Two Problems Related to Biosurveillance

Myron J. Katzoff
Office of Research and Methodology
Centers for Disease Control and Prevention
National Center for Health Statistics (NCHS)
Hyattsville, MD

May, 2008

A Definition and Objective

• **Biosurveillance**(HSPD-21): the process of active data-gathering with appropriate analysis and interpretation of biosphere data that might relate to disease activity and threats to human or animal health – whether infectious, toxic, metabolic or otherwise and regardless of intentional or natural origin – in order to achieve early warning of health threats, early detection of health events and overall situational awareness of disease activity.

•**Objective**: To describe how extreme value theory might be used in a biosurveillance problem.

The Problem

• Soon after exposure to a variety of different pathogens, victims will present symptoms of influenza-like illness(IFI) .

• Such exposures would very likely be “hidden” if they occurred during the “flu season”.

• A clue to the possibility of such exposures might be a sudden increase in the incidence of influenza.

• How can we assess the extremeness of the number of influenza cases in a population?

This could be our earliest indication that something out-of-the-ordinary is happening.

Outline

Consider EV theory from the perspective of what is needed for *statistical modeling* as an aid in decision-making when extreme natural or man-made catastrophic events occur

- Block maxima models
- Threshold excess models
- Time/spatial-location (nonstochastic) dependence
- Stochastic dependence
- SUMMARY

A First Fundamental Result: Extremal Types

Theorem 1:Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of IID r.v.'s and let $M_n = \max\{X_1, X_2, \ldots, X_n\}$. If there exist sequences of constants $\{\alpha_n > 0\}$ and $\{\beta_n\}$ such that

$$\Pr\left[\frac{M_n - \beta_n}{\alpha_n} \leq z\right] \to G(z) \text{ as } n \to \infty$$

where $G(\cdot)$ is a nondegenerate DF, then G is a member of the generalized extreme value (GEV) family of DF's:

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

where $\sigma > 0$ and $-\infty < \mu < \infty$; and the support is $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$.

Block Maxima Models

(creating a straw man)

• Suppose we have several years (say, m) of weekly observations and that for a crude and very quick analysis pertinent to the current year we are willing to regard each previous year as providing 52 independent observations from a common distribution.

• Let X_j denote the *maximum number of cases* in each year (a year providing a block of observations) for $j = 1, 2, \ldots, m$.

Block Maxima Models

(creating a straw man) *continued*

We might then consider maximum likelihood estimation of μ, σ and ξ in $G(\cdot)$ of Theorem 1 – that is, we would find values for these parameters which maximize the log-likelihood

$$l(\mu, \sigma, \xi) = -m \ln \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{m} \ln \left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right] - \sum_{j=1}^{m} \left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

Some Problems/Issues

- Systematic components of seasonality and trends

- Dependence among the observed numbers of cases in adjacent weeks

- Changes in the population base: numbers of people; mix by age, race and sex

- Spatial distribution and clustering of a population at risk

etc.

Some Problems/Issues *continued*

From Coles (2001)

• "... modeling only block maxima is a wasteful approach to extreme value analysis if other data on extremes [for example, the five largest values] are available"

• " If an entire time series of ... observations is available, then better use is made of data by avoiding altogether the procedure of blocking."

A Second Fundamental Result:

Excess Above A Threshold, u

<u>Theorem 2</u>: As before, let $\{X_i\}_{i=1}^{\infty}$ be a sequence of IID r.v.'s and let $M_n = \max\{X_1, X_2, \ldots, X_n\}$. Suppose that the conditions of Theorem 1 are again satisfied so that for the DF common to all the r.v.'s of the sequence there exist sequences of constants $\alpha_n > 0$ and β_n such that

$$\Pr\left[\frac{M_n - \beta_n}{\alpha_n} \leq z\right] \rightarrow G(z) \quad \text{(nondegenerate)}$$

as $n \rightarrow \infty$ and

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

for some μ, σ and ξ. Then, for large enough u (a real number), the DF of $(X-u)$ conditional on $X > u$ is given approximately by the generalized Pareto distribution function – that is,

$$H(y) = \Pr[(X-u) \leq y] = 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{-\frac{1}{\xi}},$$

for y in $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$.

Important Value Added by Threshold Excess Models

• It is reasonable to expect that some level of incidence in IFI-symptoms is normal and nothing to be concerned about but, beyond a certain threshold, there is, indeed, reason for alarm!

• EV theory supplies the statistical models and the diagnostic procedures for determining what a useful threshold might be: Q-Q plots, Gumbel plots, mean-excess plots and Z- and W-statistics.

• EV theory provides estimates of the probabilities of various exceedances δ of the threshold u

$$\zeta_u \left[1 + \frac{\xi\delta}{\tilde{\sigma}}\right]^{-\frac{1}{\xi}},$$

where $\zeta_u = \Pr[X > u]$. If $\xi = 0$, the second factor is replaced by its limit as $\xi \to 0$, $\exp(-\frac{\delta}{\tilde{\sigma}})$.

Systematic Nonrandom Variation

- For the GEV distribution

$$G_Z(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

one finds that $E[Z] = \mu - \frac{\sigma}{\xi}(1+\Gamma(1-\xi))$, $\Gamma(\cdot)$ being the gamma function, and $Var[Z] = (\frac{\sigma}{\xi})^2\{\Gamma(1-2\xi) - \Gamma^2(1-\xi)\}$.

- It may be difficult to model the shape parameter, ξ, as a function of time. However, to account for the systematic components of trend and seasonal variation, it is not unreasonable to consider μ and, possibly, σ^2 as functions of time, t.

Systematic Nonrandom Variation *continued*

For subsequent analyses, it is then useful to note that the standardized form of Z, defined by

$$Z^* = \frac{1}{\xi} \ln \left[1 + \xi \left(\frac{Z - \mu}{\sigma} \right) \right],$$

has a Gumbel distribution $G(z) = \exp\{-e^{-z}\}$ so that assuming accurate modeling of $\mu(t)$ and $\sigma^2(t)$ ($t \equiv j$ for $j = 1, 2, \ldots$) the

$$Z_j^* = \frac{1}{\xi} \ln \left[1 + \xi \left(\frac{Z_j - \widehat{\mu}(j)}{\widehat{\sigma}(j)} \right) \right]$$

are approximately Gumbel. Here the "hatted" variables are m.l.e. or other consistent estimators.

Stochastic Dependence

• Use is made of a property frequently assumed in the analysis of some wide-sense stationary time series: the "strength" of the dependence weakens somewhat monotonically with the time-separation of r.v.'s

• One example is that of the $D(u_n)$ condition: if for all $i_1 < i_2 < \ldots < i_p < j_1 < j_2 < \ldots < j_q$ with $j_1 - i_p > \ell$

$$\Big|\mathrm{Pr}\Big\{X_{i_1} \le u_n, \ldots, X_{i_p} \le u_n, X_{j_1} \le u_n, \ldots, X_{j_q} \le u_n\Big\} - \mathrm{Pr}\Big\{X_{i_1} \le u_n, \ldots, X_{i_p} \le u_n\Big\} \cdot \mathrm{Pr}\Big\{X_{j_1} \le u_n, \ldots, X_{j_q} \le u_n\Big\}\Big| \le \alpha(n, \ell)$$

where $\alpha(n, \ell_n) \to 0$ for some sequence ℓ_n such that $\ell_n/n \to 0$ as $n \to \infty$. Examples: Gaussian m-dependent series, ARMA series.

Stochastic Dependence *continued*

Theorem 3: Let $X_1, X_2, \ldots$ be a stationary process and define M_n as before. Then, if $\{\alpha_n > 0\}$ and β_n are sequences of real numbers such that

$$\Pr\{(M_n - \beta_n)/\alpha_n\} \to G(z),$$

a nondegenerate DF, and the sequence satisfies the $D(u_n)$ condition, $u_n = \alpha_n z + \beta_n$ for all z, then G belongs to the GEV family.

Poisson Models

• Since Poisson models are generalized linear models, for an observed process Y_j, we have two basic possibilities for the inverse-link relationship when considering stochastic time-dependence. This can be exemplified by

$$g(\mu_j) = \overrightarrow{X_j}'\overrightarrow{\beta} + d_j$$

where, at time j, μ_j is the mean, $\overrightarrow{X_j}$ is a vector of covariates, $\overrightarrow{\beta}$ is a vector of regression coefficients, and d_j is either a latent process or an explicit function of the past observables: $Y_{j-1}, Y_{j-2}, \ldots, Y_1$.

• In the first specification of d_j, the process is called parameter-driven; in the second, observation-driven because the conditional expectation of the outcome given the past values of outcomes depends explicitly on those values.

Poisson Models *continued*

• The most useful class of models would provide for both positive and negative serial dependence, which is the case for parameter-driven models. However, methods for estimating the parameters of those models are computationally intensive.

• Observation-driven models are easier to deal with. But, for some of the observation-driven models, the requirement of stationarity imposes constraints on the values of model coefficients which exclude the possibility of positive dependence.

SUMMARY

Primary interests:

- Models of threshold excess

- Methods for dependent sequences that exploit stationarity and "weak" dependence

- Account for systematic variation with time

- Concentrate on methods for observation-driven series

- Account for differences in populations due to spatial distribution, indicators of health status and access to medical services

- Multivariate models to account for possible interactions due to simultaneously occurring events at several locations

Some Referencees

1. Coles, Stuart (2004). *An Introduction to Statistical Modeling of Extremes*. Springer. [Note: R code can be found on the internet.]

2. Smith, Richard L. (2003). *Statistics of Extremes, with applications in environment, insurance and finance*. [Excellent introductory tutorial. A preprint is available as a PDF file at http://www.stat.unc.edu/postscript/rs/semstatrls.pdf.]

3. Davis,R.A.;Dunsmuir,Wm.T.M. and Streett, S.B. (2003).

Observation-driven models for Poisson counts, *Biometrika*, v.90, pp.777-790.

4. Zeger, S.L. and Qaqish, B. (1988). Markov Models for Time Series: a quasi-likelihood approach. *Biometrics*, v.44,pp.1019-1031.

5. Davis,R.A.;Dunsmuir,Wm.T.M. and Wang, Y. (2000). On autocorrelation in a Poisson regression model,*Biometrika*, v.87,pp.491-505.

6. Zeger, S.L.(1988), A regression model for time series of counts,*Biometrika*,v.75,pp.621-9.