The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.3.331
32(3), 331–348

Statistical methods for testing tumor heterogeneity

Dong Neuck Lee[a] · Changwon Lim[a,1]

[a]Department of Applied Statistics, Chung-Ang University

(Received October 25, 2018; Revised December 8, 2018; Accepted January 3, 2019)

Abstract

Understanding the tumor heterogeneity due to differences in the growth pattern of metastatic tumors and rate of change is important for understanding the sensitivity of tumor cells to drugs and finding appropriate therapies. It is often possible to test for differences in population means using t-test or ANOVA when the group of N samples is distinct. However, these statistical methods can not be used unless the groups are distinguished as the data covered in this paper. Statistical methods have been studied to test heterogeneity between samples. The minimum combination t-test method is one of them. In this paper, we propose a maximum combinatorial t-test method that takes into account combinations that bisect data at different ratios. Also we propose a method based on the idea that examining the heterogeneity of a sample is equivalent to testing whether the number of optimal clusters is one in the cluster analysis. We verified that the proposed methods, maximum combination t-test method and gap statistic, have better type- Ⅰ error and power than the previously proposed method based on simulation study and obtained the results through real data analysis.

Keywords: heterogeneity, k-means clustering, gap statistic, determining the number of clusters

1. 서론

데이터의 이질성을 검정하는 것은 통계학의 중요한 이슈이다. William Sealy Gosset에 의해 Student (1908)라는 필명으로 발표한 t-test와 Ronald A. Fisher (1918)가 제안한 분산분석은 집단 간의 평균의 이질성이 존재하는 지 검정하는 가장 널리 알려진 방법이다. 모집단의 평균을 비교하기 위한 t-검정과 분산분석 모두 중요한 전제조건이 하나 만족되어야 하는데 그것은 모든 표본이 어떤 집단에 속하는지 명확해야 한다는 것이다. 집단별로 추출된 표본으로부터 모평균과 모분산을 추정하고 그 추정값을 사용하여 가설검정을 위한 검정통계량을 계산한다.

그러나 추출된 표본의 집단이 구분되지 않은 경우에 표본 간의 이질성이 존재하는지를 검정하는 것 또한 여러 과학 분야에서 중요한 주제이다. 이를테면, 시간에 따른 종양의 변화 정도를 측정한 값이 종양별로 동일한지 혹은 종양 이질성(tumor heterogeneity)이 존재하는 지를 판단할 때 서로 비교할 집단이 구분되어 있지 않지만 투여되는 약물의 민감성이 다른 종양세포가 모여 조직을 형성하기 때문에 약물에 대한 반응정도가 달라 치료가 어려우므로 이를 연구하는 것이 중요하다.

본 논문의 목적은 전이성 종양의 성장패턴 차이와 변화율에 따른 종양 이질성을 파악하기 위한 연구 (Yoo 등, 2017)로부터 얻어진 실제자료에서 종양 이질성을 검정하는 최적의 통계적 방법을 찾는 것이

[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: clim@cau.ac.kr

다. 이 자료는 2010년 1월부터 2014년 12월까지 폐암 전이성을 지닌 10명의 환자들을 대상으로 각각 흉부 CT를 통해 종양의 최대 지름(maximum diameter)과 부피(volume)를 측정하고, 항암치료 후에도 후속 촬영을 실시하여 생성되었다. 10명의 환자들은 1명의 여성 환자와 9명의 남성 환자로 구성되어 있고, 각 환자들이 지닌 종양(nodule)의 수는 4개부터 52개까지 존재한다. 이 자료는 최소 조합 t-검정(minimum combination t-test)을 제안한 Heo와 Lim (2017)의 앞선 연구에서 활용되었다.

이전의 논문에서는 이 데이터를 종양의 관측값인 반응변수와, 종양의 종류인 요인으로 구성된 데이터로 해석하였다. 해당 요인의 유의성을 검정한다면 종양별로 성장패턴 차이가 존재하는지를 검정할 수 있다. 그러나 각 요인(종양)의 표본의 개수가 1이므로 전통적인 ANOVA를 통해 분석할 수 없다. Heo와 Lim (2017)은 이러한 문제를 해결하기 위해 최소 조합 t-검정을 제안하였다. 최소 조합 t-검정은 N개의 표본인 을 임의의 두 집단으로 나누고 평균의 동일성에 대한 가설검정을 실시하는 것이다. 이는 N개의 표본에서 이질성이 존재하지 않는다면 임의로 구분된 집단별 평균차이도 존재 하지 않는 점에 착안한 방법이다. 그러나 이질성을 갖는 표본이 소수인 경우 최소 조합 t-검정 방법의 검정력(power)이 낮다는 한계가 있다. 본 논문에서는 이러한 한계의 원인이 데이터를 크기가 동일한 두 집단으로 분류한다는 점에 있다는 아이디어에 착안하여 이전에 제안된 방법보다 검정력이 대폭 향상된 최대 조합 t-검정(maximum combination t-test)을 제안한다.

한편, 본 논문에서는 전이성 종양의 성장패턴의 이질성 검정에 대해 새로운 관점에서 접근해 보았다. 우리는 이 데이터를 N개의 표본으로 이루어진 범주화 되지 않은 데이터로 해석하고 군집분석을 이용하여 이와 같은 데이터에서 이질성의 존재를 검정하는 것을 제안한다. 군집분석을 통해 최적의 군집의 개수(k)를 결정했을 때, 데이터가 2개 이상의 군집으로 명확히 구분된다면 이는 데이터의 이질성이 존재함을 의미한다. 즉 데이터의 이질성을 검정하는 것이 군집분석에서 군집의 개수가 1개인지 그 이상인지를 검정하는 것으로 치환된다.

군집분석의 결과를 평가하고 적절한 군집의 개수를 결정하는 방법에 대한 연구는 활발히 이루어져 왔다. 순열 검정(permutation test), 교차타당성 검정(cross-validation), 재표집(resampling)을 이용한 방법도 그중 하나이다. 그러나 본 연구에서 다루는 데이터와 같이 표본의 크기가 작을 때 적절한 방법은 아니다. 한편 군집의 개수에 따른 군집분석 결과를 평가하는 지표들이 연구되었다. DB index (Davies와 Bouldin, 1979), Dunn index (Dunn, 1974), Gamma index (Baker와 Hubert, 1976) 등 수십가지의 지표들은 그 결과물이다. gap 통계량(gap statistic) (Tibshirani 등, 2001), silhouette index (Rousseeuw, 1987), elbow 방법 (Thorndike, 1953)은 데이터셋 내에서 최적의 군집의 개수를 결정하는 방법으로써 널리 사용되며 그 중에서도 gap 통계량이 성능이 우수한 것으로 알려져 있다. 이 방법들은 오차를 나타내는 함수를 설정하여 군집의 개수에 따른 함수값의 변화를 통계적으로 평가함으로써 최적의 군집의 개수를 찾는 방법이다. 그러나 모든 지표들이 군집의 개수가 1개인지 그 이상인지를 검정하는데 사용될 수 있는 것은 아니다. 같은 군집내 데이터와 군집이 서로 다른 데이터의 평균 거리를 비교하는 silhouette index는 군집의 개수(k)가 1개일 때 항상 0의 값을 갖으며 2보다 큰 값들 중에서 최적의 군집 개수 k를 찾는 방법이다. 따라서 군집의 개수가 1개인지 그 이상인지를 결정하는 데 이용될 수 없다. 또한 k가 증가함에 따라 군집내 제곱합의 감소 정도에 따라 k를 선택하는 elbow 방법도 마찬가지로 $k = 1$은 선택되지 않으므로 본 연구의 목적에 맞지 않는다. 따라서 $k = 1$이 선택되는 것이 가능하며 k를 선택하는 명확한 기준이 존재하는 gap 통계량만을 본 연구의 대상으로 한다.

모의실험을 통해 최소 조합 t-검정 방법, 최대 조합 t-검정 방법, 군집분석을 통한 접근의 성능을 비교하였다. 특히 우리는 데이터의 이질성이 존재하는 경우 서로 다른 모집단에서 추출된 표본의 비율이 상이한 경우에 높은 성능을 갖는 방법을 찾는 것에 집중하였다. 이러한 경우에 높은 성능을 갖는다는 것은 이질적인 표본이 소수더라도 이를 잘 감지함을 의미하기 때문이다.

본 논문은 총 5장으로 구성되어 있다. 2장에서 최대 조합 t-검정과 군집분석을 통한 이질성 검정 방법을 제안한다. 3장에서는 제안한 방법의 성능을 모의실험을 통해 살펴보겠다. 4장에서는 본 연구의 배경이 된 종양 성장 패턴 데이터를 분석하여 결과를 도출한다. 마지막으로 5장에서는 본 논문의 전체적인 내용을 정리하겠다.

2. 방법론

2.1. 최소 조합 t-검정(minimum combination t-test)

Heo와 Lim (2017)이 제안한 최소 조합 t-검정은 N개의 표본인 $X_1, \ldots, X_N$을 크기가 비슷한 임의의 두 집단으로 나누고 평균의 차이를 검정하는 방법이다. 이때 조합의 수(m)는 집단의 크기에 따라 다음과 같이 정의된다:

$$m = \begin{cases} {}_NC_{\frac{N}{2}}, & \text{if } N = \text{짝수}, \\ {}_NC_{\frac{N-1}{2}}, & \text{if } N = \text{홀수}, \end{cases}$$

여기서 조합의 수는 최대 50,000번으로 제한하였다. j번째 조합에서의 귀무가설과 대립가설은 다음과 같이 설정된다:

$$H_{0j} : \mu_{Aj} = \mu_{Bj} \quad \text{vs.} \quad H_{1j} : \mu_{Aj} \neq \mu_{Bj}, \quad j = 1, \ldots, m,$$

여기서 μ_{Aj}와 μ_{Bj}는 j번째 조합에서 나누어진 두 집단의 모평균들이다. m번의 두 집단의 모평균의 동일성에 대한 검정결과 얻어진 p-값 중 적어도 한 개의 p-값이 유의수준에 대해 유의하다면 귀무가설을 기각하여 나누어진 두 집단이 서로 다른 평균을 가진다고 할 수 있으므로 데이터의 이질성이 존재함을 의미한다. 이 방법은 조합의 개수만큼 나누어진 두 집단에 대하여 t-검정을 실시하기 때문에 m개의 p-값 중 최솟값을 본페르니 교정을 이용한 유의수준과 비교한다 (Dunn, 1961). 따라서 유의수준은 $0.05/m$로 설정된다. 표본 X_i의 차원(p)가 2 이상인 경우 μ_{Aj}와 μ_{Bj}는 벡터가 되며 이 경우 Hotelling's t-squared statistic을 이용하여 검정하였다.

2.2. 최대 조합 t-검정(maximum combination t-test)

최소 조합 t-검정은 N개의 표본인 $X_1, \ldots, X_N$을 크기가 비슷한 임의의 두 집단으로 나누고 평균의 차이를 검정하는 방법인 반면, 최대 조합 t-검정은 표본을 임의의 두 집단으로 나누는 모든 경우의 수를 고려하는 방법이다. 이때 조합의 수(M)는 다음과 같이 정의된다.

$$M = \begin{cases} {}_NC_{\frac{N}{2}} + {}_N C_{\frac{N}{2}-1} + {}_N C_{\frac{N}{2}-2} + \cdots + {}_N C_1, & \text{if } N = \text{짝수}, \\ {}_NC_{\frac{(N-1)}{2}} + {}_N C_{\frac{(N-1)}{2}-1} + {}_N C_{\frac{(N-1)}{2}-2} + \cdots + {}_N C_1, & \text{if } N = \text{홀수}. \end{cases}$$

단, 조합 ${}_NC_c$ ($c = 2/N, N/2-1, \ldots, 1$ if $N = $ 짝수, $c = (N-1)/2, (N-1)/2-1, \ldots, 1$ if $N = $ 짝수)의 수는 최대 50,000으로 제한하였다. 최소 조합 t-검정과 마찬가지로 총 M 가지의 조합에 따라 두 집단의 모평균의 동일성에 대한 검정을 시행한다. 단, 표본이 1개인 집단과 나머지 $N-1$개의 집단으로 나뉘는 경우에는 이를 시행할 수 없고 이러한 조합에서의 귀무가설과 대립가설은 다음과 같이 설정되며 단일 모평균 검정을 시행하게 된다.

$$H_{0j} : \mu_{A(-j)} = X_j \quad \text{vs.} \quad H_{1j} : \mu_{A(-j)} \neq X_j, \quad j = 1, \ldots, N,$$

여기서 $\mu_{A(-j)}$는 X_j를 제외한 조합으로 이루어진 집단의 모평균이다. 조합에 따라 시행하는 검정이 다르므로 각 조합에 따른 기각역과 유의수준을 다르게 설정해야 한다. 경험적 p-값 검정(empirical p-value approach)을 통해 이를 검정한다.

① 부트스트랩을 통해 $N(0,1)$을 따르는 N개의 표본을 500번 생성한다.

② 양분된 집단의 크기가 1이 아닌 $M-N$ 가지의 조합에 대해 두 집단의 모평균의 동일성 검정을 시행하여 얻은 $M-N$개의 p-값 중 최솟값을 p_{1b} $(b=1,\ldots,500)$라 한다.

③ 한 집단의 크기가 1인 N 가지의 조합에 대해 단일 모평균 검정을 시행하여 얻은 N개의 p-값 중 최솟값을 p_{2b} $(b=1,\ldots,500)$라 한다.

④ 이질성을 검정하고자 하는 데이터로부터 2, 3과 마찬가지의 반복 검정에 의해 계산된 p_{10}, p_{20}을 p_{1b}, p_{2b}과 비교하여 다음과 같은 경험적 p-값을 도출한다.

$$\text{경험적 } p\text{-값} = \min\left[\sum_{b=1}^{500} I(p_{10} \geq p_{1b}), \sum_{b=1}^{500} I(p_{20} \geq p_{2b})\right] /500$$

경험적 p-값을 유의수준 0.05와 비교하여 데이터의 이질성을 검정한다.

표본 X_i의 차원(p)가 2 이상인 경우 $N(0_p, I_p)$로부터 N개의 표본을 500번 생성한다. 이 때, μ_{Aj}와 μ_{Bj}는 벡터가 되며 이 경우 Hotelling's t-squared statistic을 이용하여 검정하였다.

2.3. Gap 통계량(gap statistic)

군집분석 방법에 제한되지 않고 최적의 군집의 개수를 찾는 방법으로서 Tibshirani 등 (2001)에 의해 제안된 gap 통계량은 군집 내 객체간의 거리를 최소화 하되, 군집의 개수를 과대추정 하지 않도록 패널티 항을 추가한 통계량이다.

m번째 군집 내에서 서로 다른 두 객체의 거리의 합은 다음과 같다:

$$D_m = \sum_{i,i' \in C_m} d_{ii'}$$

여기에서, C_m은 m번째 군집에 할당된 개체의 집합이고 n_m은 C_m의 객체 수를 나타낸다 $(m = 1,\ldots,k)$. $d_{ii'}$는 i번째 객체와 i'번째 객체간의 거리이다. 본 논문에서 $d_{ii'}$는 Euclidean distance를 사용하였다.

W_k는 k개의 분할에 따른 군집내의 동질성 측도로서 다음과 같이 정의한다:

$$W_k = \sum_{m=1}^{k} \frac{1}{2n_m} D_m$$

그러나 이 값은 k값이 증가함에 따라 반드시 감소하는 값이므로 최적의 k를 찾는 방법으로서는 부적절하다. 이에 대한 패널티 항으로 샘플의 크기가 n인 참조 분포(reference distribution)로부터 추정된 기댓값 $E_n^* \log(W_k)$를 고려한 것이 gap 통계량이다.

$$\text{Gap}_n(k) = E_n^* \log(W_k) - \log(W_k),$$

여기서 $E_n^* \log(W_k)$는 참조 분포로 이용된 균일분포에서 B번의 몬테카를로 샘플링을 통해 추정되며 이때의 표준편차를 $\mathrm{sd}(k)$로 정의한다. $E_n^* \log(W_k)$와 $\mathrm{sd}(k)$는 다음과 같이 계산된다:

$$E_n^* \log(W_k) = \frac{1}{B}\sum_{b=1}^{B} \log(W_{kb}),$$
$$\mathrm{sd}(k) = \frac{1}{B}\sqrt{\sum_{b=1}^{B} (\log(W_{kb}) - E_n^* \log(W_k))^2},$$

여기서 $E_n^* \log(W_k)$의 시뮬레이션 오차(simulation error) S_k는 다음과 같이 정의되며,

$$S_k = \mathrm{sd}(k)\sqrt{1 + \left(\frac{1}{B}\right)}$$

$\hat{k}$는 $\mathrm{Gap}_n(k) \geq \mathrm{Gap}_n(k+1) - S_{k+1}$을 만족하는 k의 최솟값으로 한다. 본 논문에서 군집분석 방법으로는 초기 배치(initial configuration)를 25번으로 하는 k-평균 군집화 방법을 이용하였다 (Hartigan과 Wong, 1979). 샘플링 횟수(B)는 500으로 하였다.

이와 같은 방법을 통해 $H_0 : k = 1$ vs. $H_1 : k > 1$을 검정하였으며 이것은 H_0 : homogeneous data vs. H_1 : heterogeneous data를 검정한 것과 같다. gap 통계량을 이용하여 찾은 $\hat{k}$ 값이 1이면 H_0을 채택하고 그렇지 않으면 H_1을 채택하였다. 통계량은 R 패키지 cluster (Maechler 등, 2018)의 clusGap 함수를 이용하여 계산하였다.

3. 모의실험 연구

3.1. 실험 계획

본 연구에서는 모의실험을 진행하기 위해 균일분포와 다변량 정규분포를 고려하였다. 표본의 개수(N)은 20으로 하였으며 표본의 차원(p)는 2로 하였다. 먼저 데이터의 이질성이 존재하지 않는 경우를 고려하기 위해 균일분포를 따르는 동일한 분포로부터 표본을 생성하였다.

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix},$$
$$X_{i1} \sim \mathrm{unif}(0,1), \quad i = 1, \ldots, N,$$
$$X_{i2} \sim \mathrm{unif}(0,1), \quad i = 1, \ldots, N.$$

마찬가지로 동일한 다변량 정규분포를 따르는 표본을 생성하였다.

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma\right), \quad i = 1, \ldots, N,$$
$$\Sigma = \begin{bmatrix} ,\sigma_1^2 & \sigma_1\sigma_1\rho \\ \sigma_1\sigma_1\rho & \sigma_2^2 \end{bmatrix}, \quad \sigma_1^2 = 1,\ \sigma_2^2 = 2,\ \rho = 0.5.$$

두 번째로, 데이터의 이질성이 존재하는 경우를 고려하기 위해서는 평균의 차이(d)가 존재하는 서로 다른 분포로부터 $N/2$개씩 표본을 생성하였다.

- 균일분포, $k = 2$, $10:10$

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix}$$
$$X_{i1}, X_{i2} \sim \begin{cases} \text{unif}(0,1), & i = 1, \ldots, 10, \\ \text{unif}(0+d, 1+d), & i = 11, \ldots, 20. \end{cases}$$

- 다변량 정규분포, $k = 2$, $10:10$

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \sim \begin{cases} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \right), & i = 1, \ldots, 10, \\ N\left(\begin{bmatrix} 0+d \\ 0+d \end{bmatrix}, \Sigma \right), & i = 11, \ldots, 20. \end{cases}$$

그러나 데이터 셋 내의 이질성이 존재하는 경우 모집단 별로 일정한 개수의 데이터가 존재한다는 가정은 상당히 강한 가정일 수 있다. 오히려 전체 데이터 셋 중에서 아주 일부의 데이터만이 이질성을 띠는 경우를 생각하기 쉽다. 따라서 세 번째로, $8:2$의 비율로 평균이 상이한 분포로부터 표본을 생성하였다.

- 균일분포, $k = 2$, $16:4$

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix},$$
$$X_{i1}, X_{i2} \sim \begin{cases} \text{unif}(0,1), & i = 1, \ldots, 16, \\ \text{unif}(0+d, 1+d), & i = 17, \ldots, 20. \end{cases}$$

- 다변량 정규분포, $k = 2$, $16:4$

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \sim \begin{cases} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \right), & i = 1, \ldots, 16, \\ N\left(\begin{bmatrix} 0+d \\ 0+d \end{bmatrix}, \Sigma \right), & i = 17, \ldots, 20. \end{cases}$$

위와 마찬가지로 서로 다른 모집단 별 데이터의 비율이 극단적으로 다른 경우를 고려하였다. 단 한 개의 표본만이 평균이 상이한 분포로부터 추출되는 데이터 셋을 생성하였다.

- 균일분포, $k = 2$, $19:1$

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix},$$
$$X_{i1}, X_{i2} \sim \begin{cases} \text{unif}(0,1), & i = 1, \ldots, 19, \\ \text{unif}(0+d, 1+d), & i = 20. \end{cases}$$

Table 3.1. The probability of making a type I error

	균일분포	다변량 정규분포
Gap 통계량	0.030	0.027
최소 조합 t-검정	0.217	0.037
최대 조합 t-검정	0.250	0.057

- 다변량 정규분포, $k = 2$, $19:1$

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \sim \begin{cases} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \right), & i = 1, \ldots, 19, \\ N\left(\begin{bmatrix} 0+d \\ 0+d \end{bmatrix}, \Sigma \right), & i = 20. \end{cases}$$

마지막으로 우리는 3개의 서로 다른 모집단으로부터 추출된 이질성을 갖는 자료를 생성하였다.

- 균일분포, $k = 3$, $7:6:7$

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix},$$

$$X_{i1},\ X_{i2} \sim \begin{cases} \text{unif}(0-d, 1+d), & i = 1, \ldots, 7, \\ \text{unif}(0, 1), & i = 8, \ldots, 13, \\ \text{unif}(0+d, 1+d), & i = 14, \ldots, 20. \end{cases}$$

- 다변량 정규분포, $k = 3$, $7:6:7$

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \sim \begin{cases} N\left(\begin{bmatrix} 0-d \\ 0-d \end{bmatrix}, \Sigma \right), & i = 1, \ldots, 7, \\ N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \right), & i = 8, \ldots, 13, \\ N\left(\begin{bmatrix} 0+d \\ 0+d \end{bmatrix}, \Sigma \right), & i = 14, \ldots, 20. \end{cases}$$

우리는 다음과 같은 기준에 따라 최소 조합 t-검정, 최대 조합 t-검정, gap 통계량을 이용한 검정의 성능을 확인하였다: (1) 귀무가설이 참일 경우의 제 1종의 오류를 범할 확률(type I error rate); (2) 대립가설이 참일 경우의 검정력(power). 각 모의실험의 시나리오별로 자료를 300번 생성시켜서 그 중에서 귀무가설을 몇 번 기각시키는 지를 센 후 제 1종의 오류를 범할 확률 또는 검정력을 구하고, 그 과정을 다시 300번 반복하여서 그 평균값을 구하였다. 이를 식으로 표현하면 다음과 같다:

$$\text{제 1종의 오류를 범할 확률(또는 검정력)} = \frac{1}{300} \sum_{s=1}^{300} \text{귀무가설을 기각한 횟수}_s.$$

3.2. 결과

Table 3.1을 통해 제 1종의 오류를 범할 확률을 기준으로 gap 통계량과 최소 조합 t-검정, 최대 조합 t-검정의 성능을 비교해보자. 다변량 정규분포에서 생성된 데이터의 경우 세 방법은 유사한 성능을 보였

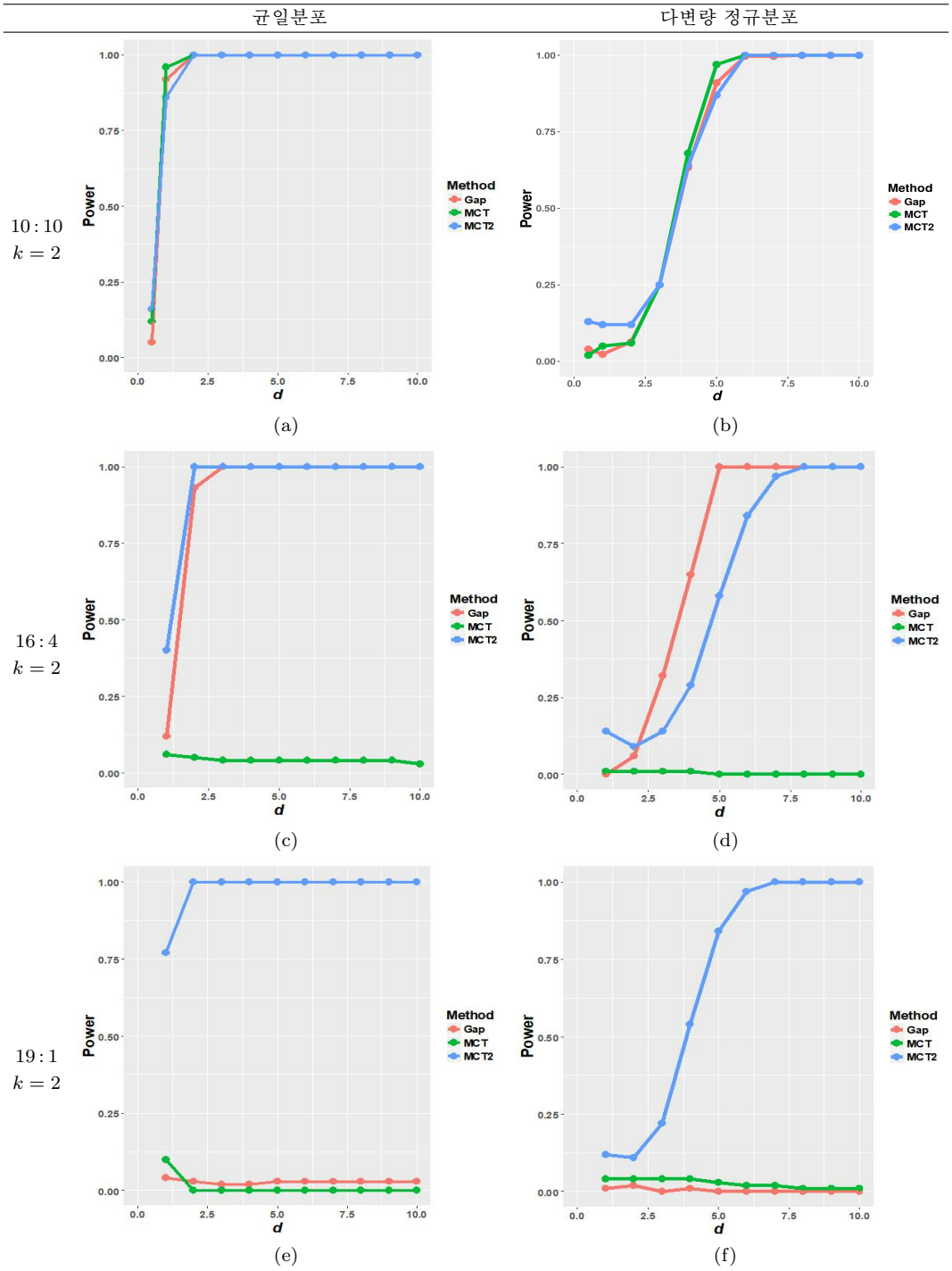
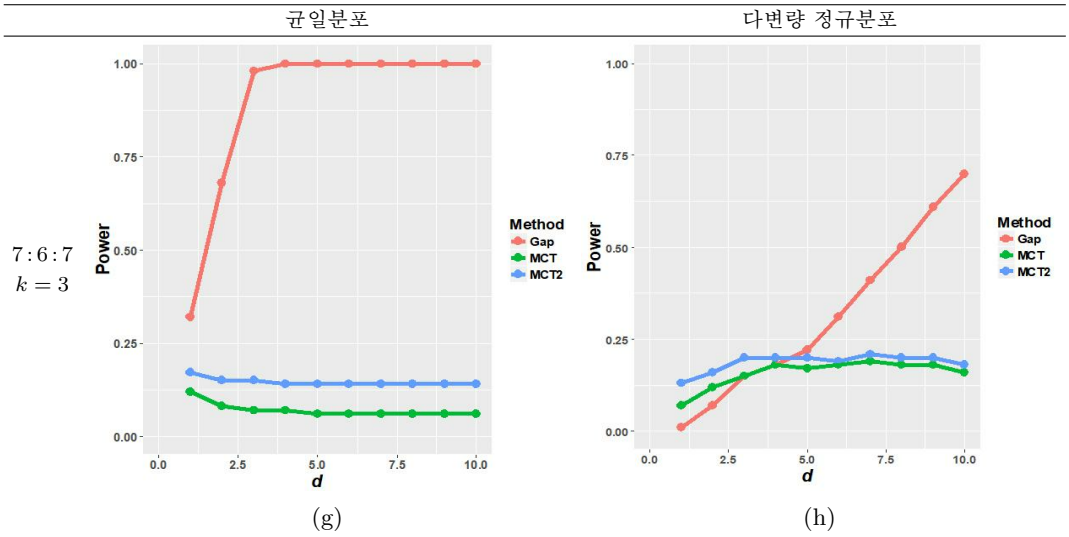


Figure 3.1. Power across differences between means ($k = 2$). red: Gap statistic; green: minimum combination t-test; blue: maximum combination t-test.



(g) (h)

Figure 3.2. Power across differences between means ($k = 3$). red: Gap statistic; green: minimum combination t-test; blue: maximum combination t-test.

으나 모집단이 균일분포를 따르는 경우 최소 조합 t-검정과 최대 조합 t-검정은 낮은 성능을 보였다. 이는 두 검정이 모집단의 분포를 다변량 정규분포로 가정하는 Hotelling's t-squared statistic를 이용함에서 기인한 것으로 보인다.

Figure 3.1과 Figure 3.2는 데이터의 이질성이 존재하는 경우 평균의 차이에 따른 검정력을 그래프로 나타낸 것이다. X축은 모집단 평균의 차이(d)를, Y축은 검정력을 나타낸다. 빨간선은 gap 통계량의 검정력을, 초록선은 최소 조합 t-검정의 검정력을, 파란선은 최대 조합 t-검정의 검정력을 의미하며 각각 Gap, MCT, MCT2로 나타내었다. 이질 데이터의 비율이 유사할 경우에는 모집단이 균일분포를 따를 때 세 방법론 간의 유의미한 성능 차이는 보이지 않았다: (a). 다변량 정규분포의 경우 평균의 차이가 미세한 구간에서 최대 조합 t-검정의 검정력이 다소 우세한 것으로 나타났다: (b). 그러나 이질 데이터의 비율이 다른 경우 최소 조합 t-검정의 경우 좋지 않은 검정력을 갖는 것으로 해석된다: (c), (d), (e), (f). 최소 조합 t-검정은 데이터를 크기가 비슷한 임의의 두 집단으로 나누어 검정하기 때문에 이질적인 분포 별로 데이터의 비율이 상이한 경우 데이터의 이질성을 감지하지 못하는 것으로 해석된다. 반면에 최대 조합 t-검정은 분포 별 데이터의 비율에 의존하지 않는 성능을 보였다. 특히 데이터의 비율이 $N-1$ 대 1인 경우 다른 두 방법은 0에 가까운 검정력을 보인 반면, 유일하게 최대 조합 t-검정만이 평균의 차이(d)가 증가함에 따라 검정력이 빠르게 1로 수렴하는 성능을 보였다. 마지막으로 3개의 서로 다른 모집단으로부터 추출된 이질성을 갖는 데이터를 통해 모의 실험한 결과는 (g), (h)이다. 균일분포의 경우 gap 통계량의 방법을 이용한 검정을 통해 높은 검정력을 얻을 수 있음을 보여준다: (g). 모집단이 다변량 정규분포를 따르는 경우, 평균의 차이가 작은 경우 최대 조합 t-검정이 상대적으로 우수한 검정력을 보였으나 평균의 차이가 증가함에 따라 gap 통계량이 보다 높은 검정력을 갖는다는 것을 알 수 있다: (h).

귀무가설이 참일 경우의 제 1종의 오류를 범할 확률과 대립가설이 참일 경우의 검정력(power)을 기준에 따라 세 가지 방법의 성능을 확인한 결과 모든 경우에서 최상의 성능을 갖는 검정 방법은 존재하지 않는 것으로 나타났다. gap 통계량을 이용한 이질성 검정은 균일분포와 다변량 정규분포 모두에서 0.05

Table 4.1. Results for the real data set using Gap statistic and Minimum combination t-test

Patient No.	No. of Nodules	Gap statistic		Minimum combination t-test			
		$\hat{k}$	$\mathrm{Gap}_n(k)$	$\min(p)$	m	$0.05/m$	Overall heterogeneity
1	27	1	0.202922	0.000781	50000	0.000001	No
2	8	2	0.693630	0.090633	70	0.000714	No
3	52	1	0.706726	0.002818	50000	0.000001	No
4	7	1	−0.048770	0.067242	35	0.001429	No
5	25	2	1.253652	0.040986	50000	0.000001	No
6	8	1	0.094498	0.032706	70	0.000714	No
7	5	2	−0.019187	0.109719	10	0.005000	No
8	10	4	0.524047	0.011917	252	0.000198	No
9	9	2	1.101800	0.005535	126	0.000397	No
10	4	1	−0.495453	0.469485	6	0.008333	No

Table 4.2. Results for the real data set using maximum combination t-test

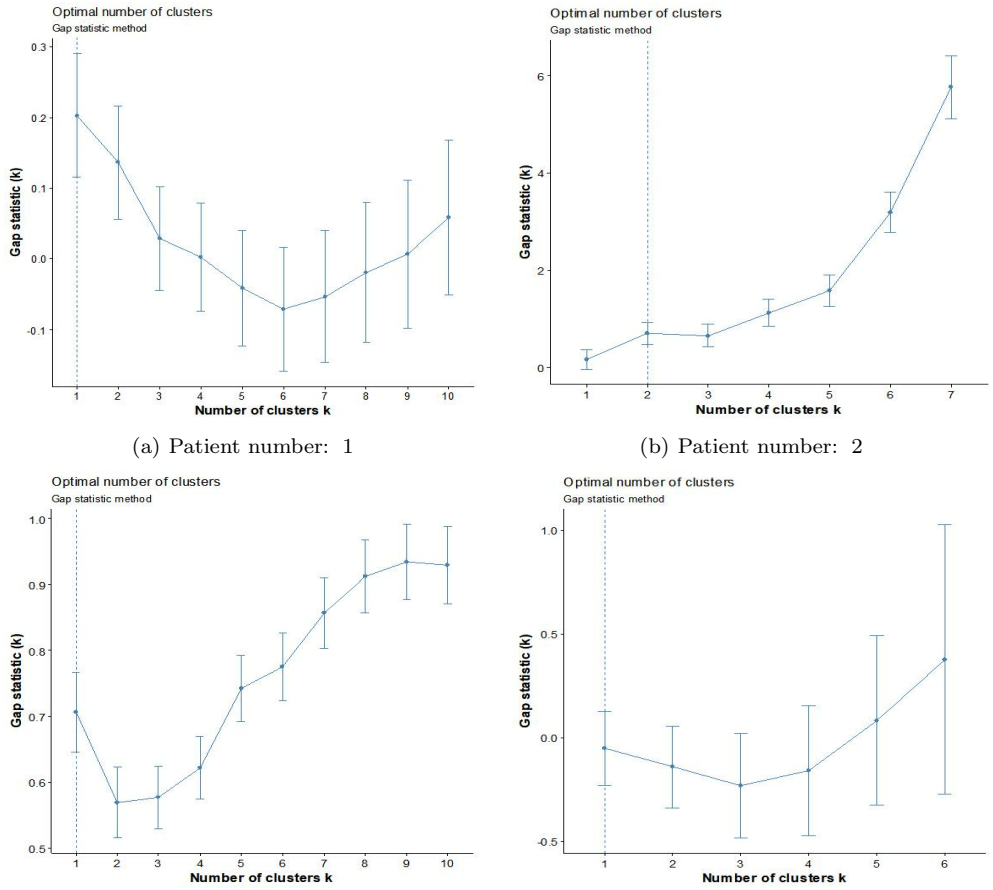
Patient No.	No. of Nodules	Maximum combination t-test				
		p_{10}	p_{20}	Empirical p-value	M	Overall heterogeneity
1	27	2.66603×10^{-6}	2.894351×10^{-13}	0.010	501492	Yes
2	8	0.012468800	3.638633×10^{-7}	0.002	162	Yes
3	52	5.090373×10^{-13}	$< 2.2 \times 10^{-16}$	< 0.002	1173478	Yes
4	7	0.014620430	0.000672962	0.142	98	No
5	25	4.037892×10^{-11}	$< 2.2 \times 10^{-16}$	< 0.002	456661	Yes
6	8	0.008527597	6.156917×10^{-5}	0.070	162	No
7	5	0.109719300	0.135352200	< 0.002	15	Yes
8	10	0.000809845	5.661957×10^{-5}	0.330	637	No
9	9	0.004716193	5.506107×10^{-10}	< 0.002	381	Yes
10	4	0.469484800	0.094351160	0.636	7	No

보다 작은 제 1종의 오류를 범할 확률을 보였으며 3개의 모집단을 갖는 경우에도 로버스트한 검정력을 보였다. 그러나 분포 별 데이터 비율의 차이가 극단적인 경우에 이질성을 감지하지 못하는 단점을 갖는다. 최대 조합 t-검정은 데이터를 2분할 하는 모든 가능한 조합을 고려하는 검정의 특성상 분포 별 데이터의 비율이 상이할 때 이에 의존하지 않는 로버스트한 검정력을 갖는다는 것을 확인하였다. 특히 데이터의 비율이 극단적일수록 다른 두 방법보다 우수한 성능을 보였으며, 최소 조합 t-검정에 비해 모든 경우에서 보다 우수한 성능을 갖는 다는 것이 확인되었다. 단, 서로 다른 3개의 모집단에서 추출된 데이터의 이질성을 감지하지 못하는 한계가 있다.

4. 실제자료 예시

이 논문에서 제안하는 방법을 적용하여 종양 이질성을 파악하기 위해, 먼저 항암치료 전과 후로 환자들의 두 번의 방문을 통해 얻어진 관측값들을 이용하여 변화율을 계산한다. 관측값은 흉부 CT를 통해 측정한 종양의 최대지름과 부피값이다. 각 측정값 각각 대해 변화율은 다음과 같이 계산된다:

$$\text{변화율} = \frac{V2 - V1}{V1} \times 100(\%),$$



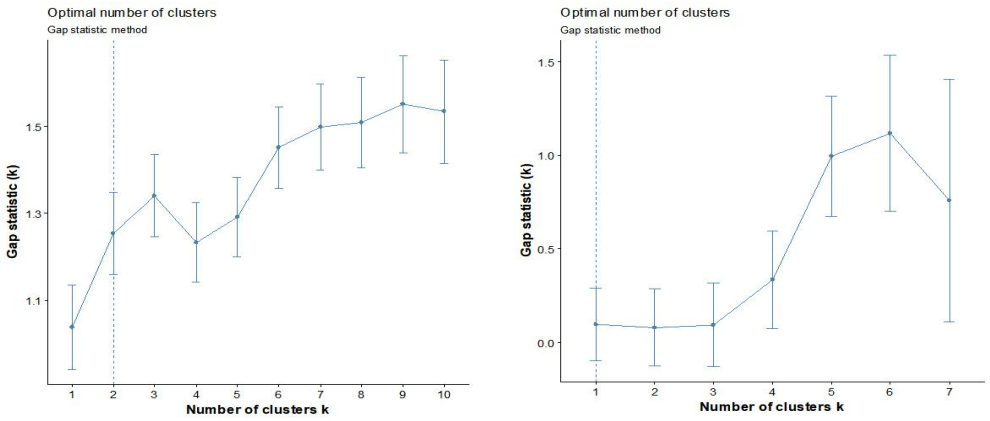
(a) Patient number: 1 (b) Patient number: 2

(c) Patient number: 3 (d) Patient number: 4

Figure 4.1. Gap statistic across the numbers of cluster (k), patient number: 1–4.

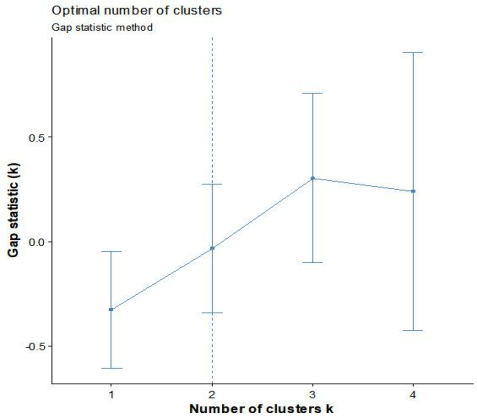
여기서 $V1, V2$는 각각 첫 번째 방문 시 관측값, 두 번째 방문 시 관측값에 해당된다. 이는 각 환자가 지닌 종양의 개수만큼 계산된다. Heo와 Lim (2017)의 연구에서 이용된 두 변수인 최대지름과 부피값에 본 논문에서 제안하는 세 가지 방법을 적용하여 얻은 결과는 Table 4.1을 보라. 여기서 7번 환자의 종양 데이터에 대한 최대 조합 t-검정은 일반적인 Hotelling's t-squared statistic을 이용할 수 없는 조합이 존재한다. 환자의 1, 2, 3, 5번 종양의 분산행렬이 비정칙 행렬이기 때문이다. 따라서 우리는 다변량 비모수 검정 을 이용하여 경험적 p-값을 도출하였다.

Table 4.1의 $\hat{k}$는 gap 통계량을 이용한 분석 결과 선택된 군집의 개수를 나타낸 것이다. 이 값이 1보다 크다면 종양 이질성이 존재함을 의미한다. gap 통계량을 이용한 분석 결과에 따르면 5명의 환자(2, 5, 7, 8, 9번)가 지닌 종양에 대하여 전체적인 이질성이 존재한다고 할 수 있다. 반면 최소 조합 t-검정을 이용한 결과에 따르면 종양 이질성을 검정한 결과 모든 환자에서 가능한 m개의 조합에 대한 p-value의 최솟값이 $0.005/m$보다 크기 때문에 이질성이 존재한다고 볼 수 없다. Table 4.2의 최대 조합 t-검정을 이용하여 종양이질성을 검정한 결과는 앞서 제시한 두 방법과 다른 결과를 보여준다. 최대 조합 t-검정 결과 5명의 환자(1, 2, 3, 5, 7, 9번)의 종양에서 이질성이 존재한다고 할 수 있다. 즉, 4명의 환자(2, 5,

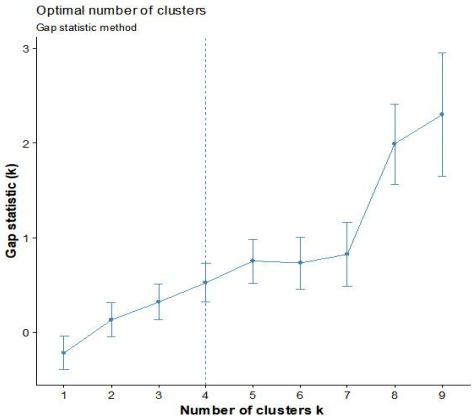


(a) Patient number: 5

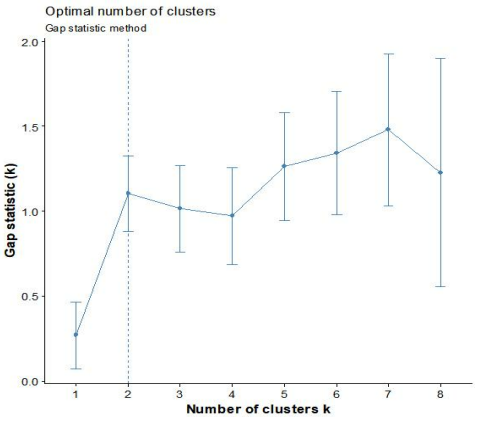
(b) Patient number: 6



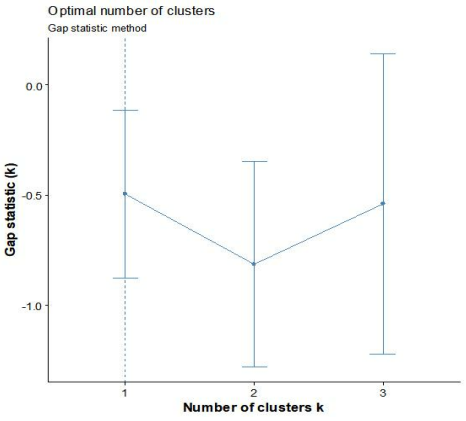
(c) Patient number: 7



(d) Patient number: 8



(e) Patient number: 9



(f) Patient number: 10

Figure 4.2. Gap statistic across the numbers of cluster (k), patient number: 5–10.

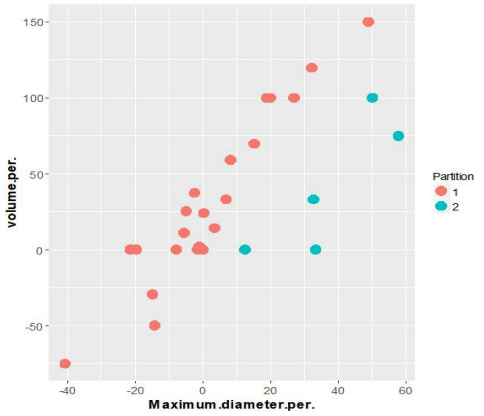
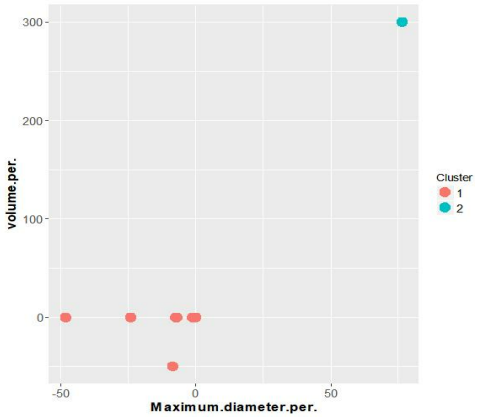
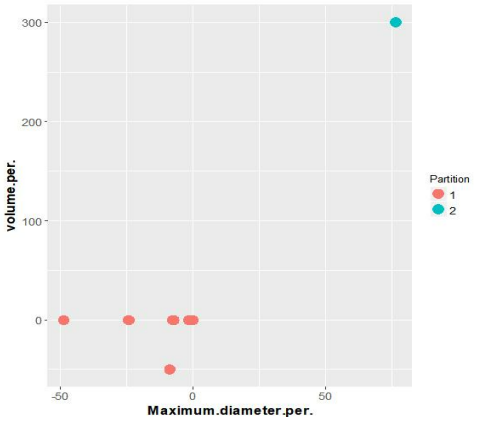
Patient No.	Gap 통계량에 따른 군집	최대 조합 t-검정에 따른 분할
No. 1	$H_0 : k = 1$채택	
No. 2		

Figure 4.3. Results for the real data set using gap statistic and maximum combination t-test, patient number: 1, 2.

7, 9번)는 gap 통계량과 최대 조합 t-검정을 이용했을 때 종양 이질성이 존재하는 것으로 판단되며, 8번 환자는 gap 통계량결과에서만, 1, 3번 환자는 최대 조합 t-검정 결과에서만 이질성이 존재하는 것으로 분석되었다.

Figure 4.1과 Figure 4.2는 10명의 환자의 종양 데이터에 대한 gap 통계량에 따른 분석 결과를 그래프로 나타낸 것이다. X축은 군집의 개수를, Y축은 이에 따른 gap 통계량 값을 나타낸 것이다. 수직 점선은 기준($\mathrm{Gap}_n(k) \geq \mathrm{Gap}_n(k+1) - S_{k+1}$을 만족하는 k의 최솟값)에 따라 선택된 군집 수를 의미한다.

Figure 4.3 부터 Figure 4.5는 gap 통계량과 최대 조합 t-검정에 따른 분석 결과를 나타낸 것이다. gap 통계량을 통해 결정한 최적의 k값에 따른 군집 분석과 데이터 분류 그래프를 왼쪽에 나타냈다. 오른쪽의 그래프는 최대 조합 t-검정결과 가장 작은 경험적 p-값을 갖는 조합을 나타낸 것이다. X축은 최대 직경의 변화량을, Y축은 부피의 변화량을 의미한다.

gap 통계량과 최대 조합 t-검정을 이용하여 전이성 종양의 성장패턴 차이와 변화율에 따른 종양 이질성을 검정한 결과, 3명의 환자의 종양 데이터에서 서로 다른 결과를 얻었다. 이중 3번 환자의 데이터의 경

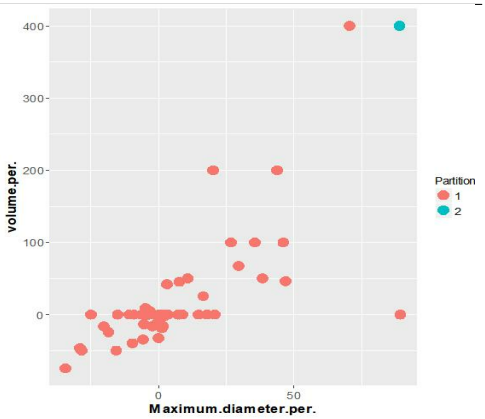
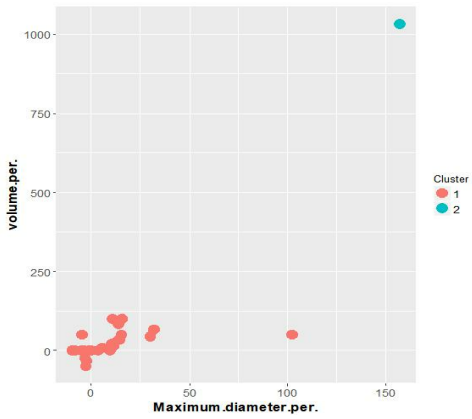
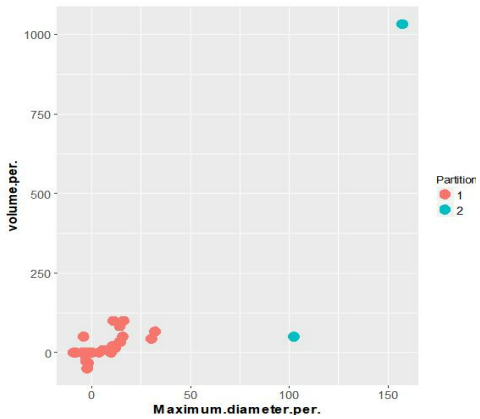
Patient No.	Gap 통계량에 따른 군집	최대 조합 t-검정에 따른 분할
No. 3	$H_0 : k = 1$채택	
No. 5		

Figure 4.4. Results for the real data set using gap statistic and maximum combination t-test, patient number: 3, 5.

우 최대 조합 t-검정에 따르면 종양 이질성을 띄는 표본이 1개 존재하지만 gap 통계량을 이용한 방법에서는 이를 감지하지 못하였다. 우리는 3장의 모의실험으로부터 서로 다른 모집단으로부터 추출된 표본의 비율이 크게 다를 때 최대 조합 t-검정은 검정력은 평균차이가 증가함에 따라 빠르게 1에 가까워지는 반면 gap 통계량을 이용한 방법에서는 상당히 낮은 검정력을 갖는 다는 사실을 확인한 바 있다. 3번 환자의 데이터는 이러한 특성에 따른 것으로 해석된다. 한 편, 8번 환자의 데이터의 경우 gap 통계량을 이용했을 때 최적의 군집개수는 4개로 선택된 반면 최대 조합 t-검정결과 이질성을 감지하지 못하였다. 이러한 결과는 이질성을 갖는 데이터들의 모집단의 개수가 2보다 큰 경우에 최대 조합 t-검정의 낮은 검정력에 기인한 것일 수 있다.

5. 결론

본 연구에서는 전이성 종양의 성장패턴 차이와 변화율에 따른 종양 이질성을 파악하기 위한 최적의 통계적 방법을 연구하고 분석 결과를 도출하였다. 기존에 연구된 검정 방법을 개선하여 소수의 데이터가 이

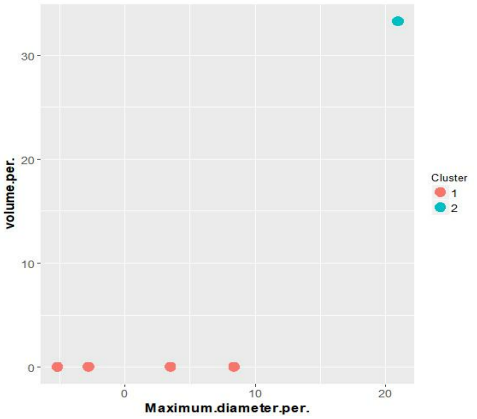
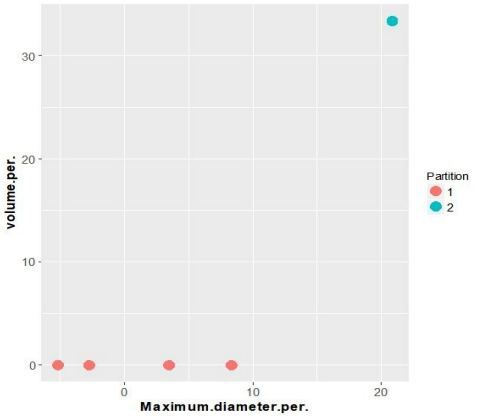
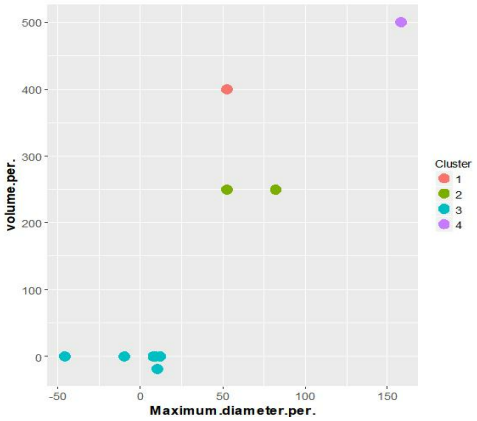
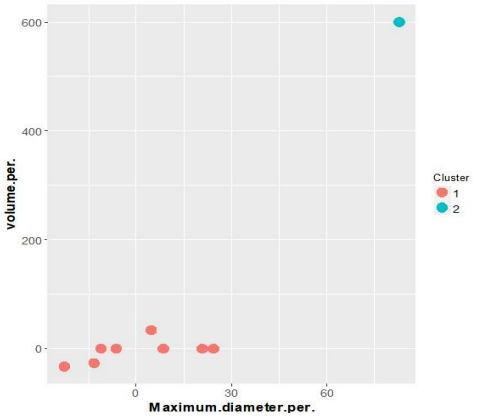
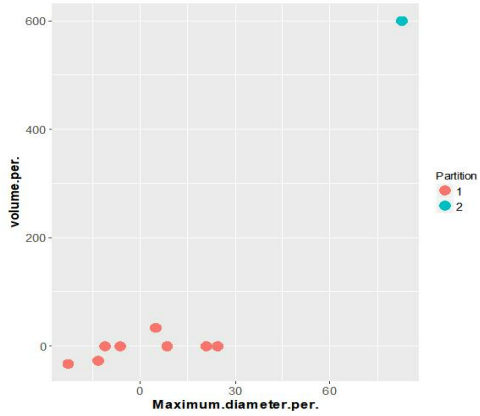
Patient No.	Gap 통계량에 따른 군집	최대 조합 t-검정에 따른 분할
No. 7		
No. 8		$H_{0j} : \mu_{Aj} = \mu_{Bj},\ j = 1, \ldots, m$ 및 $H_{0j} : \mu_{A(-j)} = Xj,\ j = 1, \ldots, N$ 채택
No. 9		

Figure 4.5. Results for the real data set using gap statistic and maximum combination t-test, patient number: 7–9.

질성을 갖는 경우의 성능이 대폭 향상된 최대 조합 t-검정 방법을 제안하였다. 한편, 이질성을 띠는 데이터가 데이터의 이질성 검정을 군집분석에서 군집 개수를 결정하는 문제로 치환 한 후 k-평균 군집화에서 gap 통계량을 기준으로 군집의 개수를 결정하는 방법을 제안하였다. 이와 함께 이전에 제안한 최소 조합 t-검정과 본 논문에서 제안하는 최대 조합 t-검정, gap 통계량을 이용한 방법의 성능을 모의실험을 통해 비교한 결과, 최대 조합 t-검정이 서로 다른 모집단 별로 추출된 데이터의 비율에 의존하지 않는 로버스트한 방법임을 확인하였다. 한편 gap 통계량을 이용한 방법이 모집단의 개수에 의존하지 않는 검정력을 보여주었다. 이 두 방법은 서로 다른 모집단으로부터 추출된 표본의 비율이 상이하거나 모집단의 개수가 2보다 클 때 최소 조합 t-검정의 검정력이 지나치게 낮아지는 단점을 해결하였다.

이러한 특징은 실제 데이터 분석 결과에서도 확인되었다. 최소 조합 t-검정 분석 결과 종양 이질성이 존재하지 않는 것으로 결론 내렸다. 그러나 이는 이질성을 갖는 표본의 수가 전체 표본에서 소수일 때 혹은 모집단의 개수가 2개 이상일 때 발생하는 오류임을 군집을 구분한 산점도에서 확인할 수 있다. 최소 조합 t-검정 분석 결과와는 달리 gap 통계량을 이용했을 때 10명의 환자 중 5명이, 최대 조합 t-검정을 이용했을 때 6명이 종양이 이질성을 갖는 것으로 분석되었다. 한 편, gap 통계량을 이용한 방법과 최대 조합 t-검정의 서로 다른 결과는 모집단 별 표본의 비율의 차이가 극심한 경우 gap 통계량의 한계와 모집단의 개수가 3개 이상일 때 낮은 검정력을 갖는 최대 조합 t-검정의 특성에 기인한 것으로 해석된다.

Gap 통계량은 통계적 가설검정 방법이 아닌 군집의 개수를 추정하는 하나의 통계량이라고 할 수 있다. 본 논문에서 제안된 방법은 gap 통계량을 통해 주어진 가설에 대해 검정통계량의 분포에 의한 p-값을 계산하지 않는다. 단순히 최솟값을 갖는 gap 통계량에 해당하는 군집의 개수를 정함으로써 귀무가설 또는 대립가설을 선택하는 것은 엄밀한 의미에서 보면 통계적인 가설검정이 아니다. 하지만 Yan과 Ye (2007)은 gap 통계량에 대하여 "it can be used to test the null hypothesis about homogeneous nonclustered data against the alternative of clustered data"라고 언급하였다. 따라서 gap 통계량이 p-값을 주지는 않지만 여전히 가설검정에 사용할 수 있다고 할 수 있다. 또한 귀무가설이 참일 때 이를 기각하게 되는 확률을 모의실험을 통해 추정할 수 있기 때문에 제 1종의 오류율이라는 표현 역시 사용할 수 있다고 할 수 있다. 이에 관해서는 추후에 더 심도있는 연구와 논의가 필요할 것이다.

본 연구는 데이터를 양분하는 가능한 모든 경우를 고려하여 데이터의 이질성을 검정하는 방법을 제안하였다. 검정의 성능은 대폭 향상된 반면, 데이터를 분석하는데 필요한 시간이 크다는 단점을 갖는다. 가능한 조합의 수를 감소시켜 분석 시간을 대폭 감소시키기 위해 표본간의 거리를 고려하여 상이한 표본쌍을 서로 다른 집단으로 분할하는 알고리즘을 고려할 수 있을 것이다. 데이터 처리에 필요한 시간문제가 해결된다면 조합할 수 있는 만큼 세 집단 이상으로 나누어 평균 비교에 대한 검정을 실시할 수 있다. 여기서 t-검정이 아니라 등분산 가정이나 정규성 가정에 따른 일원배치 분산분석이나 Kruskal-Wallis 방법 (Kruskal과 Wallis, 1952)을 고려할 수 있을 것이다. 또한 본 연구는 데이터의 이질성 검정을 군집분석에서 군집의 개수를 결정하는 것으로써 분석하였다. 그러나 본 연구에서 제안한 k-평균 군집화 방법과 gap 통계량 외에도 데이터셋 내에서 최적의 군집의 개수를 결정하기 위한 다양한 방법들이 존재한다. 데이터의 표본 개수, 다양한 모집단 분포, 모집단 별 표본의 비율, 그리고 모집단의 개수를 고려한다면 어떤 통계적인 방법이 $H_0 : k = 1$ vs. $H_1 : k > 1$을 검정하는 최적의 방법일지에 대한 연구 가능성은 무궁무진하다. 이들 방법들을 고려하여 표본간의 이질성을 분석할 수 있는 다양한 연구가 진행되어야 할 것이다.

References

Baker, F. B. and Hubert L. J. (1976). A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering, *Journal of the American Statistical Association*, **71**, 870–878.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, 224–227.

Dunn, J. C. (1974). Well-separated clusters and optimal Fuzzy partitions, *Journal of Cybernetics*, **4**, 95–104.

Dunn, O. J. (1961). Multiple comparisons among means, *Journal of the American Statistical Association*, **56**, 52–64.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transac-tions of the Royal Society of Edinburgh*, **52**, 399–433.

Hartigan, J. and Wong, M. (1979). Algorithm AS 136: A K-means clustering algorithm, *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **28**, 100–108.

Heo, M. and Lim, C. (2017). A minimum combination t-test method for testing differences in population means based on a group of samples of size one, *The Korean Journal of Applied Statistics*, **30**, 301–309.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, **47**, 583–621.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2018). Cluster: Cluster analysis basics and extensions, R package version 2.0.7-1.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Student (1908). The probable error of a mean, *Biometrika*, **6**, 1–25.

Thorndike, R. L. (1953). Who belongs in the family?, *Psychometrika*, **18**, 267.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic, *Journal of the Royal Statistical Society*, **63**, 411–423.

Yan, M. and Ye, K. (2007). Determining the number of clusters using the weighted gap statistic, *Biometrics*, **63**, 1031–1037.

Yoo, J., Kim, Y., Lim, C., Heo, M., Hwang, I., and Chong, S. (2017). *Assessment of Spatial Tumor Heterogeneity using CT Phenotypic Features Estimated by Semi-Automated 3D CT Volumetry of Multiple Pulmonary Metastatic Nodules: A Preliminary Study*, unpublished manuscript.

종양 이질성을 검정을 위한 통계적 방법론 연구

이동녘[a] · 임창원[a,1]

[a]중앙대학교 응용통계학과

(2018년 10월 25일 접수, 2018년 12월 8일 수정, 2019년 1월 3일 채택)

요 약

전이성 종양의 성장패턴 차이와 변화율에 따른 종양 이질성(tumor heterogeneity)을 파악하는 것은 종양세포의 약물에 대한 민감성을 파악하고 적절한 치료법을 찾아내기 위해 중요하다. 일반적으로 N개의 표본의 집단이 구분된다면 t-test 혹은 ANOVA 분석을 통해 집단별 평균의 차이에 대한 검정이 가능하다. 그러나 본 논문에서 다루는 데이터와 같이 집단이 구분되지 않는 경우 이러한 방법들은 사용될 수 없다. 표본들 사이의 이질성을 검정하기 위한 통계적 방법들이 연구되어 왔다. 최소 조합 t-검정 방법은 그 중 하나이다. 본 논문에서는 상이한 비율로 데이터를 양분하는 조합도 고려하는 최대 조합 t-검정 방법을 제안한다. 한편, 표본의 이질성을 검정하는 것이 군집분석에서 최적의 군집의 개수가 2개 이상인지를 검정하는 것과 같음에 착안하여 새로운 방법을 제안한다. 최대 조합 t-검정과 gap통계량을 이용하면 이전에 제안된 방법보다 개선된 제1종의 오류를 범할 확률과 검정력을 갖는다는 것을 모의실험을 통해 확인하였고 실제 자료 분석을 통해 결과를 도출하였다.

Keywords: 이질성, k-평균 군집화, gap 통계량, 최소 조합 t-검정, 최대 조합 t-검정, 최적 군집 개수

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: clim@cau.ac.kr