The Distribution of Science Majors by Gender: An Analysis of Categorical Data

Rebecca Paul

The Question

Women make up almost half the workforce, but occupy only 22% of science, engineering, and technology jobs (Hanson, Schaub, Baker, 1996). This differential may deprive society of scientific talent that could be put to better use. Some of the gender difference in the workforce stems from the choices that students make in college; if a woman decides not to major in a scientific field, it will be harder for her to move into a technological area. In fact, much of the gender difference in earnings comes from the choice of major (Jacobs 1996). Because half of all college students change majors (Jacobs 1996), college is an important time to try to retain or even gain women with interest in the sciences, so that they may go on to have scientific careers. For this reason, we look at how women and men are divided in science majors.

Women have made progress in some of the sciences in the last thirty years, but this progress has not been uniform and has slowed down over time. In 1960, 13.7% of degrees awarded to women were in sciences and engineering. This number increased to 22.7% in 1976 and has stayed constant since then. At the same times, 44.4% and then 38.9% of degrees awarded to men were in science and technology (Barber 1995). However, much of this desegregation of majors has occurred in medicine, biology, law and business, while engineering and the physical sciences have stayed male-dominated (Davies and Guppy 1997). The gender divide between life sciences such as biology and medicine, and physical sciences widened between 1976 and 1989, even after controlling for SAT scores (Turner and Bowen, 1999).

By understanding the breakdown of genders in the sciences, perhaps we can better motivate future women to enter all of them. This paper looks at two years of data on the distribution of male and female science majors to do this.

The Data

To study the distribution of science majors, we consider the number of natural science majors of each gender at Swarthmore College in the graduating classes of 2000 and 2001. The data are reproduced in a $7 \times 2 \times 2$ contingency table below:

(2000)

Department	Female Majors	Male Majors	Total Majors
Biology	24	10	**34**
Chemistry	5	7	**12**
Computer Science	1	6	**7**
Engineering	10	11	**21**
Mathematics/Statistics	2	5	**7**
Physics/Astronomy	3	3	**6**
Special Major	5	7	**12**
Total	**50**	**49**	**99**

(2001)

Department	Female Majors	Male Majors	Total Majors
Biology	25	9	**34**
Chemistry	0	2	**2**
Computer Science	0	15	**15**
Engineering	7	18	**23**
Mathematics/Statistics	0	7	**7**
Physics/Astronomy	3	8	**11**
Special Major	5	5	**10**
Total	**40**	**62**	**102**

(Total)

Department	Female Majors	Male Majors	Total Majors
Biology	49	19	**68**
Chemistry	5	9	**14**
Computer Science	1	21	**22**
Engineering	17	29	**46**
Mathematics/Statistics	2	12	**14**
Physics/Astronomy	6	11	**17**
Special Major	10	12	**22**
Total	**90**	**113**	**203**

These data come from the office of Institutional Research at Swarthmore College. Each cell contains the number of majors of that gender in that department. This means that people with two majors, both in the sciences, are counted twice, once for each department. To simplify the analysis, we ignore this, instead considering the total number of majors, instead of the total number of people in the sciences.

Notice that these are the true values for the majors in these two years. However, we may consider the number of majors in a year as realizations of a random variable. Thus, we may compute p-values for the parameters of this process.

Analysis of One Variable

We first consider the sciences as a whole, to test whether men and women major in the natural sciences equally often. Pooling the two years, we find that we have 90 female majors and 111 male majors. We test the hypothesis that majors are equally likely by testing the probability of seeing 90 female majors if the number of female majors is distributed according to the binomial distribution, Binomial(203, 0.5); this hypothesis conditions on the total number of majors we observed. With this distribution, we find a two-sided p-value of 0.09. Thus, if majors are equally likely to be male or female, a distribution this far or further from equal will occur only 9% of the time. We do not reject the two-sided hypothesis that males and females are equally likely to major in the sciences overall, though this result does cast doubt on it.

We may also find approximate confidence intervals for the true proportion of female science majors in general. Using the normal approximation to the binomial distribution, with a conservative estimate of the variance, $\sigma^2 = (0.5)(1 - 0.5)/n$, we find the 95% confidence interval:

$$90/203 \pm \text{sqrt}(1.96(0.5)^2/203) = (0.39, 0.49)$$

Notice that this confidence interval does not contain 0.5, though it is quite close. This occurs because we are using an approximation to this confidence interval, so it is not the confidence interval that would be found using the binomial distribution.

The two years taken individually yield different results. In 2000, the numbers of male and female science majors were basically equal, which yields a two-sided p-value of 0.84. We have no reason to reject the null hypothesis for the class of 2000. However, in 2001, women were less than 40% of all science majors. Under the same null hypothesis of equality, we find a two-sided p-value of 0.04, which gives evidence against the null hypothesis. This suggests a possible difference between the two years. We will model this difference in coming sections.

The Simplest Contingency Table Analysis: Homogeneity

We now consider the two-dimensional contingency table of gender and major. The most important question in a contingency table is whether the two categories are related in any way. To measure this, we test whether being a certain gender and choosing a certain major are independent. Mathematically, this is equivalent to:

$$P(\text{gender} = j \text{ and major} = i) = P(\text{gender} = j)P(\text{major} = i),$$

where i might be any of the seven majors and j might be either male or female. If this equation holds for each cell in the table, then we say the table is homogenous. The null hypothesis of homogeneity for a contingency table is:

$$H_0: \pi_{ij} = \pi_{i+}\pi_{+j}, \text{ for all i and j.}$$

We test for homogeneity by finding the difference between the observed probabilities, p_{ij}, and the probabilities we would see if homogeneity held, $p_{i+}p_{+j}$.

One way to make the comparison between the two sets of probabilities is by estimating the expected number of entries in each cell if homogeneity held. This is $p_{i+}p_{+j}Y_{++}$, which can be simplified to $Y_{i+}Y_{+j}/Y_{++}$. Since we now have observed and expected probabilities for each cell, we use Pearson's Chi-Square Test:

$$X^2 = \sum_{j=1} \sum_{i=1} (O_{ij} - E_{ij})^2/E_{ij}.$$

The distribution of X^2 is χ^2, with (I-1)(J-1) degrees of freedom, because we are assuming that the row and column totals are fixed; this means that once we have chosen what numbers are in I-1 of the rows and J-1 of the columns, we may figure out the rest of the totals by ensuring that each row or column adds up to the correct number.

We test for homogeneity in the contingency tables for 2000, 2001, and the table with both years combined. Pearson's Test yields the following test statistics:

2000: $X^2 = 11.33$
2001: $X^2 = 65.48$
Both: $X^2 = 42.12$

The 5% cutoff point for the χ^2 distribution with (7-1)(2-1) = 6 degrees of freedom is 12.59. Therefore, we reject the hypothesis of homogeneity for the total distribution over both years and for 2001, but not for the class of 2000. The p-values are:

2000: $p = 0.079$
2001: $p < 0.001$
Both: $p < 0.001$

In order to see which majors contribute most to the test statistic, we reproduce the table containing both years with $X^2 = (O_i - E_i)^2/E_i$ in each cell below:

Department	Female Majors	Male Majors	Total Majors
Biology	11.30	9.67	**20.98**
Chemistry	0.26	0.16	**0.42**
Computer Science	7.95	6.02	**13.98**
Engineering	0.37	0.73	**1.11**
Mathematics/Statistics	2.91	2.17	**5.07**
Physics/Astronomy	0.34	0.22	**0.56**
Special Major	0.002	0.01	**0.01**
Total	**23.13**	**18.99**	**42.12**

The cells with the largest entries are those in biology, followed by those in computer science. If one looks at the original data, one can see that biology has the largest percentage of women (49 of 68 majors – 72% – are female) and computer science has the smallest percentage of women (1 of 22 majors – 5% – is female). It seems reasonable that the most extreme majors would contribute more to the test statistic.

We may also perform this test on the total number of male and female science majors for the two years, to see if they are homogenous. In this case, $X^2 = 2.98$ is not significant in the χ^2_1 distribution, and in fact has a p-value of 0.08. This means that we do not reject the null hypothesis that the overall proportions of male and female science majors changed from 2000 to 2001.

Breaking Down Tables

We may break down a large contingency table to better understand the data. Because the distribution of χ_k^2 is equal to the sum of k independent χ_1^2 variables, we may consider the contingency table as k distinct 2×2 contingency tables, using the method outlined in Iversen (1979). In order for the test statistics to add up exactly, we must use the maximum likelihood form of the chi-square statistic:

$$L^2 = 2 \sum O_{ij} \ln(O_{ij}/E_{ij})$$

The asymptotic distribution of L^2 is also χ_k^2, so that our results should not be changed by using L^2 instead of X^2.

We focus on the table with both years combined for this analysis.

One way of thinking of the sciences at Swarthmore is to divide them into the theoretical sciences, such as biology, chemistry, physics, and math, and the applied sciences, such as engineering and computer science. Special majors might be either, so we keep them as a separate category. We can then break up the theoretical sciences into biology and everything else, since biology has already been shown to be a great contributor to the lack of homogeneity. Then, we compare mathematics to physics and chemistry. The final tables, then, must have computer science and engineering in one and chemistry and physics in the other.

The tables that result from this decomposition are reported here, along with the maximum likelihood chi-squared (L^2) statistic for the table, its p-value, and the percentage of the overall chi-square statistic that it makes up:

Department	Female Majors	Male Majors	Total
Biology	49	19	**68**
Other Theoretical Sciences	13	32	**45**
Total	**62**	**51**	**113**

$L^2 = 20.91$, $p < 0.0001$, 44% of non-homogeneity

Department	Female Majors	Male Majors	Total
Theoretical Sciences	62	51	**113**
Applied Sciences	18	50	**68**
Total	**80**	**101**	**181**

$L^2 = 14.30$, $p = 0.0002$, 30% of non-homogeneity

Department	Female Majors	Male Majors	Total
Computer Science	1	21	**22**
Engineering	17	29	**46**
Total	**18**	**50**	**68**

$L^2 = 9.86$, $p = 0.002$, 21% of non-homogeneity

Department	Female Majors	Male Majors	Total
Chemistry & Physics/Astro	11	20	**31**
Math/Stat	2	12	**14**
Total	**13**	**32**	**45**

$L^2 = 2.30$, $p = 0.13$, 5% of non-homogeneity

Department	Female Majors	Male Majors	Total
Special Major	10	12	**22**
All Other Sciences	80	101	**181**
Total	**90**	**113**	**203**

$L^2 = 0.01$, $p = 0.91$, 0.03% of non-homogeneity

Department	Female Majors	Male Majors	Total
Chemistry	5	9	**14**
Physics/Astronomy	6	11	**17**
Total	**11**	**20**	**31**

$L^2 = 0.0006$, $p = 0.98$, 0.001% of non-homogeneity

Thus, we see that the difference between biology and the other theoretical sciences is the biggest contributor to the differences in the overall contingency table, followed by the difference between the theoretical sciences and the applied sciences; women are more likely to major in theoretical sciences than in applied sciences. This result agrees with other findings, such as Camp (1998), that found women less likely to major in applied sciences than in theoretical sciences. The difference between engineering and computer science is also significant. Other test statistics are both much smaller and statistically insignificant.

2 x 2 Tables: Measures of Association

In order to understand how closely categorical variables are related, we consider one measure of association, the odds ratio. The odds ratio does not change when the rows and columns are

interchanged. This means that the odds ratio does not assume that one variable causes the other. This differentiates the odds ratio from other measures of association, like relative risk. In addition, the odds ratio has a more intuitive explanation than some measures of association, like Yule's Q and Yule's Y. We now offer this explanation.

If the two categorical variables are associated, then the odds of being in a certain row depends on which column contains the observation. To describe how the odds change, we consider the ratio of the odds:

$$OR = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}.$$

As always, we do not know the probability of falling into any given cell. Therefore, we estimate π_{ij} by p_{ij} again, to find the estimated odds ratio:

$$OR = p_{11}p_{22}/p_{12}p_{21}$$

If the estimated odds ratio is much greater than 1, then the odds of being in the first column given that one is in the first row is OR times greater than the odds of being in the first column given that one is in the second row.

We now find the odds ratios for the three tables that test as being significantly not homogenous, since they contain approximately 95% of the non-homogeneity. To find the strength of the relationship, we estimate the odds ratio for each:

Biology vs. Other:	OR = 6.35
Theoretical vs. Applied:	OR = 3.38
Computer Science vs. Engineering:	OR = 1/12.31

Thus, we see that the difference between computer science and engineering is stronger than the difference in the other tables, despite the fact that the sub-table containing them contributes less to the overall chi-square statistic. These odds ratios show us that being female increases the odds of majoring in a theoretical science by a factor of 3. Once a person has chosen between the theoretical and applied sciences, being female increases the odds of majoring in biology, given that one is majoring in a theoretical science, another 6 times, while being male increases the odds of majoring in computer science, given that one has chosen to major in an applied science, another 12 times.

Tables in More Dimensions: Homogeneity Again

We now use the data from all three categories. As before, we test for homogeneity before proceeding with our analysis. We now define homogeneity, also called complete independence, by the null hypothesis,

$$H_0: \pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k},$$

which states that being in a certain row is independent of what column or layer one is in, and that being in a certain column is independent of layer. Using the estimation methods of before, we find that the expected count in one cell, E_{ijk} is given by:

$$E_{ijk} = p_{i++}p_{+j+}p_{++k}Y_{+++} = (Y_{i++}Y_{+j+}Y_{++k})/Y_{+++}^2$$

Pearson's chi-square test can be used in this situation as well; X^2 is now distributed $\chi^2_{(I-1)(J-1)(K-1)}$.

When we use this method in our $2 \times 2 \times 7$ table, we find:

$X^2 = 62.17$

$p < 0.001$

We reject the null hypothesis of overall homogeneity. However, this does not give us information about what variables are causing this hypothesis to be rejected. To understand that better, we need a more detailed model.

Log-Linear Models

One model that describes contingency tables more explicitly is a log-linear model. In the saturated form of this model, the count in each cell, Y_{ijk}, is described by an overall table effect, T, the effect of being in its row, R_i, the effect of being in its column, C_i, the effect of being in its layer, L_i, the two-way interactions of each pair of the row, column, and layer, $(RC)_{ij}$, $(RL)_{ik}$, and $(CL)_{jk}$, and the three-way interaction of the cell's row, column and layer, $(RCL)_{ijk}$. We may consider the cell count as the product of these effects:

$$Y_{ijk} = T(R_i)(C_i)(L_i)(RC_{ij})(RL_{ik})(CL_{jk})(RCL_{ijk}),$$

where we assume $\Pi(R)_i = \Pi(C)_i = \Pi(L)_i = \Pi(RC)_{ij} = \Pi(RL)_{ik} = \Pi(CL)_{jk} = \Pi(RCL)_{ijk} = 1$. This allows us to exactly describe all the cells in the table with exactly IJK parameters. In order to find deeper statistical insights, however, we remove some of these parameters, testing to see whether the model has lost a significant amount of explanatory power as we do so. This allows us to model a table more succinctly.

One way to test a model is by considering L^2. Under the null hypothesis that the model is correct, L^2 is distributed approximately as a chi-squared variable with the degrees of freedom equal to the number of cells less the number of parameters estimated. If the null hypothesis is rejected, the model is incorrectly specified, and needs more terms. However, failing to reject the null hypothesis does not mean that all work is done. It might be possible to remove more parameters without losing significant explanatory power.

The most common way to test whether more parameters can be removed is through hierarchical models. In these models, whenever an interaction term, say the interaction of row and column, is included, any terms which depend on the same variables, in this case the overall row term and the overall column term, are also included. This means that a model can be completely specified by giving the highest order interaction terms containing each variable. For example, the model {Row*Column, Layer} includes the terms {Overall, Row, Column, Layer, Row*Column}, while {Row*Column, Column*Layer} includes all of these terms and the interactions of column and layer as well.

Since we see that models are often contained in each other, we remove terms one at a time and compare the change in explanatory power, as measured by the change in maximum likelihood chi-square statistic, ΔL^2. We may consider $L_{old}^2 = \Delta L^2 + L_{new}^2$. Since L_{old}^2 and L_{new}^2 are both distributed as chi-squares and since they are independent under the hypothesis that the new model is correct, their difference is a chi-square variable as well. Thus, ΔL^2 also follows a chi-square distribution, with degrees of freedom equal to the difference of the degrees of freedom of L_{old}^2 and L_{new}^2.

These models are estimated by taking logarithms, so that we estimate:

$$\ln Y_{ijk} = \ln T + \ln R_i + \ln C_i + \ln L_i + \ln (RC)_{ij} + \ln (RL)_{ik} + \ln (CL)_{jk} + \ln (RCL)_{ijk}.$$

This simplifies the estimation process by making the model linear in the parameters.

We now consider a log-linear model for the entire $2 \times 2 \times 7$ table. Because there are zeroes in the table, we adjust by adding a fixed amount less than one (in particular, 0.5); this allows us to take logarithms of all the cells in the table. Because we are adding equal amounts to all cells, this should make results less significant than they would be otherwise.

We now find the results of applying different models to the data (formulas from Freeman (1987)):

Model	df	L^2	P-Value	Δdf	ΔL^2	P-Value
Major * Gender, Gender * Year, Major * Year	28	8.91	> 0.999	28	8.91	> 0.999
Major * Gender, Major * Year	32	20.41	0.94	4	11.49	0.02
Major * Gender, Year * Gender	42	29.84	0.37	14	20.93	0.10
Year * Gender, Year * Major	42	97.72	< 0.001	14	88.80	< 0.001
Major * Gender, Year	46	41.33	0.67	14	20.93	0.10
Major * Gender	48	116.35	< 0.001	2	13.78	0.32
Major, Gender, Year	60	130.14	< 0.001	14	88.80	< 0.001

We reject all models that exclude the major-gender interaction term, as well as the model without a year term. Thus, without comparisons, we see that there is a significant interaction between gender and major, just as we found in previous analysis. We also see that the overall number of majors changed over the two years.

If we consider the models hierarchically, we see that we may remove the overall interaction term and the year-major interaction term. However, we find that we cannot remove the major-gender interaction term at any stage of the process, just as we rejected any model that did not contain it. We also marginally reject removing the year and gender interaction term, even though the model without it is not rejected. This suggests that the interaction between year and gender exists, but is not very strong. In addition, we reject the model that removes the year term, though we do not reject the change in L^2 that it causes. Overall, then, we choose the {Major*Gender, Year} model.

The parameters estimated by these models show what these effects actually are. We report the parameters used in the two best models we chose, with those used only in the model with the Year*Gender italicized:

Variable	**Level**	**Female**	**Male**	**Main Effects**
Year	2000	*0.23*	*-0.23*	**0.16**
		(1.258)	*(0.795)*	**(1.169)**
	2001	*-0.23*	*0.23*	**-0.16**
		(0.795)	*(1.258)*	**(0.856)**
Major	Biology	0.84	-0.84	**1.12**
		(2.316)	(0.432)	**(3.056)**
	Chemistry	-0.10	0.10	**-0.66**
		(0.906)	(1.104)	**(0.518)**
	Computer Science	-0.84	0.84	**-0.56**
		(2.316)	(2.325)	**(0.570)**
	Engineering	0.13	-0.13	**0.79**
		(1.142)	(0.876)	**(2.220)**
	Mathematics and Statistics	-0.49	0.49	**-0.66**
		(0.612)	(1.637)	**(0.518)**
	Physics and Astronomy	0.16	-0.16	**-0.17**
		(1.173)	(0.853)	**(0.845)**
	Special Major	0.30	-0.30	**0.14**
		(1.355)	(0.738)	**(1.149)**
Main Effects		**-0.38**	**0.38**	**1.64**
		(0.683)	**(1.464)**	**(5.171)**

Thus, we see that some of the results we have seen before – being male raises one's chances of being a science major overall, and being female raises the chances of being a biology major and lowers the chances of being a computer science major. This also shows some other results which we did not identify before – once we control for the fact that fewer women are science majors, being female raises the odds of being an engineering or physics major as well, though these effects are much smaller than the effect in biology. In addition, this model gives us a better idea of the overall distribution of majors; once someone is a science major, it is more likely that he or she is a biology major or an engineering major than any other major.

Also, we notice that the Year*Gender interaction term shows that female science majors are more likely to have come from 2000 than 2001; this agrees with the previous results that the percentage of female science majors is significantly less than half in 2001, but not in 2000.

Conclusion

We have used various models to better understand two years of data on the relationship between major and gender. Whether we combine years in a 2×7 contingency table or consider them separately in a $2 \times 2 \times 7$ contingency table, we see that gender and major are not independent. In further analysis, we have seen that there is some divide between the theoretical and applied sciences, but that more of the difference is caused by individual majors, biology and computer science in particular.

Analysis of two years of observed data is not enough to determine a cause of these disparities. However, knowing where there are disparities is the first step to understanding them. Thus, this analysis may provide useful information about what to study next.

Bibliography

Barber, Leslie A. (March-April 1995): "U.S. Women in Science and Engineering, 1960-1990" *Journal of Higher Education*, Vol. 66, No. 2, pp. 213-234.

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland (1975), *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: The MIT Press.

Camp, Tracy (1998): "Computer Science Programs in Engineering Colleges = Fewer Females" *Journal of Women and Minorities in Science and Engineering*, Vol. 4, pp. 15-25.

Davies, Scott and Neil Guppy (June 1997): "Fields of Study, College Selectivity, and Student Inequalities in Higher Education" *Social Forces*, Vol. 75, No. 4, pp. 1417-1438.

Freeman, Daniel H. (1987), *Applied Categorical Data Analysis*. New York: Marcel Dekker, Inc.

Hanson, Sandra L., Maryellen Schaub and David P. Baker (June 1996): "Gender Stratification in the Science Pipeline: A Comparative Analysis of Seven Countries" *Gender and Science*, Vol. 10, No. 3, pp. 271-290.

Iversen, Gudmund R. (1979): "Decomposing Chi-Square: A Forgotten Technique" *Sociological Methods & Research*, Vol. 8, No. 2, pp. 143-157.

Jacobs, Jerry A. (1996): "Gender Inequality and Higher Education" *Annual Review of Sociology*, Vol. 22, pp. 153-185.

Knoke, David, and Peter J. Burke (1980): *Log-Linear Models*. Beverly Hills: Sage Publications (Series: Quantitative Applications in the Social Sciences).

Reynolds, H.T. (1984): *Analysis of Nominal Data*. Beverly Hills: Sage Publications (Series: Quantitative Applications in the Social Sciences).

Turner, Sarah E., and William G. Bowen (Jan 1999): "Choice of Major: the Changing" *Industrial and Labor Relations Review*, Vol. 52, No. 2, pp. 289-303.