

Stat 525 Homework 4

Kenny Flagg

October 3, 2016

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.

Generally,

$$\widehat{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}.$$

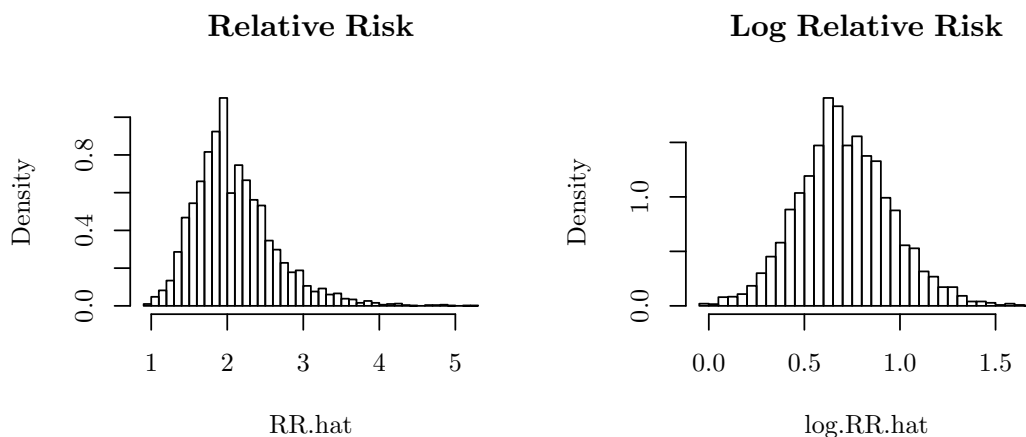
Replacing the cell counts with their expectations under independence,

$$\begin{aligned}\widehat{RR} &= \frac{\frac{\widehat{E}_{11}}{\widehat{E}_{11} + \widehat{E}_{12}}}{\frac{\widehat{E}_{21}}{\widehat{E}_{21} + \widehat{E}_{22}}} \\ &= \frac{\frac{n\widehat{P}(D)\widehat{P}(E)}{n\widehat{P}(D)\widehat{P}(E) + n\widehat{P}(\overline{D})\widehat{P}(E)}}{\frac{n\widehat{P}(D)\widehat{P}(\overline{E})}{n\widehat{P}(D)\widehat{P}(\overline{E}) + n\widehat{P}(\overline{D})\widehat{P}(\overline{E})}} \\ &= \frac{\frac{\widehat{P}(D)}{\widehat{P}(D) + \widehat{P}(\overline{D})}}{\frac{\widehat{P}(\overline{D})}{\widehat{P}(D) + \widehat{P}(\overline{D})}} \\ &= 1.\end{aligned}$$

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(1, 2))
hist(RR.hat, prob = TRUE, main = 'Relative Risk', breaks = 50)
hist(log.RR.hat, prob = TRUE, main = 'Log Relative Risk', breaks = 50)
```



The distribution of \widehat{RR} values has a noticeable long right tail, so even for this large of a sample, the normal distribution is a poor approximation of the sampling distribution. In contrast, the distribution of $\log(\widehat{RR})$ values is much more symmetric (although the right tail is heavier than the left). If we want to use the normal distribution to create an approximate confidence interval or make inference, we get a better approximation by working on the log scale.

- (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.

```
var(log.RR.hat)
```

```
[1] 0.05920074
```

The variance of the simulated values is 0.0592, which is slightly larger than the delta-method approximation,

$$\begin{aligned}
 \text{Var}\left(\log\left(\widehat{RR}\right)\right) &= \text{Var}\left(\log\left(\widehat{p}_1\right)\right) + \text{Var}\left(\log\left(\widehat{p}_2\right)\right) \\
 &= \frac{1-p_1}{n_E p_1} + \frac{1-p_2}{n_{\overline{E}} p_2} \\
 &= \frac{1-0.4}{100 \times 0.4} + \frac{1-0.2}{100 \times 0.2} \\
 &= 0.055,
 \end{aligned}$$

but the difference is probably not large enough to affect our inferences.

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	\bar{D}	
E	30	12	42
\bar{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).

```
DE <- cbind(D = c(E = 30, E.bar = 70), D.bar = c(E = 12, E.bar = 88))
twoby2(DE, alpha = 0.05)
```

2 by 2 table analysis:

Outcome : D
Comparing : E vs. E.bar

	D	D.bar	P(D)	95% conf. interval
E	30	12	0.7143	0.5614 0.8300
E.bar	70	88	0.4430	0.3676 0.5213

	95% conf. interval
Relative Risk: 1.6122	1.2442 2.0892
Sample Odds Ratio: 3.1429	1.5004 6.5832
Conditional MLE Odds Ratio: 3.1250	1.4299 7.2184
Probability difference: 0.2712	0.1023 0.4078

Exact P-value: 0.0029
Asymptotic P-value: 0.0024

The odds of disease given exposure are estimated to be 3.14 times larger than the odds of disease given no exposure. An approximate 95% Wald confidence interval for the true odds ratio is (1.50, 6.58).

(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	\bar{D}	
< 40	50	80	130
≥ 40	50	20	70
	100	100	200

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

Disease and age:

```
D.age <- cbind(D = c(` $< 40`$  = 50, ` $\geq 40`$  = 50), D.bar = c(` $< 40`$  = 80, ` $\geq 40`$  = 20))
twoby2(D.age, alpha = 0.05)
```

2 by 2 table analysis:

Outcome : D

Comparing : < 40 vs. ≥ 40

	D	D.bar	P(D)	95% conf. interval
< 40	50	80	0.3846	0.3051 0.4709
≥ 40	50	20	0.7143	0.5981 0.8077

	95% conf. interval
Relative Risk:	0.5385 0.4139 0.7005
Sample Odds Ratio:	0.2500 0.1335 0.4682
Conditional MLE Odds Ratio:	0.2518 0.1263 0.4880
Probability difference:	-0.3297 -0.4514 -0.1864

Exact P-value: 0

Asymptotic P-value: 0

The odds of disease given that the person is under 40 are estimated to be 0.250 times the odds of disease given that the person is at least 40, with an approximate 95% CI of (0.136, 0.468).

Exposure and age:

```
E.age <- cbind(E = c(` $< 40`$  = 12, ` $\geq 40`$  = 30), E.bar = c(` $< 40`$  = 118, ` $\geq 40`$  = 40))
twoby2(E.age, alpha = 0.05)
```

2 by 2 table analysis:

Outcome : E

Comparing : < 40 vs. ≥ 40

	E	E.bar	P(E)	95% conf. interval
< 40	12	118	0.0923	0.0532 0.1555
≥ 40	30	40	0.4286	0.3184 0.5463

	95% conf. interval
Relative Risk:	0.2154 0.1178 0.3937
Sample Odds Ratio:	0.1356 0.0634 0.2898
Conditional MLE Odds Ratio:	0.1372 0.0580 0.3064
Probability difference:	-0.3363 -0.4592 -0.2106

Exact P-value: 0

Asymptotic P-value: 0

The odds of exposure given that the person is under 40 are estimated to be 0.136 times the odds of exposure given that the person is at least 40, with an approximate 95% CI of (0.0634, 0.290).

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

Age < 40		
	D	\bar{D}
E	5	6
\bar{E}	45	74
	50	80
	130	

Age \geq 40		
	D	\bar{D}
E	25	6
\bar{E}	25	14
	50	20
	70	

Age < 40:

```
DE.under40 <- cbind(D = c(E = 5, E.bar = 45), D.bar = c(E = 6, E.bar = 74))
twoby2(DE.under40, alpha = 0.05)
```

2 by 2 table analysis:

```
-----
Outcome      : D
Comparing    : E vs. E.bar
```

	D	D.bar	P(D)	95% conf. interval
E	5	6	0.4545	0.2028 0.7319
E.bar	45	74	0.3782	0.2957 0.4683

	95% conf. interval
Relative Risk: 1.2020	0.6046 2.3896
Sample Odds Ratio: 1.3704	0.3953 4.7512
Conditional MLE Odds Ratio: 1.3669	0.3109 5.7260
Probability difference: 0.0764	-0.1815 0.3541

```
Exact P-value: 0.7483
Asymptotic P-value: 0.6194
-----
```

For individuals under 40, the odds of disease given exposure are estimated to be 1.37 times the odds of disease given no exposure, with an approximate 95% CI of (0.395, 4.75).

Age \geq 40:

```
DE.atleast40 <- cbind(D = c(E = 25, E.bar = 25), D.bar = c(E = 6, E.bar = 14))
twoby2(DE.atleast40, alpha = 0.05)
```

2 by 2 table analysis:

```
-----
Outcome      : D
Comparing    : E vs. E.bar
```

	D	D.bar	P(D)	95% conf. interval	
E	25	6	0.8065	0.6309	0.9104
E.bar	25	14	0.6410	0.4814	0.7745

	95% conf. interval		
Relative Risk:	1.2581	0.9401	1.6836
Sample Odds Ratio:	2.3333	0.7725	7.0478
Conditional MLE Odds Ratio:	2.3056	0.6927	8.5633
Probability difference:	0.1654	-0.0489	0.3523

Exact P-value:	0.1837
Asymptotic P-value:	0.133

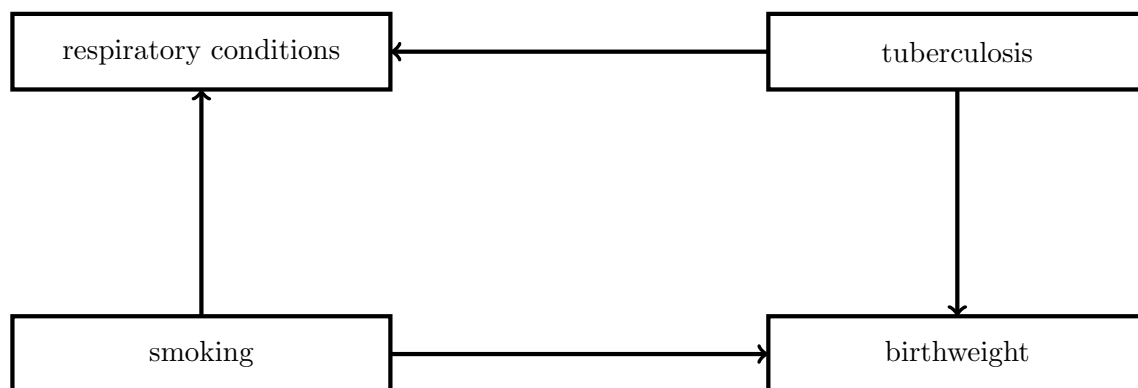
For individuals who are at least 40, the odds of disease given exposure are estimated to be 2.33 times the odds of disease given no exposure, with an approximate 95% CI of (0.773, 7.05).

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

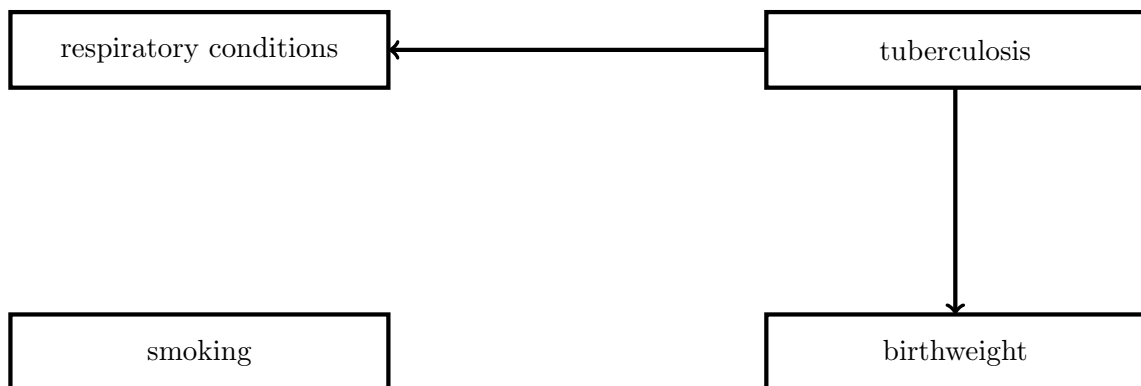
There is evidence of confounding. Age appears to be related to both disease and exposure because the confidence intervals in (b) do not include 1. Furthermore the pooled results disagree with the results from stratifying by age – in part (a) there appeared to be an association between exposure and disease, but this association disappears in part (c) when conditioning on age.

4. *Problem 8.3 on page 119.*

Assuming no direct effect of respiratory conditions on birthweight:

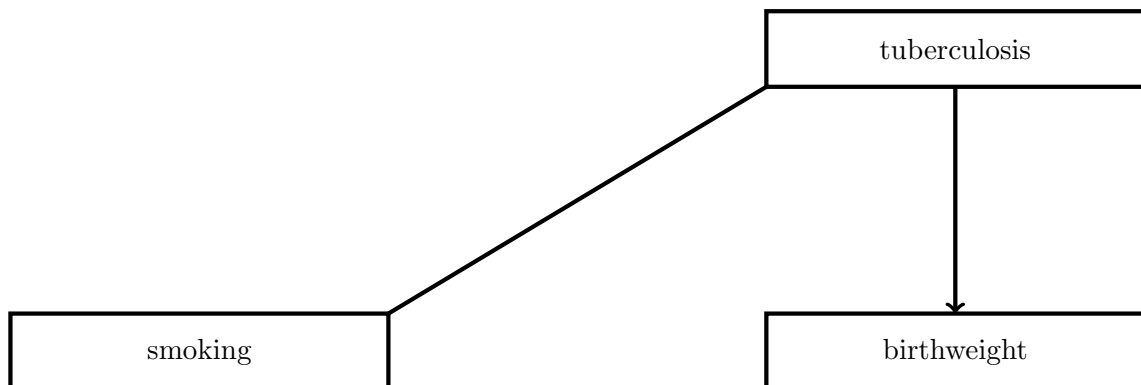


Removing the paths leaving smoking:



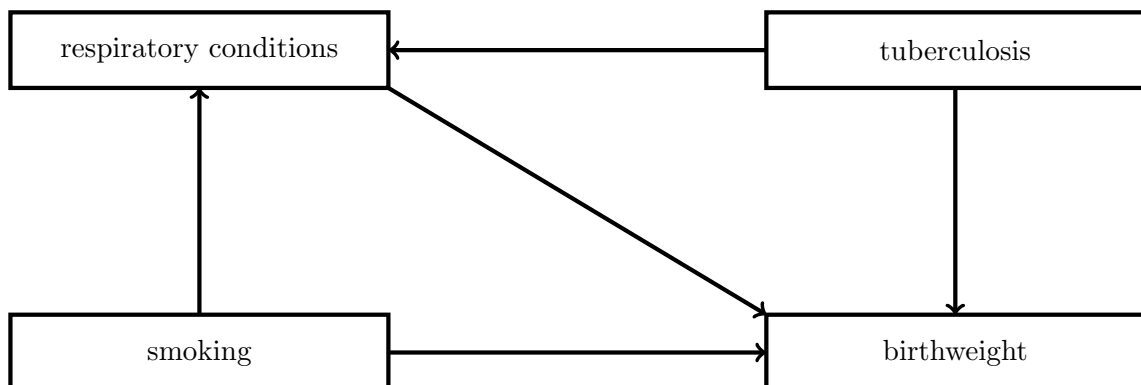
There are no unblocked backdoor paths from birthweight to smoking so the association is not confounded.

Stratifying on respiratory conditions:

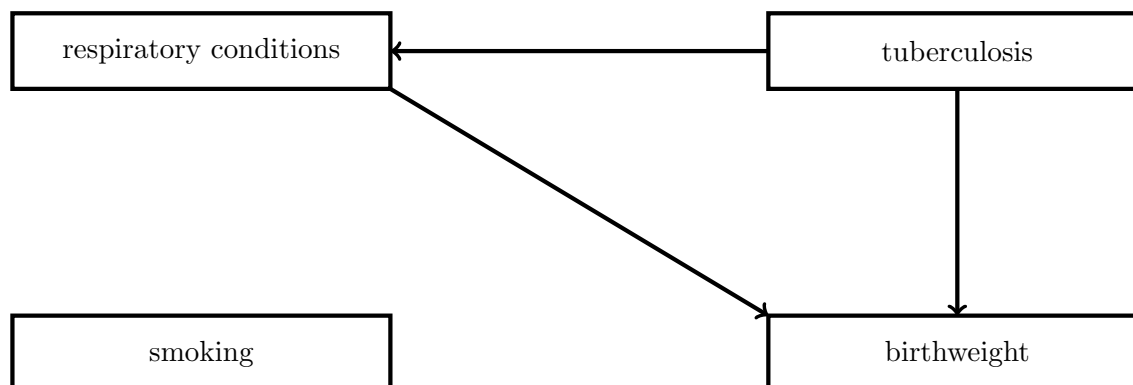


Respiratory conditions was a collider in the original graph, so stratifying on respiratory conditions results in tuberculosis becoming a confounder.

Assuming a direct effect of respiratory conditions on birthweight:

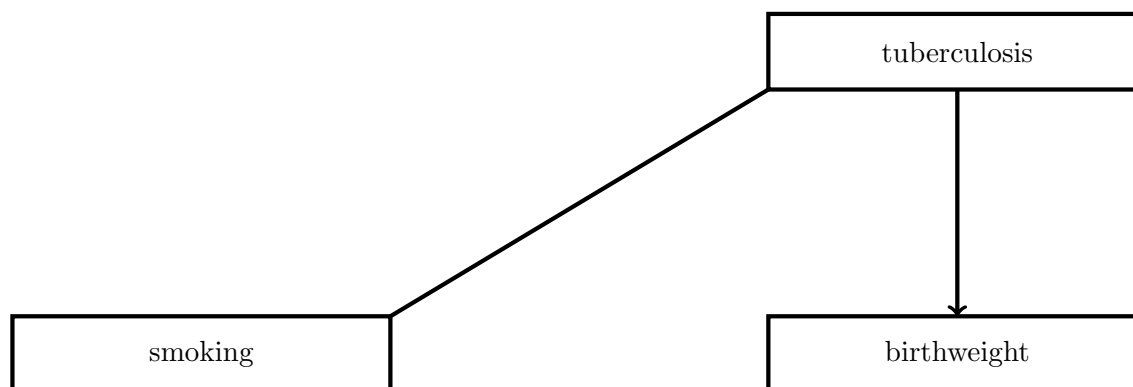


Removing the paths leaving smoking:



There are no unblocked backdoor paths from birthweight to smoking so that association is not confounded by another variable. Note however that tuberculosis confounds the association between respiratory conditions and birthweight.

Stratifying on respiratory conditions:



Again, stratifying on respiratory conditions leads to confounding of smoking with tuberculosis. To avoid confounding, we would need to condition on tuberculosis as well.

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

First, note that

$$\begin{aligned}
 \frac{d}{dp} \left[\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) \right] &= \frac{d}{dp} [\log(\hat{p}) - \log(1 - \hat{p})] \\
 &= \frac{1}{\hat{p}} + \frac{1}{1 - \hat{p}} \\
 &= \frac{1}{\hat{p}(1 - \hat{p})},
 \end{aligned}$$

so

$$\begin{aligned}\widehat{\text{Var}} \left[\log \left(\frac{\hat{p}}{1 - \hat{p}_1} \right) \right] &\approx \left(\frac{1}{\hat{p}(1 - \hat{p})} \right)^2 \widehat{\text{Var}}(\hat{p}) \\ &= \left(\frac{1}{\hat{p}(1 - \hat{p})} \right)^2 \frac{\hat{p}(1 - \hat{p})}{n} \\ &= \frac{1}{n\hat{p}(1 - \hat{p})}.\end{aligned}$$

Then

$$\begin{aligned}\widehat{\text{Var}} \left[\log(\widehat{OR}) \right] &= \widehat{\text{Var}} \left[\log \left(\frac{\hat{p}_1}{1 - \hat{p}_1} \right) \right] + \widehat{\text{Var}} \left[\log \left(\frac{\hat{p}_2}{1 - \hat{p}_2} \right) \right] \\ &= \frac{1}{n_E \hat{p}_1 (1 - \hat{p}_1)} + \frac{1}{n_E \hat{p}_2 (1 - \hat{p}_2)} \\ &= \frac{1}{(a + b) \left(\frac{a}{a+b} \right) \left(\frac{b}{a+b} \right)} + \frac{1}{(c + d) \left(\frac{c}{c+d} \right) \left(\frac{d}{c+d} \right)} \\ &= \frac{a + b}{ab} + \frac{c + d}{cd} \\ &= \frac{b}{ab} + \frac{a}{ab} + \frac{d}{cd} + \frac{c}{cd} \\ &= \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.\end{aligned}$$

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

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

This can be reexpressed as

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

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

Note that

$$\frac{n_{Dn_E}}{n^2} = \left(\frac{n_{11}}{n} + \frac{n_{21}}{n} \right) \left(\frac{n_{11}}{n} + \frac{n_{12}}{n} \right)$$

so the terms inside the summation depend only on the n_{ij}/n . Thus the sum is constant, so X^2 is proportional to n . This means that, as n increases, it becomes easier to reject the null hypothesis of independence, so with very large samples we should expect to find evidence of an association.