비모수적 방법을 이용한 공간검색통계량

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

조 호 진

비모수적 방법을 이용한 공간검색통계량

지도 정 인 경 교수

이 논문을 석사 학위논문으로 제출함

2014년 12월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

조 호 진

조호진의 석사 학위논문을 인준함

심사위원 정 인 경 

심사위원 남 정 모 

심사위원 송 기 준 

연세대학교 대학원

2014년 12월 일

감사의 글

떨리는 마음으로 처음 학교에 오던 날이 엊그제 같은데 어느덧 2년이 지나 석사과정을 마치고 졸업을 앞두게 되었습니다. 지난 대학원에서의 시간은 학문적 성장과 더불어 다양한 경험을 쌓을 수 있는 소중한 시간이었습니다. 많이 부족했던 제가 눈부신 발전을 해 논문을 무사히 쓸 수 있도록 도움을 주신 많은 분들에게 감사의 말을 전하고 싶습니다.

먼저 아낌없는 조언과 애정으로 논문을 지도해주시고 늘 의지할 수 있도록 많은 관심을 쏟아주신 든든한 버팀목 정인경 교수님, 항상 열정적이시고 유쾌한 모습으로 본보기가 되어 주시고 여러 경험을 할 수 있도록 용기를 북돋아 주신 남정모 교수님, 여러 방면으로 풍부한 지식을 채워주시고 사랑으로 세심하고 꼼꼼하게 지도해 주신 송기준 교수님께 진심으로 감사드립니다. 또한 항상 밝고 에너지 넘치는 모습으로 미숙했던 저를 정성으로 가르쳐주신 박소희 교수님께도 감사드립니다. 교수님들 덕분에 제가 이렇게까지 성장할 수 있었습니다.

다방면으로 챙겨주시고 철없었던 저를 관심과 사랑으로 지켜봐주시고 이끌어 주신 선배님들과 많이 서툴렀던 저를 이해해주고 따라준 후배들, 기쁜 일, 슬픈 일들을 모두 함께 나눈 동기 우현이에게도 감사의 마음을 전합니다. 많은 시간을 함께 보내며 좋은 추억을 쌓을 수 있어서 즐거웠습니다.

항상 따뜻한 말과 위로로 힘을 실어 준 아영이, 힘든 일, 즐거운 일 있을 때마다 함께 시간 보내주고 신나게 해준 경희, 도경이, 늘 응원해주고 내편에 서줘서 큰 힘이 되어준 성식이, 은미언니, 대현오빠, 정호오빠, 환희, 각자의 대학원생활을 하며 서로 의지할 수 있었던 소래, 애경이, 윤지, 그 외의 언급하지 못한 소중한 친구들에게도 너무 사랑하고 항상 고맙다는 말을 하고 싶습니다.

마지막으로 할 수 있다는 자신감을 가질 수 있도록 용기를 주시고 물심양면으로 지원해 주시며, 큰 사랑으로 키워주신 부모님, 항상 격려해 주시고 다독여 주시는 할머니, 언제 어디서나 삶의 활력소가 되어준 내 분신 호성이. 저를 믿어주는 가족들에게 사랑과 감사의 마음을 전합니다. 앞으로 겸손하고 노력하는 자세로 더 성장해 나가는 모습 보여드리겠습니다. 진심으로 감사합니다.

2014년 12월

조 호 진 올림

차 례

그 림 차 례

표 차 례

국 문 요 약

비모수적 방법을 이용한 공간검색통계량

지리학적 공간 역학이나 질병 감시 연구에서 공간 검색 통계 방법은 공간 데이터에서 결과 변수의 비율이 현저하게 높거나 낮은 특정 군집을 찾아내고 통계학적 유의성을 평가한다. 기존의 연구 방법들은 자료가 어떠한 분포 가정을 따른다고 생각하는 모수적 방법이다. 하지만 자료가 해당 분포를 따르지 않고 비대칭적 분포나 꼬리가 두꺼운 분포를 따르는 경우에는 분포 가정이 적합하지 않으므로 기존의 방법들을 적용시키기 어렵다. 그러므로 기존 Kulldorff, Huang, Konty (2009)가 제안한 normal 방법을 기반으로 분포 가정이 필요 없는 비모수적 방법을 적용하여 scanning window 안과 밖을 비교할 때 우도비 검정 대신 윌콕슨 순위합 검정을 사용한 방법을 제안한다.

자료가 비대칭적이거나 꼬리가 두꺼운 분포를 따를 때 기존에 제시된 대부분의 방법과 같이 모수 통계분석 방법을 사용하는 경우 검정력과 정확도가 떨어지는 문제점이 발생할 수 있다. 기존 normal 방법에서 정규분포를 따르지 않는 경우 검정력과 정확도를 확인해 본 바가 없으므로 본 논문에서 모의실험을 통해 여러 가지 분포 하에서 normal 방법과 비모수적 방법의 검정력, 정확도를 비교한다. 또한 우리나라 여성 유방암 사망률 자료에 비모수적 방법을 적용하여 기존 방법들을 적용한 결과와 비교해 본다.

모의실험 결과를 통해 자료가 비대칭적이거나 꼬리가 두꺼운 분포를 따를 경우에 normal 방법의 검정력과 정확도는 떨어지는 반면 비모수적 방법은 높게 유지되는 것을 확인하였다. 실제 자료 분석에서도 normal 방법과 weighted normal 방법으로 찾은 군집은 포아송 방법을 이용해서 찾아낸 군집과 일치하지 않고 유의성에도 차이를 보였지만 비모수적 방법은 비슷한 군집을 찾았고 유의성 또한 일치하였다.

이를 통해 연속적 자료의 분포가 비대칭적이거나 꼬리가 두꺼운 경우에는 비모수적 방법을 사용하는 것이 검정력과 정확도를 높일 수 있는 방법이 될 것으로 기대된다.

핵심되는 말 : 공간검색통계량, 비모수적 방법, 윌콕슨 순위합 검정

제1장 서론

1.1 연구 배경 및 목적

지리학적 공간 역학이나 질병 감시 연구에서 공간 검색 통계 방법(spatial scan statistic)은 공간 데이터에서 결과 변수의 비율이 현저하게 높거나 낮은 특정 군집(spatial cluster)을 찾아내고, 찾아낸 군집의 통계학적 유의성을 평가한다.

기존에 가장 널리 쓰이는 연구 방법은 인구 집단의 자료와 개인의 자료를 이용하여 포아송 분포나 베르누이 분포를 가정해 군집을 찾아내는 방법이다(Kulldorff 1997). 주로 가산 자료인 질병의 유병률이나 발생률, 사망률 등이 높거나 낮은 군집을 찾아낸다. 또한 연속적 자료가 개인이나 지역 단위로 주어졌을 경우에는 결과 변수의 평균이 크거나 작은 군집을 찾기 위해 정규 분포를 이용한다(Kulldorff, Huang, and Konty 2009). 신생아의 태어날 때의 몸무게, 혈중 납 농도 등이 높거나 낮은 군집을 찾는 경우 등에 사용한다. 이 방법에 불확실성을 고려해 가중치를 주는 weighted normal 방법 또한 제시되어 있다(Huang et al. 2009). 이외에도 자료의 형태가 순서형일 때 사용하는 방법(Jung, Kulldorff, and Klassen 2007), 다항자료에서 사용하는 방법 등이 제시되어 있다(Jung, Kulldorff, and Richard 2010).

상기 연구 방법들은 모두 분포 가정을 하는 모수적인 방법이다. 하지만 자료가 해당 분포를 따르지 않고 비대칭적 분포나 꼬리가 두꺼운 분포를 따르는 경우에는 분포 가정이 적합하지 않으므로 위의 방법들을 적용시키기 어렵다. 그러므로 분포 가정이 필요 없는 비모수적 방법을 새롭게 제안한다.

1.2 연구 내용 및 방법

Kulldorff, Huang, and Konty (2009)가 제안한 연속적 자료에서 정규 분포를 이용한 방법을 기반으로 비모수적 방법을 적용하고자 한다. 기존의 방법은 특정군집을 찾아내기 위해 scanning window를 넓혀가며 scanning window 안과 밖을 비교하는 우

도비 검정(likelihood ratio test)을 이용한다. 그러나 자료가 해당 분포를 따르지 않는 경우에는 분포 가정을 하지 않는 비모수적 방법이 더 적합하므로 본 연구에서는 scanning window 안과 밖을 비교할 때 비모수적 방법인 윌콕슨 순위합 검정(Wilcoxon rank sum test)을 사용한 방법을 제안한다.

검정통계량은 윌콕슨 순위합 검정의 유의확률 중 최솟값으로 하며, 순열 검정법(permutation test)을 이용하여 통계학적 유의성을 평가한다. 본 연구에서는 기간을 고려하지 않은 2차원 공간 군집 자료를 이용했지만 시공간 자료의 경우로 확장할 수 있다. Scanning window의 경우 원형 window를 사용하였다.

모의실험을 통해 여러 가지 분포 하에서 기존의 방법과 비모수적 방법의 검정력(power), 민감도(sensitivity), 양성예측도(positive predicted value; PPV), extended power를 비교하여 각각 방법의 장·단점을 살펴본다. 또한 우리나라 암 사망률 자료에 비모수적 방법을 적용한다.

1.3 논문의 구성

제 1장에서는 연구의 배경 및 목적과 연구 내용 및 방법을 소개한다. 2장에서는 본 연구의 이론적 배경이 되는 공간 검색 통계 방법과 윌콕슨 순위합 검정 방법, 기존의 정규 분포를 이용해 군집을 찾아내는 방법에 대해 살펴보고, 이 논문을 통해 새롭게 제안하고자 하는 비모수적 방법을 사용한 공간검색통계량에 대해 소개한다. 3장에서는 모의실험을 통해 자료가 여러 가지 분포를 따르는 경우 기존 방법과 비모수적 방법을 비교·평가한다. 4장에서는 실제자료를 이용해서 비모수적 방법이 특정 군집을 잘 찾아내는지 확인하고 기존 방법을 적용한 결과와 비교한다. 마지막으로 5장에서는 결론 및 고찰을 제시한다.

제2장 이론적 배경

2.1 공간검색통계량 (spatial scan statistic)

Kulldorff(1997)가 제안한 공간검색통계량이란 공간 군집 분석(spatial cluster analysis)의 한 방법으로 천문학, 지리학, 역학 등 다양한 분야에서 통계적으로 유의한 공간 군집을 찾기 위해 쓰인다. 이 때, 공간 군집이란 지리적으로 가까운 지역들의 집합으로 결과 변수의 비율이 높거나 낮은 특정 지역들의 집단을 뜻한다. 인구 집단의 자료와 개인의 자료를 이용하여 포아송 분포나 베르누이 분포를 가정하여 분석하는 방법(Kulldorff, 1997), 정규 분포를 이용하는 방법(Kulldorff, Huang and Konty, 2009), 정규분포를 이용할 때 가중치를 주는 방법(Huang et al, 2009) 등 여러 가지 공간검색통계량이 제안되고 있다.

공간검색통계량은 scanning window 안과 밖의 결과 변수 비율이 같다는 귀무가설 하에서 우도비 검정을 통해 통계학적 검정을 한다. Scanning window란 각각의 지역의 중심점(centroid)을 기준으로 가까운 지역을 하나씩 추가로 포함해 나가는 과정을 반복할 때 생성되는 수많은 지역들의 집단을 뜻한다. 대부분의 경우 원형 scanning window를 사용하며 최대 전체 지역의 50%를 포함하는 경우로 maximum scanning window size를 제한한다. 이렇게 생성된 많은 scanning window의 안과 밖을 비교하는 우도비 검정통계량을 각각 계산하고, 그 중 최대가 되는 값이 나오는 scanning window에 속하는 지역을 most likely cluster라고 한다. 위의 과정을 식으로 나타내면 다음과 같다.

$$\lambda = \frac{\max_{Z, H_a} L(Z, \theta)}{\max_{Z, H_0} L(Z, \theta)} = \frac{\max_Z L(Z, \hat{\theta})}{L(\hat{\theta_0})}$$

각각의 scanning window 안에 속하는 지역을 Z라고 할 때 우도비가 최대가 되는 지역 Z가 most likely cluster이다. 몬테카를로(Monte Carlo) 가설 검정(Dwass, 1957)을 이용하여 찾아낸 지역 Z에 대한 통계적 유의성을 평가한다.

2.2 연속적 자료에서의 normal 방법

Kulldorff, Huang, and Konty (2009)가 제안한 정규분포 모형을 사용한 공간검색통계량은 연속적 자료가 정규분포를 따른다고 가정하여 scanning window 안과 밖의 결과변수의 평균의 차이가 있는지를 분석하는 방법이다. 자료가 체중, 키와 같은 몇몇의 연속적인 값 $x_i,\ i=1,\dots,N$로 이루어져 있을 경우 사용한다. 각각의 값은 공간 지역 $s, s=1,\dots,S$ 에서 관측된 값으로, 한 지역에서 여러 개의 값이 관측될 수 있기 때문에 $S \leq N$ 이다. 이 때, 각 지역 s에서 관측된 모든 값의 합을 $x_s(=\sum_{i \in s} x_i)$라고 하고 각 지역에서 관측된 관측치의 개수를 n_s, 모든 지역에서 관측된 값들의 합을 $X(=\sum_i x_i)$라 한다.

2.2.1 검정통계량

Scanning window를 넓혀가며 결과변수의 평균값이 높거나 낮은 지역을 찾을 경우 scanning window 안의 지역을 Z, scanning window 밖을 Z^c 으로 표현한다. 각각의 Z마다 우도비(likelihood ratio; LR(Z))나 로그 우도비(log likelihood ratio; LLR(Z))를 계산하여 모든 값 중 최대가 되는 LR 또는 LLR 값을 통계량으로 사용한다.

$$H_0 : \mu_Z = \mu_{Z^c} \ (\text{for } all\ Z) \ \ vs. \ \ H_a : \mu_Z \neq \mu_{Z^c} \ (\text{for } some\ Z)$$

귀무가설과 대립가설이 위와 같을 경우 검정통계량은 다음과 같다.

$$\max_Z \frac{L_Z}{L_0} = \max_Z \frac{\ln L_Z}{\ln L_0}$$

이 때, 우도비나 로그 우도비가 최대가 되는 Z를 most likely cluster라고 한다.

여기서 $\ln L_0$와 $\ln L_Z$는 각각 귀무가설 하에서의 로그 우도함수, Z에서의 로그 우도함수를 나타내며 아래와 같이 계산한다.

$$\ln L_0 = -N\ln(\sqrt{2\pi}) - N\ln(\sigma) - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2}$$
$$\ln L_Z = -N\ln(\sqrt{2\pi}) - N\ln(\sqrt{\sigma_Z^2}) - N/2$$

이 때, 귀무가설 하에서 결과변수의 평균과 분산은

$$\mu = X/N \ , \ \sigma^2 = \frac{\sum_i (\mu - x_i)^2}{N}$$

로 나타낼 수 있고, 대립가설 하에서 결과변수의 평균과 분산은

$$\mu_Z = x_Z/n_Z$$
$$\sigma_Z^2 = \frac{1}{N}\left(\sum_{i\in Z} x_i^2 - 2x_Z\mu_Z + n_Z\mu_Z^2 + \sum_{i\notin Z} x_i^2 - 2(X-x_Z)\lambda_Z + (N-n_Z)\lambda_Z^2\right)$$
$$(\lambda_Z = (X-x_Z)/(N-n_Z))$$

로 나타낼 수 있다. 여기서 λ_Z는 Z 밖의 지역 Z^c에서 관측된 값들의 평균을 뜻한다.

$$H_0 : \mu_Z = \mu_{Z^c} \quad vs. \quad H_a : \mu_Z > \mu_{Z^c} \quad , \quad H_0 : \mu_Z = \mu_{Z^c} \ , H_a : \mu_Z < \mu_{Z^c}$$

만약 대립가설이 위와 같이 Z 안의 값이 Z 밖보다 크거나 작다고 할 경우에 검정통계량은 각각 다음과 같다.

$$\max_Z \frac{\ln L_Z I(\mu_Z > \lambda_Z)}{\ln L_0} \quad , \quad \max_Z \frac{\ln L_Z I(\mu_Z < \lambda_Z)}{\ln L_0}$$

2.2.2 추론

로그 우도비가 최대가 되는 군집 Z를 찾았다면 그 군집이 통계적으로 유의한지를 평가해야 한다. 이 경우 정규분포에서 임의의 자료를 생성하지 않고 관측된 자료를 순열검정법을 사용하여 p-값을 계산하는 몬테카를로 기반 가설 검정을 하였다. 몬테카를로 기반 가설 검정이란 M개의 자료를 무작위로 생성했다고 할 때, most likely cluster의 p-값을 $R/(M+1)$로 계산하는 방법이다. 이 때, R은 모든 자료와 비교하였을 때 실제 자료의 로그 우도비의 순위이다. 대부분의 경우 p-값을 계산하기 위해 M은 999, 4999 등과 같이 '9'로 끝나는 수를 사용한다.

2.2.3 한계점

대부분의 통계분석을 하는 경우 분포를 가정하는 모수 통계분석 방법을 사용한다. 하지만 모집단이나 표본이 가정한 분포를 따르지 않고 비대칭적이거나 꼬리가 두꺼운 분포를 따르는 경우에 모수 통계분석 방법을 사용하면 검정력과 정확도가 떨어지는 문제점이 발생할 수 있다. 기존 normal 방법에서는 순열검정법을 사용하였기 때문에 정규분포를 따르지 않는 경우에도 반복수를 늘리면 유의수준 α는 유지되지만 검정력과 정확도의 경우 의미 있는 값이 나오는지 확인해 본 바가 없다. 그러므로 다른 분포에서도 검정력과 정확도가 유지되는지 확인할 필요가 있다.

또한 normal 방법의 경우 관심이 있는 결과가 개인의 수준이 아닌 지역 수준의 값일 때, 각 지역의 표본 수가 달라서 발생하는 불확실성 문제가 생길 수 있다. 모든 지역 안에서 관측된 표본 수가 같다면 normal 방법이 유용하지만 관측된 표본 수가 다를 경우 표본 수가 작거나 값의 변동이 심한 지역의 평균값을 대푯값으로 쓰기에는 어려움이 있다.

2.3 연속적 자료에서의 weighted normal 방법

Huang et al.(2009)이 제안한 가중치를 사용한 정규분포를 이용하는 방법은 기존 normal 방법에서 불확실성 문제를 해결하기 위해 제안된 방법이다. Normal 방법에서 각 지역의 표본 수가 다를 경우 발생하는 불확실성 문제를 해결하기 위해 지역 z에서 가중치 δ_z(불확실성의 역수)를 사용하여 관측된 값(w_z)을 보정한다. 이를 식으로 나타내면 다음과 같다.

$$w_z \,|\delta_z \sim N(\mu_Z, \frac{\sigma_G^2}{\delta_z}) \quad , when \;\; z \in Z$$

$$w_z \,|\delta_z \sim N(\mu_{Z^c}, \frac{\sigma_G^2}{\delta_z}) \quad , when \;\; z \in Z^c (= G - Z)$$

전체 지역을 G, 공간 군집 안을 Z, 공간 군집에 속하는 지역을 z라고 할 때 관측치 w_z의 분산에 δ_z의 가중치를 사용하여 불확실성을 보정하는 방법이다.

2.3.1 검정통계량

Scanning window 안의 지역을 Z, scanning window 밖을 Z^c 로 표현할 때, 우도함수나 로그우도함수가 최대가 되는 Z를 찾는다.

$$H_0 : \mu_Z = \mu_{Z^c} \;(\text{for}\, all\, Z) \quad vs. \quad H_a : \mu_Z \neq \mu_{Z^c} \;(\text{for}\, some\, Z)$$

귀무가설과 대립가설이 위와 같을 경우, 우도비 검정통계량은 다음과 같다.

$$\lambda = \frac{\max_{Z,\, \mu_Z \neq \mu_{Z^c},\, \sigma_G^2} L(\theta_Z, \theta_{Z^c})}{\max_{Z,\, \mu_Z = \mu_{Z^c},\, \sigma_G^2} L(\theta_Z, \theta_{Z^c})} = \frac{\max_{Z,\, \mu_Z \neq \mu_{Z^c},\, \sigma_G^2} L(\theta_Z, \theta_{Z^c})}{L_0} = \frac{L(\hat{Z})}{L_0}$$

여기서 $\theta_Z = (\mu_Z, \sigma_G^2)$, $\theta_{Z^c} = (\mu_{Z^c}, \sigma_G^2)$, $where\ Z \subset G$ 로 정의하고, 주어진 지역 Z에서의 우도함수는 다음과 같다.

$$L(\theta_Z, \theta_{Z^c}) = \prod_{z \in Z} L(\theta_Z) \prod_{z \in Z^c} L(\theta_{Z^c})$$
$$\propto \prod_{z \in Z} \frac{\sqrt{\delta_z}}{\sigma_G} exp\left(-\frac{\delta_z}{2\sigma_G^2}(w_z - \mu_Z)^2\right) \prod_{z \notin Z} \frac{\sqrt{\delta_z}}{\sigma_G} exp\left(-\frac{\delta_z}{2\sigma_G^2}(w_z - \mu_Z)^2\right)$$

이 때, $\mu_Z, \mu_{Z^c}, \sigma_G^2$의 최대우도추정량(maximum likelihood estimator)은 다음과 같다.

$$\hat{\mu}_Z = \frac{\sum_{z \in Z} \delta_z w_z}{\sum_{z \in Z} \delta_z} \quad , \ \hat{\mu}_{Z^c} = \frac{\sum_{z \in Z^c} \delta_z w_z}{\sum_{z \in Z^c} \delta_z} ,$$
$$\hat{\sigma}_G^2 = \frac{\sum_{z \in Z} \delta_z (w_z - \hat{\mu}_Z)^2 + \sum_{z \in Z^c} \delta_z (w_z - \hat{\mu}_{Z^c})^2}{n_G}$$

여기서 n_G는 전체 공간 G에 속해있는 총 지역 z의 개수이다.

$$H_0 : \mu_Z = \mu_{Z^c} = \mu_G \quad vs. \quad H_a : \mu_Z > \mu_{Z^c}$$
$$H_0 : \mu_Z = \mu_{Z^c} = \mu_G \quad vs. \quad H_a : \mu_Z < \mu_{Z^c}$$

만약 대립가설이 위와 같이 Z 안의 값이 Z 밖보다 크거나 작다고 할 경우의 검정통계량은 각각 다음과 같다.

$$\lambda = \frac{\max_{Z, \mu_Z > \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c}) I(\hat{\mu}_Z > \hat{\mu}_{Z^c})}{\max_{Z, \mu_Z = \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c})}$$

$$\lambda = \frac{\max_{Z, \mu_Z < \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c}) I(\hat{\mu}_Z < \hat{\mu}_{Z^c})}{\max_{Z, \mu_Z = \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c})}$$

이러한 경우에서 검정통계량의 우도비가 최대가 되는 Z를 most likely cluster라고 한다. 즉, 찾고자 하는 발생률이 가장 높거나 낮은 군집을 뜻한다.

2.3.2 추론

검정통계량 λ의 근사적인 분포를 찾을 수 없으므로 관측된 자료를 순열검정법을 사용하여 p-값을 계산하는 몬테카를로 기반 가설 검정을 사용하였다.

2.4 윌콕슨 순위합 검정

윌콕슨 순위합 검정은 윌콕슨(Wilcoxon, 1945)에 의해 제안된 방법으로 독립적인 두 개 모집단에서 얻어진 확률표본으로부터 각 모집단의 위치모수에 대한 추정과 검정을 하는 비모수 통계분석 방법이다. 두 모집단으로부터 얻어진 각각 크기 m과 n인 확률표본을 $X_1, X_2, \cdots, X_m$, $Y_1, Y_2, \cdots, Y_n$이라 한다 ($N = m + n$, $m > n$).

$$X_i = \theta + e_i \ , \qquad i = 1,\ldots,m$$
$$Y_j = \theta + \Delta + e_{m+j} \ , \qquad j = 1,\ldots,n$$

단, θ는 미지의 상수, Δ는 이동모수(두 모집단의 위치모수의 차), e는 오차항이고 N개의 오차항 e들은 서로 독립이고 동일한 연속분포를 따른다고 가정한다. 이 경우 관심 있는 모수 Δ에 대한 검정은 다음과 같은 귀무가설로 표현할 수 있다.

$$H_0 : \Delta = 0$$

2.4.1 검정통계량

두 모집단으로부터 얻어진 $N = m + n$ 개의 X, Y 값들에 작은 값부터 순서대로 순위를 부여한다. 이 때, Y_j의 순위를 R_j라 한다. 혼합표본에서 Y표본에 부여된 순위의 합인 윌콕슨 순위합 통계량은 다음과 같다.

$$W = \sum_{j=1}^{n} R_j$$

2.4.2 추론

유의수준 α에서 위의 귀무가설과 각각의 대립가설에 따른 기각역은 다음과 같다.

표 1. 윌콕슨 순위합 검정의 정확한 기각역

대립가설	정확한 기각역
$H_a : \Delta > 0$	$W \geq w(\alpha, m, n)$
$H_a : \Delta < 0$	$W < w(1-\alpha, m, n)$
$H_a : \Delta \neq 0$	$W \geq w(\alpha/2, m, n)$ 또는 $W < w(1-\alpha/2, m, n)$

여기서 $w(\alpha, m, n)$는 귀무가설 하에서 윌콕슨 순위합 통계량 W의 분포의 상위 100α 백분위수를 나타낸다. 즉, $P_0\{W \geq w(\alpha, m, n)\} = \alpha$를 만족하는 상수이다.

2.4.3 대표본 근사

두 표본에서 m, n이 충분히 크다면 표준화를 시켜 가설 검정을 할 수 있다. 귀무가설 하에서 표준화된 W통계량의 평균과 분산은 각각 다음과 같다.

$$E_0(W) = \frac{n(m+n+1)}{2}, \quad Var_0(W) = \frac{mn(m+n+1)}{12}$$

따라서 이를 이용한 표준화된 통계량은 다음과 같다.

$$Z_W = \frac{W - E_0(W)}{\sqrt{Var_0(W)}} = \frac{W - n(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}$$

표본의 크기가 커질 때 귀무가설 하에서 Z_W는 표준정규분포를 따르게 된다.

유의수준 α에서 Z_W를 이용한 기각역은 다음과 같다.

표 2. 윌콕슨 순위합 검정의 대표본 근사 기각역

대립가설	대표본 근사 기각역
$H_a : \Delta > 0$	$Z_W > z_\alpha$
$H_a : \Delta < 0$	$Z_W < - z_\alpha$
$H_a : \Delta \neq 0$	$\lvert Z_W \rvert > z_{\alpha/2}$

여기서 z_α는 귀무가설 하에서 표준화된 통계량 Z_W 분포(표준정규분포)의 상위 100α 백분위수를 나타낸다. 즉, $P_0\{Z_W > z_\alpha\} = \alpha$를 만족하는 상수이다.

2.5 비모수적 방법을 이용한 공간검색통계량

위의 2.4에서 소개한 윌콕슨 순위합 검정 방법은 독립적인 두 개의 모집단에서 두 군을 비교하는 비모수적 통계량으로 두 군 간의 차이가 있는지를 검정하는 방법이다. 이 때, 두 군을 scanning window의 안과 밖으로 확장한다. 한 군을 scanning window 안으로, 다른 한 군을 scanning window 밖으로 간주하여 scanning window 안의 결과변수의 비율이 scanning window 밖의 비율보다 높거나 낮은지를 검정한다.

2.5.1 검정통계량

Scanning window 안을 Z라고 할 때, Z와 Z^c에서 얻어진 모든 연속적인 값 $x_i,\ i=1,\dots,N$에 작은 값부터 순서대로 순위를 부여한다. Z에 속하는 지역을 2.4에서 소개한 Y_j에 대응시키고 Z^c에 속하는 지역을 X_i에 대응시킨다. 즉, Z에서 관측된 값들의 순위의 합인 W를 구하고, 대표본 근사를 이용하여 각각의 Z마다 설정한 대립가설에 해당하는 검정을 한다. 각각의 검정 결과 나오는 유의확률(p-value) 중 가장 작은 값을 검정통계량으로 사용한다.

$$H_0 : \mu_Z = \mu_{Z^c}\ (\text{for all } Z) \quad vs. \quad H_a : \mu_Z \neq \mu_{Z^c}\ (\text{for some } Z)$$

귀무가설과 대립가설이 위와 같을 경우, 검정통계량은 다음과 같다.

$$\lambda = \min_Z(p-value) = \min_Z 2\min\{\Pr(Z \le Z_W|H_0), \Pr(Z \ge Z_W|H_0)\}$$

이 때, 검정통계량 λ가 최소가 되는 Z를 찾고자 하는 군집이라고 한다. 즉, 모든 scanning window 안과 밖을 윌콕슨 순위합 검정 방법으로 비교하였을 때 가장 유의하게 차이가 나는 Z를 결과변수의 비율이 가장 높거나 낮은 군집으로 생각한다.

$$H_0 : \mu_Z = \mu_{Z^c} \ (\text{for all } Z) \quad vs. \quad H_a : \mu_Z > \mu_{Z^c} \ (\text{for some } Z)$$
$$H_0 : \mu_Z = \mu_{Z^c} \ (\text{for all } Z) \quad vs. \quad H_a : \mu_Z < \mu_{Z^c} \ (\text{for some } Z)$$

만약 대립가설이 위와 같이 Z 안의 값이 Z 밖보다 크거나 작다고 할 경우의 검정통계량은 각각 다음과 같다.

$$\lambda = \min_Z(p-value) = \min_Z \Pr(Z \geq Z_W | H_0)$$
$$\lambda = \min_Z(p-value) = \min_Z \Pr(Z \leq Z_W | H_0)$$

2.5.2 추론

검정통계량 λ의 근사적인 분포를 찾을 수 없으므로 기존의 방법들과 같이 관측된 자료를 순열검정법을 사용하여 p-값을 계산하는 몬테카를로 기반 가설 검정을 사용한다.

제3장 모의실험

이 장에서는 새로 제시한 비모수적 방법의 검정력과 정확도를 비교하기 위해 모의실험을 시행하였다. 자료가 여러 가지 분포를 따를 때 군집 안과 밖의 평균의 차이에 따른 검정력과 정확도를 평가한다. 본 논문에서는 각 지역마다 1개의 값을 가지는 경우만을 가정하였으며 실제 군집 안의 값이 군집 밖의 값보다 큰 경우($H_a : \mu_Z > \mu_{Z^c}$)를 평가하였다.

3.1 모의실험 설계

8×8 행렬을 통해 그림 1과 같은 64개의 연구 지역을 만들었다. True cluster는 임의로 22번 지역을 중심으로 반지름 $\sqrt{2}$ 이내에 있는 9개의 지역(id - 13, 14, 15, 21, 22, 23, 29, 30, 31)으로 설정하였다. Maximum scanning window size는 전체 지역의 약 20%로 최대 13개 지역까지 포함한 경우로 제한하였다. 여러 가지 다른 분포 하에서의 검정력과 정확도를 보기 위해 정규분포, 이중지수분포, 로지스틱분포, 균등분포, 로그정규분포, 코시분포, t분포를 이용했다. 각각 분포 하에서 군집 안과 밖의 평균을 다르게 하는 경우에 대해 검정하였다.

그림 1. 임의로 생성한 연구 지역

57	58	59	60	61	62	63	64
49	50	51	52	53	54	55	56
41	42	43	44	45	46	47	48
33	34	35	36	37	38	39	40
25	26	27	28	29	30	31	32
17	18	19	20	21	22	23	24
9	10	11	12	13	14	15	16
1	2	3	4	5	6	7	8

군집 안과 밖의 평균을 다르게 설정하는 경우 정규분포, 이중지수분포, 로지스틱분포, 균등분포 하에서 true cluster 안은 $D(\mu, \sigma^2) = D(0 + c\sqrt{2}, 1)$ ($c = 0.5, 1, 1.5$) 를 따르게 설정하고 true cluster 밖은 $D(\mu, \sigma^2) = D(0, 1)$를 따르게 설정한다. 이 때, D는 해당 분포를 나타낸다. 로그정규분포의 경우 평균이 0이 될 수 없으므로 true cluster 안은 $D(\mu, \sigma^2) = D(2 + c\sqrt{2}, 1)$ ($c = 0.5, 1, 1.5$), 밖은 $D(\mu, \sigma^2) = D(2, 1)$를 따르게 한다. 코시분포의 경우에는 평균과 분산을 지정할 수 없으므로 scale을 1로 고정하고 location을 각각 2, 4, 6으로 설정하였고, t분포의 경우에는 자유도 3을 따르는 경우에 평균 차이가 $c\sqrt{2}$ ($c = 0.5, 1, 1.5$)만큼씩 나도록 설정하였다.

각 경우마다 1000개의 자료를 생성하였고 유의수준 5%하에서 검정하였다. 이 자료들을 이용해 normal 방법을 사용하여 분석했을 경우와 비모수적 방법을 사용해 분석한 경우의 검정력을 비교하였으며, 정확도 평가를 위해 민감도와 양성예측도를 계산하였다.

$$\text{검정력} = \frac{\text{유의확률이 } 0.05 \text{ 미만인 표본의 수}}{1000\text{개의 표본}}$$

으로 계산할 수 있고, L을 유의한 표본의 총 개수라고 할 때,

$$\text{민감도} = \frac{1}{L}\sum_{l=1}^{L} \frac{l\text{번째 유의한 표본에서 찾아낸 } cluster\text{중 } true\,cluster\text{에 속하는 지역의 수}}{true\,cluster\text{의 지역의 수}}$$

$$\text{양성예측도} = \frac{1}{L}\sum_{l=1}^{L} \frac{l\text{번째 유의한 표본에서 찾아낸 } cluster\text{중 } true\,cluster\text{에 속하는 지역의 수}}{l\text{번째 유의한 표본에서 찾아낸 } cluster\text{의 지역의 수}}$$

로 표현할 수 있다.

그러나 이 경우에 민감도와 양성예측도는 찾아낸 군집의 크기에 영향을 받기 때문에 찾아낸 군집의 크기에 의존적이지 않은 Takahashi and Tango (2006)가 제안한 extended power의 profile을 함께 제시한다. Extended power를 통해 검정력과 정확도를 동시에 평가해 하나의 값으로 제시할 수 있다.

여기서 extended power는 다음의 식으로 나타낼 수 있다.

$$I(w^-, w^+) = \sum_{l \geq 1} \sum_{s \geq 0} W(l,s;w^-,w^+)P(l,s)$$

$l \geq 1, s \geq 0$일 때, $P(l,s)$는 bivariate power distribution으로 다음과 같은 식을 따른다. 여기서 MLC는 most likely cluster를 뜻한다.

$$P(l,s) = \frac{l\text{개의 찾아낸 지역 중 } true\ cluster\text{에 속하는 지역이 } s\text{개인 유의한 } MLC\text{의 수}}{1000\text{개의 표본}}$$

또한 $W(l,s;w^-,w^+)$는 가중함수로 다음의 식으로 표현 가능하다.

$$W(l,s;w^-,w^+) = \begin{cases} \sqrt{(1-\min w^-(s^*-s),1)(1-\min w^+(l-s),1)} \\ \quad (s \leq l; 0 \leq s \leq s^*, 1 \leq l) \\ 0 \quad (otherwise) \end{cases}$$

여기서 s^*는 true cluster에 속하는 지역의 수이고, w^-와 w^+는 true cluster가 아닌 지역을 잘못 찾아내는 것에 대한 사전 정의된 페널티로 $0 \leq w^+ \leq w^- \leq 1$ 로 정할 수 있다. 페널티 값에 따라 extended power 값이 달라지므로 모든 페널티에 대한 전체적인 경향성을 볼 수 있는 extended power의 profile로 결과를 제시한다.

$$Q(r|s^*) = I(1/s^*, r/s^*), \quad (0 \leq r \leq 1)$$

이를 식으로 나타내면 위와 같다. 여기서 $r = w^+/w^- \ (0 \leq r \leq 1)$, $w^- = 1/s^*$을 나타낸다. 그러므로 extended power의 profile을 통해 r에 따른 모든 연속적인 extended power의 값을 나타낼 수 있다.

3.2 모의실험 결과

표 3~5는 군집 안과 밖의 평균을 다르게 설정하였을 때 모의실험 결과이다. 각각 true cluster 안과 밖의 분산이 동일할 때, 평균이 $c\sqrt{2}$ 만큼 차이가 나도록 설정했고 이 때 c는 0.5, 1.0, 1.5 세 가지 경우로 설정하였다. 각각의 c마다 두 방법의 검정력, 민감도, 양성예측도를 비교하였다.

먼저 검정력을 비교해 보면, 두 방법 모두에서 c가 증가함에 따라 군집 안과 밖의 평균 차이가 커지므로 검정력이 증가하였다. 하지만 대부분의 경우에서 비모수적 방법을 사용하였을 때 검정력이 더 높다는 것을 확인할 수 있다. 특히 이중지수분포, 로그정규분포, 코시분포, t분포처럼 비대칭적인 모양을 가지거나 꼬리가 두꺼운 분포의 경우 normal 방법을 사용하였을 때의 검정력은 매우 낮은 반면 비모수적 방법을 사용하였을 때는 $c=0.5$일 때를 제외하고 대부분 60% 이상으로 검정력이 더 좋았다.

민감도와 양성예측도 또한 검정력과 비슷하게 c가 증가함에 따라 모든 경우에서 증가하는 양상을 보였다. 대부분의 경우에서 민감도는 비모수적 방법을 사용한 경우가 더 좋았다. 하지만 양성예측도의 경우에는 두 방법이 거의 비슷하거나 몇몇 경우에서는 normal 방법이 더 좋은 양성예측도를 나타내는 경우도 있다. 하지만 이중지수분포, 로그정규분포, 코시분포, t분포에서는 비모수적 방법을 사용하였을 때의 양성예측도가 더 크거나 비슷했고, 나머지 경우에서도 normal 방법을 사용했을 때의 결과와 차이가 크지 않았다.

군집 안과 밖의 평균 차이가 달라지는 경우에 여러 가지 분포 하에서 두 방법의 검정력과 정확도를 비교해 본 결과, 검정력과 민감도는 대부분 분포에서 비모수적 방법을 사용한 경우가 더 좋았고 양성예측도는 두 방법 간의 차이가 크지는 않았지만 비대칭적이거나 꼬리가 두꺼운 분포를 따르는 경우에는 비모수적 방법을 사용하는 것이 더 좋은 결과를 얻을 수 있었다.

그러나 민감도와 양성예측도의 경우 찾아낸 군집의 크기에 따라 값이 결정된다. 대부분 normal 방법을 사용했을 때 더 작은 크기의 군집을 찾아내므로 상대적으로 normal 방법의 민감도가 작게, 양성예측도는 크게 구해지는 경향을 보인다. 이를 보완하기 위해 검정력과 정확도를 동시에 볼 수 있도록 제안된 extended power의 profile을 그림 2를 통해 확인해 보면, 균등분포의 경우 모든 평균 차이에서 normal 방법의

extended power가 더 높았다. 정규분포와 로지스틱 분포의 경우에는 평균 차이가 작게 날 때 ($c = 0.5, 1.0$)에는 비모수적 방법의 extended power가 조금 더 높지만 평균 차이가 크게 날수록 ($c = 1.5$) normal 방법의 extended power가 더 높게 나타나는 것을 볼 수 있다. 그러나 위의 결론과 마찬가지로 이중지수분포, 로그정규분포, 코시분포, t분포에서는 비모수적 방법을 사용하였을 때의 extended power가 normal 방법을 사용했을 때 보다 훨씬 더 높은 값을 가지는 것을 볼 수 있다.

이를 통해 자료가 비대칭적이거나 꼬리가 두꺼운 분포를 따르는 경우에는 normal 방법을 사용하는 것보다 비모수적 방법을 사용하는 것이 검정력과 정확도를 모두 높일 수 있는 방법임을 알 수 있다.

표 3. cluster 안과 밖의 평균 차이에 따른 검정력과 정확도 ($c = 0.5$)

	비모수적 방법	Normal 방법
	검정력 (%)	
정규분포	17.3	14.8
이중지수분포	24.0	13.5
로지스틱분포	17.7	12.9
균등분포	13.4	15.4
로그정규분포	19.7	7.6
코시분포	31.4	5.7
t분포	13.9	7.6
	민감도	
정규분포	0.71	0.65
이중지수분포	0.81	0.67
로지스틱분포	0.72	0.64
균등분포	0.65	0.66
로그정규분포	0.74	0.50
코시분포	0.83	0.44
t분포	0.66	0.44
	양성예측도	
정규분포	0.63	0.65
이중지수분포	0.74	0.72
로지스틱분포	0.64	0.65
균등분포	0.62	0.69
로그정규분포	0.64	0.52
코시분포	0.76	0.38
t분포	0.59	0.55

표 4. cluster 안과 밖의 평균 차이에 따른 검정력과 정확도 ($c = 1.0$)

	비모수적 방법	Normal 방법
	검정력 (%)	
정규분포	71.8	69.8
이중지수분포	76.9	62.1
로지스틱분포	76.9	66.7
균등분포	62.2	74.8
로그정규분포	83.2	45.0
코시분포	76.1	16.9
t분포	45.8	25.9
	민감도	
정규분포	0.90	0.87
이중지수분포	0.93	0.89
로지스틱분포	0.91	0.89
균등분포	0.88	0.86
로그정규분포	0.93	0.86
코시분포	0.92	0.79
t분포	0.86	0.75
	양성예측도	
정규분포	0.85	0.89
이중지수분포	0.88	0.91
로지스틱분포	0.88	0.91
균등분포	0.85	0.89
로그정규분포	0.87	0.87
코시분포	0.88	0.74
t분포	0.80	0.80

표 5. cluster 안과 밖의 평균 차이에 따른 검정력과 정확도 ($c = 1.5$)

	비모수적 방법	Normal 방법
	검정력 (%)	
정규분포	98.6	98.4
이중지수분포	97.6	94.1
로지스틱분포	98.8	96.8
균등분포	98.4	99.1
로그정규분포	99.8	87.9
코시분포	90.9	30.4
t분포	83.8	58.8
	민감도	
정규분포	0.97	0.96
이중지수분포	0.97	0.96
로지스틱분포	0.97	0.96
균등분포	0.97	0.96
로그정규분포	0.99	0.96
코시분포	0.94	0.87
t분포	0.92	0.86
	양성예측도	
정규분포	0.92	0.96
이중지수분포	0.93	0.96
로지스틱분포	0.93	0.97
균등분포	0.93	0.96
로그정규분포	0.93	0.95
코시분포	0.91	0.85
t분포	0.87	0.89

그림 2. 여러 가지 분포 하에서의 extended power의 profile

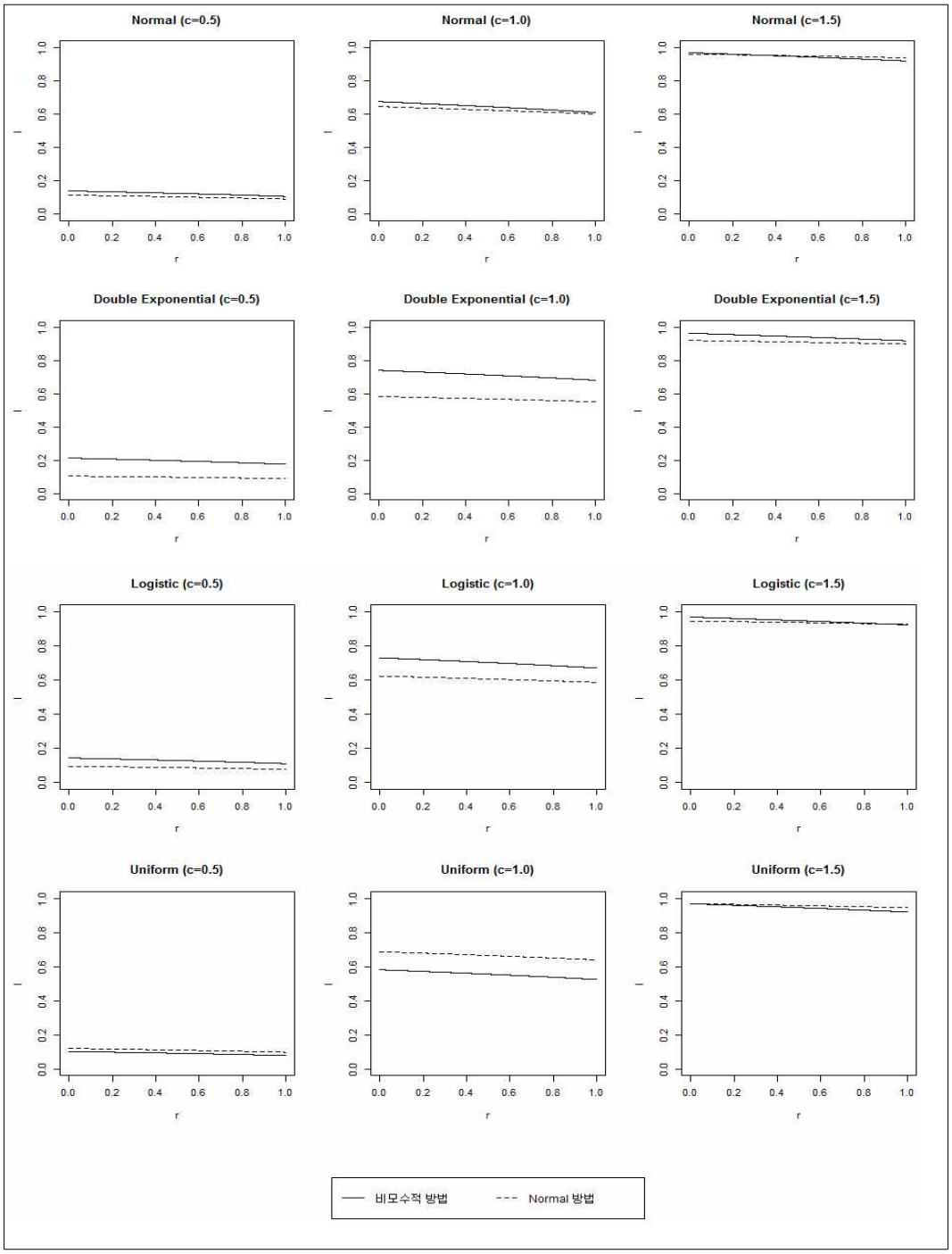
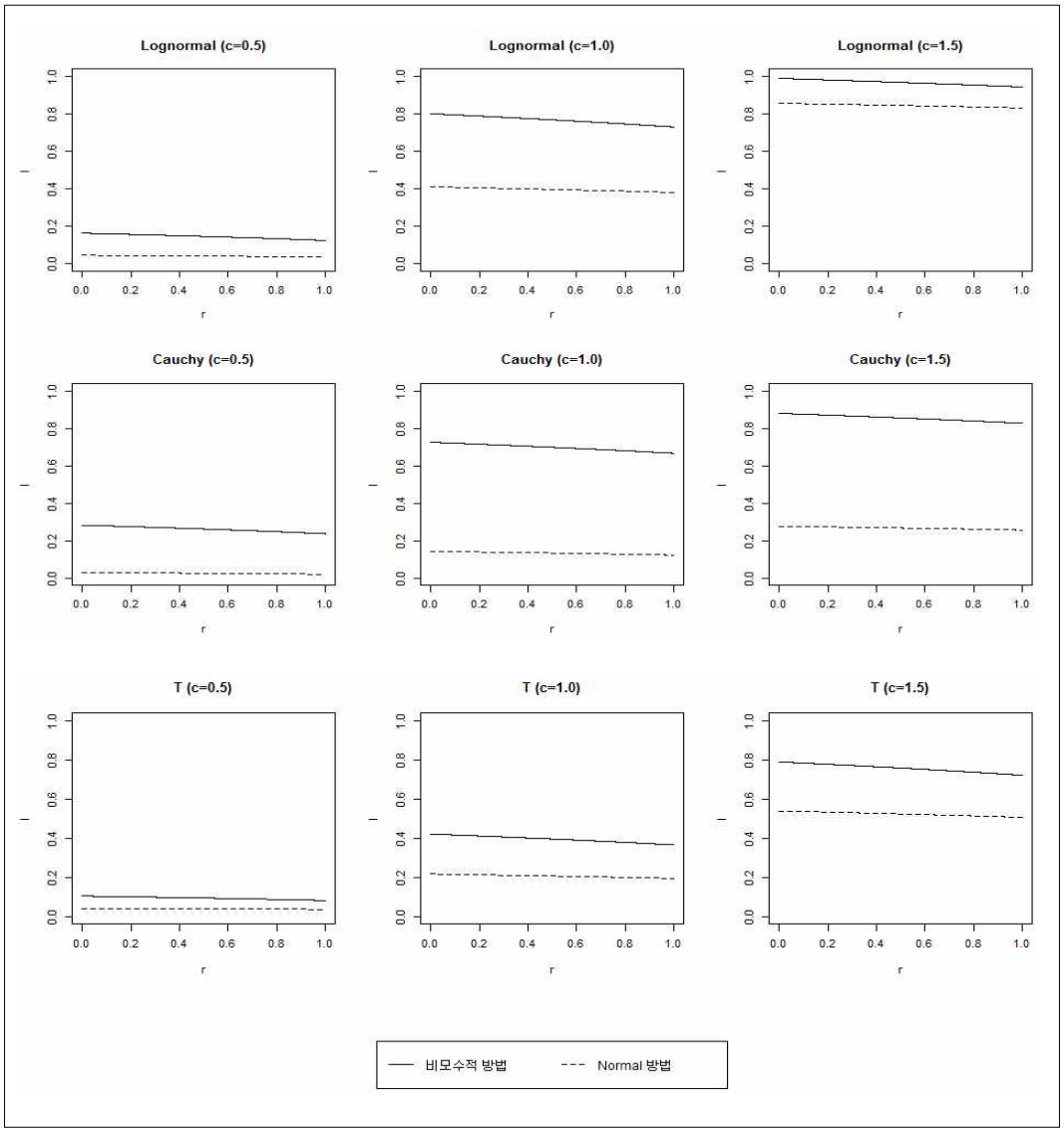


그림 2. 계속


Lognormal (c=0.5)
Lognormal (c=1.0)
Lognormal (c=1.5)
Cauchy (c=0.5)
Cauchy (c=1.0)
Cauchy (c=1.5)
T (c=0.5)
T (c=1.0)
T (c=1.5)
r
비모수적 방법
Normal 방법

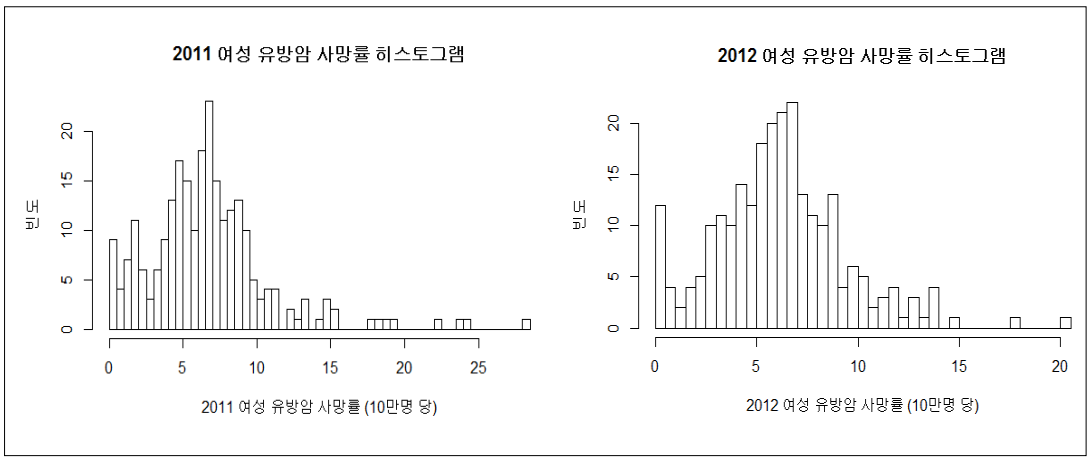
제4장 실제자료 분석

이 장에서는 우리나라 여성의 유방암 사망률 자료를 사용하여 앞에서 소개한 세 가지 방법을 적용한 결과를 비교하고자 한다.

4.1 자료 설명

통계청에서 제시한 2011년, 2012년 여성의 유방암 사망률 자료로 우리나라 248개의 시군구 별 연령 표준화 사망률을 사용하였다. 다른 암종과는 다르게 여성 유방암의 경우 사망률이 비대칭적인 경향을 보이는 것을 그림 3을 통해 확인할 수 있다. 앞서 소개한 normal 방법, weighted normal 방법, 본 논문에서 제시한 비모수적 방법을 각각 적용하여 도출된 군집과 개인 수준의 자료와 인구 집단의 자료로 포아송 분포를 이용한 방법을 사용하였을 때 도출된 군집 간의 차이를 비교해 보고자 한다. 이 때 인구 집단의 자료는 2010년 자료를 이용하였고, 유의수준 0.05하에서 검정하였다.

그림 3. 우리나라 2011-2012년 시군구별 여성 유방암 연령 표준화 사망률 히스토그램



4.2 분석 결과

모든 경우에서 maximum scanning window는 20%로 최대 50개의 지역을 포함하도록 제한하였으며 사망률이 더 높은 군집을 찾았다($H_a : \mu_Z > \mu_{Z^c}$). 세 가지 방법 모두 몬테카를로 가설 검정 기반으로 하여 찾아낸 군집의 통계학적 유의성을 평가하였다. weighted normal의 경우 일반적인 경우와 같이 각 지역의 가중치를 $1/var_z$로 주었으며 여기서 $var_z = mortality^2 / number\ of\ death$로 계산하였다.

표 6은 2011년 여성 유방암 연령표준화 사망률 자료를 분석한 결과이다. 이를 통해 공간 군집 탐색(spatial cluster detection) 분석 결과를 볼 수 있으며, 그림 4에서 총 4가지 방법을 사용해 찾아낸 공간 군집을 지도상에 각각 표시하였다. 이 때, 개인의 자료와 인구 집단의 자료로 포아송 방법을 사용하여 찾아낸 군집과 지역 별로 각각 주어진 사망률 자료로 나머지 방법들을 이용하여 찾아낸 군집들의 차이를 비교해 본다.

표 6과 그림 4를 통해 2011년 자료를 분석한 결과를 보면 포아송 방법을 사용한 경우 유의한 군집을 찾아낸 것을 알 수 있다. 반면 normal 방법을 이용하였을 때에는 p값 0.08로 유의수준 0.05하에서 유의한 군집을 찾을 수 없었다. Weighted normal 방법을 사용하였을 때는 유의한 군집을 찾을 수 있었지만, 찾아낸 군집이 포아송 방법으로 찾아낸 군집과 다른 지역이고 겹치는 부분이 거의 없는 것을 볼 수 있다. 반면 비모수적 방법을 사용했을 경우에는 유의한 군집을 찾을 수 있었고, 찾아낸 지역의 개수가 더 많긴 하지만 포아송 방법으로 찾아낸 군집과 비슷한 지역을 찾아낸 것을 볼 수 있다. 2012년 자료를 분석한 결과에서는 포아송 방법을 사용한 경우 유의한 군집이 존재하지 않았고 normal 방법과 비모수적 방법에서도 모두 사망률이 높은 유의한 군집을 찾지 못하였다. 하지만 weighted normal 방법을 사용했을 경우에는 유의한 군집을 찾을 수 있었다.

이를 통해 자료가 비대칭적 경향을 보일 때 normal 방법과 weighted normal 방법은 포아송 방법으로 찾아낸 군집과 유의성도 일치하지 않고, 찾아낸 지역도 다른 것을 볼 수 있다. 하지만 비모수적 방법을 사용했을 경우에는 찾아낸 지역과 유의성이 포아송 방법을 사용한 경우와 거의 비슷한 경향성을 보이는 것을 확인할 수 있다.

그림 4. 우리나라 2011년 시군구별 여성 유방암 연령 표준화 사망률이 높은 군집

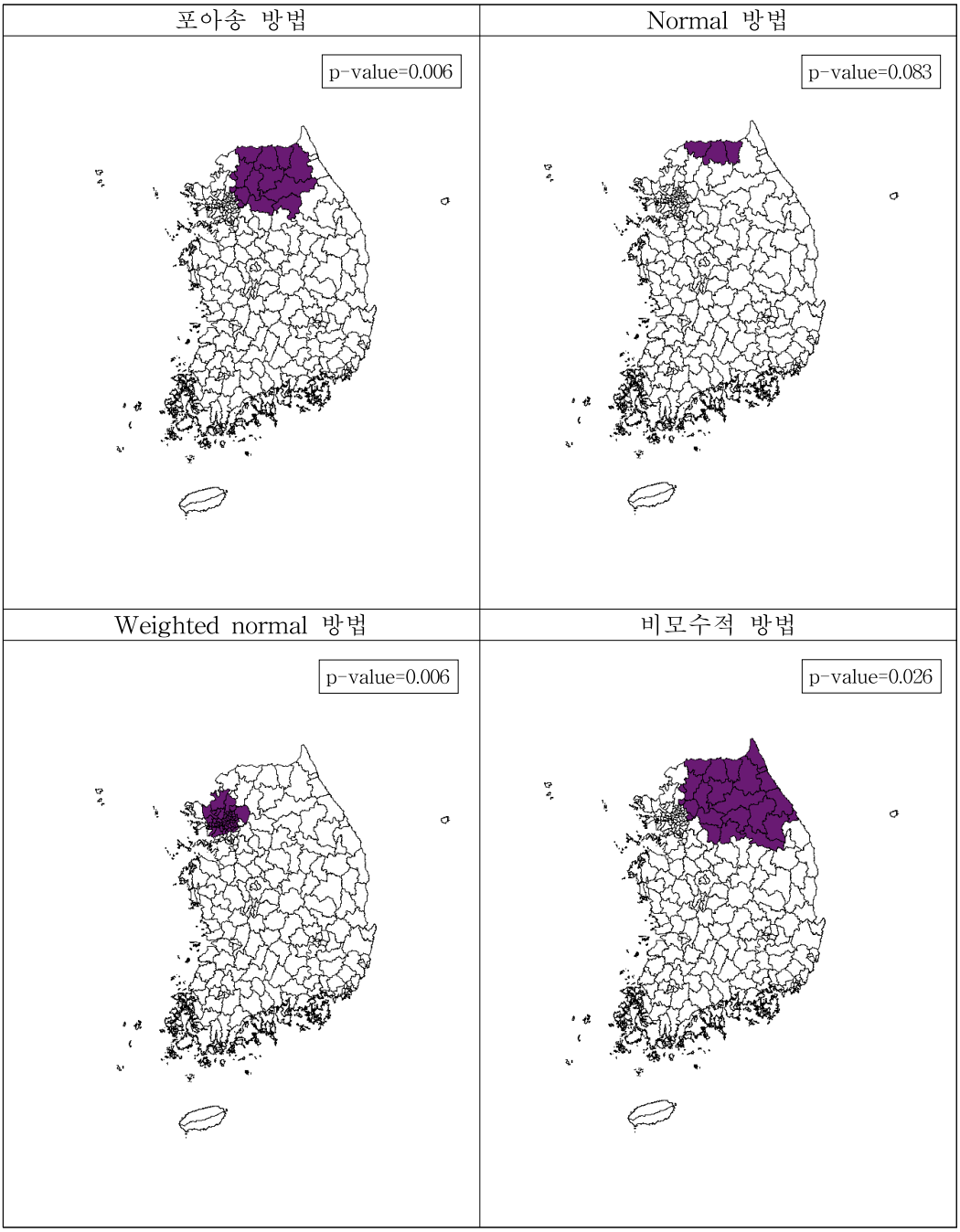


표 6. 우리나라 2011년 시군구별 여성 유방암 연령 표준화 사망률이 높은 군집 분석 결과

사용한 방법	군집 안의 사망률 평균 (10만 명 당)	군집 밖의 사망률 평균	군집 안에 속하는 지역의 개수	p 값
포아송 방법*	12.40	8.10	15	0.006
Normal 방법	19.63	6.55	3	0.083
Weighted normal 방법†	6.04	3.80	50	0.006
비모수적 방법	10.35	6.32	24	0.026

*포아송 방법의 경우 10만 명 당 발생자수를 표기함

† Weighted normal 방법의 경우 weighted mean을 표기함

제5장 결론 및 고찰

본 논문은 연속적인 자료의 공간 군집을 탐색할 때 자료의 분포가 비대칭적이거나 꼬리가 두꺼운 경우에 비모수적 방법을 적용하는 것에 대해 연구하였다.

현재 제안된 normal 방법은 자료가 정규분포를 따르거나 다른 대칭적인 분포를 따를 때에는 검정력과 정확도가 높지만 그렇지 않은 경우 검정력과 정확도가 떨어지는 단점이 있다. 또한 분산이 커지는 경우에 검정력이 크게 감소하지 않는다는 문제가 있다. 군집 내에서 분산이 커지면 불확실성이 증가하기 때문에 찾아낸 지역을 결과변수의 비율이 높거나 낮은 지역이라고 말하기 어렵다. 불확실성 문제는 weighted normal 방법을 통해 해결할 수 있지만 검정력과 정확도의 경우 확인된 바가 없다. 따라서 본 논문에서는 normal 방법과 비모수적 방법을 이용하여 여러 분포 하에서의 검정력과 정확도를 비교해 보았다.

모의실험을 통해 normal 방법과 비모수적 방법을 비교한 결과 자료가 대칭적이지 않거나 꼬리가 두꺼운 분포에서는 비모수적 방법을 적용했을 경우의 검정력, 민감도, 양성예측도가 모두 높았으며, 대칭적인 분포 하에서도 비모수적 방법을 사용했을 때와 normal 방법을 사용했을 때의 결과가 큰 차이를 보이지 않았다. 또한 찾아낸 군집의 크기에 의존하지 않는 extended power로 검정력과 정확도를 동시에 고려했을 때에도 같은 양상을 보였다.

실제 자료 분석에서도 자료가 비대칭적으로 분포했을 경우 분포 가정을 기반으로 한 normal 방법과 weighted normal 방법은 포아송 방법을 이용해서 찾아낸 군집과 다른 군집을 찾는 경향을 보이며 유의성에 있어서도 차이를 보이고 있다. 하지만 비모수적 방법을 사용하였을 경우에는 찾아낸 군집과 유의성이 포아송 방법과 거의 일치하는 것을 확인할 수 있다. 포아송 방법으로 찾아낸 군집을 true cluster로 보기는 어렵지만 사망률 등의 자료가 지역 단위로 주어졌을 경우 개인 수준의 원자료를 이용하여 찾아낸 군집과 비슷한 군집을 찾는 데 의미가 있다고 본다.

또한 모의실험과 실제 자료 분석의 결과를 통해 볼 때, normal 방법의 경우 너무 작은 군집을 찾아내는 경향을 보인다는 것을 알 수 있었다. 추가적인 비교가 필요하겠지만 이는 true cluster 내에 이상치가 존재한다면 이상치를 포함하는 작은 지역만을 찾아내어 true cluster를 온전히 찾아내지 못하는 경향을 보인다고 할 수 있다.

위의 내용들을 종합하여 볼 때 연속적 자료의 분포가 비대칭적이거나 꼬리가 두꺼운 경우에는 기존의 모수적 방법들을 사용하는 것보다 비모수적 방법을 사용하는 것이 검정력과 정확도를 높일 수 있는 방법이 될 것으로 기대된다.

참고문헌

Kulldorff, M., Huang, L., Konty, K. 2009. "A scan statistic for continuous data based on the normal probability model". *International Journal of Health Geographics*, 8(58).

Huang, L., Tiwari, R. C., Zou, Z., Kulldorff, M., Feuer, E. J. 2009. "Weighted normal spatial scan statistic for heterogeneous population data". *Journal of the American Statistical Association*, 104(487): 886-898.

Kulldorff, M. 1997. "A spatial scan statistic". *Communications in Statistics - Theory and Methods*, 26(6): 1481-1496.

Jung, I., Kulldorff, M, Klassen A. C. 2007. "A spatial scan statistic for ordinal data". *Statistics in medicine,* 26(7): 1594-1607.

Jung, I., Kulldorff, M., Richard, O. J. 2010. "A spatial scan statistic for multinomial data". *Statistics in medicine*, 29(18): 1910-1918.

Frank Wilcoxon. 1945. "Individual comparisons by ranking methods". *Biometrics Bulletin*, 1(6): 80-83.

Dwass, M. 1957. "Modified randomization tests for nonparametric hypothesis". *Annals of Mathematical Statistics*, 28(1): 181-187.

Takahashi, K., Tango, T. 2006. "An extended power of cluster detection tests". *Statistics in medicine*, 25: 841-852.

Statistics Korea. 2012. "시군구/사망원인(50항목)/성/사망자수, 사망률, 연령표준화 사망률". [database online]. *Korea Statistical Information System (KOSIS)* [cited 2014.9.5.] <http://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B34E13&conn_path=I2>.

ABSTRACT

A nonparametric spatial scan statistic

Cho, Ho Jin
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

In spatial epidemiology or geographical disease surveillance, spatial scan statistics are often used to detect spatial clusters with unusually high or low rates of outcome and to assess the statistical significance of detected clusters. The existing models for the spatial scan statistic are parametric methods, which assume a specific distribution for the data. Kulldorff, Huang, and Konty (2009) proposed a normal model for continuous outcome data. When the data follow a skewed or heavy tailed distribution, however, the assumption on the distribution may not be suitable and hence it is difficult to apply the existing methods. Therefore, we propose a nonparametric method that does not require any distributional assumption. The proposed method is based on Wilcoxon rank sum test instead of the likelihood ratio test.

Through a simulation study, we compared the statistical power and accuracy of the normal and nonparametric methods when data were simulated from various distributions. In addition, we applied the nonparametric method to Korean female breast cancer mortality data and compared the result with those from parametric methods.

Simulation study results showed that the statistical power and accuracy of the normal model are low when the underlying distribution is skewed or heavy tailed. On the other hand, the nonparametric method seems to have higher statistical

power and accuracy than the normal model in almost all cases. In the actual data analysis, the nonparametric method identified a cluster similar to the one detected using the poisson model, while the normal and weighted normal models found a very different cluster.

In conclusion, the proposed nonparametric scan statistic can be a good alternative to the normal model for continuous data following a skewed or heavy tailed distribution.

Keywords : spatial scan statistic, nonparametric method, Wilcoxon rank sum test