Stat 525 Homework 3

Kenny Flagg

September 16, 2016

1. *A 20-year cohort study of British male physicians noted that the proportion per year who died from lung cancer was 0.00140 for cigarette smokers and 0.00010 for nonsmokers. The proportion who died from coronary heart disease was 0.00669 for smokers and 0.00413 for nonsmokers.*

   (a) *Describe the association of smoking with each of lung cancer and heart disease, using the ER, NNH, RR, and OR. Interpret each of these.*

   Excess Risk:

$$\begin{aligned} ER_{\text{lung cancer}} &= P(\text{lung cancer}|\text{smoker}) - P(\text{lung cancer}|\text{nonsmoker}) \\ &= 0.00140 - 0.00010 \\ &= 0.00130 \end{aligned}$$

   We expect the lung cancer mortality percentage to be 0.130% higher among smokers than nonsmokers assuming the populations are otherwise similar.

$$\begin{aligned} ER_{\text{CHD}} &= P(\text{CHD}|\text{smoker}) - P(\text{CHD}|\text{nonsmoker}) \\ &= 0.00669 - 0.00413 \\ &= 0.00256 \end{aligned}$$

   We expect the coronary heart disease mortality percentage to be 0.256% higher among smokers than nonsmokers assuming the populations are otherwise similar.

   Number Needed to Harm:

$$NNH_{\text{lung cancer}} = \frac{1}{ER_{\text{lung cancer}}} = \frac{1}{0.00130} = 769.231$$

   For every additional 769.231 smokers in the population, we expect one to die of lung cancer.

$$NNH_{\text{CHD}} = \frac{1}{ER_{\text{CHD}}} = \frac{1}{0.00256} = 390.625$$

   For every additional 390.625 smokers in the population, we expect one to die of coronary heart disease.

Relative Risk:

$$\begin{aligned} RR_{\text{lung cancer}} &= \frac{P(\text{lung cancer}|\text{smoker})}{P(\text{lung cancer}|\text{nonsmoker})} \\ &= \frac{0.00140}{0.00010} \\ &= 14.0 \end{aligned}$$

The risk of death from lung cancer for smokers is 14 times as high as the risk of death from lung cancer for nonsmokers.

$$\begin{aligned} RR_{\text{CHD}} &= \frac{P(\text{CHD}|\text{smoker})}{P(\text{CHD}|\text{nonsmoker})} \\ &= \frac{0.00669}{0.00413} \\ &= 1.620 \end{aligned}$$

The risk of death from coronary heart disease for smokers is 1.62 times as high as the risk of death from coronary heart disease for nonsmokers.

Odds ratio:

$$\begin{aligned} OR_{\text{lung cancer}} &= \frac{P(\text{lung cancer}|\text{smoker})/P(\text{no lung cancer}|\text{smoker})}{P(\text{lung cancer}|\text{nonsmoker})/P(\text{no lung cancer}|\text{nonsmoker})} \\ &= \frac{0.00140/(1-0.00140)}{0.00010/(1-0.00010)} \\ &= 14.02 \end{aligned}$$

The odds of dying from lung cancer for smokers are 14.02 times as high as the odds of dying from lung cancer for nonsmokers.

$$\begin{aligned} OR_{\text{CHD}} &= \frac{P(\text{CHD}|\text{smoker})/P(\text{no CHD}|\text{smoker})}{P(\text{CHD}|\text{nonsmoker})/P(\text{no CHD}|\text{nonsmoker})} \\ &= \frac{0.00669/(1-0.00669)}{0.00413/(1-0.00413)} \\ &= 1.624 \end{aligned}$$

The odds of dying from coronary heart disease for smokers are 1.624 times as high as the odds of dying from coronary heart disease for nonsmokers.

(b) *Which response (lung cancer or coronary heart disease deaths) is more strongly related to cigarette smoking in terms of the reduction in number of deaths that would occur with elimination of cigarettes? Explain. (Do not try to get AR for this one. Just answer it using the information in (a).)*

Cigarette smoking is more strongly related to CHD than to lung cancer because CHD has a larger excess risk. For every 10,000 people who give up smoking, we expect 25.6 fewer CHD mortalities, but only 13.0 fewer lung cancer mortalities.

(c) *Why are RR and OR so close to one another in both responses?*

RR and OR are similar because lung cancer and CHD are very rare in both populations.

2. *Problem 4.5 on page 42.*

Incidence is computed by counting only those individuals who were diagnosed with high blood pressure during the follow-up period or who were never diagnosed with high blood pressure, ignoring those who were diagnosed with high blood pressure before the follow-up period began. Prevalence is computed using the entire population, including those who had high blood pressure at the start of the follow-up period. If those with high stress really do have a larger risk of high blood pressure, then it is possible that a large portion of the high stress group already had high blood pressure at the beginning of the study and were counted in the prevalence calculation but not in the incidence calculation.

3. *A study looked at the relationship between the presence of squamous cell carcinoma (yes, no) and smoking behavior. There were 3 categories of smoking behavior: nonsmoker, light smoker ($\leq$ 20 cigarettes per day) and heavy smoker ($>$ 20 cigarettes per day). The estimated OR for females between the presence of carcinoma and smoking behavior (nonsmoker, light smoker) was 11.7. The corresponding OR for females between the presence of carcinoma and smoking behavior (nonsmoker, heavy smoker) was 26.1. Find the estimated OR between carcinoma and smoking behavior (light smoker, heavy smoker). You must justify any calculations.*

First, let's define notation for the exposure events: let N denote a nonsmoker, L denote a light smoker, and H denote a heavy smoker. C denotes carcinoma. Then the OR for carcinoma between light smokers and heavy smokers is

$$
\begin{aligned}
OR_{LH} &= \frac{P\left(C|H\right)/P\left(\overline{C}|H\right)}{P\left(C|L\right)/P\left(\overline{C}|L\right)} \\
&= \frac{\left(\frac{P(C|H)/P\left(\overline{C}|H\right)}{P(C|N)/P\left(\overline{C}|N\right)}\right)}{\left(\frac{P(C|L)/P\left(\overline{C}|L\right)}{P(C|N)/P\left(\overline{C}|N\right)}\right)} \\
&= \frac{OR_{NH}}{OR_{NL}} \\
&= \frac{26.1}{11.7} \\
&= 2.23
\end{aligned}
$$

4. *A study exploring the relationship between exposure to a risk factor and a disease used a simple random sample of 100,000 exposed individuals and a simple random sample of 200,000 unexposed individuals. Disease status was subsequently determined for the individuals in the samples. The results are shown below.*

	D	$\overline{D}$	
E	250	99,750	100,000
$\overline{E}$	125	199,875	200,000
	375	299,625	300,000

They estimate $P(D) = 375/300{,}000 = 0.00125$ and $P\left(D|\overline{E}\right) = 125/200{,}000$ and then estimate attributable risk as

$$AR = \frac{0.00125 - 0.000625}{0.00125} = 0.5 \tag{1}$$

which they interpret by writing "The disease could be reduced by 50% if we could eliminate the risk factor."

Give two major flaws with this interpretation for this particular problem. There is one you MUST give to get credit.

First, they used a cohort design (sampling conditional on exposure), so the distribution of exposure status in their sample is not representative of the population, and therefore it is not appropriate to use the marginal proportion to estimate $P(D)$ and compute the AR.

Second, their interpretation assumes that there are no differences between the exposed and unexposed populations except for exposure status for this one risk factor. This risk factor E could be associated with another factor F not considered in this study. If there is an association between D and F, eliminating E would not have the result the authors expect.

5. *All of the problems at the end of Chapter 5 BUT you only need to identify the type of study: Population-based, Cohort, or Case-Control.*

Question 5.1 — **Population-based**

Question 5.2 — **Case-Control**

Question 5.3 — **Population-based**

Question 5.4 — **Case-Control**

Question 5.5 — **Cohort**

6. *The data are given in Problem 6.5 on page 72.*

   (a) *Carry out a test of independence. Show the tables of expected counts and the standardized Pearson residuals. What do the results indicate about the association between heart disease and hair loss? Dont just say there is a lack of independence – I want some idea of where the independence is breaking down.*

```
table.6.8 <- rbind(Balding = c(`Heart Disease` = 127, `No Heart Disease` = 1224),
                   `No Hair Loss` = c(`Heart Disease` = 548, `No Heart Disease` = 7611))
test <- chisq.test(table.6.8, correct = TRUE)
print(test)
```

```

Pearson's Chi-squared test with Yates' continuity correction

data:  table.6.8
X-squared = 12.258, df = 1, p-value = 0.0004632
```

```
xtable(test$expected, caption = 'Expected counts under independence.')
```

	Heart Disease	No Heart Disease
Balding	95.89	1255.11
No Hair Loss	579.11	7579.89

Table 1: Expected counts under independence.

```
xtable(test$residuals, caption = 'Standardized residuals.')
```

	Heart Disease	No Heart Disease
Balding	3.18	-0.88
No Hair Loss	-1.29	0.36

Table 2: Standardized residuals.

The continuity-corrected Pearson test indicates very strong evidence of an association between hair loss and coronary heart disease ($\chi^2_1 = 12.258$, p-value $= 0.0004632$). Examining the standardized residuals, we see a very large residual of 3.18 for the subjects with hair loss and heart disease. This means that male physicians with hair loss are more likely to get CHD than are male physicians without hair loss.

(b) *Estimate ER, RR, and OR. Construct approximate 95% confidence intervals in each case and interpret them.*

Large-Sample Wald Intervals:

```
twoby2(table.6.8)

2 by 2 table analysis:
------------------------------------------------------
Outcome   : Heart Disease
Comparing : Balding vs. No Hair Loss

             Heart Disease No Heart Disease    P(Heart Disease) 95% conf.
Balding                127             1224              0.0940    0.0796
No Hair Loss           548             7611              0.0672    0.0619
             interval
Balding        0.1108
No Hair Loss   0.0728

                                   95% conf. interval
             Relative Risk: 1.3996    1.1641   1.6827
         Sample Odds Ratio: 1.4411    1.1772   1.7641
Conditional MLE Odds Ratio: 1.4410    1.1676   1.7683
    Probability difference: 0.0268    0.0113   0.0444

             Exact P-value: 0.0006
        Asymptotic P-value: 0.0004
------------------------------------------------------
```

Excess Risk:

We are 95% confident that the probability of heart disease is between 0.0113 and 0.0444 higher for male physicans with hair loss than for male physicians without hair loss.

Relative Risk:

We are 95% confident that the probability of heart disease is between 1.16 and 1.68 times higher for male physicans with hair loss than for male physicians without hair loss.

Odds Ratio:

We are 95% confident that the odds of heart disease are between 1.18 and 1.76 times higher for male physicans with hair loss than for male physicians without hair loss.