NUMB3RS Activity: Stylometry
Episode: "Killer Chat"

Topic: Box-and-whisker plots **Grade Level:** 9 - 12
Objective: Compare writing samples to identify an author
Time: 30 minutes
Materials: TI-83 Plus/TI-84 Plus graphing calculator and the following files:
Sample A - Julia.pdf, Sample B - Aaron.pdf, Sample C.pdf. To download these files, go to **http://education.ti.com/exchange** and search for "7772."

Introduction

In "Killer Chat," the FBI has obtained a collection of online chat dialogues from a suspected killer. Charlie proposes he use Statistical Linguistic Analysis to identify the author. The style and language can essentially fingerprint the writing. Charlie says to "think of it like a jeweler beading a necklace. The jeweler chooses certain beads, decides what pattern to string them in, depending on his personal style. Other jewelers will have different styles, exhibiting different patterns. Just like different IM senders will exhibit their own styles – patterns of speech unique to them."

Stylometry, the practice of applying statistical analysis to linguistic style, has been used to answers questions of authorship, notably for the *Federalist Papers*, Civil War letters, and Shakespeare's plays. Works can be analyzed with many different statistics. This activity will focus on two statistics: sentence length and paragraph length.

Three writing samples are provided for students to analyze. Students will count the number of words per sentence and the number of words per paragraph for each sample. The amount of time needed can be greatly reduced if the work is shared among classmates. Six pairs of students can compile the data in less than 10 minutes. With more students, the time can be reduced, or numbers can be verified.

Each student can create the box-and-whisker plots if the data are displayed in the classroom. This approach will require each student to enter data into six lists in their calculator. Students can save time by entering the data into one calculator with a display screen. Alternatively, the TI-Navigator™ system can be used to compile data from the entire class.

As students enter data into lists, they may need a reminder that lists are cleared by placing the cursor on the list name and pressing [CLEAR] and then [ENTER], not the [DEL] key. Lists that are accidentally deleted can be replaced by placing the cursor on another list name and pressing [2nd] [INS].

The statistics calculated in this activity were purposely chosen to show a difference between the writing samples. When stylometry is used to analyze writing, many different statistics may be computed, and the size of the sample would be much larger. This activity is intended to model the process.

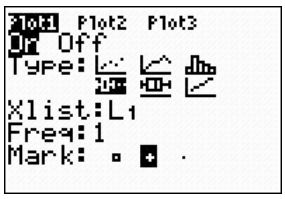
The writing samples used for this activity, originally published in *Maroon Reflections*, the Monument Mountain Regional High School newspaper, are reproduced with the permission of the authors, Julia Latino and Aaron Moser. Because the purpose of this activity is to analyze writers' styles, these writing samples have been left unedited and are reproduced in their original forms.

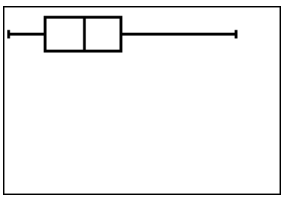
Kathy Erickson, Monument Mountain Regional High School, Great Barrington, MA

Before the Lesson

This activity assumes a familiarity with some statistics. However, a review of box-and-whisker plots will be useful. Consider this small data set of word lengths: 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 6, 6, 8.

To enter the data into a TI-83 Plus/TI-84 Plus calculator, press [STAT], select **1:Edit…**, and enter the data into L_1. Press [2nd] [STAT PLOT] and select **1:Plot1**. Use the settings shown below. To view the plot in an appropriate window, press [ZOOM], then select **9:ZoomStat**. The result is shown below, with the axes turned off.





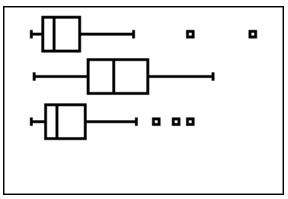
Student Page Answers:

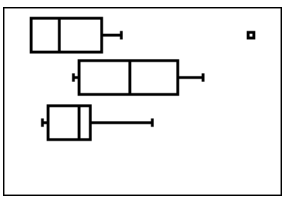
Note: For answers to questions 1, 2, and 3, the slash (/) is used to separate paragraphs in the words-per-sentence set.

1a. *WPS = (13, 4, 8, 3, 6, 4, 7, 2, 3, 5 / 42, 22, 4 / 16, 27, 18, 16, 22, 5, 4, 25, 21, 10, 13, 13, 6, 5, 21, 4, 5, 10, 13, 5, 10, 7 / 11, 8, 4, 11, 58, 14, 7 / 6, 3, 3, 2, 2, 2, 2 / 2 / 2 / 5)*
1b. *WPP = (55, 28, 276, 113, 20, 2, 2, 5)*

2a. *WPS = (14, 33, 15 / 14, 13, 18, 22, 18, 28 / 29, 29, 36, 40 / 19, 17, 25, 32, 25, 25, 16, 14, 43 / 19, 36, 17, 22, 16, 31, 30, 14 / 19, 36)*
2b. *WPP = (62, 113, 134, 216, 185, 55)*

3a. *WPS = (19, 20, 11, 12 / 9, 14 / 29, 7, 42, 6, 1, 10, 9, 16, 3, 25, 4 / 3, 12, 20, 3, 14, 6, 4 / 6, 5, 7, 9, 7, 32, 6 / 6, 3, 9, 5 / 27 / 4, 10, 38, 22, 6, 16 / 15, 33, 8, 4, 9, 6 / 18)*
3b. *WPP = (62, 23, 152, 62, 72, 23, 27, 96, 75, 18)*

4a.  **4b.** *Sample C seems to match more closely with Sample A, suggesting Julia is the author.*

5a.  **5b.** *Again, Sample C seems to match more closely with Sample A, suggesting Julia is the author (in fact, she is).* **6.** *Possible statistics are length of reply, use of abbreviations and time between replies. There are many others.*

Kathy Erickson, Monument Mountain Regional High School, Great Barrington, MA

Name _____ Date _____

NUMB3RS Activity: Stylometry

In "Killer Chat," the FBI has obtained a collection of online chat dialogues from a suspected killer. Charlie proposes he use Statistical Linguistic Analysis to identify the author. The style and language can essentially fingerprint the writing. Charlie says to "think of it like a jeweler beading a necklace. The jeweler chooses certain beads, decides what pattern to string them in, depending on his personal style. Other jewelers will have different styles, exhibiting different patterns. Just like different IM senders will exhibit their own styles – patterns of speech unique to them."

Stylometry, the practice of applying statistical analysis to linguistic style, has been used to answers questions of authorship, notable for the *Federalist Papers*, Civil War letters, and Shakespeare's plays. Works can be analyzed with many different statistics. This activity will focus on two statistics: sentence length and paragraph length. Because the analysis of chat dialogues present more challenges, the samples used in the activity are short articles written in prose.

The Situation

An anonymous piece of writing is discovered. Your goal is to use a statistical analysis to determine which of two known samples the anonymous writing most resembles. To save time and tedious work, your teacher may ask you to split up questions 1a through 3c. Each part asks you to count the number of words in a document. A few guidelines follow.

- If you can open the file on a word processor, the computer may evaluate some of these values during or after a grammar check. You may need to turn this feature on. The rules that follow agree with one such program and were used to generate the results given in this activity. Your class can agree on your own rules to apply consistently.
- A hyphenated word is one word.
- An abbreviation is one word.
- A URL (Web site address) is one word.
- A number is one word, even at the beginning of a list (in sample C: "2 Everyone procrastinates." is one sentence with 3 words).
- In Sample A, "Q: So…" is one sentence with two words.

1. Compile statistics from Sample A, written by Julia Latino.

a. Record the number of words per sentence in Sample A. Enter the data into list L_1 on the calculator.

b. Record the number of words per paragraph in Sample A. Enter the data into list L_4.

Kathy Erickson, Monument Mountain Regional High School, Great Barrington, MA

2. Compile statistics from Sample B, written by Aaron Moser.

a. Record the number of words per sentence in Sample B. Enter the data into list L_2.

b. Record the number of words per paragraph in Sample B. Enter the data into list L_5.

3. Sample C was written by either Julia or Aaron.

a. Record the number of words per sentence in Sample C. Enter the data into list L_3.

b. Record the number of words per paragraph in Sample C. Enter the data into list L_6.

4. a. Set up a box-and-whisker plot for the numbers of words per sentence. The data are in L_1, L_2, and L_3. Use three separate plots so the calculator will display the three diagrams on one screen, with Sample A on the top, followed by Sample B and then Sample C.

b. Do the numbers of words per sentence suggest who might be the author of Sample C? If so, who is the author? Explain.

5. a. Set up a box-and-whisker plot for the numbers of words per paragraph. The data are stored in L_4, L_5, and L_6. The calculator will display the three plots on one screen, with Sample A on the top, followed by Sample B and then Sample C.

b. Do the numbers of words per paragraph suggest who might be the author of Sample C? If so, who is the author? Explain.

6. The writing sample that Charlie analyzes in the episode is online chat dialogue. What statistics might you compile when attempting to "fingerprint" the author of such a sample?

Kathy Erickson, Monument Mountain Regional High School, Great Barrington, MA

The goal of this activity is to give your students a short and simple snapshot into a very extensive math topic. TI and NCTM encourage you and your students to learn more about this topic using the extensions provided below and through your own independent research.

Extensions

This activity can be applied to many different situations. Explore the topic further with other writing samples. Keep in mind that true stylometric analysis involves very large samples, so results may not always be clear. Suggested areas to explore are listed below.

- Try this activity with writing samples from your own class or school newspaper. Experiment with other statistics. Are some statistics better predictors than others? You can save time by using a word processor on your computer to evaluate some of these values during or after a grammar check.

- Compare works by different authors or in different genres to determine if some statistics can "fingerprint" work. Statistics to experiment with may be parts of speech, punctuation and length of chapters.

- Can you distinguish between authors of online chats? Statistics may include computer abbreviations such as LOL or BRB, length of responses, and time between responses.

- Research more about the use of stylometry to analyze the *Federalist Papers*, Civil War letters, and Shakespeare's plays.

Additional Resources

An instructive article detailing the question of authorship in a book from the Oz series can be found at the Science News Web site:
http://www.sciencenews.org/articles/20031220/bob8.asp

An article from the BBC on stylometry applied to music can be found at the following Web site:
http://news.bbc.co.uk/1/hi/5083986.stm?ls

Kathy Erickson, Monument Mountain Regional High School, Great Barrington, MA