**INVESTIGATION 07**

**Due Thursday midnight**

**Names: *>>***

**PART 1: The Chi-Square Goodness of Fit Test**

***A blue star shaped object with black lines

Description automatically generated with medium confidenceDeer Habitat and Fire***

***STEP 1: Ask a research question.*** *The researchers were curious about the observed patterns in deer behavior in the presence of fire.*

***STEP 2: Design a study and collect data.*** *To test this, six months after a fire burned 730 acres of homogenous deer habitat, they surveyed a 3,000-acre parcel surrounding the area, which they divided into four regions as given in the Figure.*

***STEP 3: Explore the data.***

1. ***Calculate the proportion for each region.***

|  |  |  |
| --- | --- | --- |
| **Region** | **Acres** | **Proportion** |
| 1. Inner Burn | 520 | 520/3000 = 0.173 |
| 2. Inner Edge | 210 |  |
| 3. Outer Edge | 240 |  |
| 4. Outer Unburned | 2,030 |  |
| **TOTAL** | **3,000** | **1.000** |

**STEP 4: Draw inferences beyond the data.** Under the null hypothesis, if deer were randomly distributed over 3,000 acres, then we would expect the counts of deer in the regions to be in proportion to the sizes of the regions.

1. ***Write the null and the alternative hypotheses within the context of this study.***

The researchers observed a total of 75 deer and the table below represents the observed counts.

1. ***Calculate the proportion for each region.***

|  |  |  |
| --- | --- | --- |
| **Region** | **Observed Counts** | **Expected Counts** |
| 1. Inner Burn | 2 | (0.173) × 75 = 13 |
| 2. Inner Edge | 12 |  |
| 3. Outer Edge | 18 |  |
| 4. Outer Unburned | 43 |  |
| **TOTAL** | **75** | **75** |

**Assumptions and Conditions**

We have three conditions to decide whether we can use theory-based approach.

1. **Counted Data:** The data must be *counts* for the categories of a categorical variable. This might seem a simplistic, even silly condition. But many kinds of values can be assigned to categories, and it is unfortunately common to find the methods of this chapter applied incorrectly to proportions, percentages, or measurements just because values happen to be organized in a table. So, check to be sure the values in each cell really are counts.
2. **Independence of Observations:** The counts in the cells should be independent of each other. The easiest case is when the individuals who are counted in the cells are sampled independently from some population. That’s what we’d like to have if we want to draw conclusions about that population.
3. **Sample Size:** Different books have different criteria for the Chi-square test to be valid. During the lecture, we mentioned that all observed counts should be at least 10. Alternatively, we can check the Expected Cell Counts *instead of* observed counts. This alternative approach states that one might expect at least 5 observations in each expected count cell. For this investigation, let’s use this alternative approach.

Based on the information given above

1. **Are the assumptions and conditions met for performing a goodness-of-fit test? Explain your reasoning.**

**Chi-Square Statistic**

When dealing with counts and investigating how far the observed counts are from the expected counts, we can use a new test statistic called the **chi-square (χ2)** statistic.

A black text on a white background

Description automatically generated

***Notation Alert:*** *The only use of the Greek letter (chi) in statistics is to represent this statistic and the associated sampling distribution. This is another violation of our “rule” that Greek letters represent population parameters. Here we are using a Greek letter simply to name a family of distribution models and a statistic.*

1. ***Let’s calculate expected counts and Chi-Square Statistic.***

|  |  |  |  |
| --- | --- | --- | --- |
| **Region** | **Observed Counts** | **Expected Counts** | A black and white math equation  Description automatically generated |
| 1. Inner Burn | 2 | (0.173) × 75 = 13 | A number with numbers and lines  Description automatically generated with medium confidence |
| 2. Inner Edge | 12 | 5.25 |  |
| 3. Outer Edge | 18 | 6 |  |
| 4. Outer Unburned | 43 | 50.75 |  |
| **TOTAL** | **75** | **75** |  |

1. ***Calculate df and find the p-value by using our*** [***applet***](https://www.rossmanchance.com/applets/2021/test/chiCalc.htm)***.***

***A screenshot of a calculator

Description automatically generatedPS:*** This figure is an example of df = 3 and chi-square statistics is equivalent to 4. The p-value is calculated as 0.2615. Put the df and chi-square statistics that we calculated and find the p-value.

43.2

**STEP 5: Formulate conclusions.**

1. **Use the p-value obtained in #6 to state a conclusion in the context of the problem. Be sure to comment on statistical significance.**
2. **Upon comparing the observed and expected frequencies *what can we say about deer’s movements?* Why did we see such a pattern?**

1. **To what population are you willing to generalize the results?**

A stop sign on the corner of a street

Description automatically generated**PART II: Chi-Square Test for Independence**

***STEP 1. Ask a research question.***What influences the likelihood of a driver coming to a complete stop at a stop sign?

A screenshot of a test results

Description automatically generated***STEP 2. Design a study and collect data.*** Students at Virginia Tech studied which vehicles come to a complete stop at an intersection with four-way stop signs in an observational study. They looked at several factors to see which (if any) were associated with coming to a complete stop.

1. **What are the variables in this study? Are they categorical or quantitative?**

A graph of a bar graph

Description automatically generated**STEP 3. Explore the data.** The segmented bar graph shows the proportion of vehicles that came to a complete stop (yes) for the three arrival positions to the intersection.

1. **Based on the graph, can you make an educated guess about the association between the two variables? Do they appear to be associated with each other? Explain your reasoning by referring to the graph.**

**STEP 4: Draw inferences beyond the data.**

Under the null hypothesis, if deer were randomly distributed over 3,000 acres, then we would expect the counts of deer in the regions to be in proportion to the sizes of the regions.

1. **Write the null and the alternative hypotheses within the context of this study.**

**Assumptions and Conditions**

We have three conditions to decide whether we can use theory-based approach.

1. **Counted Data:** The data must be *counts* for the categories of a categorical variable. This might seem a simplistic, even silly condition. But many kinds of values can be assigned to categories, and it is unfortunately common to find the methods of this chapter applied incorrectly to proportions, percentages, or measurements just because values happen to be organized in a table. So, check to be sure the values in each cell really are counts.
2. **Independence of Observations:** The counts in the cells should be independent of each other. The easiest case is when the individuals who are counted in the cells are sampled independently from some population. That’s what we’d like to have if we want to draw conclusions about that population.
3. **Sample Size:** Different books have different criteria for the Chi-square test to be valid. During the lecture, we mentioned that all observed counts should be at least 10. **For this investigation, let’s use this alternative approach.**

Based on the information given above

1. **Are the assumptions and conditions met for performing a goodness-of-fit test? Explain your reasoning.**

**Chi-Square Statistic**. Instead of calculating chi-square statistic by hand, we will use R programming language. Below, you will find necessary codes and outputs.

A screenshot of a computer code

Description automatically generated

**The anatomy of the code.** Below, you will find the elements of the code used above, along with explanations for each element.

* chi\_square\_result <-
  + This part creates a new object, named ***chi\_square\_result***, that will store the result of the Chi-Square test.
  + The <- symbol is the assignment operator in R, meaning the result of the test will be saved in chi\_square\_result.
* chisq.test(complete\_stop\_data)
  + The **chisq.test() function** performs a Chi-Square test on a data set.
  + In this case, the function is applied to **complete\_stop\_data**, which is the table that contains counts to analyze whether there’s an association between variables.
* print(chi\_square\_result)
  + By using print(), we can see the test statistic, degrees of freedom, and p-value, which help us determine if there’s an association between the variables.

1. **List the chi-square statistic, degrees of freedom, and p-value from the output provided above.**

**STEP 5: Formulate conclusions.**

1. **Use the p-value obtained in #14 to state a conclusion in the context of the problem. Be sure to comment on statistical significance.**

1. **To what population are you willing to generalize the results?**
2. **Can you draw a cause-and-effect conclusion? Explain.**