Entropy in Biosurveillance

Joshua Smith
Andrew McConvey
July 17, 2009

Mentor: Nina Fefferman

The Problem

- Create an algorithm that will determine parameters to best detect when disease outbreaks occur

- Determine threshold for when disease patterns should be considered an outbreak

WHY IS IT IMPORTANT?!?!?!?!?!?

- Failure to detect can lead to an otherwise avoidable increase in disease spread
- Such spread can lead to increase in the mortality rate of disease
- Since our method only depends on amount of cases, it can be applied to natural epidemics as well as acts of bioterrorism.

Entropy

- Quantifies the average amount of information received, when a given event is observed

- We will calculate entropy based on disease incidence data

- Previous groups were able to detect significant changes in Entropy when disease outbreaks occurred.

Entropy (continued)

$$E = \sum_{x_i \in X} -P(x_i)\log_2 P(x_i)$$

- Able to calculate for any string of numbers
- Independent of the relationship between numbers, depends only on the probability of a symbol occurring
- Can use this to detect abnormal shifts in incidence data

Entropy (Sample Calculation)

- Suppose we are given the following string of letters on the alphabet {1,2,3}:

  1 3 1 2 2 1 2 3 2 1

- Then we have the following probabilities

$$P(1) = \frac{2}{5} \qquad P(2) = \frac{2}{5} \qquad P(3) = \frac{1}{5}$$

- Which gives us the following equation for Entropy

$$E = \sum_{x_i \in \{1,2,3\}} -P(x_i)\log P(x_i) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{1}{5}\log\left(\frac{1}{5}\right) = 1.52$$

Parameters

- Window Size
- Step Size
- Binning

Window Size

- The size of the string of data which we will compute

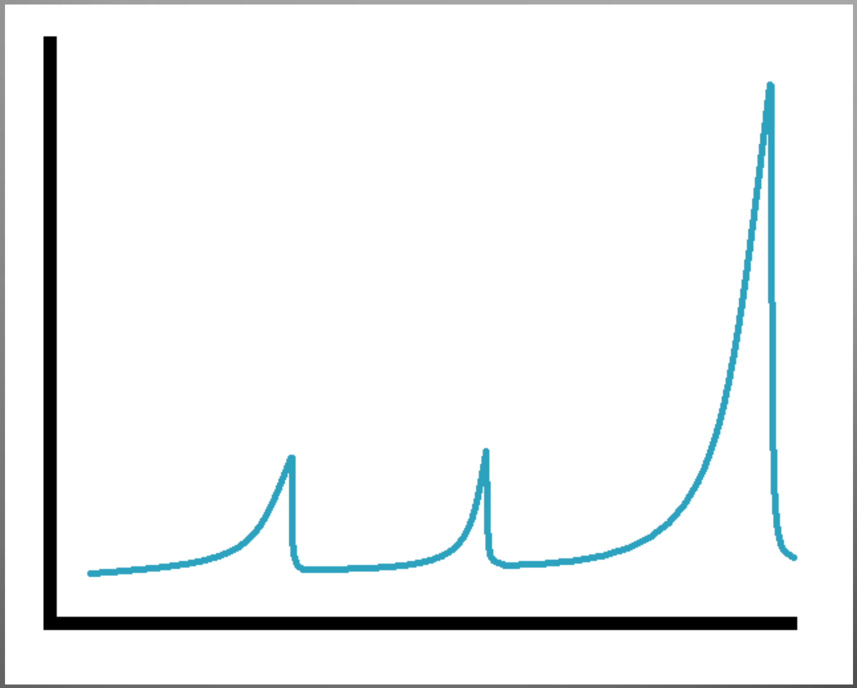
- Will ensure the window size is large enough to detect changes in patterns, but sensitive enough to not overlook meaningful data

Step Size

- How fast the process moves through the data
- How much overlap between successive windows

Step Size (Example)

- Often times data from weekends is skewed, perhaps resulting in greater incidence every Monday
- By using a step size of seven, we can account for this variation, and detect when a true outbreak occurs



Window Size and Step Size

- Suppose we are given the following string:
  1 2 4 2 1 2 1 2 1 1 3 2 2 1 1
- With a window size of 5 and a step size of 3, we would break the string into the following four segments
  1 2 4 2 1 2 1 2 1 1 3 2 2 1 1
- Alternatively, a window size of 4 and a step size of two would create six completely different bins
  1 2 4 2 1 2 1 2 1 1 3 2 2 1 1

Binning

- Allows us to account for differences between large and small variations in entropy readings
- Group together results with similar (although not necessarily the same) entropy to detect meaningful changes in pattern

Binning (Continued)

- Suppose we have the following string:

  101 120 110 113 108 100 115 106 425 723

  E = 2.00 E = 2.00 E = 2.00 E = 2.00

- Without binning, entropy would be constant since each number is distinct from the others, for example consider a step size of 2 and a window size of 4
- However, suppose we bin (into groups of 100) to accentuate meaningful changes:

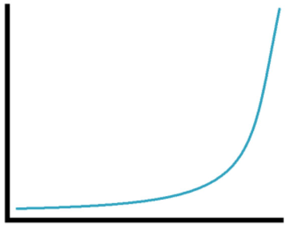
  1 1 1 1 1 1 1 1 4 7

  E = 0 E = 0 E = 0 E = 1.5

Our Progress

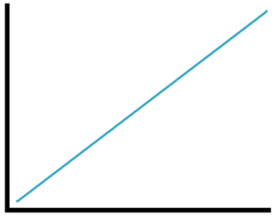
- We considered three different types of outbreaks
  - typical exponential outbreak
  - steady, linearly increasing outbreak
  - one day outbreak
- We wanted to define methods of detecting these outbreaks as early as possible

Exponential Outbreak



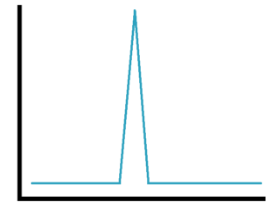
- For this outbreak, static bins should not overlook relative changes in incidence
- Increasing window size makes entropy calculations less sensitive, but do not want so small to detect normal variations in the data

Linear Outbreak



- Outbreak looks the same, regardless of where we are in the outbreak (since entropy does not take values into account)
- One way around this is to vary the size of bins
- As incidence grows, bins get smaller, leading to increase in entropy which we should then be able to detect

One Day Outbreak



- Since this outbreak lasts only one day, it must be detected as soon as possible
- Must bin the data such that when the outbreak occurs, it is put in a separate bin
- This new bin will lead to an increase in entropy as soon as our window reaches the break

Current Progress

- We are trying to define the explicit relationship between window size and step size
- Looking to calculate the probability of a certain symbol if we know
  - Step Size
  - Window Size
  - Length of the String
  - Frequency of Symbol in Each Window

Further Goals

- Still looking for the ideal combination of parameters that will lead to the automatic early detection of diseases
    - Must improve our binning methods
    - Gain a better understanding of how the binning method is affected by changes to step size and window size
- Define a threshold that will determine when an outbreak has actually occurred

Thank You!