Assessing the Effectiveness of Cumulative Sum Poisson- and Normal-based Tests for Detecting Rare Diseases

16 November 2010

LCDR Manuel Ganuza
Thesis Research

Prof. Ronald Fricker
Advisor

Graduate School Operations & Information Sciences
Department of Operations Research

Monterey, California
• Motivating Problem: Detecting Tularemia

• Research Question

• Background
  – Normal and Poisson Distributions
  – CUSUM

• Applying the CUSUM
  – Measures of Performance

• Results & Conclusions
Motivating Problem: Detecting Tularemia

- Highly virulent, highly infectious disease[^1]
  - Caused by bacterium *Francisella tularensis*

- Also known as rabbit fever and deer fly fever
  - Primary vectors are ticks and deer flies
  - Primary reservoirs are small to medium-sized mammals

- Extremely rare in the United States[^1]
  - From 1990 to 2000, rate less than 1 per 1,000,000

- Can be weaponized for aerosol release[^2]

- Category A agent by CDC
  - Category A agents have the "potential to pose a severe threat to public health and safety"

Sources:
Motivating Problem: Detecting Tularemia

• **Modes of Transmission**\[^3\]**
  - Bites by infected arthropods
  - Direct contact with infected animals
  - Handling of infectious animal tissues or fluids
  - Ingestion of contaminated food, water, or soil
  - Possibly direct contact with contaminated soil or water
  - Inhalation of infectious aerosols
  - Exposure in laboratory setting

• **Clinical Syndromes**\[^3\]**
  - Glandular and Ulceroglandular tularemia
  - Pneumonic tularemia
  - Oculoglandular tularemia
  - Oropharyngeal tularemia
  - Typhiodial tularemia

Source:
Some Tularemia Outbreaks in US

- **Martha’s Vineyard 2000**\(^4\): One fatality
  - CDC investigated for possibility of aerosolized *F. tularensis*
  - Subsequently documented cases of tularemia resulted from lawn mowing

- **Washington, D.C. 2005**\(^5\): Small amounts of *Francisella tularensis* were detected
  - Morning after an anti-war demonstration, biohazard sensors were triggered at six locations surrounding the Capitol Mall
  - While thousands of people were potentially exposed, no infections were reported

Sources:
\(^5\) Center for Infectious Diseases Research & Policy News. Tularemia agent found in DC air, but no cases seen. University of Minnesota October 2005.
• **Background:**
  – Biosurveillance systems often use Cumulative Sum (CUSUM) detection algorithm
  – CUSUM derived from normal distribution is most commonly used (“normal-based CUSUM”)
    • Rare diseases do not follow normal distribution
    • Poisson distribution more likely and appropriate

• **Research question:**

  *How does normal-based CUSUM perform compared to Poisson-based CUSUM for detecting rare diseases?*
• **Univariate distribution for **continuous data
  – Applies to many phenomena … but not all!

• **Two parameters:** mean $\mu$ and variance $\sigma^2$
Background: Poisson Distribution

- Univariate distribution for discrete data
  - Often a good model for rare events

- One parameter: $\lambda$
  - Distribution has mean $\lambda$ and variance $\lambda$
• Sequential test for change (increase) in distribution mean (Page, 1954)

• Basic form: \( C_t = \max \left( 0, C_{t-1} + \ln \frac{f_1(X)}{f_0(X)} \right) \)

where

– \( C_t \) is the current value of the CUSUM on day \( t \)
– \( X_t \) is the observed number of cases on day \( t \)
– \( C_{t-1} \) is the value of the CUSUM on day \( t-1 \)
– \( \ln[f_1(X)/f_0(X)] \) is the log-likelihood ratio for \( X \) for the outbreak distribution \( f_1 \) and the non-outbreak distribution \( f_0 \)

• Threshold \( h \): CUSUM signals at time \( t \) when \( C_t \geq h \)
Forms of the CUSUM

- The basic form:

\[ C_t = \max \left( 0, C_{t-1} + \ln \frac{f_1(X)}{f_0(X)} \right) \]

- For \( f_0 = N(\mu_0, \sigma^2) \) and \( f_1 = N(\mu_1, \sigma^2) \):

\[ C_t = \max \left( 0, C_{t-1} + X_t - \frac{\mu_1 - \mu_0}{2} \right) \]

- For \( f_0 = \text{Poisson}(\lambda_0) \) and \( f_1 = \text{Poisson}(\lambda_1) \):

\[ C_t = \max \left( 0, C_{t-1} + X_t - \frac{\lambda_1 - \lambda_0}{\ln(\lambda_1) - \ln(\lambda_0)} \right) \]
Comparing the Forms

• Normal-based and Poisson-based CUSUMs are exactly the same when the reference values match:

\[ \frac{\mu_1 - \mu_0}{2} \equiv \frac{\lambda_1 - \lambda_0}{2} = \frac{\lambda_1 - \lambda_0}{\ln(\lambda_1) - \ln(\lambda_0)} \]

• So, the CUSUMs are the same when

\[ \ln(\lambda_1) - \ln(\lambda_0) = 2 \]

which is the same as when

\[ \lambda_1 / \lambda_0 = e^2 \approx 7.3891 \]

and which is also the same as when

\[ \lambda_1 \approx 7.40 \lambda_0 \]
Three cases of interest

- Case #1: $\lambda_1 = 7.40\lambda_0$
- Case #2: $\lambda_1 \ll 7.40\lambda_0$
- Case #3: $\lambda_1 \gg 7.40\lambda_0$
• **Average Time to False/First Signal (ATFS)**
  – When there is no outbreak, it is the average time between (false) signals
  – When there is an outbreak, it is the average delay from start of outbreak to first (real) signal

• **The ideal algorithm:**
  – Has large time between false signals when there is no outbreak
  – Has a small delay to first real signal when there is an outbreak

• **Comparison methodology**
  – Match time between false signals for a given $\lambda_0$
  – Assess ATFS across variety of outbreaks ($\lambda^* > \lambda_0$)
    • Method with smaller ATFS values is preferred
Scenario

- **Disease: Tularemia**
  - Incubation period from 3 to 5 days (range 1 to 14 days)
  - Cases occur according to Poisson(\(\lambda^*\)) distribution

- **Biosurveillance Systems**
  - **System A**
    - Uses normal-based CUSUM
    - Assume disease distribution is approximate normal
      *exempli gratia*, for \(X \sim \text{Poisson}(\lambda)\) it assumes \(X \sim \text{N}(\lambda, \lambda)\)
    - Set threshold \(h_A\), so ATFS is large when no outbreak
  - **System B**
    - Uses Poisson-based CUSUM
    - Set threshold \(h_B\), so ATFS is large when no outbreak
Results: Case #1 ($\lambda_1 \approx 7.40\lambda_0$)

Biosurveillance Systems
- System A
- System B

User set parameters:
- No outbreak, mean $\lambda_0 = 0.100$
- Outbreak, mean $\lambda_1 = 0.740$
- Threshold, $h_A = 0.64$
- Threshold, $h_B = 1.36$

Actual occurrence:
- Data Poisson
- Outbreak, mean $\lambda^*$
Results: Case #1 ($\lambda_1 \approx 7.40\lambda_0$)

User set parameters:
- No outbreak, mean $\lambda_0 = 0.100$
- Outbreak, mean $\lambda_1 = 0.740$
- Threshold, $h_A = 0.64$
  $h_B = 1.36$
  $h_A' = 1.08$

Actual occurrence:
- Data Poisson
- Outbreak, mean $\lambda^*$

Biosurveillance Systems
- System A
- System B
- System A'
Results: Case #2 \( (\lambda_1 \ll 7.40\lambda_0) \)

Biosurveillance Systems

- System A
- System B

User set parameters:
- No outbreak, mean \( \lambda_0 = 0.100 \)
- Outbreak, mean \( \lambda_1 = 0.105 \)
- Threshold, \( h_A = 8.11 \) \( h_B = 2.39 \)

Actual occurrence:
- Data Poisson
- Outbreak, mean \( \lambda^* \)

\[ \text{for } \lambda^* = \lambda_0, \text{ ATFS is too high} \]
Results: Case #2 \((\lambda_1 \ll 7.40\lambda_0)\)

- User set parameters:
  - No outbreak, mean \(\lambda_0 = 0.100\)
  - Outbreak, mean \(\lambda_1 = 0.105\)
  - Threshold, \(h_A = 8.11\), \(h_B = 2.39\), \(h_A' = 7.73\)

- Actual occurrence:
  - Data Poisson
  - Outbreak, mean \(\lambda^*\)

For \(\lambda^* = \lambda_0\), ATFS match

Biosurveillance Systems

- System A
- System B
- System A'

Graph showing the relationship between average time to signal and \(\text{lambda}^{\text{star}}\).
Results: Case #3 (\( \lambda_1 \gg 7.40 \lambda_0 \))

Biosurveillance Systems
- System A
- System B

User set parameters:
- No outbreak, mean \( \lambda_0 = 0.100 \)
- Outbreak, mean \( \lambda_1 = 1.480 \)
- Threshold, \( h_A = 0.12 \) \( h_B = 0.49 \)

Actual occurrence:
- Data Poisson
- Outbreak, mean \( \lambda^* \)
Results: Case #3 ( $\lambda_1 \gg 7.40\lambda_0$ )

Biosurveillance Systems
- System A
- System B
- System A'

User set parameters:
- No outbreak, mean $\lambda_0 = 0.100$
- Outbreak, mean $\lambda_1 = 1.480$
- Threshold, $h_A = 0.12$, $h_B = 0.49$, $h_A' = 0.31$

Actual occurrence:
- Data Poisson
- Outbreak, mean $\lambda^*$
• When
  – Rate of disease, \( \lambda_0 \), is very low,
  – Occurrence counts are Poisson distributed, \( X \sim \text{Pois}(\lambda_0) \),
  – Outbreak, \( \lambda_1 \), manifests as only a small increase in the rate of disease, \( \lambda_1 \ll 7.40\lambda_0 \)

• Then *incorrect use of normal-based CUSUM can result in unacceptable delay in detection*

• To monitor rare diseases, such as Tularemia, include Poisson-based CUSUM in biosurveillance systems
  – Example here shows potential delays in signaling on the order of weeks
• **When**
  – Rate of disease, $\lambda_0$, is very low,
  – Occurrence counts are Poisson distributed, $X \sim \text{Pois}(\lambda_0)$,
  – Outbreak rate, $\lambda_1$, is significantly larger than rate of disease, $\lambda_1 \geq 7.40\lambda_0$

• **Then** *the normal-based CUSUM performs as well as Poisson-based CUSUM*
  – **If** threshold $h$, is set appropriately to achieve equivalent ATFS performance at $\lambda^* = \lambda_0$
  – With threshold $h$, set incorrectly, Normal-based CUSUM has excessively high false alarm rate
    • A real problem with existing biosurveillance systems
Assessing the Effectiveness of Cumulative Sum Poisson- and Normal-based Tests for Detecting Rare Diseases

Photo Source:
Francisella tularensis

- **Subspecies**[3]
  - *F. tularensis subsp. tularensis* (type A)
    - Highly infectious, more virulent, more genetically diverse
    - Found in North America
  - *F. tularensis subsp. holarctica* (type B)
    - Found in North America, Europe, Siberia, Far East, and Kazakhstan
  - *F. tularensis subsp. mediaasiatica*
    - Found in Central Asia

Source:
Background: Normal Distribution

- Probability density function (pdf)

\[ f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right) \]
Background: Poisson Distribution

- Probability mass function (pmf)

\[
\Pr(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}
\]

\(\lambda = 5\)
Applying the CUSUM: Normal Data

• For $f_0 = N(\mu_0, \sigma^2)$ and $f_1 = N(\mu_1, \sigma^2)$

$$C_t = \max \left( 0, C_{t-1} + X_t - \frac{\mu_1 - \mu_0}{2} \right)$$

• In words, at each time period add the observed data minus one-half the difference between $\mu_1$ and $\mu_0$ to the cumulative total
  – If the cumulative total is negative, set to zero

• This is the most common form of the CUSUM
  – Often applied to data without realizing that the underlying assumption is normality
Applying the CUSUM: Poisson Data

• For $f_0 = \text{Poisson}(\lambda_0)$ and $f_1 = \text{Poisson}(\lambda_1)$

$$C_t = \max \left( 0, C_{t-1} + X_t - \frac{\lambda_1 - \lambda_0}{\ln(\lambda_1) - \ln(\lambda_0)} \right)$$

• Similar idea, at each time period add the observed data minus a quantity based on a function of $\lambda_1$ and $\lambda_0$ to the cumulative total
  – If the cumulative total is negative, set to zero

• Much less commonly known form
  – But it is the correct form for Poisson data