Proposal for Forest Fire Analysis

November, 2018

Davy Guo and Jim Pushor

Data Set Identification

After reviewing several datasets, we have chosen the "Forest Fire Data Set" from the UCI Machine Learning Repository.

Question Identification

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

More specifically, we generated an inferential question: Do fires that begin on weekends burn more area on average than fires that begin on weekdays? We wish to use the sample data to make an inference about the all forest fires of this region (i.e. population).

Analysis Plan

We are focusing our analysis on two attributes from the dataset: "day" and "area". Further, we will create two classifications for days of the week: "weekday" and "weekend day". The "area" attribute describes the land area that burned during one event of a 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 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 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 can determine a p-value that represents 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. To provide the context of our analysis, a small set of sample data with the difference of means between the two groups can be produced along with the high and low confidence intervals. Additionally, we can provide a p-value and a plot (i.e. null hypothesis distribution) with the confidence intervals and test statistic overlayed clearly to assist with our supporting our conclusion from our data. To aid in communicating our conclusion, the null hypothesis distribution may be represented with a histogram or kernel density plot along with visually representing where our observed test statistic and confidence intervals sit.

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.