Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Difference between categorical distributions

When we have data from two samples where each one of them has its particular categorical distribution and we want to test if these distributions are the same we have to follow the next steps.

Step 1: Consider two samples with n_A and n_B data points where each data point has only n possibles values.

Step 2: We going to define two categorical random variables X_A and X_B which represents the possibles values can take these random variables.

$$X_A = \begin{cases} v_1^A & with & probability = p_1^A \\ v_2^A & with & probability = p_2^A \\ \vdots & \vdots & \vdots & \vdots \\ v_n^A & with & probability = p_n^A \end{cases}$$

$$X_B = \begin{cases} v_1^B & with & probability = p_1^B \\ v_2^B & with & probability = p_2^B \\ \vdots & \vdots & \vdots & \vdots \\ v_n^B & with & probability = p_n^B \end{cases}$$

Step 3: We going to define $(2 \cdot n)$ new random variables $Y_A^{(1)}, ..., Y_A^{(n)}, Y_B^{(1)}, ..., Y_B^{(n)}$.

$$Y_A^{(1)} = \begin{cases} v_1^A & with & probability = p_1^A \\ other\ value & with & probability = (1 - p_1^A) \end{cases}$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$Y_A^{(n)} = \begin{cases} v_n^A & with & probability = p_n^A \\ other\ value & with & probability = (1 - p_n^A) \end{cases}$$

$$\boxed{Y_B^{(1)} = \begin{cases} v_1^B & with \quad probability = p_1^B \\ other\ value & with \quad probability = (1 - p_1^B) \end{cases}}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\boxed{Y_B^{(n)} = \begin{cases} v_n^B & with \quad probability = p_n^B \\ other\ value & with \quad probability = (1 - p_n^B) \end{cases}}$$

Step 4: We have to test n hypothesis which are:

$$\boxed{H_0^{(1)} : Y_A^{(1)} = Y_B^{(1)} \qquad i.e. \qquad \mu_A^{(1)} = \mu_B^{(1)}}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\boxed{H_0^{(n)} : Y_A^{(n)} = Y_B^{(n)} \qquad i.e. \qquad \mu_A^{(n)} = \mu_B^{(n)}}$$

Step 5: For $k \in \{1, ..., n\}$ we going to define the next $(2 \cdot n)$ estimators.

$$\boxed{T_A^{(k)} = \frac{1}{n_A} \cdot \sum_{i=1}^{n_A} Y_{A,i}^{(k)} = \hat{p}_i^A}$$

$$\boxed{T_B^{(k)} = \frac{1}{n_B} \cdot \sum_{i=1}^{n_B} Y_{B,i}^{(k)} = \hat{p}_i^B}$$

You can prove that (unbiased estimators):

$$\boxed{\mathbb{E}\left[T_A^{(k)}\right] = \mu_{A,i}^{(k)} = p_i^A}$$

$$\boxed{\mathbb{E}\left[T_B^{(k)}\right] = \mu_{B,i}^{(k)} = p_i^B}$$

Considering the fact of these are Bernoulli Random Variable:

$$\boxed{\mathbb{V}\left[T_A^{(k)}\right] = \frac{p_i^A \cdot (1 - p_i^A)}{n_A}}$$

$$\boxed{\mathbb{V}\left[T_B^{(k)}\right] = \frac{p_i^B \cdot (1 - p_i^B)}{n_B}}$$

Using the estimator for the respective probabilities:

$$\boxed{\hat{\mathbb{V}}\left[T_A^{(k)}\right] = \frac{\hat{p}_i^A \cdot (1 - \hat{p}_i^A)}{n_A}}$$

$$\boxed{\hat{\mathbb{V}}\left[T_B^{(k)}\right] = \frac{\hat{p}_i^B \cdot (1 - \hat{p}_i^B)}{n_B}}$$

Step 6: To test every of the n hypothesis we going to define the next estimators:

$$T^{(k)} = T_A^{(k)} - T_B^{(k)}$$

$$\mathbb{E}\left[T^{(k)}\right] = \mathbb{E}\left[T_A^{(k)}\right] - \mathbb{E}\left[T_B^{(k)}\right] = p_i^A - p_i^B$$

$$\hat{\mathbb{V}}\left[T^{(k)}\right] = \hat{\mathbb{V}}\left[T_A^{(k)}\right] + \hat{\mathbb{V}}\left[T_B^{(k)}\right] = \frac{\hat{p}_i^A \cdot (1 - \hat{p}_i^A)}{n_A} + \frac{\hat{p}_i^B \cdot (1 - \hat{p}_i^B)}{n_B}$$

Now we define:

$$Z^{(k)} = \frac{(\hat{p}_i^A - \hat{p}_i^B) - (p_i^A - p_i^B)}{\sqrt{\frac{\hat{p}_i^A \cdot (1 - \hat{p}_i^A)}{n_A} + \frac{\hat{p}_i^B \cdot (1 - \hat{p}_i^B)}{n_B}}}$$

And to test the hypothesis:

$$H_0^{(k)} : Y_A^{(k)} = Y_B^{(k)}$$

We get:

$$\boxed{Z^{(k)} = \frac{(\hat{p}_i^A - \hat{p}_i^B)}{\sqrt{\frac{\hat{p}_i^A \cdot (1 - \hat{p}_i^A)}{n_A} + \frac{\hat{p}_i^B \cdot (1 - \hat{p}_i^B)}{n_B}}}}$$

Now if $|Z^{(k)}| \geq Z_{1-\frac{\alpha}{2}}$ we can reject the null hypothesis with $(1-\alpha)\,\%$ of confidence.
