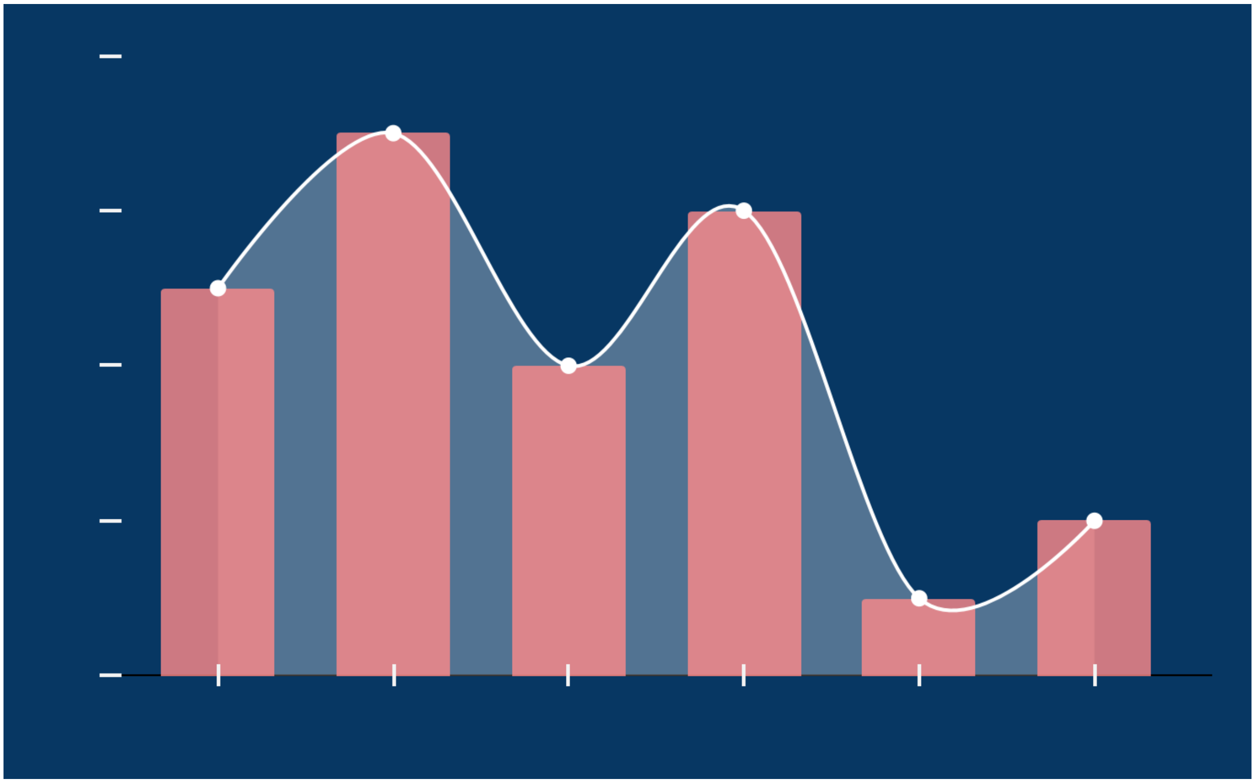
Post-Conflict Environmental Assessment - Section II (HW6)

Human Interaction Effects Analysis



Kya Allen

03.15.2021
301-919-7902

Arendelle Department of Environmental Safety

INTRODUCTION

In the Aftermath of the Arendellian-Atohallan Conflict, there have been reports of leftover radioactive hot spots across the border that separated Arendell and the formerly-Atohallan territory of the continent. The Radioactivity of these sites pose a serious risk to the health of people living in the area, or passing through. In this study we look at time-to-event (TTE) data on the time it takes for people to report unusual health effects after exposure to the radioactive sites. In preparation for further analysis, we'll be testing the assumption that the data involved is exponentially distributed, and conducting a preliminary test for the average TTE, as well as using power analysis to determine what sample sizes to strive for in future analyses.

Methodology

Our data is split into two sections. One North-Border sample and one South-Border sample. For The Larger North data set we conduct a Chi-Square goodness of fit test to compare the empirical distribution to the theoretical distribution. For the Small South Dataset we Test for linearity with a Q-Q plot regressing the observed data on the expected data from the theoretic exponential distribution. Then a standard Student's T-test is used to test whether or not the TTE is below 3.2 weeks, which is a threshold commonly used by other Health and Environmental Agencies to declare an environmental health emergency. (Analysis process is shown in the python notebooks attached to this submission)

Analysis I - North Border Data (N = 50)

To conduct this analysis, We created bins for the data and the theoretic distribution. We compiled counts of the frequency of observed data within each bin, then we compiled the expected counts from the theoretic exponential distribution for each of those bins. These theoretical frequencies were obtained first using the following formula from class:

$$n\left(1 - e^{-c_1/\theta}\right) \text{ and } \frac{n}{\theta}\left[e^{-c_i/\theta} - e^{-c_{i+1}/\theta}\right]$$

(Where the C term represents the interval) Using this Equation returned highly un-intuitive results. I expect I did something wrong but I can't figure out why. I will report these results along with the results of finding the integral of the distribution between the intervals.

$$\frac{n}{\theta} e^{-\frac{x}{\theta}}$$

After obtaining the counts, A Chi-Square test statistics was found using the following equation

$$X^2 = \frac{(Observed - Expected)^2}{Expected}$$

The results of these statistical tests are reported in the results section.

Analysis II -South Border data (N=20)

To test the smaller data set, we used a test for linearity on the Q-Q plot. First the observed data was ordered, and set to the Y-axis, then the Expected data for the Exponential distribution was calculated using the following Equation, and set as the X-axis

$$\theta\left[\frac{1}{n} + \frac{1}{n-1} + \quad \cdots\cdot \quad + \frac{1}{n-i+1}\right]$$

If the Data were perfectly exponential, we would expect it to have a perfectly linear relationship to a theoretical exponential distribution. The further off it is, the more likely the data does not come from an exponential distribution. At a glance, There appears to be some nonlinearity in the data, But we can do better than just intuiting from a visual representation.

Next, We use a polynomial ordinary least squares (OLS) regression. We'll conduct a Hypothesis test on the significance of the coefficient of the quadratic term to see if the nonlinearity of the data is significant.

$$H_0 \colon \widehat{Y} = \beta_0 + \beta_1 X + \epsilon$$

$$H_1 \colon \widehat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

To test this Hypothesis by performing a quadratic regression (the second equation) on the data with ordinary least squares. This can be done with calculus by taking the partial derivatives of the loss function with respect to beta-1 and beta-2, setting the system of equations to zero, and then solving for the Beta's. The loss function is the Sum of Squared Errors (SSE), and this method minimizes this quantity.

$$SSE = \sum_{i=0}^{n} (Y_i - \widehat{Y}_i)^2$$

Alternatively, the Beta's can be found algebraically with Matrix Algebra

$$\beta = (X^T X)^{-1} X^T Y$$

Where each term here is a Vector or Matrix of the data it represents. For the purpose of this Analysis, We computed the regression using the OLS function in the Statsmodels api for python. The package automatically computes the following T-Test on each predictor variable.

$$T = \frac{\widehat{\beta}_i - 0}{SE(\widehat{\beta}_i)}$$

Once again, the results of this will be reported in the Results Section

Analysis III - North Border Mean TTE (N=50)

For this section, We simply used a Student's T-test to test whether to reject the null hypothesis that the TTE is < 3.2, the results are reported in the results section

DATA - Analysis IV - Power Analysis

For the power Analysis, we used the TTestPower Package from the Statsmodels python API. 3 sample sizes were computed with respect to a desired Power of 95%, A desires Alpha of 0.05, and three separate Effect Sizes. (0.2, 0.5, and 0.8 for Small, Medium, and Large effect sizes)

RESULTS

Table 1				
Chi-Square Goodness of Fit test for an Exponential fit on North Border TTE data				
	df	Statistic	R.R	Alpha
Method from Class	16	146.00	> 26.296	0.05
Integration Method	16	18.16	> 26.296	0.05

Using the Method from class which I seem to have mishandled, We would reject the null hypothesis that the data comes from an exponential distribution. Using the integral based method for finding the expected frequencies, A more intuitive result is given, where we fail to reject the null hypothesis that the data comes from an exponential distribution.

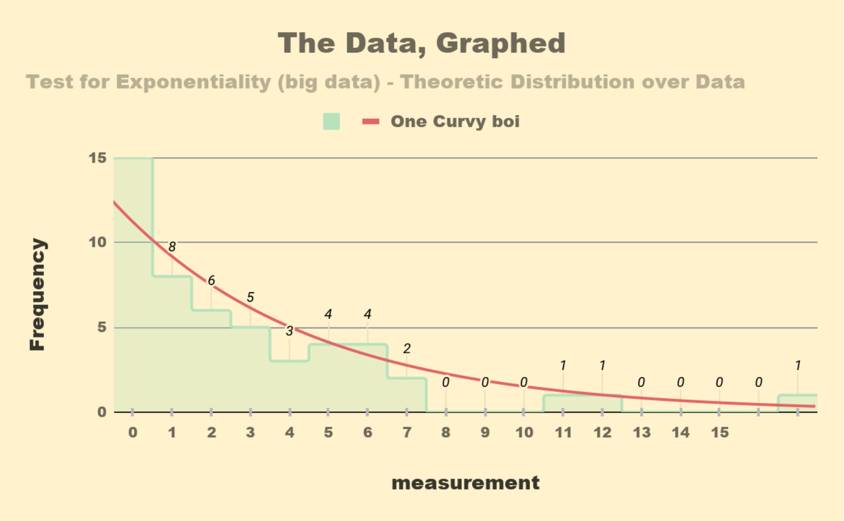
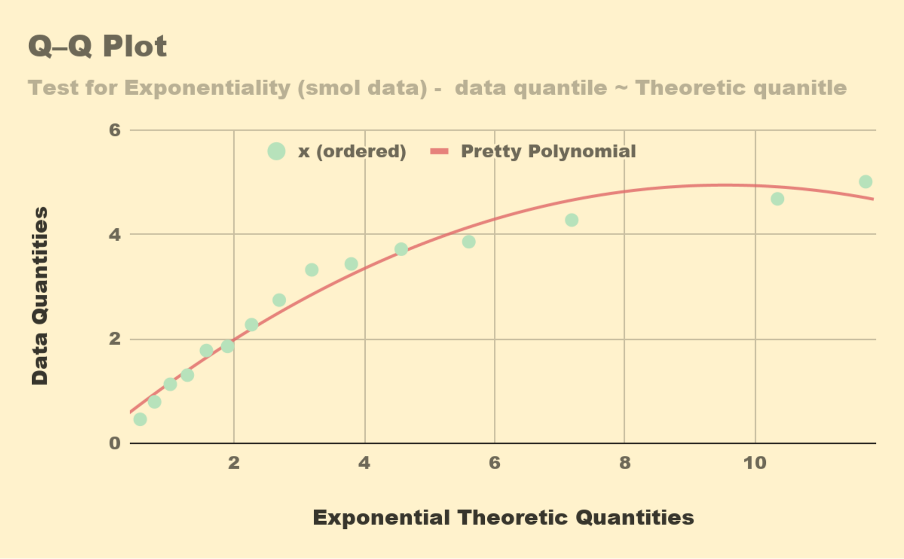
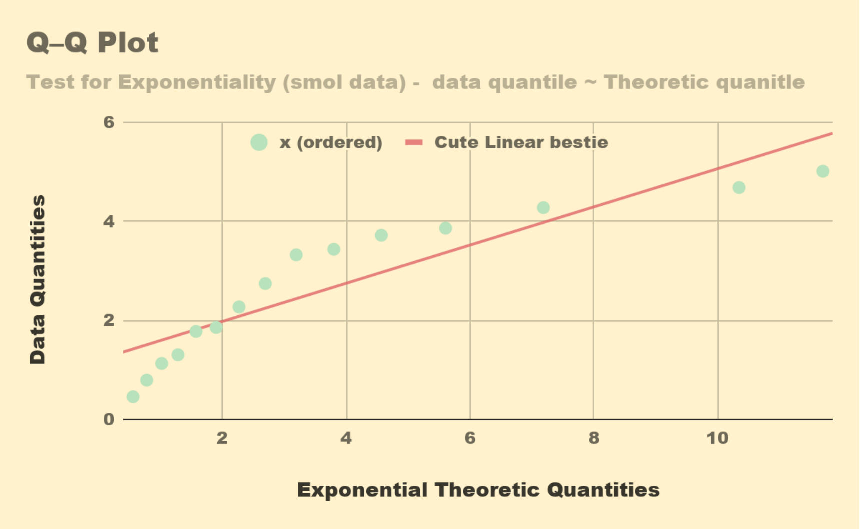


Table 2				
Regression Analysis Summary for South border data ~ Theoretic exponential Quantile				
Predictor Variable	Coefficient	Standard Error	T-Statistic	P-value
Intercept	0.2086	0.165	1.261	0.231
Exponential Quantile	0.9932**	0.08	12.363	< 0.001
Exponential Quantile ^ 2	-0.052**	0.007	-7.827	< 0.001

* indicates p<0.05. ** indicates p<0.01





The Quadratic term in the polynomial regression was found to be statistically significant, so we reject the null hypothesis that the data and the exponential theoretical quantiles have a linear relationship, and thus we reject the hypothesis that the data is from an exponential distribution.

Table 3 - Null: TTE < 3.2					
One Sample T-Test on the North Border TTE Data (N=50)					
Sample	Mean	Std Deviation	df	T-statistic	R.R
North Border TTE	3.381	3.502	49	0.126	1.68

We fail to reject the hypothesis that the mean Time-To-Event is less than 3.2 at a 0.05 alpha.

Table 4

Power Analysis for Ideal Sample Size

Statistical Test	Desired Power	Desired Alpha	Sample Size by effect size		
			Small (0.2)	Medium (0.5)	Large (0.8)
one sample T-test	95%	0.05	272	45	19

CONCLUSION

Our results drive us toward a conclusion that the Time until people interacting with radioactive zones start reporting adverse health effects is exponentially distributed to the North of the Border, but not South of the border. There could be some differences in the distribution of radioactive hot spots, or a difference in the behavior of inhabitants, leading to this difference. Alternatively The small sample size may have led to a distorted distribution in our south border sample, by chance. Further Analysis should be done to come to a more robust conclusion.

Our data on the mean TTE for the North Border suggests that the agency should declare an Environmental Health Emergency for this zone.

For our power analysis for ideal sample sizes, we should pay the most attention to the small effect size sample since the mean TTE we collected was only slightly above the threshold. There may be a true difference that is small, and difficult to detect with the sample size we used.