BIOL 417: Biostatistics
Laboratory #4
22 February 2011

Binomial and Poisson Distributions, Goodness-of-Fit Tests, Contingency Tests

Generating Binomial and Poisson Distributions Using Excel

A Binomial Variable: sex ratio

Excel has several probability distribution tools built in to it, including the binomial and Poisson distributions. Recall that we have been thinking of *n* as the number of trials in an experiment using these distributions, and *X* as the number of successes out of those *n* trials; *p* is taken as the probability of a success on any given trial.

In general, you have tables of expected and observed frequencies of X successes out of n trials. The observed frequencies are the data that you generate or that are given to you, but the tricky part is to get the correct expected frequencies as specified by a null hypothesis; this is where Excel (or another spreadsheet program) can be quite helpful.

For example, to find the expected relative frequencies of X = # of males produced in families of 6 offspring that we discussed in lecture, start by typing the numbers 0,1,2,3,4,5,6 into Column C in Excel. In these examples, n = 6 and p = 0.5. Click on the cell next to "0" in Column D and click on the little "f_x" symbol next to the data entry line immediately below the tool bars; select "Statistical" in the "Select a category" pull down menu, and double-click "BINOMDIST" from the menu below.

The menus to be filled in are fairly self-explanatory: Number_s is the number of successes (X), trials is n, Probability_s is p; enter "False" in the Cumulative menu. Click the Hot Key for Number_s, and click on the cell containing 0; fill in the other menus directly, and click OK.

Now copy and paste the formula for the cell you just entered (which should return 0.015625) to the cells adjacent to 1, 2, 3, 4, 5, 6. Your table should look like this:

0	0.015625
1	0.09375
2	0.234375
3	0.3125
4	0.234375
5	0.09375
6	0.015625

In the example in lecture, we sampled 325 families; thus, to find the expected number of families out of those 325 who would have X number of male offspring, multiply 0.015625 x 325 in the cell next to 0.015625 using the " = " command that you learned before. Copy and paste this formula to the remaining cells. Your table should now look like this:

0	0.015625	5.078125
1	0.09375	30.46875
2	0.234375	76.17188
3	0.3125	101.5625
4	0.234375	76.17188
5	0.09375	30.46875
6	0.015625	5.078125

This last column (E) is the expected absolute frequencies of families of six offspring containing X males in 325 such families, assuming that the probability of any given offspring being male is 0.5. In the lecture examples, there were two studies, each of which yielded their own observed frequencies of X males in families of 6 offspring; enter these data into Columns F and H:

1	20
8	40
75	58
169	86
62	55
8	43
2	23

These are the observed frequencies for Study 1 and Study 2 respectively. To determine if either one of these is significantly different from the expected frequency distribution in Column E, calculate $(obs.-exp)^2/exp$ for each row in Columns G and I (e.g., $(1 - 5.078125)^2/5.078125 = 3.2749$). Sum the entries in Columns G and I; these are your chi-squared test statistic values for Study #1 and Study #2 respectively. Your table should look like this:

0	5.078	1	3.274928	20	43.84916975
1	30.469	8	16.5695	40	2.981389642
2	76.172	75	0.018033	58	4.335209578
3	101.563	169	44.77762	86	2.384795339
4	76.172	62	2.636738	55	5.884755343
5	30.469	8	16.5695	43	5.15363028
6	5.078	2	1.865712	23	63.252872
			85.71202		127.8418219

The final step is to find out the P value for the chi-squared test for each study. Click the cell next to 85.71 and click f_x then select CHIDIST from the Statistical pull-down menu. For X, click on the Hot Key and then the 85.71 cell, and enter 6 for the degrees of freedom (Deg_freedom). Click OK and 2.35×10^{-16} should appear; you should get a P value of 3.66×10^{-25} for the other test. These are outrageously small P values that are far less than 0.05, suggesting that in each Study, the observed frequency distribution differs very significantly from the expected frequency distribution.

A Rare Event: Using the Poisson distribution

In lecture, we looked at the occurrence of transposable elements in a series of 57 genes; each gene is potentially 1000s of bases long, but each transposable element (TE) is only 2 base pairs long. Thus, many TEs could fit into each gene, but the mean number of TEs per gene is relatively very low (2 – 3). So we consider the occurrence of a TE in a gene to be a rare event (n is large and p is very small). This changes how we calculate an expected frequency distribution for the random variable X, which is the number of TEs in a gene, compared with the male offspring studies above.

Assume you screen 57 genes for the occurrence of TEs and count the number of TEs in each gene. Enter the results of this study into Excel, starting with Column C:

0	12
1	14
2	11
3	7
4	4
5	4
6	2
7	2
8	1
9	0
10	0
11	0

Column C is the number of TEs observed (X), and Column D is the number of genes containing X TEs; you never found any more than 8 TEs in a single gene. For convenience, truncate this table to:

0	12
1	14
2	11
3	7
4	4
5	4
6	2
>6	3

We now need the expected number of genes out of 57 that we would have expected to contain X TEs if TEs were distributed at random through these genes. Because neither n nor p are specified in this kind of problem, you need to first calculate the mean number of TEs per gene; do this by calculating a weighted mean in Mystat as you did in the previous lab (should be 2.228). Then, click on the cell next to 12 in Column E and click f_x and select POISSON from the Statistical pull-down menu. For X, Hot Key in the cell with 0, for Mean enter 2.228, and enter "False" for the Cumulative menu and click OK. Copy and paste to remaining cells in the

column. You should see expected relative frequencies of X in Column E; multiply these by 57 in Column F to get the absolute expected frequencies. Your table should now look like this:

0	12	0.107744	6.14
1	14	0.240053	13.68
2	11	0.267419	15.24
3	7	0.198603	11.32
4	4	0.110622	6.31
5	4	0.049293	2.81
6	2	0.018304	1.04
>6	3	0.007962	0.45

Note that Column D should add to 57, Column E to 1.0, and Column F to 57; this is a good check to do to make sure your math and entries are correct. Now check to see if the observed and expected frequency distributions are the same by calculating the chi-squared test statistic for this table, and find the P value in Excel; you should get chi-squared = 25.1, and P = 0.000327 (remember that you have 6 degrees of freedom here, not 7; do you remember why??).

Contingency Analysis

Contingency analysis takes the goodness-of-fit approach to determine whether an association exists between two (or possibly more) categorical variables. The logic is exactly the same as in other goodness-of-fit tests, but generation of the "expected" frequency distribution is quite different.

Analysis revolves around recognizing a "row" variable and a "column" variable; it does not matter which one becomes which, but once you decide, it helps to keep those terms separate. For example, in the plant herbivory example in lecture, we had the following data showing whether plants were damaged or undamaged on three types of soil:

	Dry	Mesic	Wet	Totals
Damaged	17	42	71	130
Not damaged	46	12	10	68
Totals	63	54	81	198

Make sure you understand what the table is telling you: there were 198 plants observed, 130 were damaged, 68 were undamaged, and 63 were on dry soil, 54 were on mesic soil, and 81 were on wet soil. The question here is whether or not soil type and damage status are associated; that is, does the rate of plant damage differ in the different soil types?

The trick here is to come up with the expected frequencies that we "should" have seen in each cell if the two variables were not associated. To do that, the simplest thing is to enter the data into Mystat.

Open Mystat, go to the data spreadsheet, and name the first column "Soil_type," the second column "Status," and the third column "Count." Remember that Soil_type and Status need to be string variables. Then enter the data from the table above in the following format:

```
Dry    Damaged     17
Dry    Undamaged   46
Mesic  Damaged     42
Mesic  Undamaged   12
Wet    Damaged     71
Wet    Undamaged   10
```

In the original data table, the damage status of the plants is the row variable, and the soil type is the column variable. It's not essential that they be this way; you could have done it the other way around. But things are easier to follow if you remember which is which…

Remember that you need to tell Mystat if a variable represents the frequency of something occurring instead of being some measurement itself. Click the Data menu and select Case Weighting… then By Frequency… and click on and add Count to the Selected Variable window and click OK.

Click the Analyze menu, then select Tables, then Two-way… Click on Status and make it the Row Variable, and the click on Soil-Type and make it the Column Variable. In the "Tables" options below these windows, make sure that "Counts" and "Expected counts" are both selected. Click OK.

In the output window, you should see the original data table pop up as well as the table of expected frequencies, shown here:

	Dry	Mesic	Wet	Totals
Damaged	41.36	35.46	53.18	130
Not damaged	21.64	18.55	27.82	68
Totals	63	54	81	198

Following this is the chi-squared test of association showing the test statistic value (62.686), the degrees of freedom (2; remember that df is [#rows-1][#columns-1]), and the P value of the test (here much less than 0.05). Note: you can change the number of decimal places that show in the output window under the Edit menu: Options…. and the Output tab.

Note that you get exactly the same test result if you reverse the Row and Column variable designations.