Statistical modelling in infectious disease epidemiology

Michael Höhle[1]

[1] Department for Infectious Disease Epidemiology
Robert Koch Institute, Berlin, Germany

Module on Infectious Disease Epidemiology
Charité – Universitätsmedizin Berlin
23 June 2011

ROBERT KOCH INSTITUT

Outline

1. Introduction
2. Mathematical models for communicable diseases
   - Discrete time stochastic Reed-Frost model
   - Continuous time deterministic SIR model
3. Statistical Surveillance of routine collected infectious disease data
   - Farrington algorithm
   - Cumulative Sum
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

Outline

1. Introduction
2. Mathematical models for communicable diseases
3. Statistical Surveillance of routine collected infectious disease data
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

Definitions and aim of this lecture

Infectious disease epidemiology

Characterizes the epidemiological analysis of *infectious diseases*. Interest lies in the detection and understanding of epidemics. One possible aim would be the ability to better control outbreaks.

Aims of this lecture:

- Give a taste of how statistical modelling can be of use in infectious disease epidemiology.
- Illustrate this by plenty of examples from theory and practice.

Statistical modelling of infectious diseases

Three reasons that classical statistical inference is not immediately applicable for infectious disease data:

1. Data are rarely a result of planned experiments
2. Individuals are not independent (a case may also be a risk factor)
3. The infection process is only partially observable

A small outbreak experiment...

Outline

1. Introduction
2. Mathematical models for communicable diseases
   - Discrete time stochastic Reed-Frost model
   - Continuous time deterministic SIR model
3. Statistical Surveillance of routine collected infectious disease data
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

Outline

1. Introduction
2. Mathematical models for communicable diseases
   - Discrete time stochastic Reed-Frost model
   - Continuous time deterministic SIR model
3. Statistical Surveillance of routine collected infectious disease data
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

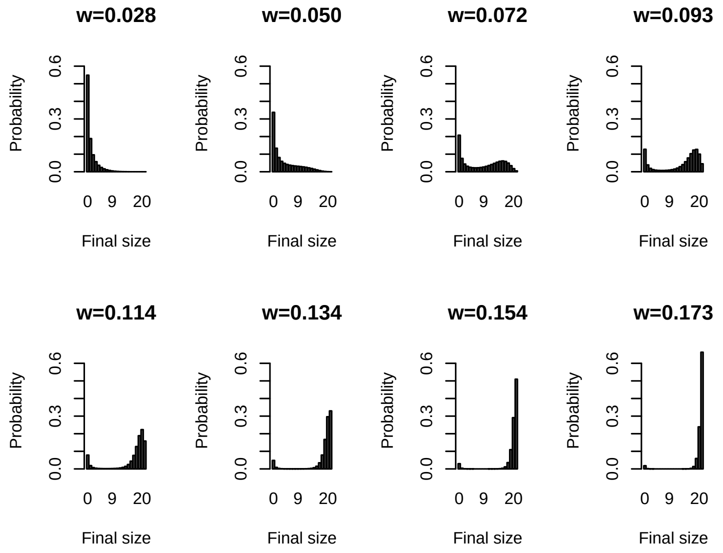
The Reed-Frost epidemic model

- SIR compartmental model, where individuals are either *S*usceptible, *I*nfectious or *R*ecovered
- Closed population with initially $x_0 = n$ susceptible and $y_0 = m$ infectious individuals
- Dynamics are described in discrete time by evolution of a Markov chain

  $$\begin{aligned} Y_{t+1} | x_t, y_t &\sim \text{Bin}(x_t, 1-(1-w)^{y_t}), \\ X_{t+1} &= X_t - Y_{t+1}, \end{aligned}$$

  where w is the probability for a contact of two individuals during one unit of time and $t = 1, 2, \ldots$
- The *final size* of the epidemic is $Z = Y_1 + Y_2 + Y_3 + \ldots$

Final size subject to the contact probability w



Estimation of model parameters

- Estimation of w from data for times $0, 1, 2, \ldots, K$ using e.g. maximum likelihood

  $$L(w) = \prod_{t=0}^{K-1} \theta_t^{y_{t+1}} (1-\theta_t)^{x_t - y_{t+1}},$$

  where $\theta_t = 1 - (1-w)^{y_t}$.
- Contribution of statistics: Interest is not only in point estimator $\hat{w}$, but also in a quantification of the uncertainty of $\hat{w}$, e.g. by stating a 95% confidence interval for w.

Research questions

- What effect does a vaccination have?
- What effect does an isolation measure have?
- How could the model take different age categories into account?
- Not every infected does actually become infectious.
- The population is not closed, what now?
- It's the rodent, stupid!

Mathematical challenges

- Mathematical abstractions of real world phenomena → *equations*
- No outbreaks are similar → *stochasticity*
- Different modes of transmission: person-to-person, air-borne, water-borne, food-borne and vector-borne → *direct and indirect transmission*
- Population heterogeneity (e.g. different places of residence, contact behaviour, susceptibility) needs to be taken into account
- Conflict between observation frequency and speed of the epidemic → *time unit of a model*
- Not all relevant events for the course of the epidemic are observable → *partial observability*

Statistical challenges

Statistics in a nutshell:

Stochastic model + data →
Parameter estimation + quantification of uncertainty

- Only one realization of the epidemic is observed.
- The data used for estimation can contain serious problems, e.g. under-reporting, changes in the test behaviour.
- The analysis is conducted using available covariables, but no information about the central risk factors is available.

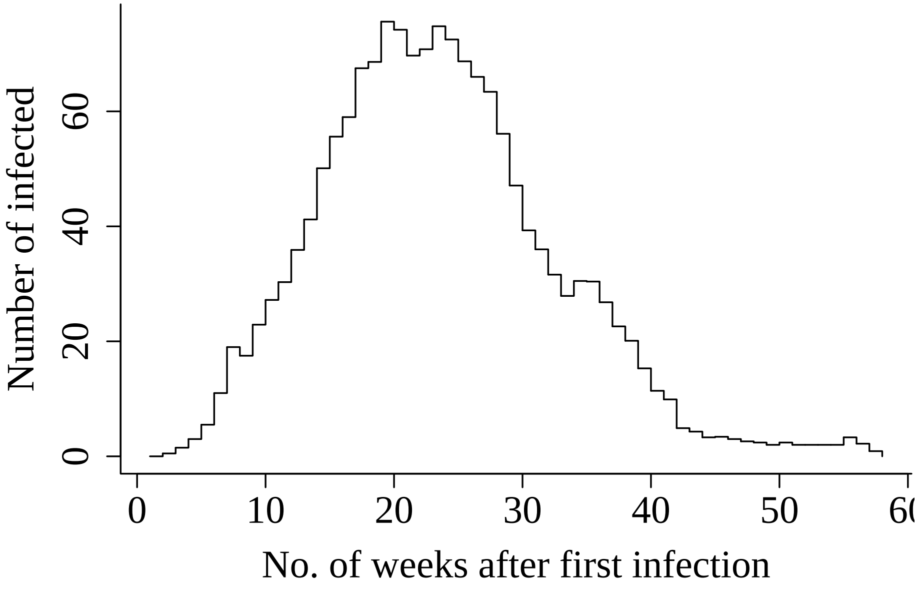
Outline

1. Introduction
2. Mathematical models for communicable diseases
   - Discrete time stochastic Reed-Frost model
   - Continuous time deterministic SIR model
3. Statistical Surveillance of routine collected infectious disease data
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

CSFV in The Netherlands (1)

- Classical Swine Fever Virus (CSFV) Characteristics
  - Symptoms after infection: dullness and anorexia.
  - Acute form: rapid mortality often without clinical symptoms.
  - Secondary symptoms: diarrhea or respiratory problems.
- Epidemic in the Netherlands lasted from 4 February 1997 to May 1998.
  - 429 infected farms detected and stamped out (∼ 700,000 pigs)
  - 1286 herds pre-emptively-slaughtered (∼ 1.1 million pigs)

CSFV in the Netherlands (2)



The basic SIR model (1)

- When the considered population is large, it can be sufficient to disregard the stochasticity of the epidemic process and use deterministic models.
- Can formulate a continuous time deterministic *SIR model* by using ordinary differential equations.
- The deterministic system intends to model the mean behaviour of the underlying stochastic system.
- We assume a closed population (i.e. no demographics turnover) of size N.

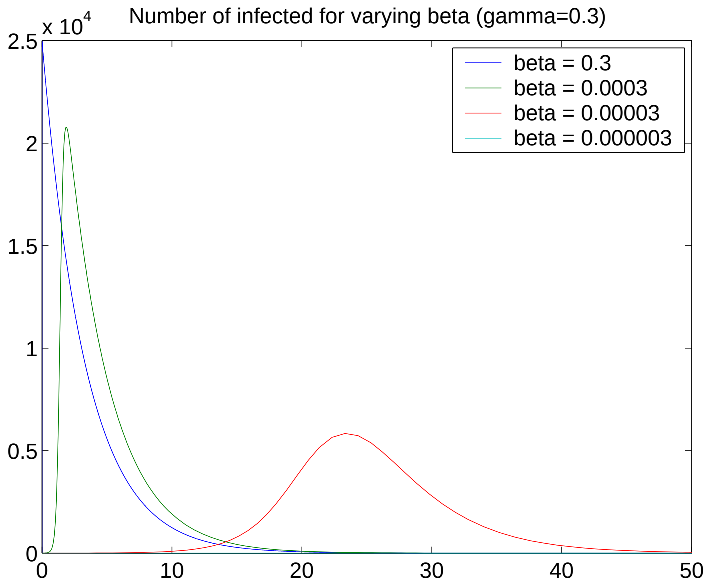
The basic SIR model (2)

- Divide population into three groups *(S)uscpetibles*, *(I)nfectious*, and *(R)ecovered*. At all times $S(t) + I(t) + R(t) = N + a$.
- Describe dynamics using an ODE-system.

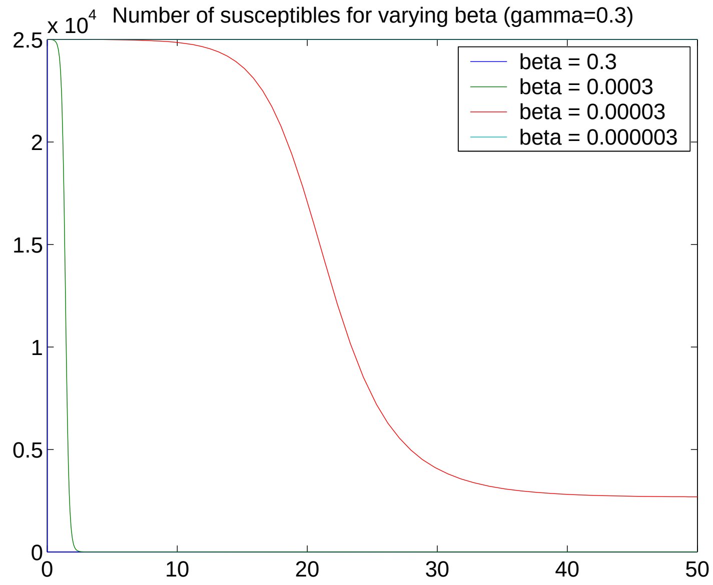
$$
\begin{aligned}
\frac{dS(t)}{dt} &= -\beta S(t)I(t) \\
\frac{dI(t)}{dt} &= \beta S(t)I(t) - \gamma I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t)
\end{aligned}
$$

- Solve ODE with initial condition $(N, a, 0)$ using numerical routines.

The basic SIR model (3)



The basic SIR model (4)



The basic reproduction rate R_0

- By definition: R_0 is the number of new infections produced by one infection in a virgin population, i.e. the initial growth rate.
- If $R_0 < 1$ the number of infected is expected to fade out right after introduction. If $R_0 > 1$ an epidemic will result.
- In a simple SIR model

$$R_0 = \frac{\beta N}{\gamma}$$

The final size of an Epidemic

- In a closed population the number of susceptibles can only decrease, hence it must have a limit for $t \to \infty$.
- Is the limit zero? Or will some fraction of the population escape from ever getting infected?
- Let f be the fraction of the (initially susceptible) population that got infected. This is also called the *final size* of the epidemic.
- It can be found as the solution to the equation

$$1 - f = \exp(-R_0 f).$$

How to estimate the parameters from data?

Depends on the available data from the epidemic.

- Final size data $\Rightarrow$ use Equation (9), i.e.

$$R_0 = -\frac{\log(1-f)}{f}$$

- Some function of recovery and infection times is observed at k discrete time points $\Rightarrow$ Least squares fit.

Least squares fit

- We have k observations of type $\mathbf{y}_i = g(\mathbf{x}(t_i, \boldsymbol{\theta}))$, where $\boldsymbol{\theta} = (\beta, \gamma)'$, $\mathbf{x}(t) = (S(t), I(t))'$ and $g(\cdot)$ is a function indicating that we might only observe part of the state.
- Least squares aims at finding $\boldsymbol{\theta}$, which minimizes the function

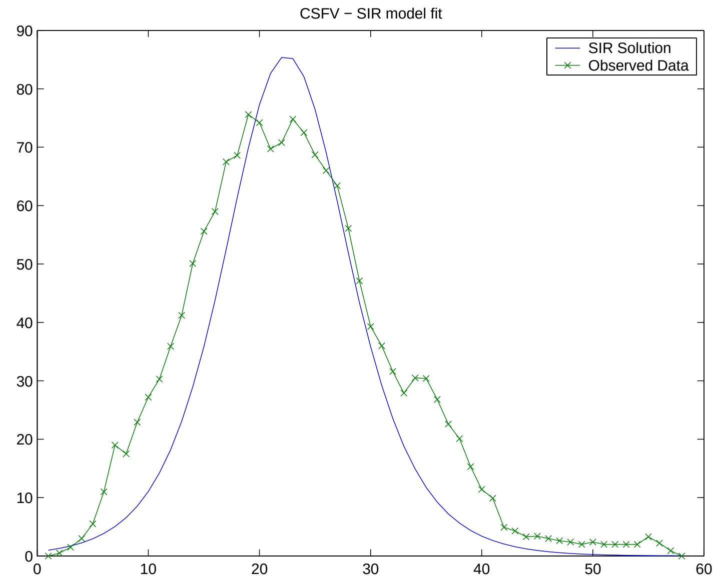
$$l(\boldsymbol{\theta}) = \sum_{i=1}^{k} \left(\mathbf{y}_i - g(\mathbf{x}(t_k, \boldsymbol{\theta}))\right)^2 ,$$

- Solution $\hat{\boldsymbol{\theta}}$ is found using numerical optimizing routines.
- If $g((S(t), I(t))') = I(t)$ least squares corresponds to assuming Gaussian distributed observations, i.e.

$$y_i \sim N(I(t), \sigma^2)$$

where σ is variance of the observation noise.

Parameter Estimation for the CSFV Data



Outline

1. Introduction
2. Mathematical models for communicable diseases
3. Statistical Surveillance of routine collected infectious disease data
    - Farrington algorithm
    - Cumulative Sum
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

Aims of statistical surveillance

Public health surveillance

Ongoing systematic collection, analysis, interpretation and dissemination of health data for the purpose of preventing and controlling disease, injury, and other health problems (Thacker, 2000).

Course view:

- Real-time online monitoring within a setting of statistical process control.
- Detect aberrations for public health events in a statistical setting with a little less heuristics involved than sometimes applied at the moment.
- Provide formal tool as a supplement to gut instinct.

Monitoring routine collected public health data

- The development of automated algorithms for the detection of abnormalities is demanded by the vast amount of data resulting from public health reporting
- This part is about prospective statistical monitoring of routinely collected surveillance data seen as multiple time series of counts and categories
- The statistical methods of this talk are implemented in the R-package `surveillance` available from the Comprehensive R Archive Network (CRAN) (H., 2007), but basic ideas can just as well be implemented in, e.g., Excel.

Examples of disease surveillance applications

In human epidemiology

- Monitoring of congenital malformations (Chen, 1978)
- Surveillance of notifiable diseases (Robert Koch Institute, 2009; Widdowson et al., 2003)
- Monitoring surgical outcomes (Steiner et al., 2000)

In veterinary epidemiology

- Salmonella in livestock reports, Veterinary Laboratories Agency, UK (Kosmider et al., 2006)
- Rabies Surveillance (WHO Collaboration Centre for Rabies Surveillance and Research, 2007)
- Monitoring of abortions in dairy cattle (Carpenter et al., 2007)

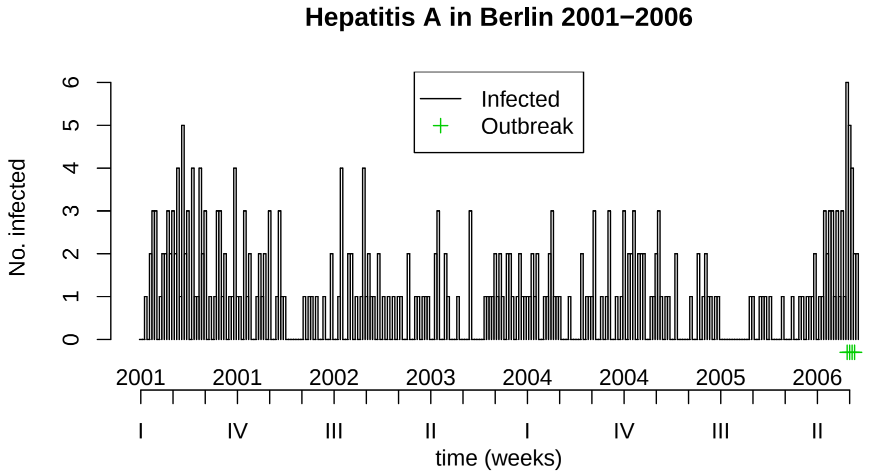
The quality of surveillance data

Issues complicating the statistical analysis of the time series

- Lack of clear case definitions
- Under-reporting and reporting delays
- Often no denominator data
- Seasonality
- Low number of disease cases
- Presence of past outbreaks

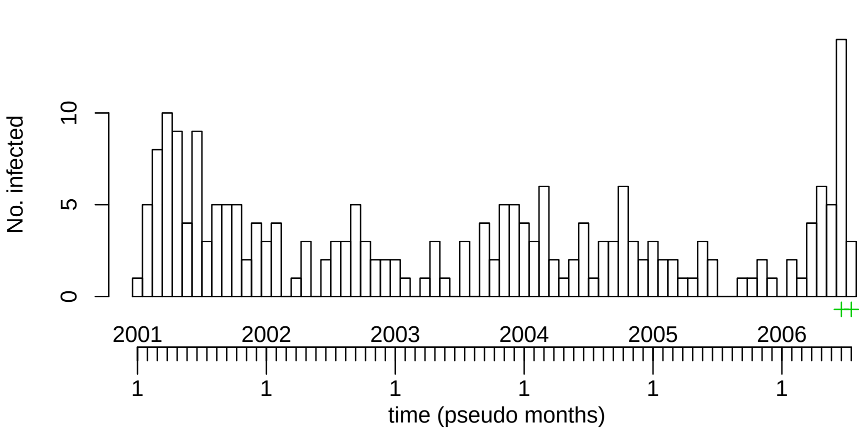
Example of surveillance data

- Weekly number of adult male hepatitis A cases in the federal state of Berlin during 2001-2006
- During the summer 2006 health authorities noticed an increased amount of cases (Robert Koch Institute, 2006).



Example of surveillance data

- Weekly number of adult male hepatitis A cases in the federal state of Berlin during 2001-2006
- During the summer 2006 health authorities noticed an increased amount of cases (Robert Koch Institute, 2006).



Six city districts aggregated into 4 week blocks

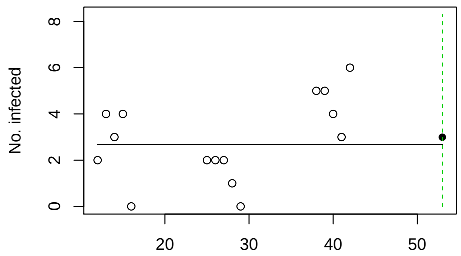
Outline

1. Introduction
2. Mathematical models for communicable diseases
3. Statistical Surveillance of routine collected infectious disease data
   - Farrington algorithm
   - Cumulative Sum
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

Farrington algorithm (1) – basic model

- Predict value y_{t_0} at time $t_0 = (t_0^m, t_0^y)$ using a set of reference values from window of size $2w + 1$ up to b years back.

Prediction at time t=53 with b=3,w=2



- Fit overdispersed Poisson generalized linear model (GLM) to the $b(2w + 1)$ reference values where $\mathrm{E}(y_t) = \mu_t$ and $\log \mu_t = \alpha + \beta t$.

Farrington algorithm (2) – outbreak detection

Predict and compare:

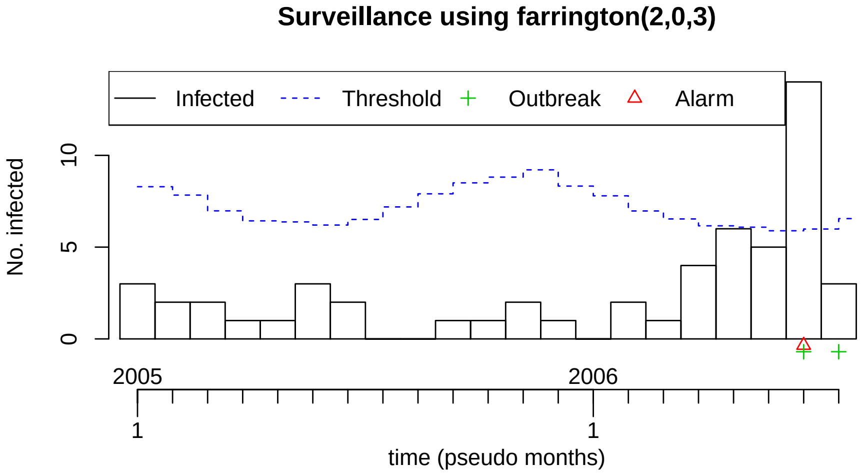
- An approximate $(1 - \alpha)\%$ prediction interval for y_{t_0} based on the GLM has upper limit $U = \hat{\mu}_{t_0} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\mathrm{Var}(y_{t_0} - \hat{\mu}_{t_0})}$
- If observed y_{t_0} is greater than U, then flag t_0 as outbreak

Remark:

- A description of the refinements of the algorithm is not given at this point.

Farrington algorithm (3) – example in surveillance

- Results for $w = 2, b = 3$ and $\alpha = 0.01$:



Idea for improvement

Cast disease monitoring into context of statistical process control

Outline

1. Introduction
2. Mathematical models for communicable diseases
3. Statistical Surveillance of routine collected infectious disease data
   - Farrington algorithm
   - Cumulative Sum
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

CUSUM as Surveillance Algorithm (1)

- A control chart known from statistical process control

Cumulative Sum (CUSUM)

In control situation $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0,1)$. Monitor shift to $N(\mu,1)$ by

$$S_t = \max(0, S_{t-1} + X_t - k), \quad t = 1, \ldots, n$$

where $S_0 = 0$ and k is the *reference value*. Raise alarm if $S_t > h$, where h is called the *decision interval*.

- CUSUMs are better to detect sustained shifts
- Given h and k we can determine the *average run length* (ARL)

CUSUM as Surveillance Algorithm (2)

- CUSUM for count data $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Po}(m)$ by transforming data to normality (Rossi et al., 1999)

$$X_t = \frac{Y_t - 3m + 2\sqrt{m \cdot Y_t}}{2\sqrt{m}}$$

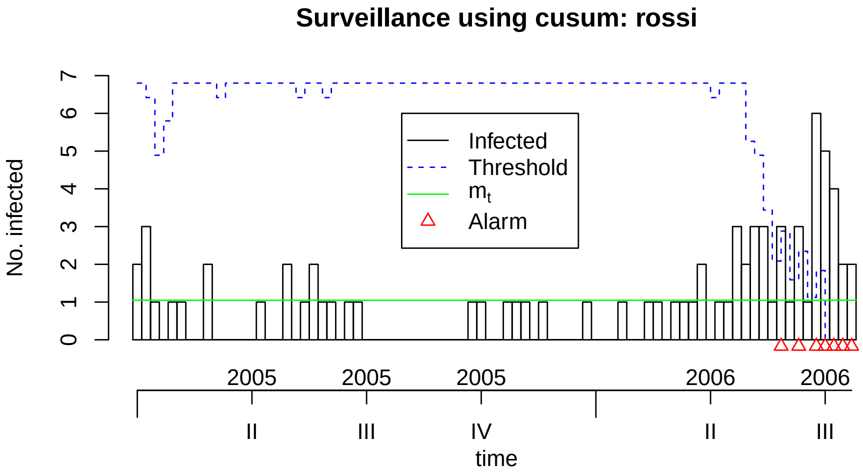
- Risk-adjust the chart by letting m be time varying, e.g. as output of a Poisson GLM model

$$\log(m_t) = \alpha + \beta t + \sum_{s=1}^{S} (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)),$$

where $\omega_s = \frac{2\pi}{52} s$ are the Fourier frequencies.

CUSUM as Surveillance Algorithm (3)

- Results for $k = 0.59$ and $h = 3.80$:



CUSUM as Surveillance Algorithm (4)

- Simulation studies show: For low counts it is better to use CUSUM directly on the counts instead of on transformed residuals
- Proposals for this setting implemented in `surveillance` are:
  - Function `rogerson`, which uses a reweighted Poisson CUSUM (Rogerson and Yamada, 2004)
  - Function `glrnb`, which uses a likelihood ratio and generalized likelihood ratio detector (H. and Paul, 2008)
- More flexibility to model the time series and to tune the detection algorithm → more work for each time series

Evaluating performance of a surveillance algorithm

Choice of threshold in surveillance algorithms should be based on performance measure:

- Let N be a discrete random variable denoting the first time that $S_t > h$, aka. the run-length
- Location parameters of the run length distribution, e.g. the ARLs $\mathrm{E}(N|\tau = 0)$ or $\mathrm{E}(N|\tau = \infty)$.
- Conditional expected delay $E(N - \tau|\tau, N \geq \tau)$
- Probability of false alarm within first m time points, i.e. $P(N \leq m|\tau = \infty)$.
- Sensitivity, Specificity, ROC-Curves

Computation of measures rarely available as closed formulas. Instead Monte-Carlo sampling is used.

Outline

1. Introduction
2. Mathematical models for communicable diseases
3. Statistical Surveillance of routine collected infectious disease data
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

EHEC/HUS Outbreak in Germany 2011

The lecture contained an additional number of slides on the EHEC/HUS break, i.e.

- Epicurve and animation of reporting delays
- Now-casting
- Backprojection from disease onset to exposure times

These slides are not available in the internet accesible version of the slides.

Outline

1. Introduction
2. Mathematical models for communicable diseases
3. Statistical Surveillance of routine collected infectious disease data
4. Applying methods in practice – EHEC/HUS outbreak
5. Summing up

Other tasks of a statistician...

Statistical analysis of which factors have an influence on the disease incidence on district level.

- Example 1: Association between Q fever 5 year incidence and the density of Sheep, Goat and Cattle per ha agricultural used area in the district.
- Example 2: Hantavirus incidence per year and district in connection with proxy variables for the bank vole population.
- Example 3: Spatio-temporal association between invasive meningococcal disease (continuous in space and time) and influenza cases (notified cases in the respective district). Characterize spread dynamics for different disease finetypes.

Addendum to initial statement

Addendum to why application of classical statistical inference is less trivial in outbreak situations:

- There is sometimes a need for urgency! (Giesecke, 2002)
- Consistency is sometimes more important than accuracy:

  > *Although this has the virtue of brevity, it is (aptly) not quite true. Both are important, but in surveillance absolute accuracy is unachievable. Consistency can be achieved and failure to do so is extremely damaging to credibility among the ignorant classes (especially journalists) as has been observed. (Cowden, 2010)*

- Afterwards, you (and unfortunately also many others) always know better.

Still there is lots to gain from mathematical and statistical modelling. A precondition is a multi-disciplinary approach!

Additional information

Books which might provide further information:

- Modern infectious disease epidemiology, J. Giesecke – nice overview without too much mathematical detail, but little on modelling.
- Analysis of infectious disease data, Niels G. Becker – nicely describes a statistical oriented view to modelling. Nothing on SIR modelling. Equations!
- Modeling Infectious Diseases in Humans and Animals, M. J. Keeling & P. Rohani – Very good overview on mathematical modelling. Little on the statistical aspects. Equations!

Another outbreak experiment...

Literature I

Carpenter, T. E., Chriel, M., and Greiner, M. (2007). An analysis of an early-warning system to reduce abortions in dairy cattle in Denmark incorporating both financial and epidemiologic aspects. *Preventive Veterinary Medicine*, 78:1–11.

Chen, R. (1978). A surveillance system for congenital malformations. *Journal of the American Statistical Association*, 73:323–327.

Cowden, J. M. (2010). Some haphazard aphorisms for epidemiology and life. *Emerging Infectious Diseases*, 16(1):174–177.

Giesecke, J. (2002). *Modern infectious disease epidemiology*. Hodder Arnold, 2nd edition.

Höhle, M. (2007). surveillance: An R package for the monitoring of infectious diseases. *Computational Statistics*, 22(4):571–582.

Höhle, M. and Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics & Data Analysis*, 52(9):4357–4368.

Kosmider, R., Kelly, L., Evans, S., and Gettinby, G. (2006). A statistical system for detecting salmonella outbreaks in British livestock. *Epidemiology and Infection*, 134:952–960.

Robert Koch Institute (2006). Epidemiologisches Bulletin 33. Available from http://www.rki.de.

Robert Koch Institute (2009). SurvStat@RKI. http://www3.rki.de/SurvStat. Last access: 9 June 2009.

Rogerson, P. and Yamada, I. (2004). Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report*, 53:79–85.

Literature II

Rossi, G., Lampugnani, L., and Marchi, M. (1999). An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine*, 18:2111–2122.

Steiner, S. H., Cook, R. J., Farewell, V. T., and Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, 1(4):441–452.

Thacker, S. B. (2000). Principles and practice of public health surveillance. chapter Historical Development, pages 1–16. Oxford University Press, 2nd edition.

WHO Collaboration Centre for Rabies Surveillance and Research (2007). WHO Rabies - Bulletin - Europe. http://www.rbe.fli.bund.de/.

Widdowson, M.-A., Bosman, A., van Straten, E., Tinga, M., Chaves, S., van Eerden, L., and van Pelt, W. (2003). Automated, laboratory-based system using the internet for disease outbreak detection, the Netherlands. *Emerging Infectious Diseases*, 9(9):1046–1052.