The Association Between Weather and the Severity of Forest Fires in Montesinho Park, Portugal

STAT 302: Accelerated Introduction to Statistics

Group 15
Leader
Hunter Abraham
Members
Luke Becker
Manav Kalathil
Andrew Fix

December 6, 2019

Contents

Abstract **5**

Introduction **5**

Background . . . 5

Broader Impact and Greater Significance . . . 5

Statement of Purpose . . . 6

Specific Aims . . . 6

Methods **6**

Population of Interest . . . 6

Type of Study . . . 6

Data Description and Data Collection . . . 7

Data Description . . . 7

Data Collection . . . 7

Sampling Scheme . . . 7

List of Proxy Variables . . . 7

Explanatory Variables . . . 7

Response Variables . . . 8

Statistical Analysis . . . 8

Randomization Test: Wind Level and Spread of Fires . . . 8

T Test: Mean Burn Area of Follow-Up Fires vs. Isolated Fires . . . 8

Chi-Square Test: Drought Code vs. Burn Area . . . 9

Linear Regression Test: Temperature vs. Burn Area . . . 9

Computational Requirements . . . 9

Results **10**

Randomization Difference in Proportions Test: Windy Conditions vs. Fire Spread . . . 10

Hypotheses . . . 10

Summary Figure . . . 11

Check for Assumptions . . . 12

Calculate Test Statistic . . . 12

Find Confidence Interval . . . 13

Compute the P Value . . . 13

Interpretation . . . 13

T Test: Difference in Means among Isolated vs. Follow-up Fires . . . 13

Hypotheses . . . 14

Summary Figure . . . 14
Check for assumptions . . . 14
Calculate Test Statistic . . . 14
Confidence Interval . . . 15
Compute the P Value . . . 15
Interpretation . . . 15
Chi-Square test: Drought code vs. Area . . . 15
Hypotheses . . . 15
Summary Figure . . . 16
Check for Assumptions . . . 16
Calculate Test Statistic . . . 17
Compute the P Value . . . 17
Interpretation . . . 17
Linear Regression Test: Temperature vs. Burn Area . . . 17
Hypotheses . . . 17
Summary Figure . . . 18
Check for Assumptions . . . 19
Calculate Test Statistic . . . 19
Compute the P Value . . . 20
Confidence and Prediction Intervals . . . 21
Interpretation . . . 22

Discussion **22**

Summary of Findings . . . 22
Randomization Test: Difference in Proportions Between Windy and Non-Windy Fires . . . 22
T Test: Difference in Means Between Follow-Up and Isolated Fires . . . 23
Chi Square Test . . . 23
Linear Regression Test . . . 23
Overall Findings . . . 24
Error Analysis . . . 24
Further Studies . . . 24

References **24**

Appendix **25**

Packages . . . 25

Randomization Test . . . 25

Check for Assumptions . . . 26

Compute the P Value . . . 27

Summary Figure . . . 28

Calculate Test Statistic . . . 29

Chi-Square test: Drought code vs. Area . . . 29

Summary Figure . . . 30

Check for Assumptions . . . 30

Calculate Test Statistic . . . 30

Compute the P Value . . . 30

Linear Regression Test: Temperature vs. Burn Area . . . 30

Summary Figure . . . 30

Check for Assumptions . . . 31

Calculate Test Statistic . . . 31

Compute the P Value . . . 31

Confidence and Prediction Intervals . . . 33

Abstract

Forest fires are a common occurrence in Montesinho Park, Portugal. In an effort to control these fires, the study was performed to determine the relationship between external weather variables and the severity of fires in Montesinho Park. The data in the study was taken from a data set compiled from the years 2000 to 2003. The data set describes forest fires in Montesinho Park, which is located in the northeast region of Portugal. The data consists of 517 recorded fires and contains metrics for the severity of the fires and weather conditions at the time of the fire. Explanatory variables in the study included temperature, wind, and rain. Response variables in the study included metrics for the severity of a fire. To test the relationship between these variables, the study uses a difference in means t-test, a randomization test, a chi square test, and linear regression test. The explanatory variables in the four tests are wind, temperature, time since previous fire, and drought level. The response variables are area and the spread of the fire. The tests show that there is an association between weather and metrics for the severity of a forest fire in Montesinho Park. The difference in means t-test produced an important result: follow-up fires tend to have smaller burn areas. This test is performed to control for the confounding variable of fuel level. The chi square test shows an association between the level of drought and the burn area of a fire. The linear regression test shows that the temperature effects the burn area of a fire. These metrics can be used to predict how severe a fire will be based on the weather conditions when it begins.

Introduction

Background

Montesinho Natural Park is one of Portugal's most coveted parks. It lies in the Trás-os-Montes northeast region in Portugal (BePortugal). The Natural Park is home to an abundance of wildlife and stunning views for the everyday tourist to enjoy. It lies in a supra-Mediterranean climate, with an average annual temperature between 8 and 12 degrees Celsius. Forest fires constantly threaten the natural beauty the park offers. The fires can be attributed to multiple causes such as human negligence or lightning strikes. Although the Portuguese government is devoting more resources to mitigate these fires, they continue to destroy forests every year. Last year alone, the country recorded 12,200 fires, compared to an average of 20,000 per year over the past 10 years (Silva).

While the forest fires threaten the beauty within the park, they also endanger the people who enjoy the park's beauty by engaging in outdoor activities. Forest fires' unpredictable nature has the capacity to burn a large area of forest in a short amount of time. The project is designed to predict when a forest fire is likely to happen and its incoming severity. The ability to predict when a fire will occur and with what severity will allow the park staff and visitors to be more prepared.

Broader Impact and Greater Significance

Forest fires are natural disasters that affect more than just the ecosystems in which they occur. They have the capability to ruin homes, schools, parks, and entire communities in a very short amount of time. For fires occurring in Montesinho Natural Park, finding out what natural conditions precede these fires is vital to preventing them. There are currently eighteen wildfires burning in California, most of which are in the greater Los Angeles area and North of San Francisco. Hundreds of thousands of homes are without power and the air quality in the state's largest cities is low. Natural disasters will always be a threat and society would benefit from determining which variables lead to more destructive fires.

Currently, researchers have devoted their efforts to investigating the effect of climate change on the frequency and severity of forest fires. They have shown that forest fires are happening with increased severity and frequency. Specifically, they found that 50% less forest area would have burned between 1984 and 2015 if climate change did not exist (Pierre-Louis and Popovich, 2018). However, they have not created a reasonable model to predict when or with what severity forest fires will occur. The goal of the project is to find the

association between weather and proxy variables for the severity of a forest fire in Montesinho Park. Other studies have provided an association for other regions such as California, parks in Canada, and areas in the Midwest United States (Collins). However, an association has not been found for Montesinho Park in Portugal. The study will find a specific association for the park. The findings can be used to predict when a fire is likely to start and how severe it will be once it begins burning. While the study cannot be generalized to the general population of wildfires, it will show predictive conditions for the park. A future study could use stratified sampling to ensure that data are sampled so that the explanatory variables are consistent with the general population of forest fires.

Statement of Purpose

The purpose of the study is to determine whether there is an association between the weather of Montesinho Natural Park, Portugal and the severity of forest fires that plague the park. The study aims to find the relationship between wind, rain, and temperature measurements and the area of forest affected by each fire. More specifically, The goal is to find evidence that a lack of rainfall and an increase in temperature and wind speeds results in a more severe burn. Such a study could be used to help fire fighting forces predict when a forest fire is likely to happen.

Specific Aims

The objective of the project is to find evidence for an association between the weather in Montesinho Park and the number and severity of forest fires it experiences. The project's overall objective can be broken down into several sub-tasks:

- Find evidence that the time between fires is associated with the area burned
- Find evidence that wind is associated with the area burned
- Find evidence that temperature is associated with the area burned

The main challenge with the study is that the data cannot be randomized. Because the project is an observational study, the explanatory variables cannot be randomized before data collection. Therefore, the study cannot draw conclusions about whether the weather caused the severity of the fires to increase. There could also be confounding variables with the study that create an association among variables without causation. For example, if there was a fire closely preceded by a different fire, it would be less severe because most of the brush would be burned. A solution to this problem is outlined in the "Randomization Test" subsection in "Methods". Also, the density of plant life could be higher in the region of a fire compared to a different one, increasing its severity. The association between variables could be skewed by human factors as well. For example, firefighters let some fires burn longer as opposed to others. Another challenge with the study is that forest fires are infrequent, so there are a limited number of cases with explanatory variables that are in fringe cases (i.e. with a high amount of rain a forest fire is unlikely to start).

Methods

Population of Interest

The population of interest in the study is all the forest fires that occur in Montesinho Park, Portugal.

Type of Study

The project is an observational study, as the explanatory variables were not controlled by the research team.

Data Description and Data Collection

Data Description

The data-set comprises readings from 517 forest fires that occurred between 2000-2003 in the northeast region of Portugal. The data-set uses the Fire Weather Index (FWI), the Canadian system for rating the potential fire danger. The system has six main components, all of which are recorded in the data-set. The six components are the following: the Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and FWI. The first three components are measures of fuel for the fire. The FFMC measures the moisture content in the surface litter and influences the ignition and spread potential. The DMC and DC are two measures of surface and deep layers of organic fuel and impact the intensity of the fire. The ISI denotes quickly a fire spread. The BUI represents the amount of available fuel. The FWI combines the BUI and ISI to indicate fire intensity. For all components, higher values represent more probable burning conditions. Several preliminary statistics for variables of concern in the data-set are listed below:

Name of Variable	Minimum	Mean	Maximum
FFMC	18.70	90.64	96.20
DMC	1.10	110.87	291.30
DC	7.90	547.94	860.60
ISI	0.00	9.02	56.10
temp	2.20	18.89	33.30
wind	0.40	4.02	9.40
rain	0.00	0.02	6.40
area	0.00	12.85	1090.84

Data Collection

The data-set was collected from 517 forest fires that occurred between 2000-2003 in the northeast region of Portugal. It is built using two separate sources. The first database was built by the inspector responsible for reviewing forest fires in the park. Every time a forest fire occurred, several observations were recorded such as: the date, time, and location on a 9x9 grid. The six components of the FWI system and the total burned area was also recorded. The second database was collected by the Bragança Polytechnic Institute in Portugal. The institute observed the weather conditions during the fire. The separate recordings were stored in many spreadsheets, but were eventually manually compiled into one database.

Sampling Scheme

The study uses convenience sampling. Fires were sampled as they occurred from 2000-2003. Because they were sampled as they occurred, the explanatory variables could not be randomized before response variables were recorded. Therefore, the study cannot draw conclusions about causality.

List of Proxy Variables

Explanatory Variables

Name of Variable	Categorical / Quantitative	Levels of Variable
Wind	Categorical	Not Windy, Windy
Rain	Quantitative	N/A
Temperature	Quantitative	N/A

Name of Variable	Categorical / Quantitative	Levels of Variable
Month	Categorical	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
DC	Categorical	Wet, Neutral, Dry

Response Variables

Name of Variable	Categorical / Quantitative	Levels of Variable
Area	Quantitative	N/A
FFMC	Categorical	Dry, Damp, Wet
ISI	Quantitative	N/A
DMC	Quantitative	N/A

Statistical Analysis

Randomization Test: Wind Level and Spread of Fires

A hypothesis test is necessary to identify the effect that wind has on the ISI (initial spread index) of a fire. The variable "wind" will be converted into a categorical variable with the labels "windy" and "not windy". The null and alternative hypotheses are stated as follows:

$$H_0 : p_w = p_n$$
$$H_a : p_w > p_n$$

To perform the test, the variable "wind" will be converted into a categorical variable. Any value in wind greater than or equal to 3.75 will be converted into the category "windy." Any value in wind less than 3.75 will be converted into the category "not windy." The value 3.75 is an industry cutoff for windy conditions. The quantitative variable "ISI" was also converted to a categorical variable based on its mean value. The two levels were "Large Spread" and "Small Spread." The two sample proportions being compared are p_w, the proportion of fires that occurred in windy conditions that are large fires and p_n, the proportion of fires that occurred in non-windy conditions that are large fires. The sample statistic will then be compared to the randomization distribution to determine a p value. The p value will be compared to a significance level, $\alpha = 0.05$, to determine whether the difference in means is statistically significant. If it is, the null hypothesis will be rejected and there will be evidence that wind has an effect on the area of forest burned by a fire.

T Test: Mean Burn Area of Follow-Up Fires vs. Isolated Fires

To solve the issue of closely preceding fires influencing the burn of a following fire, a hypothesis test is necessary. The data will be split into two groups. The first group will be "follow-up fires." These fires occur under 6 months after any prior fire. The second group will be "isolated fires" which occur at least 6 months after any prior fire. The threshold of 6 months was chosen because this is the time required by the flora of northeast Portugal to recover from a forest fire. The hypotheses for the test will be the following:

$$H_0 : \mu_f = \mu_i$$
$$H_a : \mu_f \neq \mu_i$$

Where μ_i is the mean of the sample of the means of the *isolated* fires' burn area and μ_f is the mean of the sample of the mean of *follow-up* fires' burn area. A t test will be used because the distribution of the

difference mean areas is normal, except for a few outliers. The significance level at which the null hypothesis is rejected will be set at $\alpha = 0.05$. If the hypothesis test results in H_0 is rejected, then the data-set will be split into two sets: one of isolated fires and one of follow-up fires. Doing so will account for the confounding variable of brush burn off from prior fires which could decrease the fire's burn area. If H_0 is not rejected, then the data-set will not be split

Chi-Square Test: Drought Code vs. Burn Area

A hypothesis test will be used to determine if the Drought Code is associated with the Burn Area. The quantitative variable "Drought Code" will be converted into a categorical variable with three levels: "Wet," "Neutral," and "Dry." The quantitative variable "Burn Area" will also be converted into a categorical variable with two levels: "Small" and "Large." The hypotheses for the test are stated as follows:

$$H_0 : \text{Burn area is not associated with drought code.}$$
$$H_a : \text{Burn area is associated with drought code.}$$

The hypothesis test will determine whether there is an association between drought code and burn area. The value of the chi-square test, X^2, will be compared to a chi-square distribution to determine a p-value. The resulting p-value will be compared to a significance level of $\alpha = 0.05$ to determine if there is a statistically significant argument for association between drought code and burn area.

Linear Regression Test: Temperature vs. Burn Area

A hypothesis test will be used to determine the effect that temperature has on the burn area of a fire. This can be done by creating linear models with temperature as the explanatory variable and area burned as the response variable. From the linear model, the slope of the linear regression line will be recorded. The alternative and null hypotheses are stated as follows:

$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0$$

The hypothesis test will determine whether the slope of area over temperature, β, is statistically significant. If there is a statistically significant slope, then correlation between temperature and burn area, r, is implied. β is a better test of effect than r because it is denoted in units that represent how area is affected when temperature is affected. In contrast, r is unit less, and therefore provides weaker evidence of effect among the variables. Whether β is statistically significant will be determined with respect to a significance level, α, of 0.05.

Computational Requirements

To perform these tests, a computer with an R distribution installed and the necessary computational power to perform randomization procedures and linear regression is required. Moreover, RStudio, dplyr, tidyverse, grid, gridExtra, ggplot2, knitr, ggfortify, xtable, and reshape2 are required.

Results

Randomization Difference in Proportions Test: Windy Conditions vs. Fire Spread

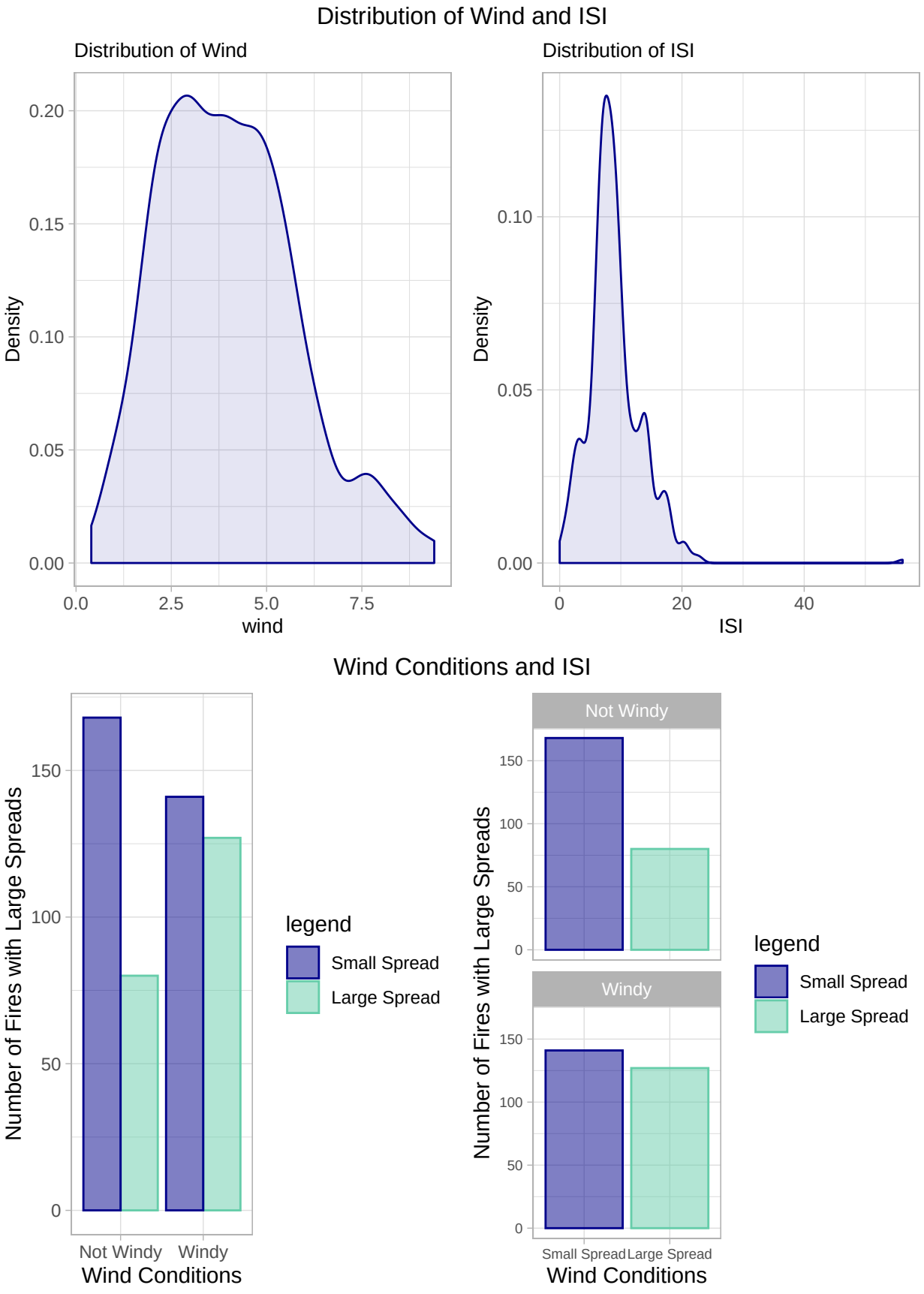
Hypotheses

The hypotheses for this test are:

$$H_0 : p_i = p_f$$
$$H_a : p_i > p_f$$

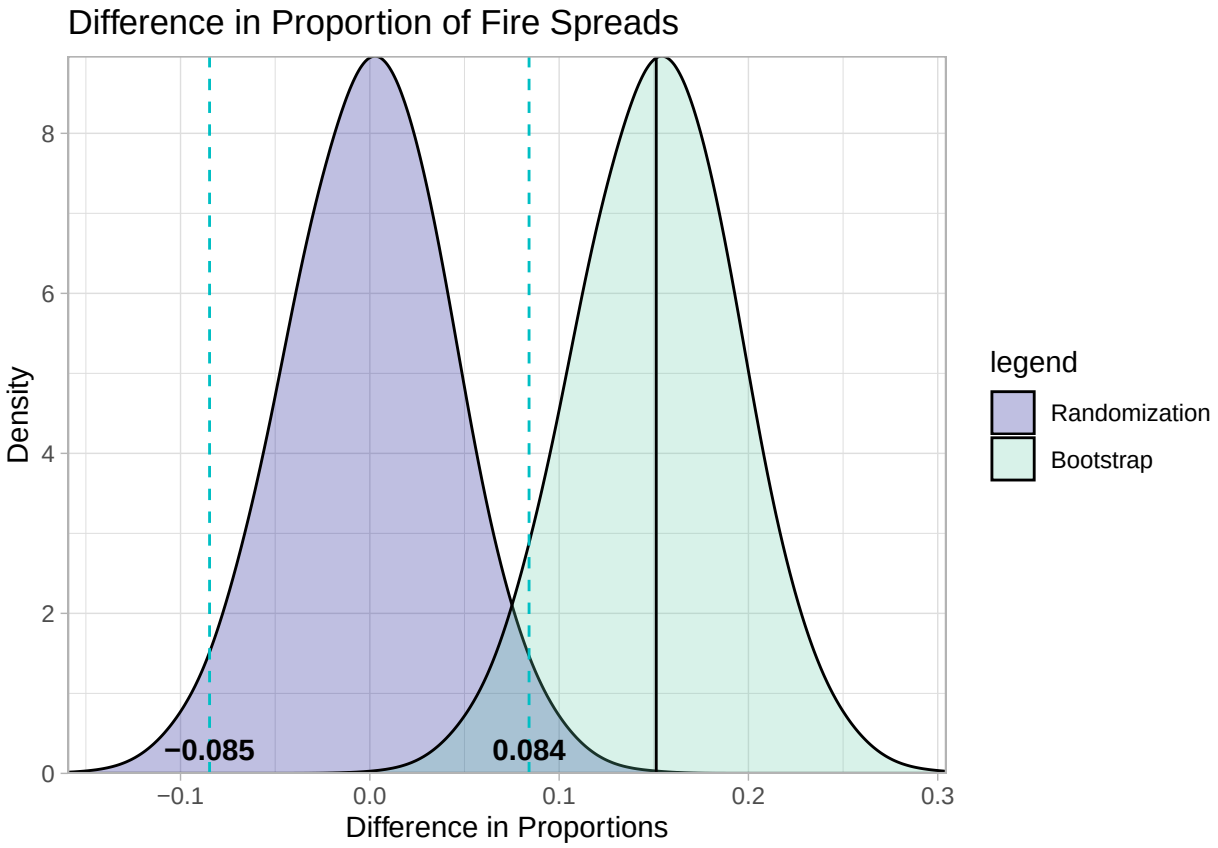
where p_i is the proportion of isolated fires (fires without a preceding fire in the past 6 months) that have large burn areas and p_f is proportion of follow-up fires (fires with a preceding fire in the past 6 months) that have large burn areas.

Summary Figure



Because the distributions of both wind and ISI are abnormal, a randomization test will be used.

Check for Assumptions



The randomization distribution is approximately symmetric and bell-shaped.

Calculate Test Statistic

The sample statistic is computed by:

$$\hat{p}_w - \hat{p}_n = \frac{n_{wl}}{n_w} - \frac{n_{nl}}{n_n}$$

where n_{wl} is the number of windy fires that also had large spreads, n_w is the total number of windy fires, n_{nl} is the number of non-windy fires that also had large spreads, and n_n is the total number of non-windy fires.

The sample statistic for difference in proportions, $\hat{p}_w - \hat{p}_n$, is 0.1513.

The test statistic, z, computed by:

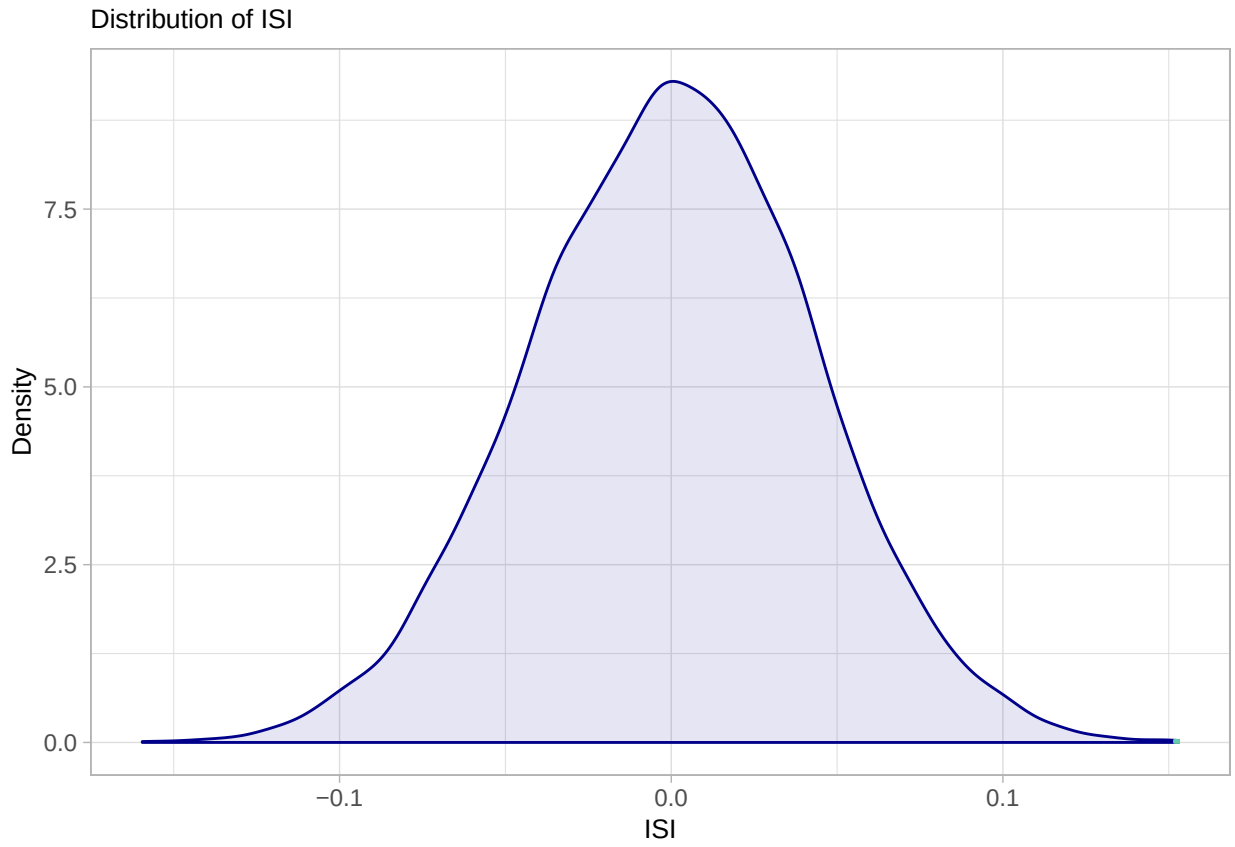
$$z = \frac{\text{Sample Statistic - Null Parameter}}{SE} = \frac{0.1513}{0.09691413} = 3.525979$$

Therefore, the test statistic, z, is 3.525979.

Find Confidence Interval

The confidence interval for the population of difference in proportions between the number of large spreading fires in windy conditions vs. non-windy conditions is computed using the percentile method. The result is $CI = [0.0668, 0.2356]$. In other words, with 95% confidence, large-spreading fires occur in windy conditions between 0.0668 and 0.2356 more of the time than in non-windy conditions.

Compute the P Value



The estimate of p value is the portion of sample difference proportions that have a value of at least 0.1513, the test statistic. The computed p value is 0.0001. That is, the proportion of samples that had a difference in means of at least 0.1513 is 0.0001. The figure above shows the corresponding area in a density plot (The area is rather small and may be invisible).

Interpretation

There is significant evidence to reject the null hypothesis that the average wind spread is not higher in windy conditions. (randomization test for difference in proportions, $\hat{p}_w - \hat{p}_n = 0.1513$, $p = 0.0001$, $\alpha = 0.05$).

T Test: Difference in Means among Isolated vs. Follow-up Fires

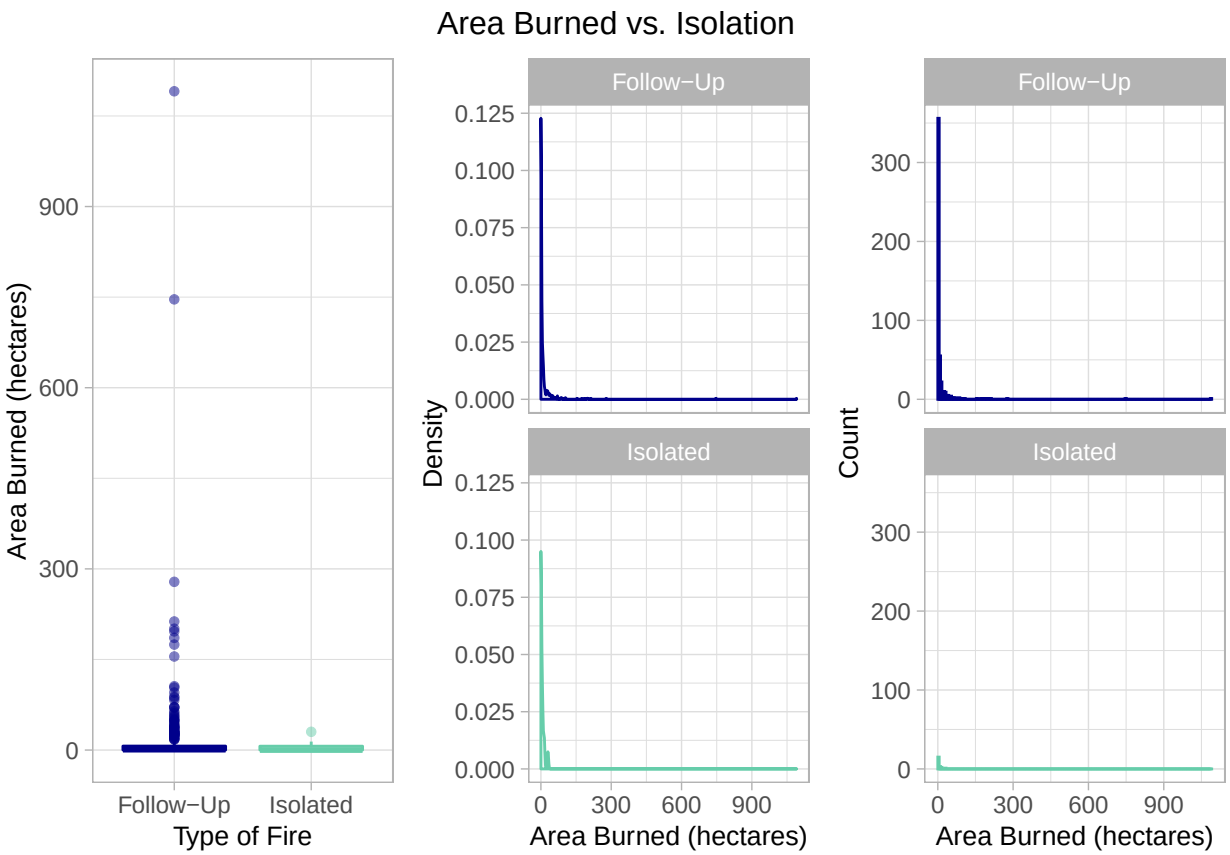
The second test in the study is used to determine if follow-up fires have different mean areas that isolate fires. An isolated fire is defined as a fire without a previous fire in the preceding 6 months. A follow-up fire is defined as a fire with another fire in the preceding 6 months.

Hypotheses

$$H_0 : \mu_i = \mu_f$$
$$H_a : \mu_i \neq \mu_f$$

where μ_i is the mean burn area of isolated fires and μ_f is the mean burn area of followup fires.

Summary Figure



Check for assumptions

In reference to the figures above, although the data is skewed, it is due to only one outlier. Outside of the one outlier, the data is approximately normally distributed. Thus, randomization is not required.

Calculate Test Statistic

The test difference in means, $\bar{x}_i - \bar{x}_f = 8.050202$ where $\bar{x}_i$ is the mean area of all isolated fires and $\bar{x}_f$ is the proportion of fires that occurred under "Not Windy" conditions.

The test statistic (t-score) is:

$$t = 2.38$$

The t-score is computed by comparing the sample difference in proportions to a t-distribution with 494.5 degrees of freedom.

Confidence Interval

The confidence interval is computed by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where s_1 is the standard deviation of the isolated fires' burn areas, n_1 is the number of isolated fires, s_2 is the standard deviation of the followup fires' burn areas, and n_2 is the number of followup fires. The confidence interval is (1.40435, 14.69605). This interval means that with 95% confidence, the difference in mean burn areas between isolated and followup fires is between 1.404 and 14.696. In other words, with 95% confidence, followup fires have burn areas between 1.40435 and 14.6905 hectares larger than isolated fires

Compute the P Value

The p-value, computed using technology, is $p = 0.01769$

Interpretation

There is significant evidence that fires with large spreads occur more often in windy conditions than in non-windy conditions. (t test for a difference in means, $\bar{x}_i - \bar{x}_f = 8.050202$, $t = 2.38$, $p = 0.01769$, $\alpha = 0.05$).

Chi-Square test: Drought code vs. Area

The third test in the study is a Chi-Square test to determine in drought code is associated with the area burned by a fire. To perform the test, the quantitative variable "Drought Code" is converted into a categorical variable with three categories: "Wet," "Neutral," and "Dry." Also, the quantitative variable "area" is converted into a categorical variable with two categories, "Large" and "Small." A table with the observed counts is shown below.

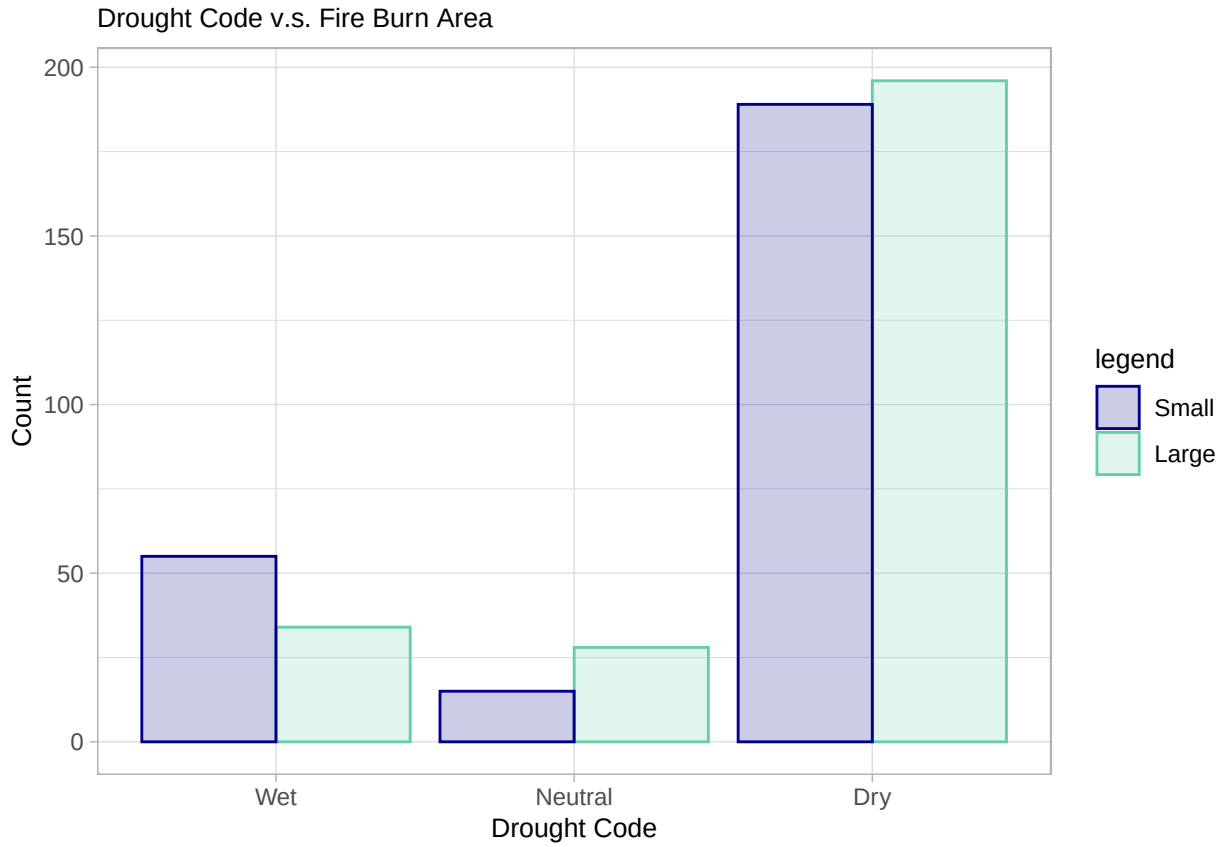
	Small	Large
Wet	55	34
Neutral	15	28
Dry	189	196

Hypotheses

H_0 : Drought Code is not associated with burn area

H_a : Drought Code is associated with burn area

Summary Figure



Check for Assumptions

Expected Count Table

	Small	Large
Wet	44.58607	44.41393
Neutral	21.54159	21.45841
Dry	192.87234	192.12766

All expected counts are over 5. Therefore, a χ^2 distribution is appropriate.

Calculate Test Statistic

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected} =$$
$$\frac{(55 - 44.58607)^2}{44.58607} + \frac{(15 - 21.54159)^2}{21.54159} +$$
$$\frac{(189 - 192.87234)^2}{192.87234} + \frac{(34 - 44.41393)^2}{44.41393} +$$
$$\frac{(28 - 21.45841)^2}{21.45841} + \frac{(196 - 192.12766)^2}{192.12766} =$$
$$\mathbf{9.010661}$$

The test statistic, X^2, is 9.010661.

Compute the P Value

The p-value, computed using technology by comparing the test statistic to a χ^2 distribution with degrees of freedom: $(rows - 1) \times (columns - 1) = 2$, is 0.01104994

Interpretation

There is statistically significant evidence that the drought code before a fire is related to how large it will be (its burn area). (chi-square test for association, $X^2 = 35.26534, df = 2, p = 0.01104994, \alpha = 0.05$).

Linear Regression Test: Temperature vs. Burn Area

A linear model is defined between the air temperature and the area burned by the fire as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

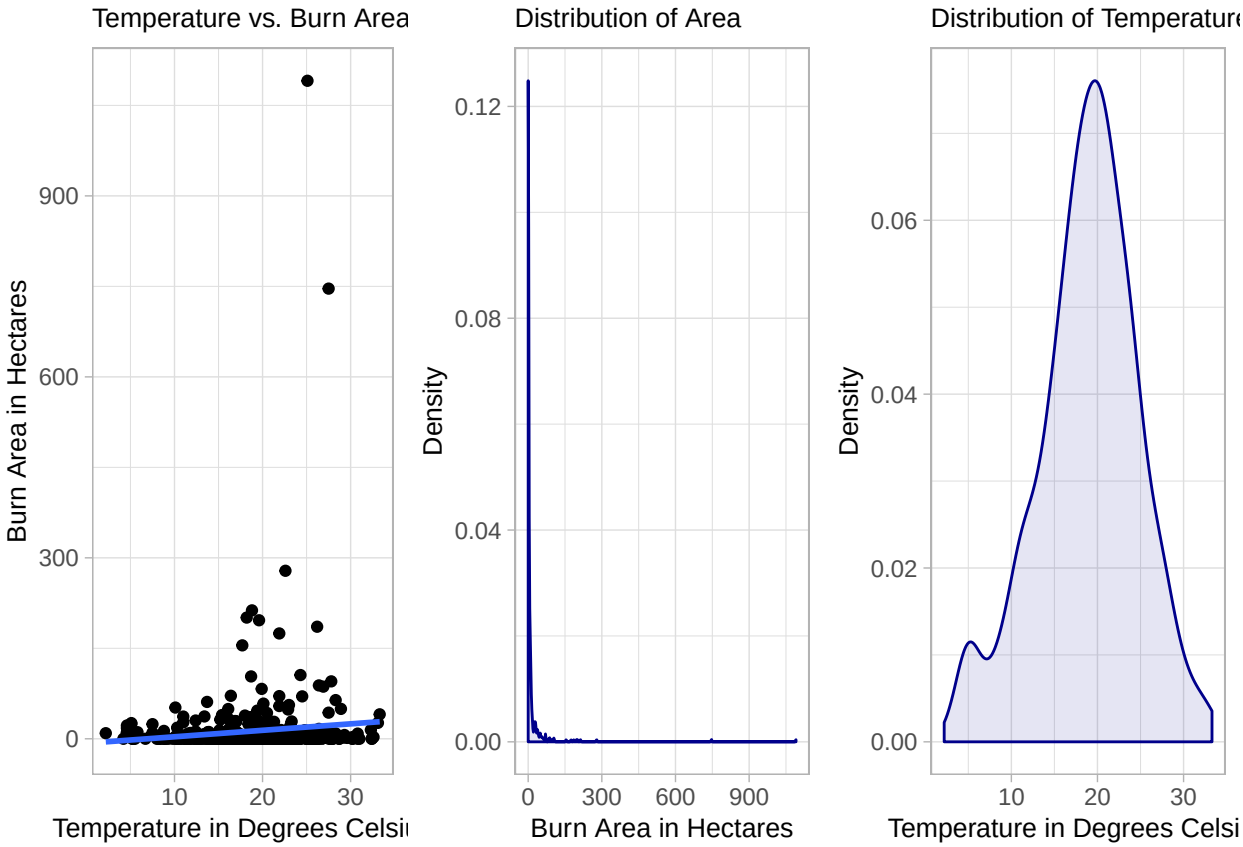
where X is the temperature, in degrees Celsius, and Y is the area burned by the fire. The hypotheses are as follows:

Hypotheses

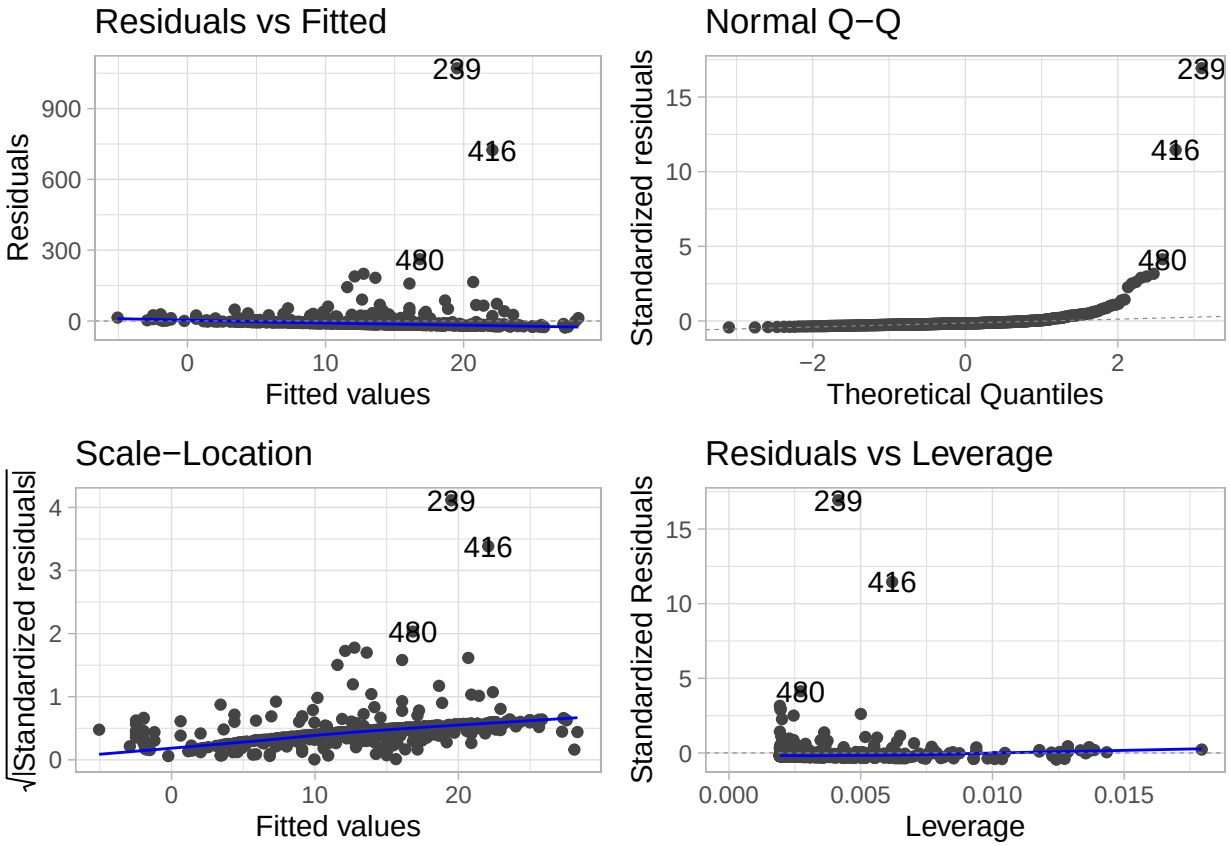
$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0$$

where β is the slope of the linear regression line between temperature and area.

Summary Figure



Check for Assumptions



Linearity
The linearity assumption can be confirmed with the plot of residual vs. fitted values in the top left section of the chart above. The trend line is centered on the y-axis at y=0. It is also approximately horizontal. Thus, the linearity assumption is satisfied.

Normality
The normality assumption can be confirmed with the Q-Q plot in the top right section of the chart above. While the chart has several outliers at the end, the rest of the chart approximately follows the line $y = x$. The outlying points are noted as a potential source of error and the test proceeds as if the normality assumption is satified.

Constant Variance
The constant variance assumption can be confirmed with the Scale-Location plot in the lower left section of the chart above. There is a slight increasing trend in the regression line. This trend is noted as a potential source for error and the test proceeds.

Distribution
Both of the variables' data is not normally distributed. While temperature is normally distributed, the area variable contains outliers that skew the data to the right. However, the sample size is large enough to justify an F test.

Calculate Test Statistic

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.413752	9.4996115	-0.7804268	0.4354982

	Estimate	Std. Error	t value	Pr(>|t|)
temp	1.072628	0.4807528	2.2311417	0.0261015

The test statistic, t^*, is computed using a t-test for slope between temperature and burn area.

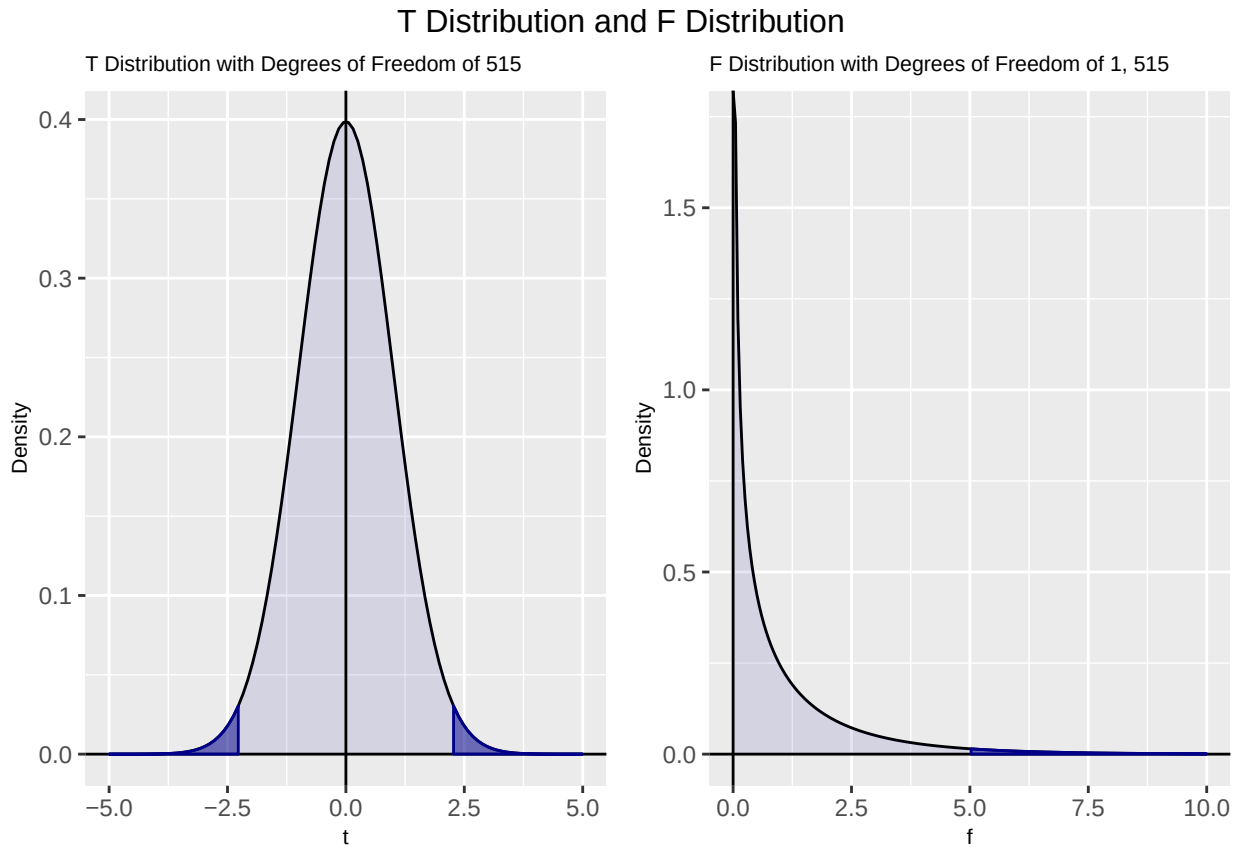
$$t = \frac{b_1 - 0}{SE} = \frac{1.072628}{0.4807528} = 2.231142$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
temp	1	20016.83	20016.828	4.977993	0.0261015
Residuals	515	2070847.80	4021.064	NA	NA

The F statistic is determined from the ANOVA table above.

$$F = \frac{MSModel}{MSE} = \frac{20016.828}{4021.064} = 4.977993$$

Compute the P Value



t:

$T \sim t_{n-2}$

$p - value = P(|T| \geq |t|) = P(|T| \geq 2.231142) = 0.02610141$

F:

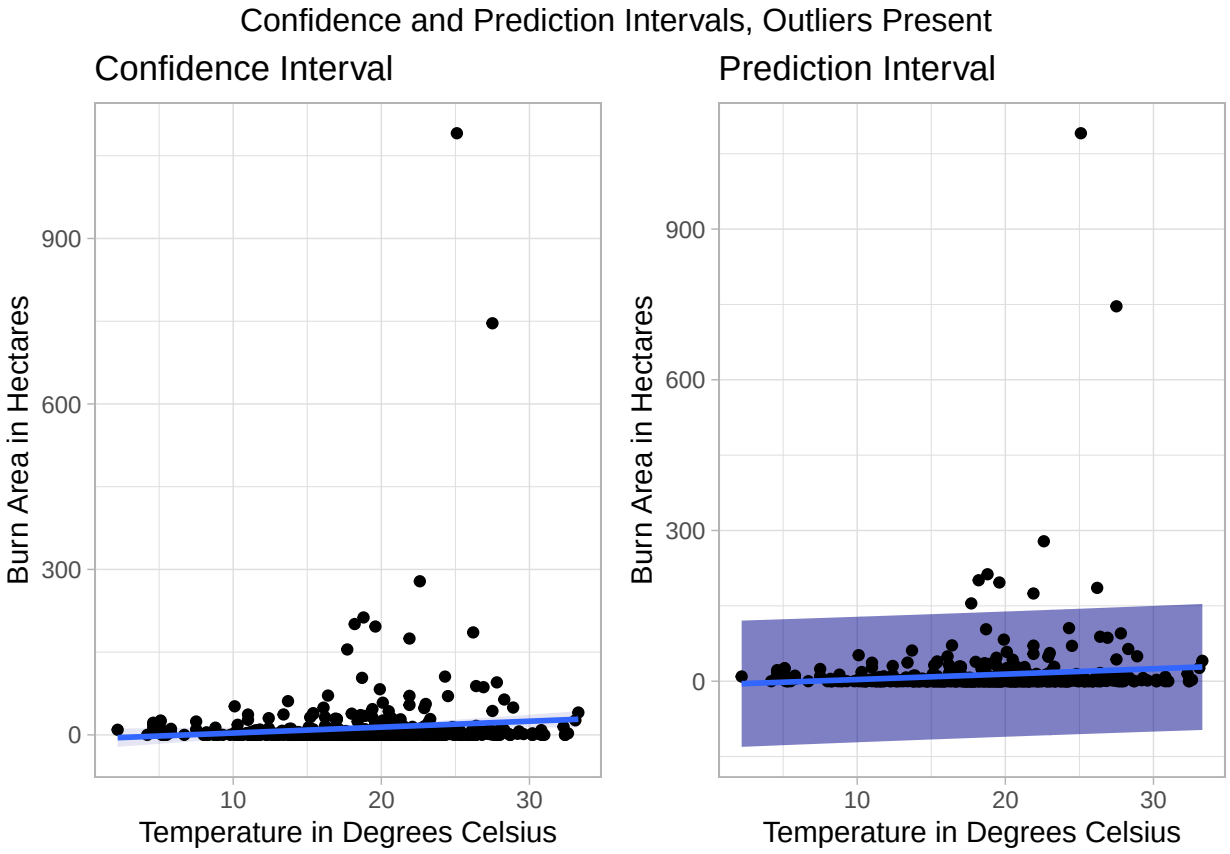
$F \sim F_{(1,n-2)}$

$p - value = P(F \geq f) = P(F \geq 4.977993) = 0.02681081$

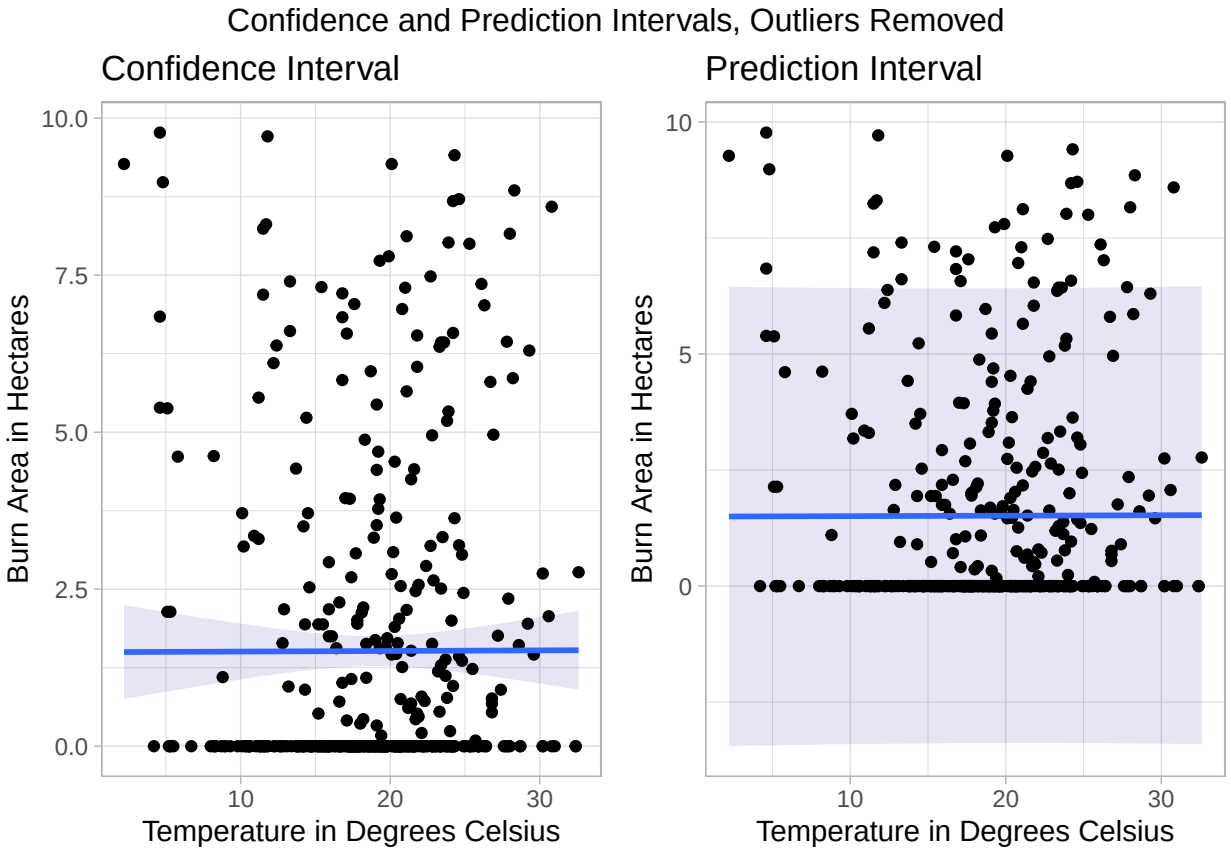
The p values from both the t and f distributions are, as expected, similar. They only differ by numerical error. The p values are represented in the t and f distributions plotted above.

Confidence and Prediction Intervals

Confidence and Prediction intervals with Outliers:



Confidence Intervals without outliers:



Interpretation

There is significant evidence that the model estimating burn area from temperature is effective. (two-tail t-test, $df = 515, p = 0.02610141, \alpha = 0.05$).

Discussion

The objective of the project is to find evidence for an association between the weather in Montesinho Park and the severity of the forest fires it experiences. Many tests are performed in the study to determine the significance of association between the weather and severity variables. The overall results of the tests showed an association between weather and the severity of forest fires.

Summary of Findings

Randomization Test: Difference in Proportions Between Windy and Non-Windy Fires

Summary Table

Test Statistic ($\hat{p}_w - \hat{p}_n$)	CI	p-value
0.1513	[0.0668, 0.2356]	0.0001

The randomization test examines the association between wind level and ISI, an index for how quickly a fire spreads. The quantitative variable "wind" was separated into two categories "Windy" and "Not Windy." Each case was separated depending on whether the wind speed was above or below 3.75 kilometers per hour, which is an industry standard metric for "Windy" conditions. The quantitative variable "ISI" was also converted to a categorical variable based on its mean value. The two levels were "Large Spread" and "Small Spread." The wind level would help a fire spread by blowing embers and the base of the fire to adjacent fuel. This spreading process would continue as long as wind conditions remained adequately fast to push the fire. The wind pushing the fire would increase the ISI, resulting a faster spreading fire. The result of the difference in proportions test was a test statistic of $p_w - p_n = 0.1513$ with a p-value of 0.0001, which was found to be significant at a significance level of $\alpha = 0.05$. Thus, the first conclusion of the study is that the notion that windy conditions are not associated with a higher initial spread index could be rejected. In other words, it could conclude that windy conditions are associated with a faster spreading fire.

T Test: Difference in Means Between Follow-Up and Isolated Fires

Summary Table

Difference in Means	SE for distribution	CI	p-value
8.050202	3.04657	[-5.147, 6.689]	0.01769

The difference in means test determines whether the difference of means between isolated fires and follow-up fires is significant. Specifically, it tests if the burn area of fires with precedent burns within the past 6 months is smaller than the burn area of fires without such preceding fires. If there is a prior burn, much of the fuel is already used, so the area of the fire is smaller. While this explanatory variable is not weather related, it helps to detect whether a confounding variable is influencing the study. The test found that follow-up fires are influencing the study. The sample difference in means of 8.0502 is found to be significant to a significance level of $\alpha = 0.05$ with a p-value of 0.01769. Thus, the second conclusion of the study is that follow-up fires tend to have smaller burn areas. Because of follow-up fires tend to have smaller burn areas, whenever a test had a response variable of area, the data was split into two groups, "follow-up fires" and "isolated fires." Testing these two categories of fires separately controls for the confounding variable of closely preceding fires.

Chi Square Test

The chi-square test determines the association between the drought code (DC) and the burn area of the fire. The drought code is a measurement of the level of drought an area is experiencing. The burn area is the area that is directly burned by a fire in hectares. If the DC is high, or the area is in a drought, then there won't be as much natural moisture in the fuel, resulting in a larger burn. From the t test, the p-value of 0.01104994 was determined from a test statistic of $t = 2.38$. This result was proven significant at a significance level of $\alpha = 0.05$. Thus, the third conclusion of the study is that there is an association between drought code and burn area.

Linear Regression Test

The linear regression determines the temperature effects the burn area of a fire. The temperature is defined as the temperature in degrees Celsius at the time of the fire. The burn area of the fires is the area that is directly burned by the fires, and is defined in hectares. The temperature could help a fire, as it would not have as high of a differential between the outside temperature and the temperature it needs to continue burning. From the t test, the p-value of 0.02610141 was determined from a test statistic of $t = 2.231142$. This result was proven significant at a significance level of $\alpha = 0.05$. The result implies that temperature effects area. This conclusion comes with a not of caution, however. The Q-Q showed that the relationship

between temperature and area may be quadratic rather than linear. Therefore, the linear regression test would not be appropriate.

Overall Findings

The overall findings of the study partially agreed with the overarching hypothesis that weather is associated with the severity of a fire. Both the level of drought and temperature were shown to have a statistically significant association with the burn area of a fire. Moreover, the randomization test concluded that windy conditions result in faster spreading fires. The explanatory variables were metrics for weather conditions in the region. The response variables were all metrics of the severity of the fire. The resultant relationship showed that weather is associated with the severity of a fire.

Error Analysis

One possible error is that the data set is incomplete. There are several cases with N/A values for certain variables. These cases were removed for statistical analysis, resulting in a data set that is a smaller sample than before. Thus, the data set is not as representative of the population as it could have been.
Another potential error is that if a fire was large enough, it could have been to dangerous to collect data accurately for the fire. The result would be inaccurate recordings for large fires, potentially resulting in inaccurate conclusions when comparing large fires to small fires.
A further error is directly related to the linear regression test. The Q-Q plot, determined when checking assumptions, showed a potentially quadratic relationship between temperature and burn area. If the relationship between these variables is quadratic, then the linear regression test performed is inappropriate for the data and the relationship could not be determined.

Further Studies

Future studies are necessary to provide results with respect to the population of all forest fires. These studies could be achieved using the same data set, but applying stratified sampling to the randomization distributions. If a randomization distribution was developed for each test with explanatory variables consistent to the population of all forest fires, then the results would be representative of the population of all forest fires. Such results would benefit not only Montesinho Park, but all forest fires globally.
Future studies could also reanalyze the relationship between the temperature and the burn area of a fire. It would be most appropriate to recheck for a quadratic relationship between the variables.

References

Collins, Brandon M., et al. “A Quantitative Comparison of Forest Fires in Central and Northern California under Early (1911–1924) and Contemporary (2002–2015) Fire Suppression.” International Journal of Wildland Fire, vol. 28, no. 2, 2019, p. 138., doi:10.1071/wf18137./newline

“Explore Montesinho, One Of The Best Natural Parks in Portugal.” BePortugal, 13 Sept. 2019, beportugal.com/montesinho/.

“Fire Weather Index (FWI) System.” NWCG, www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system. UCI Machine Learning Repository: Forest Fires Data Set, archive.ics.uci.edu/ml/datasets/Forest+Fires.

Pierre-louis, Kendra, and Nadja Popovich. “Climate Change Is Fueling Wildfires Nationwide, New Report Warns.” The New York Times, The New York Times, 27 Nov. 2018,

www.nytimes.com/interactive/2018/11/27/climate/wildfire-global-warming.html.

Silva, Daniel. "Why Is Portugal so Prone to Wildfires?" Phys.org, Phys.org, 24 July 2019, phys.org/news/2019-07-portugal-prone-wildfires.html

"Wildfires Burn across California: Live Updates." CNN, Cable News Network, 30 Oct. 2019, www.cnn.com/us/live-news/california-fires-10-30-2019/index.html.

Appendix

Packages

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(knitr)
library(ggfortify)
library(xtable)
library(reshape2)
```

Randomization Test

```
forestfires = na.omit(read.csv("forestfires.csv"))
midpoint = 3.75
forestfires.copy = forestfires
forestfires = forestfires%>%
  mutate(wind = cut(wind, breaks=c(-1, midpoint, 100),labels =c("Not Windy", "Windy")),
         ISI = cut(ISI, breaks = c(0, mean(ISI, na.rm=T), Inf), labels=c("Small Spread", "Large Spread"))


forestfires.summary = forestfires %>%
  summarize(n.wind = nrow(filter(forestfires, wind == "Windy")),
            n.no.wind = nrow(filter(forestfires, wind=="Not Windy")),
            n.wind.large = nrow(filter(forestfires, wind=="Windy" & ISI=="Large Spread")),
            n.no.wind.large = nrow(filter(forestfires, wind=="Not Windy" & ISI == "Large Spread")),
            diff.prop = (n.wind.large / n.wind) - (n.no.wind.large / n.no.wind)
  )
forestfires = na.omit(forestfires)
forestfires.copy = na.omit(forestfires.copy)


big.plot.ISI = forestfires.copy %>%
  ggplot(aes(x=ISI)) +
  geom_density(color="darkblue",fill="darkblue",alpha = 0.1) +
  labs(x = "ISI",
       y = "Density",
       title = "Distribution of ISI") +
```

```
  theme_light() +
  theme(plot.title=element_text(size = 10),
        axis.title=element_text(size = 10),
        legend.title=element_text(size = 10),
        legend.position="none")

big.plot.wind = forestfires.copy %>%
  ggplot(aes(x=wind)) +
  geom_density(color="darkblue",fill="darkblue",alpha = 0.1) +
  labs(x = "wind",
       y = "Density",
       title = "Distribution of Wind") +
  theme_light() +
  theme(plot.title=element_text(size = 10),
        axis.title=element_text(size = 10),
        legend.title=element_text(size = 10),
        legend.position="none")

side.by.side.plot = forestfires %>%
  ggplot(aes(x=wind, fill = ISI, color=ISI)) +
  geom_bar(position='dodge', alpha = 0.5) +
  labs(x="Wind Conditions", y="Number of Fires with Large Spreads") +
  scale_fill_manual("legend", values = c("Small Spread" = "darkblue", "Large Spread" = "mediumaquamarine
  scale_color_manual("legend", values = c("Small Spread" = "darkblue", "Large Spread" = "mediumaquamarin
  theme_light()

stacked.plot = forestfires %>%
  ggplot(aes(x=ISI, fill=ISI, color=ISI)) +
  geom_bar(alpha = 0.5) +
  facet_wrap(wind~., nrow = 2) +
  labs(x = "Wind Conditions",y = "Number of Fires with Large Spreads") +
  theme_light() +
  scale_fill_manual("legend", values = c("Small Spread" = "darkblue", "Large Spread" = "mediumaquamarine
  scale_color_manual("legend", values = c("Small Spread" = "darkblue", "Large Spread" = "mediumaquamarin
  theme(axis.text = element_text(size=7))


gridExtra::grid.arrange(big.plot.wind, big.plot.ISI, ncol = 2, nrow = 1, top = grid::textGrob("Distribu
gridExtra::grid.arrange(side.by.side.plot, stacked.plot, ncol = 2, nrow = 1, top = grid::textGrob("Wind
```

Check for Assumptions

```
bootstrap.diffprop = c()
set.seed(302)
for (i in 1:10000) {
  sample.forestfires = sample_n(forestfires, size = nrow(forestfires), replace = T)
  sample.forestfires
  n.large.windy = filter(sample.forestfires, sample.forestfires$wind == "Windy" & sample.forestfires$ISI
  n.large.not.windy = filter(sample.forestfires, sample.forestfires$wind == "Not Windy" & sample.forest
  p.hat.windy = nrow(n.large.windy) / nrow(filter(sample.forestfires, sample.forestfires$wind == "Windy
  p.hat.not.windy = nrow(n.large.not.windy) / nrow(filter(sample.forestfires, sample.forestfires$wind ==
```

```
  diff.prop = p.hat.windy - p.hat.not.windy
  bootstrap.diffprop = append(bootstrap.diffprop, diff.prop)
}

bootstrap.diffprop = data.frame(bootstrap.diffprop)
bootstrap.correct = -mean(bootstrap.diffprop$bootstrap.diffprop)

randomization.diffprop = bootstrap.diffprop$bootstrap.diffprop + bootstrap.correct
randomization.diffprop = data.frame(randomization.diffprop)

p.hat.t1 = (nrow(filter(forestfires, forestfires$wind=="Windy" & forestfires$ISI == "Large Spread")) / r

df.melt = reshape2::melt(data.frame(Randomization = randomization.diffprop$randomization.diffprop,Boots
ci.95.ptest = quantile(randomization.diffprop$randomization.diffprop, c(.025, .975))

suppressWarnings(print(ggplot(df.melt,aes(x = value, fill = variable)) +
  geom_density(alpha = 0.25, adjust = 2) +
  theme_light() +
  labs(title = "Difference in Proportion of Fire Spreads",x = "Difference in Proportions", y = "Density
  scale_y_continuous(expand =c(0,0)) +
  scale_x_continuous(expand =c(0, 0)) +
  geom_vline(xintercept = ci.95.ptest, color = "#00BFC4",linetype = "dashed") +
  geom_vline(xintercept = p.hat.t1) +
  annotate("text", x = ci.95.ptest, y = 0.3, label =round(ci.95.ptest,3), fontface = "bold")  +
  scale_fill_manual("legend", values = c("Randomization" = "darkblue", "Bootstrap" = "mediumaquamarine"
  ))

z.score = p.hat.t1 / (sd(randomization.diffprop$randomization.diffprop))
```

Compute the P Value

```
p.value = mean(randomization.diffprop$randomization.diffprop>=p.hat.t1)
gg_randomization = randomization.diffprop %>%
  ggplot(aes(x=randomization.diffprop)) +
  labs(x = "ISI",
       y = "Density",
       title = "Distribution of ISI") +
  geom_density(color="darkblue",fill="darkblue",alpha = 0.1) +
  theme_light() +
  theme(plot.title=element_text(size = 10),
        axis.title=element_text(size = 10),
        legend.title=element_text(size = 10),
        legend.position="none")

gg.data = ggplot_build(gg_randomization)$data
temp = gg.data[[1]]%>%
  filter(x>=p.hat.t1)

gg_randomization = gg_randomization +
  geom_area(data = temp, aes(x=x, y=y, color="mediumaquamarine", fill="mediumaquamarine"), alpha = 0.5)
  theme(legend.position="none") +
  scale_fill_manual(values=c("mediumaquamarine", "mediumaquamarine")) +
```

```
  scale_color_manual(values = c("mediumaquamarine", "mediumaquamarine"))

gg_randomization
```

Summary Figure

```
# Read in dataset
forestfires = read.csv("forestfires.csv")
# Omit null values
forestfires = na.omit(forestfires)

# Separating data
forestfires.followup = data.frame()
forestfires.isolated = data.frame()
forestfires = na.omit(forestfires)

forestfires.testst = forestfires %>%
  mutate(
    month = recode(month,`jan`="1", `feb` = "2", `mar` = "3", `apr` = "4", `may` = "5", `jun` = "6", `ju
    diffmonth = abs(strtoi(month) - strtoi(lag(month)))
    )
forestfires.testst = forestfires.testst%>%
  mutate(followup = cut(diffmonth, breaks=c(-1, 6.5, Inf),labels =c("Follow-Up", "Isolated"))) %>%
  filter(!is.na(diffmonth))


forestfires.followup = filter(forestfires.testst, forestfires.testst$diffmonth>=6)
forestfires.isolated = filter(forestfires.testst, forestfires.testst$diffmonth<6)

# Summary Plots
# Area of both dotplot
boxplot.diffmeans = forestfires.testst %>%
  ggplot(aes(x = followup, y = area)) +
  geom_boxplot(aes(color=followup,fill=followup),alpha = 0.25,outlier.alpha = 0.5,outlier.shape = 19, ou
  labs(x = "Type of Fire",y = "Area Burned (hectares)") +
  theme_light() +
  theme(plot.title=element_text(size = 10),
        axis.title=element_text(size = 10),
        legend.title=element_text(size = 10),
        legend.position="none") +
  scale_fill_manual("legend", values = c("Follow-Up" = "darkblue", "Isolated"="mediumaquamarine")) +
  scale_color_manual("legend", values=c("Follow-Up"="darkblue","Isolated"="mediumaquamarine"))

# Histogram
histogram.diffmeans = forestfires.testst %>%
  ggplot(aes(x=area)) +
  geom_histogram(aes(color=followup,fill=followup), bins = 200, boundary = 0, alpha = 0.1) +
  facet_wrap(followup~., nrow = 2)+labs(x = "Area Burned (hectares)",y = "Count") +
  theme_light() +
  theme(plot.title=element_text(size = 10),
        axis.title=element_text(size = 10),
        legend.title=element_text(size = 10),
```

```
        legend.position="none") +
  scale_fill_manual("legend", values = c("Follow-Up" = "darkblue", "Isolated"="mediumaquamarine")) +
  scale_color_manual("legend", values=c("Follow-Up"="darkblue","Isolated"="mediumaquamarine"))

# Density
density.diffmeans = forestfires.testst %>%
  ggplot(aes(x=area)) +
  geom_density(aes(color=followup,fill=followup),alpha = 0.1) +
  facet_wrap(followup~., nrow = 2) +
  labs(x = "Area Burned (hectares)",y = "Density") +
  theme_light() +
  theme(plot.title=element_text(size = 10),
        axis.title=element_text(size = 10),
        legend.title=element_text(size = 10),
        legend.position="none") +
  scale_fill_manual("legend", values = c("Follow-Up" = "darkblue", "Isolated"="mediumaquamarine")) +
  scale_color_manual("legend", values=c("Follow-Up"="darkblue","Isolated"="mediumaquamarine"))

# Arrange them
gridExtra::grid.arrange(boxplot.diffmeans, density.diffmeans, histogram.diffmeans, ncol = 3, nrow = 1,
```

Calculate Test Statistic

```
forestfires.ptest.summary = forestfires %>%
  summarize (
     mu.followup = mean(forestfires.followup$area),
     mu.isolated = mean(forestfires.isolated$area),
     diff.means = mu.isolated - mu.followup,
     se = sqrt(((sd(forestfires.followup$area) ** 2) / nrow(forestfires.followup)) + ((sd(forestfires.is

  )
df = min(nrow(forestfires.followup), nrow(forestfires.isolated))
#t.test(forestfires.isolated$area, forestfires.followup$area, alternative = "two.sided")
```

Chi-Square test: Drought code vs. Area

```
forestfires = read.csv("forestfires.csv")
forestfires = na.omit(forestfires)
forestfires = forestfires %>%
  mutate(DC = cut(DC,breaks=c(-Inf, median(DC, na.rm=T) / 3, median(DC, na.rm=T) * 0.666, Inf),labels =
         area = cut(area, breaks=c(-Inf, median(area, na.rm=T), Inf), labels=c("Small", "Large")),
         count.dry = nrow(filter(forestfires, DC == "Dry" & area =="Large")),
         count.neutral = nrow(filter(forestfires, DC=="Neutral" & area == "Large")),
         count.wet = nrow(filter(forestfires, DC=="Wet" & area == "Large")))

forestfires.table = table(forestfires$DC, forestfires$area)
dimnames(forestfires.table) = list(DC = c("Wet", "Neutral", "Dry"), area = c("Small", "Large"))
kable(forestfires.table)
```

Summary Figure

```
melt(forestfires.table) %>%
  ggplot(aes(x=DC,y=value,fill=area)) +
  geom_bar(aes(color=area), position="dodge", stat="identity",alpha = 0.2) +
  labs(x = "Drought Code", y = "Count", title = "Drought Code v.s. Fire Burn Area", fill = "Fire Burn A
  theme_light() +
  scale_fill_manual("legend", values=c("Small"="darkblue", "Large"="mediumaquamarine")) +
  scale_color_manual("legend", values=c("Small"="darkblue", "Large"="mediumaquamarine")) +
  theme(plot.title=element_text(size = 10),axis.title=element_text(size = 10),legend.title=element_text
```

Check for Assumptions

```
rows = nrow(forestfires.table)
cols = ncol(forestfires.table)
row.sums = rowSums(forestfires.table)
col.sums = colSums(forestfires.table)
n = sum(forestfires.table)
exp.counts = outer(row.sums,col.sums,"*")/n
dimnames(exp.counts) = list(DC = c("Wet", "Neutral", "Dry"), area = c("Small", "Large"))
kable(exp.counts)
```

Calculate Test Statistic

```
## Calculate Test Statistic
chi.sq = sum((forestfires.table-exp.counts)^2/exp.counts)
```

Compute the P Value

```
## Compute p-value
pchisq(chi.sq,(rows-1)*(cols-1),lower.tail=FALSE)
chisq.test(forestfires.table, correct = FALSE)
```

Linear Regression Test: Temperature vs. Burn Area

Summary Figure

```
forestfires = read.csv("forestfires.csv")
forestfires = na.omit(forestfires)
summary.fig.lr = forestfires %>%
  ggplot(aes(x = temp, y=area)) +
  geom_point() +
  labs(x = "Temperature in Degrees Celsius", y = "Burn Area in Hectares", title = "Temperature vs. Burn
  geom_smooth(method ="lm", se = FALSE) +
```

```
  theme_light() +
  theme(plot.title = element_text(size = 10),
        axis.title = element_text(size = 10),
        legend.title = element_text(size = 10))


summary.fig.temp.distribution = forestfires %>%
  ggplot(aes(x = temp)) +
  labs(x = "Temperature in Degrees Celsius", y = "Density", title = "Distribution of Temperature") +
  geom_density(color = "darkblue", fill = "darkblue", alpha = 0.1) +
  theme_light() +
  theme(plot.title = element_text(size = 10),
        axis.title = element_text(size = 10),
        legend.title = element_text(size = 10))

summary.fig.area.distribution = forestfires %>%
  ggplot(aes(x = area)) +
  labs(x = "Burn Area in Hectares", y = "Density", title = "Distribution of Area") +
  geom_density(color = "darkblue", fill="darkblue", alpha = 0.1) +
  theme_light() +
  theme(plot.title = element_text(size = 10),
        axis.title = element_text(size = 10),
        legend.title = element_text(size = 10))

gridExtra::grid.arrange(summary.fig.lr, summary.fig.area.distribution, summary.fig.temp.distribution, ni
```

Check for Assumptions

```
reg.line = lm(area ~ temp, forestfires)

autoplot(reg.line) +
  theme_light()
```

Calculate Test Statistic

```
kable(xtable(summary(reg.line)))
```

```
kable(xtable(anova(reg.line)))
```

Compute the P Value

```
t = 1.072628 / 0.4807528
f = 20016.828 / 4021.064

x = seq(-5, 5, length=100)
y = dt(x, 515)
t.dist.plot = ggplot(data = data.frame(x, y), mapping = aes(x = x, y = y)) +
```

```
    geom_line() +
    geom_area(mapping = aes(x = x, y = y), fill = "darkblue", alpha = 0.1) +
    labs(x = "t", y="Density", title = "T Distribution with Degrees of Freedom of 515") +
    geom_vline(xintercept = 0) +
    geom_hline(yintercept = 0) +
    theme(plot.title = element_text(size = 8),
          legend.title = element_text(size = 8),
          axis.title = element_text(size = 8))

t.data = ggplot_build(t.dist.plot)$data

lower = t.data[[1]] %>%
  filter(x <= -t)

upper = t.data[[1]] %>%
  filter(x >= t)

t.dist.plot = t.dist.plot +
  geom_area(data = lower, aes(x = x, y = y, color = "darkblue", fill = "darkblue"), alpha = 0.5, na.rm =
  geom_area(data = upper, aes(x = x, y = y, color = "darkblue", fill = "darkblue"), alpha = 0.5, na.rm =
  theme(legend.position="none") +
  scale_fill_manual(values=c("darkblue", "darkblue")) +
  scale_color_manual(values = c("darkblue", "darkblue"))

f = 20016.828 / 4021.064
x = seq(0, 10, length=200)
y = df(x, df1 = 1, df2 = 515)


f.plot = ggplot(data = data.frame(x, y), mapping=aes(x=x, y=y)) +
    geom_line() +
    geom_area(mapping=aes(x=x, y =y), fill="darkblue", alpha = 0.1) +
    labs(y = "Density", x = "f", title="F Distribution with Degrees of Freedom of 1, 515") +
    geom_vline(xintercept=0) +
    geom_hline(yintercept=0) +
    theme(plot.title = element_text(size = 8),
          legend.title = element_text(size = 8),
          axis.title = element_text(size = 8))

f.plot.data = ggplot_build(f.plot)$data

f.plot.area = f.plot.data[[1]] %>%
  filter(x >= f)

f.plot = f.plot +
  geom_area(data = f.plot.area, aes(x=x, y=y, color="darkblue", fill="darkblue"), alpha = 0.5) +
  theme(legend.position="none") +
  scale_fill_manual(values=c("darkblue", "darkblue")) +
  scale_color_manual(values = c("darkblue", "darkblue"))

gridExtra::grid.arrange(t.dist.plot, f.plot, ncol = 2, nrow = 1, top = grid::textGrob("T Distribution a
```

```
pt(t, 515, lower.tail = FALSE) + (1- pt(t, 515, lower.tail = TRUE))
pf(f, df1 = 1, df2 = 515, lower.tail = FALSE)
```

Confidence and Prediction Intervals

```
forestfires = read.csv("forestfires.csv")
forestfires = na.omit(forestfires)
conf.set = predict(reg.line, forestfires, interval="confidence")
conf.set = data.frame(conf.set)
forestfires = forestfires %>%
  mutate(lower.ci = conf.set$lwr, upper.ci = conf.set$upr)

plot.confidence = forestfires %>%
  ggplot(aes(x=temp,y=area, ymin=lower.ci, lwr, ymax=upper.ci)) +
  geom_ribbon(fill="darkblue",alpha = 0.1) +
  geom_point() +
  geom_smooth(method="lm",se = FALSE) +
  labs(x = "Temperature in Degrees Celsius", y="Burn Area in Hectares", title = "Confidence Interval") +
  theme_light()

pred.set = predict(reg.line, forestfires,interval="predict")
pred.set = data.frame(pred.set)

forestfires = forestfires = forestfires %>%
  mutate(lower.ci = pred.set$lwr, upper.ci = pred.set$upr )

plot.prediction = forestfires %>%
  ggplot(aes(x=temp,y=area, ymin=lower.ci, lwr, ymax=upper.ci)) +
  geom_ribbon(fill="darkblue",alpha = 0.5) +
  geom_point() +
  geom_smooth(method="lm",se = FALSE) +
  labs(x = "Temperature in Degrees Celsius", y="Burn Area in Hectares", title = "Prediction Interval") +
  theme_light()

gridExtra::grid.arrange(plot.confidence, plot.prediction, nrow = 1, ncol = 2, top = grid::textGrob("Con
```

```
forestfires = na.omit(read.csv("forestfires.csv"))

forestfires = filter(forestfires, forestfires$area<(1.5 * quantile(forestfires$area, 0.75)))

reg.line = lm(area~temp, forestfires)
#reg.line = lm(area~temp, forestfires)

conf.set = predict(reg.line, forestfires, interval="confidence")
conf.set = data.frame(conf.set)
forestfires = forestfires %>%
  mutate(lower.ci = conf.set$lwr, upper.ci = conf.set$upr)

plot.confidence = forestfires %>%
  ggplot(aes(x=temp,y=area, ymin=lower.ci, lwr, ymax=upper.ci)) +
  geom_ribbon(fill="darkblue",alpha = 0.1) +
```

```
  geom_point() +
  geom_smooth(method="lm",se = FALSE) +
  labs(x = "Temperature in Degrees Celsius", y="Burn Area in Hectares", title = "Confidence Interval") +
  theme_light()

pred.set = predict(reg.line, forestfires,interval="predict")
pred.set = data.frame(pred.set)

forestfires = forestfires %>%
  mutate(lower.ci = pred.set$lwr, upper.ci = pred.set$upr )

plot.prediction = forestfires %>%
  ggplot(aes(x=temp,y=area, ymin=lower.ci, lwr, ymax=upper.ci)) +
  geom_ribbon(fill="darkblue",alpha = 0.1) +
  geom_point() +
  geom_smooth(method="lm",se = FALSE) +
  labs(x = "Temperature in Degrees Celsius", y="Burn Area in Hectares", title = "Prediction Interval") +
  theme_light()

gridExtra::grid.arrange(plot.confidence, plot.prediction, nrow = 1, ncol = 2, top = grid::textGrob("Con
```