STAT E-102
Homework 5: Solutions

Triin Umbleja

1. A study examined risk factors for allergic diseases among 13- and 14-year-old schoolchildren in Japan. One risk factor of interest was a family history of eating an unbalanced diet. The following table shows the cases and non-cases of children exhibiting symptoms of rhinitis in the presence and absence of the risk factor.

<table>
<thead>
<tr>
<th></th>
<th>Rhinitis</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cases</td>
<td>Non-cases</td>
<td>Total</td>
</tr>
<tr>
<td>Family History</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unbalanced diet</td>
<td>656</td>
<td>1451</td>
<td>2107</td>
<td></td>
</tr>
<tr>
<td>Balanced diet</td>
<td>677</td>
<td>1662</td>
<td>2339</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1333</td>
<td>3113</td>
<td>4446</td>
<td></td>
</tr>
</tbody>
</table>


First note that to use the formulas as specified in the textbook we need to rearrange our table so that the presence/non-presence of "exposure" (unbalanced diet) is in the columns and the "disease" (rhinitis cases and non-cases) is in the rows:

<table>
<thead>
<tr>
<th></th>
<th>Rhinitis</th>
<th>Unbalanced diet</th>
<th>Balanced diet</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>656</td>
<td>677</td>
<td>1333</td>
<td></td>
</tr>
<tr>
<td>Non-cases</td>
<td>1451</td>
<td>1662</td>
<td>3113</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2107</td>
<td>2339</td>
<td>4446</td>
<td></td>
</tr>
</tbody>
</table>

a) Calculate the relative risk of rhinitis for children with family history of unbalanced diet versus children with family history of balanced diet and interpret the result.

Relative risk (RR) is the proportion of children with rhinitis among those who eat unbalanced diet divided by the proportion of children with rhinitis among those who eat balanced diet:

\[ RR = \frac{a/(a + c)}{b/(b + d)} = \frac{656/2107}{677/2339} = 1.08 \]

⇒ that children eating unbalanced diet are 8% more likely to have rhinitis than the children eating balanced diet.
b) Calculate the two-sided 95% confidence interval for the difference between the proportion of rhinitis cases among children with family history of unbalanced diet and the proportion of rhinitis cases among children with family history of balanced diet. Does the confidence interval include 0? Would your conclusion be different if it would/would not include 0? Why?

The proportion of children suffering from rhinitis among those eating unbalanced diet is

\[ \hat{p}_1 = \frac{656}{1451} = 0.31. \]

and proportion of children suffering from rhinitis among those eating balanced diet is

\[ \hat{p}_2 = \frac{677}{1662} = 0.29. \]

The difference between the latter proportions is

\[ \hat{p}_1 - \hat{p}_2 = 0.31 - 0.29 = 0.02. \]

Now, the two-sided 95% confidence interval for the difference between the proportions is:

\[
\left( \hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} ; \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right) =
\left( 0.02 - 1.96 \sqrt{\frac{0.31(1-0.31)}{2107} + \frac{0.29(1-0.29)}{2339}} ; 0.02 + 1.96 \sqrt{\frac{0.31(1-0.31)}{2107} + \frac{0.29(1-0.29)}{2339}} \right) =
\left( -0.007 ; 0.047 \right).
\]

When checking whether this interval includes 0, we test hypotheses:

\[
\left\{ \begin{array}{l}
H_0 : p_1 - p_2 = 0 \\
H_1 : p_1 - p_2 \neq 0
\end{array} \right.
\]

Since the obtained confidence interval includes 0, we do not reject the null hypothesis that the proportions of children suffering from rhinitis among the children eating unbalanced diet and among those children eating balanced diet are not different. However, if 0 would not be in this interval we would reject the null hypothesis and conclude that the two proportions are significantly different.

c) What is the estimated odds ratio of having rhinitis among subjects with a family history of an unbalanced diet compared to those eating a balanced diet? Interpret the result?

\[ \hat{OR} = \frac{ad}{bc} = \frac{656 \cdot 1662}{677 \cdot 1451} = 1.1 \]

The children eating unbalanced diet have 1.1 times greater odds for rhinitis than children eating balanced diet.

d) Calculate the 95% two-sided confidence interval for the odds ratio. Are the conclusions consistent with the ones drawn in part b)?
As the distribution of OR is not symmetric, but skewed to the right, and the distribution of $ln(OR)$ appears to be symmetric and approximately normal, we will first find the two-sided 95% confidence interval for $ln(OR)$.

$$ln(\hat{OR}) = ln(1.1) = 0.1$$

The two-sided 95% CI for $ln(OR)$ is,

$$\left( ln(\hat{OR}) - 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} ; ln(\hat{OR}) + 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right) =$$

$$\left( 0.1 - 1.96 \sqrt{\frac{1}{656} + \frac{1}{677} + \frac{1}{1451} + \frac{1}{1662}} ; 0.1 + 1.96 \sqrt{\frac{1}{656} + \frac{1}{677} + \frac{1}{1451} + \frac{1}{1662}} \right) =$$

$$(-0.03 ; 0.23).$$

We now transform the lower and upper bound of the obtained confidence interval back to get the confidence interval for $OR$:

$$\left( exp(-0.03) ; exp(0.23) \right) = (0.97 ; 1.27).$$

We now test the hypotheses:

$$\left\{ \begin{array}{l} H_0 : \ OR = 1 \\ H_1 : \ OR \neq 1 \end{array} \right.$$  

Since the obtained confidence interval for $OR$ includes 1 we do not reject the null hypothesis: there was not enough evidence to show that there is an association between family history of diet and occurrence of rhinitis. The same conclusion was drawn in part b).

2. A group of children five years of age and younger who were free of respiratory problems were enrolled in a cohort study examining the relationship between parental smoking and the subsequent development of asthma. The association between maternal cigarette smoking status and a diagnosis of asthma before the age of twelve was examined separately for boys and for girls (§ 16.4 Problem 6 from the textbook).

<table>
<thead>
<tr>
<th>Boys</th>
<th>Smoking Status</th>
<th>Asthma diagnosis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\geq \frac{1}{2}$ pack/day</td>
<td>$&lt; \frac{1}{2}$ pack/day</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>17</td>
<td>41</td>
<td>58</td>
</tr>
<tr>
<td>No</td>
<td>63</td>
<td>274</td>
<td>337</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>337</td>
<td>395</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Girls</th>
<th>Smoking Status</th>
<th>Asthma diagnosis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\geq \frac{1}{2}$ pack/day</td>
<td>$&lt; \frac{1}{2}$ pack/day</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>8</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>No</td>
<td>55</td>
<td>261</td>
<td>316</td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>281</td>
<td>344</td>
</tr>
</tbody>
</table>
a) Estimate the relative odds of developing asthma for boys whose mothers smoke at least one half pack of cigarettes per day versus those whose mothers smoke less than this.

The estimated relative odds of developing asthma for boys are
\[ \hat{OR}_b = \frac{17 \cdot 274}{41 \cdot 63} = 1.8, \]
i.e. the boys whose mothers smoke at least one half pack per day have 1.8 times greater odds of developing asthma than those boys whose mothers smoke less than that.

b) Estimate the corresponding odds ratio for girls.

The estimated odds ratio for girls is
\[ \hat{OR}_g = \frac{8 \cdot 261}{20 \cdot 55} = 1.9, \]
i.e. the girls whose mothers smoke at least one half pack per day have 1.9 times greater odds of developing asthma than those whose mothers smoke less than that.

c) Conduct a test of homogeneity to determine whether it is appropriate to combine the information in the two 2x2 tables using the Mantel-Haenszel method. What do you conclude?

For the test of homogeneity the null hypothesis is that the odds ratios of developing asthma are identical for boys and girls:

\[ \begin{cases} 
H_0 : & \hat{OR}_b = \hat{OR}_g \\
H_1 : & \hat{OR}_b \neq \hat{OR}_g 
\end{cases} \]

We again transform our point estimates for odds ratios to the log-scale:
\[ y_b = \ln(\hat{OR}_b) = \ln(1.8) = 0.59 \]
and
\[ y_g = \ln(\hat{OR}_g) = \ln(1.9) = 0.64. \]

Since
\[ w_b = \left[ \frac{1}{a_b} + \frac{1}{b_b} + \frac{1}{c_b} + \frac{1}{d_b} \right]^{-1} = \left[ \frac{1}{17} + \frac{1}{41} + \frac{1}{63} + \frac{1}{274} \right]^{-1} = 9.7 \]
and
\[ w_g = \left[ \frac{1}{a_g} + \frac{1}{b_g} + \frac{1}{c_g} + \frac{1}{d_g} \right]^{-1} = \left[ \frac{1}{8} + \frac{1}{20} + \frac{1}{55} + \frac{1}{261} \right]^{-1} = 5.1, \]
the weighted average \( Y \) is
\[ Y = \frac{\sum_{i \in \{b,g\}} w_i y_i}{\sum_{i \in \{b,g\}} w_i} = \frac{9.7 \cdot 0.59 + 5.1 \cdot 0.64}{9.7 + 5.1} = 0.61. \]
The test statistic is
\[ \chi^2 = \sum_{i \in \{b,g\}} w_i (y_i - Y)^2 = 9.7(0.59 - 0.61)^2 + 5.1(0.64 - 0.61)^2 = 0.01. \]

For a chi-square distribution with 1 degrees of freedom, \( p \) is substantially greater than 0.01. Therefore, we do not reject the null hypothesis and may proceed with the Mantel-Haenszel method to find an estimate for the summary OR.

d) If it makes sense to do so, find a point estimate for the summary odds ratio and construct a 95\% interval.

The Mantel-Haenszel estimate for the summary odds ratio is
\[ \hat{OR}_{MH} = \sum_{i \in \{b,g\}} \left( \frac{a_i d_i / T_i}{b_i c_i / T_i} \right) = \frac{17 \cdot 274/395 + 8 \cdot 261/344}{41 \cdot 63/395 + 20 \cdot 55/344} = 1.83. \]

Before constructing a confidence interval, we must first have to check the assumptions that the strata are large enough to justify the use of normal distribution. Since
\[ \sum_{i \in \{b,g\}} M_{1i} N_{1i} / T_i = 80 \cdot 58 / 395 + 63 \cdot 28 / 344 = 16.9, \]
\[ \sum_{i \in \{b,g\}} M_{1i} N_{2i} / T_i = 80 \cdot 337 / 395 + 63 \cdot 316 / 344 = 126.1, \]
\[ \sum_{i \in \{b,g\}} M_{2i} N_{1i} / T_i = 315 \cdot 58 / 395 + 281 \cdot 28 / 344 = 69.1, \]
and
\[ \sum_{i \in \{b,g\}} M_{2i} N_{2i} / T_i = 315 \cdot 337 / 395 + 281 \cdot 316 / 344 = 526.9 \]
are all greater than 5, we may proceed constructing the confidence interval. From our previous calculations \( Y = 0.61 \). Since
\[ se(Y) = \frac{1}{\sqrt{w_b + w_g}} = \frac{1}{\sqrt{9.7 + 5.1}} = 0.26, \]
a 95\% confidence interval for \( \ln(OR) \) is
\[ (0.61 - 1.96 \cdot 0.26 ; 0.61 + 1.96 \cdot 0.26) = (0.097 ; 1.12). \]
and a 95\% confidence interval for the summary odds ratio itself is
\[ (\exp(0.097) ; \exp(1.12)) = (1.1 ; 3.1). \]
Since the obtained confidence interval does not contain 1 we conclude that the odds of developing asthma are significantly greater for those children whose mothers smoke at least half a pack per day than for those children whose mothers smoke less, controlling for gender.

e) What would you do if the results of the test of homogeneity led you to reject the null hypothesis that the odds ratio is identical for boys and girls?

In that case it would not make sense to combine the two odds ratios into a common one. Instead, separate odds ratios for boys and girls should be reported.