Comparing two multinomial distributions

Assume we have two independent samples, $A$ and $B$, from the multinomial distributions with $k$ categories and probabilities $\pi^A_1, \ldots, \pi^A_k$ and $\pi^B_1, \ldots, \pi^B_k$, respectively.

Sample $A$: $Y^A_1, \ldots, Y^A_k$ and sample $B$: $Y^B_1, \ldots, Y^B_k$.

$H_0$: samples $A$ and $B$ come from the same distribution: $\pi^A_1 = \pi^B_1 = \pi_1, \ldots, \pi^A_k = \pi^B_k = \pi_k$.

Under $H_0$, we have one sample, $Y^A_1 + Y^B_1, \ldots, Y^A_k + Y^B_k$, from the multinomial distribution with probabilities $\pi_1, \ldots, \pi_k$. Restricted MLE:

$$\hat{\pi}_1 = \frac{Y^A_1 + Y^B_1}{n_A + n_B}, \ldots, \hat{\pi}_k = \frac{Y^A_k + Y^B_k}{n_A + n_B},$$

where

$$n_A = \sum_{i=1}^{k} Y^A_i, \quad n_B = \sum_{i=1}^{k} Y^B_i.$$
Comparing two multinomial distributions

Let $H_a$: samples $A$ and $B$ come from two different multinomial distributions.

**Unrestricted MLE** for the two samples:

$$\hat{\pi}_1^A = \frac{Y_1^A}{n_A}, \ldots, \hat{\pi}_k^A = \frac{Y_k^A}{n_A}, \quad \hat{\pi}_1^B = \frac{Y_1^B}{n_B}, \ldots, \hat{\pi}_k^B = \frac{Y_k^B}{n_B}.$$ 

Expected frequencies under $H_0$:

$$E_1^A = n_A\hat{\pi}_1, \quad E_1^B = n_B\hat{\pi}_1, \quad \ldots, \quad E_k^A = n_A\hat{\pi}_k, \quad E_k^B = n_B\hat{\pi}_k.$$ 

The approximate distribution of Pearson’s statistic

$$\chi^2 = \sum_{i=1}^k \left[ \frac{(Y_i^A - E_i^A)^2}{E_i^A} + \frac{(Y_i^B - E_i^B)^2}{E_i^B} \right] \sim \chi_{k-1}^2,$$

because we have $k - 1$ free parameters under $H_0$ and $2(k - 1)$ free parameters in the unrestricted model.
Comparing two multinomial distributions

**Example 1:** Sample A is 400 patients with type 2 diabetes, and sample B is 600 patients with no diabetes. There are $k = 3$ categories (low, medium and high sugar intake).

Sample A: $y_1^A = 51$, $y_2^A = 137$, $y_3^A = 212$.
Sample B: $y_1^B = 119$, $y_2^B = 206$, $y_3^B = 275$.

Under $H_0$: $\pi_1^A = \pi_1^B = \pi_1$, $\pi_2^A = \pi_2^B = \pi_2$ and $\pi_3^A = \pi_3^B = \pi_3$, the restricted MLE is

$$\hat{\pi}_1 = \frac{51 + 119}{1000} = 0.170, \quad \hat{\pi}_2 = \frac{137 + 206}{1000} = 0.343, \quad \hat{\pi}_3 = \frac{212 + 275}{1000} = 0.487.$$

Expected frequencies are

$$e_1^A = 400 \cdot 0.170 = 68, \quad e_2^A = 400 \cdot 0.343 = 137.2, \quad e_3^A = 400 \cdot 0.487 = 194.8,$$
$$e_1^B = 600 \cdot 0.170 = 102, \quad e_2^B = 600 \cdot 0.343 = 205.8, \quad e_3^B = 600 \cdot 0.487 = 292.2.$$

We find that the observed value of Pearson’s statistic is $x^2 = 9.61$. 
Comparing two multinomial distributions: CLICKER

QUESTION 1

Example 1 (contd): Sample A is 400 patients with type 2 diabetes, and sample B is 600 patients with no diabetes. There are $k = 3$ categories (low, medium and high sugar intake).

What is the approximate distribution of Pearson’s statistic under the null in this example?

A $\chi^2_5$
B $\chi^2_4$
C $\chi^2_3$
D $\chi^2_2$
E $\chi^2_1$
Comparing two multinomial distributions: CLICKER

QUESTION 2

Example 1 (contd): Sample A is 400 patients with type 2 diabetes, and sample B is 600 patients with no diabetes. There are $k = 3$ categories (low, medium and high sugar intake).

The $p$-value here is

$$\Pr(X^2 > 9.61 | X^2 \sim \chi^2_2) = 0.0082.$$ 

Select statement which is not correct

A. $H_0$ is not rejected at the 1% significance level.
B. $H_0$ is not rejected at the 0.5% significance level.
C. $H_0$ is rejected at the 5% significance level.
D. $H_0$ is rejected at the 2% significance level.
E. $H_0$ is rejected at the 1% significance level.
Comparing two multinomial distributions

**Example 2:** Assume now that we have the same frequency data but this time we randomly select 1000 patients and categorize them using two factors: sugar intake (L/M/H) and diabetes status (Y/N).

We get a two-way contingency table using six different combinations of these two factors (i.e., $k = 6$ categories):

<table>
<thead>
<tr>
<th>factor 2</th>
<th>factor 1</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>137</td>
<td>212</td>
</tr>
<tr>
<td>119</td>
<td>206</td>
<td>275</td>
</tr>
<tr>
<td>total</td>
<td>170</td>
<td>343</td>
</tr>
<tr>
<td>400</td>
<td>600</td>
<td></td>
</tr>
</tbody>
</table>

Let $H_0$: factor 1 and factor 2 are independent. For example, under $H_0$, we have

$$\Pr(\text{patient with low sugar intake has diabetes}) = \Pr(\text{patient has low sugar intake}) \Pr(\text{patient has diabetes}).$$
Comparing two multinomial distributions

**Example 2 (contd):** Assume now that we have the same frequency data but this time we randomly select 1000 patients and categorize them using two factors: sugar intake (L/M/H) and diabetes status (Y/N).

We get a multinomial distribution with \( n = 1000 \), \( k = 6 \) categories and probabilities \( \pi_{LY}, \pi_{MY}, \pi_{HY}, \pi_{LN}, \pi_{MN}, \pi_{HN} \).

Under \( H_0 \),

\[
\begin{align*}
\pi_{LY} &= \pi_L \pi_Y, & \pi_{MY} &= \pi_M \pi_Y, & \pi_{HY} &= \pi_H \pi_Y, \\
\pi_{LN} &= \pi_L \pi_N, & \pi_{MN} &= \pi_M \pi_N, & \pi_{HN} &= \pi_H \pi_N,
\end{align*}
\]

where \( \pi_L/\pi_M/\pi_H \) is the probability that randomly selected patient has lower/medium/high sugar intake, and \( \pi_Y/\pi_N \) is the probability that randomly selected patient has/doesn’t have diabetes.
Comparing two multinomial distributions: CLICKER
QUESTION 3

**Example 2 (contd):** Assume now that we have the same frequency data but this time we randomly select 1000 patients and categorize them using two factors: sugar intake (L/M/H) and diabetes status (Y/N).

How many free parameters does the multinomial distribution have under $H_0$?

- A 5 parameters
- B 4 parameters
- C 3 parameters
- D 2 parameters
- E 1 parameter
Comparing two multinomial distributions

**Example 2 (contd):** Assume now that we have the same frequency data but this time we randomly select 1000 patients and categorize them using two factors: sugar intake (L/M/H) and diabetes status (Y/N).

We find maximum likelihood estimates for $\pi_L, \pi_M, \pi_H$:

$$\hat{\pi}_L = \frac{170}{1000}, \quad \hat{\pi}_M = \frac{343}{1000}, \quad \hat{\pi}_H = \frac{487}{1000},$$

and for $\pi_Y, \pi_N$:

$$\hat{\pi}_Y = \frac{400}{1000}, \quad \hat{\pi}_N = \frac{600}{1000}.$$

**The expected frequencies** under $H_0$: $e_{LY} = 1000 \cdot \hat{\pi}_L \cdot \hat{\pi}_Y = 68$, etc. We can see that we get the same expected frequencies as in example 1!
Example 2 (contd): Assume now that we have the same frequency data but this time we randomly select 1000 patients and categorize them using two factors: sugar intake (L/M/H) and diabetes status (Y/N).

Because the expected (and observed) frequencies are the same, we get the same observed value of Pearson’s statistic: $x^2 = 9.61$. Under $H_0$: $X^2 \sim \chi^2_2$ so that $H_0$ is rejected at the 1% significance level.

Conclusion: Testing if two independent samples (for patients who have/don’t have diabetes) come from the same multinomial distribution (where categories are sugar intake levels) is equivalent to testing if the two factors (diabetes status and sugar intake level) are independent or not.