Inferences Concerning Contingency Tables

Suppose we asked 300 students whether he or she preferred taking liberal arts courses in the area of math-science, social science, or humanities. The results are broken down in the following chart.

	Favourite Subject Area			
	Math-Science	Social Science	Humanities	
Geneder	(MS)	(SS)	(H)	Total
Male (M)	37	41	44	122
Female (F)	35	72	71	178
Total	72	113	115	300

We're interested in know if preference in a subject area is related to gender. A better way of asking this is "is preference for math-science, social science, or humanities independent of the gender of the college student?"

Phrasing the question like this is good because we have a precise definition for independence. For example, we want to know if P(MS|M)=P(MS|F)=P(MS), the chances of a student choosing MS as their favourite is independent of gender.

Let's make some hypotheses:

H_0

H_a

We'll complete the hypothesis test using the 0.05 level of significance.

Notice that the statement P(MS|M)=P(MS|F)=P(MS) is really a statement about the first column. We assume the null hypothesis and therefore this statement is true and so a good estimate for P(MS) is

Again, since we are assuming P(MS|M)=P(MS|F)=P(MS) we should expect the frequency of males that said MS to be

And frequency of females that said MS to be

Expected values

	Math-Science	Social Science	Humanities	
Geneder	(MS)	(SS)	(H)	Total
Male (M)	$\frac{72}{300}\cdot 122 = 29.28$	$\frac{113}{300}\cdot 122 = 45.95$	$\frac{115}{300}\cdot 122 = 46.77$	122
Female (F)	$\frac{72}{300}\cdot 178 = 42.72$	$\frac{113}{300}\cdot 178 = 67.05$	$\frac{115}{300}\cdot 178 = 68.23$	178
Total	72	113	115	300

We want to test how far away our observed frequencies are from our expected frequencies. The probability of getting a sample as extreme as our sample is given by a χ^2 as long as each expected value is greater than 5.

So our test statistic is

$$\chi^2 = \sum \frac{f_{ij} - e_{ij}}{e_{i}j}$$

But what is $\chi^2(\alpha)$? It depends on the degrees of freedom. One way of thinking of the degrees of freedom is as the number of cells that may be filled freely when you know the totals of each row and column. For example

	Math-Science	Social Science	Humanities	
Geneder	(MS)	(SS)	(H)	Total
Male (M)				122
Female (F)				178
Total	72	113	115	300

we can see that we are only free to fill two cells, the rest are then determined.

In this case $df =$

An arrangement of data in a two way classification like we have been using is called a **contingency table.**

If we have a contingency table with size r x c the number of degrees of freedom is determined by

Back to our hypothesis test!

We have $df = 2$ so to use the classical approach

$$\chi^2(\alpha) = \chi^2(0.05) = 5.99$$

Example: Test whether the following data to determine whether voter preference is independent of gender. Use $\alpha = 0.05$.

	Women	Men
Liberals	180	143
Con.	130	141
NDP	92	71
BQ	84	96
Other	39	24

Example: A television commercial was shown in two different test regions 6 times in one week. A week later a survey was conducted in each region to identify peoplle who would remember the primary message of the commercial.

	Saw in region 1	Saw in region 2
Remembered message	63	60
Forgot message	87	140

Test at 5% if there is a difference in proportions of people who remember the commercial in region 1 or region 2.

A **test of homogeneity** is a similar test with the difference being that the row or column totals are controlled by the experimenter.

Example: 200 voters in urban areas, 200 in suburban and 100 in rural areas were asked whether or not they supported a new proposal by their governor.

	Favour	Oppose	Total
Urban	143	57	200
Suburban	98	102	200
Rural	13	86	100

Test at 5% significance if voters within different residence groups have different opinions about the governor's proposal.