Biostatistical Methods I (BIOS 5710)
Breheny

Assignment 9

Due: Wednesday, November 12

1. Early in this course, we discussed the clinical trial of Nexium. The results of the trial were that 2,430/2,624 individuals who took Nexium were healed from erosive esophagitis, compared with 2,324/2,617 individuals who took Prilosec.

   (a) Is Nexium more effective than Prilosec at treating erosive esophagitis, or could the results of this trial be explained by chance variability? Use an exact test.

   (b) Same as (a), but use an approximate test.

   (c) In the sample, what was the observed increase in the odds of healing for patients on Nexium compared with patients on Prilosec?

   (d) Calculate an approximate confidence interval for the quantity in (c).

   (e) Calculate an exact 95% confidence interval for quantity in (c).

2. Millions of American women underwent breast augmentation/reconstruction surgery when the procedure was pioneered in the early 1960s. In response to case reports of connective tissue and autoimmune diseases following the surgery, the FDA issued a moratorium on these procedures in 1992 (this moratorium is no longer in effect). To investigate whether these anecdotes were statistically significant, researchers at the Mayo clinic conducted a retrospective study and obtained the following results:

	Connective Tissue Disease	
	Yes	No
Augmentation	5	744
No augmentation	10	1488

   (a) Conduct an appropriate hypothesis test of the null hypothesis that breast augmentation/reconstruction surgery has no impact on connective tissue disease. What is your conclusion?

   (b) Is the following statement true or false: "Based on my results in part (a), there is a high probability that the null hypothesis is true."

   (c) Calculate a 95% confidence interval for the odds ratio of developing connective tissue disease for women who received this surgery compared to women who did not.

3. Comparing the data and your analysis of it in problems 1 and 2, for which study is an odds ratio of 2 more plausible? Do the confidence intervals contradict the hypothesis tests? Discuss.

4. In the prospective CDC breast cancer study we discussed in class, about 3/4 of women gave birth to their first child before age 25. Of the early-birth cohort, about 1.5% developed breast cancer, compared with 2% in the late-birth cohort.

   (a) Taking the above percentages as the true parameter values, how large a cohort is required (i.e., total sample size) in order to achieve 80% power to detect a significant difference?

(b) Based on the above percentages, what is the probability that a woman will have given birth at age 25 or older, given that she develops breast cancer?

(c) What is the probability that a woman will have given birth at age 25 or older, given that she does not develop breast cancer?

(d) Based on your calculations in (b) and (c), how large a sample is required to detect a significant association between age at first labor and breast cancer risk if we plan a case-control study with an equal number of cases and controls?

5. In class, we derived an approximate confidence interval for the odds ratio. Consider now the problem of obtaining a confidence interval for the relative risk.

   (a) Show that for $X \sim \text{Binom}(n, \pi)$

   $$\log(\hat{\pi}) \dot{\sim} \text{N}\left(\log(\pi), \frac{b}{a(a+b)}\right),$$

   where $\hat{\pi} = x/n$, a is the expected number of successes, and b is the expected number of failures.

   (b) Using your result from (a), show that for two independent binomial samples $X \sim \text{Binom}(n_1, \pi_1)$ and $Y \sim \text{Binom}(n_2, \pi_2)$,

   $$\log(\widehat{\text{RR}}) \dot{\sim} \text{N}\left(\log(\text{RR}), \frac{b}{a(a+b)} + \frac{d}{c(c+d)}\right),$$

   where a, b, c, and d are expected cell counts as in class.

   (c) Using your result from (b), calculate a confidence interval for the relative risk of death in the Lister study, comparing control patients to patients who received sterile surgery.

6. Examine the table on page 193 of our textbook.

   (a) In this situation, why might a researcher care more about the difference in proportions than the odds ratio?

   (b) In the same situation, why might a researcher care more about the odds ratio (or relative risk) than the difference in proportions?

7. This problem concerns the relationship between the relative risk and the odds ratio.

   (a) Plot the odds ratio versus π_1 while keeping the relative risk fixed at $\pi_2/\pi_1 = 2$. Obviously, your plot cannot extend past $\pi_1 = 0.5$, or π_2 will not be defined.

   (b) Which is larger, the odds ratio or the relative risk?

   (c) Describe what happens to the relationship between RR and OR as π_1 goes to 0.

   (d) Describe what happens to the relationship between RR and OR as π_1 approaches 0.5.

8. Consider a disease D and exposure E. Let $\pi_1 = P(D|E)$, $\pi_2 = P(D|E^C)$, and $\pi = P(D)$.

   (a) Show that

   $$\text{odds}(E|D) = \text{odds}(E)\frac{\pi_1}{\pi_2}$$

(b) Using your result from part (a), show that the retrospective odds ratio ($E|D$ vs. $E|D^C$) is

$$\text{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)};$$

i.e., that the retrospective odds ratio is equal to the prospective odds ratio.

9. A common use of the Poisson distribution in epidemiological studies is to account for different duration of followup among subjects. In a classic study led by the epidemiologist Richard Doll, two large cohorts of male British doctors – one group who smoked, the other who did not – were followed for a number of years to see whether or not they had died from coronary heart disease.

   In principle, this data could be modeled using a binomial distribution in which each man has a certain probability of developing the disease. However, if one man is followed for 5 years and another for 25 years, it is unrealistic to make the binomial assumption that the two men have the same probability π of developing the disease while on study.

   An alternative is to consider deaths from coronary heart disease as Poisson counts in which the size of the set is the person-years of follow-up. Specifically, let X denote the number of deaths in the smoking cohort and Y denote the number of deaths in the non-smoking cohort. We may assume that $X \sim \text{Pois}(t_x\lambda)$ and $Y \sim \text{Pois}(t_y\mu)$, where t_x and t_y are the person-years of follow-up and λ and μ are the respective rates of disease.

   In the study, the smoking cohort was followed for 142,247 person-years, while the non-smoking cohort was followed for 39,220 person-years. In that time, 630 coronary heart disease deaths were observed in the smoking cohort, compared with 101 deaths in the non-smoking cohort.

   (a) Test the hypothesis that the rate of death from coronary heart disease in the same in the two groups.
   (b) What is the observed rate ratio comparing incidence of death from coronary heart disease in smokers compared with nonsmokers.
   (c) Calculate a 95% confidence interval for the quantity in part (b).
   (d) Calculate a point estimate and 95% confidence interval for the incidence of death from coronary heart disease in each group. Express your answers as a rate per 1,000 person-years.