SPATIO-TEMPORAL STATISTICAL METHODS FOR QUICKLY DETECTING DISEASE OUTBREAKS

Sesha Dassanayaka and Joshua French
Truman State University
and
University of Colorado Denver
September 14, 2016

Overview

Section I: Background information on disease surveillance

Section II: The CUSUM method and an improved testing procedure

Section III: Case study

Section IV: Future research directions/Take Home points

SECTION I: Background information

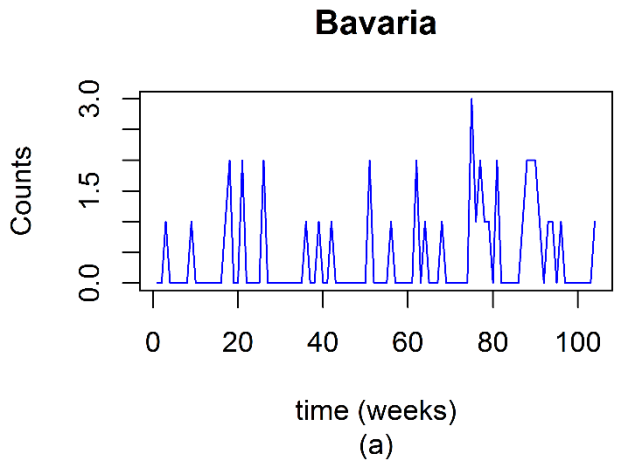
Infectious disease outbreaks currently being reported on by CDC

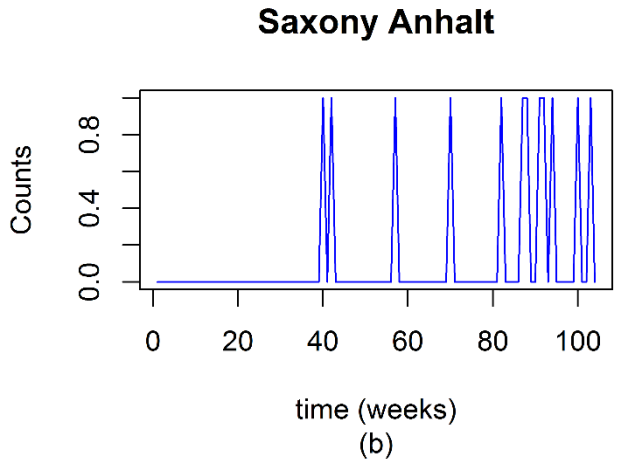
- Alfalfa Sprouts – Salmonella Reading and Salmonella Abony - Announced August 2016
- Live Poultry – Salmonella Enteritidis and 7 more - Announced June 2016
- Flour – E. coli O121 & O26 - Announced June 2016
- Raw Milk – Listeria monocytogenes - Announced March 2016
- Elizabethkingia anophelis in the Midwest - Announced January 2016
- Small turtles – Salmonella Sandiego and Salmonella Poona - Announced October 2015

Data preview: Salmonella Newport outbreak in Germany

- Weekly disease counts available for 16 German states between 2004-2014.

Salmonella case counts (2004 - 2005): Bavaria and Saxony Anhalt





What is a disease outbreak?

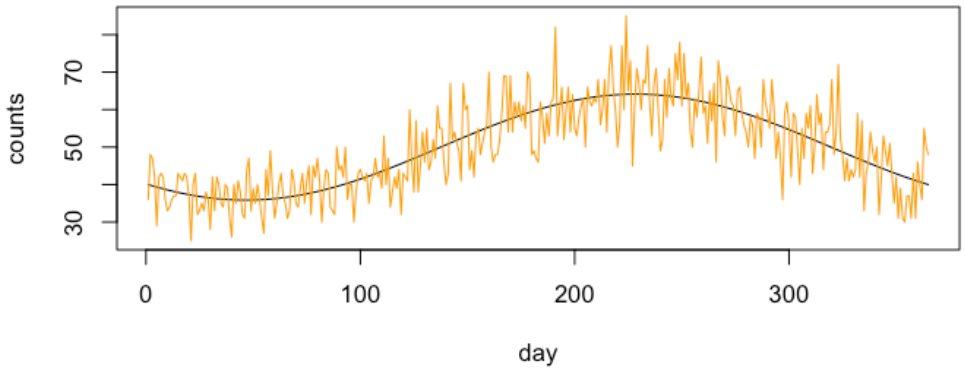
From the *World Health Organization:*

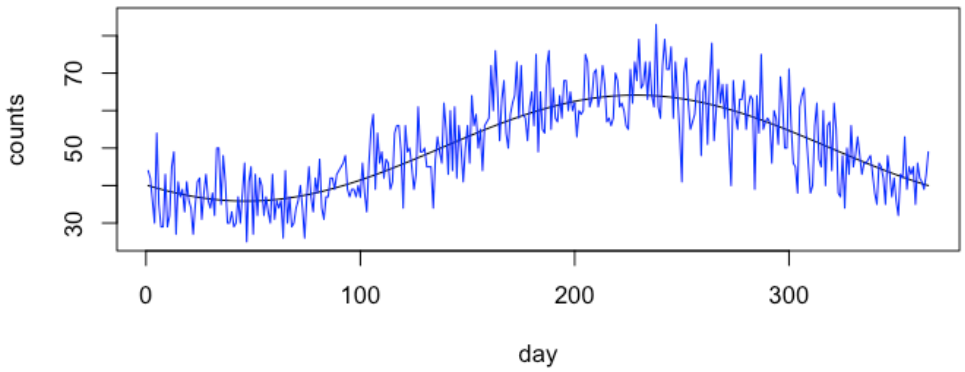
- A **disease outbreak** is the occurrence of cases of **disease** in excess of what would normally be expected in a defined community, geographical area or season.
- An **outbreak** may occur in a restricted geographical area, or may extend over several countries.
  - It may last for a few days or weeks, or for several years.

What is disease surveillance?

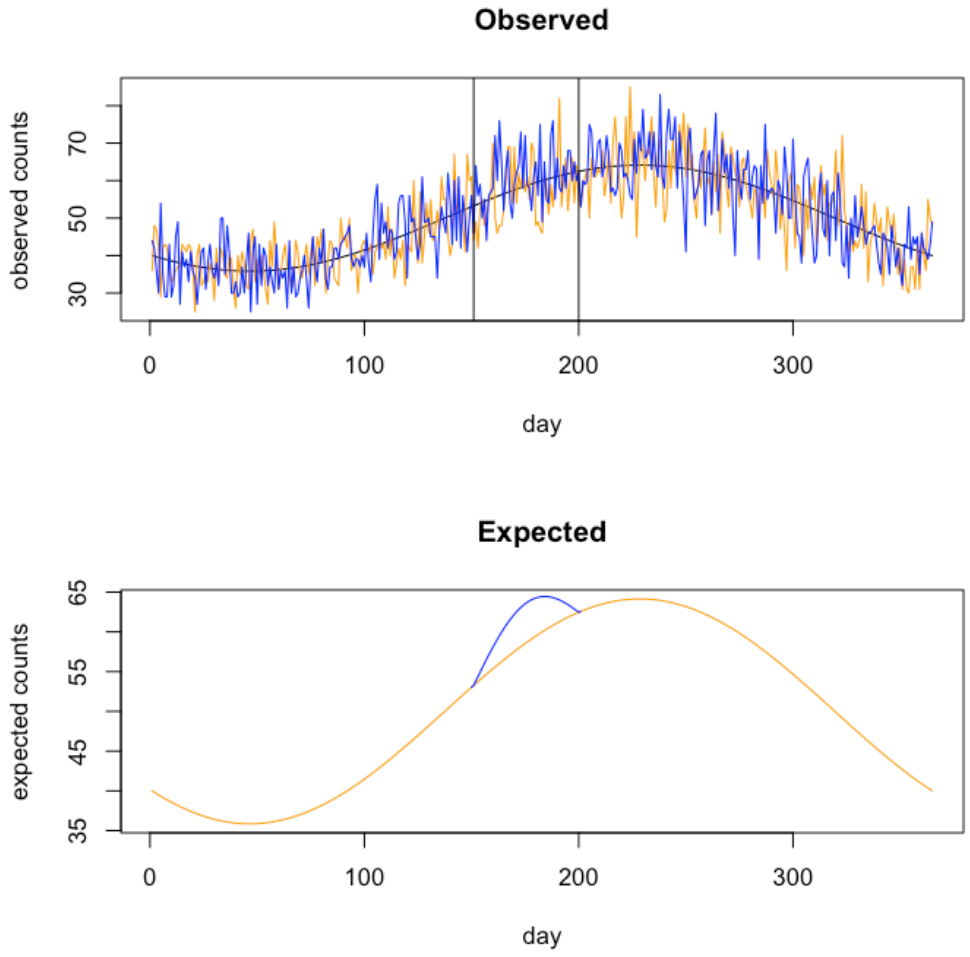
- Disease surveillance is an epidemiological practice by which the spread of disease is monitored in order to **establish patterns of progression**.
- The main role of disease surveillance is to **predict**, **observe**, and **minimize the harm** caused by outbreak, epidemic, and pandemic situations.

Which time series has an outbreak?





Solution


Observed
observed counts
30
50
70
0
100
200
300
day
Expected
expected counts
35
45
55
65
0
100
200
300
day

Recent events encouraging rapid development in disease surveillance

1. Threat of bioterrorism

- Anthrax attacks in USA (October, 2001)

2. Heighted public awareness of emerging diseases

- Avian Influenza - H5N1 Influenza (June 2008)
  China, Indonesia, Vietnam (June 2008)
- Swine influenza – H1N1 (June 2009)
- Ebola outbreak
  West Africa
  (December 2013)

Types of outbreaks (categorized by WHO)

- Communicable disease outbreaks
  - Illnesses that result from the infection, presence and growth of pathogenic (capable of causing disease) biologic agents in an individual human or other animal hosts.
- Disease outbreaks caused by chemicals or toxins.
- Disease outbreaks of unknown etiology

SECTION II: The CUSUM method and an improved testing procedure

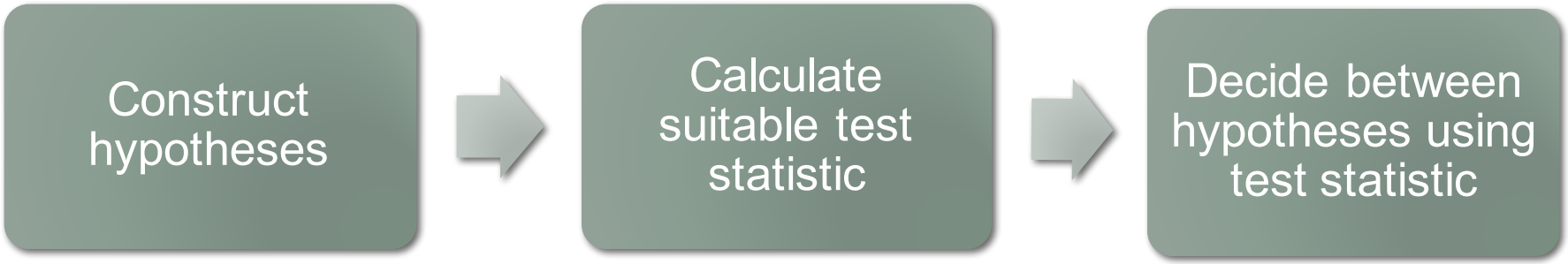
Statistical tests

- **Decide between two claims**: a null hypothesis and a research hypothesis.
  - The null hypothesis is what you assume is true if nothing interesting is going on.
    - There is no relationship between dehydration and kidney disease.
    - There is no relationship between heavy metals and kidney disease.
  - The research hypothesis is typically what we believe is true and would like to conclude.
- The observed data are used to compute a **test statistic**, which is a number used to assess the compatibility of the data with the null hypothesis.

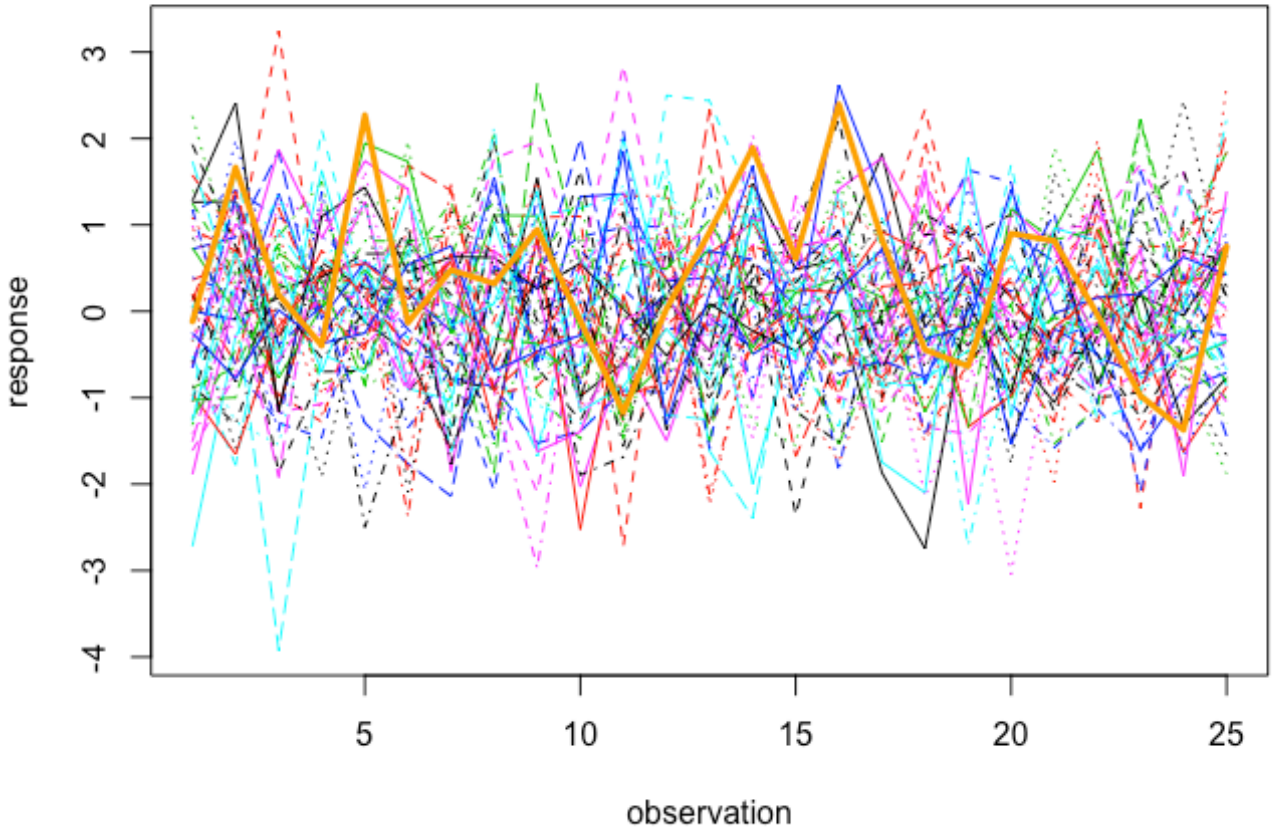
Statistical tests (cont.)

- Assuming the null hypothesis is true, **construct a decision rule** to choose between the two hypotheses.
  - The rule is chosen to ensure that the probability of an error is not too large.
  - Usually, you conclude the research hypothesis when the test statistic is “large”.

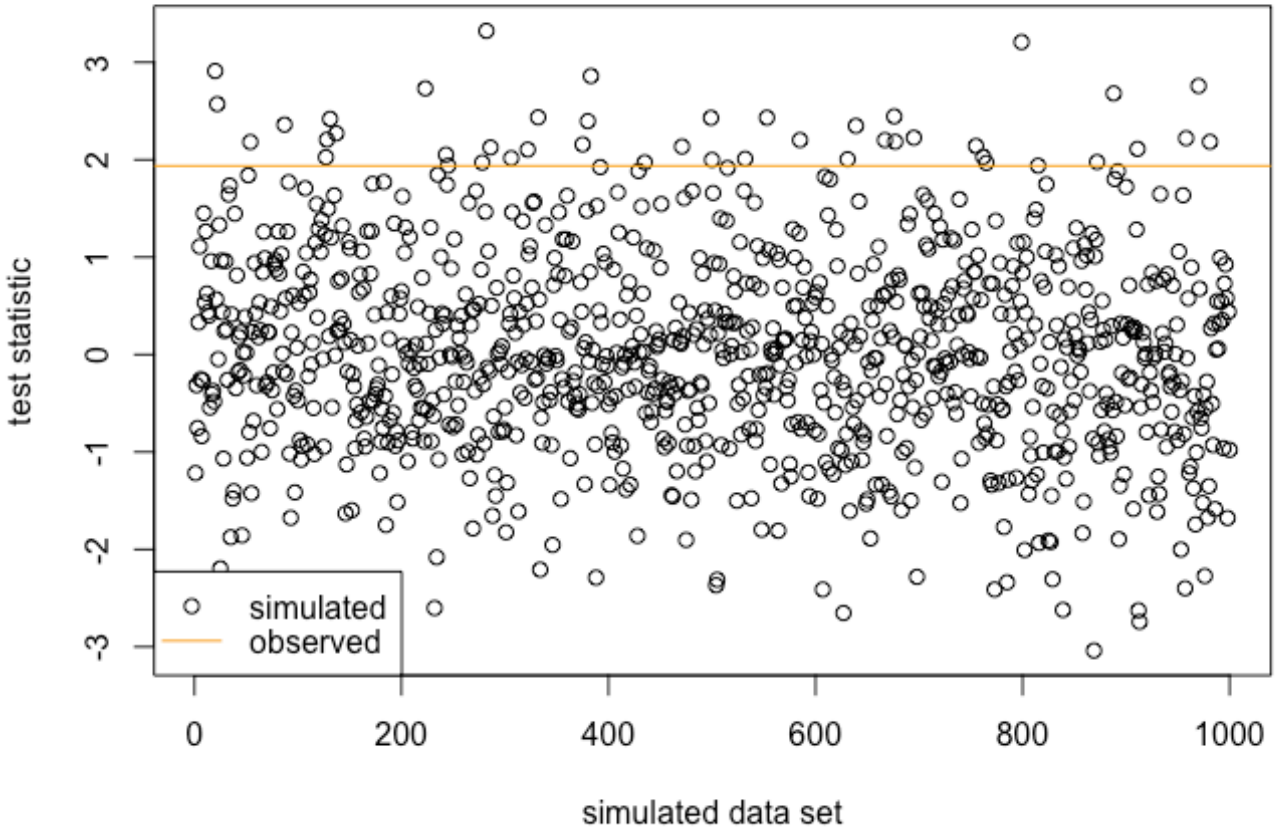
Steps for a statistical test



Illustrated example



Illustrated example (cont.)

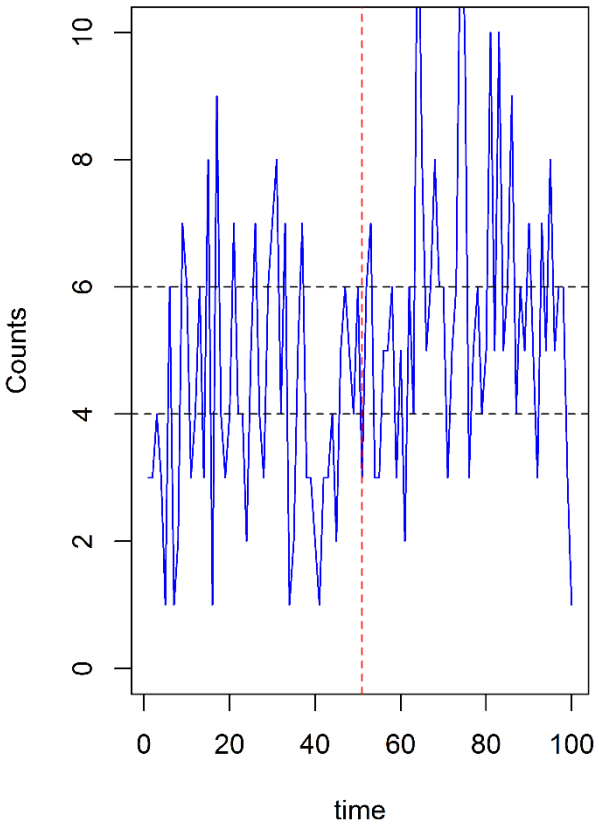


Development of CUSUM methods

- The **Cumulative Sum (CUSUM)** method has been a popular method for detecting outbreak of disease over time.
  - The **Exponentially-Weighted Moving Average (EWMA)** method is another.
- The original CUSUM method was developed for normally-distributed data in an industrial process control setting by Page in 1954.
- Later, the Poisson CUSUM was developed for count data by Lucas in 1985.

A shift in the mean of Poisson counts

- Mean of Poisson distribution for days 1-50, $\lambda_0 = 4$.
- A shift in mean occurs at day 51.
- Mean of Poisson distribution for days 51-100, $\lambda_1 = 6$.



Poisson CUSUM Method

- Cumulate recent observations using a statistic C_t (t denotes the time) assuming they show evidence of a disease outbreak.
- $C_0 = 0$
- $C_t = \max(0, C_{t-1} + Y_t - k)$
  - Y_t is the number of observed cases at time t.
  - k is a parameter related to the mean number of cases you expect to see and the mean number of cases you wish to detect if there is an outbreak.
- Signal that an outbreak has occurred if $C_t > h$, where h is some threshold chosen to control an error rate.

Determining the threshold

- The threshold *h* can be determined from the value of the parameter *k* and the desired in-control average run length ARL_0.
  - ARL_0 is the average time between false alarms.
  - **Type I** error rate is usually controlled in statistical tests.
    - This is the probability of incorrectly concluding the research hypothesis is true.
- The CUSUM method is only designed for a time series at a **single location**.

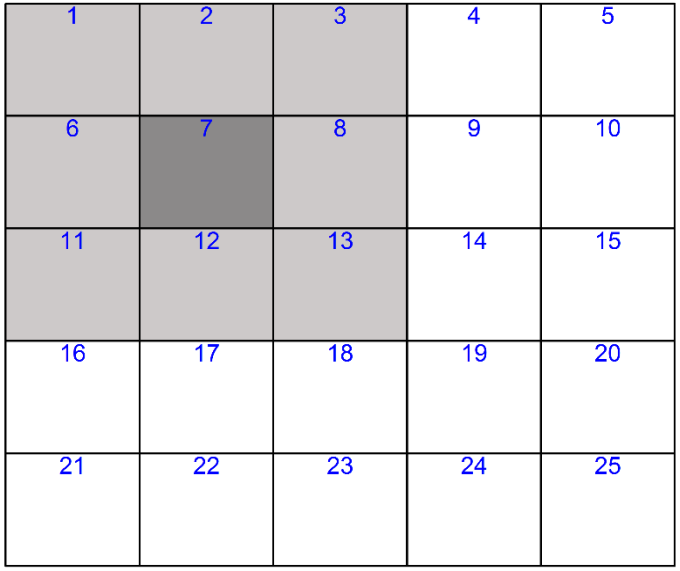
Room for a better approach?

- Should use spatial relatedness of data from different regions.
- Why control average run length (ARL_0)?
  - We can't stop the process like in industrial process control.
- Why not control the proportion of alarms that are false (sounding an alarm when there is no outbreak)?
  - This is known as the **False Discovery Rate (FDR)**.
  - There are powerful statistical procedures that can be used to control the FDR when performing multiple statistical tests simultaneously.

Objective

- To develop a computationally simple and fast algorithm for rapid detection of outbreaks producing easily interpretable results.
- The Center for Disease Surveillance (CDC) uses a seven-point guideline (CDC 1988) for evaluating a surveillance system: (1) simplicity, (2) flexibility, (3) acceptability, (4) sensitivity, (5) predictive value positive, (6) representativeness, and (7) timeliness.

The proposed method: the “big picture”



Proposed improved testing procedure

- Pool disease counts among connected regions before computing the CUSUM statistic.
- Simulate new data under the assumption that the null hypothesis is true.
- Use the simulated data to determine how extreme each observed statistic is.
- Use an FDR-controlling procedure to decide whether to sound an alarm.

SECTION III: Case study

Burden of foodborne diseases (Worldwide)

- Foodborne diseases outbreaks have devastating health and economic consequences in both developed and developing countries.
- The WHO identified the need to estimate the full extent of the disease burden associated with unsafe food.
- WHO launched an initiative to estimate the Global Burden of Foodborne Diseases in 2006.

Burden of foodborne diseases (United States)

- Even in the United States, foodborne illnesses are an important public health problem.

CDC estimates that each year:

- roughly 1 in 6 Americans (or 48 million people) get sick
- 128,000 are hospitalized, and
- 3,000 die of foodborne diseases.

- National surveillance for foodborne and waterborne disease outbreaks has been a core function of CDC since the 1970s.

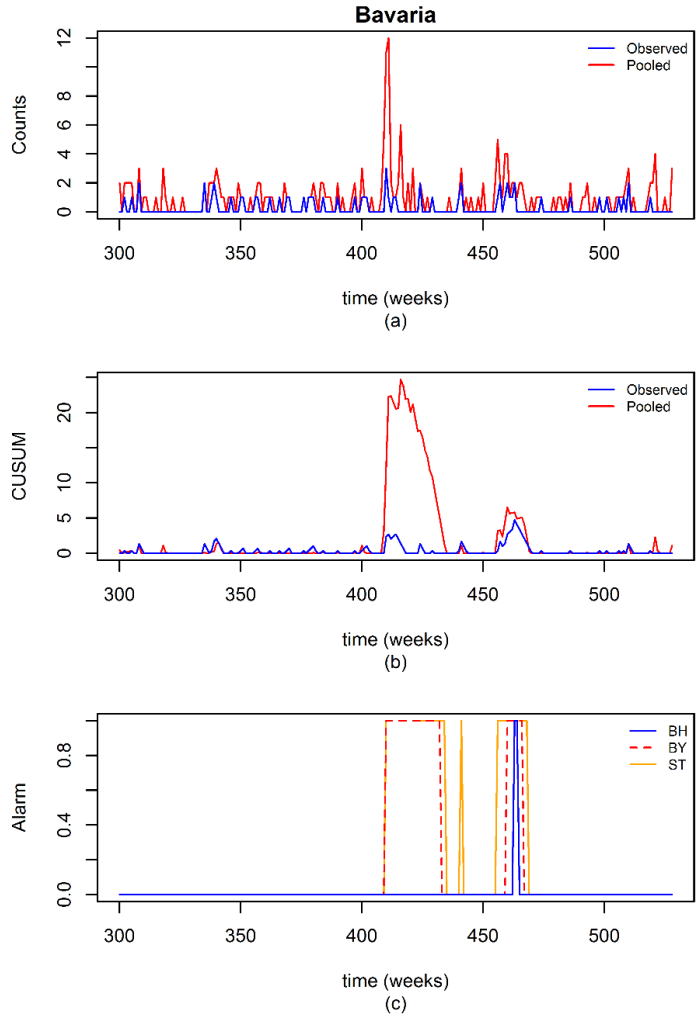
Salmonella Newport outbreak in Germany

- Weekly disease counts available for 16 German federal states between 2004-2014.
- The first two years of data (2004-2005) were used to estimate the null distribution.
- The actual outbreak occurred in the first week of November, 2011(week 410).

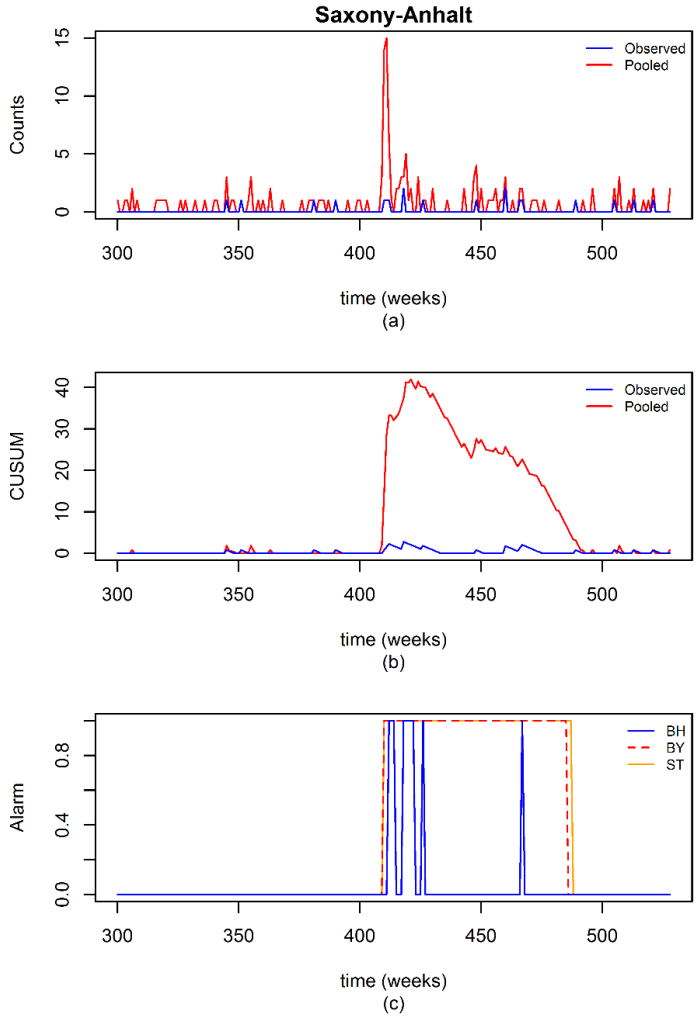
Other details

- An adjustment has to be made since larger populations are generally expected to have larger disease counts and smaller populations are expected to have smaller disease counts.
- Three different FDR-controlling procedures were compared:
  - The Storey-Tibshirani (ST) procedure (for correlated data)
  - The Benjamini-Yekutieli (BY) procedure (for correlated data)
  - The Benjamini-Hochberg (BH) procedure (for independent data)

Results: state of Bavaria



Results: state of Saxony-Anhalt



Comparison: No pooling model using BH procedure and the pooled model using ST procedure

State	Conventional method: Multiple Poisson CUSUM statistics Time of detection	Proposed method: Pooled Model using the ST procedure Time of detection
Baden Wuerttemberg	412	410
Bavaria	463	410
Berlin	410	410
Brandenburg	410	410
Bremen	Not detected	410
Hamburg	410	410
Hesse	410	410
Mecklenburg Vorpommern	410	410
Lower Saxony	410	410
North Rhine Westphalia	410	410
Rhineland Palatinate	412	410
Saxony	410	410
Saxony Anhalt	412	410
Schleswig Holstein	412	410
Thuringia	410	410

Pooling vs no pooling

- The proposed method using the ST procedure detected the outbreak consistently in all 15 states in week 410 (when the outbreak actually occurred)

- With the independent model using the BH procedure:
  - In 1 state it was not detected at all.
  - In 1 state it was detected 53 weeks later.
  - In 4 states the outbreak was detected 2 weeks later.
  - The outbreak was detected immediately for the other regions.

Comparison: Traditional multiple Poisson CUSUM vs. the Proposed method using ST procedure

State	Conventional method: Multiple Poisson CUSUM statistics Time of detection	Proposed method: Pooled Model using the ST procedure Time of detection
Baden Wuerttemberg	Not detected	410
Bavaria	Not detected	410
Berlin	410	410
Brandenburg	411	410
Bremen	Not detected	410
Hamburg	410	410
Hesse	Not detected	410
Mecklenburg Vorpommern	Not detected	410
Lower Saxony	411	410
North Rhine Westphalia	411	410
Rhineland Palatinate	Not detected	410
Saxony	Not detected	410
Saxony Anhalt	411	410
Schleswig Holstein	Not detected	410
Thuringia	412	410

State-of-the-art Poisson CUSUM vs. Proposed method using ST procedure

- The proposed method using the ST procedure detected the outbreak consistently in all 15 states in week 410 (when the outbreak actually occurred)

- With the conventional multiple Poisson CUSUM method:
  - No change was detected in 8 out of the 15 states.
  - One week delay in detection in 4 states.
  - Two week delay in detection in 1 state.
  - Zero delay in detection in 2 states.

SECTION IV: Future directions

Work in Progress/Future Work

- Construct a procedure based on the EWMA method, which requires fewer assumptions.
  - Fewer false alarms after outbreak ends.
- Formally account for correlation among regions geographically close to one another.
- The described methodology is for outbreaks with no trend or seasonality.
  - A future research direction is to extend the method to a broader class of settings, encompassing outbreaks with trend and seasonality.

Take Home Thoughts

- Disease outbreaks seem likely to become more common as the globe is energized.
  - We need to detect outbreaks as quickly as possible in order for public health agencies to respond quickly in order to mitigate adverse impacts.
- It is often difficult to distinguish disease outbreak from random variation when looking at disease counts.
- Good statistical methodology can help to correctly identify disease outbreaks WHILE controlling for testing error.
- Are you using the best tool for the job?

Questions? Comments? Poems?