Stats 117 Problem Set 6

Due: Monday, May 11 5:00 p.m. on Gradescope

Please show your work for each exercise. If you collaborate with someone else—this is fine—be sure to note that in your homework submission. You must each write up separate answer sets. Any starred exercise is optional: they are extra challenging theoretical exercises for further developing your mastery.

Question 6.1: Recall Question 5.4 from prior homework, where we wish to collect all 50 rare baseball players (and one rare baseball player appears in each pack we purchase, uniformly at random). Give the variance of the number of packs we purchase until we open all 50 rare players. (You do not need to give a precise numerical value, but should feel free to do so.)

Question 6.2: Consider the following three scenarios:

(a) A fair coin is tossed 3 times. Let X be the number of heads and Y the number of tails.

(b) A fair coin is tossed 4 times. Let X be the number of heads in the first 3 tosses and Y the number of tails in the last 3.

(c) A fair coin is tossed 6 times. Let X be the number of heads in the first 3 tosses and Y the number of tails in the last 3.

For each scenario, compute $\text{Cov}(X, Y)$. *Hint.* When possible, write X and Y as sums of H_i and $1 - H_i$, where H_i is 1 if the ith flip is heads (so $1 - H_i = 1$ if it is tails). Use the properties of covariance (Proposition 15.4) to make your life easier. In each case, do your covariance values make sense?

Question 6.3: Consider *randomized response*, a strategy for eliciting private or sensitive data from groups of individuals without compromising the privacy of any individual. The setting is as follows: we have a sensitive question (e.g., "Have you ever done drugs?" or "Have you cheated on your spouse?") with a yes/no answer; let the random variable $X = 1$ if the answer is Yes and $X = 0$ otherwise. Randomized response flips this answer to some $Z \in \{0, 1\}$ with a prescribed probability (e.g., before answering, the question respondent rolls a die, unseen to the data collector, and if the die comes up $\{1, 2, 3, 4\}$ answers truthfully, and on $\{5, 6\}$ flips their answer).

Formally, for a value $\varepsilon \in (0, 1)$, randomized response flips the answer X with probability $\frac{1-\varepsilon}{2}$: this gives a random variable Z satisfying

$$Z = \begin{cases} X & \text{with probability } \frac{1+\varepsilon}{2} \\ 1 - X & \text{with probability } \frac{1-\varepsilon}{2}. \end{cases}$$

Then the surveyor must transform Z to obtain a guess of what the "correct" answer was.

(a) Give $\mathbb{E}[Z \mid X]$.

(b) Give $\text{Var}(Z \mid X)$.

(c) Give scalars a and b such that the transformed variable $aZ + b$ satisfies

$$\mathbb{E}[aZ + b \mid X] = X.$$

(d) Use the law of total variance to compute $\mathrm{Var}(aZ + b)$ using your values a, b above. (Your answer should involve the variance of X.) When ε is near 0—corresponding to higher privacy—what happens to $\mathrm{Var}(aZ + b)$?

(e) Assume that $\mathbb{E}[X] = \frac{1}{2}$, i.e., $X \sim \mathsf{Bernoulli}(\frac{1}{2})$. When using a six-sided die to implement randomized response, so that $Z = X$ on a roll of $\{1, 2, 3, 4\}$, and $Z = 1 - X$ otherwise, give the explicit values for a and b above, and give $\mathrm{Var}(aZ + b)$.

Question 6.4: In financial problems, the *risk* of a random variable X is its variance (as opposed to its mean returns). Suppose that there is a financial crisis, continuing over several days. Each day the crisis continues, with probability p, a bank fails because of a run on the bank. The crisis will end when the central bank insures deposits. On each day (independently), the central bank decides to insure deposits with probability q.

(a) Let $N \geq 1$ be the number of days the crisis continues. What distribution does N have?

(b) Let X be the total number of failed banks. Give the distribution of X conditional on $N = n$.

(c) Give the expected number of failed banks, $\mathbb{E}[X]$.

(d) Give the *risk* of the failed banks, that is, $\mathrm{Var}(X)$. It may be useful to recall that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, and the mean of a $\mathsf{Geom}(q)$ is $\frac{1}{q}$ and its variance is $\frac{1-q}{q^2}$.

Question 6.5 (Recurrence of a random walk)**:** An ant walks randomly on the integers $\mathbb{Z}$, where at time t, the ant's position is $A_t \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. The ant is equally likely to move one position left and one position right, so that

$$P(A_{t+1} = A_t + 1 \mid A_t) = \frac{1}{2} \quad \text{and} \quad P(A_{t+1} = A_t - 1 \mid A_t) = \frac{1}{2}.$$

At time $t = 0$, the ant begins at position 0.

(a) Write the ant's position at time n as $A_n = 2B_n - n$, where B_n is a random variable we have seen in class. What distribution does B_n have?

(b) Define the binomial coefficient

$$C_n := \binom{2n}{n}.$$

Give the ratio C_{n+1}/C_n as a function of n.

(c) Define the normalized value $v_n = \frac{\sqrt{n}}{4^n} \cdot C_n$. Use the result of part (b) to show that

$$v_{n+1} \geq v_n.$$

(d) Use induction to show that $C_n \geq \frac{1}{2\sqrt{n}} \cdot 4^n$ for $n \geq 1$.

(e) Demonstrate that the expected number of times the ant visits 0, that is, returns to the origin (i.e., $A_n = 0$), is infinite. *Hint.* Use indicator random variables.