**Week 7 – CPaT Stats: Categorical Variables & Contingency Tables (χ2 analyses)**

***Advanced work – Logistic regression and CART modeling***

**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Please enter your answers to the questions below in bold text, deleting all extraneous text; just leave the question and answer. Bring hardcopy of the completed lab report to next week’s lab; please upload your Excel file to** the program fileshare, as usual:

Workspace/\_StatsLabReports/week\_7-chi-square/

Wk7\_stats\_data\_cushingSkomraWeiss.xlsx (or .xls)

where you insert your last names in place of “cushingSkomraWeiss”

The chi-square distribution is used when you are interested in comparing some observed value from your data to some expected value, or when you are interested in testing whether two components are associated (like the cactus/shrub example from lecture). The chi-square test is a common statistical method for when your dependent variable is categorical (i.e., you are testing to see what determines mortality [Y/N] or what determines the presence/absence of a certain? species); however it can also be used when your dependent variable is continuous (see example below).

For example, Carri LeRoy, an Evergreen faculty member whose area is community ecology, uses chi-square tests to determine if mixtures of leaves decompose faster or slower than expected based on the decomposition rates of each litter type in isolation. She first calculates an average decomposition rate using litter bags that contain species in isolation and then compares that average to the decomposition rate found in mixed bags (Expected vs. Observed). This research design is similar to that of Tilman et al. (2001), but occurs in aquatic systems and is based on detrital food webs instead of plant production.

**Problem 1. protectedSpecies EXAMPLE from Class (and the book):**

Before you start this lab, review the chi-square calculations for protected species that we did in class, and make sure you understand it. You can find my calculations in the worksheet “tables”.

1. To make sure you understand how the habitat-fish and bird data were reorganized for JMP, please complete the following table for the protected species study. The objective is to have in your JMP table *count* number of rows for each of the four cases. You fill in count for each case in the table below

|  |  |  |  |
| --- | --- | --- | --- |
| species | Protected? | population stable/increasing | count |
| … | No | No |  |
| … | No | Yes |  |
| … | Yes | No |  |
| … | Yes | Yes |  |

1. Now fill in the appropriate rows on your spreadsheet in the “Protected species” worksheet, and run a chi-square in JMP – do your results match what we did in class? (if not, you probably don’t have the right number of rows!).

**Problem 2. Habitat-Fish Problem**

Using the salmon preference dataset that we reviewed in lecture, calculate a chi-square goodness of fit test by hand and then analyze the data in JMP to check your answers. Remember: if salmon have no preference for habitat type, we would expect to find them in all habitats proportal to the availability of that habitat [sand (50%), gravel (30%), silt (20%)].

3. What is your null hypothesis?

4. What is your research hypothesis?

5. Using the data from the Excel worksheet “tables,” create a table of observed and expected values and paste it below:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |



6. Calculate your chi-square statistic:

7. What is your critical chi-square (from the Chi-square-table.pdf)?

8. What are the degrees of freedom?

9. Do you reject your null hypothesis?

10. Interpret your results:

In JMP, let’s do the same thing. Paste the substrate data table from the worksheet labeled “habitat-fish” into JMP. Note how the data are organized differently in this worksheet. Click “Analyze” and click “Distribution” (Analyze – Distribution). You only have one choice for the Y-variable, substrate. Then ask JMP to create a mosaic plot for you. For this example, the variable was “substrate”, so you will see mosaic plot under the “substrate” tab once your distribution plot appears. Then ask JMP to “Test Probabilities” (also under the “substrate” tab). This is where you can enter the expected probabilities of numbers of fish seen in each habitat type. (Hint: if 50% of the substrate is sand, we would expect to find 50% of the fish in sandy habitats if the null hypothesis is true. Make sure you type the correct percentage in the correct cell!)

11. Do the degrees of freedom (dfs) and chi-square that you calculated by hand match the JMP output? If not, try to figure out where you went wrong.

**Problem 3. Forage-Bird Problem**

Another chi-square method involves the use of contingency tables. Using the worksheet labeled “tables” calculate a chi-square statistic in Excel to analyze seasonal bird-foraging. The data represent the numbers of birds found in trees, shrubs, and on the ground during spring vs. fall.

12. What is your null hypothesis?

13. What is your research hypothesis?

14. Your observed values are pasted below in the first table; fill in the second table below with expected values. Remember: to calculate an expected value for a contingency table you need to know each row and column total and the total sample size: Y-expected = (row total \* column total)/grand total

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Observed | # birds in trees | # birds in shrubs | # birds on ground | Row Totals |
| spring | 30 | 20 | 9 |  |
| fall | 13 | 22 | 26 |  |
| Column Totals |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Expected | # birds in trees | # birds in shrubs | # birds on ground |
| spring |  |  |  |
| fall |  |  |  |

15. Calculate your chi-square statistic by comparing the observed vs. expected values for each cell:

16. What is your critical chi-square (from table)?

17. What are the degrees of freedom?

18. Do you reject your null hypothesis?

19. Interpret your results:

Paste the data from the worksheet labeled “bird data” into JMP. Using the “Fit Y by X” function, enter forage as your Y variable and season as your X variable (because forage location is likely to depend on the season, the season does not depend on where birds forage, right?!).

20. Do your dfs (degrees of freedom) and chi-square match the JMP output? If not, try to figure out where you went wrong.

**Problem 4: Species-AgeClass association data from the 1kcs.**

Finally, we will analyze the association of Species and Age Class for the 1kcs. You can use either the contingency table method or JMP to determine if Age Class and Species are associated. This is a harder problem because you are being asked to take the data spreadsheet (see the worksheet: 1kcsSpeciesSite).

21. First, state your null and alternative hypotheses:

22. Which method (JMP or contingency table or both) did you use to do this problem?

23. If you used contingency tables, insert below the table(s) you developed, and say what is your critical chi-square (from table):

24. (Optional) If you used JMP, try to reverse engineer the contingency table from the JMP output, and insert that table below:

25. What is your chi-square statistic?

26. What are the degrees of freedom?

27. Do you reject your null hypothesis?

28. Interpret your results:

The rest of this lab is relatively advanced, but if you are doing *Machine Learning* you might find it interesting, since we will use statistical techniques on an Iris data set similar to what we used in Data Mining.

***Problem 5:* Classification and Regression Trees (*CART) model of the Iris Data Set:***

This part of the lab optional. If you are interested in learning about logistic regression and CART modeling, please continue; otherwise, consider yourself finished!

To try a CART model, paste the data from R. A. Fisher’s famous Iris data set[[1]](#footnote-1) (Advanced\_iris-data worksheet) into JMP. Choose “Analyze – Modeling” then “Partition” modeling (JMP calls a CART model “Partition modeling”). Choose “SPECIES” as your Y variable and enter all the continuous trait variables as ‘X’s. Now, ask JMP to “split” your tree – to make the first dichotomous divide.

29. Describe the result of your first split. What was your R-squared value?

Ask JMP to split the data again. You can use “split” to keep adding branches to your tree or “prune” to remove those branches. When to stop is somewhat up to you. Create your tree, then describe it:

30. How many splits did you decide to use?

31. What was the R-squared value for the tree?

32. How well were the Iris species virginica, versicolor and setosa described by this model?

33. Paste the tree below:

Now, for logistic regression you would have a categorical Y variable (just like all other cases today), but you would have a continuous X variable.

For example, in a study of rabbits exhibiting some type of infection, you might want to know what dose of penicillin results in a 0.5 probability of curing a rabbit. Think of Chester Bliss’ probit analysis here….

Using the data provided in the tab “Advanced\_iris-data” import the data into JMP. Using the “Fit Y by X” platform, use “response” as Y and “ln(dose)” as X. JMP will automatically fill in the column “count” as a frequency – make sure this happened, and if not, place “Count” in the frequency box.

34. What was your null hypothesis for this logistic regression?

35. What are the degrees of freedom, chi-square value and p-value?

36. Do you reject your null hypothesis?

37. Interpret your results:

1. As per UCI’s machine Learning Repository (archive.ics.uci/edu/ml/datasets/Iris): This is perhaps the best known database in the pattern recognition literature. Fisher's paper is a classic in both statistics and machine learning and is referenced frequently to this day. (e.g., Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1, p. 218.) The data set contains 3 classes of 50 instances each, where each class refers to an iris subspecies. One class is linearly separable from the other two; the latter two are NOT linearly separable from each other. Fisher,R.A. "The use of multiple measurements in taxonomic problems" *Annual Eugenics*, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950). [↑](#footnote-ref-1)