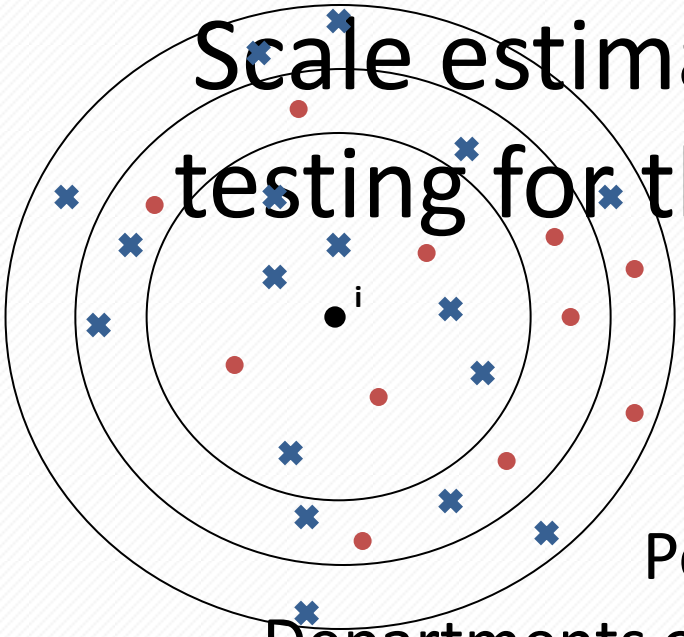
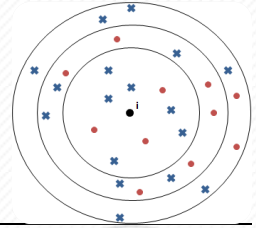


Scale estimation and significance testing for three focused statistics

Peter A. Rogerson
Departments of Geography and Biostatistics
University at Buffalo
Buffalo, NY

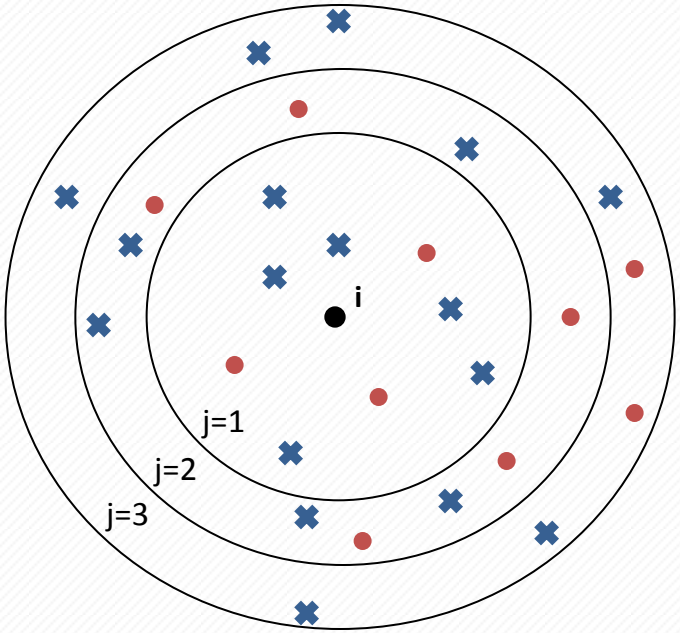
Introduction



- Focused statistical tests examine a H_o of no raised incidence in a variable of interest, around a prespecified site of interest
- Problem: Tests require specification of a definition for the surrounding neighborhood – preferably matching the risk associated with the H_a
- Possible Resolution: Test several possible neighborhood definitions around the site of interest
- Problem: Multiple testing raises the probability of rejecting a true H_o by chance alone, complicating significance testing
- **Purpose:** Describe several existing methods that may be used to carry out exact significance testing for three focused statistics in the presence of multiple testing

Introduction

Conceptual Illustration of Focused Scan Statistics



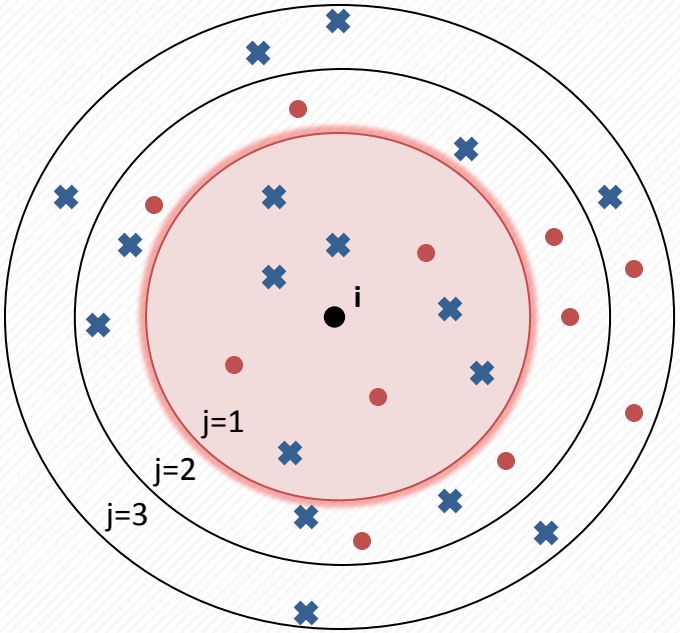
Subregion (j)	Cases	Controls	Total
1	3	6	9
2	5	5	10
3	2	6	8
Total	10	17	27

- Focal Site of Interest
- Case (10)
- Control (17)

Introduction

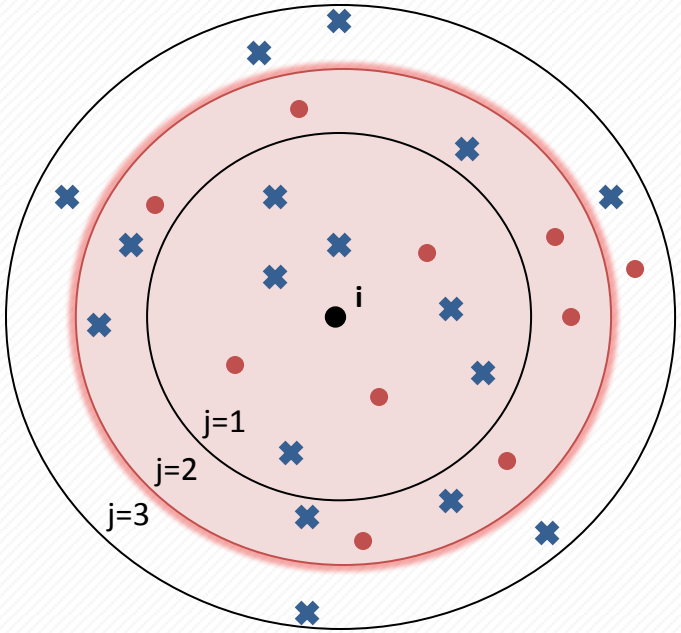
Conceptual Illustration of Focused Scan Statistics

(a) Changepoint k=1



Subregion (j)	Case (c)	Controls	Total
1	3	6	9
2	5	5	10
3	2	6	8
Total	10	17	27

(b) Changepoint k=2

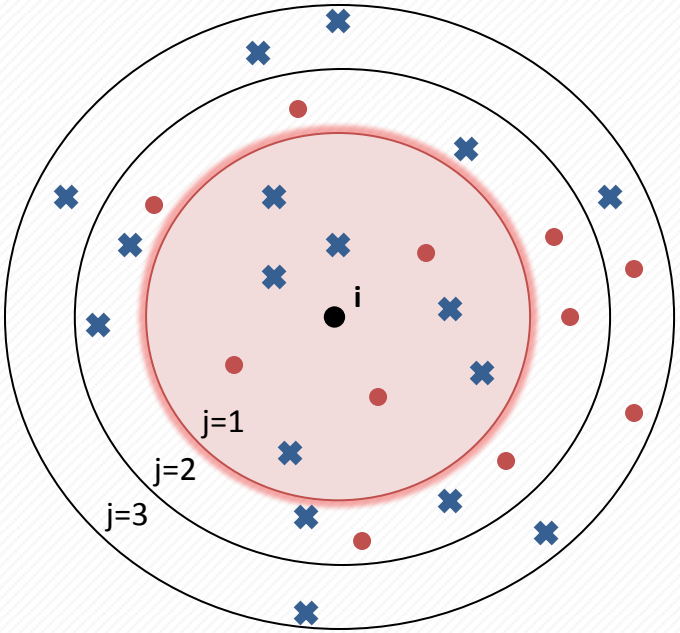


Subregion (j)	Case (c)	Controls	Total
1	3	6	9
2	5	5	10
3	2	6	8
Total	10	17	27

Introduction

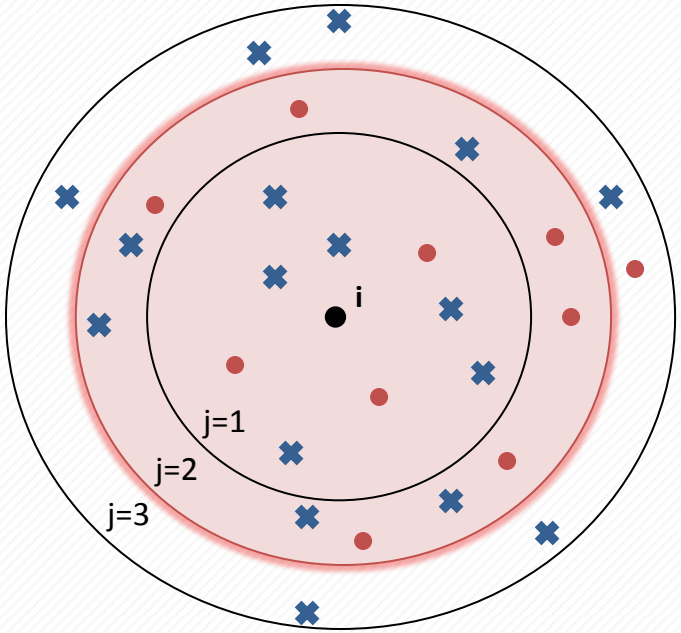
Conceptual Illustration of Focused Scan Statistics

(a) Changepoint k=1



Zone	Cases	Controls	Total
z_{ij}	3	6	9
$\sim z_{ij}$	7	11	18

(b) Changepoint k=2



Zone	Cases	Controls	Total
z_{ij}	8	11	19
$\sim z_{ij}$	2	6	8

Introduction

Local Statistic	Statistic	Data Type	Significance Approximation
Local Spatial Scan	Maximum Likelihood	Case-Control Obs. & Exp.	Worsley (1983)
Maximum χ^2	Chi-Squared	Case-Control	Boulesteix (2006)
Difference Between Obs. and Expected	K-S Statistic	Observed-Expected	Conover (1972)

- Step 1: How to use theses three focused statistics when progressively adding zones around a site of interest
- Step 2: How to assess statistical significance of those tests when the procedure creates the issue of multiple testing

Focused Statistic 1

LOCAL SPATIAL SCAN STATISTIC

Local Spatial Scan Statistic

Step 1: Implementing the Statistic with Multiple Neighborhoods

$$\Lambda = \max_{\{j=1,\ldots,m\}} L_j = \ln(L_{1i}(z_{ij}) / L_0)$$

z_{ij}: Neighborhood around site *i* consisting of the first *j* subregions

m: Number of potential neighborhood definitions

L_0: Likelihood of data under null hypothesis, when the probability that an observation is a case is the same for all zones

L_{1j}: Likelihood of data under alternative hypothesis, when the probability of being a case is higher inside z_{ij} than outside

Purpose: Determine the "best" neighborhood – the one associated with the value of *j* that maximizes the log likelihood ratio

Local Spatial Scan Statistic

Step 1: Implementing the Statistic with Multiple Neighborhoods

$$L_0 = p_0^{C}(1-p_0)^{N-C}$$

p_0: overall probability an observation is a case

C: total number of cases

N: Total number of observations

$$L_1(z) = p^{C_z}(1-p)^{N_z-C_z}\, q^{C-C_z}(1-q)^{(N-N_z)-(C-C_z)}$$

p: probability observation is a case inside zone z

q: probability observation is a case outside zone z

C_z: total number of cases inside zone z

N_z: total number of observations in zone z

$$\Lambda = \max_{\{j=1\ldots m\}} L_j = \ln(L_{1i}(z_{ij}) / L_0)$$

m: total number of possible zone definitions

(a) Changepoint k=1

Zone	Cases	Noncases	Total
z_{ij}	3	6	9
$\sim z_{ij}$	7	11	18
Total	10	17	27

p=3/9=0.333 and q=7/18=0.389

$L_{j=1}$=0.08

(a) Changepoint k=2

Zone	Cases	Noncases	Total
z_{ij}	8	11	19
$\sim z_{ij}$	2	6	8
Total	10	17	27

p=8/19=0.421 and q=2/8=0.25

$L_{j=2}$=0.733

$\Lambda=\max(L_{j=1}, L_{j=2})=(0.08, 0.733)=0.733$

Local Spatial Scan Statistic

Step 2: Assessing Statistical Significance of the "Best" Neighborhood

- Problem: Multiple testing that occurs when we examine *m* neighborhoods implies that using the usual chi-square distribution to assess the null hypothesis (for a single neighborhood) is inappropriate
- Solution: When examining one focal region we can adapt the approach of Worsley (1983) to estimate the null distribution, conditional upon C
- Calculation: Probability is derived through an iterative calculation of:

$$pr(\Lambda < x) = F_c(C)$$

Purpose: Determine the statistical significance associated with the "best" neighborhood in the presence of multiple testing

Local Spatial Scan Statistic

Step 2: Assessing Statistical Significance of the "Best" Neighborhood

$$F_{k+1}(v) = \sum_{u=a_k}^{b_k} F_k(u) h_k(u,v); \quad a_{k+1} \le v \le b_{k+1}$$

$$F_1(v) = 1; \quad a_1 \le v \le b_1$$

k: changepoint

a_k: smallest value of C_z where $L_j < x$

b_k: largest value of C_z where $L_j < x$

$$h_k(u,v) = \frac{\binom{N_k}{u}\binom{n_{k+1}}{v-u}}{\binom{N_{k+1}}{v}}$$

N_k : Total observations within changepoint k

n_{k+1} : Total observations in zone k+1

N_{k+1} : Total observations within changepoint k+1

k=0

$a_1=3 \le v \le b_1=4$

$F_1(3)=1$

$F_1(4)=1$

k=1

$a_2=7 \le v \le b_2=7$

$F_2(7)= F_1(3)h(3,7)+F_1(4)h(4,7)$

$F_2(7)= (1)(0.35)+(1)(0.3)=0.65$

$$h(3,7) = \frac{\binom{9}{3}\binom{10}{4}}{\binom{19}{7}} = 0.35; \quad h(4,7) = \frac{\binom{9}{4}\binom{10}{3}}{\binom{19}{7}} = 0.3$$

k=2

$a_3=10 \le v \le b_3=10$

$F_3(10)= F_2(7)h(7,10)$

$F_3(10)= 0.65(0.3345)=0.2174$

$$h(7,10) = \frac{\binom{19}{7}\binom{8}{3}}{\binom{27}{10}} = .3345$$

$pr(\Lambda<0.733)=0.2174$

$pr(\Lambda>0.733)=1-0.2174=0.7826$

Local Spatial Scan Statistic

Step 2: Assessing Statistical Significance of the "Best" Neighborhood

- Expansion to problems with larger numbers: For larger numbers of observations and cases the probabilities associated with *h* can be derived using:

$$h_k(0, v+1) = \frac{n_{k+1} - v}{N_{k+1} - v} h_k(0, v)$$

$$h_k(u+1, v) = \frac{(v-u)(N_k - u)}{(u+1)(n_{k+1} - v + u + 1)} h_k(u, v)$$

Purpose: Determine the statistical significance associated with the "best" neighborhood in the presence of multiple testing

Focused Statistic 2

MAXIMUM LOCAL CHI-SQUARE STATISTIC

Maximum Local Chi-Square Statistic

Introduction

$$\text{pr}\left(\max_{k=1,2,\ldots K-1} \chi_k^2 > d\right)$$

χ^2_k: Chi-square goodness-of-fit statistic for changepoint *k*

k: Zones arranged according to increasing distance from focal point

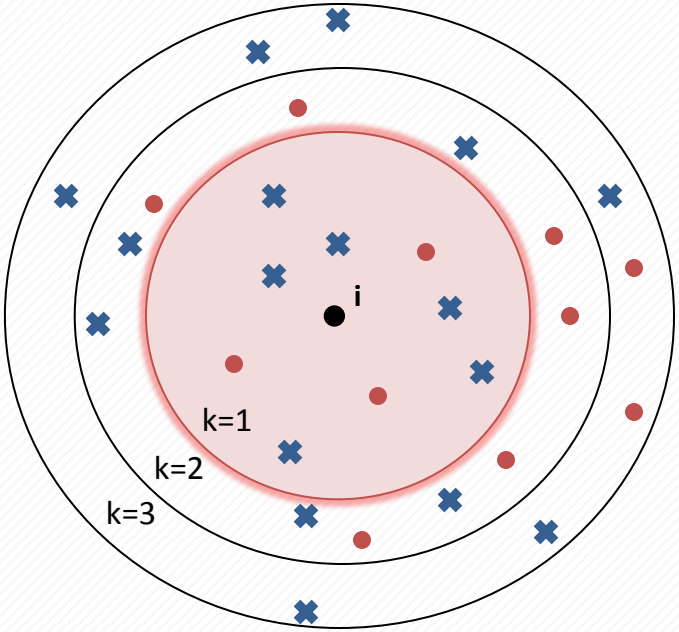
d: Observed maximal chi-square statistic

- Miller and Siegmund (1982) present the asymptotic null distribution of maximal chi-square across possible changepoints
- Koziol (1991) derives the exact null distribution for the small sample case
- Boulesteix (2006) notes that the Koziol approach is only appropriate when the variable takes on as many values as there are observations

Purpose: Determine the "best" neighborhood – the one associated with the value of *k* that maximizes the chi-square statistic

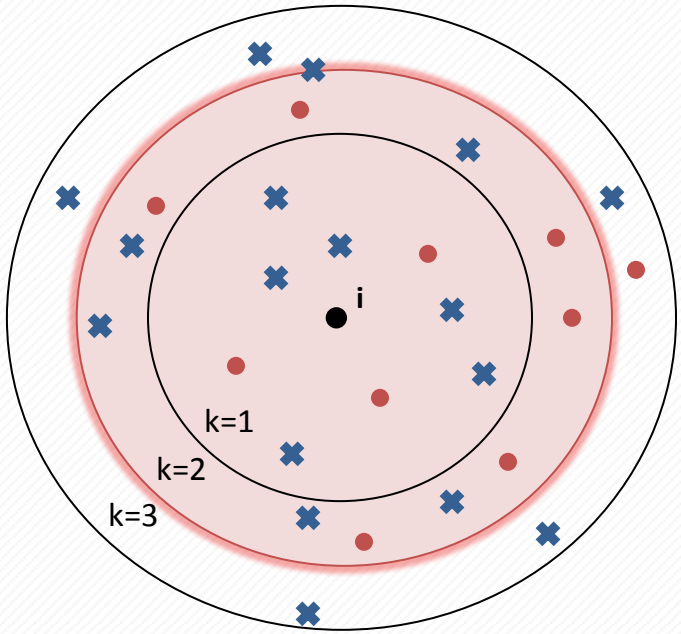
Maximum Local Chi-Square Statistic

Step 1: Implementing the Statistic with Multiple Neighborhoods



Zone	Case (c)	Controls	Total
z_{ij}	3	6	9
$\sim z_{ij}$	7	11	18
Total	10	17	27

$X^2_{k=1} = 0.0794$



Zone	Case (c)	Controls	Total
z_{ij}	8	11	19
$\sim z_{ij}$	2	6	8
Total	10	17	27

$\chi^2_{k=2} = 0.706$

$$\max(X^2_{k=1}; X^2_{k=2}) = (0.0794; 0.706) = 0.706$$

Maximum Local Chi-Square Statistic

Step 2: Assessing Statistical Significance of the "Best" Neighborhood

$$lower_d(x) = \frac{N_2 x}{N} - \frac{N_1 N_2 \sqrt{d}}{N}\sqrt{\frac{x}{N}(1-\frac{x}{N})(\frac{1}{N_1}+\frac{1}{N_2})}$$

$$upper_d(x) = \frac{N_2 x}{N} + \frac{N_1 N_2 \sqrt{d}}{N}\sqrt{\frac{x}{N}(1-\frac{x}{N})(\frac{1}{N_1}+\frac{1}{N_2})}$$

N: Number of observations

N_1: Number of cases

N_2: Number of controls

$i=a_k=x$: Total number observations prior to changepoint k

$$\max(0\ ,\ i\text{-}N_1) \le j < lower_d(i)$$

$$upper_d(x) < j \le \min(N_2\ ,\ i)$$

N_1: Number of cases

N_2: Number of controls

$i=a_k=x$: Number observations prior to changepoint k

Form coordinate pairs (i,j), for all i

$\chi^2_{k=1}=0.0794$, k=1

$x = i = 9$ $N_1=10$

N=27 $N_2=17$

$lower_d(x)=4.68$ $upper_d(x)=6.66$

$\max(0,9-10=-1) \le j < 4.68$ → **j=(0,1,2,3,4)**

$6.66 < j \le \min(17,9)$ → **j=(7,8,9)**

$\chi^2_{k=2}=0.706$, k=2

$x = i = 19$ $N_1=10$

N=27 $N_2=17$

$Lower_d(x)=11$ $upper_d(x)=12.92$

$\max(0,19-10=9) \le j < 11$ → **j=(9,10)**

$12.92 < j \le \min(17,19)$ → **j=(13,14,15,16,17)**

Maximum Local Chi-Square Statistic

Step 2: Assessing Statistical Significance of the "Best" Neighborhood

$$b_1 = \binom{i_1}{j_1}; \quad b_s = \binom{i_s}{j_s} - \sum_{r=1}^{s-1} \binom{i_s - i_r}{j_s - j_r} b_r; \quad s = 2,\ldots,q.$$

B: coordinate pairs $\{i_s, j_s\}$ s=1,2,…q: order as increasing j within each value of i

$$pr(\chi^2_{\max} > d) = \binom{N}{N_2}^{-1} \sum_{s=1}^{q} \binom{N - i_s}{N_2 - j_s} b_s$$

N: Number of observations

N_2: Number of controls

Use coordinate pairs (i_s, j_s) to find $\text{pr}(\chi^2_{\max} > d)$

	s	*i*	*j*	*b*
k=1				
	1	9	0	1
	2	9	1	9
	3	9	2	36
	4	9	3	84
	5	9	4	126
	6	9	7	36
	7	9	8	9
	8	9	9	1
k=2				
	9	19	9	36,540
	10	19	10	49,392
	11	19	13	15,750
	12	19	14	5,040
	13	19	15	966
	14	19	16	84
	15	19	17	0

$\text{pr}(\chi^2_{\max} > 0.706) = 0.6244$

Focused Statistic 3

DISTRIBUTION OF THE MAXIMUM DIFFERENCE BETWEEN OBSERVED AND EXPECTED PROPORTIONS

Maximum Difference: Observed and Expected

Introduction

$$\text{pr}\,(D^{+} \geq d^{+})$$

D^{+}: One-sided Kolmogorov-Smirnov statistic

d^{+}: Observed value of the Kolmogorov-Smirnov statistic

- Stone (1988): the maximal observed/expected ratio among cumulative observed and expected values
- Weakness: large ratios can result from small expected values
- Conover (1972): proposed a recursive approach to estimating the statistical significance associated with the observed statistic

Purpose: Determine the "best" neighborhood – the one associated with

Maximum Difference: Observed and Expected

Step 2: Assessing Statistical Significance of the "Best" Neighborhood

$$D^+ = \frac{1}{n} \max_{k=1,\ldots,K-1} \{\sum_{i=1}^{k} O_i - \sum_{i=1}^{k} E_i\}$$

n: Total number of observations

O_i: Observed count in zone i

E_i: Expected count in zone i

k: Number of zones considered

$$H_0 = 0 \qquad H_k = \sum_{i=1}^{k} E_i / n$$

n: Total number of observations

E_i: Expected count in zone i

$$f_j = 1 - d^+ - j/N$$

$$0 \le j \le n(1-d^+)$$

d^+: observed value

- Take value f_i if f_i equals any of the H_k
- Take value as max (H_k)< f_i), if $f_i \neq$ any H_k

Find values of f_j

			Cumul. Prop.			
Zone (k)	Obs (O_i)	Exp (E_i)	Obs	Exp	Diff	Cum Ratio
1	7	3	7/9	3/9	4/9	2.133
2	1	2	8/9	5/9	3/9	1.6
3	1	4	9/9	9/9	0	1.0
Total (n)	9	9				

D^+ = 1/9(max{4,3,0}) = 4/9

H_0=0, H_1=3/9, H_2=5/9, H_3=1

$0 \le j \le 9(1-(4/9))$ → j = (0,1,2,3,4,5)

f_0 = 1-(4/9)-(0/9)=5/9 → 5/9

f_1 = 1-(4/9)-(1/9)=4/9 → 3/9

f_2 = 1-(4/9)-(2/9)=3/9 → 3/9

f_3 = 1-(4/9)-(3/9)=2/9 → 0

f_4 = 1-(4/9)-(4/9)=1/9 → 0

f_5 = 1-(4/9)-(5/9)=0/9 → 0

Maximum Difference: Observed and Expected

Step 2: Assessing Statistical Significance of the "Best" Neighborhood

$$e_0=1$$

$$e_k = 1-\sum_{j=0}^{k-1}\binom{k}{j} f_j^{k-j} e_j \qquad k \geq 1$$

k: Number of zones considered

- Continue recursion for all k where $f_j>0$

$$p(D^+ \geq d^+) = \sum_{j=0}^{[n(1-d^+)]}\binom{n}{j} f_j^{n-j} e_j$$

d^+: observed value

n: Total number of observations

f_0= 1-(4/9)-(0/9)=5/9 → 5/9 f_3= 1-(4/9)-(3/9)=2/9 → 0

f_1= 1-(4/9)-(1/9)=4/9 → 3/9 f_4= 1-(4/9)-(4/9)=1/9 → 0

f_2= 1-(4/9)-(2/9)=3/9 → 3/9 f_5= 1-(4/9)-(5/9)=0/9 → 0

$e_0=1$

$e_1= \binom{1}{0} f_0^1 e_0 = 1 - f_0 = 4/9,$

$e_2= 1-\{\binom{2}{0} f_0^2 e_0 + \binom{2}{1} f_1^1 e_1\} = 1-\frac{25}{81}-\frac{24}{81}=\frac{32}{81}.$

$$pr(D^+ \geq 4/9) = \binom{9}{0} f_0^9 e_0 + \binom{9}{1} f_1^8 e_1 + \binom{9}{2} f_2^7 e_2 = 0.01215.$$

$$pr(D^+ \geq d^+ = 4/9) = 0.01215$$

Maximum Difference: Observed and Expected

Comments

Zone 1	0	0	0	0	…	8	8	9
Zone 2	0	1	2	3		0	1	0
Zone 3	9	8	7	6		1	0	0

- 55 possible distributions of observed counts

$$pr(x_1, x_2, x_3) = \frac{n!}{x_1!x_2!x_3!} {p_1}^{x_1} {p_2}^{x_2} {p_3}^{x_3}$$

p_i: probability of observation being in region i

- 13 of those 55 outcomes have a value of $D^+ > 4/9$
- Adding the probabilities associated with those 13 outcomes yields a p-value of 0.01215

Summary

Three focused statistics are considered here, and the primary issue addressed is how to control for multiple testing in examining several spatial scales.

The exact tests involve iteration and combinatorics. Although they are therefore easiest to implement for small numbers of cases, it is also possible to use known relationships (e.g., for the hypergeometric distribution) to apply them in other instances as well.

Next steps: to illustrate their use with the well-known dataset on leukemia in central New York State, which also contains data on the locations of eleven potential “focal” sites.

Acknowledgements: The support of an NSF Grant and the assistance of Peter Kedron in producing the graphics are gratefully acknowledged.