Unit 2 – Introduction to Probability
Week #4 - Practice Problems

SOLUTIONS
corrected 0-14-2008

1. 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?
Answer: Yes, but the question is a poor one as it does not specify time zero nor among whom. 2950/3000=0.9833

(b) Is it possible to estimate the relative risk of 10 year mortality that is associated with current cigarette use?
Answer: Yes, but this question too is poor. Without a meaningful time zero, the interpretation of the answer is non-existent. (40/1500)/(10/1500)=4

(c) Is it possible to estimate the probability that a randomly selected person with vital status of "Alive" at 10 years is a current smoker?
Answer: Yes. Even though the study design called for fixed numbers of enrollments of current smokers and never smokers, it is possible to estimate the conditional probability of "current smoker" for a randomly selected person from among the vital status at 10 years of "Alive".

(d) Using these data, estimate the relative risk of 10 year mortality that is associated with current cigarette use.
Answer: (40/1500)/(10/1500)=4

(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**).*
Answer: The event of interest is 10 year mortality. For this event,
OR = (Odds among current smokers)/(Odds among never smokers)
= (40/1460)/(10/1490)=4.082

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

    Answer: Here, the event of interest is current smoker. For this event,
              OR = (Odds among non-survivors)/(Odds among survivors)
                  = (40/10)/(1460/1490)=4.082

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

    Answer: They are the *same!*
              This confirms what is noted in the lecture notes, that the OR for
              the event of disease in a cohort study comparison of exposure groups
              is equal to the OR for the event of history of exposure in a case-control
              comparison of disease/non disease groups.

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

    Answer: The OR is bigger than the RR; for rare diseases, the discrepancy
              becomes negligible..

2.   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?

    Answer: No, because the case-control study design calls for fixed numbers of
              enrollment into the "tumors=yes" and "tumors=no" groups.

(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?

    Answer: No, for a reason that relates to the answer to "a". It is not possible
              to estimate the component probabilities.

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

     Answer: For the event high coffee consumption,

         OR = (Odds among "tumor=yes")/(Odds among "tumors=no")

             = (20/10)/(44/56)=2.545

(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.

     Answer: For the event "tumors=yes",

         OR = (Odds among "high coffee")/(Odds among "non-coffee")

             = (20/44)/(10/56)=2.545

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

     Answer: They are the same.

3. 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?

     Answer: [a/c] / [b/d] = (ad)/(bc).

(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?

     Answer: [a/b] [c/d] = (ad)/(bc).

(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?

     Answer: [a/(a+b)] [c/(c+d)].

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

     Answer: As "a" gets smaller and smaller, (a+b) → b

         As "c" gets smaller and smaller, (c+d) → d. As a result,

         RR =[a/(a+b)] [c/(c+d)] → [a/(b)] [c/(d)] = (ad)/(bc) = OR.

> **So … it is sometimes possible to estimate RR from a case-control study because**
>
> **OR $_{event=disease}$ = OR $_{event=exposure}$**
>
> **When the disease is rare, OR $_{event=disease}$ ≈ RR $_{event=disease}$**
>
> **When the disease is rare, OR $_{event=exposure}$ ≈ RR $_{event=disease}$**