BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2014

Homework 1 (version 3) - Key

Assigned September 3; V2 posted September 4 (minor correction); V3 posted September 8 (minor correction); Due 11:59PM September 8

Problem 1 (Easy)

a. Consider a case where we have a system of interest, where we have defined a question that we would like to answer, and an experiment that we will use to answer this question. For the sample space for this experiment and the associated Sigma-algebra, we have many options for how we could define a probability model. Explain why we might justify assuming one specific probability model over another (many acceptable answers)?

   Among the acceptable answers: We have previously observed the system (or a similar system) and given our previous observations of outcomes resulting from similar experiments, we have observed a frequency of outcomes that we have used to define our probability model.

   We have previously observed the system (or a similar system) and given how we think the system works, we have used this knowledge to define our probability model.

   We have not observed the system but based on the types of outcomes, we choose a probability model where the possible outcomes have reasonable (e.g. non-zero) probabilities of occurrence.

   etc.

b. Using two sentences at most, provide an intuitive explanation as to why disjoint events (that have non-zero probabilities) cannot be independent. Do not write down any formulas for this intuitive explanation. Among the acceptable answers: an example of disjoint sets are H and T of a single coin flip because only one of these outcomes can occur at a time. These events are not independent because knowing that the result of a flip is H provides us information concerning the probability that the flip is T, specifically that the probability is zero.

   Independence is defined as a specific relationship between two events where if both have a non-zero probability, then the probability of one event and the other is non-zero. For disjoint sets, knowing that one has occurred means the probability of the other is zero.

   etc.

Problem 2 (Medium)

Assume that the system we are interested in is a coin. The experiment we will consider is two flips of the coin. Note: for parts (d)-(j) use the probability model in part (c).

a. What is the sample space of this experiment?

$\Omega = \{HH, HT, TH, TT\}$

b. What is the Sigma-algebra (containing all events) for this sample space? Which of these events is the event 'the first flip is heads'? Which of these events is the event 'the second flip is heads'?

$\mathcal{F} = \emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\},$
$\{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\}.$
$\{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{TH, HT, TT\}$
$\{HH, TH, HT, TT\}$

$\{H_{1st}\} = \{HH, HT\}$

$\{H_{2nd}\} = \{HH, TH\}$

c. Define a probability model such that the probability of a 'heads' on the first flip and the second flip is $Pr(H_{1st}) = Pr(H_{2nd}) = 0.5$, where the probability of heads on both the first and second flip is $Pr(H_{1st} \cap H_{2nd}) = 0.3$. Write out the probabilities for all possible outcomes of an experimental trial. Write down the formulas or relationships you used to calculate these probabilities as part of your answer.

	H_{2nd}	T_{2nd}	
H_{1st}	$Pr(H_{1st} \cap H_{2nd})$	$Pr(H_{1st} \cap T_{2nd})$	$Pr(H_{1st})$
T_{1st}	$Pr(T_{1st} \cap H_{2nd})$	$Pr(T_{1st} \cap T_{2nd})$	$Pr(t_{1st})$
	$Pr(H_{2nd})$	$Pr(T_{2nd})$	

	H_{2nd}	T_{2nd}	
H_{1st}	0.3	$Pr(H_{1st})$ - $Pr(H_{1st} \cap H_{2nd})$	0.5
T_{1st}	$Pr(H_{2nd})$ - $Pr(H_{1st} \cap H_{2nd})$	Eq	0.5
	0.5	0.5	

Eq = 1 - $Pr(H_{1st} \cap H_{2nd})$ + $Pr(H_{1st} \cap T_{2nd})$ + $Pr(T_{1st} \cap H_{2nd})$

	H_{2nd}	T_{2nd}	
H_{1st}	0.3	0.2	0.5
T_{1st}	0.2	0.3	0.5
	0.5	0.5	

d. Define the random variables X_1 that is the 'number of heads on the first flip' and X_2 that is the 'number of heads on the second flip'. For X_1 and X_2, write out the values that these random variables take for all possible outcomes of an experimental trial.

ANS: $X_1(HH) = 1, X_1(HT) = 1, X_1(TH) = 0, X_1(TT) = 0$
$Pr(X_1 = 0) = Pr(H_{1st} \cap H_{2nd}) + Pr(H_{1st} \cap T_{2nd}) = 0.5,$
$Pr(X_1 = 1) = 1 - Pr(X_1 = 0) = 0.5$
ANS: $X_1(HH) = 1, X_1(HT) = 0, X_1(TH) = 1, X_1(TT) = 0$
$Pr(X_2 = 0) = Pr(H_{1st} \cap H_{2nd}) + Pr(T_{1st} \cap H_{2nd}) = 0.5,$
$Pr(X_2 = 1) = 1 - Pr(X_2 = 0) = 0.5$

e. Are X_1 and X_2 independent? Demonstrate that your answer is correct.
$Pr(X_1 = 0) * Pr(X_2 = 0) = 0.5 * 0.5 \neq Pr(T_{1st} \cap T_{2nd}) = 0.3$
Note that writing out one relationship that does not conform to independence demonstrates non-independence.

f. Calculate $\mathrm{E}X_1$, $\mathrm{E}X_2$, $Var(X_1)$, $Var(X_2)$, $Cov(X_1, X_2)$, and $Corr(X_1, X_2)$. Show your work for each calculation.

$$\mathrm{E}X_1 = 0 * Pr(X_1 = 0) + 1 * Pr(X_1 = 1) = 0 * 0.5 + 1 * 0.5 = 0.5 \tag{1}$$

$$\mathrm{E}X_2 = 0 * Pr(X_2 = 0) + 1 * Pr(X_2 = 1) = 0 * 0.5 + 1 * 0.5 = 0.5 \tag{2}$$

$$Var(X_1) = \sum_{X_1=0}^{1} (X_1 - \mathrm{E}X_1)^2 Pr(X_1) = (0 - 0.5)^2 * 0.5 + (1 - 0.5)^2 * 0.5 = 0.25 \tag{3}$$

$$Var(X_2) = \sum_{X_2=0}^{1} (X_1 - \mathrm{E}X_1)^2 Pr(X_1) = (0 - 0.5)^2 * 0.5 + (1 - 0.5)^2 * 0.5 = 0.25 \tag{4}$$

$$Cov(X_1, X_2) = \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} (X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2) Pr(X_1, X_2) \tag{5}$$

$$= (0-0.5)(0-0.5)*0.3 + (0-0.5)(1-0.5)*0.2 + (1-0.5)(0-0.5)*0.2 + (1-0.5)(1-0.5)*0.3 = 0.05 \tag{6}$$

$$Corr(X_1, X_2) = Cov(X_1, X_2)/\sqrt{Var(X_1) * Var(X_2)} = 0.2 \tag{7}$$

g. Define a new the random variables X_3 that is 'the number of heads observed for both flips'. Write out the values that this random variable takes for all possible outcomes of an experimental trial and the probabilities of each outcome.

$X_3(HH) = 2, X_3(HT) = 1, X_3(TH) = 1, X_3(TT) = 0$
$Pr(X_3 = 0) = 0.3, Pr(X_3 = 1) = 0.4, Pr(X_3 = 2) = 0.3$

h. Are X_1 and X_3 independent? Demonstrate that your answer is correct.
$Pr(X_1 = 1) * Pr(X_3 = 0) = 0.5 * 0.3 \neq Pr(T_{1st} \cap T_{2nd}) = 0$
Note that writing out one relationship that does not conform to independence demonstrates non-independence.

i. Calculate $Cov(X_1, X_3)$ and show your work.

$$\mathrm{E}X_3 = 0 * Pr(X_3 = 0) + 1 * Pr(X_3 = 2) + 1 * Pr(X_3 = 2) = 0 * 0.3 + 1 * 0.4 + 2 * 0.3 = 1 \quad (8)$$

$$Cov(X_1, X_3) = \sum_{X_1=0}^{1} \sum_{X_3=0}^{1} (X_1 - \mathrm{E}X_1)(X_3 - \mathrm{E}X_1) Pr(X_1, X_3) \quad (9)$$

$$= (0 - 0.5)(0 - 1) * 0.3 + (0 - 0.5)(1 - 1) * 0.2 + (0 - 0.5)(2 - 1) * 0 + \quad (10)$$

$$(1 - 0.5)(0 - 1) * 0 + (1 - 0.5)(1 - 1) * 0.2 + (1 - 0.5)(2 - 1) * 0.3 = 0.3 \quad (11)$$

j. Your answer to (i) should be a positive number. Using no more than two sentences, provide an intuitive explanation as to why this is the case.

A positive covariance means there is a higher probability of large numbers occurring with large numbers AND small numbers occurring with small numbers COMPARED to large numbers occurring with small AND small occurring with small. In this case, the probability of large occurring with small = small occurring with large = 0, which is smaller than large with large AND small with small.

Problem 3 (Difficult)

Consider two random variables X_1 and X_2 that have a joint probability distribution function $Pr(X_1, X_2)$. Prove that if

$$Cov(X_1, X_2) \neq 0 \quad (12)$$

then X_1 and X_2 are NOT independent. Note that for full credit, you must demonstrate this for the general case, i.e. you will not get full credit for a specific example.

We will prove this by contradiction, specifically, we will prove that $Cov(X_1, X_2) = 0$ if these two random variables are independent.

When we have:

$$Pr(X_1, X_2) = Pr(X_1)Pr(X_2) \quad (13)$$

if two random variables are independent and from the definition of covariance, we have:

$$Cov(X_1, X_2) = \sum \sum (X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2) Pr(X_1, X_2) \quad (14)$$

such that assuming these are independent we have:

$$Cov(X_1, X_2) = \sum \sum (X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2) Pr(X_1) Pr(X_2) \quad (15)$$

$$Cov(X_1, X_2) = \sum\sum Pr(X_1)(X_1 - \mathrm{E}X_1)Pr(X_2)(X_2 - \mathrm{E}X_2) \tag{16}$$

(check that the following step is true!)

$$Cov(X_1, X_2) = (\sum Pr(X_1)(X_1 - \mathrm{E}X_1))(\sum Pr(X_2)(X_2 - \mathrm{E}X_2)) \tag{17}$$

$$Cov(X_1, X_2) = (\sum Pr(X_1)(X_1) - \sum Pr(X_1)(\mathrm{E}X_1))(\sum Pr(X_2)(X_2) - \sum Pr(X_2)(\mathrm{E}X_2)) \tag{18}$$

by the definition of expected values and the sum of probabilities is one:

$$Cov(X_1, X_2) = (\mathrm{E}X_1 - \mathrm{E}X_1)(\mathrm{E}X_2 - \mathrm{E}X_2) \tag{19}$$

$$Cov(X_1, X_2) = 0 * 0 = 0 \tag{20}$$

Thus, $Cov(X_1, X_2) = 0$ if X_1 and X_2 are independent, if $Cov(X_1, X_2) \neq 0$, the two random variables are not independent.

NOTE (!!) this does not prove the opposite case, that is if $Cov(X_1, X_2) = 0$ we have not demonstrated that this means X_1 and X_2 are independent (why?).