概率论与数理统计

> 整理重邮常考的知识点

独立、相关、互斥、概率的计算

- 独立
  如果事件A,B相互独立，有$P(AB)=P(A)P(B)$, $E(AB)=E(A)E(B)$
- 互斥
  如果事件A,B互斥，有$P(A+B)=P(A)+P(B)$,即$P(AB)=0$

> 独立与互斥没有任何关系

- 不相关
  如果事件A，B不相关，有$Cov(A,B)=0$

> 独立可以推出不相关，不相关不能推出独立

- 概率中常用的计算公式
  - 和事件的概率

$$P(A+B)=P(A)+P(B)-P(AB)$$

  - 德摩根律

$$P(\overline{AB})=1-P(A+B)$$

  - 概率拆分

$$P(AB)=P(A(1-P(\overline{B})))=P(A)-P(A\overline{B})$$

条件概率

- 乘法公式
  在A的条件下，B发生的概率：

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

  满足性质：

$$P(B|A)=1-P(\overline{B}|A)$$

- 全概率公式
  将A事件（条件）划分为多个事件A_i,那么事件B发生的概率：

$$P(B)=\Sigma P(A_i)P(B|A_i)$$

- 贝叶斯公式
  全概率公式的逆公式，表示已知B事件发生的概率下，在A中某一个划分下发生的概率：

$$P(A_i|B)=\frac{P(A_i)P(B|A_i)}{P(B)}=\frac{P(A_i)P(B|A_i)}{\Sigma P(A_i)P(B|A_i)}$$

连续性随机变量的分布

- 概率密度函数
  表示连续性随机变量在数轴上分布的稠密程度,有

$$\int_{-\infty}^{\infty}f(x)dx=1$$

  性质：

$$P(a\le x\le b)=\int_a^b f(x)dx$$

- 概率分布函数
  表示连续性随机变量在数轴左端分布的概率情况，即$P(X\le x)$的概率

$$F(x)=\int_{-\infty}^{x}f(x)dx$$

> 分段的概率分布函数不要忘记了还要加上前一段的概率

  性质：

$$P(a\le x\le b)=F(b)-F(a)$$

$$P(x>a)=1-F(a)$$

- 正态分布
  分布函数特性:

  1. $\Phi(a)=P(x\le a)$，$P(x>a)=1-\Phi(a)$
  2. $\Phi(0)=0.5$
  3. $\Phi(-a)=1-\Phi(a)$

  数值特征：

1. $X \sim N(\mu, \sigma^2), E(\overline{X}) = \mu, D(\overline{X}) = Cov(X, \overline{X}) = \frac{\sigma^2}{n}$
2. 见中心极限定理

- 随机变量之间的函数关系
  倘若随机变量X，Y之间存在某种函数关系$Y = g(X)$,给定X的分布函数$F_x(x)$,求Y的概率密度函数

$$F_y(y) = P(Y \leq y) = P(g(x) \leq y) = P(x \leq h(y)) = F_X(h(y))$$

$$f_Y(y) = F_y'(y) = F_x'(h(y))h'(y) = f_x(h(y))h'(y)$$

二元随机变量

- 二元连续性随机变量的分布函数

$$F(x, y) = \iint f(x, y)dA = P(A)$$

  其中A是由x,y围成的区域，对应x和y的一组规划

- 边缘分布概率密度和分布函数
  对于$F(x, y)$，

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y)dx$$

$$F_x(x) = F(x, \infty) = \int_{\infty}^{x} f_x(x)dx$$

$$F_y(y) = F(\infty, y) = \int_{\infty}^{y} f_y(y)dy$$

- 条件分布函数
  X在Y=y条件下的概率密度：

$$f_{X|Y}(x, y) = \frac{f(x, y)}{f_Y(y)}$$

- 二元随机变量的分布函数

  - Z=X+Y

$$f_z(z) = \int f(z - y, y)dy = \int f(x, z - x)dx$$

> 注意需要考虑$Z - Y$的取值范围，必要的时候要对X，$Z - Y$两者的取值范围大小进行分类讨论，始终取最小的区间

  - Z=max{X,Y}

$$F_z(z) = F_x(z)F_y(z)$$

- Z=min{X,Y}

$$F_z(z) = 1 - [(1 - F_x(x))(1 - F_y(y))]$$

统计特征

- 数学期望
  离散型：略
  连续型：

$$E(x) = \int xf(x)dx$$

$$E(g(x)) = \iint g(x)f(x,y)dA$$

  性质：

  - $E(X+Y) = E(X) + E(Y)$
  - X,Y相互独立时：$E(XY) = E(X)E(Y)$

- 方差

$$D(x) = E(x^2) - [E(x)]^2$$

$$D(x) = \Sigma[E(x - E^2(x))]$$

  性质:

  - $D(X \pm Y) = D(X) + D(Y) \pm 2Cov(X,Y)$

  > 当X,Y独立的时候才有$D(X \pm Y) = D(X) + D(Y)$，在使用公式之前一定要注意X,Y是否是独立的

  - $D(Cx) = c^2D(x)$

- 协方差和相关系数

  - 协方差

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

  协方差反映了两个随机变量的相关性，如果$Cov(X,Y) = 0$，则X,Y不相关。
  相关系数是标准化的协方差:

$$\rho_{(X,Y)} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

- 性质
  $Cov(X,X)=D(X)$
  $Cov(aX+bY,cX+dY)=acD(X)+(ad+bc)Cov(X,Y)+bdD(Y)$

- 方差矩阵
  - 原点矩
    指 $E(X^k)$
  - 中心距
    指 $E[(X-E(X))^k]$

中心极限定理

中心极限定理的使用条件都是n比较大的时候。

- 李雅普诺夫定理
  对于n个独立的随机变量$X_1,X_2,...X_n$,当n以概率趋近于无穷时，他们的和服从正态分布。

$$\frac{\Sigma X_i-\Sigma\mu_i}{\sqrt{\Sigma\sigma^2}}\sim N(0,1)$$

  当n个随机变量都是正态分布的时候，他们的和也服从正态分布，即：

$$\frac{\Sigma X_i-n\mu}{\sigma\sqrt{n}}\sim N(0,1)$$

- 棣莫弗-拉普拉斯定理
  当n足够大时，二项分布可视为正态分布。
  若$X\sim B(n,p)$，
  有$E(X)=np$,$D(x)=np(1-p)$ 。
  那么可以标准化X，有：

$$\frac{X-E(X)}{\sqrt{D(X)}}\sim N(0,1)$$

> 二项分布的极限可以是正态分布，也可以是泊松分布，优先选择泊松分布。

统计样本的数值特征和分布

- 样本的数值特征
  - 样本均值

$$\overline{X}=\frac{1}{n}\Sigma x_i$$

  - 样本方差

$$S^2=\frac{1}{n-1}\Sigma(x_i-\overline{x})^2$$

> n-1

- 样本的分布
    - χ^2分布
      如果样本$X_1,X_2,...X_n$ 服从正态分布，那么$\chi^2=\Sigma x_i^2$服从自由度为n的χ^2分布。
      χ^2统计量：
      $$\chi^2=\frac{(n-1)S^2}{\sigma^2}$$
      分位点：
      双侧α分位点：$\chi^2_{1-\frac{\alpha}{2}}$,$\chi^2_{\frac{\alpha}{2}}$
      单侧α分位点：$\chi^2_{1-\alpha}$，χ^2_{α}
    - t分布
      如果样本$X\sim N(0,1)$,且$Y\sim\chi^2_n$, X,Y独立，那么$t=\frac{X}{\sqrt{\frac{Y}{n}}}$服从自由度为n的t分布。
      t统计量：
      $$t=\frac{x-\overline{\mu}}{S}$$
      分位点：
      双侧α分位点：$t_{\frac{\alpha}{2}}$,$t_{-\frac{\alpha}{2}}$
      单侧α分位点：t_{α},$-t_{\alpha}$
    - F分布
      设$U\sim\chi^2(n_1)$,$V\sim\chi^2(n_2)$,U,V相互独立，则称$F=\frac{\frac{U}{N_1}}{\frac{V}{N_2}}$服从自由度为$(n_1,n_2)$的F分布。

样本的点估计法

- 矩估计法
  设X的概率密度函数为$f(x:\theta_1,...,\theta_k)$，用样本1~k阶矩（$E(x^k)$）代替总体1~k阶矩建立k个方程，联立求解，结果是含有A的式子。
    - 结论
      $$\mu=\overline{x},\quad \sigma^2=B_2=A_2-A_1^2$$
- 最大似然估计法
  设X的概率密度函数为$f(x:\theta_1,...,\theta_k)$，通过如下方法找出参数的估计量：
    1. 求解似然函数$L(\theta)=\Pi f(x)$
    2. 对似然函数取对数$ln(L(\theta))=ln(\Pi f(x))$
    3. 求导或者是偏导数，使每一个结果都等于0：$\frac{\partial ln(L(\theta)}{\partial\theta}=\frac{\partial ln(\Pi f(x))}{\partial\theta}=0$
    4. 求解θ

> 特殊情况： 均匀分布估计a,b的值，需要设最大值最小值函数

- 统计量的选取
  - 无偏性
    如果θ的估计量 θ 的数学期望存在，且$E(\theta) = \theta$，称θ 是θ的无偏估计量。
  - 有效性
    如果θ的估计量 θ_1,θ_2的方差存在，且$D(\theta_1) < D(\theta_2)$,称$\theta_1$比$\theta_2$更有效。
  - 相合性*
    如果θ的估计量 θ 以概率趋近于θ，称θ 是θ的相合估计量。

区间估计

区间估计遵循以下方法：

1. 选取一个合适的统计量
2. 找到α分位点
3. 反解出关于参数的不等式
4. 代值求出不等式，即为置信区间

假设检验

假设检验遵循以下方法：

1. 提出原假设和备择假设，备择假设中不能有等于符号
2. 找到α分位点，根据备择假设的符号来判断是单边检验还是双边检验

   > 拒绝域的符号和备择假设的符号是相同的

3. 找到拒绝域
4. 计算统计量的值，并与分位点的值进行比对，看统计量是否落在了拒绝域

随机过程的数值特征和平稳性

- 数值特征
  均值函数：$\mu_x(t) = E(X(t))$
  例如：若 $X(t) = At + B$,那么$\mu_x(t) = E(X(t)) = tE(A) + E(B)$
  自相关函数：$R_{xx}(t_1, t_2) = E[X(t_1).X(t_2)]$
  方差函数：$D_x(t) = R_{xx}(t, t) - [\mu_x(t)]^2$
- 平稳性
  验证随机过程是否具有宽平稳性需要有三步：

1. 验证该过程是否是二阶矩过程*
2. 求均值函数是否是一个参数
3. 求自相关函数$R(t, t+\tau)$ 是否只与τ有关

马尔科夫链

- 概率转移矩阵
  表示状态由现态转移到n次态的概率 $P = [现态（竖）\backslash次态（横）]$

$$P(n) = p^n$$

p表示一步转移矩阵，$P(n)$为n步转移矩阵

- 转移概率

$$P_{ij}(m, m+n) = P(X_{m+n} = a_j | X_m = a_i)$$

表示马尔科夫链在时刻m，状态a_i下在时刻m+n下转移到状态a_j的概率

  - 利用全概率公式可以推出$P(X_{m+n})$的概率
  - 利用乘法公式可以推出$P(X_m, X_{m+n})$的概率