

# A RISK-ADJUSTED SCAN STATISTIC FOR NON-HOMOGENEOUS DICHOTOMOUS EVENTS

NSF-CMMI 032856, NSF-DMI 032856

Aysun Taseli, James Benneyan PhD

Quality and Productivity Lab, Mechanical and Industrial Engineering, Northeastern University, Boston, MA

## BACKGROUND

### Problem Definition:

Many processes produce binary events that do not have identical failure probabilities, unlike common i.i.d. assumption of many statistical methods (e.g., binomial distribution). Examples:

- Disease occurrence based on age, gender, etc.
- Cancer rates based on at-risk group
- Drug abuse based on socio-economic group

Random variable of interest is the total  $T$  or fraction  $F$  of all events (e.g., patients or field goal attempts) resulting in the outcome of interest (e.g., mortality or scoring percentages):

$$T = X_1 + X_2 + \dots + X_J, \text{ where } X_i \sim \text{Bin}(n_i, p_i) \text{ and } p_i \neq p_j \forall (i, j) \rightarrow f_T(t) = ?$$

### Probability Distribution of $T$ and $F$ :

What is PDF of  $T$ ?

Convolution of independent non-identical Binomial random variables, can be significantly non-binomial and non-normal.

$$P(T=t) = P(F=f) = \sum_{x_1=0}^{\min(t, n_1)} P(X_1=x_1) \left[ \sum_{x_2=0}^{\min(t-x_1, n_2)} P(X_2=x_2) \left[ \sum_{x_3=0}^{\min(t-x_1-x_2, n_3)} P(X_3=x_3) \dots \right] \right] \quad (1)$$

where  $T_i = \sum_{k=1}^J x_{ik}$  and  $N_i = \sum_{k=1}^J n_{ik}, i=1, \dots, J; \hat{p}_k = \sum_{i=1}^J x_{ik} / \sum_{i=1}^J n_{ik}, k=1, \dots, J$

### Example:

$T_k$  = Total number of patients who developed cancer in the scanning window  $k$

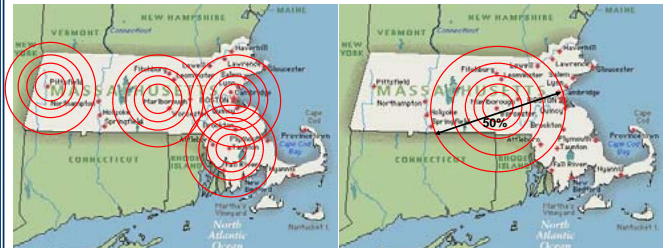
$X_{jk}$  = Number of cancer patients in risk category  $j$  in scanning window  $k$

$n_{jk}$  = Number of people under the risk of developing a cancer in risk category  $j$  in scanning window  $k$

$p_j$  = Risk of a cancer for patients in category  $j$

Scanning Window	Category 1 $p_1: 0.032764$ Low Risk		Category 2 $p_2: 0.081633$		Category 3 $p_3: 0.267606$		Category 4 $p_4: 0.464286$ High Risk		Total	
	$n_{1m}$	$x_{1m}$	$n_{2m}$	$x_{2m}$	$n_{3m}$	$x_{3m}$	$n_{4m}$	$x_{4m}$	$N_m$	$T_m$
1	41	2	38	2	17	5	1	0	97	9
2	60	5	34	0	17	6	1	1	112	12
3	50	1	46	3	15	5	2	2	113	11
...	...	...	...	...	...	...	...	...	...	...
17	2	0	4	0	4	2	0	0	10	2

### Kulldorf's Scan Statistic:



- Detects spatial clusters by scanning a window over a geographical space.
- Determines the window that maximizes the likelihood ratio based on a set of hypotheses stating that there is not and there is cluster over the scanned region, respectively.
- The p-values for the likelihood ratios are calculated based on Monte Carlo estimation and the region that maximizes the likelihood ratio is determined as the first likely cluster.

## OBJECTIVE

- Derive and investigate a scan statistic that can model the non-homogeneity in the population when the number of cases can be described by convolution of non-identical binomial random variates.
- Compare performance of new models to conventional approaches.

## METHODOLOGY

### Kulldorf's Bernoulli Model:

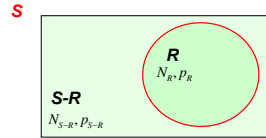
Assumes that the risks  $p_R$  and  $p_{S-R}$  are the same for all individuals inside and outside the region  $R$ , respectively.

$$H_0: p = p_S$$

$$H_1: p = p_R \forall s_i \in R$$

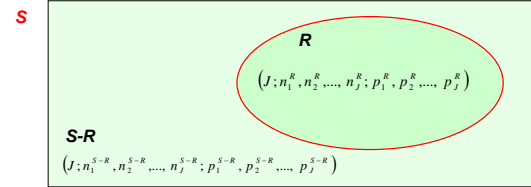
$$p = p_{S-R} \forall s_i \in S-R$$

$$s.t. p_R > p_{S-R}$$



$$LR = \frac{P_R^{M(R)} (1-P_R)^{N(R)-M(R)} P_{S-R}^{M(S-R)} (1-P_{S-R})^{N(S-R)-M(S-R)}}{P_S^{M(S)} (1-P_S)^{N(S)-M(S)}} \text{ , if } p_R > p_{S-R}; \text{ 1, otherwise.}$$

### A Risk Adjusted Scan Statistic:



$$H_0: T_R: T_{S-R} \sim \text{JB}(J, n^S, p^S); (J; n_1^S, n_2^S, \dots, n_J^S; p_1^S, p_2^S, \dots, p_J^S) \text{ through the study area } S.$$

$$H_1: T_R \sim \text{JB}(J, n^R, p^R); (J; n_1^R, n_2^R, \dots, n_J^R; p_1^R, p_2^R, \dots, p_J^R) \\ T_{S-R} \sim \text{JB}(J, n^{S-R}, p^{S-R}); (J; n_1^{S-R}, n_2^{S-R}, \dots, n_J^{S-R}; p_1^{S-R}, p_2^{S-R}, \dots, p_J^{S-R}), \text{ where } p_i^R > p_i^{S-R}, \forall i = 1, 2, \dots, J$$

When we know the individual cases ( $X_j$ 's):

$$LR = \frac{\prod_{i=1}^J (p_i^R)^{m_i^R} (1-p_i^R)^{n_i^R - m_i^R} (p_i^{S-R})^{m_i^{S-R}} (1-p_i^{S-R})^{n_i^{S-R} - m_i^{S-R}}}{\prod_{i=1}^J (p_i^S)^{m_i^S} (1-p_i^S)^{n_i^S - m_i^S}} \text{ , if } p_i^R > p_i^{S-R}; \text{ 1, otherwise.}$$

When we only know total cases (T):

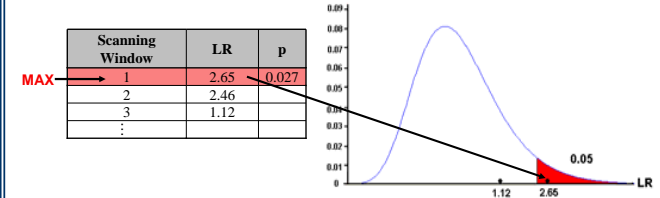
$$LR = \frac{P_R(T=M(R)) P_{S-R}(T=M(S-R))}{P_S(T=M(S))} \text{ , if } p_1^R > p_1^{S-R}; \text{ and 1, otherwise.}$$

$$P_R(T=M(R)) \sim \text{JB}(J, n^R, p^R), P_{S-R}(T=M(S-R)) \sim \text{JB}(J, n^{S-R}, p^{S-R}),$$

$$P_S(T=M(S)) \sim \text{JB}(J, n^S, p^S), P_S(T=M(S-R)) \sim \text{JB}(J, n^{S-R}, p^{S-R})$$

### Statistical Significance:

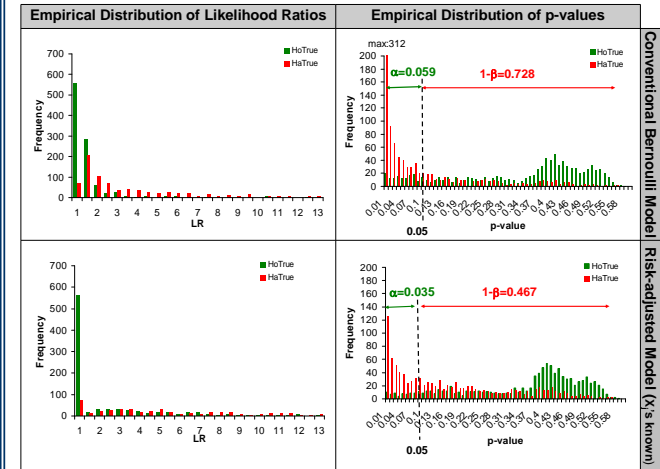
The significance of the likelihood ratio is determined by Monte Carlo Estimation, also called Randomization Testing where data are generated given the total number of cases in region  $S$  by hypergeometric randomization.



## PRELIMINARY RESULTS AND DISCUSSION

- The following results are obtained for one likelihood ratio for one location versus the maximum of likelihood ratios when the scanning feature is incorporated.
- The risk adjusted approach produces smaller type I error probabilities than the conventional Bernoulli model, but also less power.
- These counter intuitive results need further investigation.
- As sample sizes increase, type I error probabilities become closer to the desired value and power increases, as expected.
- Next step in the research will be incorporating the scan feature to the risk-adjusted model.

$$J=4; n_1^{S-R}=30, n_1^R=50; p^{S-R}:(0.25, 0.05, 0.15, 0.35); H_0: p^R=p^{S-R}; H_1: p^R=1.5 \cdot p^{S-R}$$



### References

- Benneyan JC, Borgman DA, "A Useful J-Binomial Type Distribution for Non-homogeneous Dichotomous Events", *Industrial Engineering Research Conference Proceedings*, 1-6, 2004.
- M. Kulldorf, "A Spatial Scan Statistics," *Communications in Statistics*, vol. 26, pp. 1481-1496, 1997.