**INVESTIGATION 07 RUBRIC**

**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. ***(2 pt. each; 6 pt. total) Calculate the proportion for each region.***

|  |  |  |
| --- | --- | --- |
| **Region** | **Acres** | **Proportion** |
| 1. Inner Burn | 520 | 520/3000 = 0.173 |
| 2. Inner Edge | 210 | 0.07 |
| 3. Outer Edge | 240 | 0.08 |
| 4. Outer Unburned | 2,030 | 0.677 |
| **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. ***(2 pt. each; 4 pt. total) Write the null and the alternative hypotheses within the context of this study.***

***H0:*** *The observed counts follow the same distribution as the expected counts.*

***HA:*** *The observed counts do not follow the same distribution as the expected counts.*

**OR**

***H0:*** *Pr {inner burn} = 0.173; Pr {inner edge} = 0.07; Pr {outer edge} = 0.08; Pr {outer unburned} = 0.677*

***HA:*** *At least one of them will be different.*

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

1. ***(2 pt. each; 6 pt. total) Calculate the proportion for each region.***

|  |  |  |
| --- | --- | --- |
| **Region** | **Observed Counts** | **Expected Counts** |
| 1. Inner Burn | 2 | (0.173) × 75 = 13.00 |
| 2. Inner Edge | 12 | 5.25 |
| 3. Outer Edge | 18 | 6.00 |
| 4. Outer Unburned | 43 | 50.75 |
| **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. **(4 pt.) Are the assumptions and conditions met for performing a goodness-of-fit test? Explain your reasoning.**

Yes because the data are counted (# of deers) and we assume that the counts in each cells are independent of each other. For the sample size, there are at least 5 observations in each expected count cell.

**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. ***(5 pt. each; 15 pt. total) Let’s calculate expected counts and Chi-Square Statistic.***

|  |  |  |  |
| --- | --- | --- | --- |
| **Region** | **Observed Counts** | **Expected Counts** |  |
| 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 | (12 – 5.25)2/5.25 |
| 3. Outer Edge | 18 | 6 | (18 – 6)2/6 |
| 4. Outer Unburned | 43 | 50.75 | (43 – 50.75)2/50.75 |
| **TOTAL** | **75** | **75** | 43.2 |

1. ***A screenshot of a calculator

   Description automatically generated(5 pt.) Calculate df and find the p-value by using our*** [***applet***](https://www.rossmanchance.com/applets/2021/test/chiCalc.htm)***.***

***PS:*** 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.

**STEP 5: Formulate conclusions.**

1. ***(5 pt.)* Use the p-value obtained in #6 to state a conclusion in the context of the problem. Be sure to comment on statistical significance.**

The upper tail area beyond 43.2 is less than 0.0001 so we have very strong evidence in favor of HA. The deer show preference for some areas over the others.

1. ***(5 pt.)* Upon comparing the observed and expected frequencies *what can we say about deer’s movements?* Why did we see such a pattern?**

They moved from 1st and 4th areas to 2nd and 3rd areas. We might explain this situation due to having fresh vegetation after fire in those areas. They just wanted to have some fresh food over there 😊

1. ***(5 pt.)*** **To what population are you willing to generalize the results?**

**Deer lives in this particular area or an area having similar characteristics after the presence of fire.**

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. ***(5 pt.)* What are the variables in this study? Are they categorical or quantitative?**

* **The arrival position of vehicles approaching an intersection all traveling in the same direction. It has three groups (categorical variable): (1) whether the vehicle arrives alone, (2) is the lead in a group of vehicles, or (3) is a follower in a group of vehicles.**
* **Whether a driver coming to a complete stop at a stop sign or not (cateogorical variable)**

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. ***(5 pt.)* 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.**

**Any response having solid reasoning can be considered as fully correct answer.**

**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. ***(4 pt.)* Write the null and the alternative hypotheses within the context of this study.**

***Null hypothesis:*** There is no association between the arrival position of the vehicle and whether it comes to a complete stop.

***Alternative hypothesis:*** There is an association between the arrival position of the vehicle and whether it comes to a complete stop.

***OR***

******

**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 approach.**

Based on the information given above

1. ***(5 pt.)* Are the assumptions and conditions met for performing a chi square test for independence? Explain your reasoning.**

**If they are too conservative, they will say no because one of the observed counts was below 10. If they are a little bit flexible, they will say yes because only one cell violated this sample size condition. We can accept both as fully correct as long as they have these reasons. At this stage, this answer is okay.**

**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. ***(2 pt. each; 6 pts. total)* List the chi-square statistic, degrees of freedom, and p-value from the output provided above.**

**Chi square is 4.7276, degrees of freedom is 2 and p value is 0.09406**

**STEP 5: Formulate conclusions.**

1. ***(5 pt.)* Use the p-value obtained in #14 to state a conclusion in the context of the problem. Be sure to comment on statistical significance.**

**With a p-value of 0.094 we have moderate evidence against the null hypothesis, but this evidence is not extremely strong. *(Again, both answers (reject or fail to reject H0 can be correct as long as they have a solid reasoning)***

**So, these sample data provide only moderate evidence that the probability of coming to a complete stop is related to the arrival position of the vehicle.**

1. ***(7.5 pt.)* To what population are you willing to generalize the results?**

**It is probably unwise to generalize these results to intersections beyond the one in northern Virginia that the students studied.**

**Each intersection probably has many different factors that would influence someone to come to a complete stop or not.**

1. ***(7.5 pt.)* Can you draw a cause-and-effect conclusion? Explain.**

**We also would not want to draw any cause-and-effect conclusions between arrival position and coming to a complete stop, because this is an observational study and not a randomized experiment.**