MAT 243 Final Project Part 1: Summary Report

Milestone 1

Scenario A: National Weather Service

Jason Lima

Jason.lima1@snhu.edu

Southern New Hampshire University

**1. Statement of the Problem**

The purpose of this report is to analyze monthly historical weather data for the city of Manchester in New Hampshire. The findings in this analysis will be used to update weather pattern models based on historical trends and finding will published in academic journals. Using statistical models and certain variables of interest, we can calculate descriptive statistics to gather information from the imported data. Variables of interest were imported from an external spreadsheet to be inserted in to statistical methods. These variables consisted of information like ‘Highest monthly maximum precipitation’, ‘Highest monthly maximum temperature’, ‘Lowest monthly minimum temperature’, ‘Average monthly maximum temperature’, ‘Average monthly maximum temperature’, and ‘Average monthly temperature’.

**2. Descriptive Statistics**

Below is a table (table 1) that contains the basic statistical information that will be use the analyze the weather data. The count describes the number of months that have been recorded. The mean describes the sum of all the information in the extreme maximums divided by the count. Standard deviation is the measure of the spread. Spread is the difference between values in a dataset and the center of the dataset. Center of the dataset is not purely considered as sets with different data can have the same mean and median. Variance is the standard deviation squared. Minimum is the smallest maximum value of precipitation or temperature. The 1st Quartile is all the data at or below the 25th percentile. The 2nd Quartile is all the data at or below the 50th percentile to the 25th percentile. The 3rd Quartile is all the data at or below the 75th percentile. Maximum is the largest maximum value of precipitation or temperature.

Table 1. Descriptive Statistics for Variable EMXT

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Statistic 1 *(e.g., Mean, Median, and  Standard Deviation)* | 9.9999  *\*Round off to 4 decimal places* |
| Count | 347 |
| Mean | 219.8559 (Celsius) |
| Standard Deviation | 91.5754 |
| Variance | 8386.0543 |
| Minimum | 39 (Celsius) |
| 1st Quartile | 133 (Celsius) |
| 2nd Quartile | 233 (Celsius) |
| 3rd Quartile | 306 (Celsius) |
| Maximum | 387 (Celsius) |

Table 2. Descriptive Statistics for Variable EMXP

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Statistic 1 *(e.g., Mean, Median, and  Standard Deviation)* | 9.9999  *\*Round off to 4 decimal places* |
| Count | 347 |
| Mean | 290.6311 (mm) |
| Standard Deviation | 180.1321 |
| Variance | 32447.5572 |
| Minimum | 3 (mm) |
| 1st Quartile | 179 (mm) |
| 2nd Quartile | 254 (mm) |
| 3rd Quartile | 363 (mm) |
| Maximum | 1760 (mm) |

**3. Confidence Intervals**

Confidence intervals are a combination of probability and interval estimate. An interval estimate is “the use of sample data to calculate an interval of plausible values of an unknown population parameter” (Wikipedia). The percentage in which the confidence interval contains the parameter is called the confidence level. Based on a 99% confidence interval for the proportion of months with the highest monthly maximum temperature above the tested 19 degrees based on a range of values, there is a 54.0565% to 67.55% to be above 19 degrees Celsius (See Table 3). Based on a 95% confidence interval for average highest monthly precipitation, there is a 95% chance to get an average precipitation between 271.6118 and 309.6505 (See Table 4).

Table 3. Confidence Interval for EMXT

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| (Lower Limit, Upper Limit) | (array([ 0.541]), array([ 0.6756])) |

Table 4. Confidence Interval for EMXP

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| (Lower Limit, Upper Limit) | (array([ 271.6118]), array([ 309.6505])) |

**4. Hypothesis Tests (1-sample and 2-sample)**

A hypothesis is a statement that makes a claim about parameters of one or more populations. The variable being tested is EMXT (Extreme Maximum Monthly Temperature). The condition being tested is, of the EMXT population, if the proportion of monthly maximum temperatures above 19 degrees Celsius is 67%. The alternative hypothesis is a statement the contradicts the hypothesis. So, if we are looking to test if the proportion above the 19 degrees is 67%, then the alternative is the proportion above the hypothesis is not 67%. Thus, not equal. Therefore, since the level of significance is 5%. Since the p-value for the hypothesis test for EMXT is less than the significance level then the null hypothesis, 67%, is rejected (See Table 5).

Table 5. Hypothesis Test for EMXT

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Test Statistics | 67%  19 (Celsius) |
| P-Value | 0.01812 |
| Null Value | 0.67 |
| Alternative | ‘not-equal’ |

The hypothesis is making a claim that the average highest monthly maximum precipitation is 28 millimeters. The alternative to this claim is the average highest monthly maximum is not actually 28 millimeters and is either above or below that test statistic. The level of significance established is a 1% level of significance. Since the p-value is much larger than the significance level, there is significance evidence that to reject the hypothesis that the monthly maximum it 28 millimeters.

Table 6. Hypothesis Test for EMXP

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Test Statistics | 28 mm |
| P-Value | 0.2724 |
| Null Value | 280 |
| Alternative | ‘not-equal’ |

**5. Results**

The objective was to update important weather pattern models based on the trends we found in our data. A critical component in delivering the information needed is having enough information to describe the dataset. This includes the count, mean, standard deviation, min, etc. The confidence interval of the proportion of months with the highest monthly maximum above 19 degrees Celsius has an lower limit and upper limit of 0.541 and 0.6756 respectively for the entire population of temperatures. This can be added compared year over year with other temperatures where the monthly maximum is above 19 degrees. Using this as an example, we can present this to the city of Manchester. The data has been analyzed by being imported from a csv file. Then, using the correct statistical functions and parameters accurate analysis can be executed free of errors.

**6. References**

Berrier, J., Nestler, S., Pardoe, I., Sturdivant, R. X., Watts, K., Chan, C., . . . Vahid, F. (2016). MAT 243: Applied Stats for STEM. Los Gatos, CA: Zyante.

Khan Academy. (n.d.). Retrieved from <https://www.khanacademy.org/>

Wikimedia Foundation, Inc. (n.d.). Retrieved from https://www.wikipedia.org/wiki/interval\_estimation

MAT 243 Final Project Part 2: Summary Report

Milestone II

Scenario (A): Title *National Weather Service*

Jason Lima

Jason.lima1@snhu.edu

Southern New Hampshire University

**1. Statement of the Problem**

The purpose of this analysis it to compare two hypotheses of data across the same data pool to determine if the observed values differ from each other and how. Weather can be different in different months and the within the same month, temperature can be different in different years. Questions analyst need to ask is, “Is there a trend that occurs in a single month? And does this trend carry over to the next?” This would lead to investigating the cause and effect of this relationship depending on the statistical significance of the correlation. To answer these types of questions we will implement the workflow for two sample hypothesis tests for proportion and mean hypothesized differences.

**2. Hypothesis Test for Population Proportions (2-sample)**

Table 1. Hypothesis Test for EMXT Proportion

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Count of July Months | 29 |
| Count of August Months | 29 |
| Sum count of July values greater than 32.5 Celsius | 15 |
| Sum count of August values greater than 32.5 Celsius | 11 |
| Significance Level | 1% (0.01) |
| z-score | 1.0561 |
| P-value | 0.2909 |

In the scenario for Table 1, we are looking to find if the proportion of Extreme Maximum Temperature over 32.5 degrees Celsius is the same for months July and August. Given the results we are using a 1% significance level. There is the count of data point in each given month. Then the count of the months that meet the hypothesis criteria which is months with an EMXT greater than 32.5 degrees Celsius. The z-score represent the number of standard deviations from the mean. The positive z-score indicates that the quantity is above the mean. About one standard deviation above the mean. The p-value is the probability of obtaining a result that is as extreme or more extreme than the data if the null hypothesis were true. Since the p-value is well above the 1%, there is reasonable evidence to reject the hypothesis.

Table 2. Hypothesis Test for EMXP Proportion

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Count of February Months | 29 |
| Count of August Months | 29 |
| Sum count of February values greater than 20.0mm | 13 |
| Sum count of August values greater than 20.0 mm | 21 |
| Significance Level | 5% (0.05) |
| z-score | -2.1328 |
| P-value | 0.0329 |

The counts for Table 2 are the same as Table 1. The difference between these tables is Table 2 is evaluating the difference of two population proportions for Extreme Maximum Precipitation over 20.0mm. Also, the population proportions will be evaluated using months February and August. The z-score represent a value a little over 2 standard deviations below the mean of the populations. The threshold value for the hypothesis is 5% or 0.05. Since the p-value of 0.0329 is less than 0.05, there is enough evidence to reject the null hypothesis. The hypothesis is to check if the proportion of Extreme Maximum Precipitation of over 20.00mm is the same for February and August. The null hypothesis is the alternative, the proportion of EMXP above 20.0mm is not the same.

**3. Hypothesis Tests for Population Means (2-sample)**

Table 1. Hypothesis Test for EMXT Mean

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| jul\_data | An array of EMXT values where the ‘Month’ equals 7 |
| aug\_data | An array of EMXT values where the ‘Month’ equals 8 |
| Statistic | 1.2931 |
| P-Value | 0.2013 |

Using the previously used significance level of 1% (0.01), if we look at the p-value, we can analyze that since the p-value is greater then the significance level, there is insufficient evidence to reject the null hypothesis of the average EMXT is the same for July and August.

Table 1. Hypothesis Test for EMXP Mean

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| feb\_data | An array of EMXP values where the ‘Month’ equals 2 |
| aug\_data | An array of EMXP values where the ‘Month’ equals 8 |
| Statistic | 2.2438 |
| P-Value | 0.0299 |

Using the previously used significance level of 5% (0.05), since the p-value is less than the significance level, we have enough evidence to reject the null hypothesis of the EMXP in February is more than August.

**4. Results**

After performing a hypothesis test for the difference of two population proportions of Extreme Maximum Temperature, the results determine the z-score was 1.0561 standard deviations above the random normalized mean. Around 34.1% above the mean. The p-value, which gives evidence to accept or reject the hypothesis was 0.2909. With a provided statistical significance of 1% (0.01), since the p-value was above what was considered a statistical significance then we have enough evidence to reject the hypothesis. This means that the proportion of temperature above 32.5 degrees Celsius is not the same for months July and August. Now we can investigate factors that could have cause this hypothesis to not be true.

After conducting the analysis of the hypothesis test for the difference of two population proportions of Extreme Maximum Precipitation. The hypothesis goal was to determine if the proportion of precipitation above 20.0mm is the same for months February and August in the data set. Based on a z-value of -2.1328, means the comparison of the proportion is slightly greater than 2 standard deviation below the mean of the normally distributed dataset. Further the p-value was below the 5% level of significance which determine the hypothesis had enough evidence to be accepted. We can look at factors that determine this hypothesis to be accepted such as the February and months and the August months experienced similar seasonal chances. One lead to cold weather and the other leads to warm weather.

**5. References**

Berrier, J., Nestler, S., Pardoe, I., Sturdivant, R. X., Watts, K., Chan, C., . . . Vahid, F. (2016). MAT 243: Applied Stats for STEM. Los Gatos, CA: Zyante.

Khan Academy. (n.d.). Retrieved from https://www.khanacademy.org/