概率论易错题、综合题、难题整理

1.注意多条件概率原条件一直在条件中，即

$$P(C|B)=P(C|AB)P(A|B)+P(C|\bar{A}B)P(\bar{A}|B)$$

易错点：容易写成

$$\boldsymbol{P(C|B)=P(C|A)P(A|B)+P(C|\bar{A})P(\bar{A}|B)}$$

这是错误的。

2.（1）若$\boldsymbol{X_i \sim Geo(\theta)}$且相互独立，求证：$\boldsymbol{\sum_{i=1}^{n} X_i \sim NB(n,\theta)}$。

（2）若$\boldsymbol{X_i \sim Exp(\lambda)}$且相互独立，求证：$\boldsymbol{\sum_{i=1}^{n} 2\lambda X_i \sim {\chi_{2n}}^2}$。

(3) 若$\boldsymbol{X_i \sim Geo\left(\frac{n-i+1}{n}\right)}$且相互独立，$\boldsymbol{\sum_{i=1}^{n} X_i}$的分布函数为$\boldsymbol{F(t)}$，求$\boldsymbol{\lim\limits_{t\to+\infty} F(t)}$。

解（1）几何分布的概率函数为

$$P(X_i=j_i)=\theta(1-\theta)^{j_i-1}$$

于是

$$P\left(\sum_{i=1}^{n} X_i=t\right)=\sum_{\sum_{i=1}^{n} j_i=t} P(X_i=j_i,X_2=j_2,\cdots,X_n=j_n)$$

$$=\sum_{\sum_{i=1}^{n} j_i=t} P(X_i=j_i)P(X_2=j_2)\cdots P(X_n=j_n)$$

$$=\sum_{\sum_{i=1}^{n} j_i=t} \theta^n(1-\theta)^{\sum_{i=1}^{n} j_i-n}=\sum_{\sum_{i=1}^{n} j_i=t} \theta^n(1-\theta)^{t-n}$$

由隔板法知满足$\sum_{i=1}^{n} j_i=t$的情况共有C_{t-1}^{n-1}种，故

$$P\left(\sum_{i=1}^{n} X_i=t\right)=C_{t-1}^{n-1}\theta^n(1-\theta)^{t-n}$$

故

$$\sum_{i=1}^{n} X_i \sim NB(n,\theta)$$

（2）令$Y_i=2\lambda X_i$，易知Y_i的概率密度函数为

$$f_i(y_i)=\frac{1}{2}e^{-\frac{y_i}{2}}I(y>0)$$

于是$Y_1,Y_2,\cdots,Y_n$的联合概率密度函数为

$$f(y_1,y_2,\cdots,y_n)=\frac{1}{2^n}e^{-\frac{\sum_{i=1}^{n} y_i}{2}}I(y_1,y_2,\cdots,y_n>0)$$

令$Z_i=Y_i(1\le i\le n-1),Z_n=\sum_{i=1}^{n}Y_i$，则$Z_1,Z_2,\cdots,Z_n$的联合概率密度函数为

$$g(z_1,z_2,\cdots,z_n)=f(y_1,y_2,\cdots,y_n)\left|\frac{\partial(y_1,y_2,\cdots,y_n)}{\partial(z_1,z_2,\cdots,z_n)}\right|$$

$$=\frac{1}{2^n}e^{-\frac{z_n}{2}}I(z_1,z_2,\cdots,z_{n-1}>0,z_1+z_2+\cdots+z_{n-1}<z_n)$$

在z_n固定时，区域

$$V=\{(z_1,z_2,\cdots,z_n)|z_1,z_2,\cdots,z_{n-1}>0,z_1+z_2+\cdots+z_{n-1}<z_n\}$$

是一个$n-1$维棱锥，其测度为$z_n^{n-1}/(n-1)!$，故z_n的概率密度函数为

$$h(z_n)=\frac{z_n^{n-1}}{2^n(n-1)!}e^{-\frac{z_n}{2}}I(z_n>0)$$

故

$$\sum_{i=1}^{n}2\lambda X_i=\sum_{i=1}^{n}Y_i=z_n\sim{\chi_{2n}}^2$$

（3）由题意知

$$P(X_i=j_i)=\frac{n-i+1}{n}\left(\frac{i-1}{n}\right)^{j_i-1}$$

$$P(X_1=j_1,X_2=j_2,\cdots,X_n=j_n)=\prod_{m=1}^{n}\frac{n-m+1}{n}\prod_{k=1}^{n}\left(\frac{k-1}{n}\right)^{j_k-1}$$

$$F(t)=P\left(\sum_{i=1}^{n}X_i\le t\right)=\prod_{m=1}^{n}\frac{n-m+1}{n}\prod_{k=1}^{n-1}\left(\frac{k-1}{n}\right)^{j_k-1}\sum\cdots\sum\left[\sum_{j_n=1}^{t-\sum_{i=1}^{n-1}j_i}\left(\frac{n-1}{n}\right)^{j_n-1}\right]$$

$$=n\prod_{m=1}^{n}\frac{n-m+1}{n}\prod_{k=1}^{n-2}\left(\frac{k-1}{n}\right)^{j_k-1}\sum\cdots\sum\left\{\sum_{j_{n-1}=1}^{t-\sum_{i=1}^{n-2}j_i}\left(\frac{n-2}{n}\right)^{j_{n-1}-1}\left[1-\left(\frac{n-1}{n}\right)^{t-\sum_{i=1}^{n-1}j_i}\right]\right\}$$

$$=n\prod_{m=1}^{n}\frac{n-m+1}{n}\prod_{k=1}^{n-3}\left(\frac{k-1}{n}\right)^{j_k-1}\sum\cdots\sum\left\{\sum_{j_{n-2}=1}^{t-\sum_{i=1}^{n-3}j_i}\left(\frac{n-3}{n}\right)^{j_{n-2}-1}\left[\frac{n}{2}-u_3(t,n,j_i)\right]\right\}$$

$$=n\prod_{m=1}^{n}\frac{n-m+1}{n}\prod_{k=1}^{n-4}\left(\frac{k-1}{n}\right)^{j_k-1}\sum\cdots\sum\left\{\sum_{j_{n-3}=1}^{t-\sum_{i=1}^{n-4}j_i}\left(\frac{n-4}{n}\right)^{j_{n-3}-1}\left[\frac{n}{2}\times\frac{n}{3}-u_4(t,n,j_i)\right]\right\}$$

$$=\cdots\cdots\cdots\cdots\cdots\cdots$$

$$=\prod_{m=1}^{n}\frac{n-m+1}{n}\left[\prod_{k=1}^{n}\frac{n}{k}-u_n(t,n,j_i)\right]$$

其中$u_n(t,n,j_i)$是最终求和完毕后的余项，故不含j_i，即$u_n(t,n,j_i)=u_n(t,n)$，其关于t是指数型函数，且底数均在$(0,1)$区间内，当$t\ge n$时所有指数均大于0，因此

$$\lim_{t\to+\infty} u_n(t,n) = 0$$

从而

$$\lim_{t\to+\infty} F(t) = \prod_{k=1}^{n} \frac{n-k+1}{k}$$

易错点：1.坐标变换下概率密度函数的变换；2.（3）中的每个X_i分布不同，所以和分布不是负二项分布。

难点：（3）中规律的寻找。

3.某卡片系列由n种卡片组成，每种数量均非常多且相等。设集齐这套卡片需要抽取X_n次，每次抽取相互独立。

（1）求$E(X_n)$,$Var(X_n)$。

(2) 某人连续抽取卡片若干次，但他抽取卡片后并未打开卡片，因此他不知道是否已集齐卡片。设$n=5$，为使其集齐卡片的概率超过95%，求他至少应抽取卡片的次数。

（3）卡片老板将卡片种数调整为每人集齐卡片所需平均次数附近。随机抽取2000次后，发现收集到的每种卡片数量如下表所示：

数量	2	3	4	5	6	7
种数	2	6	48	38	4	1

试在$\alpha=0.05$下检验假设H_0:$n=100$。

解（1）令$X_n=\sum_{i=1}^{n} X_i$，其中X_i表示得到第$i-1$张不同卡片之后，抽到第i张卡片还需要抽取的次数。显然有$X_1=1$，其余各随机变量均服从几何分布，且参数分别为$\theta=(n-i+1)/n$，由于几何分布的期望为$1/\theta$，方差为$(1-\theta)/\theta^2$，即$E(X_i)=n/(n-i+1)$，$Var(X_i)=n(i-1)/(n-i+1)^2$，故

$$E(X_n) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{n}{n-i+1}$$

$$Var(X_n) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \frac{n(i-1)}{(n-i+1)^2}$$

（2）当$n=5$时，$E(X_5)=137/12$，$Var(X_5)=3625/144$，由切比雪夫不等式

$$P(|X_n - E(X_n)| \geq c) \leq \frac{Var(X_n)}{c^2}$$

令$Var(X_5)/c^2=5\%$，解得$c\approx 22.44$，故$X_5\geq 34$，即他至少应抽取34次卡片。

（3）首先，由$Stolz$定理可证明$E(X_n)\sim n\ln n\,(n\to+\infty)$，而表中的实际值很大，故可用$n\ln n$近似计算$E(X_n)$。

在H_0成立的条件下，收集到的每张卡片数量的理论值均为$2000/100\ln 100=4.3429$。拟合优度为

$$\chi=1\times\frac{(4.3429-0)^2}{4.3429}+2\times\frac{(4.3429-2)^2}{4.3429}+6\times\frac{(4.3429-3)^2}{4.3429}$$
$$+48\times\frac{(4.3429-4)^2}{4.3429}+38\times\frac{(4.3429-5)^2}{4.3429}+4\times\frac{(4.3429-6)^2}{4.3429}$$
$$+1\times\frac{(4.3429-7)^2}{4.3429}=18.595<43.773={\chi_{30}}^2(0.05)<{\chi_{99}}^2(0.05)$$

故接受H_0。

易错点：（1）中对X的合理拆分。

难点：1.不等式的应用；2.等价量的寻找。

4.已知$X_1,X_2,\cdots,X_n$是从正态总体$N(\mu,\sigma)$中随机选取的样本。

（1）分别在μ未知和已知的条件下求σ的极大似然估计$\hat{\sigma}_{L1},\hat{\sigma}_{L2}$。

(2)判断$\hat{\sigma}_{L1},\hat{\sigma}_{L2}$是否为$\sigma$的无偏估计，若否，修改为无偏估计。

(3)比较$\hat{\sigma}_{L1},\hat{\sigma}_{L2}$的无偏估计在$\mu$已知时的有效性。

(4)若μ已知，求σ^2的最小方差无偏估计。

(5)利用σ无偏估计的方差下界证明：

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\geq\frac{\sqrt{2n+1}}{n}$$

解（1）若μ未知，似然函数为

$$L(x_i,\mu,\sigma)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}}$$

对数似然函数为

$$l(x_i,\mu,\sigma)=\ln L(x_i,\mu,\sigma)=-n\ln\sqrt{2\pi}-n\ln\sigma-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}$$

令

$$\begin{cases}\dfrac{\partial l}{\partial\mu}=\dfrac{\sum_{i=1}^n(x_i-\mu)}{\sigma^2}=0\\ \dfrac{\partial l}{\partial\sigma}=-\dfrac{n}{\sigma}+\dfrac{\sum_{i=1}^n(x_i-\mu)^2}{\sigma^3}=0\end{cases}$$

解得

$$\begin{cases} \hat{\mu}_L = \bar{X} \\ \hat{\sigma}_{L1} = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \end{cases}$$

若μ已知，同理可得

$$\hat{\sigma}_{L2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

（2）若μ未知，令

$$S_1 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

由于$X = (n-1)S_1^2/\sigma^2 \sim {\chi_{n-1}}^2$，令$s = S_1/\sigma$，则$X = (n-1)s^2$，于是$s$的概率密度函数为

$$g(s) = \left|\frac{\partial X}{\partial s}\right| k_{n-1}(x) = \frac{(n-1)^{\frac{n-1}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)} s^{n-2} e^{-\frac{(n-1)s^2}{2}} I(s > 0)$$

于是

$$E(S_1) = \sigma \int_0^{+\infty} s g(s)\, ds = \sigma \frac{(n-1)^{\frac{n-1}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)} \int_0^{+\infty} s^{n-1} e^{-\frac{(n-1)s^2}{2}} ds$$

令$t = (n-1)s^2/2$，则计算可得

$$E(S_1) = \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma$$

因此

$$E(\hat{\sigma}_{L1}) = \sqrt{\frac{n-1}{n}} E(S_1) = \sqrt{\frac{2}{n}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma \neq \sigma$$

所以$\hat{\sigma}_{L1}$不是σ的无偏估计，无偏估计应为$\hat{\sigma}'_{L1} = c_{n1} S_1$，其中

$$c_{n1} = \sqrt{\frac{n-1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

同理，$\hat{\sigma}_{L2}$也不是σ的无偏估计，无偏估计应为$\hat{\sigma}'_{L2} = c_{n2} S_2$，其中

$$E(S_2)=\sqrt{\frac{2}{n}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sigma$$

$$S_2=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

$$c_{n2}=\sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}$$

（3）由于$E(S_1^2)=E(S_2^2)=\sigma^2$，故

$$Var(\hat{\sigma}'_{L1})=c_{n1}^2[E(S_1^2)-E^2(S_1)]=(c_{n1}^2-1)\sigma^2$$
$$Var(\hat{\sigma}'_{L2})=c_{n2}^2[E(S_2^2)-E^2(S_2)]=(c_{n2}^2-1)\sigma^2$$

分别在n为奇数和偶数时对c_{n1}与c_{n2}作商并判断商的单调性可知$c_{n1}>c_{n2}$，即$\hat{\sigma}'_{L2}$更有效。

（4）待估参数为$g(\sigma^2)=\sigma^2$。令

$$f(x,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

则

$$\frac{\partial f}{\partial\sigma^2}=\frac{(x-\mu)^2-\sigma^2}{2\sqrt{2\pi}\sigma^5}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

费歇尔信息量为

$$I(\sigma^2)=\int_{-\infty}^{+\infty}\left[\left(\frac{\partial f}{\partial\sigma^2}\right)^2/f\right]dx=\frac{1}{2\sigma^4}$$

由克拉美——劳不等式得对σ^2的任意无偏估计$\hat{g}$，有

$$Var_{\sigma^2}(\hat{g})\geq\frac{g'(\sigma^2)}{nI(\sigma^2)}=\frac{2\sigma^4}{n}$$

另一方面，σ^2的极大似然估计为

$$\hat{\sigma}_L^2=\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

且易知其是无偏的。经计算得

$$Var(\hat{\sigma}_L^2)=\frac{E(x_i-\mu)^4-E^2[(x_i-\mu)^2]}{n}=\frac{3\sigma^4-\sigma^4}{n}=\frac{2\sigma^4}{n}$$

故$\hat{\sigma}_L^2$是σ^2的最小方差无偏估计。

（5）同（4）有

$$\frac{\partial f}{\partial\sigma}=\frac{(x-\mu)^2-\sigma^2}{\sqrt{2\pi}\sigma^4}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$I(\sigma)=\int_{-\infty}^{+\infty}\left[\left(\frac{\partial f}{\partial\sigma^2}\right)^2/f\right]dx=\frac{2}{\sigma^2}$$

$$Var_\sigma(\hat{g}')\ge\frac{g'(\sigma)}{nI(\sigma)}=\frac{\sigma^2}{2n}$$

于是

$$Var(\hat{\sigma}_{L2}')=(c_{n2}^2-1)\sigma^2\ge\frac{\sigma^2}{2n}$$

整理得

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\ge\frac{\sqrt{2n+1}}{n}$$

易错点：较为复杂的积分。

难点：1.Γ函数的应用；2.求S的分布时对已知分布的应用；3.（3）中两数大小的比较；4.方差下界的估计。