BST 140.651
Problem Set 1
Due in class on September 15

Reading: Rice chapters 1, 2, 4.

Problem 1. Let A and B be two sets. Argue the following.

a. If $A \subset B$ then $A \cup B = B$.

b. If $A \subset B$ then $A \cap B = A$.

c. $(A \cup B)^c = A^c \cap B^c$ (where a superscript c denotes the converse).

d. $(A \cap B)^c = A^c \cup B^c$.

Problem 2. Let A and B be events of the sample space Ω with an associated probability P. Using the 3 axioms that define the probability demonstrate the following.

a. $P(A^c) = 1 - P(A)$.

b. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

c. If $A \subset B$ then $P(A) \leq P(B)$.

Problem 3. Suppose that an influenza epidemic strikes a city. In 10% of the families the mother has contracted influenza while in 2% of the cities both the mother and father have contracted influenza.

a. Do the events $A = \{$mother contracts influenza$\}$ and $B = \{$father contracts influenza$\}$ appear to be independent?

b. What is the conditional probability that the father has influenza given that the mother has influenza?

c. What is the conditional probability that the father has influenza given that the mother does *not* have influenza?

Problem 4. Assume that an act of intercourse between an HIV infected person and a non-infected person results in a $1/500$ probability of spreading the infection.

a. How many acts of intercourse would an uninfected person have to have with an infected persons to have a 50% probability of obtaining an infection? State the assumptions of your calculations.

b. A related problem. How many people have to be in a room to have a 50% probability of at least two people having the same birthday? State the assumptions of your calculations.

c. Another related problem. I have 200 songs on my mp3 player. The documentation claims that its shuffle option samples randomly with replacement from the list of songs. How many songs do I have to listen to to have a 50% probability of hearing at least one song twice?

Problem 5. The Chinese Mini-Mental Status Test (CMMS) is a test consisting of 114 items intended to identify people with Alzheimer's disease (AD) and dementia among people in China. An extensive clinical evaluation was performed of this instrument, whereby participants were interviewed by psychiatrists and nurses and a definitive (clinical) diagnosis of AD was made. The table below show the counts obtained on the subgroup of people with at least some formal education. Suppose a cutoff value of ≤ 20 on the test is used to identify people with AD.

	Clinical diagnosis of AD	
CMMS score	No	Yes
0-5	0	2
6-10	0	1
11-15	3	4
16-20	9	5
21-25	16	3
26-30	18	1

(a) What is the sensitivity and specificity of the CMMS test using the 20 cutoff?

(b) Graph the positive predictive value as a function of the prevalence of AD. Do the same for the negative predictive value.

(c) What would be the sensitivity and specificity if a cutoff of 15 had been used?

Problem 6. A study was performed where a large number of patients were diagnosed in one of 7 possible disease states based on a detailed medical exam and history. The presence or absence of a specific symptom (X) was also recorded from this study and from estimates based on other published data. Two sets of probabilities were generated.

(1) The unconditional probability of each of the disease states (referred to as the prevalence in the column below).

(2) The conditional probability of the specific symptom (X) given each of the disease states (the final column of the table).

The proportion of persons with disease Y_4 is .020, for example. Similarly, the probability that a person has symptom X given that he or she has disease Y_4 is .95. Assume that these represent an exhaustive list of possible diagnoses and that a patient can have one and only one diagnosis. Disease Y_1 represents the normal (non-diseased) state.

Disease Diagnosis	Prevalence of Y_i	Probability of Symptom X given Y_i	Probability of Y_i given X
Y_1	.155	.01	?
Y_2	.126	.70	?
Y_3	.084	.95	?
Y_4	.020	.95	?
Y_5	.098	.40	?
Y_6	.391	.30	?
Y_7	.126	.70	?

a. What is the probability of each diagnosis given that you have the symptom? (Fill in the missing column above). What is the most likely diagnosis given that you have symptom X?

b. Suppose that symptom X is used as a screening criterion for diagnosing disease Y_4. What is the sensitivity of this test? What is the specificity?

Problem 7. Given below are the sexes of the children of 7,745 families of 4 children recorded in the archives of the Genealogical Society of the Church of Jesus Christ of Latter Day Saints in Salt Lake City, Utah. M indicates a male child and F indicates a female child.

Sequence	Freq	Sequence	Freq
MMMM	537	MFFM	526
MMMF	549	FMFM	498
MMFM	514	FFMM	490
MFMM	523	MFFF	429
FMMM	467	FMFF	451
MMFF	497	FFMF	456
MFMF	486	FFFM	441
FMMF	473	FFFF	408

a. Estimate the probability distribution of the number of male children, say X, in these families using the data below by calculating proportions.

b. Find the expected value of X.

c. Find the variance of X.

d. Find the probability distribution of $\hat{p}$, where $\hat{p}$ is the proportion of children in each family who are male. Find the expected value of $\hat{p}$ and the variance of $\hat{p}$

Problem 8. A number of clinical characteristics were ascertained in a large group of subjects with insulin-dependent diabetes mellitus (IDDM). Suppose that the distribution of the percentage of ideal body weight in this group of patients is normally distributed with a mean 110% and standard deviation 13%. Do the questions below using both a normal table and using R (or equivalent).

(a) What proportion of subjects with IDDM are above their ideal body weight, (i.e. above 100% of the ideal body weight)?

(b) What proportion of subjects with IDDM have a normal body weight (with 10% of their ideal body weight)?

Problem 9. The exponential distribution has density $f(x) = c\exp\{-\beta x\}$ for $x \geq 0$ and some constant c.

(a) Calculate what the constant, c, must be in relation to β.

(b) Find the associated distribution function for the exponential density.

(c) Find the expected value of an exponentially distributed random variable.

(d) Find the variance of an exponential distributed random variable. What is the relationship between the mean and the variance?

(e) Analytically find the probability that an exponential random variable is between a and b where $0 \leq a \leq b$.