Homework #1 - STAT 453

By: Jordan Poles

In [1]:
```
from IPython.display import display
import numpy as np
import pandas as pd
from contingency import *
from strata import *
```

Question 7.2

Using the data from Table 6.7, estimate the Odds Ratio, with an associated 95% confidence interval, for the association between current exposure to biomass fuel and tuberculosis. Repeat your analysis using the small sample adjustments for an estimate and confidence interval for an Odds Ratio. Compare the results from the two approaches and comment.

In [2]:
```
biomass = ContTable(50, 238, 21, 524, 0)
print(biomass.num_tot, "total participants in this study.")
biomass.display()
```

```
833 total participants in this study.
2x2 Contingency Table:
```

	E	~E
D	50	238
~D	21	524

In [3]:
```
_ = biomass.odds_ratio(CI=.95, verbose=2)
```

```
Odds Ratio (OR): 5.2420968387354945
=> 95% CI using logOR: 1.656722
=> => logOR Var: 0.07372912523788142
=> => Upper Bound: 8.925503208782592
=> => Lower Bound: 3.078770868587116
=> Odds Ratio (OR) indicates a positive relationship between D & E.
```

In [4]:
```
_ = biomass.ss_odds_ratio(CI=.95, verbose=2)
```

```
Small Sample Odds Ratio (ss_OR): 4.98288322556105
=> 95% CI using ss_logOR: 1.641997
=> => ss_logOR Var: 0.07241305791543796
=> => Upper Bound: 8.753174602530487
=> => Lower Bound: 3.0482772379608516
=> Small Sample Odds Ratio (OR) indicates a positive relationship bet
ween D & E.
```

In [5]:
```
biomass.OR - biomass.ss_OR
```

Out[5]:
```
0.25921361317444447
```

The results of the small sample adjustment for the Odds Ratio (OR) was not largely noticeable. The OR was reduced by the adjustment, by around .26, with a concomitant tiny reduction in the logOR variance of approximately .0013. Given there were 833 participants in this study, the small sample adjustment does not seem particularly necessary in this circumstance.

Question 7.3

Tuyns et al. (1977) carried out a case-control study of esophageal cancer in the region known as Ille-et-Vilaine in Brittany, France. The data set, oesoph, can be found at http://www.crcpress.eom/ejtroducts/downloads/ (http://www.crcpress.eom/ejtroducts/downloads/) and is also examined in detail in Breslow and Day (1980). One risk factor of interest was daily alcohol consumption, measured in grams per day, given in the data set in four levels: 0 to 39, 40 to 79, 80 to 120, and >120g/day. Dichotomize this risk factor according to whether an individual's alcohol. consumption is less than or greater than or equal to 80 g/day. With this binary measure of alcohol consumption, estimate the Odds Ratio with an associated 95% confidence interval. Also, examine the relationship between incidence of esophageal cancer and the dichotomized measure of alcohol consumption using the chisq test.

I sourced my dataset from http://forge.scilab.org/index.php/p/rdataset/source/tree/master/csv/datasets/esoph.csv (http://forge.scilab.org/index.php/p/rdataset/source/tree/master/csv/datasets/esoph.csv), as the original link was broken.

In [6]:
```
esoph = pd.read_csv("data/esoph.csv")
esoph.drop(esoph.columns[[0]], axis=1, inplace=1)
alcbinary = esoph["alcgp"].isin(["80-119", "120+"])
exposed = esoph.loc[alcbinary, ["ncases", "ncontrols"]].sum()
unexposed = esoph.loc[~alcbinary, ["ncases", "ncontrols"]].sum()
#third + fourth argument adjusted to fit book data
esoph_alc_table = ContTable(exposed["ncases"], unexposed["ncases"], e
xposed["ncontrols"]-exposed["ncases"], unexposed["ncontrols"]-unexpos
ed["ncases"], 0)
esoph_alc_table.display()
esoph_alc_table.odds_ratio(verbose=2)
_ = esoph_alc_table.chisq_indep(verbose=2)
```

```
2x2 Contingency Table:
```

	E	**~E**
D	96	104
~D	109	666

```
Odds Ratio (OR): 5.64008468596
=> 95% CI using logOR: 1.729899
=> => logOR Var: 0.0307078647102
=> => Upper Bound: 7.951467464104598
=> => Lower Bound: 4.000589250771749
=> Odds Ratio (OR) indicates a positive relationship between D & E.

ChiSquared Independence Test (n=975)
=> test statistic = 110.255387414
=> p-value ≈ 0.0
=> ChiSquared Test suggests an association between D & E.
```

Question 9.1

In Perez-Padilla et al. (2001) the authors were concerned that the variable, monthly family income (an indicator of economic status), might confound the observed association between indoor air pollution and tuberculosis. The data in Table 6.7, stratified by income, are shown in Table 9.13, with income information coded as "<1000 pesos per month" and "1000 or more pesos per month."

Based on the income strata, (1) what is the Odds Ratio associating biomass fuel exposure, for the low income (<1000 pesos/month) stratum? (2) Similarly, what is the Odds Ratio for the high income (1000 + pesos/month) stratum? Now estimate the Odds Ratio, associated with biomass fuel exposure, adjusting for income, using the Mantel-Haenszel method.

Part 1

We calculate the OR for the individual groups (wealthy and poor) using simple Contingency Tables.

In [7]:
```
poor_exposure = ContTable(38, 102, 12, 141, 0)
_ = poor_exposure.odds_ratio(verbose=1)
```

```
Odds Ratio (OR): 4.377450980392157
=> 95% CI using logOR: 1.476467
=> => logOR Var: 0.1265452429572053
=> => Upper Bound: 8.790648269155707
=> => Lower Bound: 2.1798252528168405
```

In [8]:
```
wealthy_exposure = ContTable(12, 136, 9, 383, 0)
_ = wealthy_exposure.odds_ratio(verbose=1)
```

```
Odds Ratio (OR): 3.7549019607843137
=> 95% CI using logOR: 1.323062
=> => logOR Var: 0.2044083516783563
=> => Upper Bound: 9.108354899055
=> => Lower Bound: 1.5479511823331233
```

Part 2

We now turn to the Mantel-Haenszel method to correct for the effect of this confounding variable on the Odds Ratio.

In [9]:
```
biofuel_strata = Strata([poor_exposure, wealthy_exposure])
_ = biofuel_strata.mantel_haenszel_correction()
```

```
Mantel-Haenszel adj. OR: 4.15847515476779
=> 95% CI using logOR: 1.425148
=> => logOR Var: 0.07841897932703254
=> => Upper Bound: 7.199451163453156
=> => Lower Bound: 2.401976931325776
```

Part 3 - Discussion

Based on your calculations and those reported from Question 6.4, what is your assessment of confounding by monthly income? Does the income variable fulfill the criteria required for a variable to be a confounder? In addition to income, the authors considered many other confounding factors: age, sex, urban or rural residence, smoking, crowding, level of education, and socioeconomic status. After controlling for all these confounders, their estimate of the adjusted Odds Ratio between indoor air pollution and tuberculosis was 2.2. Based on the crude and adjusted Odds Ratio, was the association of Question 6.4 confounded?

In [10]:
```
sixptfour = ContTable(50, 238, 21, 524, 0)
sixptfour.display()
_ = sixptfour.odds_ratio(verbose=1, CI=0)
```

```
2x2 Contingency Table:
```

	E	~E
D	50	238
~D	21	524

```
Odds Ratio (OR): 5.2420968387354945
```

In [11]:
```
low_income_disease = ContTable(
    poor_exposure.a+poor_exposure.b,
    wealthy_exposure.a+wealthy_exposure.b,
    poor_exposure.c+poor_exposure.d,
    wealthy_exposure.c+wealthy_exposure.d,
    0
)
_ = low_income_disease.odds_ratio()
```

```
Odds Ratio (OR): 2.4236000706588943
=> 95% CI using logOR: 0.885254
=> => logOR Var: 0.022986582020195465
=> => Upper Bound: 3.2622286786261805
=> => Lower Bound: 1.8005596422417085
```

In [12]:
```
low_income_exposure = ContTable(
    poor_exposure.a+poor_exposure.c,
    wealthy_exposure.a+wealthy_exposure.c,
    poor_exposure.b+poor_exposure.d,
    wealthy_exposure.b+wealthy_exposure.d,
    0
)
_ = low_income_exposure.odds_ratio()
```

```
Odds Ratio (OR): 5.085243974132863
=> 95% CI using logOR: 1.626343
=> => logOR Var: 0.07366105623009926
=> => Upper Bound: 8.656308944249604
=> => Lower Bound: 2.9873825487286045
```

Given the above calculations it seems quite clear that income/wealth is an important confounding factor mediating the occurence of tuberculosis in populations exposed to biofuel emissions. This claim is supported by the reasoning that income is shown to mediate both biofuel polution exposure as well as tuberculosis incidence (as shown by low_income_disease/low_income_exposure).

Based upon the crude OR and the OR adjusted for all other factors (5.242 - 2.2), it appears that the association found in 6.4 was also highly confounded, though further investigation of these individual confounders is neccessary to achieve more detailed understanding of the interactions.

Question 9.2

Refer again to the data set esoph. The data, associating the binary measure of alcohol consumption with esophageal cancer incidence, can be stratified into two age groups, 25 to 54 years old and 55 years old and above. Use the Cochran-Mantel- Haenszel method to examine the association between alcohol consumption and incidence of esophageal cancer, adjusting for this dichotomous measure of age.

Give a summary estimate of the Odds Ratio using both the Woolf and Mantel-Haenszel methods. Provide two 95% confidence intervals based on your summary estimates. Compare these confidence intervals and comment.

Part 1

We construct the relevant contingency tables and perform a summary analysis.

In [13]:
```
agebinary = esoph["agegp"].isin(["55-64", "65-74", "75+"])
#old fok
exposed = esoph.loc[np.logical_and(alcbinary, agebinary), ["ncases",
"ncontrols"]].sum()
unexposed = esoph.loc[np.logical_and(~alcbinary, agebinary), ["ncase
s", "ncontrols"]].sum()
esoph_old = ContTable(exposed["ncases"], unexposed["ncases"],
exposed["ncontrols"]-exposed["ncases"], unexposed["ncontrols"]-unexpo
sed["ncases"], 0)
#young folk
exposed = esoph.loc[np.logical_and(alcbinary, ~agebinary), ["ncases",
 "ncontrols"]].sum()
unexposed = esoph.loc[np.logical_and(~alcbinary, ~agebinary), ["ncase
s", "ncontrols"]].sum()
esoph_young = ContTable(exposed["ncases"], unexposed["ncases"], expos
ed["ncontrols"]-exposed["ncases"], unexposed["ncontrols"]-
unexposed["ncases"], 0)
print("Elderly Patients:")
esoph_old.display()
esoph_old.odds_ratio(verbose=2)
esoph_old.chisq_indep(verbose=2)
print("Young Patients:")
esoph_young.display()
esoph_young.odds_ratio(verbose=2)
esoph_young.chisq_indep(verbose=2)
```

Elderly Patients:
2x2 Contingency Table:

	E	~E
D	66	78
~D	45	258

Odds Ratio (OR): 4.85128205128
=> 95% CI using logOR: 1.579243
=> => logOR Var: 0.0540702191865
=> => Upper Bound: 7.652202484565185
=> => Lower Bound: 3.075576945142571
=> Odds Ratio (OR) indicates a positive relationship between D & E.

ChiSquared Independence Test (n=447)
=> test statistic = 50.1955284271
=> p-value ≈ 1.39166456137e-12
=> ChiSquared Test suggests an association between D & E.

Young Patients:
2x2 Contingency Table:

	E	~E
D	30	26
~D	64	408

Odds Ratio (OR): 7.35576923077
=> 95% CI using logOR: 1.995485
=> => logOR Var: 0.089870852187
=> => Upper Bound: 13.23747872156092
=> => Lower Bound: 4.087435539231687
=> Odds Ratio (OR) indicates a positive relationship between D & E.

ChiSquared Independence Test (n=528)
=> test statistic = 54.768292101
=> p-value ≈ 1.35558231307e-13
=> ChiSquared Test suggests an association between D & E.

Out[13]: 1.3555823130673161e-13

In [14]:
```
esoph_strata = Strata([esoph_young, esoph_old])
_ = esoph_strata.cmh_test(verbose=2)
```

Cochran-Mantel-Haenszel Test:
=> test statistic = 98.7203476955
=> p-value ≈ 0.0
=> Cochran-Mantel-Haenszel Test suggests an association between D & E
.

Part 2

We calculate adjusted ORs using two methods and compare the results with the baseline results.

In [15]:
```
_ = esoph_strata.woolf_correction()
```

```
Woolf adj. OR: 5.676410286915972
=> 95% CI using logOR: 1.736319
=> => logOR Var: 0.033387404575
=> => Upper Bound: 8.120956836093036
=> => Lower Bound: 3.9677139524001195
```

In [16]:
```
_ = esoph_strata.mantel_haenszel_correction()
```

```
Mantel-Haenszel adj. OR: 5.56856910509
=> 95% CI using logOR: 1.717138
=> => logOR Var: 0.034240697901
=> => Upper Bound: 8.002985105769879
=> => Lower Bound: 3.874674445639857
```

In [17]:
```
print("Pooled Odds Ratio:", esoph_strata.pool().odds_ratio(verbose=0)
["OR"])
```

```
Pooled Odds Ratio: 5.64008468596
```

Analysis of the stratified measures using both the Woolf and Mantel-Haenszel yields a value near to our baseline/pooled OR, with a wide 95% confidence interval which also encompases this value. This suggests that there is unlikely to be a significant effect when stratifying based upon the confounding variable of age.