Population - Proportion

A difference of proportion test is generally carried out when rather than comparing the values of some group, wee are comparing the count of group. As an example, if you make a assumption about a big population, like there are 45% international student and 65% native student in a University that has around 80,000 students. Now, one cannot go to each and every individual to and get their detail, rather one can select a **random sample** out of that population and saw that out of that sample 55% were International student. In this kind of situation, you make a null hypothesis that population proportion of international student is equal to 45% and alternative hypothesis that population proportion of international student is greater than 45 % based on the sample collected.

Problem statement

Scientists want to study the effect of fire burns on land. One of the issues was to determine how much land get burned down when fire occurs. Some local people claim that almost 50% of the land that has got burned was bigger than 12 acres of land. However, scientist wanted to test the result based on hypotheses, so they collected a random sample of 100 burned areas and found out that only 28% of burned area was bigger than 12 acres. Hence, they came up with this hypothesis.

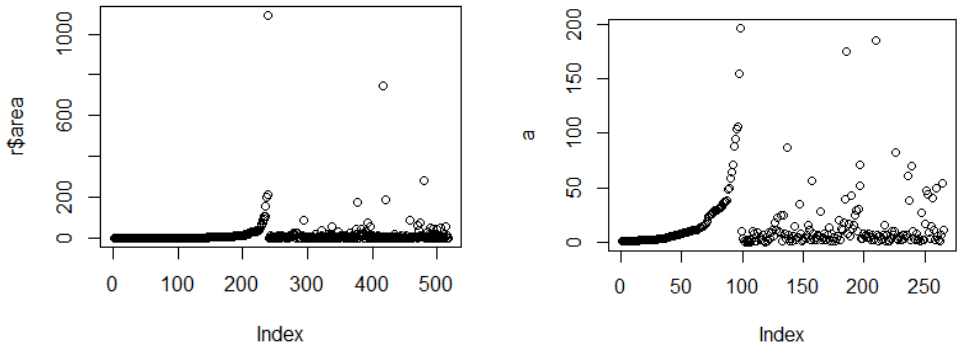
H0: 50% of land that got burned was greater than 12 acres.
H1: Less than 50% of land that got burned was greater than 12 acres.

Dataset

File Name: forest_fire

The dataset if from about the forest fire in northeastern region of Portugal. The dataset has 13 columns and 517 rows. However, for our hypothesis we will only be concerned about one column signifying land burned by fire.

The data have some outliers, since the focus of our work is to run proportion hypotheses, we can remove those outliers. Also, it has around many zeros, so I am going remove them as well.



Code

```r
df <- file.choose() # please choose file named "forest_fire"
df <- read.csv(df) # read the file as csv file

a <- df$area # extract the column that has area value
plot(a) # plot the area variable to get the distribution
a <- a[a>0 & a<=200] # extract only values that are greater than zero and less than 200 (removing outliers)
plot(a) #plot again

value <- a>12 # extracting boolian value where area is greater than 12

result <- prop.test(sum(value),length(a),alternative="less") # running the propertion test
result # printing the result
```

I have extracted the area column out of dataset and stored it into a variable called a. To get an idea of outliers as well as the distribution of the data points I have drawn two plots.

Since we need the count of values that we are going to investigate, I have used a Boolean value and passed the sum of that Boolean value as the first argument of the test function. The second argument is the total length of the data set or in other words number of trials. The last argument is telling us that the alternative hypotheses is that the investigated value is less than 50%.

Interpretation

```
        1-sample proportions test with continuity correction

data:  sum(value) out of length(a), null probability 0.5
X-squared = 50.777, df = 1, p-value = 5.173e-13
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.3285919
sample estimates:
        p
0.2792453
```

We can see in the picture above that says alternative hypothesis: true p is less than 0.5. It indicates that our alternative hypotheses were that the proportion of area that got burnt in category bigger than 12 acres was less than 50%. Hence, by p-value so small we can conclude that we have enough evidence to reject the null hypothesis that more than 50% of the times that land was bigger than 12 acres.