Homework 2 Solutions,

1. The file `crash.dat` was obtained from a national data base of automobile crashes. Data were selected from serious accidents in New Jersey in 1999. The data file has one line for every individual involved in the crashes. Individuals in a given car are numbered, and given in the variable `pn`. Since individuals in the same car are likely to have related outcomes, in order to obtain a set of independent observations, the code below retains only the subject labeled 1. Autos in a crash are numbered, and the number is contained in the variable `vh`. Again, in order to obtain a set of independent observations, retain only the vehicles labeled 1. The following code reads in crash data, and creates categorical variables reflecting car body type, restraint use, and whether the crash was a fatality. Explore the relationship between proper restraint use and surviving the accident.

```plaintext
/**************************************************************/
/* Data obtained from http://www-fars.nhtsa.dot.gov by re- */
/* questing violations, accident severity, body type, speed */
/* limit, and restraint use for NJ, 1999. Variables are id, */
/* id=line in the data set */
/* state=34 for New Jersey */
/* case=a number of the accident */
/* vh=number of the vehicle within accident */
/* pn=number of passenger within vehicle */
/* sl=speed limit where the accident occurred */
/* bt=body type of vehicle */
/* v1-v3=codes for violations cited. */
/* rs=whether proper restraints (ex. seat belts) were used */
/* is=injury severity */
/* Each accident potentially involves multiple vehicles, and */
/* each vehicle potentially involves multiple passengers. */
/* Select one vehicle (vh=1) from each crash. Select one */
/* passenger (pn=1) from this vehicle. */
/**************************************************************/

data crash ;
  input id state case vh pn sl bt v1 v2 v3 rs is;
  if vh=1 then ;
  if pn>1 then delete;
  fatal="Not fatal";
  if is=4 then fatal="Fatal";
  vt="Car";
  if bt>10 then vt="Light truck";
  if bt>20 then vt="Van";
  if bt>30 then vt="Truck";
  beltuse="Proper restraint use";
  if rs>12 then beltuse="Improper or no restraint use";
  if rs=0 then beltuse="Improper or no restraint use";
  if rs=. then beltuse="";
  keep vt fatal beltuse;
run;

This code and the crash data set can be found on the course web page.
```
a. Test whether the tables stratified on body type have a common odds ratio not equal to 1. 

```
Running the command

proc freq data=crash; table vt*fatal*beltuse/cmh norow nopct notot ; run;
```

gives a test statistics of 32.79707, and a p-value of < 0.0001. Reject the null hypothesis of no association. These calculations might have been done in R using

```
crash<-read.table("crash.dat",na.strings=".")
names(crash)<-c("id","state","case","vh","pn","sl","bt",
              "v1","v2","v3","rs","is")
crash<-crash[crash$pn==1,]
crash<-crash[crash$vh==1,c("sl","bt","v1","v2","v3","rs","is")]
crash$vt<-(crash$bt>10)+(crash$bt>20)+(crash$bt>30)
crash$notfatal<-(crash$bt>10)+(crash$bt>20)+(crash$bt>30)
crash$beltuse<-(crash$rs>0)&(crash$rs<13)
mantelhaen.test(crash$notfatal,crash$beltuse,crash$vt)
```

b. Estimate a common odds ratio stratified based on car body type. Give a confidence interval for this value.

The SAS code in the part above gives

```
Estimates of the Common Relative Risk (Row1/Row2)
Type of Study Method Value 95% Confidence Limits
--------------------------------------------------------------
Case-Control Mantel-Haenszel 2.5617 1.8673 3.5143
(Odds Ratio) Logit ** 2.5050 1.7987 3.4888
```

c. What assumption underlies parts (a) and (b)?
Both (a) and (b) assume that the tables have a common odds ratio. If for example, vans had no association between belt use and fatalities, but trucks had association, then the assumption would be violated.

d. The variables vh and pn represent the number of the person and vehicle in the accident. Why did I remove those whose value for either is greater than one? I am looking for a statistical, rather than computational, issue here.
Distributional assumptions about the fitted odds ratio rely on independence between individuals in the data set. Individuals involved in the same accident are likely to share conditions linked to fatality. For example, if the driver of one car is killed, it is likely that the accident was physically more violent, and others involved are more likely to be killed as well. I have selected individuals so only one person per accident contributes to the odds ratio.

Some students noted that position in the car might be a confounder, and removing everyone from the study except the driver gets rid of this confounder. This is right; I hadn’t thought of it.

2. In each of six labs, twenty chicks were randomly divided into a treatment group and a control group. The treatment (T) group were exposed to pulsed electro-magnetic radiation, and the control (C) chicks were placed in the presence of a similar apparatus
which was not turned on. The chicks were examined for deformities, and the results were tabulated below:

<table>
<thead>
<tr>
<th>Lab</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Healthy Chicks

<table>
<thead>
<tr>
<th>Lab</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Unhealthy Chicks

These data may be found at http://www.stat.rutgers.edu/~kolassa/960-553/hen.dat. Note that most of the labs have fewer than 20 chicks classified, because some of the chicks were lost to causes unrelated to the presence or absence of radiation during the experiment. This data is a subset of that collected by Berman, et. al. (1990). Please use a method covered in class so far.

a. Ignoring the fact that these data were collected in different labs, test the null hypothesis that chick deformities are unrelated to radiation.

```Use
data hen; infile 'hen.dat'; input lab sick treat count; run;
proc freq data=hen;
  table sick*treat/chisq relrisk norow nocol nopercent;
  exact or;
  weight count; run;

or
hen<-read.table("hen.dat")
names(hen)<-c("lab","sick","treat","count")
chisq.test(xtabs(count~sick+treat,data=hen))
```

to obtain

<table>
<thead>
<tr>
<th></th>
<th>sick</th>
<th>treat</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>46</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>57</td>
<td>58</td>
<td>115</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>0.0348</td>
<td>0.8520</td>
</tr>
</tbody>
</table>

There is no evidence at all that radiation influences deformities.

b. Again ignoring the fact that these data were collected in different labs, calculate a 95% confidence interval for the odds ratio associating deformities with radiation.
The estimate is \((12 \times 46)/(11 \times 46) = 12/11 = 1.090909\), and its log is \(\log(1.090909) = 0.0870\).

The standard error of \(\log\) odds ratio is
\[
\sqrt{\frac{1}{46} + \frac{1}{46} + \frac{1}{11} + \frac{1}{12} = 0.4666}.
\]

The 95\% confidence interval for \(\text{the odds ratio}\) is
\[
\exp(0.0870 \pm 1.96 \times 0.4666) = (0.437, 2.722).
\]

This corresponds to the confidence intervals given by SAS. In R,

\[
\text{fisher.test(xtabs(count~sick+treat,data=hen))}
\]

c. Test the null hypothesis that chick deformities are unrelated to radiation, accounting for the fact that the data were collected from different labs.

Use the commands

\[
\text{proc freq data=hen; table sick*treat/cmh relrisk; weight count; run;}
\]
or

\[
\text{mantelhaen.test(xtabs(count~sick+treat+lab,data=hen))}
\]

to find

Summary Statistics for sick by treat
Controlling for lab
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Nonzero Correlation</td>
<td>1</td>
<td>0.0080</td>
<td>0.9288</td>
<td></td>
</tr>
</tbody>
</table>

\(\text{Hence we find no impact of treatment on deformities.}\)

d. Allowing for the fact that these data were collected in different labs, calculate a 95\% confidence interval for the odds ratio associating deformities with radiation. Compare the width of the interval with that obtained in part (b).

The contingency table confidence intervals computed from the previous input are

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Method</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control</td>
<td>Mantel-Haenszel</td>
<td>0.4063 2.6885</td>
</tr>
<tr>
<td>(Odds Ratio)</td>
<td>Logit</td>
<td>0.4034 2.7324</td>
</tr>
</tbody>
</table>

\(\text{respectively. Again, there is little difference in widths.}\)

e. Estimate odds ratios for each table separately. Describe what you see.

\(\text{Odds ratios for the individual labs are}\)

\[
1.11111 1.75000 1.00000 1.00000 0.50000 1.00000
\]

\(\text{You could have done this using the commands}\)
proc sort data=hen; by lab; run;
proc freq data=hen noprint; output out=outset relrisk;
   by lab; table treat*sick/relrisk; weight count; run;
proc print data=outset noobs; var LAB _RROR_; run;

Hence most of the labs show no effect; lab 1 shows a small effects indicating treatment damages chicks, lab 2 shows a larger effect in the same direction, and lab 5 shows a larger protective effect of treatment. None of these are significant.