14.4 χ^2 Test of Association in Contingency Tables

- We may have two categorical variables that can be used to classify individuals.
- We can test whether these categorical variables are _____ or whether they are _____ ( ).

Example 3 (General Social Survey 2006):

- Is Marital Status associated with happiness?

		Marital Status		
		Married	Divorced/ Separated	Never Married
Happiness	Very/Pretty Happy	1320	467	603
	Not too Happy	93	119	127

- This two-way table giving counts for two types of classification is called a

_____ _____.

- Suppose our contingency table has r rows and c columns (r×c table).
- Let _____ be the _____ _____ and let _____ be the _____ _____.
- Let _____ be the probability an individual falls in the (i, j) cell: row i and column j.
- If the two classifications are <u>independent</u>, then

H_0:

H_a:

- If H_0 is true, each p_{ij} = _____ and each expected cell count is _____

- These probabilities must be estimated from the data:
- The MLE of any row probability is

- The MLE of any column probability is

- So the estimated expected cell counts are

- The test statistic has a similar form:

- Specifically,

- How many degrees of freedom?
- There are cells. The cell probabilities add to 1 so we

- We estimate

- In all, we have

- Reject H_0 if

- Same large-sample rule of thumb:

Example 3 again:

		Marital Status			
		Married	D/S	Never	
Happiness	Happy	1320	467	603	2390
	Not	93	119	127	339
		1413	586	730	2729

Expected Counts:

- Comparing the observed cell counts with the expected cell counts under independence, we see that married people tend to be _____ happy than would be expected if the classifications were independent, and the other types of people _____ happy.

14.5 Tests of Homogeneity in r x c Tables

- In Section 14.4, we considered measuring _____ categorical variables on a single sample of n individuals.
- We tested whether the two categorical variables were _____ or _____.
- Now, suppose we take independent samples from several populations, and measure the same categorical variable for each sample.
- The column categories are the different samples.

- The sample sizes are determined ahead of time, as usual, so the column totals are _____, not _____.

Example 4: Suppose we wish to compare the political ideologies of students at Clemson, U of SC, and College of Charleston. We will sample 50 students from each school and ask them to identify themselves as "Liberal", "Moderate", or "Conservative".

- Our null hypothesis is

		School			
		Clemson	U of SC	C of C	
	Liberal	13	21	16	
Ideology	Moderate	17	13	12	
	Conserv	20	16	22	
		50	50	50	150

- Note the column totals (sample sizes) are fixed beforehand, but the row totals are random.
- Under H_0, the MLE's of the expected cell frequencies are the same as in the χ^2 test for independence:

- To find the degrees of freedom, note there are probabilities.
- The sum of column probabilities must equal :

- This is linear constraints, so we lose degrees of freedom from these.
- Also, we estimate row probabilities ( ).
- So the d.f. for the X^2 statistic is

- So the mechanics of this test for homogeneity are exactly the same as with the test for independence.

Example 4:

Related Applications

(1) An interesting application of the two-way table is whether a binomial (or multinomial) population changes over time.

- We could measure binomial (or multinomial) counts over c time periods to get a $2 \times c$ (or $r \times c$) table.
- We could test whether the category probabilities differ over time with a χ^2 test.

Example: Has the proportion of American golfers in the Masters changed over time?

Hypothetical data (yearly counts):

	1990	1991	. . . .	2021	2022
American Internat'l					

② Note: If more than 2 categorical variables are measured for each sampled unit, we have a 3-way (or more) contingency table.

- This is typically analyzed with a log-linear model.

③ Finally, in some applications, both the row totals and column totals are fixed.

Example: It is known that 12 White applicants, 8 African-American applicants, and 10 Asian applicants have applied to a program. The program has room to accept 10 applicants and must reject the remaining 20.

Question: Is acceptance status associated with applicant's race?

- Fisher's Exact Test is an approach to test for association when both row and column totals are fixed.
- Fisher's Exact test may be used as a small-sample alternative to the χ^2 test even when both margins are not fixed.