随机变量的数字特征

期望

X~P(λ) , 则 E(X) = λ;

指数分布，E(X)=1/λ;

超几何分布，E(X)=nM/N

期望相加不需要求边缘函数，也不需要相互独立；期望相乘需要

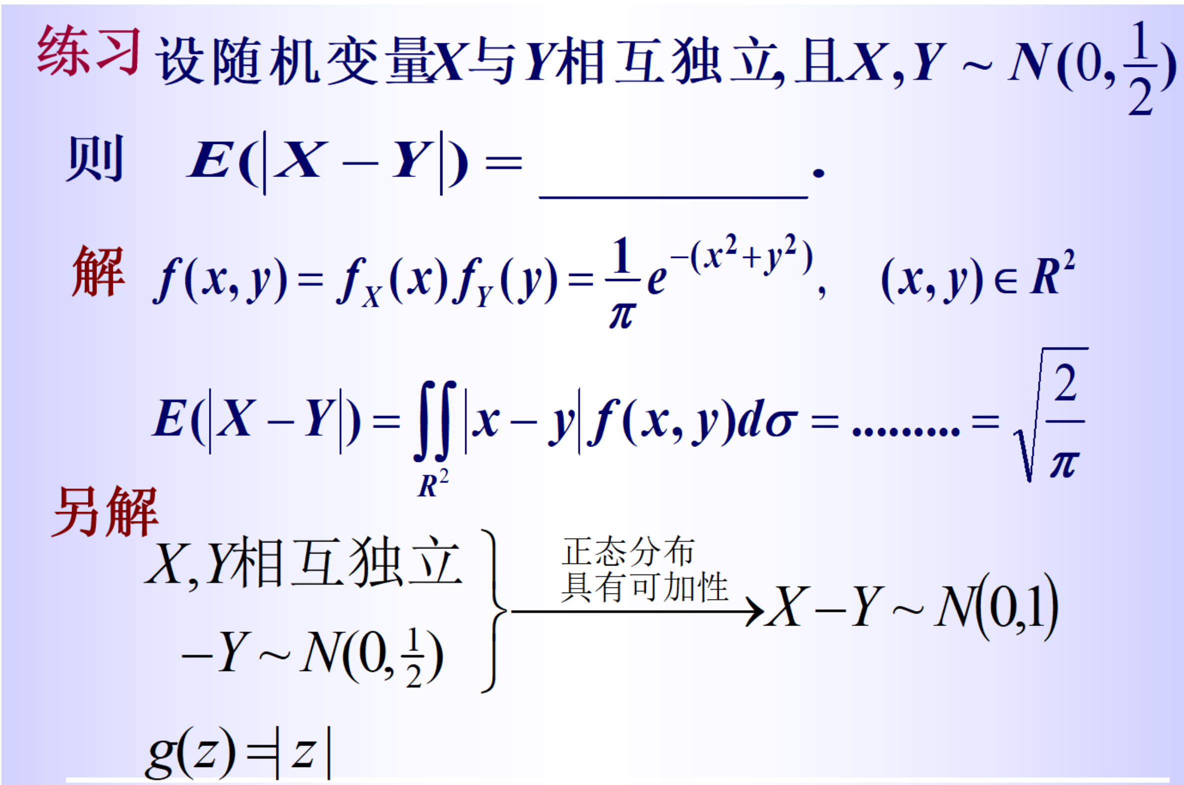
T：求期望

1.定义（X出现正负交替的时候注意可能用定义式能求出值，但是不是绝对收敛，期望不存在）

PS：变形成已知的分布，比如利用抽样分布定理构造凯方分布，利用凯方的方差和均值

2.函数变形式，一重二重

PS：Y=g(x)，无法写出具体的表达式，比如极值函数，就用E(y)的定义式，直接求f(y)，不通过f(x)，这里也可以利用分布的可加性化简，比如正态分可加，就直接求函数的分布，不用用定义式算了



PS：求函数的期望的反向操作：待求的随机变量的概率密度很不好表示，能不能找到好求概率密度的随机变量，再找到关联函数

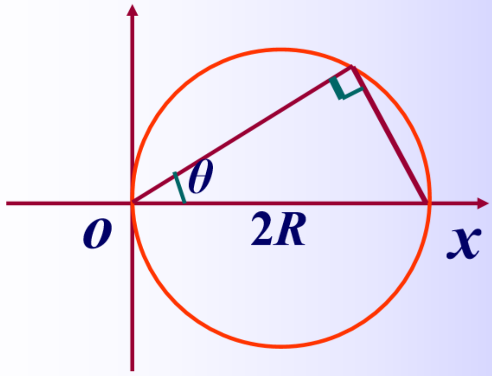
例4.1.3 过半径为R的圆周上的已知点，与圆周上的任意点相连，求这样得到的弦的平均长度.

解 以已知点为原点, 过已知点的直径为 x 轴正向，如图所示.

设弦与直径的夹角为θ

则 θ均匀分布于区间

$$\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$



设弦长为 L，则有

$$L = 2R\cos\theta$$

3.应用题里面利用性质，拆分成多个随机变量的和，注意是X=X1+X2+...+Xn，而不是根据X的取值范围分类（拆分+等可能性）

例4.1.13抛掷硬币直到出现k次正面为止，抛掷次数，第i-1次抛中到第i次抛中之间的次数是Xi，E(X)=kE(Xi)

例4.1.8 随机变量X 的分布为

$$P\{X=m\} = C_M^m C_{N-M}^{n-m} / C_N^n \quad m=0,1,2,\dots,n \quad n \le M \le N$$

试求 $E(X)$.

原始模型 N个球中有M个红球, 余下为白球, 从中任取n个球，n个球中的红球数为X。

分析： 1）直接求解很困难, 应利用数学期望的性质求解.

2）设想这n个球是逐个不放回抽取的，共取了n 次.

解 设想这n个球是逐个不放回抽取的,共取了n次.令X_i表示第i次取到红球的个数.$i=1,2,\ldots,n$

则 $X=X_1+X_2+\ldots+X_n$

从而 $E(X_i)=1\times M/N+0\times(1-M/N)=M/N$

$$E(X)=\sum_{i=1}^{n}E(X_i)=\frac{nM}{N}$$

有放回--二项分布，每次模球独立

无放回--超几何，每次不独立，但根据抽签的公平性，每次取到红球的概率是相等的

PS：分布可加性是在相互独立的基础之上的，因为推导时用到了概率随机变量的概率相乘。但期望相加不需要相互独立。

两种方式的每一次摸球都可以看作一重实验，两种抽签方式的概率都是M/N，所以求出的期望相同，但是由于每次实验之间的关系不同，分布律不同

方差

公式里的E(X)是一个常数

求期望和方差，注意当前情况下谁是随机变量，其他参数都可以当常量，提出来

协方差

相关系数

例4.3.4 设二维随机变量（X, Y）在矩形$G=\{(x, y)|0\le x\le 2，0\le y\le 1\}$上服从均匀分布．记

$$U=\begin{cases}0, & X\le Y;\\ 1, & X>Y.\end{cases}\qquad V=\begin{cases}0, & X\le 2Y;\\ 1, & X>2Y.\end{cases}$$

求ρ_{UV}.

三分之根号三

注意U，V是由X，Y的关系决定的离散型随机变量，用定义求E

T：求方差，协方差，相关系数

计算E，D的过程中注意熟练运用方差、协方差 、期望的性质，$E(X^2)$、D(X)之间，cov（X，Y）、E(XY)之间可以互推，能不用定义积分算的就尽量不用

add:

- 描述高阶的相关关系，模仿矩，k次方后再算相关系数
- 描述多变量之间的相关关系

协方差矩阵/相关矩阵

！ 正态分布的性质

联合分布是正态，才能由不相关推出相互独立

相互独立才可加，相互独立的正态分布的平方和才是凯方分布（所以S^2，凯方n-1）

4 单选 (3分) 设随机变量X和Y都服从正态分布，且它们不相关，则

A. $X+Y$服从一维正态分布

B. X和Y一定独立

C. X和Y不一定独立

D. (X,Y)一定服从二维正态分布

---

1 单选 (3分) 设随机变量X的分布函数为F（X）=$0.3\phi(x)+0.7\phi[(x-1)/2]$，其中$\phi(x)$为标准正态分布的分布函数，则E（X）=?

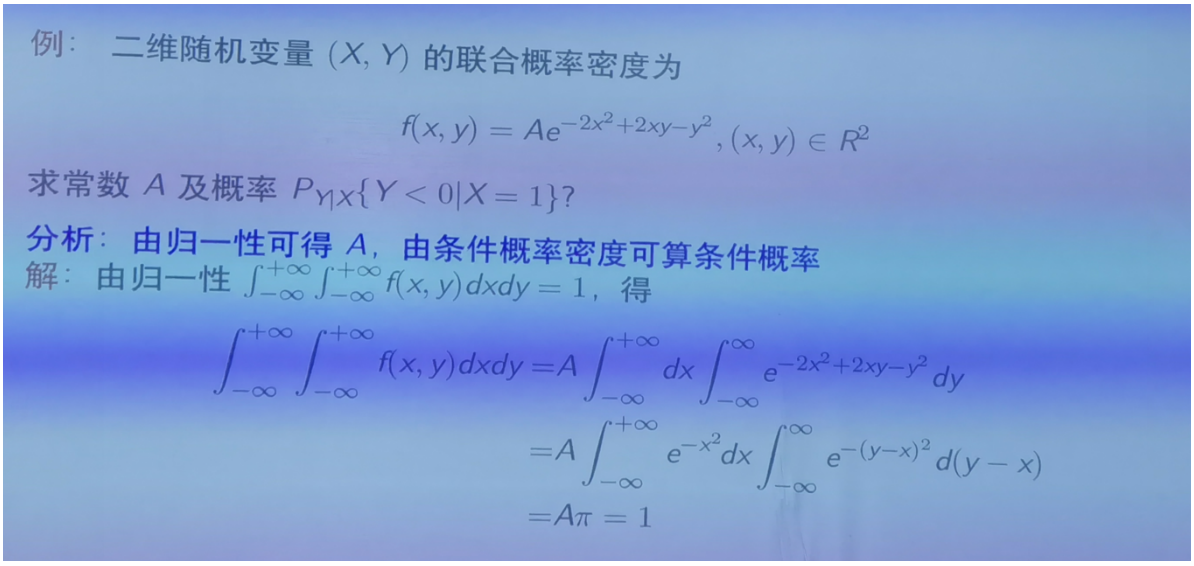
A. 0.7

B. 0.3

C. 1

D. 0

最好直接求概率密度再积分，也可以构造变量，注意U=1/2(V-1)，V~N(0,1)的分布函数不是第二个，非标准到标准是(x-μ)/σ，所以构造出的变量的均值是1，方差是4.



指数函数积分积不出来，配方凑成正态分布

PS：多元回归分析中的R就是y和y-hat之间的相关系数

$$
\begin{aligned}
\rho(y_i,\hat{y}_i) &= \frac{cov(y_i,\hat{y}_i)}{\sqrt{var(y_i)var(\hat{y}_i)}} \\
&= \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}} \\
&= \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}} \\
&= \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}} \\
&= \frac{0+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \\
&= \sqrt{R^2}
\end{aligned}
$$

R复相关系数：为了测定一个变量y与其他多个变量x1,x2,...xk之间的相关系数，可以考虑构造一个y关于x1,x2,...xk的线性组合，通过计算该线性组合（y拔）与y之间的简单相关系数作为变量y与x1,x2,...xk之间的复相关系数。

$$
R=corr(y,x_1,\cdot\cdot\cdot,x_p)=corr(y,\hat{y})=\frac{cov(y,\hat{y})}{\sqrt{var(y)var(\hat{y})}}
$$

决定系数R^2：

残差平方和，$SSR=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

总平方和 SST，$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$

$$SST=SSR+SSE$$

SSE= $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

决定系数

$$R^2\equiv 1-\frac{SS_{\rm res}}{SS_{\rm tot}}.$$

$$\overline{\hat{y}}=\frac{\sum_{i=1}^{n}\hat{y}_i}{n}=\frac{\sum_{i=1}^{n}y_i}{n}=\bar{y}$$

统计上的相关关系并不代表因果，有可能受到同一**共同潜在因子**的影响，但本质上是没有因果关系的。

赤池信息准则：

$$AIC=2k+n\ln(RSS/n)$$

k是参数的个数，AIC鼓励数据拟合的优良性但是尽量避免出现过度拟合（Overfitting）的情况

概率论的定理

泊松分布与指数分布的联系 P7 T18

指数分布是泊松过程的事件间隔的分布：泊松分布表示的是t时间内事件发生的次数的分布律，“次数”是离散变量随机变量的分布；指数分布是两件事情发生的平均间隔时间的分布，“时间”是连续随机变量的分布。

二项、泊松与正态分布的近似关系

- 二项分布什么时候趋近于泊松分布，什么时候趋近于正态分布？

  二项分布有两个参数，一个 n 表示试验次数，一个 p 表示一次试验成功概率。

  现在考虑一列二项分布，其中试验次数 n 无限增加，而 p 是 n 的函数。

如果 np 存在有限极限 λ，则这列二项分布就趋于参数为 λ 的 泊松分布。反之，如果 np 趋于无限大，则根据德莫佛-拉普拉斯(De'Moivre-Laplace)中心极限定理，这列二项分布将趋近于正态分布。

- 当λ>20时，工程上也认为泊松分布近似正态分布

所以可以看作np为定值且较小时（n极大，p极小）用泊松分布近似，np较大是用正态分布近似

独立同分布的随机变量序列，大数定理说明随机变量序列的前n项和依概率收敛于期望之和；中心极限定理说明依分布收敛于正态分布

T：中心极限定理估算概率

1.定义Xi，Yn=Xi的前n项之和，写出Xi的分布律

2.求E(Xi),D(Xi)

3.由独立同分布的中心极限定理，Y~N（）

（二项分布由demorve-laplace公式）

4.代入公式求P