THE 2-BY-2 CHI-SQUARED INDEPENDENCE STATISTIC

RUSS WOODROOFE

1. General setting

We are interested in determining whether two discrete random variables A and B (on the same sample space) are independent.

Let A have a outcomes $\alpha_1, \ldots, \alpha_a$, and B have b outcomes $\beta_1, \ldots, \beta_b$. We get a joint probability distribution:

$$p_{A,B}(i,j) = P(A = \alpha_i, B = \beta_j).$$

For short, we denote $p_{A,B}(i,j)$ as $p_{i.j}$. We recall that A and B independent means that $p_{i,j} = p_A(i)p_B(j)$, where $p_A(i) = P(A = \alpha_i)$ is the pmf of A, and similarly for B. For short, we denote $p_A(i) = p_{i,\bullet}$ and $p_B(j) = p_{\bullet,j}$. Thus, A and B are independent if and only if $p_{i,j} = p_{i,\bullet}p_{\bullet,j}$.

As usual, to do statistics, we perform n repeated trials of the experiment underlying the sample space, and for each resulting ω find $A(\omega)$ and $B(\omega)$. We count the number of times each possible (α_i, β_j) outcome occurs, and denote this $O_{i,j}$. To say this differently, $O_{i,j}$ is the number of experiments where $A = \alpha_i$ and $B = \beta_j$.

Similarly to our notation for $p_{i,\bullet}$, we let $O_{i,\bullet}$ be the number of experiments where $A = \alpha_i$, and $O_{\bullet,j}$ be the number of experiments where $B = \beta_j$. Thus $O_{i,\bullet} = \sum_{j=1}^{b} O_{i,j}$ and $O_{\bullet,j} = \sum_{i=1}^{a} O_{i,j}$.

By our work on the χ^2 goodness-of-fit statistics, we have that if we knew $p_{i,\bullet}$ and $p_{\bullet,j}$ for all i and j, and if A and B are independent then

$$\sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(O_{i,j} - np_{i,\bullet}p_{\bullet,j})^2}{np_{i,\bullet}p_{\bullet,j}} \text{ is approximately } \chi^2(ab-1).$$

In typical circumstances, however, the probabilities $p_{i,\bullet}$ and $p_{\bullet,j}$ will be unknown, so we estimate them as $\hat{p}_{i,\bullet} = \frac{O_{i,\bullet}}{n}$ and $\hat{p}_{\bullet,j} = \frac{O_{\bullet,j}}{n}$. Similarly to the situation with estimating σ^2 with S^2 (which uses $\overline{X}$ in place of μ), this will require an adjustment to our statistics.

2. Theorem and proof

Theorem 1. *Let A and B be random variables on a common sample space, where A has a possible values $\alpha_1, \ldots, \alpha_a$, and B has b possible values $\beta_1, \ldots, \beta_b$. Repeat the random trial n times, and let*

$$O_{i,j} = \#\textit{trials with } A = \alpha_i, B = \beta_j.$$

As above, denote by $O_{i,\bullet}$ the number of trials with $A = \alpha_i$, similarly for $O_{\bullet,j}$.

If A and B are independent, then

$$X = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(O_{i,j} - n\frac{O_{i,\bullet}}{n}\frac{O_{\bullet,j}}{n}\right)^2}{n\frac{O_{i,\bullet}}{n}\frac{O_{\bullet,j}}{n}} \textit{ has approximate distribution } \chi^2((a-1)(b-1)).$$

Proof sketch for $a = 2$. In this case, we have the simple expression $O_{\bullet,j} = O_{1,j} + O_{2,j}$. To make the notation simpler, we write $n_1 = O_{1,\bullet}$ and $n_2 = O_{2,\bullet}$, so that $n_1 + n_2 = n$.

Then we expand the "square" term for $i = 1$:

$$\begin{aligned}\left(O_{1,j} - n\frac{O_{1,\bullet}}{n}\frac{O_{\bullet,j}}{n}\right)^2 &= \left(O_{1,j} - n_1\frac{O_{\bullet,j}}{n}\right)^2 = \left(\frac{nO_{1,j} - n_1O_{\bullet,j}}{n}\right)^2 \\ &= \left(\frac{(n_1+n_2)O_{1,j} - n_1(O_{1,j}+O_{2,j})}{n}\right)^2 \\ &= \left(\frac{n_2O_{1,j} - n_1O_{2,j}}{n}\right)^2.\end{aligned}$$

For $i = 2$, we reverse the role of 1 and 2, which merely reverses the sign inside the square, hence also

$$\left(O_{2,j} - n\frac{O_{2,\bullet}}{n}\frac{O_{\bullet,j}}{n}\right)^2 = \left(\frac{n_2O_{1,j} - n_1O_{2,j}}{n}\right)^2.$$

Thus, our statistic is

$$\begin{aligned}X &= \sum_{j=1}^{b} \frac{\left(\frac{n_2O_{1,j}-n_1O_{2,j}}{n}\right)^2}{n_1\cdot\frac{O_{\bullet,j}}{n}} + \frac{\left(\frac{n_2O_{1,j}-n_1O_{2,j}}{n}\right)^2}{n_2\cdot\frac{O_{\bullet,j}}{n}} \\ &= \sum_{j=1}^{b} \frac{(n_1+n_2)\left(\frac{n_2O_{1,j}-n_1O_{2,j}}{n}\right)^2}{n_1n_2\cdot\frac{O_{\bullet,j}}{n}} \\ &= \sum_{j=1}^{b} \frac{(n_2O_{1,j}-n_1O_{2,j})^2}{nn_1n_2\frac{O_{\bullet,j}}{n}}.\end{aligned}$$

To simplify notation, at this point we switch to writing $\frac{O_{\bullet,j}}{n}$ as $\hat{p}_j$ – we've seen this kind of notation before for estimates of probabilities for Bernoulli trials. We notice that

$$O_{1,j} \text{ has approximate distribution } N(n_1\hat{p}_j, n_1\hat{p}_j\hat{q}_j),$$

and similarly (but reversed) for $O_{2,j}$. Using independence, and our formulas for expected value and variance of the linear combination of independent random variables, we get that

$$n_2O_{1,j} - n_1O_{2,j} \text{ has approximate distribution } N(0, (n_2^2n_1 - n_1^2n_2)\hat{p}_j\hat{q}_j).$$

We notice that $(n_2^2n_1 - n_1^2n_2)\hat{p}_j\hat{q}_j = nn_1n_2\hat{p}_j\hat{q}_j$, and then an argument similar to the χ^2 goodness-of-fit statistic gives us that X has approximate distribution $\chi^2((2-1)\cdot(b-1))$. □

Proof for $a = 2, b = 2$. Proceed exactly as above for $a = 2$, to get down to

$$X = \frac{(n_2O_{1,1} - n_1O_{2,1})^2}{nn_1n_2\hat{p}_1} + \frac{(n_2O_{1,2} - n_1O_{2,2})^2}{nn_1n_2\hat{p}_2}.$$

We notice that $O_{1,1} + O_{1,2} = n_1$ and $O_{2,1} + O_{2,2} = n_2$, hence

$$n_2O_{1,2} - n_1O_{2,2} = n_2(n_1 - O_{1,1}) - n_1(n_2 - O_{2,1}) = -(n_2O_{1,1} - n_1O_{2,1}).$$

Using this, and putting over a common denominator, we get that

$$X = \frac{(n_2O_{1,2} - n_1O_{2,2})^2(\hat{p}_2 + \hat{p}_1)}{nn_1n_2\hat{p}_1\hat{p}_2} = \left(\frac{n_2O_{1,2} - n_1O_{2,2}}{nn_1n_2\hat{p}_1\hat{q}_1}\right)^2,$$

which (since $n_2O_{1,j} - n_1O_{2,j}$ has approximate distribution $N(0, nn_1n_2\hat{p}_j\hat{q}_j)$) is the square of an approximate standard normal, hence has an approximate $\chi^2(1)$ distribution, as desired. □