数据科学与工程数学基础 作业5 _

2021年6月12日 上午

6.2k 字 52 分钟

一

> 求随机变量 $X \sim b(n,p)$ 的期望与方差。

由 $X \sim b(n,p)$ 可知

$$P(X=k)=\binom{n}{k}p^k(1-p)^{1-k}, k \in \{0,1,2,\cdots,n\}$$

故其期望为

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k(1-p)^{n-k} \\
&= np\sum_{k=1}^{n} \binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} \\
&= np\sum_{m=0}^{n-1} \binom{n-1}{m}p^{m}(1-p)^{n-1-m} \\
&= np \cdot (p+1-p)^{n-1} \\
&= np
\end{aligned}$$

又由

$$\begin{aligned}
E(X^2) &= \sum_{k=0}^{n} k^2\binom{n}{k}p^k(1-p)^{n-k} \\
&= n \cdot \sum_{k=1}^{n} k\binom{n-1}{k-1}p^k(1-p)^{n-k} \\
&= n \cdot \sum_{k=1}^{n} (k-1+1)\binom{n-1}{k-1}p^k(1-p)^{n-k} \\
&= n \cdot \left[\sum_{k=1}^{n} (k-1)\binom{n-1}{k-1}p^k(1-p)^{n-k} + \sum_{k=1}^{n} \binom{n-1}{k-1}p^k(1-p)^{n-k}\right] \\
&= n \cdot \sum_{k=2}^{n} (n-1)\binom{n-2}{k-2}p^k(1-p)^{n-k} + n \cdot \sum_{k=1}^{n} \binom{n-1}{k-1}p^k(1-p)^{n-k} \\
&= n(n-1)p^2 + np
\end{aligned}$$

可知其方差为

$$\begin{aligned}
Var(X) &= E(X^2) - E^2(X) \\
&= n(n-1)p^2 + np - n^2p^2 \\
&= n^2p^2 - np^2 + np - n^2p^2 \\
&= np(1-p)
\end{aligned}$$

二

> 设连续性随机变量 X 的分布函数为
>
> $$F_X(x)=\begin{cases}0, & x<1\\ \ln x, & 1\le x<e\\ 1, & x\ge e\end{cases}$$
>
> 1. 求 $P(X<2), P(0<X<3)$
> 2. 求概率密度函数 $f_X(x)$

1.

$$P(X<2)=F_X(2)=\ln 2$$
$$P(0<X<3)=F_X(3)-F_X(0)=1-0=1$$

2.

由

$$f_X(x)=\frac{d}{dx}F_X(x)$$

可知

当 $x<1$ 时，$f_X(x)=0$

当 $1<x<e$ 时，$f_X(x)=\frac{1}{x}$

当 $x>e$ 时，$f_X(x)=0$

又

$$F'_-(1)=\lim_{x\to1^-}\frac{f(1)-f(x)}{1-x}=0\neq 1=F'_+(1)$$
$$F'_-(e)=\lim_{x\to e^-}\frac{f(e)-f(x)}{e-x}=\frac{1}{e}\neq 0=F'_+(e)$$

故 $f_X(x)$ 在 $x=1$ 和 $x=e$ 处不存在

因此

$$f_X(x)=\begin{cases}0, & x<1\\ \frac{1}{x}, & 1<x<e\\ 0, & x>e\end{cases}$$

三

> 下表为二维离散随机变量 (X,Y) 的联合分布列，其中最后一列为随机变量 Y 的边缘分布列，最后一行为随机变量 X 的边缘分布列，且 X,Y 独立。试将下表补充完整，并给出 X,Y 的协方差 $\mathrm{Cov}(X,Y)$

	$X=1$	$X=2$	$X=3$	$P_Y(Y)$
$Y=1$	0.03	0.15	0.12	0.3
$Y=2$	0.03	0.15	0.12	0.3
$Y=3$	0.02	0.1	0.08	0.2
$Y=4$	0.02	0.1	0.08	0.2
$P_X(X)$	0.1	0.5	0.4	/

由于 X,Y 独立，故 $Cov(X,Y)=0$

四

> 已知所有的胰腺癌患者都有某症状，若一个人有该症状的概率为万分之一，并且胰腺癌的发病概率也为万分之一。问若一个人有该症状，则他也是胰腺癌患者的概率为多少。

设 $A=\{\text{有该症状}\}, B=\{\text{有胰腺癌}\}$

由于 $B\subset A$

故

$$P(B|A)=\frac{P(A\cap B)}{P(A)}=\frac{P(B)}{P(A)}=1$$

五

> 一个不透明的箱子中有一些红球和白球，有放回地在箱子中随机摸出5个球，分别为红、白、白、白、红，试估计箱子中红球与白球的比例。

设箱子中摸出红球的概率为 p，

$$X_i=\begin{cases}1, & \text{第}i\text{次摸出红球}\\ 0, & \text{第}i\text{次摸出白球}\end{cases}$$

故 $X_i\overset{i.i.d}{\sim}b(1,p)\ , i\in\{1,2,3,4,5\}$

于是

$$L(p)=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}$$

故

$$\frac{\partial \ln L(p)}{\partial p}=\frac{\sum\limits_{i=1}^{n} x_i}{p}-\frac{n-\sum\limits_{i=1}^{n} x_i}{1-p}$$

于是 p 的极大似然估计

$$\hat{p}=\bar{x}=\frac{1+0+0+0+1}{5}=\frac{2}{5}$$

故

$$\frac{\text{红球}}{\text{白球}}=\frac{2}{3}$$

六

随机地取8只活塞环，测得他们的直径为(以mm计)

```
74.001 74.005 74.003 74.001
74.000 73.998 74.006 74.002
```

试求总体均值 μ 以及方差 σ^2 的矩估计值。

$$\begin{aligned}\hat{\mu}&=\bar{x}\\&=\frac{74.001+74.005+74.003+74.001+74.000+73.998+74.006+74.002}{8}\\&=74.002\end{aligned}$$

$$\begin{aligned}\hat{\sigma^2}&=s^2\\&=\frac{1}{7}[(74.001-74.002)^2+(74.005-74.002)^2+(74.003-74.002)^2+(74.001-74.002)^2+(74.000-74.002)^2+(73.998-74.002)^2+(74.006-\\&\approx 6.8571\times 10^{-6}\end{aligned}$$

七

给定 N 个独立同分布样本 x_t，服从多元正态分布

$$G(x_t)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x_t-\mu)^T\Sigma^{-1}(x_t-\mu)\right\}$$

，其中 Σ 是可逆对称矩阵，$x_t,\mu\in\mathbb{R}^d$。利用极大似然估计(MLE)估计参数 μ,Σ。

似然函数

$$\begin{aligned}L(\mu,\Sigma)&=\prod_{i=1}^{n}\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}\\&=\frac{1}{(2\pi)^{\frac{nd}{2}}|\Sigma|^{\frac{n}{2}}}e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}\end{aligned}$$

故其对数似然函数

$$l(\mu,\Sigma)=-\frac{nd}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

由于

$$\begin{aligned}dl&=Tr\left[d\left(-\frac{nd}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)\right]\\&=Tr\left[-\frac{1}{2}\sum_{i=1}^{n}\left(d(x_i-\mu)^T\cdot\Sigma^{-1}\cdot(x_i-\mu)+(x_i-\mu)^T\cdot\Sigma^{-1}\cdot d(x_i-\mu)\right)\right]\\&=Tr\left[\frac{1}{2}\sum_{i=1}^{n}\left(d\mu^T\cdot\Sigma^{-1}\cdot(x_i-\mu)+(x_i-\mu)^T\cdot\Sigma^{-1}\cdot d\mu\right)\right]\\&=Tr\left[\frac{1}{2}\sum_{i=1}^{n}\left((x_i-\mu)^T\cdot\left(\Sigma^{-1}\right)^T\cdot d\mu+(x_i-\mu)^T\cdot\Sigma^{-1}\cdot d\mu\right)\right]\\&=Tr\left[\sum_{i=1}^{n}\left((x_i-\mu)^T\cdot\Sigma^{-1}\right)d\mu\right]\end{aligned}$$

故

$$\frac{\partial l}{\partial\mu}=\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu)$$

因此

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$$

又由于

$$\begin{aligned}dl&=Tr\left[-\frac{n}{2}d\ln|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\cdot d\Sigma^{-1}\cdot(x_i-\mu)\right]\\&=Tr\left[-\frac{n}{2|\Sigma|}|\Sigma|\Sigma^{-1}d\Sigma+\frac{1}{2}\sum_{i=1}^{n}\left((x_i-\mu)^T\cdot\Sigma^{-1}d\Sigma\cdot\Sigma^{-1}\cdot(x_i-\mu)\right)\right]\\&=Tr\left[-\frac{n}{2}\Sigma^{-1}d\Sigma\right]+\frac{1}{2}\sum_{i=1}^{n}Tr\left[\Sigma^{-1}\cdot(x_i-\mu)(x_i-\mu)^T\cdot\Sigma^{-1}d\Sigma\right]\\&=Tr\left[\left(-\frac{n}{2}\Sigma^{-1}+\frac{1}{2}\sum_{i=1}^{n}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)\right)d\Sigma\right]\end{aligned}$$

故

$$\frac{\partial l}{\partial \Sigma}=\frac{1}{2}\sum_{i=1}^{n}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)-\frac{n}{2}\Sigma^{-1}$$

因此

$$\begin{aligned}\hat{\Sigma}&=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\\&=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T\end{aligned}$$

八

> 证明：在多分类问题中，利用交叉熵函数作为损失函数和用KL散度作为损失函数是等价的。

对于多分类问题，若设 p_i 为第 i 个数据的目标输出，q_i 为第 i 个数据的实际输出，则

$$L_{CrossEntropy}=-\sum_{i=1}^{n}p_i\ln q_i$$
$$L_{KL}=\sum_{i=1}^{n}p_i\ln p_i-\sum_{i=1}^{n}p_i\ln q_i$$

二者仅相差一个与 q_i 无关的常数，即

$$\frac{\partial L_{CrossEntropy}}{\partial q_i}=\frac{\partial L_{KL}}{\partial q_i}=-\frac{p_i}{q_i}$$

故二者作为损失函数等价

九

> 同时抛2颗骰子，事件 A,B,C 分别表示为
>
> $$\begin{gathered}A:\text{仅有一个骰子是}3\\B:\text{至少一个骰子是}4\\C:\text{骰子上点数总和为偶数}\end{gathered}$$
>
> 试计算事件 A,B,C 发生后所提供的信息量

$$\begin{gathered}I_1=-\lg\frac{5}{18}\approx1.8480\\I_2=-\lg\frac{11}{36}\approx1.7105\\I_3=-\lg\frac{1}{2}=1\end{gathered}$$

数据科学数学基础 Mathematics DataScience

树莓派在线教学系统

操作系统实验 内存管理