ST520, Fall 2019

Homework 1, due: Tuesday, 9/10/2019

1. (10 pts) In a case-control study conducted in France a while ago to investigate the association between esophageal cancer and alcohol assumption, the following data was collected from 200 cases and 775 controls (so the total sample size is 975)

Alcohol assumption	Esophageal Cancer	No Esophageal Cancer
Heavy	96	109
Light	104	666

where "Heavy" is defined as "Average daily alcohol assumption $\geq$ 80 g", "Light" is defined as "Average daily alcohol assumption $<$ 80 g".

Do the following:

(a) Can you estimate the proportion of heavy drinkers in French at the time when the study was conducted. State your reason. If you can, please provide your estimate.

(b) Can you estimate the prevalence of esophageal cancer in French at the time when the study was conducted. State your reason. If you can, please provide your estimate.

(c) Can you estimate the prevalence of esophageal cancer among heavy drinkers in French at the time when the study was conducted. State your reason. If you can, please provide your estimate.

(d) Can you estimate the prevalence of esophageal cancer among light drinkers in French at the time when the study was conducted. State your reason. If you can, please provide your estimate.

(e) Can you estimate the relative risk of having the esophageal cancer between heavy drinkers and light drinkers in French at the time when the study was conducted. State your reason. If you can, please provide your estimate and interpret your result.

(f) Can you estimate the odds-ratio of having the esophageal cancer between heavy drinkers and light drinkers in French at the time when the study was conducted. State your reason. If you can, please provide your estimate and interpret your result. Can your odds-ratio estimate help you make inference on the relative risk in (e)?

(g) Find a 95% CI for the true odds-ratio of having the esophageal cancer between heavy drinkers and light drinkers. What is the conclusion based on this CI?

2. (20 pts) An investigator wants to investigate the association between heart attack and coffee drinking. Since heart attack is a rare event, she would like to conduct a case-control study to improve the efficiency of the estimate of the odds-ratio of having heart attack between coffee drinkers and non coffee drinkers. The cost to sample a control is \$1 and the cost to sample a case is \$4. The investigator has \$400 for this study.

   For design purpose, let us assume the following joint probabilities of having an heart attack (or no heart attack) and being a coffee drinker (or non-drinker)

	Heart Attack	No Heart Attack
Coffee Drinker	0.006	0.495
None Coffee Drinker	0.004	0.495

   Do the following:

   (a) Find the true relative risk and odds-ratio of having heart attack between coffee drinkers and non-drinkers using the given information. Are they close to each other?

   (b) Find the conditional probabilities of being a coffee drinker given cases and controls and find the odds-ratio of being a coffee drinker between cases and controls. Is it the same as the odds ratio you got in (a)?

   (c) If the investigator would like to have equal sample size for cases and controls, how many cases and controls can she have in her study given her budget constraint? and how many individuals can she expect for the following 2×2 table?

	Heart Attack	No Heart Attack
Coffee Drinker		
None Coffee Drinker		

   Based on the numbers in your table, find the variance of $\log(\widehat{\theta})$ (**Hint**: You can replace n_{ij}'s in the variance formula for $\log(\widehat{\theta})$ by their expectations).

   (d) Find the optimal design for the investigator. That is, how many cases and how many controls should the investigator sample to yield the most efficient (having the smallest variance) estimate $\log(\widehat{\theta})$ giving the budget constraint. Find the variance of $\log(\widehat{\theta})$ for this optimal design. Is it smaller than the variance you got in (c)? (**Hint**: Denote

the case sample size by x. Then the budget constraint says that the control sample size has to be $400 - 4x$ and $x \in (0, 100)$. Try x as a real number so that you can do maximization more easily. Then find out the variance of $\log(\widehat{\theta})$ for two integers around the optimal value of x. The x_0 with the smaller variance gives you the sample size for cases and $400 - 4x_0$ is the sample size for controls)

3. (10 pts) Suppose the joint distribution of an exposure variable and a disease variable in the target population is given in the following table

	D	$\bar{D}$
E	0.05	0.35
$\bar{E}$	0.02	0.58

   (a) If an investigator conducts a prospective study with equal (but large) sample sizes (n) for the exposure and non-exposure groups, find the variance estimate for the estimate of $\log(\widehat{\theta})$ in terms of n. (**Hint**: You can replace n_{ij}'s in the variance formula for $\log(\widehat{\theta})$ by their expectations).

   (b) If the investigator conducts a case-control study with the same sample size (n) as in (a) for cases and controls, find the variance estimate for the estimate of $\log(\widehat{\theta})$ in terms of n.

   (c) Based on your results, find the ratio of the variances in (a) and (b). Which study is more efficient?