Unit 2 – Introduction to Probability
Practice Problems

Due Date: Monday October 5, 2009

1. ***Before you begin***: *This exercise gives you practice with some of the basics of probability calculations. See lecture notes pp 9-12.*

Let A and B denote two independent genetic traits. Suppose the probability that an individual will exhibit trait A is ½ and the probability that an individual will exhibit trait B is ¾. What is the probability that an individual will exhibit

(a) Both traits?
(b) Neither trait?
(c) trait A but not trait B?
(d) trait B but not trait A?
(e) exactly one trait?

2. ***Before you begin***: *This exercise gives you practice with the multiplication rule. For this, see lecture notes pp 25-26. Notice that this is about the general multiplication rule and not the special case where the events are independent.*

Suppose you are told that pr (right eye is blue) = 1/3 and pr(left eye is blue) = 1/3. Using the concepts and formulae in the lecture notes for Unit 2 (Introduction to Probability), confirm for yourself what you know by intuition, namely that pr(person is blue eyed = 1/3 by solving for pr(blue right eye and blue left eye).

3. ***Before you begin***: *This exercise gives you practice with Bayes Rule. See lecture notes pp 29-30.*

A physician develops a diagnostic test that is positive for 95% of the patients who have disease and is positive for 10% of the patients who do not have disease. Of patients tested, 20% actually have disease. Suppose you evaluate a patient by administering this diagnostic test and obtain a positive result. Using the information given, calculate the probability that this patient has disease.

4. ***Before you begin***: *Parts "a" – "c" invite you to pause in your thinking. Parts "d" – "f" are straightforward application of the ideas of relative risk and odds ratio. See lecture notes pp 39-47.*

In introductory epidemiology, one of the study designs that are introduced is the **prospective cohort study**. In this type of study involving two groups, the investigator enrolls pre-set and known numbers of participants into each of the two groups that are generically described as "exposed" and "not exposed" and follows them forward to a designated end of the observation period, at which point some outcome is measured.

Consider the following prospective cohort study. A total of 1500 ***never smoker*** consenting heart attack survivors aged 60-65 are enrolled as "non-exposed". An equal number, 1500 ***current smoker*** heart attack survivors aged 60-65 are enrolled as "exposed". All are followed for a full 10 years and the occurrence of death recorded. Following are the data.

		Vital Status at 10 Years		
		Dead	Alive	
Exposure	Current Smoker	**40**	**1460**	**1500**
Status	Never Smoker	**10**	**1490**	**1500**
		50	2950	3000

(a) Is it possible to estimate the probability of 10 year survival on the basis of these data?

(b) Is it possible to estimate the relative risk of 10 year mortality that is associated with current cigarette use?

(c) Is it possible to estimate the probability that a randomly selected person with a vital status of "Alive" at 10 years is a current smoker?

(d) Using these data, estimate the relative risk of 10 year mortality that is associated with current cigarette use.

(e) Using these data, estimate the relative odds of 10 year mortality that is associated with current cigarette use. *(Note – This question is asking you to compute an **odds ratio**).*

(f) Using these data, estimate the relative odds of a ***current smoker*** notation for non-survivors relative to survivors.

(g) What do you notice about your answers to "e" and "f"?

(h) How do your answers to "e" and "f" compare to your answer to "d"?.

5. ***<u>Before you begin</u>***: *This question is really an elaboration of the thinking that was developed in question 1.*

Another study design that is introduced in introductory epidemiology is the **retrospective case-control study**. This is, by definition, a study that compares two groups. Here, the investigator enrolls pre-set and known numbers of participants into each of the two groups defined by disease status; "cases" are the enrollees with disease, "controls" do not have the disease under investigation. Retrospective review of the histories of all study participants is performed to identify the subsets in each of the case and control groups who have a history of the exposure of interest.

Consider the following retrospective case-control study of the association between coffee consumption and tumors of the lower urinary tract. The investigator enrolls 30 consenting cases that are patients with one or more tumors of the lower urinary tract. For comparison purposes, he/she also enrolls 100 consenting controls who have no such tumors. Following are the data.

		Tumors of Lower Urinary Tract		
		Yes	No	
History of	5+ cups/day	**20**	**44**	64
Coffee consumption	<1 cup/day	**10**	**56**	66
		30	**100**	130

(a) Is it possible to estimate the probability of one or more tumors of the lower urinary tract on the basis of these data?

(b) Is it possible to estimate the relative risk of one or more tumors of the lower urinary tract that is associated with consumption of 5 or more cups of coffee/day?

(c) Using these data, estimate the relative odds of **high coffee consumption (5+ cups/day)** among cases, relative to controls.

(d) Using these data, estimate the relative odds of **tumors of the lower urinary tract** among high coffee consumers (5+ cups/day), relative to non-coffee drinkers.

(e) What do you notice about your answers to "c" and "d"?

6. ***Before you begin***: *This last question is a bit of a weaving of a story. If you follow along, step by step, you will end up seeing for yourself a truly marvelous result. It is the result of page 46 of your lecture notes!*

Now consider a fully **cross-sectional study design**, this time with generic counts "a", "b", "c", and "d". In this design, the investigator does not do any formal enrollment. Counts are accumulated by observation. CDC surveillance programs are examples

		Disease	
		Yes	No
History of	Yes	**a**	**b**
Exposure	No	**c**	**d**

(a) Using the letters "a", "b", "c", and "d", what is the formula for estimating relative odds of the **event of exposure** for persons with disease, compared to that for persons without disease?

(b) Using the letters "a", "b", "c", and "d", what is the formula for estimating relative odds of the **event of disease** for exposed persons, compared to that for non-exposed persons?

(c) Using the letters "a", "b", "c", and "d", what is the formula for estimating relative risk of the **event of disease** for exposed persons, compared to that for non-exposed persons?

(d) What happens to your formula in your answer to #3b when the counts of disease (a and c) are very very small? Comment.