Proposal for Forest Fire Analysis

November, 2018

Davy Guo and Jim Pushor

Data Set Identification

After review several dataset and dicusstion, we've chosen the Forest Fire Data Set from the UCI Machine Learning Repository as our project topic.

Question Identification

This dataset represents 517 instances of forest fires in a national park located in Portugal. Each individual event instance has 13 related attributes, for example: ambient temperature, day of week fire started, wind speed and area burned. Many of these attributes are ones that we assume would be more obviously associated with burnt area (wind speed or ambient temperature), but what about an attribute that might not be immediately associated with the size of the forest that burns? Would the day of the week that a fire starts have any baring on the average size of burnt area? Sometimes resources available in public services are more scarce on weekends versus week days. Might this contribute to larger fires that start on the weekend compared to week days?

Our question: Do fires that begin on weekends burn more area on average than fires that begin on weekdays? We consider this question an **exploratory question**.

Analysis Plan

We will focus our analysis on two columns from the dataset: "day" and "area". We will create two classifications for days of the week: weekday and weekend day. The "area" attribute describes the land area that is burned during one event of forest fire.

Once we have identified our test statistic we will:

1. Define the null and alternative hypothesis:
   H_0: The average burned area of the fire which starts on a weekend is the equal to the average burned area of the fire which starts on a weekday. H_A: The average burned area of the fire which starts on a weekend is *different* from the the average burned area of the fire which starts on a weekday.
2. Compute the test statistic (difference of means) that corresponds to the null hypothesis:

   $\delta^* = S_{weekends} - S_{weekdays} = 0$
3. Use a model of the null hypothesis to generate a random dataset similar to the original dataset (δ) and calculate a test statistic from that randomly generated dataset (do this many times to generate a distribution).
4. See where the test statistic (δ^*) from our sample(s) falls on this distribution
5. If it is near the extremes (past a threshold of alpha = 0.05) we reject the null hypothesis, otherwise we cannot. More specifically, we will determine a p-value that will represent the probability of seeing our observed test statistic or one more extreme under the null hypothesis. Our goal is to determine if there is evidence to reject our null hypothesis.

Data summarization

We have already determined a fixed value of alpha to 0.05 to represent a threshold of significance for rejection of the null hypothesis. We will provide a small set of sample data in a table as well as a table with the difference of means between the two groups. The high and low confidence intervals will be displayed in a table and visually for comparison purposes. Additionally, we will privide a p-value and a plot (i.e. null hypothesis distribution) with the confidence intervals and test statistic overlayed clearly to assist with drawing a conclusion from our data. The null hypothesis distribution could be represented with a histogram or kernel density plot. The important part is to clearly represent where our observed test statistic sits on the null hypothesis distribution.

Citation

https://archive.ics.uci.edu/ml/datasets/Forest+Fires Paulo Cortez, pcortez '@' dsi.uminho.pt, Department of Information Systems, University of Minho, Portugal. Aníbal Morais, araimorais '@' gmail.com, Department of Information Systems, University of Minho, Portugal.