Ecology Counts!

Odalys Barrientos, Brianna Cirillo, Veronia Marquez

22 December 2020

1 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.

2 Results

2.1 Continent

The continent category demonstrated if any of the published articles, included in the study, completed their research in a particular continent. This allowed us to see which continents had more published articles about ecology. The following contingency table, seen in Table 1, counts the number of times a continent was represented in the data set. From Table 1, notice the number

Continent	**Count**
Africa	5
Asia	14
Australia	2
Europe	34
North America	94
South America	13
Oceania	5

Table 1: This contingency table, shows the number of journal articles that had work done on each continent.

of published articles is higher in North America in comparison to the other continents.

Under the assumption that the probability of each continent being represented in a journal entry is the same, then there should be the same number of journal articles for each continent. Figure 1 is a XChi-Square test that shows if this statement is true or not.

```
        Chi-squared test for given probabilities

data:  x
X-squared = 269.39, df = 6, p-value < 2.2e-16

     5        14         2        34        94         5        13 
(23.86)   (23.86)   (23.86)   (23.86)   (23.86)   (23.86)   (23.86) 
[ 14.91] [  4.07] [ 20.02] [  4.31] [206.23] [ 14.91] [  4.94] 
<-3.86>   <-2.02>   <-4.47>   < 2.08>   <14.36>   <-3.86>   <-2.22> 

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

.

Figure 1: This is the XChi-Square test for the continents.

From the Pearson residuals, notice that North America has the most positive residual. Thus, the observed frequency exceeds the expected frequency. Additionally, Australia has the most negative residual thus, the observed frequency does not meet the expected frequency. Through visual representation, Figure 2 clarifies this idea. The pie chart on the left shows what the pie chart should look like if every continent 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 continents had a equal chance of being represented in these published articles these pie charts would look similar. Unfortunately, this is not the case and it is obvious that North America takes up the majority of the pie chart.

If there is an assumption that the probability of each continent being represented in a journal entry is the same as each continent's square mileage. Then the continent with larger square mileage would have more published articles than those continent with a smaller square mileage. Figure 3, is a

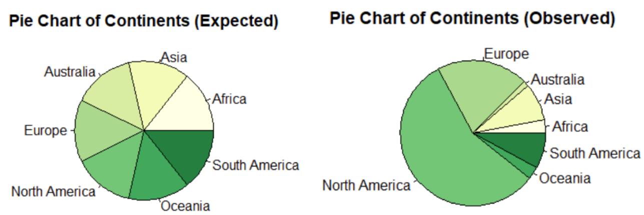

Figure 2: This is a comparison of what the expected values for continent and the actual values for continent look like as pie charts.

XChi-Square test that shows if this statement is true or not.

Looking into the Pearson residuals notice that North America has the most positive residual thus showing that the observed frequency exceeds the expected frequency. Through a visual representation, Figure 4 clarifies this idea. The bar graph on the left shows the expected number of journals for each continent, while taking into account the continent's square mileage. The bar graph on the right displays the observed number of articles from each

```
        Chi-squared test for given probabilities

data:  x
X-squared = 254.62, df = 6, p-value < 2.2e-16

     5       14        2       34       94        5       13
(35.26)  (51.73)  ( 8.93)  (11.82)  (28.67)  ( 9.89)  (20.70)
[ 25.97] [ 27.52] [   5.38] [ 41.65] [148.82] [   2.42] [   2.87]
<-5.10>  <-5.25>  <-2.32>  < 6.45>  <12.20>  <-1.56>  <-1.69>

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

Figure 3: This is the XChi-Square test for the continents, using the square mileage data.

continent.

Asia is the continent with the largest square mileage, thus giving it the largest number of expected journal articles based on it's square mileage. However it resulted in being one of the continent with the least number of journal articles. Asia and every continent besides North America did not meet the expected frequencies, therefore having less journal articles from each than predicted. From the bar graph, there is an obvious difference in the number of journal articles from each continent and the expected number based on the continents' square mileage.

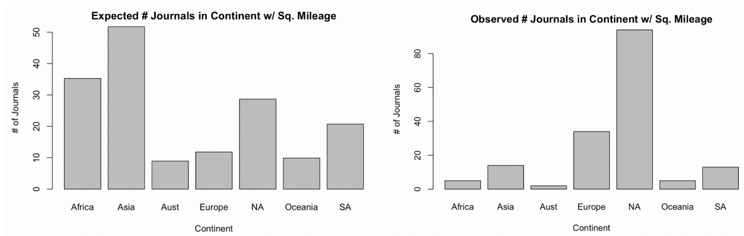

Figure 4: This is a bar graph comparing the expected and actual values for continent based on square mileage.

2.2 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 each of the 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.

From Table 2, it can 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 done for the articles, was done 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. This test was run with the assumption that every country had an equal chance. Since the data does not meet the normality assumption, the p-value for the chi-square test is virtually meaningless. Yet, the residuals can show how far away the data is from the expected values.

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

data:  x
X-squared = 995.96, df = NA, p-value = 0.0004998

    2        4        1        1        2        1        1        5
12       2        3        1        2        2        1        1
5        1        1        2        3        2        9        1
2        1        2        1        1        2        5        1
1        5        4        4        68
 (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)
(4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)
(4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)
(4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)   (4.38)
(4.38)   (4.38)   (4.38)   (4.38)   (4.38)
[  1.292] [  0.033] [  2.607] [  2.607] [  1.292] [  2.607] [  2.607]
[  0.088] [ 13.267] [  1.292] [  0.434] [  2.607] [  1.292] [  1.292]
[  2.607] [  2.607] [  0.088] [  2.607] [  2.607] [  1.292] [  0.434]
[  1.292] [  4.878] [  2.607] [  1.292] [  2.607] [  1.292] [  2.607]
[  2.607] [  1.292] [  0.088] [  2.607] [  2.607] [  0.088] [  0.033]
[  0.033] [924.477]
<-1.14>  <-0.18>  <-1.61>  <-1.61>  <-1.14>  <-1.61>  <-1.61>  < 0.30>
< 3.64>  <-1.14>  <-0.66>  <-1.61>  <-1.14>  <-1.14>  <-1.61>  <-1.61>
< 0.30>  <-1.61>  <-1.61>  <-1.14>  <-0.66>  <-1.14>  < 2.21>  <-1.61>
<-1.14>  <-1.61>  <-1.14>  <-1.61>  <-1.61>  <-1.14>  < 0.30>  <-1.61>
<-1.61>  < 0.30>  <-0.18>  <-0.18>  <30.41>

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

Figure 5: The xchi-square test above is analyzing whether the probability of each country in a journal article is equal. The expected numer of journal articles for each country is approximately 4. The expected values are in parentheses and the observed values are just above it.

From Figure 5, the expected number of articles from each country is approximately 4. Yet from the observed frequencies above, we can see that many countries only have 1 or two articles with work done there, but the

United States has 68. Therefore, we can say that the probability of each country being represented in a journal article might not be the same.

Being that the size of each country varies, it would not make sense for each country to have the same likelihood of being represented in a journal article. Thus, the square mileage of each country was researched in order to better estimate the probability of each country being represented in a journal article.

Country	**Count**
Argentina	2
Australia	4
Belgium	1
Bermuda	1
Brazil	2
Cambodia	1
Canada	5
China	12
Colombia	2
Costa Rica	3
England	1
Finland	2
France	2
French Guiana	1
Galapagos	1
Germany	5
Greece	1
Italy	1
Kenya	2
Mexico	3
New Zealand	2
Panama	9
Papua New Guinea	1
Peru	2
Portugal	1
PuertoRico	2
Republic of Congo	1
Singapore	1
South Africa	2
Spain	5
St. John	1
Svalbard	1
Sweden	5
Switzerland	4
The Netherlands	4
USA	68

Table 2: This contingency table shows the frequency of countries in journal articles. Not all countries are represented because there were not journal articles that had work done in every country in the data set.

2.3 Region

The region category exhibited if any or the amount of published articles completed their study in a particular region. This allowed us to see which region were more or less commonly seen for ecology work. The following contingency table counts the number of times a region was represented in the data set. From Table 3 notice the number of published articles is higher

Region	Count
$East - North\ Central$	7
$Mid - Atlantic$	3
$Mountain$	12
$New\ England$	3
$North\ West$	3
$Pacific$	24
$South\ Atlantic$	10
$West\ North\ Central$	3
$West\ South\ Central$	3

Table 3: This is the contingency table for the regions. It shows the number of journal articles that had work done on each region.

in the Pacific region in comparison to the other regions.

If there is an assumption that the probability of each region being represented in a journal entry is the same, then there should be the same number of counts for each region. Below, is a XChi-Square test that shows if this statement is true or not. From the Pearson residuals notice that the Pacific has the most positive residual thus, the observed frequency exceeds the expected frequency. Additionally, Mid-South and North West has the most negative residual thus, the observed frequency does not meet the expected frequency. Through visual representation, Figure 7 clarifies this idea. The pie chart on the left shows what the pie chart should look like if every region 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 regions had a equal chance of being represented in these published articles these pie charts would look similar. Unfortunately, this is not the case and it is obvious that Pacific takes up the majority of the pie chart.

```
        Chi-squared test for given probabilities

data:  x
X-squared = 68.373, df = 9, p-value = 3.171e-11

     7        3        1       12        3        1       24       10        3        3
 (6.70)   (6.70)   (6.70)   (6.70)   (6.70)   (6.70)   (6.70)   (6.70)   (6.70)   (6.70)
[ 0.013] [ 2.043] [ 4.849] [ 4.193] [ 2.043] [ 4.849] [44.670] [ 1.625] [ 2.043] [ 2.043]
< 0.12>  <-1.43>  <-2.20>  < 2.05>  <-1.43>  <-2.20>  < 6.68>  < 1.27>  <-1.43>  <-1.43>

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

Figure 6: A chi-square test was ran on the regions in order to see if each region was represented in a journal article an equal amount of times.

If there is an assumption that the probability of each region being represented in a journal entry is the same as each region's square mileage/ Then the region with a larger square mileage would have more published articles than those regions with a smaller square mileage. Figure 8, shows the Chi-Square test that proves if this statement is true or not.

Looking into the Pearson residuals notice that North West has the most positive residual thus showing that the observed frequency exceeds the expected frequency. Through visual representation, Figure 9 clarifies this idea. The bar graph on the left is the number of journals one expects to see given each region's square mileage. The bar graph on the right is what was observed using the counts from the data set. Also, due to unidentifiable region we couldn't locate square mileage for Mid-South. Due to this, we redid some data cleaning to remove this region. So our results do not include this specific region.

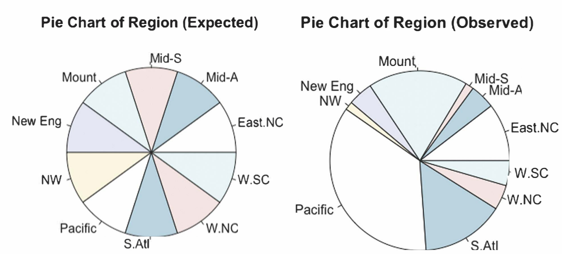

Figure 7: The pie charts above depict the expected and observed number of journal entries in each region.

```
	Chi-squared test for given probabilities

data:  x
X-squared = 26.334, df = 8, p-value = 0.0009206

    7        3       12        3        1       24       10        3        3
( 5.54)  ( 3.54)  (15.85)  ( 1.33)  ( 0.10)  (16.58)  ( 5.42)  ( 9.41)  ( 8.23)
[0.384]  [0.083]  [0.935]  [2.083]  [7.974]  [3.319]  [3.873]  [4.363]  [3.319]
< 0.62>  <-0.29>  <-0.97>  < 1.44>  < 2.82>  < 1.82>  < 1.97>  <-2.09>  <-1.82>

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

Figure 8: This is XChi-Square test on region, using square mileage in order to create expected values.

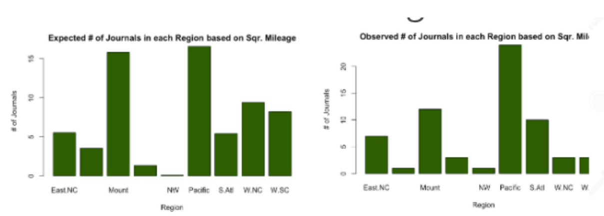

Figure 9: The bar graph above depicts the expected and observed number of journal entries in each region.

2.4 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 contingency table, in Table 4, counts the number of times a state was represented in the data set.

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

Table 4: The above table shows the number of journal articles that had research done in each state.

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. Figure 10, shows the Chi-Square test that tests if this statement is true.

```
        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 10: This is the chi-square test on the states, using square mileage to predict expected.

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 11 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

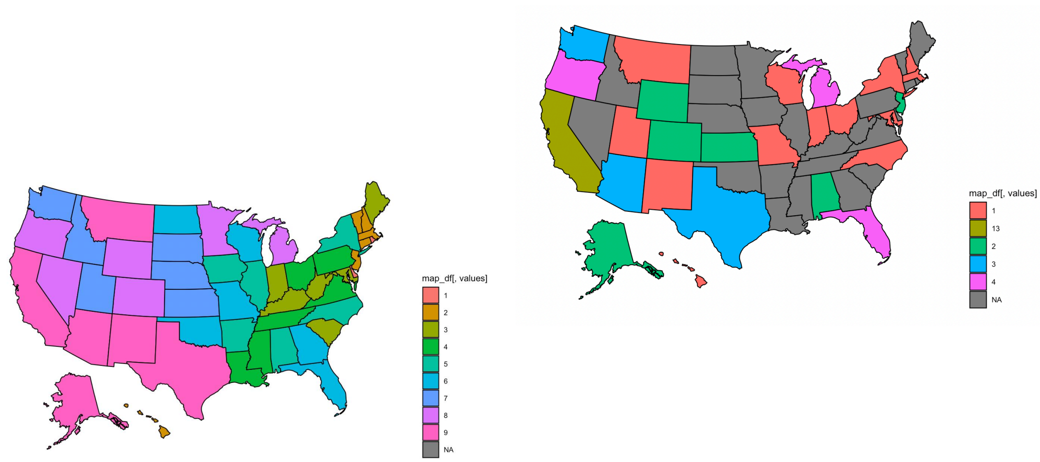

Figure 11: The charts above depict the expected and observed number of journal entries in each state.

mileage. In other words, just because a state is bigger does not mean there are more published articles from that particular state.

2.5 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 contingency table, in Table 5, counts the number of times an ecosystem was represented in the data set. From Table 5, one can tell that most of the

Ecosystem	**Count**
Marrine	20
Terrestrial	127
Freshwater	12

Table 5: The contingency table above shows the number of journal entries in each ecosystem.

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. Figure 12, shows the Chi-Square test that test if this statement is true.

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 13 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. Figure 14, shows 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 15 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.

3 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.

```
        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>
```

Figure 12: The chi-square test is looking for a difference between expected and actual values for ecosystem.

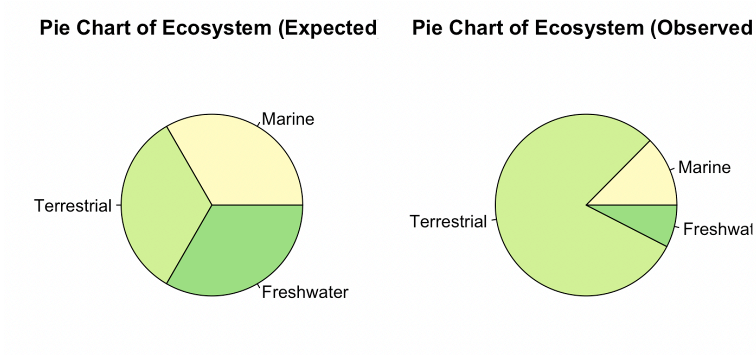

Figure 13: The pie charts above depict the expected and observed number of journal entries in each ecosystem.

```
        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 14: The chi-square test is looking for a difference between expected and actual values for ecosystem, while using square mileage.

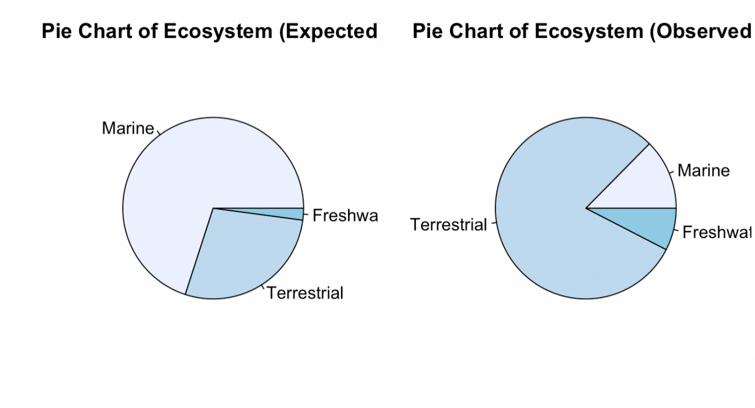

Figure 15: The pie charts above shows the expected frequency based on square mileage vs the observed.