Lab Six: Categorical data analysis

Lab Objectives
After today’s lab you should be able to:

1. Use PROC FREQ to produce contingency tables, chi-square test, Fischer’s exact test, and measures of association (odds ratio, risk ratio).
2. Interpret output from PROC FREQ.
3. Calculate a chi-square statistic, a risk ratio, and an odds ratio.
4. Use PROC SORT to sort data for use in a sub-setting “by” statement.
5. Use a “where” statement to restrict analyses, and “by” statement to stratify analyses.
6. Enter, format, and analyze a 2x2 table using a weight variable to specify cell counts.
7. Use PROC RANK to create categories based on quartiles, tertiles, etc.
8. Understand the function of a SAS MACRO.
9. Write and run a simple SAS MACRO.

Recommended reading in Walker: Chapters 16-17
LAB EXERCISE STEPS:

Follow along with the computer in front…

1. Double-click on to open the SAS editor file “data creation code” which should be saved in your stats210 folder from last week; run the libname statement:

   libname stats210 ‘C:\Documents and Settings\mitl-pc.LANE-LIB\My Documents\Stats210’;

2. Using the Explorer Browser on the left hand side of your screen, double check that a stats210 library has been properly created, and that it contains the SAS dataset “runners”.

3. Recall that these results are from a randomized trial. The variable “treatr” is 1 for those assigned to treatment and 0 for those assigned to control. In a randomized trial with a relatively small sample size, it is important to check that there aren’t any large imbalances in baseline characteristics between the treatment and control groups. For example, test whether or not the two groups differ in age using the following code:

   /**Do treatment and control differ in age?***/
   proc ttest data=stats210.runners;
     class treatr;
     var age;
   run;

4. What if you want to test the difference between treatment and control with regards to a categorical variable? Here, we could use a chi-square test or Fischer’s exact test. For example, women with prior stress fractures may be more likely to fracture during the study period. Therefore, we’d like to make sure that prior stress fracture (stressf= 1 or 0) is balanced between treatment and control:

   /**Was randomization roughly equal among those prone to stress fractures and those not*/
   proc freq data=stats210.runners;
     tables treatr*stressf/chisq nocol nopercent norow expected;
   run;
**SAS OUTPUT:**
The FREQ Procedure

Table of treatr by stressf

<table>
<thead>
<tr>
<th>treatr (TREATR)</th>
<th>stressf (STRESSF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Expected</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>0</td>
<td>11 12.162</td>
</tr>
<tr>
<td></td>
<td>7 5.8378</td>
</tr>
<tr>
<td>1</td>
<td>14 12.838</td>
</tr>
<tr>
<td></td>
<td>5 6.1622</td>
</tr>
<tr>
<td>Total</td>
<td>25 12</td>
</tr>
</tbody>
</table>

Frequencies Missing = 10

Statistics for Table of treatr by stressf

<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>0.6668</td>
<td>0.4142</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>0.6687</td>
<td>0.4135</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>0.2165</td>
<td>0.6417</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>0.6488</td>
<td>0.4206</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>-0.1342</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.1331</td>
<td></td>
</tr>
<tr>
<td>Cramer's V</td>
<td></td>
<td>-0.1342</td>
<td></td>
</tr>
</tbody>
</table>

Fisher's Exact Test

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell (1,1) Frequency (F)</td>
<td>11</td>
<td>0.3211</td>
</tr>
<tr>
<td>Left-sided Pr &lt;= F</td>
<td>0.8786</td>
<td></td>
</tr>
<tr>
<td>Right-sided Pr &gt;= F</td>
<td>0.8786</td>
<td></td>
</tr>
<tr>
<td>Table Probability (P)</td>
<td>0.1998</td>
<td></td>
</tr>
<tr>
<td>Two-sided Pr &lt;= P</td>
<td>0.4951</td>
<td></td>
</tr>
</tbody>
</table>

Effective Sample Size = 37
Frequency Missing = 10

WARNING: 21% of the data are missing.
5. A more interesting question is whether being randomized to the treatment increased or decreased subsequent risk of stress fracture (a major study outcome)? Variable “SF1” is 1 if they fractured, 0 if they did not. Use the following code to produce risk ratios and odds ratios for the increase or decrease in risk of fracture with randomization to treatment:

```sas
/**Did being randomized to treatment increase or decrease risk of subsequent stress fracture**/
proc freq data=stats210.runners;
tables treatr*sf1 / measures nocol nopercent norow;
run;
```

Asks for measures of association.

GIVES:

<table>
<thead>
<tr>
<th>treatr(TREATR)</th>
<th>sf1(SF1)</th>
<th>Frequency</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>22</td>
<td>5</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>4</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>9</td>
<td>47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. 2x2 tables come up “rotated” in SAS. Recode as follows such that “exposure”—here treatment—and “disease”—here stress fracture—are in cell 1,1:

```sas
data temp;
  set stats210.runners;
  treatr=1-treatr;
  sf1=1-sf1;
run;

proc format;
  value treat
    1=control;
run;

proc format;
  value sf
    1=no fract;
run;

proc freq data=temp;
  format sf1 sf .;
  format treatr treat .;
  tables treatr*sf1 sf1*treatr/ measures nocol nopercent norow;
run;
```

Table of treatr by sf1

<table>
<thead>
<tr>
<th>treatr (TREATR)</th>
<th>sf1 (SF1)</th>
<th>Frequency</th>
<th>fracture</th>
<th>no fract</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>treated</td>
<td></td>
<td>4</td>
<td>16</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>control</td>
<td></td>
<td>5</td>
<td>22</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>9</td>
<td>38</td>
<td></td>
<td>47</td>
</tr>
</tbody>
</table>

Measures of association:

Estimates of the Relative Risk (Row1/Row2)

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Value</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control (Odds Ratio)</td>
<td>1.1000</td>
<td>0.2544 - 4.7556</td>
</tr>
<tr>
<td>Cohort (Col1 Risk)</td>
<td>1.0800</td>
<td>0.3316 - 3.5176</td>
</tr>
<tr>
<td>Cohort (Col2 Risk)</td>
<td>0.9818</td>
<td>0.7395 - 1.3036</td>
</tr>
</tbody>
</table>

Because this is a longitudinal study, we can estimate the *risk* of getting a stress fracture in each group, and thus can calculate a *risk ratio*. If it had been a case-control study (where would have selected stress fracture cases and a fracture-free control group) we would not have been able to estimate the risk of disease, and we would have had to use the *odds ratio*.

The ratio of risk for the treatment group (column 1) vs. controls is 1.08, or 8% higher risk in the treatment group than the controls. This does not account for the fact that many women switched groups (“intention to treat analysis”).

7. Repeat with a “where” statement restrict the analysis to those who stayed on protocol (stayed in treatment or in control). Switching groups can sometimes bias the risk ratio towards the null, but you’ll see there’s not much change here when we restrict to non-switchers.

```sas
proc freq data=temp;
   format sf1 sf.;
   format treatr treat.;
tables treatr*sfl / measures nocol nopercent norow;
   where onprotocol=1;
run;
```

Restricts analysis to those who remained in the correct group.
8. You might suspect that the effect of treatment might differ by menstrual group (amenorrheic might benefit more from the treatment, for example). So, you can stratify on menstrual status, and examine the effect of treatment separately within each menstrual group as follows:

```
proc sort data=temp;
  by mencat;
run;

proc freq data=temp;
  by mencat;
  format sf1 sf.;
  format treatr treat.;
  tables treatr*sf1 / measures nocol nopercent norow;
run;
```

Use Explorer Browser to navigate between the strata.

9. To change a continuous variable into a categorical variable for the purposes of categorical analysis, you can use PROC RANK to define quartiles, quintiles, top half, bottom half, etc. Here, we will test the hypothesis that there is an association between being in the top quartile of EDI scores and having menstrual irregularity using a chi-square test.

```
proc rank data=temp out=temp groups=4;  *ranks them 0 to 3;  
  Rank 1-47; then split into 4 equal groups. Assign the lowest rank values “0”, next quartile “1”, top quartile “3”.
  ranks rankedi;
  var sumedi1;
  run;

data temp;
  set temp;
  if rankedi=3 then badedi=1;
  else badedi=0;
run;

proc freq data=temp;
  tables badedi*isIrreg / chisq measures norow nocol nopercent;
run;
```
Table of badei by isirreg

<table>
<thead>
<tr>
<th>badei</th>
<th>isirreg</th>
<th>Frequency</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>27</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>17</td>
<td>47</td>
</tr>
</tbody>
</table>

The odds ratio is the cross product (so table organization doesn’t matter):

\[
OR = \frac{27 \times 9}{8 \times 3} = 10.125
\]

We can calculate either an odds ratio or a risk ratio here since the data are cross-sectional.

Estimates of the Relative Risk (Row1/Row2)

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Value</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control (Odds Ratio)</td>
<td>10.125</td>
<td>2.2004</td>
</tr>
<tr>
<td>Cohort (Col1 Risk)</td>
<td>3.0857</td>
<td>1.1392</td>
</tr>
<tr>
<td>Cohort (Col2 Risk)</td>
<td>0.3048</td>
<td>0.1527</td>
</tr>
</tbody>
</table>

Confidence limits show highly significant increase in risk of menstrual irregularity with high EDI score. Confidence limits are calculated as:

\[
\ln(10.125) \pm 1.96 \sqrt{\frac{1}{3} + \frac{1}{8} + \frac{1}{9} + \frac{1}{27}} = (.78, 3.84)
\]

\[
e^{-78} = 2.2
\]

\[
e^{3.84} = 46.5
\]

A program to do this calculation in SAS:

```sas
data _null_; x=log(10.125); y=sqrt(1/3+1/27+1/8+1/9); low=exp(x+probit(.95)*y); high=exp(x-probit(.95)*y); put low; put high; run;
```
10. Generalize the above program into a SAS MACRO, such that you can enter any 2x2 table and get out any size confidence interval (90%, 95%, 99%, etc.)

```sas
/**MACRO to calculate XX% confidence limits for an odds ratio from the desired confidence and the 2x2 cell sizes: a,b,c,d**/

number, e.g. "95";
    data _null_
    OR=&a.*&d. /(&b. *&c.);
    lnOR=log(OR);
    error=sqrt(1/&a.+1/&b.+1/&c.+1/&d.);
    Z=-probit((1-&confidence./100)/2); *gives left hand Z score, multiply by negative;
    lower=exp(lnOR-Z*error);
    upper=exp(lnOR+Z*error);
    put lower;
    put upper;
    run;
%mend oddsratio;

/**Invoke MACRO with the above data and ask for 95% confidence limit**/
%oddsratio(95, 9, 3, 8, 27);
```

DUE FROM LAB SIX: Use the above MACRO to get a 99% confidence interval from a 2x2 table with a=20, b=20, c=10, d=20. Email me the resulting values.