(u,v)dv\right)du$

边际密度 $p_X(x)=\int_{-\infty}^{+\infty}p(x,y)dy$

独立性: $P(X_1=x_1,\ldots X_n=x_n)=P(X_1=x_1)P(X_2=x_2)\cdots P(X_n=x_n)$

$$p(x_1\cdots x_n)=\prod_{i=1}^{n}p_{X_i}(x_i) \qquad \Leftrightarrow \qquad F(x_1\cdots x_n)=\prod_{i=1}^{n}F_{X_i}(x_i)$$

验证是否独立: ① 把边缘密度求出来看乘积是否为联合密度

② 联合密度能拆成 $f_{X_1}(x)f_{X_2}(x)$ 的一定独立.

多元随机变量求方差: 多元积分. 利用对称性 $E\left((X-Y)^2\right)=Var(X-Y)$ 当 $E(X-Y)=0$ 时.

$\frac{1}{n}\ 1+\frac{1}{n}2+\cdots+\frac{1}{n}(n-1)\ \cdots\ \frac{1}{n}(1+\cdots)$ $\quad 1+\frac{(n+2)(n-1)}{2}$ $\quad \frac{n^2+n}{2}$

$\frac{(n+2)(n-1)}{2n}+\frac{1}{n}=\frac{n+1}{n}Ex$

清華大學

L7 协方差与相关系数

$Cov(X,Y)=E(XY)-EXEY=E[(X-EX)(Y-EY)]$ 双线性: $Cov(C_1X+a, C_2Y_1+C_3Y_2+b)=C_1C_2Cov(X,Y_1)+C_1C_3Cov(X,Y_2)$

和方差关系: $Var(X+Y)=VarX+VarY+\boxed{2}Cov(X,Y)$

$[Cov(X,Y)]^2\leq VarX\,VarY$ 证明用Cauchy方法, 设 $E[(X-EX)^2]+2tE[(X-EX)(Y-EY)]+t^2E[(Y-EY)^2]=g(t)$ under判别式

$Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{VarX}\sqrt{VarY}}\in[-1,1]$ 不相关 $\Leftarrow$ 独立

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

二元正态分布 $p(x,y)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right\}$ 二元来说 $n=2$ $\Sigma=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$

$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$

二元正态 $\Leftarrow$ 边缘是一元正态 (反之不成立, 如 W $\frac{-1\ |\ 1}{\frac{1}{2}\ |\ \frac{1}{2}}$ $X\sim N(0,1)$ $Y=WX\sim N(0,1)$ X,Y不相关且不独立.

二元正态下 独立 $\Leftrightarrow$ 不相关 (正态分布间仅存线性关系).

条件分布: $X\sim P(\lambda_1)$ $Y\sim P(\lambda_2)$ 则 $X|X+Y\sim b(n,\frac{\lambda_1}{\lambda_1+\lambda_2})$ $X+Y\sim P(\lambda_1+\lambda_2)$

$p_{X|Y}(x|y)=\frac{p(x,y)}{p_Y(y)}$

$\frac{\lambda_2^{n-k}\lambda_1^k}{(n-k)!k!}e^{-\lambda_1-\lambda_2}$

$\frac{1}{n!}\sum_{i=0}^{n}\frac{n!\lambda_1^i\lambda_2^{n-i}}{i!(n-i)!}e^{-\lambda_1-\lambda_2}$ $\frac{\lambda_1^k}{k!}$.

$(\lambda_1+\lambda_2)^n\frac{1}{n!}\sum_{i=0}^{n}C_n^i\frac{\lambda_1^i\lambda_2^{n-i}}{(\lambda_1+\lambda_2)^n}e^{-\lambda_2}$

$\frac{(\lambda_1+\lambda_2)^n}{n!}e^{-\lambda_2}$

L8

条件期望与随机变量函数.

$E(X|Y)=g(Y)$

离散型 $E(X|Y)\sim\frac{E(X|Y=y_1)}{P(Y=y_1)}\ \cdots\ \frac{E(X|Y=y_n)}{P(Y=y_n)}$ $\Leftarrow$ $Y=y_n$下X期望 $\Leftarrow$ $Y=y_n$的概率.

连续型 $Z=E(X|Y)\sim$ $Z=E(X|Y=y)$时 $p_Y(y)\to p_Z(z)$

重期望 $E(X)=E(E(X|Y))$ $\Rightarrow$ 应用: 递归方法求期望 方差.

随机变量函数

方法一：先求F再求p. ⇒ 注意定义域、分类讨论.

方法二：对连续性型用求导直接转换

$$p_Y(y_1\cdots y_n)=\begin{cases}\sum_s p_X(x_1^{(s)}(y_1\cdots y_n)\cdots x_n^{(s)}(y_1\cdots y_n))\cdot|J^{(s)}| \\ 0\end{cases}\quad J^{(s)}=\frac{\partial x_1^{(s)}\cdots x_n^{(s)}}{\partial y_1\cdots y_n}$$

当$(y_1\cdots y_n)$ s.t. $y=g_i(x)$ 有解时

卷积公式：$(X,Y)\sim p_{XY}(x,y)$ $Z=X+Y$ 满足 $p_Z(z)=\int_{-\infty}^{+\infty}p_{XY}(z-w,w)\,dw$

次序统计量 $p_k(x)=\frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}p(x)(1-F(x))^{n-k}$ 如：$Z=\min\{X_1\cdots X_n\}\sim Exp(\lambda_1+\cdots+\lambda_n)$ $X_i\sim Exp(\lambda_i)$

L9. 大数定律

LLN：伯努利：S_A为n次伯努利试验中A发生次数，p为每次A发生概率，则 $\forall\varepsilon>0$ $\lim_{n\to\infty}P(|\frac{S_A}{n}-p|<\varepsilon)=1$

切比雪夫：$X_1\cdots X_k\cdots$两两不相关，方差有界，$\forall\varepsilon>0$ 有 $\lim_{n\to\infty}P(|\frac{1}{n}\sum_{k=1}^n X_k-\frac{1}{n}\sum_{k=1}^n E(X_k)|<\varepsilon)=1$

依概率收敛 $\{X_n\}$为随机变量序列，若 $\forall\varepsilon>0$ $\lim_{n\to\infty}P(|X_n-X|<\varepsilon)=1$ 则 $X_n\xrightarrow{P}X$

马尔可夫条件 $\frac{1}{n^2}Var(\sum_{k=1}^n X_k)\to0$ 满足条件即满足LLN一般形式

CLT：$\{X_n\}$独立同分布 $E(X_1)=\mu$ $Var(X_1)=\sigma^2$，则$\forall y$ 有 $\lim_{n\to\infty}P(\frac{X_1+\cdots+X_n-n\mu}{\sqrt{n}\sigma}\le y)=\Phi(y)$

i.e. $\frac{X_1+\cdots+X_n-n\mu}{\sqrt{n}\sigma}\to N(0,1)$ 更一般 $X_1+\cdots+X_n\sim N(\sum_{k=1}^n E(X_k),\sum_{k=1}^n Var(X_k))$

0.5校正：对随机变量(离散取整数)的问题，一般最好用+

Stat

L1 统计量

$\bar{X}=\frac{X_1+\cdots+X_n}{n}$ $S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2$ $\bar{X},S^2$是对 $E(X)$ $Var(X)$ 的无偏估计

$Y=X_1^2+\cdots+X_n^2\sim\chi^2(n)$ $X_1\cdots X_n\sim N(0,1)$ i.i.d

$T=\frac{X_1}{\sqrt{\frac{Z_1^2+\cdots+Z_n^2}{n}}}\sim t(n)$ $X_1,Z_1\cdots Z_n\sim N(0,1)$ i.i.d

$F=\frac{X_1/m}{X_2/n}\sim F(m,n)$ 没有开根. $X_1\sim\chi^2(m)$ $X_2\sim\chi^2(n)$ X_1X_2独立.

统计抽样定理 $X_1\cdots X_n\sim N(0,\mu,\sigma)$ i.i.d.

$\bar{X}$ S^2独立

$\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$

$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$

清華大學

2 点估计

矩估计：用样本矩代替理论矩 $\bar{X} \to E(X)$ $S^2 \to Var(X)$

一般先列出 $E(X)$, $Var(X)$ 关于 $\vec{\theta}$ 的表达式，然后用 $\bar{X}$, S^2 代替后反解 $\vec{\theta}$

极大似然 MLE：$L(\theta) = L(\theta; x_1 \cdots x_n) = \prod_{k=1}^{n} p(x_k | \theta)$ 称为似然函数 $l(\theta) = \ln(L(\theta))$ 为对数似然.

$\hat{\theta} = \hat{\theta}(x_1 \cdots x_n)$ s.t. $L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$ 则 $\hat{\theta}$ 为 θ 极大似然估计

求 $\hat{\theta}$：①求导 ② $\boxed{\text{取值范围}}$ ☆

点估计评价指标：①相合性 ②无偏性 ③有效性 $Var(\hat{\theta}_1) \le Var(\hat{\theta}_2)$ 且至少一个 θ s.t. 不等号成立 ⇐ 增大样本量

正态分布矩估计与 MLE：
$$\begin{cases} \text{矩估计} & \hat{u} = \bar{X},\ \hat{\theta} = S^2 \\ \text{MLE} & \hat{u}_{MLE} = \bar{X} \quad \hat{\theta}_{MLE} = \frac{n-1}{n} S^2 \end{cases}$$

4. 区间估计

置信水平：随机区间至少以 $1-\alpha$ 覆盖 置信系数

枢轴量法：构造 $T = g(\vec{X}, \vec{\theta})$ T分布与θ无关 选取 c d 满足 $P(c \le T \le d) \ge 1-\alpha$ 再求解.

$X \sim N(u, 6^2)$ 求 u
$$\begin{cases} 6\text{已知} & \frac{\sqrt{n}(\bar{X}-u)}{6} \sim N(0,1) \\ 6\text{未知} & \frac{\bar{X}-u}{6/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{6^2}} = \frac{\sqrt{n}(\bar{X}-u)}{6} \sim t(n-1) \end{cases}$$

双总体下.

$X \sim N(u_1, 6_1^2)$ $Y \sim N(u_2, 6_2^2)$, $6_1^2 = 6_2^2 = 6^2$ 未知时 $t = \sqrt{\frac{mn(m+n-2)}{m+n}} \frac{\bar{X} - \bar{Y} - (u_1 - u_2)}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}} \sim t(m+n-2)$ 可用于估计 $u_1 - u_2$.

方差估计：$\frac{(n-1)S^2}{6^2} \sim \chi^2(n-1)$ 作为枢轴量.

估计区间长度：不给出下可以用均值不等式优化.

双总体下

$F = \frac{S_1^2/6_1^2}{S_2^2/6_2^2} \sim F(m-1, n-1)$

L4 假设检验

$H_0: \theta\in\Theta_0 \leftrightarrow H_1: \theta\in\Theta_1$，W 拒绝域　显著性水平 $\alpha = \max P(拒H_0 \mid H_0为真)$

第一类错误：拒真　第二类：受伪　α起到限制第一类错误，保护原假设作用.

U检验：σ^2已知　$H_0: \mu=\mu_0$　$H_1: \mu\neq\mu_0$　$W=\{\bar{x} \mid |\bar{x}-\mu_0| > \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\}$

$H_0: \mu\geq\mu_0$　$H_1: \mu<\mu_0$　$W=\{\bar{x} \mid \bar{x}-\mu_0 < -\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\}$

T检验：σ^2未知：$H_0: \mu=\mu_0$　$H_1: \mu\neq\mu_0$　$W=\{\bar{x} \mid |\bar{x}-\mu_0| > \frac{\sigma}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\}$

$H_0: \mu\geq\mu_0$　$H_1: \mu<\mu_0$　$W=\{\bar{x} \mid \bar{x}-\mu_0 < -\frac{\sigma}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\}$

P值：原假设成立条件下，检验统计量比观测值更异常的概率

雨课堂：

快排：X_{ij}表示 $y_i\ y_j$ 是否被比较，$X_{ij}=1 \Leftrightarrow y_i\ y_j$ 是第一次比较的（对于 i,j 之间的来说）

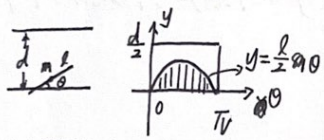
$P(X_{ij}=1)=\frac{2}{j-i+1}$　比较次数 $= E(\sum_{i=1}^{n}\sum_{j=i+1}^{n}X_{ij}) = \sum\sum E(X_{ij}) \sim n\ln n + o(n)$

二元正态分布：$X,Y \sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$　$(aX+bY, cX+dY)$ 也服从二元正态.

$\Rightarrow$ 构造线性组合，用 $\mathrm{Cov}(X',Y')=0$ 推 X',Y' 独立.

伪随机数生成：$\forall x\in R$　$F(x)$单增连续，且 $F(-\infty)=0<F(x)<1=F(+\infty)$　则 $X=F^{-1}(U)$　$(U\sim U(0,1))$ 的概率分布函数为 $F(x)$

布丰投针问题.



$P=\frac{\int_0^{\pi}\frac{l}{2}\sin\theta\, d\theta}{\frac{d}{2}\pi}=\frac{2l}{\pi d}$

贝特朗奇论：弦长大于…的概率：分析1—弦中心唯一确定 $\frac{1}{4}$　分析2—固定一个端点，在圆弧上选取另一个端点 $\frac{1}{3}$

分析3—选半径 $\frac{1}{2}$　等可能假设不同.

Monty Hall问题：换门：$\frac{2}{3}$　不换：$\frac{1}{3}$

康托尔分布：既非连续也非离散

清華大學

L_5 拟合优度检验.

$$X=\sum_{i=1}^{n}\frac{(N_i-Np_i)^2}{Np_i} \sim \chi^2(k-s-1)$$ s为用样本估计的参数数, k为取值个数.

独立性检验:

$$Y=\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{\left(n_{ij}-n\cdot\frac{c_i}{n}\frac{d_j}{n}\right)^2}{n\cdot\frac{c_i}{n}\cdot\frac{d_j}{n}} \sim \chi^2((s-1)(t-1))$$

A\B	1, 2, ⋯ t	行合计
1	n_{11} ⋯ n_{1t}	c_1
2	⋮ ⋮	⋮
⋮	⋮ ⋮	⋮
s	n_{s1} ⋯ n_{st}	c_s
列合计	d_1 ⋯ d_t	n