STAT 515 - Section 13.3 Supplement

Brian Habing - University of South Carolina

Last Updated: November 16, 2016

S13.3 - Chi-Square Test for Homogeneity

Section 13.3 discusses the *Chi-Square Test for Independence*, where the *n* observations are a random sample from a single population of interest (note that the grand total *n* is fixed in Table 13.3 and that a single population is the first condition in the box on page 787). The easiest way to check that you are performing a test of independence is that you know the grand total (*n*) in advance, and not the row totals (the r_i) or column totals (the c_j).

From the work on page 783 we see that the expected value for cell *ij* is $\hat{E}(n_{ij}) = R_i C_j / n$. We can also verify why the formula for the degrees of freedom is (*r*-1)(*c*-1). The basic formula for calculating the degrees of freedom is: # cells - # parameters estimated - # free cells.

1. The number of cells in the table is the number of rows times the number of columns, so *rc*.
2. The parameters that need to be estimated are the probabilities for each column and the probabilities for each row (we then multiply these to get the probabilities for each cell.) Note that we didn't need to estimate all *r* of the row probabilities. This is because we know that these probabilities have to add one, so if I know what *r*-1 of the probabilities are I can use subtraction to find the last one. Similarly we only need to estimate *c*-1 of probabilities for the columns. Thus the number of parameters estimated is (*r*-1)+(*c*-1).
3. To see that there is one free cell we need to notice that since we know the grand total is *n*, we don't need to know the count for the last cell to figure it out. We can just use *n* minus the total for all the other cells.

The degrees of freedom are thus *rc* - ((*r*-1)+(*c*-1)) - 1 = *rc*-*r*+1- *c*+1-1 = *rc*-*r*-*c*+1 = (*r*-1)(*c*-1).

Imagine that the values in Table 13.5 (page 785) were actually counts (and not percentages by Gender). We would get the expected values of 45.3, 45.3, 54.7, and 54.7, and would calculate that there is 1 degree of freedom. There would be something suspicious about this example though; they got exactly the same number of men and women. There is only a 5.7% chance of getting a perfect split like that for a sample of size 200. What probably happened in a case like this is that the researchers didn't get one large sample of size 200, instead they probably selected

100 men and 100 women. This would make sure that the sample was balanced in terms of gender. Unfortunately it violates the assumptions we used to construct the test of independence.

In this case then we are not testing that the assignment to the categories "male" and "female" is independent of the assignment to the categories "could identify" and "could not identify". Instead we are testing that the probability of being assigned to "could identify" and "could not identify" is the same for males and for females. We could write this null hypothesis as:

H_0: P("could identify" | Male) = P("could identify" | Female)
and P("could not identify" | Male) = P("could not identify" | Female)

We call these tests of sameness a *Test of Homogeneity*. (The book briefly mentions this kind of test on page 789 and calls it a *contingency table with fixed marginals*). The classical example of a test of homogeneity is to compare several rigged dice to see if they all work in the same way. In this case you know how many times you rolled each of the dice. If you used which die it was as the row in the table you would know the row totals then. If you used which die it was as the column total you would know the column totals. This tells you the key way of telling that you are performing a test of homogeneity: you know either the row totals or the column totals in advance.

A test of homogeneity is one in which there are random samples independently taken from several populations. The null and alternate hypotheses for tests of homogeneity can be written as:

H_0: The classification probabilities are the same for every population
H_A: The classification probabilities differ for at least two populations

In order to test these hypotheses we need to figure out how to get the expected values for the table and what degrees of freedom to use. (Below, we will set up our table so that the row totals are fixed. If your data has the column totals fixed instead just switch every mention of column and row below.)

TABLE S13.1 Contingency Table for Homogeneity of Rows

		Column (Classification)				Row Totals
		1	2	…	*c*	(Sample Sizes)
	1	n_{11}	n_{12}	…	n_{1c}	R_1 (fixed)
Row	2	n_{21}	n_{22}	…	n_{2c}	R_2 (fixed)
(Population)	⋮	⋮	⋮		⋮	⋮
	r	n_{r1}	n_{r2}	…	n_{rc}	R_c (fixed)
Column Totals		C_1	C_2	…	C_c	*n*

To get the estimated probability for any cell we would take the probability of being in that column (classification) times the number of observations (sample size) in that row: $\hat{E}(n_{ij}) = \hat{p}_j R_i$. To calculate these expected values we need to estimate the probability $\hat{p}_j$ of being classified into column *j*. According to the null hypothesis all of the populations have the same chance of being classified into each column. Because of this, if we assume the null hypothesis is true (which we always do when setting up a hypothesis test) we can estimate the probabilities by grouping all of the populations together. The best estimate for the chance of being in a column is thus the percentage of all of the observations that fall into that column: $\hat{p}_j = C_j/n$, which makes $\hat{E}(n_{ij}) = \hat{p}_j R_i = R_i C_j/n$. This is the same as for the test of independence.

To calculate the degrees of freedom we again use the equation:
degrees of freedom = # cells - # parameters estimated - # free cells.

1. As before, the number of cells in the table is the number of rows times the number of columns, so *rc*.
2. In this case we need to estimate one probability for each column (except for the last one as they must add to 1) so there are *c*-1 parameters estimated.
3. As we know each row total, we get one free cell in each row, so there are *r* free cells.

The degrees of freedom are thus $rc - (c-1) - r = rc-r-c+1=(r-1)(c-1)$ just like for the test of independence.

Even though the data comes from a different experimental set-up than the data for a test of independence (several populations instead of one) and the null hypotheses are different, the test of homogeneity is conducted in exactly the same way!