CS 70 Discrete Mathematics and Probability Theory
Summer 2011 Kamil HW 8

Due Wednesday, August 10, 4:59pm

Note the special due date.

1. **(10 pts.) The Basics: Joint Distributions.**
Consider figure 1 in Note 17 in the course reader.

   (a) What are the distributions of X and Y?

   (b) What is $\Pr[X = 1 \cap Y = 1]$?

   (c) What is $\Pr[X = 1 | Y = 1]$?

   (d) What is the distribution of the random variable Y conditioned on $Y = 1$?

   (e) What is the distribution of the random variable X conditioned on $Y = 1$?

   (f) Say X represents one of three variants of a gene, and $Y = 0$ represents the event that an individual is healthy, $Y = 1$ has a type 1 variant of a disease, and $Y = 2$ has a type 2 variant of the disease.

   What is the probability that an individual is sick? Given that X is not 1, what is the probability that the individual is sick?

2. **(10 pts.) The myth of fingerprints**
A crime has been committed. The police discover that the criminal has left DNA behind, and they compare the DNA fingerprint against a police database containing DNA fingerprints for 20 million people. Assume that the probability that two DNA fingerprints (falsely) match by chance is 1 in 10 million. Assume that, if the crime was committed by someone whose DNA fingerprint is on file in the police database, then it's certain that this will turn up as a match when the police compare the crime-scene evidence to their database; the only question is whether there will be any false matches.

   Let D denote the event that the criminal's DNA is in the database; $\overline{D}$ denotes the event that the criminal's DNA is not in the database. Assume that it is well-documented that half of all such crimes are committed by criminals in the database, i.e., assume that $\Pr[D] = \Pr[\overline{D}] = 1/2$. Let the random variable X denote the number of matches that are found when the police run the crime-scene sample against the DNA database.

   (a) Calculate $\Pr[X = 1 | D]$.

   (b) Calculate $\Pr[X = 1 | \overline{D}]$.

   (c) Calculate $\Pr[\overline{D} | X = 1]$. Evaluate the expression you get and compute this probability to at least two digits of precision.

   As it happens, the police find exactly one match, and promptly prosecute the corresponding individual. You are appointed a member of the jury, and the DNA match is the only evidence that the police present. During the trial, an expert witness testifies that the probability that two DNA fingerprints (falsely) match by chance is 1 in 10 million. In his summary statement, the prosecutor tells the jury that this means that the probability that the defendant is innocent is 1 in 10 million.

(d) What is wrong with the prosecutor's reasoning in the summary statement?

(e) Do you think the defendant should be convicted? Why or why not?

3. (15 pts.) Mendelian inheritance

Let's look at an application of the theory we've been developing, to classical inheritance of genes, as initially described by Gregor Mendel based upon his experiments with pea plants.

Here's a review of Mendel's model. Mendel determined that the height of each pea plant is determined by the genes it has. In his model, each pea plant has two genes that together determine the plant's height; each gene can be one of two possibilities, either h or H. The H gene is dominant, and if H is present in either of the plant's two genes, the plant will be tall. The h gene is recessive, and if both genes are h, the plant will be short.

We can see that there are four possibilities for the combinations of the plant's two genes: HH, Hh, hH, hh. The combinations Hh and hH are indistinguishable, so the standard convention is to write both of these cases as simply Hh. Thus, there are three possibilities for the plant's genotype (the portion of its genetic code related to height): HH, Hh, and hh.

The height of each plant is determined by its genotype. Plants with a genotype of HH or Hh will be tall; plants with the genotype hh will be short.

A new pea plant can be formed by crossing two existing pea plants: its "father" and its "mother". The "child" plant inherits one gene from its father (a gene chosen uniformly at random from its father's two genes) and one gene from its mother (randomly chosen from the mother's two genes). For each parent, it's random which gene the child inherits from that parent, and both possibilities are equally likely. For example, if the father has genotype HH and the mother has genotype Hh, the child might have genotype HH or Hh, each with probability $1/2$. As another example, if the father has genotype Hh and the mother has genotype Hh, the child's genotype could be HH, Hh, or hh; these occur with probabilities $1/4$, $1/2$, and $1/4$, respectively. (Here Hh occurs with probability $2/4$, since it can be obtained either by inheriting a H from the father and h from the mother, or by inheriting a h from the father and H from the mother.)

Mendel deduced that, in a large population of pea plants, the genotypes will be HH, Hh, and hh, in proportions $1/4$, $1/2$, and $1/4$, respectively. Thus, $3/4$ of pea plants (the ones with genotype HH or Hh) will be tall, and $1/4$ will be short. Of the tall pea plants, $1/3$ will have genotype HH and $2/3$ will have genotype Hh.

Here's the thing. Given a pea plant, you can directly measure whether it is tall or short. However, there is no easy way to determine its genotype directly. Of course, based upon the pea plant's height, you can draw some inferences about its potential genotypes, but there is no way to observe genotypes directly (without sophisticated technology that was not available to Mendel). In this problem, we're going to develop a procedure for probabilistically inferring the genotype of a pea plant, by crossing it with other plants and measuring the heights of its children.

So suppose we have a particular pea plant, let's call it Penelope; we want to infer Penelope's genotype. Let the random variable X denote Penelope's genotype. X is hidden: we cannot observe X directly. Based upon the overall frequency of genotypes in the population at large, before we observe anything, our best estimate for X can be summarized by the prior distribution: $\Pr[X = HH] = 1/4$, $\Pr[X = Hh] = 2/4$, $\Pr[X = hh] = 1/4$. Now we're going to repeatedly pick another plant at random from a large population of pea plants, cross that other plant with Penelope, and look at the height of the child. Let the random variable Y_i be an indicator r.v. for the event that the ith child obtained in this way is tall. In other words, $Y_i = 1$ if Penelope's ith child is tall, and $Y_i = 0$ if Penelope's ith child is short. Assume that the genotypes of all the other plants crossed with Penelope are independent.

(a) Calculate the conditional probabilities

$$\Pr[Y_1 = 1 | X = HH] \qquad \Pr[Y_1 = 0 | X = HH]$$
$$\Pr[Y_1 = 1 | X = Hh] \qquad \Pr[Y_1 = 0 | X = Hh]$$
$$\Pr[Y_1 = 1 | X = hh] \qquad \Pr[Y_1 = 0 | X = hh]$$

(b) What is the probability that Penelope's first child is tall? In other words, calculate $\Pr[Y_1 = 1]$.

(c) Suppose we measure and find that Penelope's first child is tall. What is the posterior distribution for X, given this observation? In other words, calculate the conditional probabilities

$$\Pr[X = HH | Y_1 = 1], \qquad \Pr[X = Hh | Y_1 = 1], \quad \text{and} \quad \Pr[X = hh | Y_1 = 1].$$

(d) In part (c), you determined a method for updating the prior distribution to the posterior distribution after observing the event that Penelope's first child is tall, under the assumption that the prior distribution is $(1/4, 2/4, 1/4)$. Now let's generalize this to an arbitrary prior distribution. Suppose the prior distribution is $\Pr[X = HH] = p$, $\Pr[X = Hh] = q$, $\Pr[X = hh] = 1 - p - q$. With this prior, calculate the posterior distribution

$$(\Pr[X = HH | Y_1 = 1], \quad \Pr[X = Hh | Y_1 = 1], \quad \Pr[X = hh | Y_1 = 1]),$$

as a function of p and q. (This provides a general update rule for updating your estimate of the distribution of X, after observing a tall child.)

Note: The new prior distribution $(p, q, 1 - p - q)$ in parts (d) and (e) only applies to Penelope. The remaining plants (from which we draw the plant to cross with Penelope) still have the distribution $(1/4, 1/2, 1/4)$.

(e) In part (d), you developed an update rule for the case where the child is observed to be tall. Now develop a general update rule for the case where the child is observed to be short. Suppose the prior distribution is $\Pr[X = HH] = p$, $\Pr[X = Hh] = q$, $\Pr[X = hh] = 1 - p - q$. With this prior, calculate the posterior distribution

$$(\Pr[X = HH | Y_1 = 0], \quad \Pr[X = Hh | Y_1 = 0], \quad \Pr[X = hh | Y_1 = 0]),$$

as a function of p and q.

(f) Suppose that, after measuring, we find Penelope's first two children are both tall. Calculate the conditional distribution for X, given that Penelope's first two children are both tall: i.e., calculate the posterior distribution

$$(\Pr[X = HH | Y_1 = 1, Y_2 = 1], \quad \Pr[X = Hh | Y_1 = 1, Y_2 = 1], \quad \Pr[X = hh | Y_1 = 1, Y_2 = 1]).$$

Plot this distribution.

Hint: Apply the update rule from part (d) to the distribution you calculated in part (c).

4. (15 pts.) Likelihood ratios

Gamblers tend to express their chances of winning a bet in terms of odds: for instance, a bet that's one to six is one where we're six times as likely to lose as to win. In contrast, a mathematician would probably describe the same bet by saying that the probability of winning is $1/7$ and the probability of losing is $6/7$. Of course, the two are equivalent: given the odds, you can compute the probabilities of winning and losing; giving the probabilities, you can compute the odds. In this class, we've used probabilities, because they tend to be easier to work with. But here we will see a case where odds can aid calculations. In particular, we will work with a ratio of probabilities, which can be viewed as a description of the odds.

Suppose we have an inference problem with a hidden r.v. X and observable r.v.'s $Y_1, \ldots, Y_n$ that are conditionally independent given X. Suppose we have two hypotheses regarding the value of X, which we will represent as two events H_0 and H_1. (For instance, maybe H_0 is the event $X = 5$ and H_1 is the event $X = 6$.)

In any probability space, we can compute the ratio of the probability of H_0 over the probability of H_1. For example, the probability ratio associated with the prior distribution is $R_0 = \frac{\Pr[H_0]}{\Pr[H_1]}$. The probability ratio associated with the posterior after observing the event $Y_1 = a_1$ is $R_1 = \frac{\Pr[H_0|Y_1=a_1]}{\Pr[H_1|Y_1=a_1]}$. And, after observing the events $Y_1 = a_1, Y_2 = a_2, \ldots, Y_k = a_k$, the probability ratio associated with the posterior is $R_k = \frac{\Pr[H_0|Y_1=a_1,\ldots,Y_k=a_k]}{\Pr[H_1|Y_1=a_1,\ldots,Y_k=a_k]}$. If the probability ratio is greater than 1, that means that hypothesis H_0 is more likely than hypothesis H_1; and a probability ratio less than 1 means that H_0 is less likely than H_1. So, the probability ratio helps us decide between the two hypotheses.

In this problem, you will develop methods for calculating with ratios and then apply them to several inference problems.

(a) Suppose we know the prior distribution on X is $\Pr[X = 0] = \frac{2}{7}$, $\Pr[X = 1] = \frac{5}{7}$, and H_0 is the hypothesis $X = 0$ and H_1 the hypothesis that $X = 1$. Find the probability ratio R_0 of the prior.

(b) Suppose we know the probability ratio of the prior is given by $R_0 = 4$, where H_0 is the hypothesis $X = 0$ and H_1 the hypothesis that $X = 1$ and where we know that X can only take on the values 0 or 1. Find the prior distribution of X. In other words, compute $\Pr[X = 0]$ and $\Pr[X = 1]$.

(c) Suppose we observe the event $Y_1 = a_1$. Show that

$$R_1 = R_0 \times L_1,$$

where $L_1 = \frac{\Pr[Y_1=a_1|H_0]}{\Pr[Y_1=a_1|H_1]}$. ($L_1$ is called a *likelihood ratio*.) In other words, show that

$$\frac{\Pr[H_0|Y_1 = a_1]}{\Pr[H_1|Y_1 = a_1]} = \frac{\Pr[H_0]}{\Pr[H_1]} \times \frac{\Pr[Y_1 = a_1|H_0]}{\Pr[Y_1 = a_1|H_1]}.$$

(d) Suppose we observe the sequence of k events $Y_1 = a_1, Y_2 = a_2, \ldots, Y_k = a_k$. Show that

$$R_k = R_0 \times L_1 \times L_2 \times \cdots \times L_k,$$

where $L_i = \frac{\Pr[Y_i=a_i|H_0]}{\Pr[Y_i=a_i|H_1]}$. ($L_1 \times \cdots \times L_k$ is another likelihood ratio.) In other words, show that

$$\frac{\Pr[H_0|Y_1 = a_1, \ldots, Y_k = a_k]}{\Pr[H_1|Y_1 = a_1, \ldots, Y_k = a_k]} = \frac{\Pr[H_0]}{\Pr[H_1]} \times \prod_{i=1}^{k} \frac{\Pr[Y_i = a_i|H_0]}{\Pr[Y_i = a_i|H_1]}.$$

Hint: Use conditional independence.

(e) Let's apply the likelihood-ratio framework to the stripped-down version of the multi-armed bandit problem, as described in Lecture Note 17. Suppose $p_1 = 2/3$, $p_2 = 1/2$, H_0 is the event that $X = 1$, and H_1 is the event that $X = 2$. We're going to compare the likelihood that the biased coin was selected (H_0) to the likelihood that the fair coin was selected (H_1), given the outcomes from a sequence of k coin tosses using the coin that was selected. Assume a uniform prior on X. Calculate the following values:

$$R_0 = \frac{\Pr[H_0]}{\Pr[H_1]} = ?$$
$$\Pr[Y_i = H | H_0] = ?$$
$$\Pr[Y_i = T | H_0] = ?$$
$$\Pr[Y_i = H | H_1] = ?$$
$$\Pr[Y_i = T | H_1] = ?$$
$$L_i(H) = \frac{\Pr[Y_i = H | H_0]}{\Pr[Y_i = H | H_1]} = ?$$
$$L_i(T) = \frac{\Pr[Y_i = T | H_0]}{\Pr[Y_i = T | H_1]} = ?$$

If after observing the first k coin tosses, we find that h of them were Heads and $k-h$ were Tails, find a formula for R_k in terms of R_0, $L_i(H)$, and $L_i(T)$. Next, answer the following question by plugging into your formula: if we observe 11 Heads and 7 Tails, is it more likely that we are dealing with the $2/3$-biased coin (H_0) or the fair coin (H_1)?

5. (10 pts.) Lunch Date

Alice and Bob agree to try to meet for lunch between 12 and 1pm at their favorite sushi restaurant. Being extremely busy they are unable to specify their arrival times exactly, and can say only that each of them will arrive (independently) at a time that is uniformly distributed within the hour. In order to avoid wasting precious time, if the other person is not there when they arrive they agree to wait exactly fifteen minutes before leaving. What is the probability that they will actually meet for lunch? Phrase your solution using the language of continuous random variables introduced in Note 18.

6. (10 pts.) James Bond

James Bond, my favorite hero, has again jumped off a plane. In the lecture we assumed the plane moves at constant *velocity* from base A to base B, distance 100 km apart. Now suppose the plane takes off from the ground at base A, climbs at an angle of 45 degrees to an altitude of 10 km, flies at that altitude for a while, and then descends at an angle of 45 degrees to land at base B. All along the plane is assumed to fly at the same *speed*. James Bond jump off at a time uniformly distributed over the duration of the journey. You can assume that James Bond, being who he is, violates the laws of physics and descends vertically after he jumps.

(a) Is there enough information to compute the probability density function of the position Bond lands? If so, compute it. If not, specify any additional information you need and then compute it. Plot the density.

(b) Compute the expectation and variance of this position. How do they compare to the case when the plane flies at constant velocity (with no ascending and descending)?

7. (10 pts.) Exponential Distribution

We begin by proving two very useful properties of the exponential distribution. We then use them to solve a problem in photography.

(a) Let r.v. X have geometric distribution with parameter p. Show that, for any positive m, n, we have

$$\Pr[X > m + n \mid X > m] = \Pr[X > n].$$

This is the "memoryless" property of the geometric distribution. Why do you think this property is called memoryless?

(b) Let r.v. X have exponential distribution with parameter λ. Show that, for any positive s,t, we have

$$\Pr[X > s+t \mid X > t] = \Pr[X > s].$$

This is the "memoryless" property of the exponential distribution.

(c) Let r.v.'s X_1, X_2 be independent and exponentially distributed with parameters λ_1, λ_2. Show that the r.v. $Y = \min\{X_1, X_2\}$ is exponentially distributed with parameter $\lambda_1 + \lambda_2$. [Hint: work with cdf's.]

(d) You have a digital camera that requires two batteries to operate. You purchase n batteries, labeled $1, 2, \ldots, n$, each of which has a lifetime that is exponentially distributed with parameter λ and is independent of all the other batteries. Initially you install batteries 1 and 2. Each time a battery fails, you replace it with the lowest-numbered unused battery. At the end of this process you will be left with just one working battery. What is the expected total time until the end of the process? Justify your answer.

(e) In the scenario of part (d), what is the probability that battery i is the last remaining working battery, as a function of i?

8. (10 pts.) Normal Distribution

If a set of grades on a Discrete Math examination in an inferior school (not UC!) are approximately normally distributed with a mean of 64 and a standard deviation of 7.1, find:

(a) the lowest passing grade if the bottom 5% of the students fail the class;

(b) the highest B if the top 10% of the students are given A's.

NOTE: You may assume that if X is normal with mean 0 and variance 1, then $\Pr[X \leq 1.3] \approx 0.9$ and $\Pr[X \leq 1.65] \approx 0.95$.

9. (10 pts.) The normal approximation to the binomial

Suppose $B \sim \text{Binomial}(n, p)$, i.e., the r.v. B is binomially distributed: it is the number of heads after flipping n coins, with heads probability p. We have seen previously $\mathbb{E}(B) = np$ and $\text{Var}(B) = np(1-p)$. It turns out that, for large n, the binomial distribution B approximates the normal distribution with the same mean and variance. A standard rule of thumb is that the normal approximation is a reasonable approximation if $np \geq 5$ and $n(1-p) \geq 5$. Use this fact to solve the following questions.

(a) Suppose that the final exam for an inferior Discrete Math class with a lazy prof (not at UC!) has 80 multiple-choice questions, where each question has 4 choices. If you guess blindly, you have a $1/4$ chance of guessing right on each question. Calculate the approximate probability that, if you answer every question by guessing blindly, you get 30 or more questions right.

Hint: Approximate the number of right answers as a normal distribution, normalize it to obtain a standard normal distribution, then use a normal table.

(b) Find a value k for which, when you flip a fair coin 10,000 times, the probability of k or more heads is approximately 0.20.

Hint: Approximate the number of heads as a r.v. X with an appropriate normal distribution, normalize it to obtain a standard normal distribution, then use a normal table.

Advice: If using an online normal calculator, we recommend always normalizing the random variable first, so that you have a standard normal distribution. We have noticed that some online normal calculators aren't always perfectly accurate when dealing with something other than the standard normal.

Optional Problems

These problems are for extra practice, and they will not be graded, so don't turn them in! We will provide solutions, however.

10. (0 pts.) How to Lie With Statistics

Here is some on-time arrival data for two airlines, A and B, into the airports of Los Angeles and Chicago. (Predictably, both airlines perform better in LA, which is subject to less flight congestion and less bad weather.)

	Airline A		Airline B	
	# flights	# on time	#flights	# on time
Los Angeles	600	534	200	188
Chicago	250	176	900	685

(a) Which of the two airlines has a better chance of arriving on time into Los Angeles? What about Chicago?

(b) Which of the two airlines has a better chance of arriving on time overall?

(c) Do the results of parts (a) and (b) surprise you? Explain the apparent paradox, and interpret it by writing down precise expressions involving conditional probabilities.

11. (0 pts.) Bayesian Learning

(a) Consider the example in class, where there are n coins with biased probabilities $p_1, p_2, \ldots, p_n$. A coin is randomly chosen and we want to learn about its identity by flipping it multiple times. Suppose you observe the outcomes $Y_1 = b_1, Y_2 = b_2, \ldots$. To keep track of what you have learnt about X, the identity of the unknown coin, you want to compute for every k, the conditional distribution of X given $Y_1 = b_1, Y_2 = b_2, \ldots, Y_k = b_k$. However, you don't want to start from scratch for each k. Give a recursive formula for computing the conditional distribution of X given $Y_1 = b_1, \ldots, Y_{k+1} = b_{k+1}$ in terms of b_{k+1} and the conditional distribution of X given $Y_1 = b_1, \ldots, Y_k = b_k$. You may find the following conditional version of Bayes' rule very useful:

$$\Pr[A|B,C] = \frac{\Pr[A|C]\Pr[B|A,C]}{\Pr[B|C]} \tag{1}$$

(b) Take $n = 3$, $p_1 = 2/3, p_2 = 0.5, p_3 = 0.2$. Simulate one run of the experiment by randomly choosing a coin and flipping it successively. Use your recursive formula in part (c) to compute the conditional distribution of X given the observations up to time k. Plot the conditional distributions for $k = 1$, $k = 5$, $k = 10$ and $k = 15$. In your simulated run, would you say that learning is taking place as you see more observations?

12. (0 pts.) Cumulative Distribution Function

In class, the statistics of a r.v. are specified by the distribution in the discrete case and specified by the probability density function (pdf) in the continuous case. To unify the two cases, we can define the *cumulative distribution function* (cdf) F for a r.v., which is valid for both discrete and continuous r.v.'s:

$$F(a) := \Pr[X \leq a], \qquad a \in \mathbb{R}.$$

(a) In the discrete case, show that the cdf of a r.v. contains exactly the same information as its distribution, by expressing F in terms of the distribution and expressing the distribution in terms of F.

(b) In the continuous case, show that the cdf of a r.v. contains exactly the same information as its pdf, by expressing F in terms of the pdf and expressing the pdf in terms of F.

(c) Compute and plot the cdf for (i) $X \sim \text{Geom}(p)$, (ii) $X \sim \text{Exp}(\lambda)$.

(d) Identify two key properties that a cdf of any r.v. has to satisfy.

13. (0 pts.) A difference between discrete and continuous r.v.'s

Discrete and continuous r.v.'s have a lot of similarities but some differences too.

(a) Suppose X is a discrete r.v. Let the r.v. $Y = cX$ for some constant c. Express the distribution of Y in terms of that of X.

(b) Suppose X is a continuous r.v. Let the r.v. $Y = cX$ for some constant c. Express the pdf of Y in terms of that of X. Is there any difference with the discrete case? [Hint: work with cdf's introduced in Q.12.]

(c) If $X \sim N(\mu, \sigma^2)$, what is the density of $Y = cX$?