Ecology Counts!

Odalys Barrientos, Brianna Cirillo, Veronia Marquez

Statistical Methodology

In this analysis, contingency tables were used to asses how many journal entries came from each of the independent variables used. A contingency table, which can also be called a cross tabulation, is a table that shows the frequency distribution of each of the variables. Data cleaning was performed, thus removing any data where the independent variable being looked at was not applicable. Therefore, we separated the data by continent, country, region, state, and ecosystem. We looked at each of these tables to determine if the number of journal entries in each category, of these variables, were equal or close in frequency.

In order to better understand the distribution of the number of journal entries in each category, pie charts and bar graphs were made. This gave a visual representation of the distribution of journal entries in each independent variable. Therefore allowing for a visual analysis based on graphs and tables made.

To further analyze these categories, chi square tests were used to determine whether or not the observed amount of journal entries for each of the independent variables were equal. The Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table. An assumption of the test is that observations are mutually exclusive and independent. Therefore, data cleaning was done to ensure this condition was met. But the data was not randomly sampled, which goes against one of the assumptions. Thus, the p-values obtained do not have any relevance or any real meaning in relation to the data.

Being that the data consists of predominantly categorical variables, some additional data was added. Square mileage of continent, country, region, and state were researched, in order to better estimate the expected number of journal entries in each category.This allowed for chi square tests to be run with expected frequencies, that match the square mileage of each of the independent variables. This allowed for more accurate expected frequencies of journal entries from each independent variable. Data cleaning was done to remove rows that did not contain applicable data for each independent variable. The assumption of random sampling is still violated, therefore the p-values obtained could not be used to draw conclusions.

Results

a. Continent

b. Country

The variable country accounted for how many of these published articles completed the work done for the study in a particular country. Thus, we can see how many published articles from this collection of data have ecology work done in countries. To better model this, a contingency table was made to depict how frequently work done in a particular country was represented in the data set.

insert figure

From the figure above, an be seen that the country where the study was done, that had the highest frequency of articles was the United States. There were 68 journal articles where the work was in the United States. This is different from the number of articles with work done in other countries, it was more common that each only had one or two journal articles.

Using a chi squared test, we analyzed whether the probability of each country being represented in a journal article is the same. Being that the size of each country varies, it would not make sense for each country to have a

c. Region

d. State

The state category noted if and how many of these published articles completed their study in a particular state. This allowed us to see which states were more or less popular for ecology work. The following contingency table counts the number of times a state was represented in the data set.

Figure 1: Contingency Table

State	Count
AK	2
AL	2
AZ	3
CA	13
CO	2
FL	4
HI	1
IN	1
KS	2
MA	1
MD	1
MI	4
MO	1
MT	1
NC	1
NH	1
NJ	2
NM	1
NY	1
OH	1
OR	4
TX	3
UT	1
WA	3
WI	1
WY	2

From the table above, one can tell that most of the journal entries are published when the work was in California. The states that are not included in the table were never represented in the data set. If there is an assumption that the probability of each state being represented in a journal entry is the same as each state's square mileage then the larger states would have more published articles than the smaller states. Below, is the Chi-Square test that test if this statement is true.

From the Pearson residuals one can note that California has the most positive residual thus, the observed frequency exceeds the expected frequency. For a visual presentation, figure *3* helps explain this idea. The map on the left is what the map is expected to look like based off of how large and small each state is. The map on the right is what the map looks like when using the counts from the data set.

California is a large state and the majority of the published articles had work done in California. However, all the other states do not meet the assumption. The gray states represent the sates that were never counted in the data set. From the maps, one can conclude that the probability of each state being represented in a journal entry is not the same as the states square mileage. In other words, just because a state is bigger does not mean there are more published articles from that particular state.

e. Ecosystem

The ecosystem variable noted how many of the published articles used a particular ecosystem. Ecosystems were broken down into 3 categories, Marine, Terrestrial, and Freshwater. When looking at ecosystems, the idea was to see if one ecosystem was counted more than a different ecosystem. The following contingency table counts the number of times an ecosystem was represented in the data set.

Figure 2: Chi-Square Test: States

```
        Chi-squared test for given probabilities with simulated p-value (based on 2000
        replicates)

data:  x
X-squared = 140.37, df = NA, p-value = 0.0004998

  2.00     2.00     3.00    13.00     2.00     4.00     1.00     1.00     2.00     1.00
1.00     4.00     1.00     1.00     1.00     1.00     2.00     1.00     1.00     1.00     4.00
3.00     1.00     3.00     1.00     2.00     0.00     0.00     0.00     0.00     0.00     0.00
0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00
0.00     0.00     0.00     0.00     0.00     0.00     0.00
(10.257) ( 0.819) ( 1.781) ( 2.558) ( 1.627) ( 0.937) ( 0.101) ( 0.569) ( 1.286) ( 0.165) (
0.194) ( 1.513) ( 1.089) ( 2.298) ( 0.823) ( 0.146) ( 0.136) ( 1.900) ( 0.851) ( 0.700) (
1.537) ( 4.197) ( 1.327) ( 1.114) ( 1.023) ( 1.528) ( 0.831) ( 0.087) ( 0.039) ( 0.929) (
1.306) ( 0.905) ( 0.879) ( 0.631) ( 0.810) ( 0.553) ( 1.358) ( 0.757) ( 1.209) ( 1.728) (
1.105) ( 1.092) ( 0.720) ( 0.024) ( 0.500) ( 0.659) ( 0.150) ( 0.668) ( 0.379) ( 1.205)
[6.6e+00] [1.7e+00] [8.3e-01] [4.3e+01] [8.6e-02] [1.0e+01] [8.0e+00] [3.3e-01] [4.0e-01]
[4.2e+00] [3.4e+00] [4.1e+00] [7.3e-03] [7.3e-01] [3.8e-02] [5.0e+00] [2.5e+01] [4.3e-01]
[2.6e-02] [1.3e-01] [3.9e+00] [3.4e-01] [8.0e-02] [3.2e+00] [5.4e-04] [1.5e-01] [8.3e-01]
[8.7e-02] [3.9e-02] [9.3e-01] [1.3e+00] [9.0e-01] [8.8e-01] [6.3e-01] [8.1e-01] [5.5e-01]
[1.4e+00] [7.6e-01] [1.2e+00] [1.7e+00] [1.1e+00] [1.1e+00] [7.2e-01] [2.4e-02] [5.0e-01]
[6.6e-01] [1.5e-01] [6.7e-01] [3.8e-01] [1.2e+00]
<-2.578> < 1.305> < 0.913> < 6.529> < 0.293> < 3.163> < 2.830> < 0.571> < 0.630> < 2.056> <
1.831> < 2.022> <-0.085> <-0.856> < 0.195> < 2.234> < 5.048> <-0.653> < 0.161> < 0.358> <
1.986> <-0.584> <-0.284> < 1.787> <-0.023> < 0.381> <-0.912> <-0.294> <-0.197> <-0.964>
<-1.143> <-0.951> <-0.938> <-0.795> <-0.900> <-0.744> <-1.166> <-0.870> <-1.099> <-1.314>
<-1.051> <-1.045> <-0.848> <-0.155> <-0.707> <-0.812> <-0.388> <-0.817> <-0.615> <-1.098>

key:
        observed
        (expected)
        [contribution to X-squared]
        <Pearson residual>
```

Figure 3: Expected Frequency based on Square Mileage vs Observed Frequency

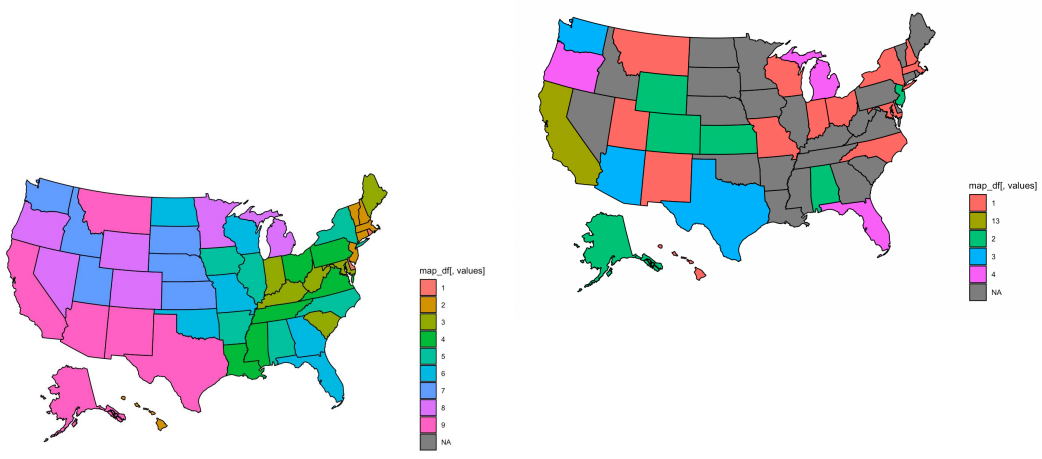


Figure 4: Contingency Table

Ecosystem	Counts
Marine	20
Terrestrial	127
Freshwater	12

From the table above, one can tell that most of the published articles have a terrestrial ecosystem. If there is an assumption that the probability of each ecosystem being represented in a journal entry is the same, then there should be the same number of counts for each ecosystem. Below, is the Chi-Square test that test if this statement is true.

Figure 5: Chi-Square Test: Ecosystem

```
        Chi-squared test for given probabilities

data:  x
X-squared = 155.58, df = 2, p-value < 2.2e-16

   12       20      127
(53.00)  (53.00)  (53.00)
[ 31.72] [ 20.55] [103.32]
<-5.63>  <-4.53>  <10.16>

key:
        observed
        (expected)
        [contribution to X-squared]
        <Pearson residual>
```

From the Pearson residuals one can note that terrestrial has the most positive residual thus, the observed frequency exceeds the expected frequency. Additionally,freshwater has the most negative residual thus, the observed frequency does not meet the expected frequency. For a visual presentation, figure *3* helps explain this idea. The pie chart on the left shows what the pie chart should look like if every ecosystem had an equal chance of being represented in a published article. While the pie chart on the right is what the pie chart looks like when using the counts from the data set.

If the ecosystems had a equal chance of being represented in these published articles these pie charts would look similar. However this is not the case and it is obvious that terrestrial takes up the majority of the pie chart.

If there is an assumption that the probability of each ecosystem being represented in a journal entry is the same as each ecosystem's square mileage then the larger ecosystem's would have more published articles than the smaller ecosystems. Below, is the Chi-Square test that test if this statement is true.

Terrestrial has the most positive residual and marine has the most negative residual. Figure *3* shows what the expected counts should look like and what the observed counts were.

These bar charts do not match up thus, we cannot conclude that the probability of each ecosystem being represented in a journal entry is the same as each ecosystem's square mileage.

Figure 6: Expected Frequency vs Observed Frequency

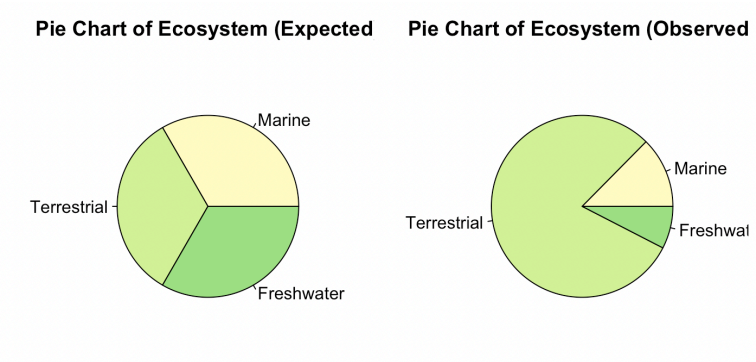


Figure 7: Chi-Square Test: Ecosystem

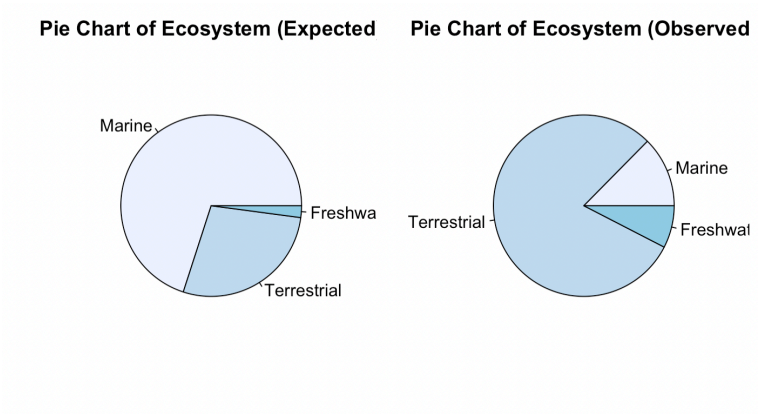
```
        Chi-squared test for given probabilities with simulated p-value (based on 2000
        replicates)

data:  x
X-squared = 251.31, df = NA, p-value = 0.0004998

   12       20      127
(  3.34) (111.30) ( 44.36)
[ 22.47] [ 74.89] [153.95]
< 4.74>  <-8.65>  <12.41>

key:
        observed
        (expected)
        [contribution to X-squared]
        <Pearson residual>
```

Figure 8: Expected Frequency based on Square Mileage vs Observed Frequency



Discussion

Due to the data not being random and there only being qualitative variables, there was not much statistical analysis to be done. For the next time the study is done, we have come up with two different ways to approach the study. If the research question was: Do publishing companies publish more articles with work done in their region, in comparison to work done outside of their region? The goal of using this question is to eliminate the need for data on journal articles that were not published. The study would focus on articles from different publishing companies, but all from the same year. The regions from which they published the most would be analyzed for any kind of bias. This would be an observational study looking at the counts of journal articles from each region.

Another approach would be to consider published articles that have significant results. The research question is: Is there a greater percentage of published articles that have a statistically significant p-value ($p < 0.05$) in comparison to those who do not? The goal of using this question would also be to eliminate the need for data on journal articles that were not published, due to the population being all published articles on ecology. This would be an observational study, that would be conducted by looking at different publishing companies from the same year and calculating the proportion of published articles that had a p-value less than 0.05. This would be the test statistic used to determine if there is any statistically significant difference in the proportion of published articles with significant p-values.

Although both of these approaches fix the need for unattainable data, they still do not fix meeting the randomness assumption. In order to fix this, the data collectors can look for articles on a database, with the criteria of the year they are looking for, the subject they want to look into, and any other specifics they would like. After getting the results of the search, they could use a random number generator to select which articles they will use for their study. This will correct the issue of the data not being randomly selected.