

Testing Independence: Another Example

Lab 02.25.05

This is data regarding the fate of passengers on the Titanic, originally collected by the British Board of Trade (see `help(Titanic)`, for more information). The data also has information on sex and age, but let's focus on two variables for now. Let's say we're interested in whether or not a passenger's class had any relationship with whether or not that passenger survived the trip. Were some classes more likely than others to survive? (Leo was in 3rd class and Kate was in 1st, and we all know – for better or worse – how that turned out.) Or were class and survival independent from one another?

The data looks like this. The total number of passengers, $n = 2201$. Let's assume the total number of passengers was fixed, but the row sums were not, for the sake of illustrating a test of independence.

These are our observed counts.

Class	Died	Survived	Sum
1 st	122	203	325
2 nd	167	118	285
3 rd	528	178	706
Crew	673	212	885
Sum	1490	711	2201

In order to test independence, we assume independence for the time being, and calculate the marginal probabilities: the probability of being in 1st class, in 2nd class, etc; the probability of dying and surviving. So the original cell entries are no longer important (until later).

These are just the marginals, or the sums in the lower and right margins, divided by the total n . So the marginal probabilities associated with class result from the following calculations:

$$325/2201 = 0.14766, \quad 285/2201 = 0.1295, \quad \text{etc.}$$

Marginal probabilities

Class	Died	Survived	Sum
1 st			0.14766
2 nd			0.129487
3 rd			0.320763
Crew			0.40209
Sum	0.676965	0.323035	1

Under the assumption of independence, the maximum likelihood estimate for each of the cells is just the product of the marginal probabilities. These are our estimates of the conditional probabilities. So, the conditional probabilities for those that died result from the following calculations:

$$0.15 * 0.677 = 0.09996, \quad 0.129 * 0.677 = 0.088, \quad \text{etc.}$$

Conditional probabilities

Class	Died	Survived	Sum
1 st	0.099961	0.047699	0.14766
2 nd	0.087658	0.041829	0.129487
3 rd	0.217146	0.103618	0.320763
Crew	0.272201	0.129889	0.40209
Sum	0.676965	0.323035	1

Finally, in order to get our expected cell counts under the assumption of independence, we multiply by the total number of passengers, 2201. So, the expected number of counts for passengers who died result from the following calculations:

$$0.09996 * 2201 = 220, \quad 0.0877 * 2201 = 193, \quad \text{etc.}$$

These are our expected counts.

Class	Died	Survived	Sum
1 st	220.0136	104.9864	325
2 nd	192.935	92.06497	285
3 rd	477.9373	228.0627	706
Crew	599.114	285.886	885
Sum	1490	711	2201

From this, we can calculate the **Pearson** chi-square statistic.

$$\begin{aligned}
 \text{Pearson}X^2 &= \sum_{i=1}^k \frac{(n_i - \mu_i)^2}{\mu_i} = \sum_{i=1}^{\#\text{cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
 &= \frac{(220 - 122)^2}{220} + \frac{(203 - 105)^2}{104} + \frac{(167 - 193)^2}{193} + \dots \\
 &= 190.4011
 \end{aligned}$$

Since we're assuming that only the total number of passengers is fixed, what are the degrees of freedom of this statistic?

A handy shortcut for this type of case: **degrees of freedom = (nrows - 1)(ncol - 1)**

In this case, that's $3 * 1 = 3$

However, after all that work, R can do it for you . . .