Review:

Common notation for a 2x2 table:

	Exposed	Not Exposed	
Disease	a	b	a+b
No Disease	c	d	c+d
	a+c	b+d	N

Warning: Rosner presents the transposed table so b and c have their meanings reversed in the text.

where $OR=\theta_1/(1-\theta_1)/\theta_2/(1-\theta_2)=ad/bc$

Here

θ_1=P(disease |exposed) and θ_2=P(disease| not exposed) or

θ_1=P(exposed |disease) and θ_2=P(exposed| not diseased)

The three statistics of interest comparing the exposed to the not exposed are:

Risk Difference = RD = a/(a+c) – b/(b+d)

Risk Ratio = RR = (a/(a+c)) / (b/(b+d))

Odds Ratio = OR = ad/bc

Confidence Intervals for RD, RR, OR

The interval for RD we've studied extensively coming up with an exact interval, an asymptotically normal (Wald) interval, and a modified Wald (Wilson) interval. For the RR, we need some calculus.

Delta Method: $Var[f(X)] \approx (f'(X))^2 * Var[X]$.

Take f = log(), and let $p_1=a/(a+c)$ and $p_2=b/(b+d)$
For simplicity write $p_1=a/n_1$ and $p_2=b/n_2$

$Var[\ \log(p_1/p_2)\ ] = Var[\ \log(p_1) - \log(p_2)\ ]$

$= Var[\ \log(p_1)\ ] + Var[\ \log(p_2)\ ]$ by independence

$\approx 1/p_1^2\ Var(p_1) + 1/p_2^2\ Var(p_2)$ by Delta Method

$= p_1*(1- p_1)/(n_1*p_1^2) + p_2*(1- p_2)/(n_2*p_2^2)$

$= (1- p_1)/(n_1*p_1) + (1- p_2)/(n_2*p_2)$

$\approx (c/n_1)/(n_1*a/n_1) + (d/n_2)/(n_2*b/n_2)$ by substituting our point estimates for p_1 and p_2

$= c/(a*n_1) + d/(b*n_2)$

$(1-\alpha)100\%$ CI give n_1 and n_2 are sufficiently large:

$\exp\{\ \log(p_1/p_2) \pm Z_{1-\alpha/2} * \mathrm{sqrt}(c/(a*n_1) + d/(b*n_2))\ \}$

Typically we test H_0: $\theta_1=\theta_2$ (or H_0: OR=1) with a Chi-square statistic, given by:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} = \left[\frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\theta_p(1-\theta_p)\left(\frac{1}{a+c} + \frac{1}{b+d}\right)}} \right]^2$$

Additionally, we can estimate the <u>Log</u> odds ratio with a (1-α)100% CI of the form:

$$\log(\hat{or}) \pm Z_{\alpha/2}\sqrt{Var(\log(\hat{or}))}$$

$$\text{where} \quad Var(\log(\hat{or})) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

To get the confidence interval for the Odds Ratio, you need to exponentiate the ends of the CI for the Log Odds ratio:

Use (e^l, e^u) where

$$l = \log(\hat{or}) - Z_{\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$u = \log(\hat{or}) + Z_{\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Remember that this exponentiated confidence interval is not symmetric about the Odds ratio!

Combining 2x2 tables

The Mantel-Haenszel method is an alternative to pooling data from multiple 2x2 tables. The primary motivation is to avoid Simpson's Paradox problems.

The basic idea is this: Instead of pooling the data and then computing the pooled OR, we will compute a weighted average of the odds ratios from each 2x2 table.

Suppose we have i=1,…,g 2x2 tables (or strata):

	Exposed	Not Exposed	
Disease	a_i	b_i	a_i+b_i
No Disease	c_i	d_i	c_i+d_i
	a_i+c_i	b_i+d_i	N_i

The odds ratio for the i^{th} strata is $OR_i = a_i d_i / b_i c_i$ and the simple summary OR is $OR_{pooled} = \sum a_i \sum d_i / \sum b_i \sum c_i$.

Let w_i be the *inverse* of the variance of the strata specific log-odds ratio:

$w_i = 1/\mathrm{Var}[\log(OR_i)]$ or $w_i = [1/a_i+1/b_i+1/c_i+1/d_i]^{-1}$

The basic idea of is to get a weighted average of the strata-specific odds ratios. This can be done on the log scale using weights that are inversely proportional to the variance of the strata specific estimate or it can be done on the correct scale using other weights.

⇨ Important side note: *Statisticians love to take weighted averages! But even more than that, they love to set the weights to be inversely proportional to the variance.*

Why? Because then estimates with a larger variance (less precise, less information) gets less weight and estimates with a smaller variance (more precise, more information) get more weight.

⇨ Clever statisticians have proven that using weights that are inversely proportional to the stratum-specific variances will always yield the weighted average with the smallest variance. (WOW!)

There are many different ways to estimate the <u>common odds ratio</u> underlying different 2x2 tables.

(Notice the implicit assumption here: There must be a common OR or else there is no reason to combine the data in the first place.)

One natural estimate of a common Odds Ratio is

$OR_{exp} = \exp\{ \sum w_i \log(OR_i)/ \sum w_i \}$ where

$w_i = [1/a_i+1/b_i+1/c_i+1/d_i]^{-1}$

The Mantel-Haenszel estimate of the common OR is

$OR_{mh} = \sum w_i OR_i/ \sum w_i$ where $w_i = b_i c_i /N_i$

Important side note (re-scaling the weights): Notice that we divide by $\sum w_i$. We do this because the weights to not sum to one, i.e. $w_1 + \ldots + w_g \neq 1$.

We have that $w_1 + \ldots + w_g = \sum w_i$ and

$(w_1 + \ldots + w_g)/\sum w_i = 1$

So therefore $w_1/\sum w_i + \ldots + w_g/\sum w_i = 1$ which we

re-write as $w^*_1 + \ldots + w^*_g = 1$ where $w^*_1 = w_1/\sum w_i$

Alternate forms of the above ORs are

$$OR_{exp} = \exp\{ \sum w^*_i \log(OR_i) \} \quad \text{where} \quad w^*_i = w_i/\sum w_i$$
$$w_i = [1/a_i+1/b_i+1/c_i+1/d_i]^{-1}$$

$$OR_{mh} = \sum w^*_i \, OR_i \quad \text{where} \quad w^*_i = w_i/\sum w_i$$
$$w_i = b_i c_i /N_i$$

The point of re-writing these estimates using weights that add to one, is to show that these estimates are subject to the same advantages and disadvantages of weighted averages just like those we explored in the context of crude and adjusted rates.

But here, instead of choosing weights according to the population distribution of interest, we choose weights according to some statistical criteria.

The advantage is often statistical. For example, OR_{exp} yields the combined estimate with the smallest possible variance and OR_{mh} works even when there are zero cells.

The disadvantage is that we can never compare two weighted odds ratios because their weights will be different (e.g., you can not compare the common OR over age between males and females because the weights are different for each gender).

Recall our Simpson's Paradox Scenario

Males	Smoking		
cancer	Yes	No	
Yes	**9**	**51**	60
No	**6**	**43**	49
	15	94	109

OR(males) = (9x43)/(6x51) = 1.264706

Females	Smoking		
cancer	Yes	No	
Yes	**14**	**7**	21
No	**19**	**12**	31
	33	19	52

OR(females) = (14x12)/(19x7) = 1.263158

combined	Smoking		
cancer	Yes	No	
Yes	**23**	**58**	81
No	**25**	**55**	80
	48	113	161

OR(combined) = (23x55)/(58x25) = 0.872

The first proposed estimate for the common OR is

$OR_{exp} = \exp\{ \sum w_i \log(OR_i)/ \sum w_i \}$ where
$w_i = [1/a_i+1/b_i+1/c_i+1/d_i]^{-1}$

= exp{ [
1/(1/9+1/51+1/6+1/43) * log[(9*43)/(6*51)] +
1/(1/14+1/7+1/19+1/12) * log[(14*12)/(19*7)]]
/ [1/(1/9+1/51+1/6+1/43) +
1/(1/14+1/7+1/19+1/12)] }

= exp{ 0.2342542 }

= 1.263966

The Mantel-Haenszel estimate of the common OR is

$OR_{mh} = \sum w_i OR_i/ \sum w_i$ where $w_i = b_i c_i /N_i$

= [(51*6/109) * (9*43)/(6*51) +
(7*19/52) * (14*12)/(19*7)]
/ [(51*6/109) + (7*19/52)]

= 1.263968

Notice that both estimators prevent Simpson's paradox.

Keep in mind that OR_{exp} and OR_{mh} will yield different estimates for the same (unknown) common OR.

Generally these differences are exacerbated when some cells of the 2x2 table are very small or zero. One work around is to add ½ to each cell (so-called `continuity correction').

Generally we use OR_{exp} to verify that the there is indeed one common underlying OR. This is called the Test of Homogeneity.

If indeed the strata are judged to be homogeneous by the above test, then it is sometimes common to estimate the common OR using Mantel-Haenszel's OR_{mh} and to check if $OR_{mh}=1$ statistically. This is done using a Test of Association.

These two steps combined is called the Mantel-Haenszel method.

Test of Homogeneity

The test of Homogeneity is based on OR_{exp} where

$OR_{exp} = \exp\{ \sum w_i \log(OR_i)/ \sum w_i \}$ and
$w_i = [1/a_i+1/b_i+1/c_i+1/d_i]^{-1}$

A $(1-\alpha)100\%$ CI for OR_{exp} is given by (e^l, e^u) where

$$l = \log\left(OR_{\exp}\right) - Z_{\alpha/2}\frac{1}{\sqrt{\sum w_i}}$$

$$u = \log\left(OR_{\exp}\right) + Z_{\alpha/2}\frac{1}{\sqrt{\sum w_i}}$$

The corresponding Hypothesis Test has the null hypothesis that H_0: $OR_1 = \ldots = OR_g$ and the test statistic is

$\chi^2 = \sum w_i [\log(OR_i) - \log(OR_{exp})]^2$

which is approximately distributed as Chi-square with g-1 df.

Test of Association

The test of association is based on OR_{mh} where

$OR_{mh} = \sum w_i OR_i / \sum w_i$ where $w_i = b_i c_i / N_i$

The null hypothesis that H_0: $OR_{mh} = 1$ and the test statistic is

$\chi^2 = \{ \sum a_i - \sum E[a_i] \}^2 / \sum Var_i$ ~ Chi-square, 1 df.

Where $E[a_i] = (a_i+b_i)(a_i+c_i)/N_i$ and

$Var_i = (a_i+b_i)(a_i+c_i)(b_i+d_i)(d_i+c_i)/N_i^2(N_i-1)$

The variance of OR_{mh} is too ugly to be put to paper.

We get a (1- α)100% CI for OR_{mh} by using the set of null hypothesis that do not reject at the α-level. This makes use of the duality of hypothesis testing and confidence intervals (sometimes it is called inverting the test), and this process can be expressed analytically:

$$\left[OR_{mh}^{\,1-\frac{Z_{\alpha/2}}{\sqrt{\chi^2}}} \;,\; OR_{mh}^{\,1+\frac{Z_{\alpha/2}}{\sqrt{\chi^2}}} \right]$$

Final thoughts

Tests of Homogeneity and Association are only used when the strata specific ORs all indicate associations in the same direction (i.e., all greater than one or less than one). If this not the case, the contribution of some tables can cancel contributions from others and make it appear there is no association when there is.

Males	Exposure		
Disease	Yes	No	
Yes	**15**	**5**	20
No	**85**	**95**	180
	100	100	200

OR(males)=(15x95)/(5x85)=3.353

Females	Exposure		
Disease	Yes	No	
Yes	**5**	**15**	20
No	**95**	**85**	180
	100	100	200

OR(Females)=(5x85)/(15x95)=0.298

Both	Exposure		
Disease	Yes	No	
Yes	**20**	**20**	40
No	**180**	**180**	360
	200	200	400

OR_{pooled}=(180x20)/(180x20)=1

$OR_{mh} = \sum w_i \, OR_i / \sum w_i$ where $w_i = b_i c_i / N_i$

OR_{mh} =
{85(5)/200 (3.353) + (15)95/200 (0.298)} /
{ 85(5)/200+(15)95/200 }
= 1

In this case the test statistic is zero (p-value=1).

$\chi^2 = \{ \sum a_i - \sum E[a_i] \}^2 / \sum Var_i = 0$

But $\sum a_i = 20$ and $\sum E[a_i]$=100(20)/200+(20)100/200=20

So $\sum a_i - \sum E[a_i] = 0$ and $\chi^2=0$ in this case.

Thus, we would fail to reject that males and females have the same OR; we fail to reject that the association between disease and exposure does not exist.