Quantitative Genomics and Genetics - Spring 2016
BTRY 4830/6830; PBSB 5201.01

Homework 2 (version 1) - Key

Assigned February 9; Due 11:59PM February 15

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 defined for probability while substituting a value for x in $f_X(x)$ will not.

Among the acceptable answers: A probability mass function directly maps values of a discrete random variable to the probability of that specific value, such that substituting a value into $P_X(x)$ returns the probability of the input. A probability density function is used to calculate the probability of intervals by taking the integral of an interval, where intervals can have non-zero values and such that a probability of zero is assigned each point value. Inputting a individual value into a probability density function $f_X(x)$ therefore does not produce a valid result for probability, e.g., a point input can map to a non-zero value.

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.

Among the acceptable answers: $F_X(0) = 0.5$ since the normal distribution with $\mu = 0$ is symmetric around zero (half of the density is on each side of $X = 0$) and since the cumulative density function integrates from $-\infty$ to the input value (in this case zero), the result is half of the density of the random variable.

Problem 2 (Medium)

Consider a coin that you plan to learn about where you perform a 'two flips of the coin' experiment. Assume the probability model 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.

$P_{X_1}(0) = 0.25, P_{X_1}(2) = 0.5, P_{X_1}(4) = 0.25$
$P_{X_2}(0) = 0.5, P_{X_2}(0.5) = 0.5$

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 each of the value of these functions for each of the possible values of X_1 and X_2, respectively.

$F_{X_1}(0) = 0.25, F_{X_1}(2) = 0.75, F_{X_1}(4) = 1.0$
$F_{X_2}(0) = 0.5, F_{X_2}(0.5) = 1.0$

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

$\text{E}X_1 = \sum_{i=0}^{2} X_i P_{X_1}(x) = 0 * P_{X_1}(0) + 1 * P_{X_1}(2) + 2 * P_{X_1}(4) = 0 * 0.25 + 2 * 0.5 + 4 * 0.25 = 2$
$\text{E}X_2 = \sum_{i=0}^{1} X_i P_{X_2}(x) = 0 * P_{X_2}(0) + 0.5 * P_{X_2}(1) = 0 * 0.5 + 0.5 * 0.5 = 0.25$

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

$\text{Var}(X_1) = \sum_{i=0}^{2} (X_i - \text{E}X)^2 P_{X_1}(x) = (0-2)^2 * 0.25 + (2-2)^2 * 0.5 + (4-2)^2 * 0.25 = 2$
$\text{Var}(X_2) = \sum_{i=0}^{2} (X_i - \text{E}X)^2 P_{X_2}(x) = (0-0.25)^2 * 0.5 + (0.5-0.25)^2 * 0.5 = 0.0625$

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 $[X_1, X_2]$.

$P_{X_1,X_2}(0,0) = 0.25, P_{X_1,X_2}(2,0) = 0.25, P_{X_1,X_2}(4,0) = 0$
$P_{X_1,X_2}(0,0.5) = 0, P_{X_1,X_2}(2,0.5) = 0.25, P_{X_1,X_2}(4,0.5) = 0.25$

f. Are X_1 and X_2 independent? Use the definition of conditional probability to justify your answer.

No - $Pr(X_1, X_2) = Pr(X_1)Pr(X_2)$ for all cases
e.g., $P_{X_1,X_2}(0,0) = 0 \neq P_{X_1}(0) * P_{X_2}(0) = 0.125$ or $P_{X_1|X_2}(0,0) = 0 \neq P_{X_1}(0) = 0.25$

g. Calculate the covariance of X_1 and X_2 (show your work!!).

$\text{Cov}(X_1, X_2) = -\text{E}X_1\text{E}X_2 + \sum_{i=0}^{2} \sum_{j=0}^{1} (X_1 = i) * (X_2 = j) P_{X_1,X_2}(x_i, x_j)$
$-2 * 0.25 + 0 * 0 * 0.25 + 2 * 0 * 0.25 + 4 * 0 * 0 + 0 * 0.5 * 0 + 2 * 0.5 * 0.25 + 4 * 0.5 * 0.25 = 0.25$

h. Provide an intuitive explanation as to why the answer to part 'g' makes sense given your answer to 'f' and why the sign of the covariance makes sense given the distribution of $[X_1, X_2]$.

The random variables X_1 and X_2 are not independent so the covariance will not be zero. Large values of X_1 have a high probability of occurring with high values of X_2 and vice versa, such that it makes sense that the covariance is positive.

i. 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 $2 * X_2$.

In both cases, the binomial distribution could represent the probability distributions with parameters $n = 2, p = 0.5$ for $0.5 * X_1$ and $n = 1, p = 0.5$ for $2 * X_2$.

j. Considered together, do the two univariate distributions in part 'i' completely describe the probability distribution $P_{X_1,X_2}(x_1, x_2)$? Explain your answer.

No - these univariate distributions do not capture the non-independence relationship between X_1 and X_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)!

If X_1 and X_2 are independent, then every possible combination of values taken by X_1 and X_2 has a non-zero probability such that we can write:

$\text{Cov}(X_1, X_2) = -\text{E}X_1\text{E}X_2 + \sum\sum X_1 * X_2 Pr(X_1, X_2)$

$= -\text{E}X_1\text{E}X_2 + \sum X_1(\sum X_2)Pr(X_1, X_2)$

and then by the independence relationship $Pr(X_1, X_2) = Pr(X_1)Pr(X_2)$ we can write:

$= -\text{E}X_1\text{E}X_2 + \sum X_1(\sum X_2)Pr(X_1)Pr(X_2)$

$= -\text{E}X_1\text{E}X_2 + \sum X_1(\sum X_2 Pr(X_2))Pr(X_1)$

$= -\text{E}X_1\text{E}X_2 + \sum X_1(\text{E}X_2)Pr(X_1)$

$= -\text{E}X_1\text{E}X_2 + (\text{E}X_2)\sum X_1 Pr(X_1)$

$= -\text{E}X_1\text{E}X_2 + (\text{E}X_2)\text{E}X_1$

$= 0$

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!).

Define X_3 to equal '0' if the two flips are the same, '1' if they are different:

$P_{X_3}(0) = 0.5, P_{X_3}(1) = 0.5$

$P_{X_1,X_3}(0,0) = 0.25, P_{X_1,X_3}(2,0) = 0, P_{X_1,X_3}(4,0) = 0.25$
$P_{X_1,X_3}(0,1) = 0, P_{X_1,X_3}(2,1) = 0.5, P_{X_1,X_3}(4,1) = 0$

These are not independent:
e.g., $P_{X_1,X_3}(0,1) = 0 \neq P_{X_1}(0) * P_{X_3}(1) = 0.125$

But the covariance is zero:

$\mathrm{E}X_3 = 0 * P_{X_3}(0) + 1 * P_{X_2}(1) = 0 * 0.5 + 1 * 0.5 = 0.5$

$\mathrm{Cov}(X_1, X_2) = -\mathrm{E}X_1\mathrm{E}X_3 + \sum_{i=0}^{2}\sum_{j=0}^{1}(X_1 = i) * (X_3 = j)P_{X_1,X_3}(x_i, x_j)$
$-2 * 0.5 + 0 * 0 * 0.25 + 2 * 0 * 0 + 4 * 0 * 0.25 + 0 * 1 * 0 + 2 * 1 * 0.5 + 4 * 1 * 0 = 0$