Homework 4, Due Friday September 30, 2016

You are welcome to use R but attach the code.

1. For 2×2 tables show that $\widehat{RR} = 1$ when computed using the estimated expected counts. $\widehat{E}_{ij}$. That is, show that the expected counts satisfy the null hypothesis of independence. You can assume population based sampling with the total sample size fixed.

2. We will examine the skewness of the estimator of relative risk and how the log transformation helps via simulation For the simulation $RR = 0.4/0.2 = 2$. We will simulate 5000 values of $\widehat{p}_1$ and $\widehat{p}_2$ based on samples of size 100.

   (a) The R-code below will do this for us and plot the results. Discuss.

   ```
   p1.hat<-rbinom(5000,100,0.4)/100
   p2.hat<-rbinom(5000,100,0.2)/100
   RR.hat<-p1.hat/p2.hat
   log.RR.hat<-log(RR.hat)
   par(mfrow=c(2,1))
   hist(RR.hat,prob=T,main="Relative Risk")
   hist(log.RR.hat,prob=T,main="Log Relative Risk")
   ```

   (b) Compute the variance of the 5000 simulated values of $\log(\widehat{RR})$ you got in part (a) and compare to the Delta Method approximate variance formula. Discuss.

3. This example involves hypothetical data collected in a case-control design.

   (a) A 2×2 table giving the relationship between exposure and disease is shown below.

	D	$\overline{D}$	
E	30	12	42
$\overline{E}$	70	88	158
	100	100	200

   Estimate the ratio of the odds of disease given exposure to the odds given no exposure (give me both the point estimate and an approximate 95% CI).

   (b) Below are two tables: one showing the distribution of cases and controls by age and one showing the distribution of exposed and unexposed by age. Investigate the association between age and exposure status and age and disease status. Again give point estimates and approximate 95% CIs for odds ratios.

	D	$\overline{D}$	
< 40	50	80	130
≥ 40	50	20	70
	100	100	200

	E	$\overline{E}$	
< 40	12	118	130
≥ 40	30	40	70
	42	158	200

(c) Stratifying on age yields the two table below. Analyze these separately. Report the estimated odds ratios and associated 95% CIs.

Age < 40

	D	$\overline{D}$	
E	5	6	11
$\overline{E}$	45	74	119
	51	80	130

Age ≥ 40

	D	$\overline{D}$	
E	25	6	31
$\overline{E}$	25	14	39
	50	20	70

(d) Is there evidence of confounding? Why or why not?

4. Problem 8.3 on page 119.

5. STAT 525 Students: Using the Delta Method confirm the variance formula for $\log\left[\widehat{OR}\right]$.

6. STAT 525 Students: We saw one version of the formula for X^2 as

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - \widehat{E}_{ij}\right)^2}{\widehat{E}_{ij}}$$

This can be reexpressed as

$$X^2 = n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(\frac{n_{ij}}{n} - \frac{n_D n_E}{n^2}\right)^2}{\frac{n_D n_E}{n^2}}\right]$$

Discuss the implications of this as n increases while all n_{ij}/n remain constant.