Quantitative Genomics and Genetics - Spring 2019
BTRY 4830/6830; PBSB 5201.01

Homework 2 (version 1)

Assigned January 31; Due 11:59PM February 6

Problem 1 (Easy)

a. Consider a probability mass function $P_X(x)$ and a probability density function $f_X(x)$. Explain why substituting a value for x in $P_X(x)$ will result in an output that is the probability of x while the output of substituting a value for x in $f_X(x)$ will not result in the probability of x.

$P_X(x)$ is a discrete function across the integers that represents the probabilities of all possible outcomes for a given experiment. Therefore, inputting a given outcome x will output a real probability value. $f_X(x)$ on the other hand is a continuous function across the reals that can only be used to assign probabilities to all possible intervals of experimental outcomes. Therefore, inputting any x into $f_X(x)$ will only return the value 0. Reading off values from the y-axis of $f_X(x)$ is not informative of a given probability at a point, it is only the result of the graph that must be drawn from the equation for $f_X(x)$.

b. Consider a random variable X that has a normal probability distribution $X \sim N(\mu, \sigma^2)$ for which $\mu = 0$. What is the value of the cumulative density function $F_X(x)$ at $x = 0$? Provide an intuitive explanation (no formulas!) as to why you know that this is the answer in terms of the 'shape' of the normal distribution and how a cumulative density function works.

$F_X(x)$ at $x = 0$ is 0.5. The normal distribution is a bell curve that has a peak at μ and curves downward evenly from this point then asymptotes at 0. Therefore, the probability at μ for a normal distribution is 0.5. In the cdf this will look like the middle of the sigmoid function.

Problem 2 (Medium)

Consider a coin that you plan to learn about where you perform a 'two flips' experiment. Assume the probability model (on the Sigma-algebra!) is defined by the following structure $Pr(HH) = Pr(HT) = Pr(TH) = Pr(TT) = 0.25$ and define the random variables X_1 that is '2 times the number of heads' and X_2 that is '0.5 times the number of heads on the first flip'.

a. Write out the probability mass functions $P_{X_1}(x_1)$ and $P_{X_2}(x_2)$. That is, write out each of the value of these functions for each of the possible values of X_1 and X_2, respectively.

Outcome	X_1	X_2
HH	4	0.5
HT	2	0.5
TH	2	0
TT	0	0

Outcome	X_1	X_2	$Pr(X_1, X_2)$
HH	4	0.5	0.25
HT	2	0.5	0.25
TH	2	0	0.25
TT	0	0	0.25

$Pr(X_1 = 0) = 0.25$ $Pr(X_1 = 2) = 0.5$ $Pr(X_1 = 4) = 0.25$ $Pr(X_2 = 0) = 0.5$ $Pr(X_2 = 0.5 = 0.5$

b. Write out the 'jumps' of the cumulative mass functions $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$. That is, write out the value of these functions for each of the possible values of X_1 and X_2, respectively. $F_{X_1}(X_1 \leq 0) = 0.25$ $F_{X_1}(X_1 \leq 2) = 0.75$ $F_{X_1}(X_1 \leq 4) = 1.0$ $F_{X_2}(X_2 \leq 0) = 0.5$ $F_{X_2}(X_2 \leq 0.5) = 1.0$

c. Calculate the expected values for X_1 and X_2 (show your work using the formula for the expected value!!).

$E(X) = \sum_{min}^{max} Pr(X = i)(X = i)$ $E(X_1) = Pr(X_1 = 0) * 0 + Pr(X_1 = 2) * 2 + Pr(X_1 = 4) * 4$
$= 0 + 0.5 * 2 + 0.25 * 4 = 2$ $E(X_2) = Pr(X_2 = 0) * 0 + Pr(X_2 = 0.5) * 0.5$
$= 0 + 0.5 * 0.5 = 0.25$

d. Calculate the variances for X_1 and X_2 (show your work using the formula for variance!!).

$Var(X) = \sum_{min}^{max}((X = i) - EX)^2 Pr(X = i)$ $Var(X_1) = (0-2)^2 * 0.25 + (2-2)^2 * 0.5 + (4-2)^2 * 0.25$
$= 4 * 0.25 + 0 + 4 * 0.25)$
$= 2$ $Var(X_2) = (0 - 0.25)^2 * 0.5 + (0.5 - 0.25)^2 * 0.5$
$= 2(0.5 * 0.0625)$
$= 0.0625$

e. Write out the values for the joint probability mass function $P_{X_1,X_2}(x_1, x_2)$ for all possible values of the random vector $[X_1, X_2]$.

This makes sense because there is only one way to achieve each of the above combinations of r.v. values based on the definitions of X_1, X_2. Also $Pr(X_1 = 4, X_2 = 0) = Pr(X_1 = 0, X_2 = 0.5) = 0$ because these events cannot occur together.

f. Are X_1 and X_2 independent? Justify your answer using an appropriate equation. No the r.v.s are not independent. Intuitively, knowing one of the variables will give you information about the other because both of them are tracking the occurrence of heads (e.g. $X_1 = 0$ tells you $X_2 = 0$ because you cannot have a head on the first flip if no flips are heads. This also follows from conditional probabilities and their ability to prove independence. Where $Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$ and $Pr(A \cap B) = Pr(A) * Pr(B)$ so $Pr(A|B) = \frac{Pr(A)Pr(B)}{Pr(B)} = Pr(A)$ This does not hold for the given problem. For example $Pr(X_2 = 0|X_1 = 2) = \frac{Pr(\{TH,TT\} \cap \{HT,TH\})}{Pr(B)} = \frac{Pr(\{TH\})}{Pr(\{TH,TT\})} = \frac{0.25}{0.5} = 0.5$

g. Calculate the covariance of X_1 and X_2 (show your work using the formula for covariance!!).
$Cov(X_1, X_2) = \sum_{minX_1}^{maxX_1} \sum_{minX_2}^{maxX_2}((X_1 = i) - EX_1)((X_2 = i) - EX_2)P_{X_1,X_2}(x_1, x_2)$
$= (4-2)(0.5-0.25)(0.25) + (4-2)(0-0.25)*0 + (2-2)(0.5-0.25)(0.25) + (2-2)(0-0.25)(0.25) + (0-2)(0.5-0.25)*0 + (0-2)(0-0.25)(0.25)$
$= (2)(0.25)(0.25) + (-2)(-0.25)(0.25) = 2*2*0.25*0.25$
$= 4 * \frac{1}{4} * 0.25 = 0.25$

h. Provide an intuitive explanation as to why the answer to part [g] makes sense given your answer to part [f].

This value falls between the variances of the two distributions for the r.v.s alone. It seems to me that since X_2 and X_1 are not independent, as X_2 can add information to X_1 its smaller variance could help shrink the spread of the values and vice versa.

i. Provide an intuitive explanation as to why the sign of the covariance in part [g] makes sense given the probability distribution of $[X_1, X_2]$ in part [e].

Covariance is related to how variables change with one another. The covariance is positive, indicating big values of X_1 correspond to big values of X_2. This makes sense because both variables represent a counting of heads.

j. Provide the formula for a family of probability distributions that could represent the (univariate) probabilities of $0.5 * X_1$ and the value(s) of the parameter(s) that produce the probability model of $0.5 * X_1$. Do the same for $2 * X_2$.

X_1 follows a normal distribution of the family $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ with $\mu = 2, \sigma^2 = 2$

X_2 follows a Binomial distribution $y = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$ with $p = 0.5, n = 2$

Problem 3 (Difficult)

a. For any bivariate, finite, discrete distribution $P_{X_1,X_2}(x_1, x_2)$, prove that if X_1 and X_2 are independent, then $Cov(X_1, X_2) = 0$. Hint: for X_1 and X_2 to be independent, there must be the same set of values of X_1 paired with every X_2 with non-zero probability (and vice versa)!

$Cov(X_1, X_2) = \sum_{minX_1}^{maxX_1} \sum_{minX_2}^{maxX_2}((X_1 = i) - EX_1)((X_2 = i) - EX_2)P_{X_1,X_2}(x_1, x_2)$

and for r.v.s to be independent, the events they represent must follow $Pr(X_1 \cap X_2) = Pr(X_1)Pr(X_2)$ So $Pr(X_1, X_2)$ is a constant and can be factored out of the Cov summation. Giving $Cov(X_1, X_2) = P_{X_1,X_2}(x_1, x_2) \sum_{minX_1}^{maxX_1} \sum_{minX_2}^{maxX_2}((X_1 = i) - EX_1)((X_2 = i) - EX_2)$ This sum will evenly subtract the respective expectations from X_1 and X_2 such that there will be a negative and positive value that will cancel each pair. For example, if X_1 = num heads on first flip X_2 = num heads on second flip. You could factor out the probability of each combination of these $Pr(X_1, X_2) = 0.25$) and would sum $(1-0.5)(1-0.5) + (1-0.5)(0-0.5) + (0-0.5)(1-0.5) + (0-0.5)(0-0.5)$ or $0.25 - 0.25 - 0.25 + 0.25 = 0$.

b. Show that the converse of the statement in part [a] need not be true (i.e., it is possible for two random variables to have a zero covariance but that are non-independent) by defining a random variable X_3 for problem 2 above, such that $Pr(X_1, X_3) \neq Pr(X_1)Pr(X_3)$ but $Cov(X_1, X_3) = 0$ (show both of these relationships as part of your answer!).

You can define X_3 in another way and still make the sum cancel to zero. For example, if $X_3 = 1$ if the flips are the same, the r.v.s are inherently not independent, but the covariance calculation still cancels. For example, $Pr(X_3 = 1|X_1 = 4) = 1$ $Pr(X_3 = 1|X_1 = 4) \neq Pr(X_3 = 1)Pr(X_1 = 4) = 0.5 * 0.25 = 0.125$

Nevertheless, $Cov(X_1, X_3) = (4-2)(1-0.5)(0.25) + 0 * 4terms + (0-2)(1-0.5)(0.25) = 0$