

Exercise 8- Chi-Squared Map Analysis (10 pts)

Introduction

The Chi-Square (χ^2) statistic is used to determine if there is a difference between observed values and expected values, therefore the null hypothesis is that the frequency of the observations found in the rows are independent of the frequency of observations found in the columns. The null hypothesis (h_0) would be “there is no difference between the observed and expected frequencies.” The alternative hypothesis (h_a) would therefore be “there is a difference between the observed and expected frequencies.” Remember that contingency tables test for differences but not the direction of the difference... that must either be implied from the data or determined using another statistical method. For this exercise we will use an alpha value (α) of 0.05 (or 5%). This number determines how “rare” the results have to be before we reject the null hypothesis.

Procedure:

1. Construct a contingency table from the stratified data.
2. Determine both the row totals and column totals (sum across rows, sum across columns).
3. Calculate the expected frequencies.
4. Calculate the Chi-Square statistic.
5. Determine the Degrees of Freedom.
6. Compare the Chi-Square statistic to the critical value.
7. Write an appropriate statement of your findings.

The equation for determining the expected frequencies (f'_{ij}) is:

$$f'_{ij} = \frac{(R_i)(C_j)}{n}$$

...where f'_{ij} is the expected frequency, R is the row, C is the column, and n total observations.

The equation for calculating the Chi-Square (χ^2) statistic is:

$$\chi^2 = \sum \sum \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

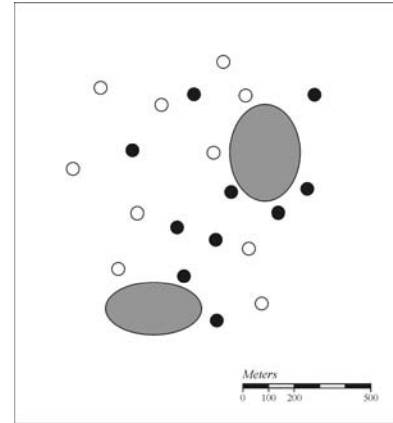
...where f_{ij} is the observed frequency and f'_{ij} is the expected frequency.

The equation for the degrees of freedom (v) is:

$$V = (r - 1)(c - 1)$$

An Example of Chi-Square Map Analysis

In a neighborhood there is an elementary school 5th grade class that has experienced a large number of childhood cancer cases. Of the 20 children in the class, ten developed some form of cancer. Parents noted that these children lived near two old factory sites and were concerned that pollutants found at these sites were making their children sick. One method of testing whether children living near the old factories were more likely to develop cancer is by spatially applying a χ^2 contingency analysis. In this analysis we will be testing whether the observed number of cancers near the old factories is different than the expected number of cancer cases. In this analysis we will use the entire 5th grade class as a study population. The first step is to determine which students live “near” the factories. Previous research suggests that the maximum distance that pollutants would migrate off-site is 200 meters, so we will use this distance as a measure of which cases are near or not near.



Using the scale provided, 200 meter radii are drawn around the locations where the students live. Any student residence whose radii intersects the factory sites will be counted as “near”, the remaining will be counted as “not near.” On the map to the right eight of the 20 student homes had radii that intersected a factory site. Of these eight locations, five were also cancer cases while three had no cancer. These data can now be used to create a contingency table as seen below.

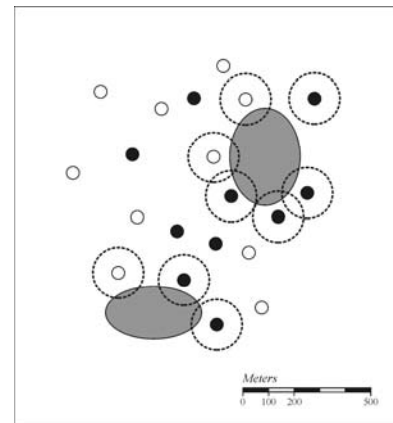


Table 1 Observed Cancer Cases

| | Near | Not Near | Total |
|-----------|------|----------|-------|
| Cancer | 5 | 5 | 10 |
| No Cancer | 3 | 7 | 10 |
| Total | 8 | 12 | 20 |

Next the formula below to calculate the expected frequencies based on what has been observed.

$$f_{ij} = \frac{(R_i)(C_j)}{n}$$

So the expected number of cancer cases near the factories would be (10x8)/20 or 4. The number of no cancer students not living near the factories would be (12x10)/20 or 6. Remember that to get the expected number of cases multiply the row total and the column total and divide by the sample size.

Table 2 Expected Cancer Cases

| | Near | Not Near |
|-----------|------|----------|
| Cancer | 4 | 6 |
| No Cancer | 4 | 6 |

In this theoretical distribution of students with and without cancer, four from each group should be found living near the factories and six from each group should live away from the factories. Now we can test to see if what we observe is different from what is expected. We do this by using the χ^2 statistic...

What is a Theoretical Distribution?

Calculating the Chi-squared statistic is simply a matter of determining whether the observed distribution differs significantly from a theoretical distribution. The theoretical distribution is derived from the observed distribution as follows for the above example:

- .4 or 40% of the students live near a factory (8 out of 20)
- .6 or 60% of the students do not live near the factory (12 out of 20)
- .5 or 50% of the students have cancer (10 out of 20)
- .5 or 50% of the students do not have cancer (10 out of 20)

We can partition these percentages out in a table as follows:

| | Near | Not Near |
|-----------|-----------------------|-----------------------|
| Cancer | $(.4 \times .5) = .2$ | $(.6 \times .5) = .3$ |
| No Cancer | $(.4 \times .5) = .2$ | $(.6 \times .5) = .3$ |

Therefore, .2 or 20% of the total students should live near the factory **and** have cancer, .3 or 30% of the total students should not live near the factory **and** have cancer. So the expected number of students living near the factory and having cancer is 20% of 20 students, or 4 students. Also note that if you add up the percentages in the above table they will always equal 100% or 1. Chi-squared analysis looks at the total difference between what is observed and what is expected and then uses probabilities to determine just **how** different the two are.

Calculating the Chi-Squared Statistics and Testing for Difference

Since we need to know the total difference between the observed and expected distributions, we simply subtract the observed from the expected frequencies (cell values). Since the observed may be less than the expected there is a chance that negative numbers will result, which when summed will lower the total difference. To keep this from happening the differences are squared (hence chi-squared) to remove the negative sign. The squared difference is then divided by the expected frequency so that the result is in “expected frequency units”... this is because the probability function found in the chi-squared table is calculated in expected frequency units. Below is the chi-squared equation:

$$\chi^2 = \sum \sum \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

So for the above example chi-squared would be calculated as follows:

$$\chi^2 = \frac{(5-4)^2}{4} + \frac{(3-4)^2}{4} + \frac{(5-6)^2}{6} + \frac{(7-6)^2}{6} = 0.33$$

As on the first page, the degrees of freedom (ν) is $(r-1)(c-1)$ where r are the rows and c are the columns. Since we have 2 rows and 2 columns, $\nu = (2-1)(2-1) = 1$. Degrees of freedom (ν) is a measure of minimum amount of information you need to calculate the expected frequencies. In a 2x2 table, if you have (2-1) or 1 column cells filled in and (2-1) or 1 row cells filled in you can fill in the rest of the table based on those numbers alone. Notice in the table below that as ν increases (e.g. the amount of preexisting information decreases) the critical χ^2 value increases... in effect as the answer becomes less certain due to lack of previous information it becomes harder to state definitively that the two distributions are different.

The other values that needs to be set is the α -level. By convention it is typically set at 0.05 (5%) but it can be set to any number up to 1. The value 0.05 is a good compromise since we are stating that we only want there to be a 5%

chance that we are incorrect in deciding that the two distributions are different. In other words we want to be 0.95 or 95% confident that our answer is correct. Alternately we could set the α -level at 0.01 which would give us a confidence level of 99%, but that may be too restrictive... our hypothesis may end up being correct, but the statistics will tell us our hypothesis is incorrect (an error of omission).

The final step is to determine the threshold value for chi-squared. This is taken from a chi-squared table are noted below:

Chi-Squared Probability Table

| α v | 0.250 | 0.100 | 0.050 | 0.025 | 0.010 | 0.005 |
|-----------------|---------|---------|---------|---------|---------|---------|
| 1 | 1.32330 | 2.70554 | 3.84146 | 5.02389 | 6.63490 | 7.87944 |
| 2 | 2.77259 | 4.60517 | 5.99146 | 7.37776 | 9.21034 | 10.5966 |
| 3 | 4.10834 | 6.25139 | 7.81473 | 9.34840 | 11.3449 | 12.8382 |
| 4 | 5.38527 | 7.77944 | 9.48773 | 11.1433 | 13.2767 | 14.8603 |
| 5 | 6.62568 | 9.23636 | 11.0705 | 12.8325 | 15.0863 | 16.7496 |
| 6 | 7.84080 | 10.6446 | 12.5916 | 14.4494 | 16.8119 | 18.5476 |
| 7 | 9.03715 | 12.0170 | 14.0671 | 16.0128 | 18.4753 | 20.2777 |
| 8 | 10.2189 | 13.3616 | 15.5073 | 17.5345 | 20.0902 | 21.9550 |
| 9 | 11.3888 | 14.6837 | 16.9190 | 19.0228 | 21.6660 | 23.5894 |

Since the degrees of freedom is 1 and our α -level is 0.05 (5%), the chi-squared threshold value is 3.84146. Our calculated chi-squared value is 0.33, which is much less than the threshold value. This means that our observed distribution *is not* different than the theoretical distribution—the cancer cases are not more frequent near the factories. Therefore we accept the null hypothesis of no difference between the distributions. We would state our finding as follows:

Based on the results of a χ^2 analysis it was found that there was no difference between cancer cases in 5th grade children living near and those not living near the old factory sites within the study area ($\chi^2 = 0.33$).

The Exercise

Using the method outlined above and the attached map, please determine if cancer is more frequent in students living near high voltage power lines. A student is considered to be “living near” a high voltage power line if their homes are within 100 meters of the nearest power line. Be sure to draw the radii on the attached map. Use an α -level of 0.05 (5%). You must show all of your work on the attached blank page. When you have completed the exercise please fill in the following information:

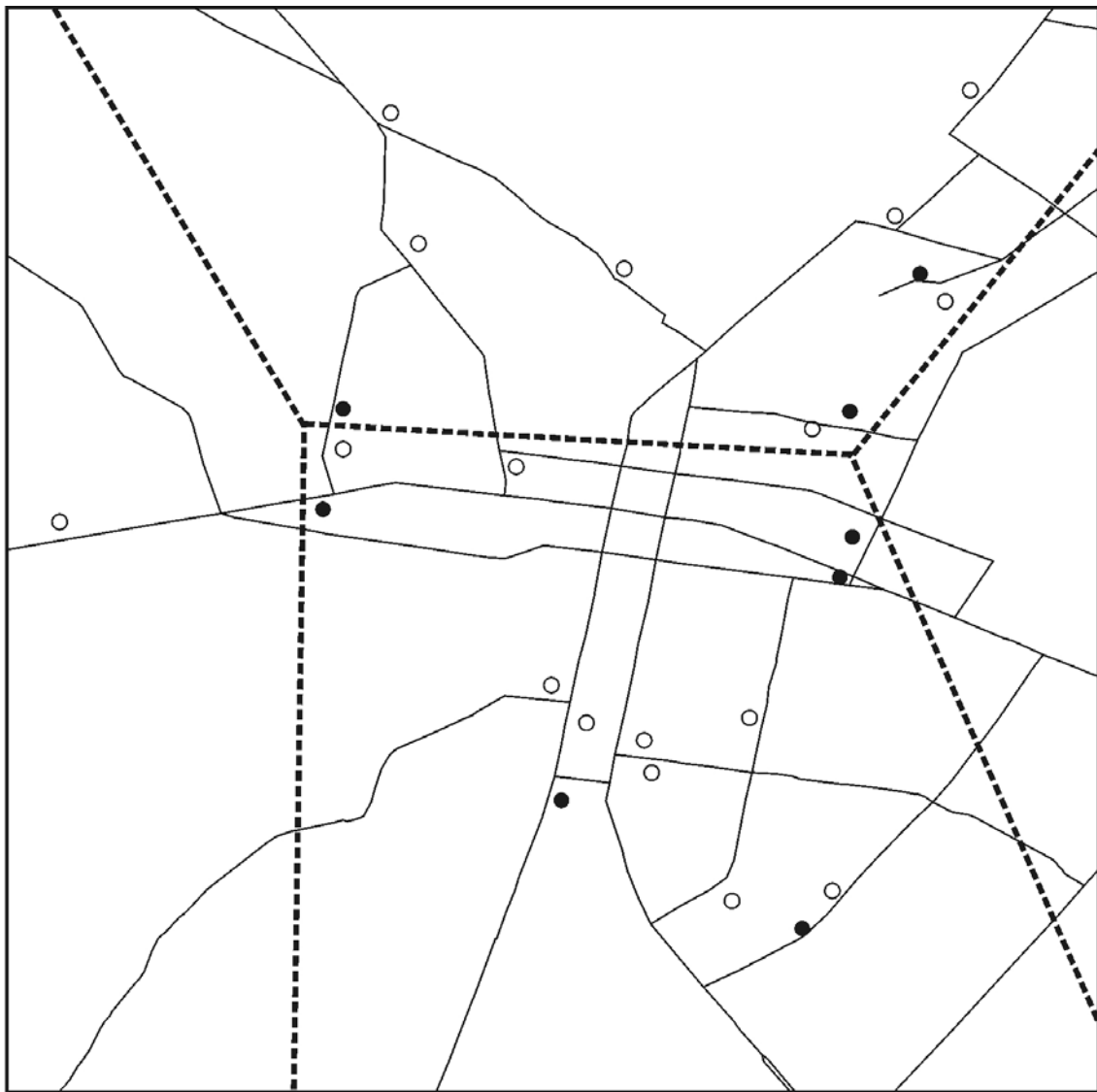
Alpha (α) level = _____

v = _____

Threshold (critical) value = _____

χ^2 = _____

Please note you findings in a succinct statement:



Streets



High voltage power lines



Student with cancer



Student without cancer

Meters

