STATISTICAL ANALYSIS OF FOREST FIRE AREAS IN PORTUGAL

STAT0032: GROUP PROJECT 2021-22

GROUP NUMBER: 14

Student IDs: 21111357,21169262,21080157,21134487,18017060

Introduction

The occurrences of forest fires due to anthropogenic climate change have become increasingly concerning, particularly for countries in the Mediterranean Basin. France, Greece, Italy, Portugal, Spain, and Turkey contribute to around 80% of the area burned in the European continent each year. The number of large forest fires (> 500 hectares) continue to grow and have been exacerbated by climate change despite a decrease in total area burned and the number of fires since the 1980s (WWF, 2019). It's been estimated that the European continent incurs losses of around 3 billion Euros per year due to forest fires. In addition to financial loss, Portugal, Spain, and Greece lost 225 people to such extreme fires between 2017-2018 alone (WWF, 2019).

Forest fires are more prominent in summer (June-September) as 90% of the area burned in Portugal occurs during this period (Mateus & Fernandes, 2014). To better understand the behavior of forest fires, we will be performing statistical analysis of the data distributions of the area of these fires for August and September. How these behave and differ will help us inform the government of how forest fires develop over the year and the appropriate actions they should take.

Discussion of the problem

To determine whether or not the distribution of forest fire sizes for both August and September follow a log-normal distribution, we will perform the Lilliefors and the Shapiro-Wilk tests for both months. Knowing the probability distribution that the forest fires' area follows will allow the government to further understand the fires' behavior and make better decisions to mitigate their effects.

We will also perform a Welch's t-test to compare the means of the distributions and a Kolmogorov–Smirnov (KS) test to determine whether the data from the two months follow the same probability distribution. In the case that the area of fires for both months follows the same distribution, the government can take actions for implementing public policy decisions regarding the impacts of fires for both months.

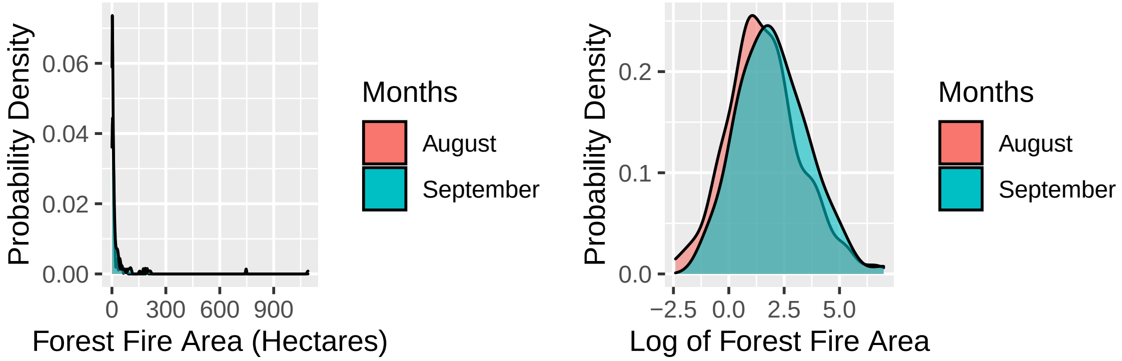
Presentation of the dataset

The dataset used for this project is the "Forest Fires Data Set" from the UCI Machine Learning Repository and contains forest fires information from the Montesinho Natural Park in Portugal between January 2000 and December 2003 (Cortez & Morais, 2007). We will be performing our analysis on the area (hectares) of given fires for the months August and September. Since our primary concern is large fires, we will not consider any fire with a size of 0.01 hectares (ha) or smaller. In total, 99 fires were recorded for August, with the smallest being 0.09 ha, the largest being 746 ha, and the average being 23.2 ha. September recorded 97 fires, with 0.33 ha being the smallest, 1091 being the largest, and the average being 31.8 ha.

To better understand the distributions of the two data sets we transformed the original data by applying the natural logarithm to the area. The graph on the left in Graph 1 shows the original data distribution, while the graph on the right shows the distribution of the area and the log of the area for forest fires in August (red) and September (blue).

For all statistical tests conducted in this report, we used a significance level of $\alpha = 0.05$, to reduce the amount of Type I errors while keeping the power of the test at a satisfactory level (Freeman & Zeegers, 2016). All statistical tests were performed using packages in R and Python.

Graph 1. Distribution of Forest Fire Areas for August and September



Description of the Log-normal distribution tests:

This section of the report aims to test whether the area of forest fires in the August and September follows a log-normal distribution or not. The log-normal distribution is continuous with the following probability density function (PDF):$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log x-\mu)^2}{2\sigma^2}\right]$ where μ is the distribution mean and σ is the standard deviation (Kissell & Poserina, 2017).

Moreover, the natural logarithm of random variable X that follows a log-normal distribution ($log(X)$) is normally distributed. Similarly, the exponential of a normally distributed random variable ($exp(y)$) follows a log-normal distribution. (Maymon, 2018). Therefore, using the natural logarithm of the area of fires allows for testing for normality instead of the log-normality of the area itself.

* Lilliefors Test

The first test we performed was the Lilliefors goodness-of-fit test (LF test) on the logarithm of the area for each month. This measures the distance between the cumulative distribution function (CDF) of the normal distribution with the EDF of the observed data(Razali & Wah, 2011). This test can be used when the parameters of the normal distribution are unknown and the estimate of the mean μ is the sample mean $\bar{X}$ and the estimate of the variance σ^2 is the sample variance (s^2) (Razali & Wah, 2011). The sampling distribution of the LF test is obtained using Monte-Carlo techniques rather than an analytical solution.

Null hypothesis: $H_0 : F(X) = F^*(X)$ for $-\infty < X < \infty$. The natural logarithm of forest fire areas ($log(area)$) follow a normal distribution. $F(X)$ is the distribution of $log(area)$ and $F^*(X)$ is the normal distribution

Alternative hypothesis: $H_1 : F(X) \neq F^*(X)$. The natural logarithm of forest fire areas ($log(area)$) do **not** follow a normal distribution.

Test statistic: $D = max_x|F^*(X) - Sn(X)|$ Where $F^*(X)$ is the normal cumulative distribution function (CDF) whose mean is the sample mean, and the variance is the sample variance. $Sn(X)$ is the sample empirical cumulative distribution function.

The test statistics and p-value for August were [0.064, 0.418], and [0.056, 0.65] for September respectively. The p-values are larger than the specified level of the test (0.05) in both months so there is not sufficient evidence to suggest that the natural logarithm of the forest fire area is not normally distributed or does not follow a log-normal distribution.

* Shapiro-Wilk test

Shapiro-Wilk test is used to test a sample of unknown mean and variance for normality (Razali & Wah, 2011). We have opted for a modified version suitable for sample sizes between 3 and 5000 based on modifications done by Royston (1995). We conducted this test on the log of the area for each month. The test statistic, W, is essentially a ratio of estimates of a normal distribution for a random sample of n observations (Royston,1995).

Null hypothesis: $H_0 = F(x)$ the distribution of the transformed area variable ($log(area)$) follows a normal distribution with unspecified variance and mean.

Alternative hypothesis: $H_1 = F(x)$ the distribution of the transformed area variable ($log(area)$) does not follow a normal distribution.

Test statistic: $W = \frac{(\sum\limits_{i=1}^{n} a_i y_{(i)})^2}{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2}$. Where, " a_i denotes the coefficients defined by $a' = (a_1, ..., ..a_n) = m'V^{-1}[(m'V^{-1})(V^{-1} - m)]^{\frac{-1}{2}}$ where $m' = (m_1, m_2 ...., m_n)$ denote the vector of expected values of standard normal order statistics, and let $V = (v_{ij})$ be the corresponding n x n covariance matrix" (Shapiro & Wilk, 1965). The denominator corresponds to sample variance S^2 and the W test statistic has a value between zero and one, where low values indicate departure from normality.

The obtained test statistics (W) and p-values were, [0.988, 0.513] for August, and [0.985,0.351] for September respectively. Since both p-values are higher than our specified significance level of 0.05, we do not have enough evidence to reject the null hypothesis H_0.

In comparison with the LF test, the SW test has a higher power. Both the LF and the SW tests are normality tests, and they do not require any knowledge about the parameters of the normal distribution. The LF test is based on the largest difference between the empirical CDF of the data and a normal distribution with parameters set as the sample mean and sample standard deviation. (Razali & Wah, 2011). In comparison, the SW test has its origins rooted in probability plotting and it's test statistic W can be interpreted as an approximation of the "straightness" of normal Q-Q plot (Royston, 1995).

Description of the two sample tests:

This section of the report aims to determine whether the area of forest fires in August and September come from the same population or not. To do this, we conducted a test of equal means, and a test to quantify the distance between the EDF for each month. We follow a standard hypothesis testing procedure for each of the tests and calculate p-values for each test statistic.

* Welch's t-test

For the test of equal means, we performed a Welch's t-test where the data is assumed to be normally and independently distributed (Welch, 1951). Each distribution has a mean and variance associated with them; however, they are not considered to be known or equal. The Welch's t-test is used when the variance of the two populations is different or when the two sample sizes differ (Rasch, Teuscher, and Guiard 2007). This test was conducted under the assumption that the log of the areas for each month is normally distributed.

Null hypothesis: $H_0 : \mu_{aug} = \mu_{sep}$

Alternative hypothesis: $H_1 : \mu_{aug} \neq \mu_{sep}$ where μ_i represents the mean for the log of the areas for each month.

Test statistic: t = $\frac{\bar{X_1}-\bar{X_2}}{\sqrt{\frac{s^2_{X_1}}{n_1}+\frac{s^2_{X_2}}{n_2}}}$. Here, n_i is the sample size, $\bar{X_i}$ is the sample mean, and s_i is the sample standard deviation for X_i. The Welch's t-test statistic follows a t distribution under the null hypothesis.

The obtained test statistic and p-value [1.868, 0.063] suggest there is not enough evidence to reject the null hypothesis that the two datasets have the same mean at the specified significance level (0.05).

* Kolmogorov-Smirnov test

To compare the empirical distribution functions (EDF) of two data samples, we performed a Kolmogorov-Smirnov test (KS test) for August and September. Since the KS test is non-parametric, we do not make assumptions about the data distribution (Conover, 1999). Instead, we assume the areas from both months are random, mutually independent, and continuous, according to the data gathering methodology of forest fires (Cortez & Morais, 2007).

Null hypothesis: $H_0 : F(x) = G(x)$. The distribution function of the area of forest fires from August is the same as that of forest fires from September.

Alternative hypothesis: $H_0 : F(x) \neq G(x)$. The distribution function of the area of forest fires in August is **not** the same as the distribution function of forest fires from September.

Test statistic: the test statistic calculates the greatest vertical distance between the two empirical distribution functions (August and September): $T_1 = sup_x|F(X) - G(X)|$

The obtained test statistic and p-value [0.176, 0.098] suggest there is not enough evidence to reject the hypothesis that the forest fires in August follow the same distribution as the forest fires in September.

Conclusion

In the first section, the Lilliefors and the Shapiro-Wilk tests resulted in a p-value greater than our specified level of significance (0.05). Therefore, there is not enough evidence to reject the null hypothesis that the natural logarithm of the area for both months follows a normal distribution. In the second section, the Welch's t-test and the Kolmogorov-Smirnoff tests resulted in a p-value greater than our prespecified significance level (0.05). Therefore we do not have enough evidence to reject the null hypothesis that both months have the same mean and follow the same distribution. Finally, it is worth noting that the Lilliefors test has lower power compared to the Shapiro-Wilk test (Razali & Wah, 2011).

The performed tests and analysis on forest fires in Portugal for August and September have helped understand the behavior of the fires. The gained knowledge of the distributions of the fire sizes will provide an insight on how to model future fires and apply solutions to mitigate the impact of these fires. Prediction of forest fire sizes can be achieved by implementing parameter estimation techniques such as Maximum Likelihood Estimation, which will, in turn, help the better government prepare for the future.

There are some limitations with our method for statistical analysis. For example, there was no prior knowledge about the parameters of the lognormal distributions. If we had this knowledge, it could have allowed further analysis or predictions of forest fire sizes. In addition, the study was constrained to two months, August and September of the same year. A more representative sample would include data from previous years.

References to relevant literature

- Abdi, H. & Molin, P. (2007) Lilliefors/Van Soest's test of normality. Encyclopedia of measurement and statistics. Neil Salkind (Ed.). Sage, Thousand Oaks, California

- Changyong, F. E. N. G., Hongyue, W. A. N. G., Naiji, L. U., Tian, C. H. E. N., Hua, H. E., & Ying, L. U. (2014). Log-transformation and its implications for data analysis. Shanghai archives of psychiatry, 26(2), 105.

- Comprehensive R Archive Network (CRAN). (2015, July 30). Tests for normality [R package nortest version 1.0-4]. The Comprehensive R Archive Network. Retrieved December 12, 2021, from https://cran.r-project.org/web/packages/nortest/index.html.

- Conover, W. J. (1999). Practical nonparametric statistics (Vol. 350). John Wiley & sons.

- Cortez, P., & Morais, A. D. J. R. (2007). A data mining approach to predict forest fires using meteorological data.

- D'Agostino, R. B. (1986). Goodness-of-fit-techniques (Vol. 68). CRC press.

- Fernandes, P. M., Barros, A. M., Pinto, A., & Santos, J. A. (2016). Characteristics and controls of extremely large wildfires in the western Mediterranean Basin. Journal of Geophysical Research: Biogeosciences, 121(8), 2141-2157.

- Freeman, M. D., & Zeegers, M. P. (2016). Forensic epidemiology: Principles and practice. Elsevier Academic Press.

- Gillett, N. P., Weaver, A. J., Zwiers, F. W., & Flannigan, M. D. (2004). Detecting the effect of climate change on Canadian forest fires. Geophysical Research Letters, 31(18).

- Hantson, S., Pueyo, S., & Chuvieco, E. (2016). Global fire size distribution: From power law to log-normal. International Journal of Wildland Fire, 25(4), 403. https://doi.org/10.1071/wf15108

- Kissell, R. L., & Poserina, J. (2017). Optimal sports math, statistics, and fantasy. Academic Press.

- Mateus, P., & Fernandes, P. M. (2014). Forest fires in Portugal: Dynamics, causes and policies. Forest Context and Policies in Portugal, 97–115. https://doi.org/10.1007/978-3-319-08455-8_4

- Maymon, G. (2018). Stochastic crack propagation: Essential practical aspects. Academic Press, an imprint of Elsevier.

- Rasch, D., Teuscher, F., & Guiard, V. (2007). How robust are tests for two independent samples?. Journal of statistical planning and inference, 137(8), 2706-2720.

- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics, 2(1), 21-33.

- Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. Journal of the Royal Statistical Society: Series C (Applied Statistics), 31(2), 115-124.

- scipy.stats.shapiro — SciPy v1.7.1 Manual. Retrieved 4 December 2021, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html

- Shapiro, S., & Wilk, M. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591. doi: 10.2307/2333709

- Simard, R., & L'Ecuyer, P. (2011). Computing the two-sided Kolmogorov-Smirnov distribution. Journal of Statistical Software, 39(11), 1-18.

- Welch, B. L. (1951). On the Comparison of Several Mean Values: An Alternative Approach. Biometrika, 38(3/4), 330–336. https://doi.org/10.2307/2332579

- World Wide Fund for Nature (2019). The Mediterranean burns (pp. 2-6). WWF. Retrieved from https://www.wwf.es/?51162%2FThe-Mediterranean-burns-2019.

Plagiarism and Collusion

All members of the group have read and understood the “Plagiarism and Collusion” section in the Taught Postgraduate Student Handbook for the Department of Statistical Science.

Contributions

All members of the group had an equal contribution to the ICA.