2022Fall Probability & Mathematical Statistics 2022/12/10

Homework 11

Professor: Ziyu Shao *Due: 2022/12/19 3:00pm*

1. Let X be the height of a randomly chosen adult man, and Y be his father's height, where X and Y have been standardized to have mean 0 and standard deviation 1. Suppose that (X, Y) is Bivariate Normal, with $X, Y \sim \mathcal{N}(0,1)$ and $\text{Corr}(X, Y) = \rho$.

   (a) Let $y = ax + b$ be the equation of the best line for predicting Y from X (in the sense of minimizing the mean squared error), *e.g.*, if we were to observe $X = 1.3$ then we would predict that Y is $1.3a + b$. Now suppose that we want to use Y to predict X, rather than using X to predict Y. Give and explain an intuitive guess for what the slope is of the best line for predicting X from Y.

   (b) Find a constant c (in terms of ρ) and an r.v. V such that $Y = cX + V$, with V independent of X.

   Hint: Start by finding c such that $\text{Cov}(X, Y - cX) = 0$.

   (c) Find a constant d (in terms of ρ) and an r.v. W such that $X = dY + W$, with W independent of Y.

   (d) Find $E(Y|X)$ and $E(X|Y)$.

   (e) Reconcile (a) and (d), giving a clear and correct intuitive explanation.

2. Let $X_1, \dots, X_n$ be i.i.d. r.v.s with mean μ and variance σ^2, and $n \geq 2$. A bootstrap sample of $X_1, \dots, X_n$ is a sample of n r.v.s $X_1^*, \dots, X_n^*$ formed from the $X_j, \forall j \in \{1, \dots, n\}$ by sampling with replacement with equal probabilities. Let $\bar{X}^*$ denote the sample mean of the bootstrap sample:

$$\bar{X}^* = \frac{1}{n}(X_1^* + \cdots + X_n^*).$$

   (a) Calculate $E(X_j^*)$ and $\text{Var}(X_j^*)$ for each $j \in \{1, \dots, n\}$.

   (b) Calculate $E(\bar{X}^*|X_1, \dots, X_n)$ and $\text{Var}(\bar{X}^*|X_1, \dots, X_n)$.

   Hint: Conditional on $X_1, \dots, X_n$, the $X_j^*, \forall j \in \{1, \dots, n\}$ are independent, with a PMF that puts probability $1/n$ at each of the points $X_1, \dots, X_n$. As a check, your answers should be random variables that are functions of $X_1, \dots, X_n$.

   (c) Calculate $E(\bar{X}^*)$ and $\text{Var}(\bar{X}^*)$.

   (d) Explain intuitively why $\text{Var}(\bar{X}) < \text{Var}(\bar{X}^*)$.

3. Properties of Conditional Expectation.

   (a) Let X, Y be jointly Gaussian random variables. Show

$$E[Y|X] = L[Y|X] = E(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - E(X)).$$

   (b) We wish to estimate the probability of landing heads, denoted by θ, of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim \mathrm{Unif}(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MMSE $E(\Theta|X)$ and LLSE $L(\Theta|X)$.

4. Let X be a discrete r.v. whose distinct possible values are $x_0, x_1, \ldots,$ and let $p_k = P(X = x_k)$. The entropy of X is defined as $H(X) = \sum_{k=0}^{\infty} p_k \log_2(1/p_k)$.

   (a) Find $H(X)$ for $X \sim \mathrm{Geom}(p)$.

   (b) Let X and Y be i.i.d. discrete r.v.s. Show that $P(X = Y) \geq 2^{-H(X)}$.

5. Previously we introduce the coupon collector problem: n toy types, collected one by one, sampling with replacement from the set of toy types each time, and all toy types were equally likely to be collected. Now suppose that at each time, the j^{th} toy type is collected with probability p_j, where the p_j, $j \in \{1, \ldots, n\}$ are not necessarily equal. Let N be the number of toys needed until we have a full set. Find $E(N)$.

   This is the generalized coupon collector problem. We can solve it by the famous embedding trick: discrete problems can be solved by embedding them in continuous Poisson processes.

   (a) Considering a continuous-time version of the previous problem. Suppose that toys arrive according to a Poisson process with parameter $\lambda = 1$. Then the inter-arrival times between toys are X_i, where $X_i \sim \mathrm{Expo}(1), i \in \{1, 2, \ldots\}$. Let $N_t^{(j)}$ denote the number of j^{th} type toys arrives by time t. Show that $N_t^{(1)}, \ldots, N_t^{(n)}$ are independent Poisson process, and $N_t^{(j)} \sim \mathrm{Pois}(p_j \cdot t)$.

   (b) Let Y_j denote the arrival time of the first toy of type j, where $j \in \{1, \ldots, n\}$. Let T denote the waiting time until all types of toys are collected. Show that

$$T = \max(Y_1, \ldots Y_n),$$

   and then find $E(T)$.

   (c) Show that

$$T = X_1 + \cdots + X_N,$$

   and then find $E(N)$.

6. This is the bonus problem.

春节期间，牛奶公司推出了新春集福活动：每盒牛奶都附赠一个红包，红包中藏有下列“虎”，“生”，“威”中的一款图案。（如下图）

集齐两个“虎”，一个“生”，一个“威”即可拼齐成为“虎虎生威”全家福。这项活动一经推出，就成为了网红爆款，很多人希望能够集齐一整套。假设红包中的图案是独立随机分布的（并且不能从红包外观上进行区别），“虎”，“生”，“威”三款红包按均匀概率$\frac{1}{3}$分布。

(1) 收集齐一整套“虎虎生威”全家福所需要购买的牛奶盒数的数学期望是多少？

(A) $6\frac{1}{3}$;

(B) $7\frac{1}{3}$;

(C) $8\frac{1}{3}$;

(D) $9\frac{1}{3}$;

(E) 以上都不对。

(2) 在市场部的周会讨论中，大家认为当前的图案投放比例，会导致在收集“虎虎生威”全家福时收集到太多的“生”和“威”，于是探讨可能的改进方案。记图案“虎”、“生”和“威”的投放比例为(p,q,r)，那么下面哪种方案下，收集齐一整套“虎虎生威”全家福所需要购买的牛奶盒数的数学期望是最小的？

(A) $(p,q,r)=(\frac{1}{3},\frac{1}{3},\frac{1}{3})$;

(B) $(p,q,r)=(\frac{1}{2},\frac{1}{4},\frac{1}{4})$;

(C) $(p,q,r)=(\frac{2}{5},\frac{3}{10},\frac{3}{10})$;

(D) $(p,q,r)=(\frac{3}{4},\frac{1}{8},\frac{1}{8})$.