Discrete Response Model
Lecture 4

datascience@berkeley

Independence

Independence

Independence between X and Y when the occurrence of X = i does not have an effect of the occurrence of Y = j for each i = 1, …, I and j = 1, …, J.

Symbolically, independence exists when $\pi_{ij} = \pi_{i+}\pi_{+j}$

for i = 1, …, I and j = 1, …, J.

➔ This means that the probability of an item being in cell (i,j) only involves knowing the separate marginal probabilities for X and for Y; thus, X and Y are independent of each other.

Why is independence important?

Independence helps to simplify the understanding of probabilities within the contingency table!

There are only $I - 1 + J - 1 = I + J - 2$ unknown probability parameters.

Note that the "-1" parts occur due to the $\sum_{i=1}^{I} \pi_{i+} = 1$ and $\sum_{j=1}^{J} \pi_{+j} = 1$.

$IJ - 1$

Without independence, there are $IJ - 1$ unknown probability parameters, where the "-1" part occurs due to the $\sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1$.

Implications of Independence

Consider the I multinomial distributions case again. Similar to Section 1.2, it is often of interest to know if these conditional probabilities are equal across the rows of the table. Thus, we want to know if $\pi_{j|1} = \cdots = \pi_{j|I}$ for $j = 1, \ldots, J$. Note that this is mathematically equivalent to $\pi_{ij} = \pi_{i+}\pi_{+j}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$!

Recall from w203 that

$$P(Y = j \mid X = i) = P(X = i, Y = j) / P(X = i).$$

Thus,

$$\pi_{j|i} = \pi_{ij} / \pi_{i+} = \pi_{i+}\pi_{+j} / \pi_{i+} = \pi_{+j}$$

under independence.

Because each $\pi_{j|I}$ is equal to π_{+j} for each i, we have $\pi_{j|1} = \cdots = \pi_{j|I}$.

Because of this equivalence, we will refer to $\pi_{j|1} = \cdots = \pi_{j|I}$ for $j = 1, \ldots, J$ as "independence" as well.

Test for Independence

The hypotheses are:

$H_0: \pi_{ij} = \pi_{i+}\pi_{+j}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$

H_a: Not all equal

Remember that a Pearson chi-square test statistic calculates

$$\frac{(\text{observed count} - \text{estimated expected count})^2}{(\text{estimated expected count})}$$

for every cell of a contingency table and sums these quantities. The Pearson chi-square test for independence then uses the statistic

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}$$

Remarks

- The estimated expected cell count is $n\hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n$
- X^2 is equivalent to the corresponding statistic used in Section 1.2 (in the text) for the Pearson chi-square test for a 2×2 contingency table.
- If the null hypothesis is true, X^2 has a $\chi^2_{(I-1)(J-1)}$ distribution for a large sample.
- Reject the null hypothesis if $X^2 > \chi^2_{(I-1)(J-1),1-\alpha}$.

Likelihood-Ratio Test (LRT) Statistic

The LRT statistic is formed the usual way with

$$\Lambda = \frac{\text{Max. lik. when parameters satisfy } H_0}{\text{Max. lik. when parameters satisfy } H_0 \text{ or } H_a}$$

The numerator of Λ uses $\hat{\pi}_{i+}\hat{\pi}_{+j}$ to estimate π_{ij}, and the denominator of Λ uses $\hat{\pi}_{ij}$ to estimate π_{ij}. The transformed statistic simplifies to

$$-2\log(\Lambda) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{n_{i+}n_{+j}/n}\right)$$

where we use $0\times\log(0) = 0$. The large sample distribution is the same as for X^2.

Degree of Freedom

A general way to find degrees of freedom for a hypothesis test is to calculate:

(Number of free parameters under H_a)
− (Number of free parameters under H_0)

Under the alternative hypothesis for the test of independence, we have IJ π_{ij} parameters with the restriction that $\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij} = 1$. When independence is true, we need the I different π_{i+} and the J different π_{+j} parameters to find π_{ij} with the restriction that $\sum_{i=1}^{I}\pi_{i+} = 1$ and $\sum_{j=1}^{J}\pi_{+j} = 1$.

Thus, the overall degrees of freedom is

$$(IJ - 1) - (I + J - 2) = (I - 1)(J - 1)$$

Example

- Fiber is often added to foods as a convenient way for people to consume it.
- The Data and Story Library (DASL) describes the results of a study where individuals are given a new type of fiber-enriched cracker.
- The participants ate the crackers and then a meal. Shortly afterward, the participants were instructed to describe any bloating that they experienced.
- Below is the data:

		Bloating severity			
		None	Low	Medium	High
Fiber source	None	6	4	2	0
	Bran	7	4	1	0
	Gum	2	2	3	5
	Both	2	5	3	2

Berkeley SCHOOL OF INFORMATION