Data Presentation and Analysis for the Gall Fly Laboratory
Winter 2006

Now that the class has collected data on the different protein forms of PGM present in our sample of gall fly larvae, you need to (a) summarize the data for your report in an effective way and (2) analyze the data to determine the answers to two of the questions this lab set out to address.

As a reminder, those questions were:

1. Is there variation in the population of gall flies?
2. Are the gall flies from the Arb different from the gall flies from McKnight Prairie?

Data Presentation

In order to present your data effectively, you will need to summarize the results. You can find the data in COURSES|Course Materials|Lab Materials.

You should use the class data to fill out the table below. **Total the values for all the lab sections.**

Source	SS	SM	SF	MM	MF	FF
Cowling Arboretum						
McKnight Prairie						

You will actually be comparing the allelic frequency, rather than the genotype frequency, between the two collection sites. Calculate the total number of each allele in each group of gall flies. (For example, if there are 8 MF larvae, 5 FF larvae, and 0 SF larvae from S. altissima, you know there are 8+(2x5)+0=18 F alleles.) You can fill these values in on the table below:

Source	S	M	F
Cowling Arboretum			
McKnight Prairie			

For the lab report, we would like you to make a bar graph containing the information from this second table. Make one set of bars for each collection site; remember that the independent variable should go on the x-axis, and the dependent variable should go on the y-axis. If you need help using Excel to make graphs, please contact your instructor or TA.

Data Analysis

Is there variation in the population of gall flies?

This is a fairly direct question; you basically want to indicate if the gene coding for the protein is polymorphic (has multiple forms) or not. If a gene is polymorphic in a population, it just means there are multiple alleles present in that population. You will not need any statistics to analyze your results for this question.

Are the gall flies from the Arb different from the gall flies from McKnight Prairie?

This question is really asking if the two rows of the second table on the previous page are really any different from each other. Do larvae from the two collection sites differ in terms of their PGM and GPDH alleles? In order to be sure, we will need to analyze the data using statistical analysis. The first important concept to understand about statistics is something called a "**p-value**." The p-value is a proportion, ranging from zero to one, and represents the likelihood that the data you have are not significant. A p-value of 0.95 means there is a 95% chance that your data look the way they do due purely to random events, not because there is some pattern to the data. A p-value of 0.01 means that there is only a 1% chance that your data can be explained by random events. In biology, it is generally agreed on that a p-value less than 0.05 (less than 5%) is "significant" and indicates a distinct, non-random pattern to your data (the particular pattern depends on the particular test you use). See pages 9-10 in the Lab Report Guide for more information about p-values.

In our case, the p-value will represent the likelihood that the two different groups of larvae are not different from one another. So a p-value below 0.05 will mean that the groups are significantly different. A p-value above 0.05 means you cannot claim the groups are different (although you can describe trends and call for more studies to see if the trends hold up with larger numbers of organisms).

The test we are using is called the **Chi Square Test for Independence**, also referred to as a contingency table analysis ("chi" is pronounced "ki" as in "kite"). The basic idea is that we will compare the distributions of alleles in the two groups by first determining what the distribution would be if the populations were identical (the "Expected" value), and then seeing how different the actual, "Observed" groups are from that Expected value. In order to demonstrate the process you'll go through, we will work a (simpler) problem as an illustration.

Say we are interested in the different types of larvae some of you found in your galls, and we want to know if the frequency of finding certain larvae varies by the collection site. We might set up a table very much like the one on the previous page:

Source	**beetle**	**gall fly**	**wasp**
Cowling Arboretum	5	15	12
McKnight Prairie	15	20	3

In other words, we opened 32 galls from the Arb and found 5 beetles, 15 gall flies, and 12 wasps (same idea for McKnight). These numbers look pretty different for the two collection sites, but we don't really know if they are significantly different until we analyze them statistically. How do we do this?

Step 1: Find the total number of each row and column in the table.

Source	**beetle**	**gall fly**	**wasp**	**Row Totals**
Cowling Arboretum	5	15	12	5+15+12=32
McKnight Prairie	15	20	3	15+20+3=38
Column Totals	5+15=20	15+20=35	12+3=15	

Step 2: Find the grand total number for the data by adding up the column totals or the row totals in the lower right hand box (the number should be the same no matter whether you choose the column totals or the row totals).

Source	beetle	gall fly	wasp	Row Totals
Cowling Arboretum	5	15	12	32
McKnight Prairie	15	20	3	38
Column Totals	20	35	15	32+38=70

Step 3: Next, we will use these row and column totals to **determine the Expected value for each of our original boxes in our table**. The Expected value formula is: $\frac{\text{(column total x row total)}}{\text{grand total}}$

For example, in the upper left box, the beetle's column total is 20, and the Arb row total is 32. The Expected value is thus (20x32)÷70=9.143. Repeat this process for all the squares of the table:

Source	beetle	gall fly	wasp
Cowling Arboretum	5	15	12
(Expected)	(20x32)÷70=9.143	(35x32)÷70=16	(15x32) ÷70=6.857
McKnight Prairie	15	20	3
(Expected)	(20x38)÷70=10.857	(35x38)÷70=19	(15x38) ÷70=8.143

Why does this equation work to give us the Expected value for each box? Maybe writing the process out in words will help answer this question. Let's focus on the beetle larvae from the Arb (the upper left corner box). We are taking the total number of beetle larvae we found from both collection sites (20, or the column total) from all the galls we collected (70, or the grand total) and looking at that as a proportion: 20/70 goldenrod galls contain beetles. So far we haven't made any predictions about what how many of those beetles should come from each collection site. For the Expected value, we are going to assume that it is equally likely for beetle larvae to show up at either site. You might think that means each site should have 10 galls containing beetle larvae (half of the 20 we mentioned above). However, this ignores the fact that we looked at more galls from McKnight than galls from the Arb. Including the row total (32) in our equation, also over 70 total galls, takes into account that the proportion of galls from the Arb is slightly less than half the total. When you look at the Expected value for that box, 9.143, it is indeed slightly less than 10, which is consistent with this reasoning.

Step 4: Now that we know our Observed and Expected values, we will compare these using **the chi square statistic**. (Remember the Observed value is just the one from our original data table, where we recorded how what species were present in the galls.) To calculate this statistic, use the following equation:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

The funny x looking thing is just the Greek letter "chi." The other Greek letter, sigma, you might remember means "sum."

All this equation means is that for each of those boxes in your table, you calculate $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ and then add all the values together.

So, for our example, here are the chi square statistics in their boxes:

Source	beetle	gall fly	wasp
Cowling Arboretum	$(5-9.143)^2 \div 9.143 =$ 1.877	$(15-16)^2 \div 16 =$ 0.063	$(12-6.857)^2 \div 6.857 =$ 3.857
McKnight Prairie	$(15-10.857)^2 \div 10.857 =$ 1.581	$(20-19)^2 \div 19 =$ 0.0526	$(3-8.143)^2 \div 8.143 =$ 3.248

And here is their sum: $\chi^2 = 1.877 + 0.063 + 3.857 + 1.581 + 0.0526 + 3.248 = 10.679$

Before we get to what this means, take a look at how much each box contributed to the total. When the Expected and Observed values were quite similar (like in the gall fly column), the value was low. As the difference between the Observed and Expected values increases, the chi square value increases. So, do you think a high chi square value will correspond to a high p-value, or a low p-value?

Step 5: Determine the degrees of freedom for your table. Basically, this takes into account the number of different boxes there are in your table (we won't go into the details here). There are two degrees of freedom for this table and also for the table you will use to analyze your data.

Step 6: Check to see if your chi square value corresponds to a p-value which is greater than 0.05 or less than 0.05. Statisticians have developed a table that correlates chi square values with p-values. (You can see one at http://www.richland.cc.il.us/james/lecture/m170/tbl-chi.html) In our example, and in the data analysis you will perform, **a chi square value of 5.991 corresponds to a p-value of 0.05**. In this example, if the chi square value is greater than 5.991, the p-value is less than 0.05 and there is a significant difference between the gall contents from the two different sites. If the chi square value is less than 5.991, the p-value is greater than 0.05, and there is no significant difference. (Note that we can only say they are not significantly different; we have not proven that they are they same.) Since our chi square value is greater than 5.991 (remember, it was 10.679), we can say confidently that the distribution of larvae in galls is different in the Arb than it is at McKnight.

Step 7: Report your statistical results in the Results section of your lab report. For the example we worked, an appropriate statement would be:

> We found that the distribution of larvae in galls from Cowling Arboretum was significantly different than the distribution of larvae in galls from McKnight Prairie ($\chi^2 = 10.679$, 2 df, $p<0.05$). Gall fly larvae were found in the greatest number at both sites. However, wasps were the next most common species in Cowling Arboretum, while beetles were more common than wasps at McKnight Prairie.

Note that you can not say (based on the chi square analysis) that there were significantly more wasps in the Arb than at McKnight—that is not what we tested. We only tested if the overall distribution of all three larval types differed between the two collection sites. But you can (and should) summarize the patterns in the results which contributed to the difference.

When you are ready to calculate the chi square statistic for PGM alleles, be sure to use the values in the second table on page 1. Then you can just repeat the steps we used in the example above to generate the chi square statistic and determine if the populations are significantly different.