Temporal Surveillance Using Scan Statistics

Joseph Naus and Sylvan Wallenstein

presented by Curtis Storlie

The Underlying Problem

- A test for
  - H_0: arbitrary model with a known background rate of occurrence.
  - H_1: A spike or pulse is superimposed on this background rate.

Data Collection Scenarios

1. Continuous
   - the time of each occurrence is reported on a continuous scale.
2. Grouped
   - For each of T disjoint intervals, the number of occurrences during each interval is reported.
3. Binary
   - for each of T trials, it is reported if an event occurs or not

Temporal Scan Statistics

1. Continuous Data

   Continuous scan statistic: S_w = max number of events in a window of length w

2. Grouped Data

   Ratchet-scan statistic: S_w = max number of events in w consecutive intervals

3. Binary Data

   Binary scan statistic: S_w = max number of events in w consecutive trials

Data Processing Scenarios

a. Prospective (Real Time)

   – $P\{\text{type I error in any interval of length } T\} = \alpha$

b. Retrospective (Batch)

   – $P\{\text{type I error over the review period}\} = \alpha$

Our Focus

- We will focus on scenario 2 a.
  - Prospective processing of grouped data
  - Only temporal, not spatio-temporal like Kulldorff's paper
- The following approaches will be presented
  - P-scan
  - GLRT
  - CUSUM

Statement of the Problem

- Assume for the moment that we are only interested in the first T intervals.

- Let U_i be the number of events in the i^{th} interval.

- Assume $U_i \sim \text{Poisson}(\lambda_i)$

  H_0: $\lambda_i = \lambda_i^{(0)}$ for $i = 1, \ldots, T$

  H_1: $\lambda_i = \theta\lambda_i^{(0)}$ for $i = b - w + 1, \ldots, b$
  $\lambda_i = \lambda_i^{(0)}$ otherwise

  – where $\lambda_i^{(0)}$ is known for all i.

- This is what's called the pulse alternative.

Statement of the Problem

- This differs from Kulldorff's Setup slightly
- Kulldorff's setup would be $U_i \sim \text{Poisson}(\lambda_i)$

H_0: $\lambda_i = p\mu_i$ for $i = 1, \dots, T$

H_1: $\lambda_i = q\mu_i$ for $i = b - w + 1, \dots, b$
$\lambda_i = p\mu_i$ otherwise

- where μ_i is known for all i, but p is an unknown nuisance parameter.

P-Scan

- Let $Y_t(w) = \sum_{i=1}^{w} U_{t-w+i}$ which is the observed number of events in the consecutive intervals $t-w+1, t-w+2, \ldots, t$, for $t = w, \ldots, T$
- Let $E_t(w) = \sum_{i=1}^{w} \lambda_{t-w+i}$ be the expected number of events in these intervals.

P-Scan

- For the constant background case, $\lambda_i^{(0)} = \lambda^{(0)}$ for all i,
  - the GLRT is to reject H_0 for large values of the ratchet-scan statistic, $S_w = \max_t\{Y_t(w)\}$
  - The p-value for this constant background case is denoted

$$P(k; \lambda, w, T) = P(S_w \geq k \mid \lambda)$$

- Simple approximations for this quantity are available (i.e. no simulation required)

P-Scan

- The P-scan approach works as follows
  - For each observed $Y_t(w)$ compute $P(Y_t(w), E_t(w), w, T)$
  - The P-scan statistic is

$$PSS = \min_t \{P(Y_t, E_t, w, T)\}$$

  - Reject H_0 if $PSS \le \alpha$
  - Equivalently if for any t $P(Y_t, E_t, w, T) \le \alpha$
- This procedure will keep the overall type I error rate at α, $P\{PSS \le \alpha \mid H_0\} \le \alpha$. Proof given in Appendix.

Mid-p-value

- Consider a randomized test based on a discrete valued test statistic.

- That is, suppose we are to reject H_0 for large values of k which is the observed value of the discrete random variable K.

- Then we would reject H_0 if:

$$P(K \geq k \mid H_0) \leq \alpha$$

OR $P(K \geq k+1 \mid H_0) \leq \alpha < P(K \geq k \mid H_0)$ and $U \leq f$

  – where $U \sim U(0,1)$ independent of K and

$$f = \frac{\alpha - P(K(X) \geq k+1 \mid H_0)}{P(K \geq k \mid H_0) - P(K \geq k+1 \mid H_0)}$$

- Under this strategy $P(\text{type I error}) = \alpha$.

Mid-p-value and P-scan

- The randomized P-scan test would be to sound an alarm if for any t

$$P(Y_t, E_t, w, T) \leq \alpha$$

OR $P(Y_t + 1, E_t, w, T) \leq \alpha < P(Y_t, E_t, w, T)$ and $U \leq f$ where

$$f = \frac{\alpha - P(Y_t + 1, E_t, w, T)}{P(Y_t, E_t, w, T) - P(Y_t + 1, E_t, w, T)}$$

- Randomized tests have the unfavorable result that two researchers could get different answers from the same data.

- The **mid-p-value** approach would be to sound an alarm if for any t, $f \geq 0.5$ or equivalently if for any t

$$[P(Y_t + 1, E_t, w, T) + P(Y_t, E_t, w, T)]/2 \leq \alpha$$

GLRT

- The likelihood is

$$L(\boldsymbol{\lambda}; \boldsymbol{u}) = \prod_{i=1}^{T} \lambda_i^{u_i} e^{-\lambda_i} / u_i!$$

  – where u_i is the observed number of events in the i^{th} interval.

- which makes for a GLR of

$$\max_{b,\theta} \sum_{i=b-w+1}^{b} \{u_i \log \theta - \lambda_i(\theta - 1)\} =$$

$$\max_{b,\theta} \{Y_b(w) \log \theta - E_b(w)(\theta - 1)\}$$

GLRT

- For a given b, the max occurs at

$$\hat{\theta}_b = Y_b(w)/E_b(w)$$

- This leads to a GLRT that rejects for large values of

$$G(w) = \max_b \{Y_b(w) \log[Y_b(w)/E_b(w)] - [Y_b(w)/E_b(w)]\}.$$

- Notes
  - If the alternative hypothesis was additive instead of multiplicative, the GLRT would stay the same
  - If the value of w is unknown, but is known to be in the range $u \leq w \leq v$, then the GLRT test statistic is $G(w)$ maximized over the values of w in that range.

GLRT

- The GLRT for the Kulldorff Setup is slightly different,

$$\begin{aligned} G(w) &= \max_b \{Y_b(w) \log[Y_b(w)/E_b(w)] + \\ &\qquad (N - Y_b(w)) \log[(N - Y_b(w))/(N - E_b(w))]\}. \end{aligned}$$

  - where N = number of events in $[0, T]$.
  - The dependency on N is introduced by the fact that p is free in his specification of $\lambda_i^{(0)} = p\mu_i$,

CUSUM

- Variant of CUSUM where the quantity being summed is the same quantity on which the GLRT is based.

$$C(t) = \max[0, C(t-1) + \log(Y_t(1)/\lambda_t) - (Y_t(1)/\lambda_t)]$$

- Sound alarm if $C(t) > h$ where h is determined by

$$P\left(\max_{0 \leq t \leq T} C(t) \geq h \mid H_0\right) = \alpha$$

Simulation Results

- Under the null it is assumed that $U_t \sim \text{Poisson}(\lambda_t)$ where

$$\lambda_t = \gamma + \beta t \qquad \text{for } t = 1, \ldots, 52$$

- For these type I error results below, $\gamma = 2$ and $\beta = 0.06$.

$w \backslash \alpha$	.01	.05	.10	.20	.30
3	.010	.046	.099	.185	.288
5	.010	.047	.096	.189	.285

Table 1: Observed Type I Error Rates for P-Scan

Simulation Results

- Under the alternative

$$\lambda_t = 2 + 0.06t + \delta + \theta(t - 20)$$

  for intervals 20 to $20 + v - 1$.

- For each of the cases in the table, the expected number of excess events was kept constant at 15.

- Notice
  - setting $\theta = 0$ gives a pulse alternative.
  - setting $\theta > 0$ gives a gradual increase over time.

Simulation Results

w	v	δ	θ	$PSS(w)$	$G(w)$	$G(w-1)$	$G(max)^*$	$CUSUM$
3	2	5	5	.890	.890	.935	.921	.900
3	3	5	0	.855	.856	.794	.828	.836
3	3	4	1	.853	.854	.789	.831	.837
3	3	3	2	.848	.848	.800	.840	.837
3	3	2	3	.858	.859	.834	.867	.857
3	3	1	4	.868	.869	.880	.900	.877

Table 2: Power of detecting an excess of 15 expected cases over v intervals when $T = 52$, $\lambda_t = 2 + 0.06t + \delta + \theta(t - 21)$ for $20 \leq t \leq 20 + v - 1$.

* $G(max) = \max[G(w-2), G(w-1), G(w)]$.

Simulation Results

w	v	δ	θ	$PSS(w)$	$G(w)$	$G(w-1)$	$G(max)^*$	$CUSUM$
3	4	1.5	1.5	.754	.756	.713	.762	.796
3	5	1	1	.670	.670	.611	.641	.715
3	5	3	0	.595	.596	.526	.558	.673
5	3	5	0	.774	.795	.810	.852	.822
5	3	3	2	.787	.808	.827	.863	.837
5	4	1.5	1.5	.757	.761	.795	.805	.791
5	5	3	0	.724	.728	.691	.698	.686
5	5	2	0.5	.725	.728	.692	.711	.692
5	5	1	1	.733	.735	.743	.740	.726
5	5	0.5	1.25	.756	.757	.778	.775	.762

Simulation Results

- As expected $PSS(w)$ and $G(w)$ are best when $v = w$ and $\theta = 0$
- CUSUM is preferable when $v \neq w$ or when θ is large relative to δ (i.e. a ramp-like increase).
- When $v < w$, $G(\max)$ was always the best method.
  - However when $v > w$, $G(\max)$ has the poorest performance of any method.
  - It seems that the best strategy is to use $G(\max)$ and make sure we include the correct value of w in the search.
  - What is the breaking point of this strategy?
    (i.e. too much flexibility will result in loss of power.)

Future Directions for Scan Statistics

- Generalize the background rate, $\lambda(x, y, t)$, to be a random process, $\lambda(x, y, t, \omega)$.
  - Each year has a different flu season spatially and temporally.
  - Allow the $\lambda(x, y, t)$ to be a latent variable that looks similar from year to year, but is not the same.
- Generalize GLRT alternatives (w.r.t. the spatio-temporal framework)
  - Kulldorff assumes a rate increase inside of a cylinder in space-time.
  - How about an elliptical cylinder or an ellipsoid alternative?