Analysis of $r \times c$ Contingency Tables

Chi-Square Test for Independence

• Suppose we have a sample of *n* observations, each one classified according to two categorical variables.

• One categorical variable is called the row variable (has *r* categories) and the other is called the column variable (has *c* categories).

Model:

Then

Example: (snoring / heart disease)
• Random sample of 2484 British adults taken.
• Observed their snoring pattern and whether they had suffered from heart disease.
• Results given in a 2 × 3 contingency table:

		Snoring Pattern			
		Never	Occasionally	≈Every Night	
Heart Disease	Yes	24	35	51	110
	No	1355	603	416	2374
		1379	638	467	2484

Question: Does the incidence of heart disease depend on snoring pattern?

• Let π_{ij} denote the probability of an observation falling in cell (i, j).

Cell Probabilities for a $r \times c$ Contingency Table:
(r = # of rows, c = # of columns)

		Column Variable				
		1	2	…	c	
	1	π_{11}	π_{12}	…	π_{1c}	$\pi_{1\cdot}$
Row	2	π_{21}	π_{22}	…	π_{2c}	$\pi_{2\cdot}$
	⋮	⋮	⋮		⋮	⋮
Variable	⋮	⋮	⋮		⋮	⋮
	r	π_{r1}	π_{r2}	…	π_{rc}	$\pi_{r\cdot}$
		$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	…	$\pi_{\cdot c}$	1

Observed Cell Counts for a $r \times c$ Contingency Table

		Column Variable				
		1	2	…	c	Row Totals
	1	T_{11}	T_{12}	…	T_{1c}	$T_{1\cdot}$
Row	2	T_{21}	T_{22}	…	T_{2c}	$T_{2\cdot}$
	⋮	⋮	⋮		⋮	⋮
Variable	⋮	⋮	⋮		⋮	⋮
	r	T_{r1}	T_{r2}	…	T_{rc}	$T_{r\cdot}$
Col. Totals		$T_{\cdot 1}$	$T_{\cdot 2}$	…	$T_{\cdot c}$	n

Of interest: Are the two categorical variables independent, or are they associated?

H_0: The classifications are independent
H_a: The classifications are dependent

<u>Note</u>: Under H_0 (if the two classifications are <u>independent</u>),

→ If H_0 is true, the expected cell count for cell (*i, j*) is estimated by:

Our χ^2 test statistic is therefore:

where

• When *n* is large, under H_0,

• Large discrepancies between observed and expected counts lead to a _____ χ^2* and provide evidence _____ H_0.

We reject H_0 if:

<u>Rule of Thumb</u>: The large-sample assumption is satisfied if <u>every</u> (estimated) <u>expected</u> cell count is at least 5.

<u>Example</u>: Expected count for cell (1, 1):

All expected counts:

(see handout for details)

Test Statistic

R example: (`chisq.test` function). R gives P-value

Chi-Square Test for Homogeneity

• Suppose we take a random sample from each of *r* populations, and observe the same categorical variable (with *c* categories) for each sample.

Of interest: Is the set of cell probabilities the same for each population *i* = 1, 2, …, *r*?

Note: This differs from the test for independence because of a different sampling process.

• With the test for independence, we:

Here, we:

Again, let T_{ij} =

Example: A survey sampled 150 voters in Upstate SC, 150 in Midlands SC, and 150 in Coastal SC. Voters were asked whether they "approved", "were neutral", or "disapproved" of the governor's performance. The 3 × 3 table of results:

• Are the true proportions the same across the 3 populations?
• We test H_0: "(π_{i1}, π_{i2}, π_{i3}) are the same for $i = 1, 2, 3$".

• The mechanics of the test for homogeneity are exactly the same as for the test for independence!

Test statistic:

Reject H_0 if

Can calculate χ^{2*} by hand, or R gives:

Note: For the special case of $r = 2$, $c = 2$, the χ^2 test of homogeneity is equivalent to the z-test of H_0: $\pi_1 = \pi_2$ vs. H_a: $\pi_1 \neq \pi_2$.
Note: If the sample size is not large enough for the χ^2 test, we could: (1) combine two categories into one (not ideal) or (2) use Fisher's Exact test on the $r \times c$ table.