TRINITY COLLGE DUBLIN

School of Computer Science and Statistics

Mock Final Exam 1 ST3009: Statistical Methods for Computer Science

Question 1. Let X and Y be independent random variables that take values in the set $\{1, 2, 3\}$. Let $V = 2X + 2Y$, and $W = X - Y$.

(a) Assume that $P(X = x)$ and $P(Y = x)$ are non-zero for any $x, y \in \{1, 2, 3\}$. Can V and W be independent? Explain.

For the remaining parts of this problem, assume that X and Y are uniformly distributed on $\{1, 2, 3\}$ i.e. the probability of each value occurring is the same.

(b) Compute the PMF of V, $E[V]$ and $var(V)$.

(c) Compute the joint PMF of V and W.

(d) Compute $E[V|W > 0]$.

Solution:

(a) For V and W to be independent we require

$$P(V = v \text{ and } W = w) = P(V = v)P(W = w)$$

Now $V = v = 10$ when $(X = 2, Y = 3)$, $(X = 3, Y = 2)$ i.e.

$$P(V = 10) = P(X = 2 \text{ and } Y = 3) + P(X = 3 \text{ and } Y = 2)$$

For $W = w = 0$ then $X - Y = 0$ when $X = Y$ i.e.

$$P(W = 0) = P(X = 1 \text{ and } Y = 1) + P(X = 2 \text{ and } Y = 2) + P(X = 3 \text{ and } Y = 3)$$

Since $P(X = x)$ and $P(Y = y)$ are non-zero for any $x, y \in \{1, 2, 3\}$ it follows that $P(V = 10) \neq 0$ and $P(W = 0) \neq 0$. But

$$P(V = 10 \text{ and } W = 0) = 0$$

since when $V = 10$ we cannot have $W = 0$, and so $P(V = 10 \text{ and } W = 0) \neq P(V = 10)P(W = 0)$ i.e. V and W are not independent.

(b) The pair (X, Y) has sample space $S = \{(1, 1), (1, 2), (1, 3), \cdots, (3, 1), (3, 2), (3, 3)\}$, and $|S| = 9$. For each of these possible pairs of values (X, Y) the value of $V = 2(X + Y)$ is as follows:

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	4	6	8
$Y = 2$	6	8	10
$Y = 3$	8	10	12

and so V has sample space $\{4, 6, 8, 10, 12\}$. Since X and Y are uniformly distributed, each pairs of values (X, Y) has the same probability $1/|S| = 1/9$. Therefore,

$$P(V = 4) = \frac{1}{9},\ P(V = 6) = \frac{2}{9},\ P(V = 8) = \frac{3}{9},\ P(V = 10) = \frac{2}{9},\ P(V = 12) = \frac{1}{9}$$

The expectation $E[V]$ is:

$$\begin{aligned} E[V] &= 4 \cdot P(V = 4) + 6 \cdot P(V = 6) + 8 \cdot P(V = 8) + 10 \cdot P(V = 10) + 12 \cdot P(V = 12) \\ &= 4 \cdot \frac{1}{9} + 6 \cdot \frac{2}{9} + 8 \cdot \frac{3}{9} + 10 \cdot \frac{2}{9} + 12 \cdot \frac{1}{9} \\ &= 8 \end{aligned}$$

The variance is:

$$var(V) = (4 - 8)^2 \cdot \frac{1}{9} + (6 - 8)^2 \cdot \frac{2}{9} + (10 - 8)^2 \cdot \frac{2}{9} + (12 - 8)^2 \cdot \frac{1}{9} = \frac{16}{3}$$

(c) For each pair of values pairs of values (X, Y) the value of pair (V, W) is

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	$(4, 0)$	$(6, 1)$	$(8, 2)$
$Y = 2$	$(6, -1)$	$(8, 0)$	$(10, 1)$
$Y = 3$	$(8, -2)$	$(10, -1)$	$(12, 0)$

Each pair of values (X, Y) has the same probability $1/9$ and the value of (V, W) is different for each value of (X, Y), so the probability of each of the pairs of values (V, W) in the above table is $1/9$. The probability of other pairs of values is zero.

(d) Conditioning on $W > 0$ the above table reduces to:

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	$-$	$(6, 1)$	$(8, 2)$
$Y = 2$	$-$	$-$	$(10, 1)$
$Y = 3$	$-$	$-$	$-$

and so

$$\begin{aligned} &P(V = 4|W > 0) = 0,\ P(V = 6|W > 0) = \frac{1}{3},\ P(V = 8|W > 0) = \frac{1}{3} \\ &P(V = 10|W > 0) = \frac{1}{3},\ P(V = 12|W > 0) = 0 \end{aligned}$$

Therefore

$$\begin{aligned} E[V|W > 0] = 4 \cdot P(V = 4|W > 0) + 6 \cdot P(V = 6|W > 0) + 8 \cdot P(V = 8|W > 0) \\ + 10 \cdot P(V = 10|W > 0) + 12 \cdot P(V = 12|W > 0) \\ = 6 \cdot \frac{1}{3} + 8 \cdot \frac{1}{3} + 10 \cdot \frac{1}{3} \\ = \frac{24}{3} = 8 \end{aligned}$$

Question 2. We want to perform a survey of n people, asking their answer to a sensitive question. We use the randomised response approach whereby each respondent tosses a fair coin and if it comes up heads they respond with "yes" and if tails they respond by truthfully answering "yes" or "no" to the sensitive question being asked. Associate with the i'th respondent a random variable X_i which takes value 1 when the person's truthful answer to the sensitive question is "yes" and value 0 otherwise. Assume that the X_i, $i = 1, 2, \cdots, n$ are independent and identically distributed with $P(X_i = 1) = p$.

(a) Let random variable Y_i take value 1 when the i'th respondent answers "yes" and 0 otherwise. Give an expression for $E[Y_i]$ in terms of $E[X_i]$, and using this then rearrange to get $E[X_i]$ in terms of $E[Y_i]$. Hint: use the linearity of expectation.

(b) Let random variable $Z = \sum_{i=1}^{n} Y_i$ be the number of respondents who answer "yes". Write an expression for the probability that out of n respondents z respond "yes".

(c) Suppose we use Z/n as an estimate of $q = E[Y_i]$, where n is the total number of respondents. Recalling that for two random variables W and V we have $E[W + V] = E[W] + E[V]$, show that $E[Z/n] = q$. Using Chebyshev's inequality explain the weak law of large numbers and the behaviour of $|\frac{Z}{n} - q|$ as n becomes large.

(d) Using Z/n as an estimate of q explain how you could use bootstrapping to estimate a confidence interval for the accuracy of this estimate.

Solution:

(a) $P(Y_i = 1) = 0.5 + (1 - 0.5)p$ (with probability 0.5 coin comes up heads and answer "yes", otherwise answer truthfully and with probability p the truthful answer is "yes"). $P(Y_i = 0) = 1 - P(Y_i = 1) = (1 - 0.5)(1 - p)$. Therefore

$$E[Y_i] = 1 \cdot P(Y_i = 1) + 0 \cdot P(Y_i = 0) = 0.5 + (1 - 0.5)p$$

Now $E[X_i] = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) = p$ and so $E[Y_i] = 0.5 + 0.5E[X_i]$. Rearranging, we get

$$E[X_i] = 2E[Y_i] - 1$$

(b) $Z = \sum_{i=1}^{n} Y_i$ with $P(Y_i = 1) = 0.5 + (1 - 0.5)p$ and $P(Y_i = 0) = 1 - P(Y_i = 1)$. The Y_i, $i, 1, 2, \cdots, n$ are independent and so Z is a binomial random variable with

$$P(Z = z) = \binom{n}{z} q^z (1 - q)^{n-z}$$

where $q = 0.5 + (1 - 0.5)p$.

(c) See notes.

(d) See notes.

Question 3. Diana the Daredevil is trying to break the world land-speed record using her rocket-powered motorcycle. In order to do so, she needs to cover more than 500 metres in 10 seconds. If her motorcycle has Z pounds of rocket fuel to start, then it travels a distance of $X = 50\sqrt{Z}$ metres in 10 seconds.

(a) Suppose that the amount of rocket fuel Z is a random variable uniformly distributed over $[50, 150]$. Compute the CDF of the distance X.

(b) Compute the probability that Diana breaks the world record on any given trial.

(c) Now suppose that Diana is allowed to take a total of n trials in order to try and break the world record. (Here n is some fixed positive integer: i.e., $n \in \{1, 2, 3, \cdots\}$). The amount of rocket fuel at the start is independent from trial to trial. How large does n have to be in order for Diana to have a better than 90% chance of breaking the world record over the n trials?

Solution:

(a) The CDF is:

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(50\sqrt{Z} \leq x) = P(Z \leq \left(\frac{x}{50}\right)^2) \\ &= \begin{cases} 0 & \text{if } x < 50\sqrt{50} \\ \frac{(\frac{x}{50})^2 - 50}{100} & \text{if } 50\sqrt{50} \leq x \leq 50\sqrt{150} \\ 1 & \text{if } > 50\sqrt{150} \end{cases} \end{aligned}$$

(b) The probability that Diana breaks the world record is equal to the probability that her motorcycle travels a distance greater than 500 metres in 10 seconds. So,

$$P(X > 500) = 1 - F_X(500) = 1 - 0.5 = 0.5$$

(c) The probability that Diana breaks the world record in n trials is equal to the probability that the maximum distance traveled by her motorcycle in the n trials is greater than 500 metres in 10 seconds. Let Y denote the maximum distance traveled by her motorcycle. We have that $Y = \max(X_1, X_2, ..., X_n)$, where the X_i are independent and equally distributed random variables with X_i the distance travelled in trial i.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\max(X_1, X_2, ..., X_n) \leq y) \\ &= P(X_1 \leq y \text{ and } X_1 \leq y \text{ and } \cdots X_n \leq y) \\ &= P(X_1 \leq y)P(X_2 \leq y) \cdots P(X_n \leq y) \\ &= F_X(y)^n \end{aligned}$$

Therefore,

$$P(Y > 500) = 1 - F_Y(500) = 1 - F_X(500)^n = 1 - 0.5^n$$

and the minimum n for which $1 - 0.5^n \geq 0.9$ is $n = 4$.