Outline

1. Some contingency tables.

<table>
<thead>
<tr>
<th>Region</th>
<th>Absolutely</th>
<th>Probably</th>
<th>Probably not</th>
<th>Absolutely not</th>
</tr>
</thead>
<tbody>
<tr>
<td>East</td>
<td>220</td>
<td>53</td>
<td>32</td>
<td>12</td>
</tr>
<tr>
<td>Midwest</td>
<td>409</td>
<td>55</td>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>South</td>
<td>373</td>
<td>45</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>West</td>
<td>266</td>
<td>78</td>
<td>43</td>
<td>29</td>
</tr>
</tbody>
</table>

   (b) The National Survey of Student Engagement. Random samples of first-year students and seniors. The question is: this institution has challenged me to critically evaluate and reconsider values that I have always held.

<table>
<thead>
<tr>
<th>Class</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>58</td>
<td>121</td>
<td>221</td>
</tr>
<tr>
<td>SR</td>
<td>31</td>
<td>58</td>
<td>202</td>
</tr>
</tbody>
</table>

   (c) Testing whether a screening test for depression helps primary care physicians to diagnose depression. Will doctors recognize depression status within twelve months of first visit?

<table>
<thead>
<tr>
<th>Group</th>
<th>Recognized</th>
<th>Not recognized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed, Doctor told</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>Depressed, Doctor not told</td>
<td>11</td>
<td>41</td>
</tr>
<tr>
<td>Not depressed</td>
<td>4</td>
<td>56</td>
</tr>
</tbody>
</table>

2. Three different data collection paradigms and corresponding research questions.
   (a) \( I \) independent populations (homogeneity)
   (b) cross-classifying a sample from one population (independence)
   (c) \( I \) experimental treatments (homogeneity)

3. Notation:
   \( I \) the number of rows
   \( J \) the number of columns
   \( n_{ij} \) the integer entry in the \( i^{th} \) row and \( j^{th} \) column
   \( n_i \) the sum of the entries in the \( i^{th} \) row
   \( n_j \) the sum of the entries in the \( j^{th} \) column
   \( n = n_{..} \) the total of all entries in the table

4. Homogeneity.
   (a) parameters \( \pi_{i,j} \)
   (b) null hypothesis: \( H_0 \): for every \( j \), \( \pi_{1,j} = \pi_{2,j} = \cdots = \pi_{I,j} \)
   (c) “expected counts” under null hypothesis: \( \hat{n}_{i,j} = \frac{n_i n_j}{n} \)
   (d) Test statistic: \( X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \)
   (e) Distribution of test statistic under null hypothesis: chi-square with \((I-1)(J-1)\) degrees of freedom.
Useful R

```r
> m=matrix(c(20,28,11,41), nrow=2,ncol=2, byrow=T)
> m
   [,1] [,2]
[1,]  20  28
[2,]  11  41
> rownames(m)=c('Told','Not Told')
> colnames(m)=c('Diagnosed','Not Diagnosed')
> m
   Diagnosed Not Diagnosed
  Told      20       28
 Not Told  11       41
> chisq.test(m)

  Pearson's Chi-squared test with Yates' continuity correction

  data:  m
  X-squared = 3.9979, df = 1, p-value = 0.04556
> chisq.test(m,correct=F)

  Pearson's Chi-squared test

  data:  m
  X-squared = 4.91, df = 1, p-value = 0.0267
```

Homework - due Thursday, April 24, 2008

1. Read Section 7.1.

2. Do problems 7.1,2.