Contingency tables, test of independence.

<table>
<thead>
<tr>
<th>Feature 1 = 1</th>
<th>Feature 2 = 1</th>
<th>F2 = 2</th>
<th>F2 = 3</th>
<th>...</th>
<th>F2 = b</th>
<th>row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 = 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1 = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1 = a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>col. total</td>
<td>N_{a1}</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>N_{ab}</td>
</tr>
<tr>
<td></td>
<td>N_{+a}</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>N_{+b}</td>
<td>n</td>
</tr>
</tbody>
</table>

\( X^1_i \in \{1, \ldots, a\} \)
\( X^2_i \in \{1, \ldots, b\} \)

Random Sample:
\( X_1 = (X^1_1, X^2_1), \ldots, X_n = (X^1_n, X^2_n) \)

Question: Are \( X^1, X^2 \) independent?
Example: When asked if your finances are better, worse, or the same as last year, see if the answer depends on income range:

<table>
<thead>
<tr>
<th></th>
<th>Worse</th>
<th>Same</th>
<th>Better</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq 20K )</td>
<td>20</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>20K - 30K</td>
<td>24</td>
<td>27</td>
<td>32</td>
</tr>
<tr>
<td>( \geq 30K )</td>
<td>14</td>
<td>22</td>
<td>23</td>
</tr>
</tbody>
</table>

Check if the differences and subtle trend are significant or random.

\( \theta_{ij} = P(i, j) = P(i) \times P(j) \) if independent, for all \( ij \)

Independence hypothesis can be written as:
\( H_1: \theta_{ij} = p_i q_j \) where \( p_1 + \ldots + p_a = 1, q_1 + \ldots + q_b = 1 \)
\( H_2: \) otherwise.
\( r = \) number of categories = \( ab \)
\( s = \) dimension of parameter set = \( a + b - 2 \)
The MLE \( p_i^*, q_j^* \) needs to be found →

\[
T = \sum_{i,j} \frac{(N_{ij} - np_i^* q_j^*)^2}{np_i^* q_j^*} \sim \chi^2_{r-s-1 = ab-(a+b-2)-1 = (a-1)(b-1)}
\]

Distribution has \((a - 1)(b - 1)\) degrees of freedom.

Likelihood:

\[
\psi(\overline{p}, \overline{q}) = \prod_{i,j} (p_i q_j)^{N_{ij}} = \prod_i p_i^{N_{i+}} \times \prod_j q_j^{N_{+j}}
\]

Note: \( N_{i+} = \sum_j N_{ij} \) and \( N_{+j} = \sum_i N_{ij} \)
Maximize each factor to maximize the product.
$\sum_i N_i \log p_i \rightarrow \max, p_1 + ... + p_a = 1$

Use Lagrange multipliers to solve the constrained maximization:
$\sum_i N_i \log p_i - \lambda(\sum_i p_i - 1) \rightarrow \max, \min \lambda$

$$\frac{\partial}{\partial p_i} = \frac{N_i}{p_i} - \lambda = 0 \rightarrow p_i = \frac{N_i}{\lambda}$$

$$\sum_i p_i = \frac{n}{\lambda} = 1 \rightarrow \lambda = n \rightarrow p_i^* = \frac{N_i}{n}$$

$$p_i^* = \frac{N_i}{n}, q_j^* = \frac{N_j}{n}$$

$$T = \sum_{i,j} \frac{(N_{ij} - N_i N_j/n)^2}{N_i N_j/n} \sim \chi^2_{(a-1)(b-1)}$$

Decision Rule:
$\delta = \{ H_1 : T \leq c; H_2 : T > c \}$

Choose $c$ from the chi-square distribution, $(a - 1)(b - 1)$ d.o.f., at a level of significance $\alpha = \text{area}$. 

From the above example:
$N_1 = 47, N_2 = 83, N_3 = 59$
$N_1 = 58, N_2 = 64, N_3 = 67$

$n = 189$

For each cell, the component of the $T$ statistic adds as follows:

$$T = \frac{(20 - 58(47)/189)^2}{58(47)/189} + ... = 5.210$$

Is $T$ too large?

$T \sim \chi^2_{(3-1)(3-1)} = \chi^2_4$

For this distribution, $c = 9.488$

According to the decision rule, accept $H_1$, because $5.210 \leq 9.488$

**Test of Homogeneity** - very similar to independence test.
1. Sample from entire population.
2. Sample from each group separately, independently between the groups.

Question: $P(\text{category } j \mid \text{group } i) = P(\text{category } j)$
This is the same as independence testing!

$P(\text{category } j, \text{group } i) = P(\text{category } j)P(\text{group } i)$

$$
\rightarrow P(C_j|G_i) = \frac{P(C_jG_i)}{P(G_i)} = \frac{P(C_j)P(G_i)}{P(G_i)} = P(C_j)
$$

Consider a situation where group 1 is 99% of the population, and group 2 is 1%. You would be better off sampling separately and independently. Say you sample 100 of each, just need to renormalize within the population. The test now becomes a test of independence.

Example: pg. 560
100 people were asked if service by a fire station was satisfactory or not. Then, after a fire occurred, the people were asked again. See if the opinion changed in the same people.

<table>
<thead>
<tr>
<th>Before Fire</th>
<th>After Fire</th>
</tr>
</thead>
<tbody>
<tr>
<td>satisfied</td>
<td>unsatisfied</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>72</td>
<td>28</td>
</tr>
</tbody>
</table>

But, you can’t use this if you are asking the same people! Not independent! Better way to arrange:

| Originally Satisfied | 70 | 10 |
| Originally Unsatisfied | 2  | 18 |
| After, Satisfied     |    |    |
| After, Not Satisfied |    |    |

If taken from the entire population, this is ok. Otherwise you are taking from a dependent population.

** End of Lecture 34