Categorical Analysis

STAT120C
Review of Tests Learned in STAT120C

• Which test(s) should be used to answer the following questions?
  – Is husband’s BMI larger than wife’s?
  – Is men’s BMI different from women’s?
  – Do gender and smoking affect human’s weight?
  – Is kid’s weight linearly related to mother’s weight?
  – Is kid’s weight linearly related to parents’?
  – Is smoking associated with gender?
Review of Tests Learned in STAT120C

- Is husband’s BMI larger than wife’s? (paired t-test or one-sample t-test on differences)
- Is men’s BMI different from women’s? (two-sample t-test or ANOVA)
- Do gender and smoking affect human’s weight? (two-way ANOVA)
- Is kid’s weight linearly related to mother’s weight? (simple linear regression)
- Is kid’s weight linearly related to parents’? (multiple linear regression)
- Is smoking associated with gender? (chi-squared test or Fisher’s exact test)
Categorical Variables

• Both Smoking and Gender are categorical variables

• Other examples of categorical variables
  – Blood type: A, B, AB, O
  – Patient condition: good, fair, serious, critical
  – Socioeconomic class: upper, middle, low

• Nominal variables are categorical variables without a natural order. E.g., Blood type

• Ordinal variables are categorical variables with a natural order. E.g, socioeconomic class, patient condition
Some Distributions from STAT120A
Suppose that each of $n$ independent, identical trials can have outcome in any of $c$ categories. Let $N_i$ denote the number of trials having outcome in category $i$. The counts $(N_1, N_2, \ldots, N_c)$ have the multinomial distribution, i.e.,

- $(N_1,\ldots,N_c) \sim \text{Multinomial}(n, (\pi_1, \ldots, \pi_c))$
- Probability mass function

\[
p(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \left( \frac{n!}{n_1!n_2!\cdots n_c!} \right) \prod_{i=1}^{c} \pi_i^{n_i}
\]

where

\[
\sum_{i=1}^{c} N_i = n
\]

\[
\sum_{i=1}^{c} \pi_i = 1
\]
Multinomial Distribution

• Some properties

\[
E(N_i) = n\pi_i \\
Var(N_i) = n\pi_i(1 - \pi_i) \\
cov(N_i, N_j) = -n\pi_i\pi_j, i \neq j
\]

Marginal distribution: \( N_i \sim \text{Binomial}(n, \pi_i) \).
Conditional distribution:

\[
(N_1, \cdots, N_{c-1})|N_c = n_c \sim \text{Multinomial}(n - n_c, \frac{\pi_1}{1 - \pi_c}, \cdots, \frac{\pi_{c-1}}{1 - \pi_c})
\]

• E.g., the counts of different blood types among 100 students. \((\pi_A, \pi_B, \pi_{AB}, \pi_O) = (0.42, 0.10, 0.04, 0.44)\)
• \((N_A, N_B, N_{AB}, N_O) = \text{Multinomial}(100, (0.42, 0.10, 0.04, 0.44))\)
Binomial distribution

• When \( c=2 \), multinomial distribution is also called binomial distribution

• E.g, the number of persons with O blood type in 100 students
  \[(N_0, N_Q) \sim \text{Binomial}(100, (0.44,0.56))\]

• Because of the constraint \((N_0+ N_Q=100)\), we don’t need to write both out. Therefore, for simplicity, we just say
  \[N_0 \sim \text{Binomial}(100, 0.44)\]
Hypergeometric Distribution

$r$ red balls, $n-r$ white balls

Random take $m$ balls out without replacement

Let $X$ denote the number of red balls out of the $m$ balls
Hypergeometric Distribution

• $X$ is random variable
• We say $X$ follows a hypergeometric distribution
• The probability mass function of $X$ is

$$\Pr(X = k) = \binom{r}{k} \frac{\binom{n-r}{m-k}}{\binom{n}{m}}$$
Fisher’s Exact Test
Fisher’s Exact Test for 2x2 Tables

• Is smoking associated with gender?

• Collect data and summarize your data into a 2x2 table.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>F</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$N_{11}$</td>
<td>$N_{12}$</td>
<td>$n_1.$</td>
</tr>
<tr>
<td>No</td>
<td>$N_{21}$</td>
<td>$N_{22}$</td>
<td>$n_2.$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{1.}$</td>
<td>$n_{2.}$</td>
<td>$n_{..}$</td>
</tr>
</tbody>
</table>

• Pearson’s chi-squared test can be used to test whether there is an association between these two variables
Fisher’s Exact Test for 2x2 Tables

• Pearson’s Chi-squares test is an asymptotic test. It is not accurate when sample size is small
• Rule of thumb: when any of the expected counts is less than 5, Pearson’s Chi-squared test should be used with caution
• When sample size is small, we can consider an exact test, which is called Fisher’s exact test
## Assumptions of Fisher’s Exact Test

- Independent observations
- Fixed marginal counts

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>F</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$N_{11}$</td>
<td>$N_{12}$</td>
<td>$n_{1.}$</td>
</tr>
<tr>
<td>No</td>
<td>$N_{21}$</td>
<td>$N_{22}$</td>
<td>$n_{2.}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{.1}$</td>
<td>$n_{.2}$</td>
<td>$n_{..}$</td>
</tr>
</tbody>
</table>

- For each characteristic (categorical variable), each observation can be categorized as one of the two mutually exclusive types.
Hypothesis in Fisher’s Exact Test

• $H_0$: The proportion of smokers is the same between men and women
  – Alternatively, we can restate it to: smoking and gender is not associated

• Two-sided alternative
  – $H_1$: The proportion of smokers is different between men and women
    • Alternatively, we can restate it to: smoking and gender is associated

• One-sided alternative can also be used, when appropriate
Test Statistic

• There are four random variables in the 2x2 table. Because the marginal counts are fixed, if we know one of them, the remaining three are fixed

• For example, we can use $N_{11}$ as the test statistic

• To find the rejection region or calculate p-value, we need to know the distribution of $N_{11}$ when the null hypothesis is true
The Null Distribution of $N_{11}$

### The 2x2 table

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>F</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$N_{11}$</td>
<td>$N_{12}$</td>
<td>$n_1.$</td>
</tr>
<tr>
<td>No</td>
<td>$N_{21}$</td>
<td>$N_{22}$</td>
<td>$n_2.$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{.1}$</td>
<td>$n_{.2}$</td>
<td>$n_{..}$</td>
</tr>
</tbody>
</table>

$n_{1.}$ smokers, $n_{2.}$ non-smokers

$n_{11}$ smokers, $N_{21}$ non-smokers

If gender is not associated with smoking, the smaller box is a random sample from larger box.
The Null Distribution of $N_{11}$

- Therefore, $N_{11}$ follows a hypergeometric distribution and

\[
Pr(N_{11} = n_{11}) = \frac{\binom{n_1}{n_{11}} \binom{n_2}{n_{11} - n_{11}}}{\binom{n_{1.}}{n_{1.}}} = \frac{\binom{n_{1.}}{n_{11}} \binom{n_{2.}}{n_{21}}}{\binom{n_{1.}}{n_{1.}}}
\]
Observed data

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Column totals</td>
<td>6</td>
<td>8</td>
<td>14</td>
</tr>
</tbody>
</table>

The distribution of $N_{11}$ under Ho

\[
p\text{Value} = 0.027972028 + 0.059940060 + 0.002997003 = 0.09091
\]

We don’t have enough evidence to reject the null hypothesis. One possible reason could be the small sample we have – only 14 students in total!
Fisher’s Exact Test in R

• Step 1: prepare the table

• Step 2: use “fisher.test” in R

```r
> matrix(c(4,2,1,7),2,2)
     [,1] [,2]
[1,]   4   1
[2,]   2   7
> fisher.test(matrix(c(4,2,1,7),2,2))

Fisher's Exact Test for Count Data

data:  matrix(c(4, 2, 1, 7), 2, 2)
p-value = 0.09091
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.6418261 779.1595463
sample estimates:
  odds ratio
   10.98111
```
Limitations of Fisher’s Exact Test

• For large samples, the calculation required for null distribution is demanding
• When one factor has $I$ levels and another factor has $J$ levels, we need to deal with $I \times J$ tables. This is not straightforward
• When sample size is large enough, one can use Pearson’s chi-squared test. This test is an asymptotic test, which based upon asymptotic theories
Pearson’s Chi-squared Test
Asymptotic Tests for Contingency Tables

• The General Two-Way Contingency Table. Consider a two-way table with $I$ rows and $J$ columns:

\[
\begin{array}{cccc}
  n_{11} & n_{12} & \cdots & n_{1J} \\
  n_{21} & n_{22} & \cdots & n_{2J} \\
  \vdots & \vdots & \ddots & \vdots \\
  n_{I1} & n_{J2} & \cdots & n_{IJ} \\
\end{array}
\]

<table>
<thead>
<tr>
<th>$n_1.$</th>
<th>$n_2.$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1.$</td>
<td>$n_2.$</td>
</tr>
</tbody>
</table>

where

- $n_{ij}$ is the observed count in row $i$ and column $j$
- $n_{i.} = \sum_{j=1}^{J} n_{ij}$ is the total number of observations in row $i$.
- $n_{.j} = \sum_{i=1}^{I} n_{ij}$ is the total number of observations in column $j$.
- $n_{..} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ is the total number of observations in the table.
Pearson’s Chi-squared Test

- Assuming that we have two factors
- Pearson’s chi-squared test can be used to answer questions such as
  - Are the two factors independent?
  - Are subpopulations homogeneous?
- The choice of which question to address depends on study design.
  - If we have a random sample from a population, we can ask whether the two factors are independent
  - If we have a random sample from each subpopulation, we can ask whether the underlying subpopulations are homogeneous or not
Pearson’s Chi-squared Test

• Test statistic

\[ X^2 = \sum_{i=1}^{c} \frac{(Obs_i - Exp_i)^2}{Exp_i} \]

• \( c \) is the total number of cells. For the \( I \times J \) table, \( c = IJ \)

• \( Obs_i \) is the observed count for cell \( i \)

• \( Exp_i \) is the expected count for cell \( i \) under a specific null distribution and it can be calculated based on the MLE of parameters under a null distribution.
Theoretical Justification (not required)

- CLT says we can construct a random vector with a limiting distribution that is multivariate normal distribution
- One can then construct quadratic forms that follow chi-squared distributions
where $c$ is the total number of cells; $Obs_i$ is the observed count for cell $i$. $Exp_i$ is the expected count for cell $i$ under a specific null distribution and it can be calculated based on the mle of model parameters under a null distribution.

Under a specific null hypothesis, the chi-squared statistic follows a chi-squared distribution asymptotically. The degrees of freedom of a $X^2$ are the number of independent parameters in the full model minus that in the reduced model. For the two-way contingency table with $I$ rows and $J$ columns, the chi-squared statistic can be written as

$$X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

of which the degrees of freedom depends on the null hypothesis.

### 3.2.1 The Theorectical Justification of the Pearson’s Chi-Squared Statistic

For a multinomial sample $(n_1, n_2, \cdots, n_c)$ of size $n$, the marginal distribution of $n_i$ is the $Binomial(n, \pi)$ distribution. For large $n$, by the normal approximation to the binomial, $n_i$ has approximate normal distribution. By the central limit theorem, $\hat{\pi} = (n_1/n, \cdots, n_{c-1}/n)^T$ has an approximate multivariate normal distribution. Let $\Sigma_0$ denote the null covariance matrix of $\sqrt{n}\hat{\pi}$, and let $\pi_0 = (\pi_{10}, \pi_{20}, \cdots, \pi_{c-1,0})^T$. Under $H_0$,

$$\sqrt{n}(\hat{\pi} - \pi_0) \rightarrow N(0, \Sigma_0)$$

Therefore, the quadratic form

$$n(\hat{\pi} - \pi_0)^T \Sigma_0^{-1}(\hat{\pi} - \pi_0) \rightarrow \chi^2_{c-1}$$

The covariance matrix of $\sqrt{n}\hat{\pi}$ has elements

$$\sigma_{jk} = -\pi_j \pi_k \text{ if } j \neq k$$
$$\sigma_j = \pi_j (1 - \pi_j) \text{ if } j = k$$

The matrix $\Sigma_0^{-1}$ has (j,k)th element $1/\pi_{c0}$ when $j \neq k$ and $(1/\pi_{j0} + 1/\pi_{c0})$ when $j = k$. It can be shown that the quadratic form is identical to the Pearson’s Chi-squared statistic.

Below is an argument used by R. A. Fisher (1922). Suppose $(n_1, n_2, \cdots, n_c)$ are independent Poisson r.v.s with means $(\mu_1, \mu_2, \cdots, \mu_c)$. For large $\{\mu_j\}$, the standardized values $\{z_i = (n_i - \mu_i)/\sqrt{\mu_i}\}$ have approximate standard normal distributions. Thuse $X^2 = \sum_{i=1}^c z_i^2$ has an approximate chi-squared distribution with $c$ degrees of freedom. Adding the linear constraint $\sum_{i=1}^n (n_i - \mu_i) = 0$, we lose a degree of freedom.
Pearson’s Chi-squared Test

• For the two-way contingency table, the chi-squared statistic can be written as

\[ X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} \]

• When “the” (will be discussed) null hypothesis is true, it follows the chi-squared distribution with \((I-1)(J-1)\) df. We will justify the df later.
Pearson’s Chi-squared Test for Independence
The chi-squared Test of Independence

• Suppose that 200 students are selected at random from UCI, i.e., we have a random sample.

• Each student in the sample is classified both according to
  – major
  – preference for candidate (A or B) in a forthcoming election.
The chi-squared Test of Independence

<table>
<thead>
<tr>
<th></th>
<th>Biology</th>
<th>Engineering and Science</th>
<th>Social Science</th>
<th>Other</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>24</td>
<td>24</td>
<td>17</td>
<td>27</td>
<td>92</td>
</tr>
<tr>
<td>B</td>
<td>23</td>
<td>14</td>
<td>8</td>
<td>19</td>
<td>64</td>
</tr>
<tr>
<td>Undecided</td>
<td>12</td>
<td>10</td>
<td>13</td>
<td>9</td>
<td>44</td>
</tr>
<tr>
<td>Totals</td>
<td>59</td>
<td>48</td>
<td>38</td>
<td>55</td>
<td>200</td>
</tr>
</tbody>
</table>

Is major associated with preference of candidate?
The general situation

Observed table

<table>
<thead>
<tr>
<th>$n_{11}$</th>
<th>$n_{12}$</th>
<th>$\cdots$</th>
<th>$n_{1J}$</th>
<th>$n_{1.}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$\cdots$</td>
<td>$n_{2J}$</td>
<td>$n_{2.}$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>$\cdots$</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>$n_{I1}$</td>
<td>$n_{J2}$</td>
<td>$\cdots$</td>
<td>$n_{IJ}$</td>
<td>$n_{I.}$</td>
</tr>
<tr>
<td>$n_{.1}$</td>
<td>$n_{.2}$</td>
<td>$\cdots$</td>
<td>$n_{.J}$</td>
<td>$n_{..}$</td>
</tr>
</tbody>
</table>
### Parameters

| $\pi_{11}$ | $\pi_{12}$ | \ldots | $\pi_{1J}$ | $\pi_1$. |
| $\pi_{21}$ | $\pi_{22}$ | \ldots | $\pi_{2J}$ | $\pi_2$. |
| \vdots | \vdots | \vdots | \vdots | \vdots |
| $\pi_{I1}$ | $\pi_{I2}$ | \ldots | $\pi_{IJ}$ | $\pi_I$. |
| $\pi_{1}$ | $\pi_{2}$ | \ldots | $\pi_{J}$ | 1 |

- $\pi_{ij}$ is the probability of being in row $i$ and column $j$.
- $\pi_{i.} = \sum_{j=1}^{J} \pi_{ij}$ is the probability of being in row $i$ (marginal probability).
- $\pi_{.j} = \sum_{i=1}^{I} \pi_{ij}$ is the probability of being in column $j$ (marginal probability).
- $\sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = \sum_{i=1}^{I} \pi_{i.} = \sum_{j=1}^{J} \pi_{.j} = 1.$
The null hypothesis ($H_0$)

- The random numbers ($N_{11}, \ldots, N_{ij}$) follow a multinomial distribution
- Likelihood function

$$L_1 = L(\pi_{11}, \ldots, \pi_{IJ}) = \left(\frac{n_{..}}{n_{11}, \ldots, n_{IJ}}\right)^{n_{11}} \cdots \pi_{IJ}^{n_{IJ}} \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{n_{ij}}$$

- Under the assumption of no association, i.e., independence, whether a subject belongs to a row is independent of which column (s)he belongs to

$$H_0 : \pi_{ij} = \pi_i \cdot \pi_j \text{ for } i = 1, \ldots, I, \ j = 1, \ldots, J$$
MLEs under $H_0$

- Under the null hypothesis $H_0$, the likelihood becomes

$$L_0 = \prod_{i=1}^{I} \prod_{j=1}^{J} \left[ \pi_{i, j} \right]^{n_{ij}}$$

$$l_0 = \log(L_0) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left[ n_{ij} \log(\pi_{i, j}) \right] + \text{Constant} = \sum_{i=1}^{I} \sum_{j=1}^{J} \left[ n_{ij} \log(\pi_{i, j}) \right] + \text{Constant}$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} \left[ n_{ij} \log(\pi_{i, j}) \right] + \sum_{i=1}^{I} \sum_{j=1}^{J} \left[ n_{ij} \log(\pi_{j, j}) \right] + \text{Constant}$$

$$= \sum_{i=1}^{I} \left[ \log(\pi_{i, i}) \sum_{j=1}^{J} n_{ij} \right] + \sum_{j=1}^{J} \left[ \sum_{i=1}^{I} n_{ij} \log(\pi_{j, j}) \right] + \text{Constant}$$

$$= \sum_{i=1}^{I} \left[ n_{i, i} \log(\pi_{i, i}) \right] + \sum_{j=1}^{J} \left[ \log(\pi_{j, j}) \sum_{i=1}^{I} n_{ij} \right] + \text{Constant} = \sum_{i=1}^{I} \left[ n_{i, i} \log(\pi_{i, i}) \right] + \sum_{j=1}^{J} \left[ n_{j, j} \log(\pi_{j, j}) \right] + \text{Constant}$$
Note that $\sum_{i=1}^{I} \pi_i = 1$. We use the set of "unique" parameters: $(\pi_{i_1}, \ldots, \pi_{i_{I_i-1}})$. By the constraint,

$\pi_{i_{I_i}} = 1 - \sum_{i=1}^{I_i-1} \pi_i$.

\[
\frac{\partial \mathcal{L}}{\partial \pi_{i}}, \quad \pi_{i} = \frac{\partial}{\partial \pi_{i}} \left[ \sum_{i=1}^{I} \eta_i \log(\pi_i) \right] \\
= \frac{\partial}{\partial \pi_{i}} \left[ \sum_{i=1}^{I} \eta_i \log(\pi_i) \right] \\
= \frac{\partial}{\partial \pi_{i}} \left[ \eta_{i_1} \log(\pi_{i_1}) + \cdots + \eta_{i_{I_i-1}} \log(\pi_{i_{I_i-1}}) + \eta_{i_{I_i}} \log(1 - \sum_{i=1}^{I_i-1} \pi_i) \right] \\
= \frac{\partial}{\partial \pi_{i}} \left[ \eta_{i_1} \log(\pi_{i_1}) + \eta_{i_{I_i}} \log(1 - \sum_{i=1}^{I_i-1} \pi_i) \right] \\
= \frac{\eta_{i_1}}{\pi_{i_1}} + \eta_{i_{I_i}} \frac{\partial}{\partial \pi_{i}} \left[ \log(1 - \sum_{i=1}^{I_i-1} \pi_i) \right] \\
= \frac{\eta_{i_1} \hat{\pi}_{i_1}}{\pi_{i_1}} + \eta_{i_{I_i}} \frac{\partial}{\partial \pi_{i}} \left[ \log(1 - \sum_{i=1}^{I_i-1} \pi_i) \right] \\
= \frac{\eta_{i_1} \hat{\pi}_{i_1}}{\pi_{i_1}} + \eta_{i_{I_i}} \frac{\partial}{\partial \pi_{i}} \left[ (1 - \pi_{i_1} - \cdots - \pi_{i_{I_i-1}}) \right] \\
= \frac{\eta_{i_1} \hat{\pi}_{i_1}}{\pi_{i_1}} - \frac{\eta_{i_{I_i}}}{1 - \sum_{i=1}^{I_i-1} \pi_i} = \frac{\eta_{i_1}}{\pi_{i_1}} - \frac{\eta_{i_{I_i}}}{\hat{\pi}_{i_i}} \triangleq 0
\]

$\Rightarrow \hat{\pi}_{i_1} = \frac{\eta_{i_1}}{\hat{\pi}_{i_1}} / \eta_{i_1}$.

Similarly, $\hat{\pi}_{i_i} = \frac{\eta_{i_i}}{\hat{\pi}_{i_i}} / \eta_{i_i}$ for $i = 1, \ldots, I_i - 1$.
MLEs under the null hypothesis

The mles of $\pi_i$ are solutions to

$$\frac{\partial l_0}{\partial \pi_i} = \frac{n_i}{\pi_i} - \frac{n_I}{1 - \sum_{i=1}^{I-1} \pi_i} = \frac{n_i}{\pi_i} - \frac{n_I}{\pi_I} = 0, \ i = 1, \cdots, I - 1$$

which implies $\hat{\pi}_i = \hat{\pi}_I \cdot \frac{n_i}{n_I}$. Because $\pi_i$ are probabilities of a multinomial distribution,

$$1 = \sum_{i=1}^I \hat{\pi}_i = \sum_{i=1}^I \frac{\hat{\pi}_I \cdot n_i}{n_I} = \frac{\hat{\pi}_I \cdot n_i}{n_I}$$

giving $\hat{\pi}_I = \frac{n_I}{n_i}$. Substituting it back leads to

$$\hat{\pi}_i = \frac{n_i}{n_i}, \ \text{for} \ i = 1, \cdots, I$$

Similarly,

$$\hat{\pi}_j = \frac{n_j}{n_i}, \ j = 1, \cdots, J$$

Therefore, under the null hypothesis of independence,

$$Exp_{ij} = n_\cdot \hat{\pi}_{ij} = n_\cdot \hat{\pi}_i \cdot \hat{\pi}_j = \frac{n_i \cdot n_j}{n_\cdot}$$
The Chi-squared Test for Independence

\[ X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_i \cdot n_j / n_{..})^2}{n_i \cdot n_j / n_{..}} \]

• Degrees of freedom
  – Full model: IJ-1 unique probability parameters, as the IJ probabilities add up to 1.
  – Reduced model: I-1 unique parameters for row marginal probabilities; J-1 unique parameters for column marginal probabilities. There are I+J-2 unique parameters in total.
  – The difference is (IJ-1)-(I+J-2)=IJ-I-J+1=(I-1)(J-1).

• Under the null hypothesis, the test statistic follows chi-squared distribution with (I-1)(J-1) degrees of freedom
# Major and Candidate Preference

## Observed counts

<table>
<thead>
<tr>
<th></th>
<th>Biology</th>
<th>Engineering and Science</th>
<th>Social Science</th>
<th>Other</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>24</td>
<td>24</td>
<td>17</td>
<td>27</td>
<td>92</td>
</tr>
<tr>
<td>B</td>
<td>23</td>
<td>14</td>
<td>8</td>
<td>19</td>
<td>64</td>
</tr>
<tr>
<td>Undecided</td>
<td>12</td>
<td>10</td>
<td>13</td>
<td>9</td>
<td>44</td>
</tr>
<tr>
<td>Totals</td>
<td>59</td>
<td>48</td>
<td>38</td>
<td>55</td>
<td>200</td>
</tr>
</tbody>
</table>

## Expected counts

<table>
<thead>
<tr>
<th></th>
<th>Biology</th>
<th>Engineering and Science</th>
<th>Social Science</th>
<th>Other</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>27.14</td>
<td>22.08</td>
<td>17.48</td>
<td>25.30</td>
<td>92</td>
</tr>
<tr>
<td>B</td>
<td>18.88</td>
<td>15.36</td>
<td>12.16</td>
<td>17.60</td>
<td>64</td>
</tr>
<tr>
<td>Undecided</td>
<td>12.98</td>
<td>10.56</td>
<td>8.36</td>
<td>12.10</td>
<td>44</td>
</tr>
<tr>
<td>Totals</td>
<td>59</td>
<td>48</td>
<td>38</td>
<td>55</td>
<td>200</td>
</tr>
</tbody>
</table>
Major and Candidate Preference

• The chi-square statistic is 6.68
• Since I=3, J=4, the null distribution is the chi-squared distribution with (3-1)(4-1)=6 df
• Since the upper 5% point of $\chi_6^2$ is 12.59, at significance level 0.05, we do not reject the null hypothesis. There is not enough evidence to support the dependence between major and candidate preference.
• We can also use p-value: $1 - pchisq(6.68, 6) = 0.35$. 
Pearson’s Chi-squared Test for Homogeneity
Major vs preference (revisited)

• We now assume that the data were NOT from a random sample of the whole population
• Instead, they were obtained in the following way
  – First, 59 students were randomly selected from all students enrolled in Biology major
  – Second, 48 students were randomly selected from Engineering and Science
  – Third, 38 students were randomly selected from Social Science
  – Last, 55 students were randomly selected from other majors
• For each student, we asked his/her preference
Test of Homogeneity

• We are interested to know whether students in different major have same preference in candidates

Observed table

<table>
<thead>
<tr>
<th></th>
<th>$n_{11}$</th>
<th>$n_{12}$</th>
<th>…</th>
<th>$n_{1J}$</th>
<th>$n_{1.}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>…</td>
<td>$n_{2J}$</td>
<td>$n_{2.}$</td>
<td></td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>$n_{I1}$</td>
<td>$n_{J2}$</td>
<td>…</td>
<td>$n_{IJ}$</td>
<td>$n_{I.}$</td>
<td></td>
</tr>
<tr>
<td>$n_{.1}$</td>
<td>$n_{.2}$</td>
<td>…</td>
<td>$n_{.J}$</td>
<td>$n_{..}$</td>
<td></td>
</tr>
</tbody>
</table>
Test of Homogeneity

Parameter table

<table>
<thead>
<tr>
<th>$Pop_1$</th>
<th>$Pop_2$</th>
<th>$\cdots$</th>
<th>$Pop_J$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{11}$</td>
<td>$p_{12}$</td>
<td>$\cdots$</td>
<td>$p_{1J}$</td>
</tr>
<tr>
<td>$p_{21}$</td>
<td>$p_{22}$</td>
<td>$\cdots$</td>
<td>$p_{2J}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$p_{I1}$</td>
<td>$p_{I2}$</td>
<td>$\cdots$</td>
<td>$p_{IJ}$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>$\cdots$</td>
<td>1</td>
</tr>
</tbody>
</table>

- $p_{ij}$ is the probability that an observation chosen at random from the $j$th population will belong to the $i$th class of type $i$.

- $\sum_{i=1}^{I} p_{ij} = 1$, for $j = 1, \cdots, J$. 
The \( H_0 \)

The \( j \)th column \((n_{1j}, n_{2j}, \cdots, n_{Ij})\) denotes the observations from the \( j \)th population with a multinomial distribution and parameters \(((p_{1j}, p_{2j}, \cdots, p_{Ij}))\). Under the null hypothesis, the distributions for all the \( J \) columns are the same. The parameter table under the null hypothesis looks like:

\[
\begin{array}{cccc}
Pop_1 & Pop_2 & \cdots & Pop_J \\
p_{11} = p_1 & p_{12} = p_1 & \cdots & p_{1J} = p_1 \\
p_{21} = p_2 & p_{22} = p_2 & \cdots & p_{2J} = p_2 \\
\vdots & \vdots & \vdots & \vdots \\
p_{I1} = p_I & p_{I2} = p_I & \cdots & p_{IJ} = p_J \\
1 & 1 & \cdots & 1 \\
\end{array}
\]

The null hypothesis is

\[ H_0 : p_{i1} = p_{i2} = \cdots = p_{iJ} = p_i \text{ for } i = 1, \cdots, I \]
MLEs under $H_0$

With the constraint $\sum_{i=1}^{I} p_i = 1$, we have

$$L_0(p_1, \cdots, p_I) \propto p_1^{n_1} p_2^{n_2} \cdots \left(1 - \sum_{i=1}^{I-1} p_i\right)^{n_I}.$$ 

The loglikelihood

$$l_0 = \log L_0(p_1, \cdots, p_I) = c + n_1 \log(p_1) + \cdots + n_{(I-1)} \log(p_{I-1}) + n_I \log\left(1 - \sum_{i=1}^{I-1} p_i\right)$$

The mles are solutions to

$$\frac{\partial l_0}{\partial p_i} = \frac{n_i}{p_i} - \frac{n_I}{1 - \sum_{i=1}^{I-1} p_i} = \frac{n_1}{p_1} - \frac{n_I}{p_I} = 0, \ i = 1, \cdots, I$$
MLEs under $H_0$

which implies $\hat{p}_i = \hat{p}_I n_i / n_I$. Because $p_i$ are probabilities of a multinomial distribution,

$$1 = \sum_{i=1}^{I} \hat{p}_i = \sum_{i=1}^{I} \frac{\hat{p}_I n_i}{n_I} = \frac{\hat{p}_I n_..}{n_I}$$

giving $\hat{p}_I = \frac{n_I}{n_..}$. Substituting it back leads to

$$\hat{p}_i = \frac{n_i}{n_..}, \text{ for } i = 1, \cdots, I$$

Therefore, under the null hypothesis of homogeneity,

$$Exp_{ij} = n_{..} \hat{p}_i = \frac{n_i n_{..}}{n_..}$$
Test Statistic and Null Distribution

The Chi-squared test statistic is

\[ X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i..}n_{..}/n_{..})^2}{n_{i..}n_{..}/n_{..}} \]

Under the null hypothesis of homogeneity,

\[ X^2 \text{ follows } \chi^2_{(I-1)(J-1)} \]

Justification of df:

- **Full model**: \((I-1)\) parameters for each subpopulation. The total number of parameters in the \(J\) subpopulations is \((I-1)J\).
- **Reduced model**: \((I-1)\) parameters, as all the subpopulations have the same probabilities
- Difference in numbers of parameters: \((I-1)J – (I-1) = (I-1)(J-1)\)
Major vs Preference

• $X^2=6.68$, $p$-value=0.35

• At significance level 0.05 we fail to reject the null hypothesis. We have not enough evidence to conclude that the candidate preferences are different across majors.
Matched-Pairs Design
Introduction – Example 1

• Johnson and Johnson (1971) selected 85 Hodgkin’s patients who had a sibling of the same sex who was free of the disease and whose age was within 5 years of the patient’s. These investigators presented the following table:

<table>
<thead>
<tr>
<th></th>
<th>Tonsillectomy</th>
<th>No Tonsillectomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hodgkin’s</td>
<td>41</td>
<td>44</td>
</tr>
<tr>
<td>Control</td>
<td>33</td>
<td>52</td>
</tr>
</tbody>
</table>

• They wanted to know whether the tonsil act as a protective barrier against Hodgkin’s disease. The Pearson’s chi-squared statistic was calculated: 1.53 (p-value 0.22), which is not significant and they concluded that the tonsil is not a protector against Hodgkin’s disease. Any problem with this?
Introduction – Example 2

• Suppose that 100 persons were selected at random in a certain city, and that each person was asked whether he/she thought the service provided by the fire department in the city was satisfactory. Shortly after this survey was carried out, a large fire occurred in the city. Suppose that after this fire, the same 100 persons were again asked whether they thought that the service provided by the fire department was satisfactory. The results are presented in the table below:
Introduction – Example 2

Suppose we want to know whether people’s opinion was changed after the fire, how should we analyze the data? You may want to consider a test of homogeneity using a chi-square test. You apply the chi-squared test for homogeneity and obtain a chi-square statistic 1.75 and the corresponding p-value 0.19. However, it would not be appropriate to do so for this table because the observations taken before the fire and the observations taken after the fire are not independent. Although the total number of observations in the table is 200, only 100 indecently chosen persons were questioned in the surveys. It is reasonable to believe that a particular person’s opinion before the fire and after the fire are dependent.

<table>
<thead>
<tr>
<th></th>
<th>Satisfactory</th>
<th>Unsatisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before the fire</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>After the fire</td>
<td>72</td>
<td>28</td>
</tr>
</tbody>
</table>
The Proper Way to Display Correlated Tables

• To take the pairing/correlation nature of data into consideration, the data in the two examples should be displayed in a way that exhibits the pairing.

Example 1

<table>
<thead>
<tr>
<th></th>
<th>Sibling No Tonsillectomy</th>
<th>Sibling Tonsillectomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>No Tonsillectomy</td>
<td>37</td>
</tr>
<tr>
<td>Patient</td>
<td>Tonsillectomy</td>
<td>15</td>
</tr>
</tbody>
</table>

Example 2

<table>
<thead>
<tr>
<th></th>
<th>After the fire Satisfactory</th>
<th>Unsatisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before the fire</td>
<td>Satisfactory</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Unsatisfactory</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18</td>
</tr>
</tbody>
</table>
Analyzing Matched Pairs

• With the appropriate presentation, the data are a sample of size $n$ from a multinomial distribution with four cells. We can represent the probabilities in the tables as follows:

\[
\begin{array}{ccc}
\pi_{11} & \pi_{12} & \pi_1. \\
\hline
\pi_{21} & \pi_{22} & \pi_2. \\
\hline
\pi_1. & \pi_2. & 1 \\
\end{array}
\]
The null hypothesis for matched pairs

- The appropriate null hypothesis states that the probabilities of tonsillectomy are the same among patients and among siblings.

The null hypothesis is $\pi_2. = \pi_2$, which is equivalent to

$$H_0 : \pi_{12} = \pi_{21}$$
The likelihood

Under the full model (no constraint was imposed on the probability parameters except that the four probabilities add up to 1), the four counts follow a multinomial distribution

\[ (N_{11}, N_{12}, N_{21}, N_{22}) \sim \text{Multinomial}(n, (\pi_{11}, \pi_{12}, \pi_{21}, 1 - \pi_{11} - \pi_{12} - \pi_{21})) \]

The likelihood is

\[
L(\pi_{11}, \pi_{12}, \pi_{21}) = \frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} (1 - \pi_{11} - \pi_{12} - \pi_{21})^{n_{22}}
\]

\[
\propto \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} (1 - \pi_{11} - \pi_{12} - \pi_{21})^{n_{22}}
\]
The likelihood under $H_0: \pi_{12} = \pi_{21}$

- The likelihood under the null

\[
L_0(\pi_{11}, \pi_{12}) = \binom{n}{n_{11}, n_{12}, n_{21}, n_{22}} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} (1 - \pi_{11} - \pi_{12} - \pi_{12})^{n_{22}}
\]

\[
= \frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}+n_{21}} (1 - \pi_{11} - 2\pi_{12})^{n_{22}} \propto \pi_{11}^{n_{11}} \pi_{12}^{n_{12}+n_{21}} (1 - \pi_{11} - 2\pi_{12})^{n_{22}}
\]

- The log-likelihood under the null

\[
l_0 = \log[L_0(\pi_{11}, \pi_{12})]
\]

\[
= \text{Constants} + n_{11} \log(\pi_{11}) + (n_{12} + n_{21}) \log(\pi_{12}) + n_{22} \log(1 - \pi_{11} - 2\pi_{12})
\]
The MLE under \( H_0 : \pi_{12} = \pi_{21} \)

• To find MLE, we take partial derivatives, set them to zero, and solve for MLEs

\[
\begin{align*}
\frac{\partial l_0}{\partial \pi_{11}} &= \frac{n_{11}}{\pi_{11}} - \frac{n_{22}}{1 - \pi_{11} - 2\pi_{12}} = 0 \\
\Rightarrow \quad \pi_{11} &= \frac{n_{11}}{n} \\
\frac{\partial l_0}{\partial \pi_{12}} &= \frac{n_{12} + n_{21}}{\pi_{12}} - \frac{2n_{22}}{1 - \pi_{11} - 2\pi_{12}} = 0 \\
\Rightarrow \quad 2\pi_{12} &= \frac{n_{12} + n_{21}}{n_2} \\
\pi_{11} - \pi_{12} &= \frac{n_1}{n_2} \quad (1 - \pi_{11} - 2\pi_{12})
\end{align*}
\]

In addition, we have

\[
\begin{align*}
\hat{\pi}_{11} &= \frac{n_{11}}{n} = \frac{n_{11}}{n} , \\
2\hat{\pi}_{12} &= \frac{n_{12} + n_{21}}{n} = \frac{n_{12} + n_{21}}{n}
\end{align*}
\]
The Test Statistic

• Under the null hypothesis, the mle’s of the cell probabilities are

\[
\hat{\pi}_{11} = \frac{n_{11}}{n} \quad \hat{\pi}_{12} = \hat{\pi}_{21} = \frac{n_{12} + n_{21}}{2n} \quad \hat{\pi}_{22} = \frac{n_{22}}{n}
\]

The test-statistic is

\[
X^2 = \frac{(n_{11} - n_{11})^2}{n_{11}} + \frac{(n_{12} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} + \frac{(n_{21} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} + \frac{(n_{22} - n_{22})^2}{n_{22}}
\]

\[
= \frac{(n_{12} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} + \frac{(n_{21} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2}
\]

\[
= \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}
\]
• Under the null hypothesis, $X^2$ follows the chi-squared distribution with 1 df
  – Full model: three parameters $\pi_{11}, \pi_{12}, \pi_{21}$
  – Reduced model: two parameters $\pi_{11}, \pi_{12}$

• The test is known as McNemar’s test
Example 1

We fail to reject the null hypothesis at significance level 0.05. There is not enough evidence that tonsillectomy changes the risk of Hodgkin’s disease.
Example 2

We reject the null hypothesis at significance level 0.05 and conclude that the opinions of the residents were changed by the fire.

<table>
<thead>
<tr>
<th></th>
<th>After the fire</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Before the fire</td>
<td>Satisfactory</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Unsatisfactory</td>
<td>2</td>
<td>18</td>
</tr>
</tbody>
</table>

\[ X^2 = \frac{(10 - 2)^2}{(10 + 2)} = 5.33 \]

p-value < 0.02
Summary of Categorical Analysis

• Small sample size: Fisher’s exact test
• Large sample size: Pearson’s Chi-squared test. To calculate $\chi^2$, we need to find the expected counts. Steps
  – Step1: Write down the likelihood function under the null
  – Step2: the MLEs under the null
  – Step3: use the MLEs to compute the expected counts
Hints for Problem 2c (hw6)

• The likelihood function is

\[
\left( \frac{n!}{n_{11}!n_{12}! \ldots n_{IJ}!} \right) \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{n_{ij}}
\]

• To calculate the likelihood ratio statistic, you need to find the maximized likelihood functions under the following two situations:
  – Under the full model
  – Under the reduced model (the null hypothesis)
Maximized Likelihood Under the Full model

• Under the full model, the counts follow a multinomial with IJ-1 probability parameters.

• Step1: find the MLEs under the full model (show it!)

$$\hat{p}_{ij} = \frac{n_{ij}}{n_\cdot} \text{ for } i = 1, \cdots, I \text{ and } j = 1, \cdots, J.$$ 

• Step2: Plugging the MLEs into the likelihood function on page 59, you will obtain the maximized likelihood under the full model

$$\max\{L_1\} = \left(\frac{n!}{n_{11}!n_{12}!\cdots n_{IJ}!}\right) \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{n_{ij}}{n_\cdot}\right)^{n_{ij}}$$
Maximized Likelihood Under the Reduced Model

• Under the reduced model, we need to estimated \((I-1)+(J-1)\) probability parameters. We showed in class that (you don’t need to do it here) the MLEs are

\[
\hat{p}_i. = \frac{n_{i.}}{n_{..}}, \text{ for } i = 1, \cdots, I
\]

\[
\hat{p}_j = \frac{n_{.j}}{n_{..}}, \text{ for } j = 1, \cdots, J
\]

• The maximized likelihood is

\[
\max_{H_0}\{L_0\} = \left(\frac{n_{..}}{n_{11}!n_{12}!\cdots n_{IJ}!}\right) \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{n_{i.} n_{.j}}{n_{..} n_{..}}\right)^{n_{ij}}
\]
The Likelihood Ratio Statistic

• The likelihood ratio

\[ \Lambda = \frac{\max_{H_0}(L_0)}{\max(L_1)} = \frac{\prod_{i=1}^{I} \prod_{j=1}^{J} \left( \frac{n_{i.} \cdot n_{..}}{n_{..} \cdot n_{..}} \right)^{n_{ij}}}{\prod_{i=1}^{I} \prod_{j=1}^{J} \left( \frac{n_{ij}}{n_{..}} \right)^{n_{ij}}} \]

• To conduct a large-sample test, we use

\[ G = -2 \log \Lambda = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \left( \frac{n_{ij}}{n_{i.} \cdot n_{.j} / n_{..}} \right) \]

\[ = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \left( \frac{Obs_{ij}}{Exp_{ij}} \right) \]
The Large Sample LRT

• When the null hypothesis is true, $G$ follows the chi-squared distribution with
• $[(IJ-1)-[(I-1)+(J-1)]=((I-1)(J-1)$ df
• Based upon the observed counts, you use either R or a hand calculator to find $G$
• Make your conclusions