University of Illinois **Spring 2012**

ECE 313: Problem Set 5
Bayes' Formula and binary hypothesis testing

Due:	Wednesday, February 22 at 4 p.m.
Reading:	*ECE 313 Notes* Sections 2.10 & 2.11
Reminder:	Exam I on Monday, February 27, 7:00 p.m. – 8:15 p.m. Location: Section E (meets 9 am) 163 Everitt Lab Section C (meets 10 am) 269 Everitt Lab Section D (meets 11am) 103 Talbot Lab Section F (meets 1pm) 165 Everitt Lab One two-sided 8.5"×11" sheet of notes allowed, with font size no smaller than 10 pt or equivalent handwriting. Bring a picture ID. The exam will cover the reading assignments, lectures, and problems associated with problem sets 1-5.

1. **[Explaining a sum]**
   Suppose $S = X_1 + X_2 + X_3 + X_4$ where X_1, X_2, X_3, X_4 are mutually independent and X_i has the Bernoulli distribution with parameter $p_i = \frac{i}{5}$ for $1 \leq i \leq 4$.

   (a) Find $P\{S = 1\}$.

   (b) Find $P(X_1 = 1|S = 1)$.

2. **[The weight of a positive]**
   (Based on G. Gigerenzer, *Calculated Risks*, Simon and Schuster, 2002 and S. Strogatz NYT article, April 25, 2010.) Women aged 40 to 49 years of age have a low incidence of breast cancer; the fraction is estimated at 0.8%. Given a woman with breast cancer has a mammogram, the probability of detection (i.e. a positive mammogram) is estimated to be 90%. Given a woman does not have breast cancer, the probability of a false positive (i.e a false alarm) is estimated to be 7%.

   (a) Based on the above numbers, given a woman aged 40 to 49 has a mammogram, what is the probability the mammogram will be positive?

   (b) Given a woman aged 40 to 49 has a positive mammogram, what is the conditional probability the woman has breast cancer?

   (c) For 1000 women aged 40 to 49 getting mammograms for the first time, how many are expected to have breast cancer, for how many of those is the mammogram positive, and how many are expected to get a false positive?

3. **[A simple hypothesis testing problem]**
   Suppose there are two hypotheses about an observation X, with possible values in $\{-4, -3, \ldots, 3, 4\}$:
   H_0 : X has pmf $p_0(i) = \frac{1}{9}$ for $-4 \leq i \leq 4$ H_1 : X has pmf $p_1(i) = \frac{i^2}{60}$ for $-4 \leq i \leq 4$.

   (a) Describe the ML rule. Express your answer directly in terms of X in a simple way.

   (b) Find $p_{false\ alarm}$ and p_{miss} for the ML rule.

   (c) Find the MAP rule for priori distribution $\pi_0 = \frac{2}{3}$ and $\pi_1 = \frac{1}{3}$.

   (d) Find p_e for the MAP rule found in part (c), assuming the prior used in part (c) is true.

   (e) For what values of $\frac{\pi_0}{\pi_1}$ does the MAP rule always decide H_0? Assume ties are broken in favor of H_1.

4. **[Matching Poisson means]**
Consider hypotheses H_0 and H_1 about a two dimensional observation vector $X = (X_1, X_2)$. Under H_0, X_1 and X_2 are mutually independent, and both have the Poission distribution with mean 4. Under H_1, X_1 and X_2 are mutually independent, X_1 has the Poission distribution with mean 2, and X_2 has the Poisson distribution with mean 6.

   (a) Describe the ML rule for H_0 vs. H_1. Display your answer by indicating how to partition the set of possible observations, $\{(i,j) : i \geq 0, j \geq 0\}$, into two sets, Γ_0 and Γ_1, for which the decision is H_0 if $(X_1, X_2) \in \Gamma_0$ and H_1 if $(X_1, X_2) \in \Gamma_1$.

   (b) Describe the MAP rule for H_0 vs. H_1, assuming the prior distribution with $\frac{\pi_0}{\pi_1} = 2$. Display your answer by indicating how to partition the set of possible observations, $\{(i,j) : i \geq 0, j \geq 0\}$, into two sets, Γ_0 and Γ_1, for which the decision is H_0 if $(X_1, X_2) \in \Gamma_0$ and H_1 if $(X_1, X_2) \in \Gamma_1$.

5. **[Field goal percentages – home vs. away]**
The Illini woman's basketball team plays some games at home and some games away. During games, the players make field goal attempts (i.e. throw the ball towards the hoop), and some of the attempts result in actual field goals (i.e. the ball goes through the hoop). Let p_h be the probability a field goal attempt at a home game is successful and p_a denote the probability a field goal attempt at an away game is successful. We assume (perhaps this is quite inaccurate) that attempts are successful independently of each other. We'd like to test the hypothesis $H_1 : p_h \neq p_a$ vs. hypothesis $H_0 : p_h = p_a$, based on actual data.[1] Specifically, the following statistics were collected from the team's website, for the games played in November and December 2011. There were five home games and nine away games.

	field goals	field goal attempts	shooting percentage
home games	119	281	42.35%
away games	212	521	40.70%

We take the two numbers of attempts, 281 and 521, as given, and not part of the random experiment.

   (a) Using the methodology of Section 2.9, use the given data to calculate 95% confidence intervals for p_h and for p_a. (Note: If the two intervals you find intersect each other, then the accepted scientific methodology would be to say that there is not significant evidence in the data to reject the null hypothesis, H_0. In other words, the experiment is inconclusive.)

   (b) Suppose an analysis of data similar to this one were conducted, and the two intervals calculated in part (a) did not intersect. What statement could we make in support of hypothesis H_1? (Hint: Refer to equation (2.11) of the notes, for which the true value p is fixed and arbitrary, and $\widehat{p} = \frac{X}{n}$ is random.)

[1] This hypothesis testing problem falls outside the scope of Section 2.11, because the hypotheses are *composite* hypotheses, meaning that they involve one or more unknown parameters. For example, H_0 specifies only that $p_h = p_a$, without specifying the numerical value of the probabilities. Problems like this are often faced in scientific experiments. A common methodology is based on the notion of p-value, and on certain functions of the data that are not sensitive to the parameters, such as in T tests or$F$ tests. While the details are beyond the scope of this course, this problem aims to give some insight into this common problem in scientific data analysis.