2023 Spring Probability & Statistics for EECS May 12, 2023

Homework 14

Professor: Ziyu Shao & Dingzhu Wen *Due: 23:59 on May 21, 2023*

1. A DNA sequence can be represented as a sequence of letters, where the "alphabet" has 4 letters: A,C,T,G. Suppose such a sequence is generated randomly, where the letters are independent and the probabilities of A,C,T,G are p_1, p_2, p_3, p_4, respectively.

   (a) In a DNA sequence of length 115, what is the expected number of occurrences of the expression "CATCAT" (in terms of the p_j)? (Note that, for example, the expression "CATCATCAT" counts as 2 occurrences.)

   (b) For this part, assume that the p_j are unknown. Suppose we treat p_2 as a $\text{Unif}(0,1)$ r.v. before observing any data, and that then the first 3 letters observed are "CAT". Given this information, what is the probability that the next letter is C?

2. Let $X_1, \ldots, X_n$ be i.i.d. r.v.s with mean μ and variance σ^2, and $n \geq 2$. A bootstrap sample of $X_1, \ldots, X_n$ is a sample of n r.v.s $X_1^*, \ldots, X_n^*$ formed from the $X_j, \forall j \in \{1, \ldots, n\}$ 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 $j \in \{1, \ldots, n\}$.

   (b) Calculate $E(\bar{X}^*|X_1, \ldots, X_n)$ and $\text{Var}(\bar{X}^*|X_1, \ldots, X_n)$.
   Hint: Conditional on $X_1, \ldots, X_n$, the $X_j^*, \forall j \in \{1, \ldots, n\}$ are independent, with a PMF that puts probability $1/n$ at each of the points $X_1, \ldots, X_n$. As a check, your answers should be random variables that are functions of $X_1, \ldots, X_n$.

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

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

3. A coin with probability p of Heads is flipped repeatedly. For (a) and (b), suppose that p is a known constant, with $0 < p < 1$.

   (a) What is the expected number of flips until the pattern HT is observed? What about the pattern HH? Solve the problems using conditional expectation.

   (b) Now suppose that p is unknown, and that we use a $\text{Beta}(a, b)$ prior to reflect our uncertainty about p (where a and b are known constants and are greater than 2). In terms of a and b, find the corresponding answers to (a) and (b) in this setting.

4. A fair 6-sided die is rolled repeatedly.

   (a) Find the expected number of rolls needed to get a 1 followed right away by a 2.

   (b) Find the expected number of rolls needed to get two consecutive 1's.

   (c) Let a_n be the expected number of rolls needed to get the same value n times in a row (*i.e.*, to obtain a streak of n consecutive j's for some not-specified-in-advance value of j). Find a recursive formula for a_{n+1} in terms of a_n.

   (d) Find a simple, explicit formula for an for all $n \geq 1$. What is a_7 (numerically)?

5. 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.

   (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.