**MST 690 – Data Science Mathematics** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Session 5 – Statistical Inference**

**In-Class Exercise**

Instructions: Solve each problem. Show all work. Attach additional sheets as necessary. Unless otherwise stated, you may use Python or other tools to assist.

1. You are a data scientist developing an algorithm for Russian event detection on Twitter, and are testing your algorithm using the 2018 Russian election as your test event. You are interested in discursive similarity over time (proceeding by argument or reasoning rather than by intuition)– the day-to-day similarity of discussion topics – within a target group of Twitter users. You assume that, given the amount of “noise” on Twitter (e.g., the high variability of topics discussed on any ordinary day), your similarity metric will be near zero. However, if a significant event occurs that becomes a hot topic of discussion, you hypothesize that your similarity metric will significantly increase for that given timeframe. You have collected discursive similarity data that bracket the test event, and want to determine if an anomaly has occurred on the date of the Russian election, 18 March 2018.

Consider the Excel data set on Blackboard. You will solve this by hand, as well as using Python. When solving by hand, you may use Excel to calculate any descriptive statistics. To infer any statistically significant difference from a specified value, the t-statistic can be calculated as follows:

Where s is the sample standard deviation, µ0 is the specified value, n is the sample size, and x-bar is the sample mean.

1. What is your null hypothesis? Explain what rejection of the null hypothesis means in terms of statistical inference.
2. Are you able to detect an anomalous event on 18 March 2018, given the above data set, and assuming a type I error rate of 5%? Explain why or why not.
3. Find the p-value using Python (see Python notebook for this exercise). Do you reject your hypothesis based on the calculated p-value?
4. Calculate by hand again assuming a type I error rate of 1%. What happens to the critical value of t? Intuitively, why is this the case? Does this change impact your conclusion?

Dataset:

|  |  |
| --- | --- |
| **Date** | **Similarity Metric (0.00-1.00)** |
| 18-Feb-18 | 0.0614 |
| 19-Feb-18 | 0.0913 |
| 20-Feb-18 | 0.0368 |
| 21-Feb-18 | 0.0213 |
| 22-Feb-18 | 0.0098 |
| 23-Feb-18 | 0.0161 |
| 24-Feb-18 | 0.0484 |
| 25-Feb-18 | 0.0394 |
| 26-Feb-18 | 0.0561 |
| 27-Feb-18 | 0.1 |
| 28-Feb-18 | 0.0683 |
| 1-Mar-18 | 0.0873 |
| 2-Mar-18 | 0.0627 |
| 3-Mar-18 | 0.0667 |
| 4-Mar-18 | 0.0572 |
| 5-Mar-18 | 0.0531 |
| 6-Mar-18 | 0.0149 |
| 7-Mar-18 | 0.0381 |
| 8-Mar-18 | 0.0622 |
| 9-Mar-18 | 0.0962 |
| 10-Mar-18 | 0.0774 |
| 11-Mar-18 | 0.076 |
| 12-Mar-18 | 0.0122 |
| 13-Mar-18 | 0.0695 |
| 14-Mar-18 | 0.0183 |
| 15-Mar-18 | 0.0427 |
| 16-Mar-18 | 0.0844 |
| 17-Mar-18 | 0.0037 |
| 18-Mar-18 | 0.8965 |
| 19-Mar-18 | 0.7814 |
| 20-Mar-18 | 0.0148 |
| 21-Mar-18 | 0.0743 |
| 22-Mar-18 | 0.0762 |
| 23-Mar-18 | 0.0851 |
| 24-Mar-18 | 0.0868 |
| 25-Mar-18 | 0.0859 |
| 26-Mar-18 | 0.0221 |
| 27-Mar-18 | 0.0412 |
| 28-Mar-18 | 0.0139 |
| 29-Mar-18 | 0.0176 |
| 30-Mar-18 | 0.0007 |
| 31-Mar-18 | 0.0223 |
| 1-Apr-18 | 0.0753 |
| 2-Apr-18 | 0.0525 |
| 3-Apr-18 | 0.0305 |
| 4-Apr-18 | 0.0156 |
| 5-Apr-18 | 0.0359 |
| 6-Apr-18 | 0.0142 |
| 7-Apr-18 | 0.0923 |
| 8-Apr-18 | 0.0916 |
| 9-Apr-18 | 0.0925 |
| 10-Apr-18 | 0.0295 |
| 11-Apr-18 | 0.0851 |
| 12-Apr-18 | 0.0905 |
| 13-Apr-18 | 0.0397 |
| 14-Apr-18 | 0.0647 |
| 15-Apr-18 | 0.0174 |
| 16-Apr-18 | 0.0043 |
| 17-Apr-18 | 0.0119 |
| 18-Apr-18 | 0.013 |