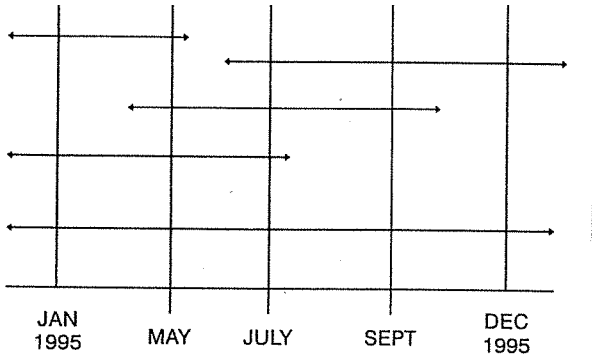
Stat 525 Midterm Exam

Kenny Flagg

October 12, 2016

1. *The figure below shows cases of a disease in a defined population of 50 individuals. There are 5 cases fo a disease. The disease is fatal so the length of the lines shows the survival time.*



   (a) *What is the incidence proportion over the specified period of time?*

   The incidence proportion is

$$\frac{\text{2 new cases}}{\text{47 at risk in January}} = 0.04255$$

   (b) *What is the point prevalence in July?*

   The point prevalence in July is

$$\frac{\text{4 diseased}}{\text{49 alive}} = 0.08163$$

   (c) *What is the point prevalence in December?*

   The point prevalence in December is

$$\frac{\text{2 diseased}}{\text{47 alive}} = 0.04255$$

2. *The data are described in Problem 9.1 on pages 144-145 and the problem continues as Problem 10.3 on page 163. Refer. to the specific questions below. R code to load these data into R is attached.*

```
# Problem 2 Data
# Pooled Data
TB.pool<-array(c(50,238,21,524),dim=c(2,2),
dimnames=list(c("BiomassYes","BiomassNo"),c("Case","Control")))
TB<-array(c(38,102,12,141,12,136,9,383),dim=c(2,2,2),
dimnames=list(c("BiomassYes","BiomassNo"),c("Case","Control"),c("< 1000",">= 1000")))
```

(a) *Calculate the ratio of the odds of TB given exposure to biomass fuel over the odds of TB given no exposure for the pooled data. Also give an approximate 95% confidence interval and interpret in terms of the problem.*

```
pool.OR <- (TB.pool[1,1] * TB.pool[2,2]) / (TB.pool[2,1] * TB.pool[1,2])
print(pool.OR)
```

```
[1] 5.242097
```

```
pool.logOR.SE <- sqrt(1/TB.pool[1,1] + 1/TB.pool[2,1] + 1/TB.pool[1,2] + 1/TB.pool[2,2])
pool.CI <- pool.OR * exp(qnorm(c(0.025, 0.975)) * pool.logOR.SE)
print(pool.CI)
```

```
[1] 3.078771 8.925503
```

The odds of TB for exposed individuals are estimated to be 5.242 times larger than the odds of TB for unexposed individuals. We are 95% confident that the true odds ratio is between 3.079 and 8.926.

(b) *Calculate the stratum specific (< 1000, ≥ 1000 peso income) odds ratios. Again provide me with the approximate 95% confidence intervals.*

```
OR1 <- (TB[1,1,1] * TB[2,2,1]) / (TB[2,1,1] * TB[1,2,1])
print(OR1)
```

```
[1] 4.377451
```

```
logOR1.SE <- sqrt(1/TB[1,1,1] + 1/TB[2,1,1] + 1/TB[1,2,1] + 1/TB[2,2,1])
OR1.CI <- OR1 * exp(qnorm(c(0.025, 0.975)) * logOR1.SE)
print(OR1.CI)
```

```
[1] 2.179825 8.790648
```

For individuals with monthly family incomes below 1,000 pesos, the odds of TB for exposed individuals are estimated to be 4.377 times larger than the odds of TB for unexposed individuals. We are 95% confident that the true odds ratio is between 2.180 and 8.791.

```
OR2 <- (TB[1,1,2] * TB[2,2,2]) / (TB[2,1,2] * TB[1,2,2])
print(OR2)
```

```
[1] 3.754902
```

```
logOR2.SE <- sqrt(1/TB[1,1,2] + 1/TB[2,1,2] + 1/TB[1,2,2] + 1/TB[2,2,2])
OR2.CI <- OR2 * exp(qnorm(c(0.025, 0.975)) * logOR2.SE)
print(OR2.CI)
```

```
[1] 1.547951 9.108355
```

For individuals with monthly family incomes of 1,000 pesos or more, the odds of TB for exposed individuals are estimated to be 3.755 times larger than the odds of TB for unexposed individuals. We are 95% confident that the true odds ratio is between 1.548 and 9.108.

(c) *Using the Cochran-Mantel-Haenzel approach provide me with an estimate of the common odds ratio adjusted for income along with an approximate 95% confidence interval. Use the* `mantelhaen.test` *function.*

```
mantelhaen.test(TB)
```

```
Mantel-Haenszel chi-squared test with continuity correction

data:  TB
Mantel-Haenszel X-squared = 27.363, df = 1, p-value = 1.686e-07
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 2.401977 7.199451
sample estimates:
common odds ratio
         4.158475
```

Adjusting for income level, the estimated odds ratio is 4.158 with an approximate 95% confidence interval of 2.402 to 7.199.

(d) *A key assumption of the CMH procedure is that there is no interaction. Is there any evidence of an interaction involving income and biomass fuel exposure? You can use the Breslow-Day test.*

```
breslowday.test(TB)
```

```
         < 1000   >= 1000
log OR 1.476467 1.3230622
Weight 0.120520 0.1979638
        OR       Stat         df     pvalue
4.15847515 0.06887353 1.00000000 0.79298404
```

The Breslow-Day statistic is $\chi^2_1 = 0.06887$ with p-value $= 0.7930$, so there is very little evidence of an interaction between income and exposure. It is safe to assume there is no interaction.

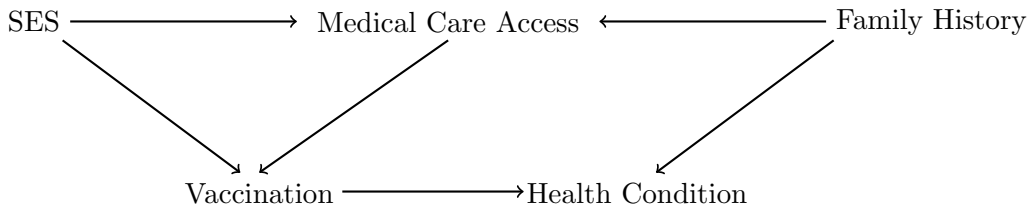
(e) *Based on the above which odds ratio would you present in, say, a final report and why? Filling in a table like the one below may help you with your answer.*

	OR	CI
Pooled	5.242	(3.079, 8.926)
CMH	4.158	(2.402, 7.199)
< 1000	4.377	(2.180, 8.791)
≥ 1000	3.755	(1.548, 9.108)

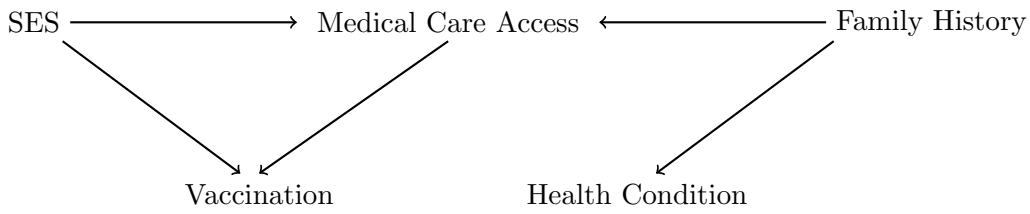
The Breslow-Day test results suggest interaction is not a problem, so we can report a single odds ratio. The pooled odds ratio is larger than the stratum-specific odds ratios. This is evidence of confounding so we should report the CMH odds ratio to adjust for confounding by income level.

3. *Refer to Figure 8.6 on page 106 in the text. Assume that access to medical care has no influence on health condition.*

(a) *Draw a new causal graph reflecting this assumption.*

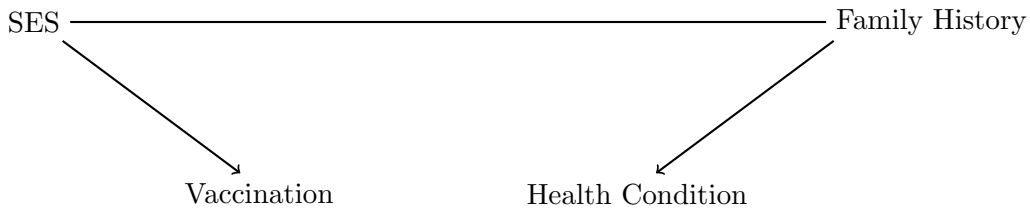


(b) *Is there any confounding of the relationship between vaccination and health condition? If so identify which of the 3 variables are confounders: socioeconomic status, access to medical care, and/or family history and explain why the confounding exists. If no confounding exists, explain why.*



The path Health Condition → Family History → Medical Care Access → Vaccination is the only unblocked backdoor path. Stratifying on Family history would remove this path because no pair of nodes have a common decendent through Family History. Therefore, Family History is a confounder, and the other variables are not confounders.

(c) *An investigation adjusts for access to medical care. Will their analysis of the relationship between vaccination and health condition be confounded? If so identify which of the remaining 2 are confounders: socioeconomic status and/or family history and explain why the confounding exists. If no confounding exists, explain why.*



If the investigator adjusts for Medical Care Access, there still exists an unblocked back-door path Health Condition → Family History → SES → Vaccination. SES and Family History are both confounders but controlling for either of these variables would remove the confounding.

4. *A retrospective study of lung cancer and tobacco smoking in several English hospitals resulted in the following data.*

```
# Problem 4
LC<-array(c(7,55,489,475,293,38,61,129,570,431,154,12),dim=c(6,2),
dimnames=list(c("None","<5","5-14","15-24","25-49","50+"),c("LCYes","LCNo")))
print(LC)
```

```
      LCYes LCNo
None      7   61
<5       55  129
5-14    489  570
15-24   475  431
25-49   293  154
50+      38   12
```

(a) *Test for independence between lung cancer and smoking using the chi-squared test for independence. Interpret the results.*

```
# row totals
n<-rowSums(LC)
# number of lung cancer cases
d<-LC[,1]
prop.test(d,n)
```

```
6-sample test for equality of proportions without continuity
correction

data:  d out of n
X-squared = 137.72, df = 5, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
```

```
   prop 1    prop 2    prop 3    prop 4    prop 5    prop 6
0.1029412 0.2989130 0.4617564 0.5242826 0.6554810 0.7600000
```

The test statistic is $\chi^2_5 = 137.72$ with p-value < 0.0001 so there is very strong evidence of an association between the number of cigarettes per day and lung cancer.

(b) *You should have found strong evidence of an association. Where does the independence break down and how?*

```
chisq.test(LC)$stdres
```

```
          LCYes      LCNo
None  -6.632073  6.632073
<5    -5.650247  5.650247
5-14  -3.187447  3.187447
15-24  1.790995 -1.790995
25-49  7.193499 -7.193499
50+    3.711301 -3.711301
```

All of the standardised residuals are large ($|r_{ij}| > 3$) except for the 15-24 cigarettes per day group. People who don't smoke or who smoke fewer than 15 cigarettes per day have lower risk of lung cancer than we would expect in the absence of association, while people who smoke 25 or more cigarettes a day have higher lung cancer risk than would be expected assuming no association.

(c) *Test for a trend in risk of lung cancer over smoking level. Interpret the results. Do the results of this test tell you anything about the direction of the trend?*

```
prop.trend.test(d,n)
```

```
Chi-squared Test for Trend in Proportions

data:  d out of n ,
 using scores: 1 2 3 4 5 6
X-squared = 129.23, df = 1, p-value < 2.2e-16
```

The trend test statistic is $\chi^2_1 = 129.23$ with p-value < 0.0001, giving very strong evidence that lung cancer rates have a linear trend with respect to the number of cigarettes smoked. This result alone does not provide any information about the direction of the trend, so we need to look at the estimated proportions output in part (a) to see that the trend is increasing.

(d) *Perform a goodness-of-fit test and give the results.*

```
pchisq(8.49, 4, lower.tail = FALSE)
```

```
[1] 0.07519092
```

The goodness-of-fit test statistic is $\chi^2_4 = 137.23 - 129.23$ with p-value $= 0.07519$, weak evidence that a more complicated model would fit better than the linear trend. The linear trend describes the estimated proportions well, but there may be room for improvement.

5. *An experiment on the efficacy of a vaccine was carried out on 215 volunteers, 104 of which were randomly assigned to receive the vaccine with the rest receiving a placebo injection. The results were*

	Caught Disease	Did Not Catch Disease	
Vaccine	10	94	104
Placebo	33	78	111
	43	172	

Estimate the Absolute Risk Reduction and the Number Needed to Treat. Give approximate 95% confidence intervals for these and interpret the results.

```
# Problem 5
vaccine<-array(c(33,10,78,94),dim=c(2,2),dimnames=list(c("Placebo","Vaccine"),
c("Disease","No Disease")))
prop.test(vaccine[,1], rowSums(vaccine))
```

```
2-sample test for equality of proportions with continuity correction

data:  vaccine[, 1] out of rowSums(vaccine)
X-squared = 12.349, df = 1, p-value = 0.0004412
alternative hypothesis: two.sided
95 percent confidence interval:
 0.08965445 0.31263245
sample estimates:
    prop 1     prop 2
0.29729730 0.09615385
```

$\widehat{ARR} = 0.29730 - 0.09615 = 0.20115$ with approximate 95% confidence interval (0.08965, 0.31263). Assuming the individuals in the vaccinated population and the individuals who receive placebos are otherwise similar, we expect a disease rate among people who receive the vaccine to be 0.20115 lower than the disease rate for individuals who take placebos.

$\widehat{NNT} = \dfrac{1}{0.20115} = 4.971$ with an approximate 95% confidence interval of $\left(\dfrac{1}{0.31263}, \dfrac{1}{0.08965}\right) = (3.198, 11.15)$. For every additional 4.971 individuals who receive the vaccine, we expect one fewer case of the disease.

6. *On Homework 4 you used the Delta Method to show that an approximation to the standard error of the MLE of the log of the odds*

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right)$$

was given by

$$SE\left[\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right)\right] = \sqrt{\frac{1}{n\widehat{p}\,(1-\widehat{p})}}$$

(a) *Assuming approximate normality and approximate unbiasedness of the estimator (not unreasonable for the asymptotic behavior of an MLE) show that an approximate 100(1 - α)% Wald CI for the log of the odds is given by*

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm z_{1-\alpha/2}/\sqrt{n\widehat{p}\,(1-\widehat{p})}$$

For large n,

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \sim \mathrm{N}\left(\log\left(\frac{p}{1-p}\right), \frac{1}{np\,(1-p)}\right)$$

and therefore

$$P\left[\left|\frac{\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) - \log\left(\frac{p}{1-p}\right)}{\sqrt{\frac{1}{np(1-p)}}}\right| < z_{1-\alpha/2}\right] \approx 1-\alpha$$

so

$$P\left[\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) - \frac{z_{1-\alpha/2}}{\sqrt{n\widehat{p}\,(1-\widehat{p})}} < \log\left(\frac{p}{1-p}\right) < \log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) + \frac{z_{1-\alpha/2}}{\sqrt{n\widehat{p}\,(1-\widehat{p})}}\right] \approx 1-\alpha$$

Thus $\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm z_{1-\alpha/2}/\sqrt{n\widehat{p}\,(1-\widehat{p})}$ is an approximate 95% confidence interval for $\log\left(\frac{p}{1-p}\right)$.

(b) *Show how to use this interval to obtain and interval for p itself.*

The inverse of $\phi = \log\left(\frac{p}{1-p}\right)$ is $p = \frac{\exp(\phi)}{1+\exp(\phi)}$. Applying the inverse to the endpoints of the interval from part (a),

$$\frac{\exp\left(\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm \frac{z_{1-\alpha/2}}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right)}{1+\exp\left(\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm \frac{z_{1-\alpha/2}}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right)} = \frac{\left(\frac{\widehat{p}}{1-\widehat{p}}\right)\exp\left(\pm\frac{z_{1-\alpha/2}}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right)}{1+\left(\frac{\widehat{p}}{1-\widehat{p}}\right)\exp\left(\pm\frac{z_{1-\alpha/2}}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right)}$$

so this is an approximate 95% confidence interval for p.