STA 250: Statistics

Notes 18. Chi-square tests for Independence, Goodness-of-fit

Book chapters: 10.4

1 Two-way tables: category counts based on two attributes

A two-way contingency table is a special kind of categorical data where n units are split into categories determined by two attributes. In the following table, 231 humans are categorized according to their hair color and eye color.

		Eye color				
		Blue	Green	Brown	Black	Total
Hair color	Blonde	20	15	18	14	67
	Red	11	4	24	2	41
	Brown	9	11	36	18	74
	Black	8	17	20	4	49
	Total	48	47	98	38	231

The table contains counts data X on 16 categories and so we can model $X \sim \mathsf{Multinomial}(n, p)$ with p in the 16-dimensional simplex.

However, as shown in the table, it is more convenient and informative to present the counts as a two-dimensional array, with the labels of one attribute along each dimension. Therefore, we would interpret, with $R =$ number of rows and $C =$ number of columns of the table,

$$X = (X_{11}, \cdots, X_{R1}, X_{12}, \cdots, X_{R2}, \cdots, X_{1C}, \cdots, X_{RC})$$
$$p = (p_{11}, \cdots, p_{R1}, p_{12}, \cdots, p_{R2}, \cdots, p_{1C}, \cdots, p_{RC}),$$

i.e., x_{ij} is the count in the (i, j)-th cell of the table and p_{ij} is the probability that any unit would land up in that cell. Observed data would be denoted in the same manner: $x = (x_{11}, \cdots, x_{R1}, x_{12}, \cdots, x_{R2}, \cdots, x_{1C}, \cdots, x_{RC})$.

To facilitate discussion, introduce the following margin counts

$$\begin{aligned} X_{i\cdot} &= X_{i1} + X_{i2} + \cdots + X_{iC}, \quad i = 1, 2, \cdots, R \\ X_{\cdot j} &= X_{1j} + X_{2j} + \cdots + X_{Rj}, \quad j = 1, 2, \cdots, C \end{aligned}$$

and use similar definitions on p to get $p_{i\cdot}$'s, $p_{\cdot j}$'s and on x to get $x_{i\cdots}$'s and $x_{\cdot j}$'s. For the hair-eye color table, $x_{1\cdot} = 67$, $x_{\cdot 3} = 98$ and so on.

The row counts $X^R = (X_{1\cdot}, \cdots, X_{R\cdot})$ are simply the count data when the units are only split according to the row attribute (hair color). Similarly, the columns counts $X^C =$

$(X_{\cdot 1}, \cdots, X_{\cdot C})$ give the one-way categorization according to the column-attribute (eye color). Under the model $X \sim \mathsf{Multinomial}(n, p)$, we must have $X^R \sim \mathsf{Multinomial}(n, p^R)$ and $X^C \sim \mathsf{Multinomial}(n, p^C)$. And if we observe data $X = x$ then the observation for X^R is simply x^R and that for X^C is x^C.

2 Test of attribute independence

Two way tables are widely used to test for independence between the two attributes. For our hair-eye color example, we could test whether what eye color we have is independent of our hair color. Under independence of the two attributes we can write:

$$\begin{aligned} p_{ij} &= P(\text{row attribute label} = i, \text{column attribute label} = j) \\ &= P(\text{row attribute label} = i) \times P(\text{column attribute label} = j) \\ &= p_{i\cdot} p_{\cdot j}. \end{aligned}$$

Therefore, if we know $p^R = (p_{1\cdot}, \cdots, p_{R\cdot})$ and $p^C = (p_{\cdot 1}, \cdots, p_{\cdot C})$, then under the assumption of independence we can construct all RC elements of$p$ by taking products $p_{ij} = p_{i\cdot} p_{\cdot j}$. We'll use $p = \text{out}(p^R, p^C)$ to denote such an "outer product" construction of $p \in \Delta_{RC}$ from a $p^R \in \Delta_R$ and a $p^C \in \Delta_C$. So, we can formalize the hypothesis of independence as

$$H_0 : p \in \Delta_{RC}^{\text{out}}, \quad H_1 : p \notin \Delta_{RC}^{\text{out}}$$

where $\Delta_{RC}^{\text{out}} = \{p = \text{out}(p^R, p^C) : p^R \in \Delta_R,\ p^C \in \Delta_C\}$.

3 Pearson's test of independence

From our discussion in the last handout, we can test the above H_0 as follows. We first construct

$$Q(x) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(x_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

where $\hat{e}_{ij} = n\hat{p}_{ij}^{\text{out}}$ is the expected cell counts under H_0, with $\hat{p}^{\text{out}}$ denoting the restricted MLE of p under H_0. To find $\hat{p}^{\text{out}}$ notice that for a $p = \text{out}(p^R, p^C)$, the likelihood function simplifies to

$$L_x(p) = \text{const} \times \left[\prod_{i=1}^{R} p_{i\cdot}^{x_{i\cdot}}\right] \times \left[\prod_{j=1}^{C} p_{\cdot j}^{x_{\cdot j}}\right]$$

and hence it is maximized at $\hat{p}_{i\cdot} = x_{i\cdot}/n$, $i = 1, 2, \cdots, R$ and $\hat{p}_{\cdot j} = x_{\cdot j}/n$, $j = 1, 2, \cdots, C$. That is, $\hat{p}_{ij}^{\text{out}} = (x_{i\cdot} x_{\cdot j})/n^2$ and consequently, $\hat{e}_{ij} = (x_{i\cdot} x_{\cdot j})/n$.

With $Q(x)$ constructed, Pearson's chi-square size-α test would reject H_0 if $Q(x) > F_{k-1-r}^{-1}(1-\alpha)$ where $k = RC$ is the total number of categories, and r is the number of free parameters in Δ_{RC}^{out}. Note that elements of this set are made of a $p^R \in \Delta_R$ with $R-1$ free elements, plus with a $p^C \in \Delta_C$ with another $C-1$ free elements. So the number of free elements in Δ_{RC}^{out} is $r = R-1+C-1$. Combining, we get $k-1-r = RC-R-C+1 = (R-1)(C-1)$.

So in summary, we'd perform a size-α Pearson's chi-square test of independence by rejecting H_0 if $Q(x) > F^{-1}_{(R-1)(C-1)}(1-\alpha)$. The p-value based on these tests is given by $1 - F_{(R-1)(C-1)}(Q(x))$.

Example (Hair-eye color)**.** To perform the test on the hair/eye color data, we'd insert in each cell on the table the expected count $x_{i\cdot}x_{\cdot j}/n$ (shown in parentheses below).

		Eye color				
		Blue	Green	Brown	Black	Total
Hair color	Blonde	20 (13.9)	15 (13.6)	18 (28.4)	14 (11.0)	67
	Red	11(8.5)	4 (8.3)	24 (17.4)	2 (6.7)	41
	Brown	9 (15.4)	11 (15.1)	36 (31.4)	18 (12.2)	74
	Black	8 (10.2)	17 (10.0)	20 (20.8)	4 (8.1)	49
	Total	48	47	98	38	231

and find $Q(x) = 30.9$. So the p-value is $1 - F_{(4-1)(4-1)}(30.9) = 1 - F_9(30.9) \approx 0$.