Probability and Statistics
Tutorial 11

Siyi Wang

Southern University of Science and Technology

11951002@mail.sustech.edu.cn

December 1, 2020

Outline

1. Law of Large Numbers (LLN)

定理 5.2.1 (大数定律) 令 $X_1, X_2, \cdots, X_i \cdots$ 是独立随机变量序列, $E(X_i) = \mu, \operatorname{Var}(X_i) = \sigma^2$. 令 $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. 那么对任意的 $\varepsilon > 0$, 当 $n \to \infty$ 时,

$$P(|\overline{X}_n - \mu| > \varepsilon) \to 0$$

证明 我们首先计算 $E(\overline{X}_n)$ 和 $\operatorname{Var}(\overline{X}_n)$:

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

因为 X_i 独立,

$$\operatorname{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(X_i) = \frac{\sigma^2}{n}$$

利用切比雪夫不等式立即得到想要的结果，具体表述为

$$P(|\overline{X}_n - \mu| > \varepsilon) \leqslant \frac{\operatorname{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0, \quad \text{当 } n \to \infty \text{ 时} \qquad \blacksquare$$

2. Convergence in Probability

1. 依概率收敛 设$\{X_n\}$为一随机变量序列，X为一随机变量. 如果对任意的$\varepsilon > 0$，有

$$\lim_{n \to +\infty} P\{|X_n - X| < \varepsilon\} = 1,$$

则称$\{X_n\}$依概率收敛于X，记作$X_n \xrightarrow{P} X$.

3. Monte Carlo

例 5.2.1 (蒙特卡罗积分) 假设我们希望计算

$$I(f)=\int_0^1 f(x)\mathrm{d}x$$

这里的积分不能通过初等方式或者积分表计算出来. 最常用的方法是利用数值计算，用和近似代替积分. 很多设计方案和软件包都可以完成这项工作. 另外一种方法，称为**蒙特卡罗方法**（Monte Carlo method），其工作原理如下所述. 生成 $[0,1]$ 上独立的均匀随机变量，即 $X_1, X_2, \cdots, X_n$，并计算

$$\hat{I}(f)=\frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

由大数定律知道，它应该接近于 $E[f(X)]$，即

$$E[f(X)]=\int_0^1 f(x)\mathrm{d}x=I(f)$$

对其进行简单修改就可以改变积分区间和积分形式. 与标准数值方法比较，蒙特卡罗方法在一维情形下不是特别有效，但是随着积分维数的增加，其有效性越来越强.

Review

4. Central Limit Theorem (CLT)

定理（独立同分布的中心极限定理） 设$\{X_n\}$为独立同分布的 r.v 列，其数学期望和方差分别为

$$E(X_k)=\mu, D(X_k)=\sigma^2>0 \quad (k=1,2,\cdots)$$

则$\{X_n\}$服从中心极限定理，即标准化 r.v

$$Z_n=\frac{\sum_{k=1}^{n}X_k-E(\sum_{k=1}^{n}X_k)}{\sqrt{D(\sum_{k=1}^{n}X_k)}}=\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\,\sigma} \quad (n=1,2,\cdots)$$

的分布函数$F_n(x)$对任意 x 满足

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\left\{\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\,\sigma}\le x\right\}$$

$$=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=\Phi(x)$$

4. Central Limit Theorem (CLT)

对于均值为μ,方差$\sigma^2>0$的独立同分布的 r.v 列

$$X_1, X_2, \cdots, X_k, \cdots$$

有

$$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\,\sigma} \overset{\text{近似}}{\sim} N(0, 1)$$

即或

$$X_1 + X_2 + \cdots + X_n \overset{\text{近似}}{\sim} N(n\mu, n\sigma^2)$$

4. Central Limit Theorem (CLT)

4. 按分布收敛、弱收敛 设$\{F_n(x)\}$是随机变量序列$\{X_n\}$的分布函数列，$F(x)$为X的分布函数. 若对$F(x)$的任一连续点x，都有$\lim_{n\to+\infty} F_n(x)=F(x)$，则称$\{F_n(x)\}$弱收敛于$F(x)$，记作$F_n(x)\xrightarrow{W}F(x)$. 也称$\{X_n\}$按分布收敛于$X$，记作$X_n\xrightarrow{L}X$.

4. Central Limit Theorem (CLT)

（1）**棣莫弗–拉普拉斯中心极限定理**　设n重伯努利试验中，事件A在每次试验中出现的概率为$p(0<p<1)$，记μ_n为n次试验中事件A出现的次数，且记

$$Y_n^* = \frac{\mu_n - np}{\sqrt{np(1-p)}}.$$

则对任意实数y，有

$$\lim_{n\to+\infty} P(Y_n^* \leqslant y) = \Phi(y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-\frac{t^2}{2}}\mathrm{d}t.$$

4. Central Limit Theorem (CLT)

（4）**二项分布的泊松近似(泊松定理)** 在 n 重伯努利试验中,记事件 A 在一次试验中发生的概率为 p_n(与试验次数 n 有关),如果当 $n \to +\infty$ 时,有 $np_n \to \lambda$,则

$$\lim_{n \to +\infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

4. Central Limit Theorem (CLT)

（4）**二项分布的泊松近似(泊松定理)**　在 n 重伯努利试验中，记事件 A 在一次试验中发生的概率为 p_n(与试验次数 n 有关)，如果当 $n \to +\infty$ 时，有 $np_n \to \lambda$，则

$$\lim_{n \to +\infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

4. Central Limit Theorem (CLT)

例 5.3.7 (粒度分布) 某些谷类的粒度分布经常偏度很大，右尾递减缓慢. 称为**对数正态** (lognormal) 的分布有时用来拟合这样的分布. 如果 $\log X$ 服从正态分布，我们就说 X 服从对数正态分布. 中心极限定理给出了某些场合下利用对数正态分布的理论依据.

假设初始的粒度 y_0 遭受重复的冲击，每次冲击后，有 X_i 比率的粒子存活下来，且假设 X_i 独立同分布. 在第一次冲击后，粒度是 $Y_1 = X_1 y_0$；第二次冲击后，粒度是 $Y_2 = X_2 X_1 y_0$；在 n 次冲击后，粒度是

$$Y_n = X_n X_{n-1} \cdots X_2 X_1 y_0$$

那么

$$\log Y_n = \log y_0 + \sum_{i=1}^{n} \log X_i$$

应用中心极限定理于 $\log Y_n$. ■

类似的构造与金融理论相关. 假设初始投资为 v_0，收益在离散时刻 (例如天) 发生. 如果第一天收益率是 R_1，那么投资变为 $V_1 = R_1 v_0$. 两天后的值是 $V_2 = R_2 R_1 v_0$，n 天之后的值是

$$V_n = R_n R_{n-1} \cdots R_1 v_0$$

因此，对数值

$$\log V_n = \log v_0 + \sum_{i=1}^{n} \log R_i$$

如果收益率是独立同分布的，那么 $\log V_n$ 的分布是近似正态的.

5. Statistics

如何收集数据？

从研究对象中任取 n 个 “个体”，观察它们的数量指标

$$X_1, X_2, \cdots, X_n$$

这一过程称为**抽样**. $X_1, X_2, \cdots, X_n$ 称为**容量**为 n 的**样本**.

抽样的特点

在相同条件下对总体 X 进行 n 次重复、独立观察

独立性：要求各次取样的结果互不影响

代表性：每次取出的样品与总体有相同的分布

样本的特点

观察前：$X_1, X_2, \cdots, X_n$ 是相互独立, 与总体同分布的r.v

观察后：样本值 $x_1, x_2, \cdots, x_n$ n个具体的观察数据

5. Statistics

样本的分布

问题 样本 $X_1, X_2, \cdots, X_n$ 作为多维随机变量，服从什么分布?

1. 若总体分布函数为 $F(x)$，则样本联合分布函数为

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$$

2. 若总体密度函数为 $f(x)$，则样本联合密度函数为

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$$

例 设 $X_1, X_2, \cdots, X_n$ 为来自总体 $X \sim N(\mu, \sigma^2)$ 的样本，则样本分布密度函数为

$$f^*(x_1, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Review

5. Statistics

和方差不同?

样本均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$

为什么不是 $\frac{1}{n}$(后续说明

样本标准差 $S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$

5. Statistics

设总体 X 的均值和方差

$$E(X) \triangleq \mu,\ D(X) \triangleq \sigma^2$$

都存在. $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本，则

$$E(\bar{X}) = \mu,\ \ D(\bar{X}) = \frac{\sigma^2}{n},\ \ E(S^2) = \sigma^2$$

Homework

1. 令 $X_1, X_2, \cdots$ 是独立随机变量序列，具有均值 $E(X_i) = \mu$ 和方差 $\operatorname{Var}(X_i) = \sigma_i^2$. 证明：如果 $n^{-2}\sum_{i=1}^{n}\sigma_i^2 \to 0$，那么依概率 $\overline{X} \to \mu$.

Homework

Solution

10. Proof. For $\forall \varepsilon > 0$, we have

$$\mathbb{P}(|\bar{X} - \mu| > \varepsilon) \leq \frac{1}{\varepsilon^2} Var(\bar{X})$$

$$= \frac{1}{\varepsilon^2} \cdot \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 \longrightarrow 0, \text{ as } n \to \infty. \qquad \square$$

Homework

2. 令 X_i 如习题 1，但是具有 $E(X_i) = \mu_i$ 和 $n^{-1}\sum_{i=1}^{n}\mu_i \to \mu$，证明依概率 $\overline{X} \to \mu$.

Solution

11. Proof.

For $\varepsilon > 0$, $\exists N$ s.t. for $\forall\, n > N$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mu_i - \mu \right| < \frac{\varepsilon}{2}.$$

For $\forall\, n > N$, we have

$$\mathbb{P}(|\bar{X} - \mu| > \varepsilon) \leq \mathbb{P}\left(\left|\bar{X} - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| > \frac{\varepsilon}{2}\right)$$

$$\left(|\bar{X} - \mu| \leq \left|\bar{X} - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| + \frac{\varepsilon}{2} \right)$$

$$\leq \frac{4}{\varepsilon^2} \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 \to 0, \quad \text{as } n \to \infty. \quad \square$$

10 / 15

Homework

1、设$\{X_k\}$为独立随机变量序列，且

$$P\left(X_k = \pm\sqrt{\ln k}\right) = \frac{1}{2},\ k = 1,2,\dots.$$

证明：$\{X_k\}$服从大数定律。

Homework

Solution

12. Proof: $\mathbb{E} X_k = 0$, $k = 1, 2, \cdots$.

$Var(X_k) = \ln k$, $k = 1, 2, \cdots$.

$$\mathbb{P}(|\overline{X} - 0| > \varepsilon) \leq \frac{1}{\varepsilon^2} \frac{1}{n^2} \sum_{i=1}^{n} \ln i$$

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} \ln i}{n^2} = \lim_{n \to \infty} \frac{\ln n}{n^2 - (n-1)^2}$$

$$= \lim_{n \to \infty} \frac{\ln n}{2n - 1} = 0.$$

Hence, $\lim_{n \to \infty} \mathbb{P}(|\overline{X}| > \varepsilon) = 0$. $\square$

Homework

2、 设$\{X_n\}$为独立随机变量序列，且

$$P(X_n = 1) = p_n,\ P(X_n = 0) = 1 - p_n\ \ n = 1,2,\ldots.$$

证明：$\{X_n\}$服从大数定律。

Solution

Hence, $\{\mu_n\}$ is a Cauchy sequence. $\square$

2° $\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{k=1}^{n} p_k(1-p_k) \leqslant \frac{1}{4n}$

3° $\mathbb{P}(|X-\mu| > \varepsilon) \leqslant \frac{1}{\varepsilon^2}\cdot\frac{1}{4n} \to 0, \quad n\to\infty.$ $\square$

Solution

Counterexample:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\mu_i \text{ may not exist.}$$

$\mu_n = \mathbb{E}X_n = p_n.$

$p_1 = 1,\ p_2 = 0,\ p_3 = p_4 = 1,\ p_5 = p_6 = 0,$

$p_7 = p_8 = p_9 = p_{10} = 1,\ p_{11} = p_{12} = p_{13} = p_{14} = 0,$

$\cdots \quad \cdots$

Let $\quad m_k = \frac{1}{k}\sum_{i=1}^{k}\mu_i = \frac{1}{k}\sum_{i=1}^{k}p_i$

$$m_{2(1+\cdots+2^n)} = \frac{1+2+\cdots+2^n}{2(1+2+\cdots+2^n)} = \frac{1}{2}$$

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

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

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

$$= \frac{2}{3}.$$

Homework

1、设某种电器元件的寿命服从均值为100小时的指数分布，现随机取得16只，设它们的寿命是相互独立的， 求这16只元件的寿命的总和大于1920小时的概率。

Homework

Solution

14. Solution. $X_i \sim \text{Exp}\left(\frac{1}{100}\right)$, $\mu = 100$, $\sigma^2 = 10^4$

$$S = \sum_{i=1}^{n} X_i, \quad n = 16.$$

$$\mathbb{P}(S > 1920) = \mathbb{P}\left(\frac{S - n\mu}{\sqrt{n\sigma^2}} > \frac{1920 - n\mu}{\sqrt{n\sigma^2}}\right)$$

$$\approx \mathbb{P}\left(Z > \frac{320}{\sqrt{16 \times 10^4}}\right) = 1 - \Phi\left(\frac{4}{5}\right). \quad \square$$

Homework

2、某保险公司的老年人寿保险有1万人参加，每人每年交200元，若老人在该年内死亡，公司付给受益人1万元。设老年人死亡率为0.017，试求保险公司在一年内这项保险亏本的概率。

Solution

15. Solution $X_i \sim Bern(p)$, $p = 0.017$

$$S = \sum_{i=1}^{n} X_i, \quad n = 10^4.$$

$$\mathbb{P}(S > 200) = \mathbb{P}\left(\frac{S - np}{\sqrt{np(1-p)}} > \frac{200 - np}{\sqrt{np(1-p)}}\right)$$

$$\approx \mathbb{P}\left(Z > \frac{30}{\sqrt{167.11}}\right) \approx 1 - \Phi(2.32). \square$$

13 / 15

Supplement Exercises

1. 某保险公司多年的统计资料表明，在索赔户中被盗索赔户占 20%，以 X 表示在随意抽查的 100 个索赔户中因被盗向保险公司索赔的户数.

(1) 写出 X 的分布列；

(2) 求被盗索赔户不少于 14 户且不多于 30 户的概率的近似值.

Supplement Exercises

Solution

解 （1）X 服从 $n=100, p=0.2$ 的二项分布 $b(100,0.2)$，即

$$P(X=k)=\binom{n}{k}0.2^k 0.8^{100-k},\qquad k=0,1,2,\cdots,n.$$

（2）利用棣莫弗–拉普拉斯中心极限定理，有

$$P(14 \leqslant X \leqslant 30)=P(13.5<X<30.5)$$
$$\approx \Phi\left(\frac{30.5-100\times 0.2}{\sqrt{100\times 0.2\times 0.8}}\right)-\Phi\left(\frac{13.5-100\times 0.2}{\sqrt{100\times 0.2\times 0.8}}\right)$$

Supplement Exercises

17. 设某生产线上组装每件产品的时间服从指数分布，平均需要 10 分钟，且各件产品的组装时间是相互独立的.

(1) 试求组装 100 件产品需要 15 小时至 20 小时的概率；

(2) 保证有 95% 的可能性，问 16 小时内最多可以组装多少件产品？

Solution

解 记 X_i 为组装第 i 件产品的时间(单位:分钟),则由 $X_i \sim Exp(\lambda)$, $E(X_i)=1/\lambda=10$,知 $\mathrm{Var}(X_i)=1/\lambda^2=100$.

(1) 根据题意所求概率如下,再用林德伯格–莱维中心极限定理可得

$$P\left(15\times 60 \leqslant \sum_{i=1}^{100} X_i \leqslant 20\times 60\right) \approx \Phi\left(\frac{1\,200-100\times 10}{\sqrt{100\times 100}}\right) - \Phi\left(\frac{900-100\times 10}{\sqrt{100\times 100}}\right)$$

$$=\Phi(2)-\Phi(-1)=\Phi(2)+\Phi(1)-1=0.818\,5.$$

(2) 设 16 小时内最多可以组装 k 件产品. 则根据题意可列出概率不等式

$$P\left(\sum_{i=1}^{k} X_i \leqslant 16\times 60\right) \geqslant 0.95,$$

再用林德伯格–莱维中心极限定理可得

$$\Phi\left(\frac{960-10k}{\sqrt{100k}}\right) \geqslant 0.95,$$

由此查表得 $\dfrac{960-10k}{10\sqrt{k}} \geqslant 1.645$,从中解得 $k=81$.

11. 设 X 与 Y 相互独立，分别服从参数为 λ_1 和 λ_2 的泊松分布，试求

$$E(X_1 \mid X_1 + X_2 = n).$$

Supplement Exercises

Solution

解　因为 $X_1+X_2\sim P(\lambda_1+\lambda_2)$,所以

$$P(X_1=i\mid X_1+X_2=n)=\frac{P(X_1=i,X_1+X_2=n)}{P(X_1+X_2=n)}$$

$$=\frac{P(X_1=i)P(X_2=n-i)}{P(X_1+X_2=n)}$$

$$=\frac{\dfrac{\lambda_1^i}{i!}e^{-\lambda_1}\dfrac{\lambda_2^{n-i}}{(n-i)!}e^{-\lambda_2}}{\dfrac{(\lambda_1+\lambda_2)^n}{n!}e^{-(\lambda_1+\lambda_2)}}=\binom{n}{i}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^i\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-i},\quad i=0,1,\cdots,n.$$

Solution

这说明：$X_1 \mid X_1 + X_2 = n$ 服从二项分布 $b(n,p)$，其中 $p = \lambda_1/(\lambda_1 + \lambda_2)$，所以

$$E(X_1 \mid X_1 + X_2 = n) = np = \frac{n\lambda_1}{\lambda_1 + \lambda_2}.$$

12. 设二维连续随机变量(X,Y)的联合密度函数为

$$p(x,y)=\begin{cases} x+y, & 0<x,y<1, \\ 0, & \text{其他}. \end{cases}$$

试求$E(X \mid Y=0.5)$.

Supplement Exercises

Solution

解 先求条件密度函数 $p(x \mid y)$. 当 $0 < y < 1$ 时, $p_Y(y) = \int_0^1 (x + y)\,\mathrm{d}x = 0.5 + y$. 所以

$$p(x \mid y = 0.5) = \begin{cases} x + 0.5, & 0 < x < 1, \\ 0, & \text{其他}. \end{cases}$$

由此得

$$E(X \mid Y = 0.5) = \int_0^1 x \cdot (x + 0.5)\,\mathrm{d}x = \frac{7}{12}.$$

66. 一座大楼有两座电梯，一快一慢，较慢电梯的平均等待时间是 3 分钟，较快电梯的平均等待时间是 1 分钟. 如果乘客以概率 $\frac{2}{3}$ 选乘较快的电梯，以概率 $\frac{1}{3}$ 选乘较慢的电梯，期望等待时间是多少? (利用 4.4.1 节定理 4.4.1.1 的全期望公式，定义合适的随机变量 X 和 Y.)

Solution

Solution. X: 等待时间, Y: 乘客的选择 $\begin{pmatrix} 1: \text{快} \\ 0: \text{慢} \end{pmatrix}$

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

$$= \mathbb{E}[X|Y=1]\mathbb{P}(Y=1) + \mathbb{E}[X|Y=0]\mathbb{P}(Y=0)$$

$$= 1 \times \frac{2}{3} + 3 \times \frac{1}{3} = \frac{5}{3}. \square$$

Supplement Exercises

73. 抛掷一枚质地均匀的硬币 n 次，记录出现正面的次数为 N. 再将硬币抛掷 N 次. 计算该过程出现正面的总期望数.

Solution

Solution. $N \sim b(n, \frac{1}{2})$.

$$M \mid N \sim b(N, \tfrac{1}{2}).$$

$$\mathbb{E}[M+N] = \mathbb{E}[N] + \mathbb{E}[M] = \frac{n}{2} + \mathbb{E}[\mathbb{E}[M|N]]$$

$$= \frac{n}{2} + \mathbb{E}[\frac{N}{2}] = \frac{n}{2} + \frac{1}{2} \times \frac{n}{2} = \frac{3}{4}n. \quad \square$$

Supplement Exercises

11. 有人给出下面的论据以说明中心极限定理必定有瑕疵:"我们知道独立泊松随机变量之和服从泊松分布,分布参数就是和项参数之和. 特别地,如果对参数为 n^{-1} 的 n 个独立泊松随机变量求和,那么得到参数为 1 的泊松分布. 中心极限定理告诉我们,当 n 趋于无穷时,和的分布趋向于正态分布,但是参数为 1 的泊松分布不是正态分布."你是怎么考虑这个论据的?

Supplement Exercises

48. 如果 $U = a + bX$，$V = c + dY$，证明 $|\rho_{UV}| = |\rho_{XY}|$.

Supplement Exercises

21. 掷一颗骰子两次,求其点数之和与点数之差的协方差.

Supplement Exercises

Solution

解 记 X 为第一次掷出的点数，Y 为第二次掷出的点数，则 X 与 Y 独立同分布，即有 $E(X)=E(Y)$，$E(X^2)=E(Y^2)$. 由此得

$$\begin{aligned}\mathrm{Cov}(X+Y,X-Y) &= E(X^2-Y^2)-E(X+Y)E(X-Y)\\ &= 0.\end{aligned}$$

23. 将一枚硬币重复掷 n 次，以 X 和 Y 分别表示正面向上和反面向上的次数，试求 X 和 Y 的协方差及相关系数.

Supplement Exercises

Solution

$$\mathrm{Cov}(X,Y)=\mathrm{Cov}(X,n-X)=-\mathrm{Cov}(X,X)=-\mathrm{Var}(X)=-\frac{n}{4},$$

$$\mathrm{Corr}(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}=\frac{-n/4}{n/4}=-1.$$

这表明:X与Y间是完全负相关. 这个结论早就蕴含在线性关系式$X+Y=n$之中.

37. 设 a 为区间(0,1)上的一个定点，随机变量 X 服从区间(0,1)上的均匀分布. 以 Y 表示点 X 到 a 的距离. 问 a 为何值时 X 与 Y 不相关.

Solution

解 由题设条件知 $X \sim U(0,1)$，$Y = |X - a|$，$E(X) = 1/2$. 又因为

$$E(Y) = \int_0^1 |x - a| \mathrm{d}x = \int_0^a (a - x) \mathrm{d}x + \int_a^1 (x - a) \mathrm{d}x = a^2 - a + \frac{1}{2}.$$

$$E(XY) = \int_0^a x(a - x) \mathrm{d}x + \int_a^1 x(x - a) \mathrm{d}x = \frac{a^3}{3} - \frac{a}{2} + \frac{1}{3}.$$

$$\mathrm{Cov}(X,Y) = \frac{a^3}{3} - \frac{a^2}{2} + \frac{1}{12},$$

1. The Exchange Paradox

4.9.1 The Exchange Paradox

The *"Exchange Paradox"* (Christensen and Utts 1992) has generated a lengthy dialog among statisticians. The problem (or the paradox) goes as follows:

> A swami puts m dollars in one envelope and $2m$ dollars in another. You and your opponent each get one of the envelopes (at random). You open your envelope and find x dollars, and then the swami asks you if you want to trade envelopes. You reason that if you switch, you will get either $x/2$ or $2x$ dollars, each with probability 1/2. This makes the expected value of a switch equal to $(1/2)(x/2) + (1/2)(2x) = 5x/4$, which is greater than the x dollars that you hold in your hand. So you offer to trade.
>
> The paradox is that your opponent has done the same calculation. How can the trade be advantageous for both of you?

Further Reading

1. The Exchange Paradox

(i) Christensen and Utts say, "The conclusion that trading envelopes is always optimal is based on the assumption that there is no information obtained by observing the contents of the envelope," and they offer the following resolution. Let $M \sim \pi(m)$ be the pdf for the amount of money placed in the first envelope, and let X be the amount of money in your envelope. Then $P(X = m|M = m) = P(X = 2m|M = m) = 1/2$, and hence

$$P(M = x|X = x) = \frac{\pi(x)}{\pi(x) + \pi(x/2)} \text{ and } P(M = x/2|X = x) = \frac{\pi(x/2)}{\pi(x) + \pi(x/2)}.$$

1. The Exchange Paradox

It then follows that the expected winning from a trade is

$$\frac{\pi(x)}{\pi(x)+\pi(x/2)}2x+\frac{\pi(x/2)}{\pi(x)+\pi(x/2)}\frac{x}{2},$$

and thus you should trade only if $\pi(x/2) < 2\pi(x)$. If π is the exponential(λ) density, it is optimal to trade if $x < 2\log 2/\lambda$.

(ii) A more classical approach does not assume that there is a pdf on the amount of money placed in the first envelope. Christensen and Utts also offer an explanation here, noting that the paradox occurs if one incorrectly assumes that $P(Y = y|X = x) = 1/2$ for all values of X and Y, where X is the amount in your envelope and Y is the amount in your opponent's envelope. They argue that the correct conditional distributions are $P(Y = 2x|X = m) = 1$ and $P(Y = x/2|X = 2m) = 1$ and that your expected winning if you trade is $E(Y) = 3m/2$, which is the same as your expected winning if you keep your envelope.

Further Reading

1. The Exchange Paradox

This paradox is often accompanied with arguments for or against the Bayesian methodology of inference (see Chapter 7), but these arguments are somewhat tangential to the underlying probability calculations. For comments, criticisms, and other analyses see the letters to the editor from Binder (1993), Ridgeway (1993) (which contains a solution by Marilyn vos Savant), Ross (1994), and Blachman (1996) and the accompanying responses from Christensen and Utts.

Thank you!