응용통계연구 (2011)
24(6), 1169–1179
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.6.1169

사회연결망분석을 이용한 확률분포들의 이용빈도 구조에 대한 연구

장대흥[1] · 이성백[2]

[1]부경대학교 통계학과, [2]부경대학교 통계학과

(2011년 7월 접수, 2011년 9월 채택)

요 약

본 논문에서 포털사이트의 정보를 이용한 사회연결망분석을 통하여 통계학 책에서 주로 언급되는 확률분포들의 종류와 쓰임새에 대한 설명이 일상생활에서 언급되는 확률분포들과 어떤 관계가 있는 지 알아본다. 이를 통하여 우리들의 일상생활을 염두에 둘 때 통계학 책에서 강조하여야 할 확률분포들에 대하여 알아본다.

주요용어: 확률분포, 포털사이트, 사회연결망분석.

1. 서론

사회연결망분석(SNA; social network analysis)은 노드와 연결선으로 구성되는 네트워크 이론을 사용하여 사회연결망을 연구하는 분야이다. 사회연결망에 대한 역사는 1800년대까지 거슬러 올라간다. 사회연결망분석이 대중들에게 전세계적으로 널리 알려지게 된 것은 페이스북과 같은 소셜네트워크서비스(SNS; social network service)들의 성공때문이라고 저자는 생각한다. 국내외 각 학문분야에서 사회연결망분석에 관련된 논문들은 시간이 갈수록 급증하고 있다(예로, Sternitzke 등 (2008), Brandes 등 (2011), Gregson 등 (2011), Sailer와 McCulloh (2011) 등이 있다). 인문사회과학자의 입장에서 저술한 국내 사회연결망분석 관련 저서로서 손동원 (2002)과 김용학 (2004)이 있다. 허명회 (2010)는 R을 활용한 사회연결망분석에 관한 책을 저술하였다.

본 논문에서는 이러한 사회연결망분석의 한 응용으로서 통계학 책에 주로 언급되는 확률분포들의 종류와 쓰임새에 대한 설명이 일상생활에서 언급되는 확률분포들과 어떤 관계가 있는 지 포털사이트에서의 정보를 이용한 사회연결망분석을 통하여 알아보고 우리들의 일상생활을 염두에 둘 때 통계학 책에서 강조하여야 할 확률분포들의 종류와 쓰임새에 대하여 알아보고자 한다. 포털사이트로서는 구글(2011년 7월 23일 현재)을 사용하였고 총 23개의 확률분포들을 분석대상으로 하였다. 통계패키지 R을 사용(구체적으로는 library(sna)를 사용함)하여 사회연결망분석을 행하였다. 2절에서는 사회연결망분석을 이용한 확률분포들의 이용빈도에 대하여 언급하였고 3절에서는 결론을 내렸다.

2. 사회연결망분석을 이용한 확률분포들의 이용빈도 구조

고등학교 수학과 교육과정에서는 확률분포로서 이항분포와 정규분포를 언급하고 대학생을 대상으로하는 기초통계학 교재에서는 이산확률분포로서 주로 초기하분포, 베르누이분포, 이항분포, 포아송분포를

[1]교신저자: (608-737) 부산광역시 남구 대연3동 599-1, 부경대학교 통계학과, 교수. E-mail: dhjang@pknu.ac.kr

다루고 연속확률분포로서 주로 정규분포, 표집분포로서 카이제곱분포, t분포, F분포를 다룬다. 이러한 분포들의 중심에는 정규분포가 있다. 각종 이산적 확률분포들은 적절한 조건이 만족되면 정규근사화함으로 통계학에서는 정규분포를 중요하게 다루게 된다. Leemis와 McQueston (2008)은 76개의 일변량 확률분포들(이산확률분포: 19개, 연속확률분포: 57개)의 관계에 대하여 하나의 그림으로 도표화하여 제시하였다. 그런데 본 저자가 20여년간 기초통계학 내용 중 확률분포에 대한 교수 가운데 계속 품어왔던 의문은 다음과 같다.

"통계학 책에 주로 언급되는 확률분포들이 실제 우리들의 일상생활에서도 중요하게 언급되는가?"

이 문제를 풀어보기 위한 하나의 시도로서 포털사이트를 이용하여 통계분포를 언급한 웹페이지수(앞으로 '웹페이지수'라고 언급하겠다)를 알아 보고, 이 정보를 사회연결망분석에 사용하여 우리들의 일상생활에서 나타나는 확률분포들의 이용빈도 구조에 대하여 알아보았다. 사회연결망분석에서 노드로서는 통상 개인, 기관, 회사, 웹페이지 주소 등이 대상이 되나 Racherla와 Hu (2010) 연구논문에서처럼 추상적 개념이 노드의 대상이 될 수 있다. Racherla와 Hu (2010) 연구논문에서는 관광학 관련 공동연구 9개의 주제(1. Segmentation Studies, 2. Marketing & Strategy, 3. Information & Communication Technology, 4. Sustainable/Eco/Green/Alternative Tourism, 5. Industry/Performance & Impact Studies, 6. Forecasting & Community Studies, 7. Hospitality & Gaming, 8. Human Resources, Training, Education & Research, 9. Travel Industry/Airlines/Others)들을 노드로 삼아 사회연결망분석을 행하였다. 마찬가지로 통계 개념인 확률분포들을 노드로 하여 사회연결망분석을 행함으로 우리들의 일상생활에서 나타나는 확률분포들의 이용빈도 구조에 대하여 알아볼 수 있다.

포털사이트로서는 구글(2011년 7월 23일 현재)을 사용하였고 총 23개의 확률분포들(1. 정규분포, 2. 지수분포, 3. 감마분포, 4. 와이블분포, 5. 코시분포, 6. t분포, 7. F분포, 8. 카이제곱분포, 9. 베타분포, 10. 균일분포, 11. 삼각형분포, 12. 로그정규분포, 13. 로지스틱분포, 14. 극단값분포, 15. 파레토분포, 16. 이중지수분포, 17. 베르누이분포, 18. 이항분포, 19. 초기하분포, 20. 포아송분포, 21. 기하분포, 22. 음이항분포, 23. 제타분포)을 분석대상으로 하였다. 일상생활이나 전반적인 학문분야에서 많이 쓰이는 확률분포들을 분석대상 확률분포로 정하였다. 포털사이트 구글을 사용하여 23개의 확률분포 각각의 웹페이지수와 확률분포 두 개씩의 동시웹페이지수를 구하였다. 서로 비교하기 위하여 한글용어와 영문용어로 나누어 조사하였다.

2.1. 한글용어를 이용한 사회연결망분석

한글용어를 이용하여 23개의 확률분포 각각의 웹페이지수와 확률분포 두 개씩의 동시웹페이지수를 구하면 이용빈도행렬을 구할 수 있다. 이러한 이용빈도행렬에서 대각선 원소는 대응되는 확률분포의 웹페이지수로 구성되고 비대각선 원소는 대응되는 두 개의 확률분포의 동시웹페이지수로 구성된다. 우리는 이렇게 구한 이용빈도행렬을 이용하여 확률분포들의 이용빈도 네트워크를 다음 그림 2.1과 같이 그릴 수 있다. 이 그림은 R library(sna)에 있는 gplot을 이용하여 그렸다. 이러한 네트워크는 무방향 네트워크이어서 가중인접행렬이 대칭행렬이 된다. 관계의 시작과 끝에 방향이 있으면 방향 네트워크가 되고 관계의 시작과 끝에 방향이 없고 단지 관계의 존재여부에 대한 정보만 있으면 무방향 네트워크가 된다. 그림 2.1에서 각 노드의 크기는 해당 확률분포의 웹페이지수에 비례하고 각 연결선의 굵기는 두 개의 해당 확률분포들의 동시웹페이지수에 비례한다. 우리의 예상과는 달리 정규분포가 네트워크 중심에 있지 않음을 알 수 있다. 오히려 t분포, F분포, 기하분포, 지수분포 등이 네트워크 중심에 위치함을 알 수 있다. 특이한 것은 이항분포는 노드의 크기(웹페이지수)는 작으나 네트워크 중심 가까이에 위치하여 있어 다른 확률분포들 사이를 연결하는 중요한 자리에 있음을 알 수 있다. 이산분포로 기초통계학책에 자주

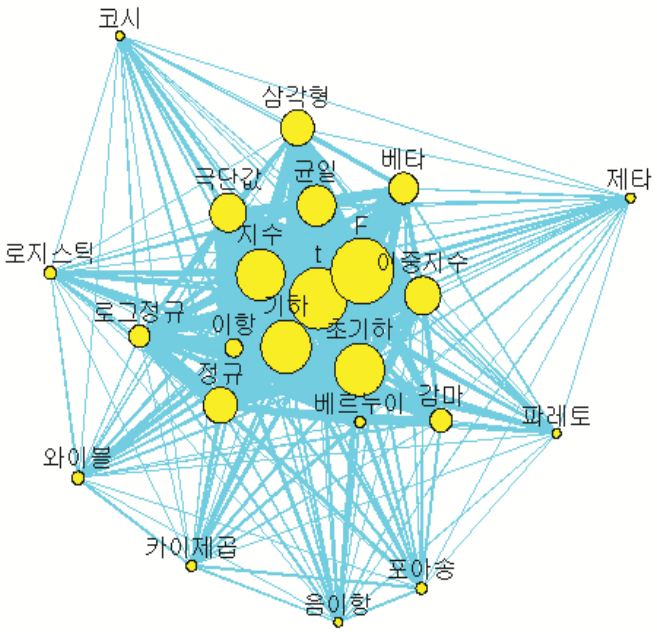

그림 2.1. 확률분포들의 이용빈도 네트워크(한글용어)

등장하여 중요한 확률분포로 인식되는 포아송분포는 우리의 예상과는 달리 네트워크의 중심이 아닌 가장자리에 위치함을 알 수 있다. 표집분포인 t분포와 F분포는 네트워크 중심에 위치함에 비하여 또 다른 표집분포인 카이제곱분포는 네트워크의 중심이 아닌 가장자리에 위치함을 알 수 있다.

중심성(centrality)은 특정 노드에 대하여 전체 네트워크 중심에 위치하는 정도를 나타내는 측도이다. 기초통계학에서 언급되는 중심경향의 측도(measure of central tendency)와 비교되는 개념이다. 사회연결망분석에서 사용하는 중심성에는 연결선수[연결정도중심성](degree), 근접중심성(closeness centrality), 중개[매개]중심성(betweenness centrality), 그래프중심성(graph centrality), 고유벡터중심성(eigenvector centrality) 등이 있다. 이 중 근접중심성은 다음과 같이 정의된다.

$$C(j) = \frac{\sum_{i \neq j} \frac{1}{d(i,j)}}{n-1}, \quad j = 1, 2, \ldots, n,$$

여기서 n은 노드의 수이고 $d(i,j)$는 노드 i에서 j에 이르는 최단경로의 길이이다. 즉, 근접중심성은 각 노드간의 거리를 근거로 하여 측정되는 중심성이다. 확률분포들의 이용빈도 네트워크에서 중심노드를 찾아보기 위하여 근접중심성을 계산하여 보니 다음 그림 2.2와 같았다. 이러한 네트워크는 무방향 네트워크이어서 외향근접중심성(out-degree closeness centrality)값과 내향근접중심성(in-degree closeness centrality)값이 같다. 중심성은 노드 간의 방향이 존재하는 경우 외향중심성과 내향중심성으로 구분되는 데 내향중심성은 교류방향이 외부에서 관심 노드로 들어오는 경우의 중심성이고 외향중심성은 관심 노드에게서 외부로 나가는 경우의 중심성이다. 정규분포보다 근접중심성 값이 큰 확률분포의 개수가 9개, 정규분포보다 근접중심성 값이 작은 확률분포의 개수가 13개였다. 근접중심성 값을 중심으로 23개의 확률분포를 나누어보면 다음과 같이 대략 3개의 그룹으로 나누어 볼 수 있다.

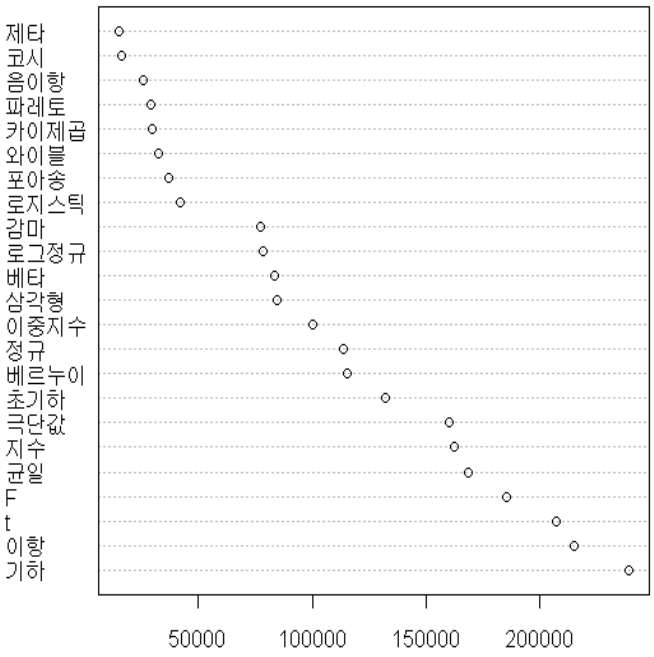

그림 2.2. 확률분포들의 근접중심성(한글용어)

제 1그룹: 기하, 이항, t, F, 균일, 지수, 극단값

제 2그룹: 초기하, 베르누이, 정규, 이중지수, 삼각형, 베타, 로그정규, 감마

제 3그룹: 로지스틱, 포아송, 와이블, 카이제곱, 파레토, 음이항, 코시, 제타

각 확률분포들의 근접중심성과 웹페이지수와의 관계를 알아보기 위하여 산점도를 그리니 그림 2.3과 같았다. 이항분포가 특이한 현상을 나타냄을 알 수 있다. 즉, 웹페이지수는 작으나 근접중심성 값은 큼을 알 수 있다. 또한, 기하분포는 초기하분포나 지수분포와 웹페이지수는 비슷하나 근접중심성 값은 23개 확률분포들 중 제일 큼을 알 수 있다. 정규분포보다 근접중심성 값이 작은 13개의 확률분포는 두 개의 그룹(1. 로지스틱, 포아송, 와이블, 카이제곱, 파레토, 음이항, 코시, 제타, 2. 이중지수, 삼각형, 베타, 로그정규, 감마)로 쉽게 구별되나 정규분포보다 근접중심성 값이 큰 9개의 확률분포는 복잡한 형태를 이룬다.

2.2. 영문용어를 이용한 사회연결망분석

영문용어를 이용하여 23개의 확률분포 각각의 웹페이지수와 확률분포 두 개씩의 동시웹페이지수를 구하면 이용빈도행렬을 구할 수 있다. 이러한 이용빈도행렬에서 대각선 원소는 대응되는 확률분포의 웹페이지수로 구성되고 비대각선 원소는 대응되는 두 개의 확률분포의 동시웹페이지수로 구성된다. 우리는 이렇게 구한 이용빈도행렬을 이용하여 확률분포들의 이용빈도 네트워크를 다음 그림 2.4와 같이 그릴 수 있다. 이 그림은 R library(sna)에 있는 gplot을 이용하여 그렸다. 그림 2.4에서 각 노드의 크기는 해당 확률분포의 웹페이지수에 비례하고 각 연결선의 굵기는 두 개의 해당 확률분포들의 동시웹페이지수에 비례한다. 영문용어를 이용하여 그린 확률분포들의 인용빈도 네트워크는 한글용어를 이용하여 그린

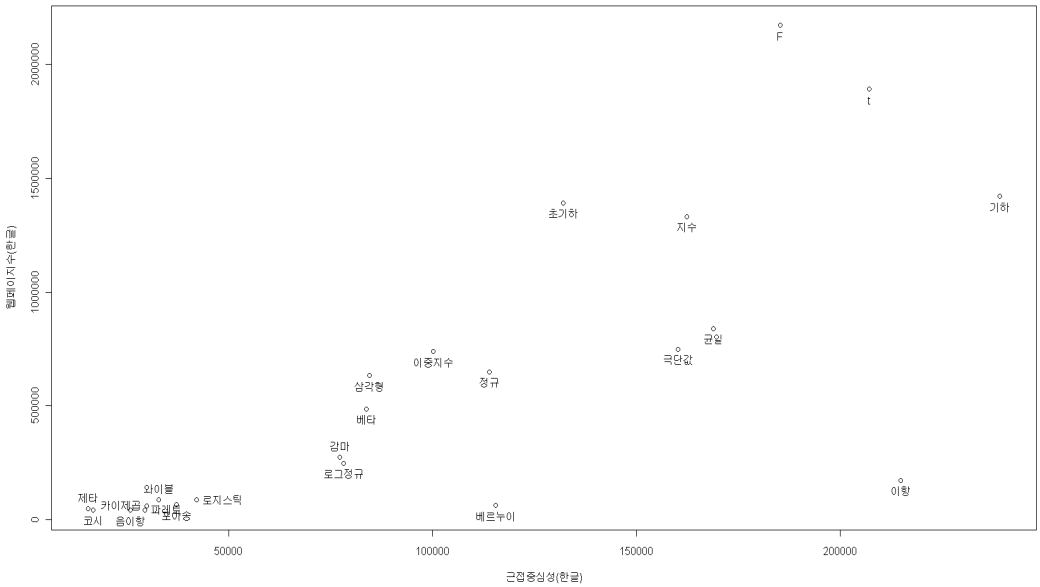

그림 2.3. 각 확률분포들의 근접중심성과 웹페이지수와의 관계를 알아보기 위한 산점도(한글용어)

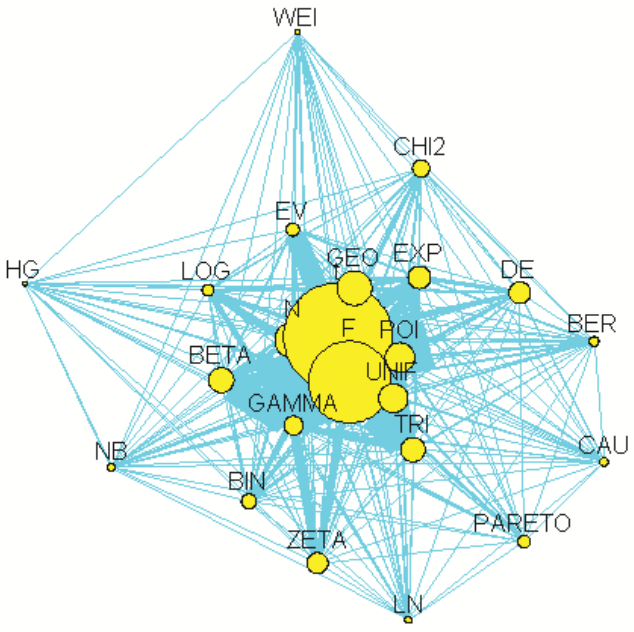

그림 2.4. 확률분포들의 이용빈도 네트워크(영문용어 1. N: 정규분포, 2. EXP: 지수분포, 3. GAMMA: 감마분포, 4. WEI: 와이블분포, 5. CAU: 코시분포, 6. t: t분포, 7. F: F분포, 8. CHI2: 카이제곱분포, 9. BETA: 베타분포, 10. UNIF: 균일분포, 11. TRI: 삼각형분포, 12. LN: 로그정규분포, 13. LOG: 로지스틱분포, 14. EV: 극단값분포, 15. PARETO: 파레토분포, 16. DE: 이중지수분포, 17. BER: 베르누이분포, 18. BIN: 이항분포, 19. HG: 초기하분포, 20. POI: 포아송분포, 21. GEO: 기하분포, 22. NB: 음이항분포, 23. ZETA: 제타분포)

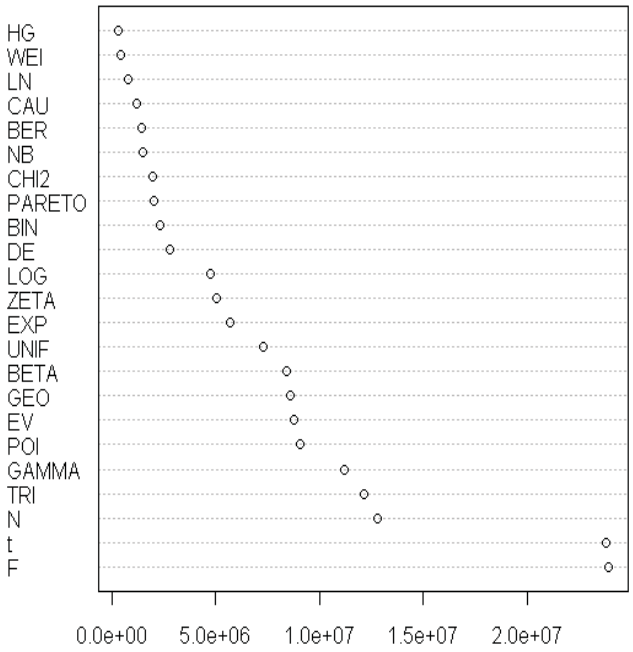

그림 2.5. 확률분포들의 근접중심성(영문용어)

확률분포들의 인용빈도 네트워크과는 다른 구조를 갖는다. 한글용어에서는 정규분포가 네트워크 중심에 있지 않았으나 영문용어에서는 정규분포가 t분포, F분포와 아울러 네트워크 중심에 있음을 알 수 있다. 특이한 것은 이항분포가 한글용어에서는 네트워크 중심 가까이에 위치하여 있었으나 영문용어에서는 네트워크의 중심이 아닌 가장자리 가까이에 위치한다는 사실이다. 반면 한글용어에서는 포아송분포가 네트워크의 중심이 아닌 가장자리에 위치했으나 영문용어에서는 중심 가까이에 위치한다. 한글용어에서처럼 영문용어에서도 표집분포인 t분포와 F분포는 네트워크 중심에 위치함에 비하여 또 다른 표집분포인 카이제곱분포는 네트워크의 중심이 아닌 가장자리에 위치함을 알 수 있다.

확률분포들의 인용빈도 네트워크에서 중심노드를 찾아보기 위하여 근접중심성을 계산하여 보니 다음 그림 2.5와 같았다. 한글용어에서는 포아송분포, 감마분포, 삼각형분포, 정규분포의 근접중심성 값이 상대적으로 작았으나 영문용어에서는 크게 나타났고, 한글용어에서는 이항분포의 근접중심성 값이 상대적으로 컸으나 영문용어에서는 작게 나타났다. 한글용어에서는 제타분포의 근접중심성 값이 제일 작았으나 영문용어에서는 초기하분포의 근접중심성 값이 제일 작았다. 근접중심성 값을 중심으로 23개의 확률분포를 나누어보면 다음과 같이 대략 3개의 그룹으로 나누어 볼 수 있다.

제 1그룹: F, t

제 2그룹: 정규, 삼각형, 감마, 포아송, 극단값, 기하, 베타, 균일, 지수, 제타, 로지스틱

제 3그룹: 이중지수, 이항, 파레토, 카이제곱, 음이항, 베르누이, 코시, 로그정규, 와이블, 초기하

각 확률분포들의 근접중심성과 웹페이지수와의 관계를 알아보기 위하여 산점도를 그리니 그림 2.6과 같았다. t분포와 F분포가 근접중심성 값이나 웹페이지수에서 다른 분포들과 확연한 차이를 보이고 있음을 알 수 있다.

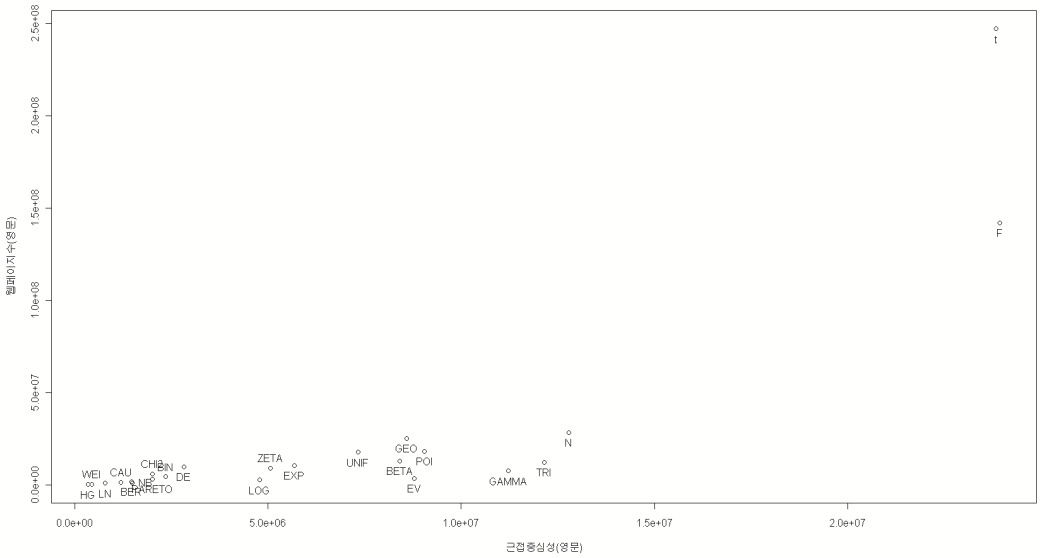

그림 2.6. 각 확률분포들의 근접중심성과 웹페이지수와의 관계를 알아보기 위한 산점도(영문용어)

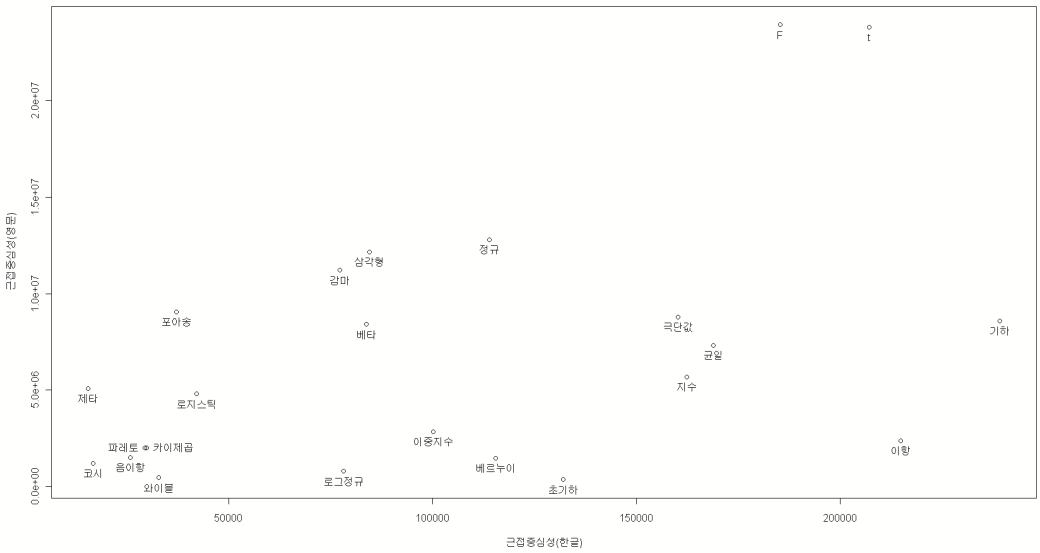

그림 2.7. 각 확률분포들의 근접중심성(한글)과 근접중심성(영문)과의 관계를 알아보기 위한 산점도

각 확률분포들의 근접중심성(한글)과 근접중심성(영문)과의 관계를 알아보기 위하여 산점도를 그리니 그림 2.7과 같았다. t분포와 F분포는 근접중심성(한글)과 근접중심성(영문) 모두 큰 값을 갖고 있어 다른 분포들과 확연한 차이를 보이고 있음을 알 수 있어 이 두 분포가 한글과 영문용어 확률분포들간의 인용빈도 네트워크에서 핵심적 위치에 있음을 알 수 있다. 한글용어에서는 포아송분포, 감마분포, 삼각형분포, 정규분포의 근접중심성 값이 상대적으로 작았으나 영문용어에서는 크게 나타났고, 한글용어에서

표 **2.1.** 기초통계학 교재(5 종류)에 언급된 확률분포

확률분포	김상익 등 (2010)	김우철 등 (2010)	김진경 등 (2010)	김주한 등 (2009)	이외숙 등 (2002)
기하분포			O	O	O
이항분포	O	O	O	O	O
t분포	O	O	O	O	O
F분포	O	O	0	O	O
균일분포	O		O		O
지수분포	O				O
극단값분포					
초기하분포	O	O	O		O
베르누이분포	O	O			
정규분포	O	O	O	O	O
이중지수분포					
삼각형분포			O		
베타분포					
로그정규분포					
감마분포					O
로지스틱분포					
포아송분포	O	O	O	O	O
와이블분포					
카이제곱분포	O	O	O	O	O
파레토분포					
음이항분포				O	
코시분포					
제타분포					

는 이항분포의 근접중심성 값이 상대적으로 컸으나 영문용어에서는 작게 나타남을 알 수 있다. 제타분포는 한글용어에서는 확률분포들 중 근접중심성 값이 제일 작았으나 영문용어에서는 확률분포들 중 중간 순위(12위)에 해당하고 초기하분포는 한글용어에서는 확률분포들 중 근접중심성 값이 8위에 해당했으나 영문용어에서는 확률분포들 중 제일 작았다.

우리는 그림 2.1~2.7을 통하여 우리는 기초통계학 내용 중 확률분포의 취급에 대하여 다음과 같은 제안을 할 수 있다.

1. 대부분의 기초통계학 책에서는 이산확률분포로서 초기하분포, 베르누이분포, 이항분포와 아울러 포아송분포를 다룬다. 그런데, 포아송분포와 매우 밀접한 관계를 갖는 지수분포에 대해서는 크게 강조하지 않는다. 한글용어의 경우 지수분포의 근접중심성 값은 포아송분포의 근접중심성 값의 4배가 넘는다. 그러므로 우리는 기초통계학에서 포아송분포와 더불어 지수분포에 대한 설명도 삽입할 필요가 있다.
2. 대부분의 기초통계학 책에서는 기하분포에 대한 언급이 빈약하다. 그러나 한글용어의 경우 기하분포의 근접중심성 값은 23개 확률분포 중 최고이므로 우리는 기초통계학에서 기하분포에 대한 설명을 할 필요가 있다. 기하분포의 활용 예로서 스포츠 경기에 대한 예들이 많이 있다.
3. 연속확률분포로서 균일분포는 균일분포-삼각형분포-정규분포로 이어지는 일련의 연속확률분포 설명에서 시발점이 되는 확률분포임으로 우리는 균일분포를 기초통계학 책에서 취급할 필요가 있다.

표 **2.2.** 균일분포의 명칭칭들과 웹페이지수(한글용어)

균일분포의 명칭	웹페이지수
균일분포	336,000
직사각형분포	227,000
균등분포	89,900
일양분포	7,240

4. 특이한 현상 중 하나가 극단값분포의 등장이다. 웹상에서는 우리는 극단값분포에 대하여 자주 언급하나 국내 기초통계학 책에서는 극단값분포에 대한 설명이 전무하다. 최근 인기있는 응용통계학의 한 분야인 금융통계학 분야에서 자주 언급되는 분포가 극단값분포이므로 우리는 기초통계학 책에서 이 분포를 언급할 필요가 있다.

국내에서 출간된 기초통계학 교재들 중 5 종류를 택하여 언급된 확률분포들을 조사하여 보니 다음 표 2.1과 같았다. 삼각형분포와 극단값분포에 대한 언급이 거의 또는 전혀 없음을 알 수 있다.

참고. 균일분포는 균일분포, 균등분포, 일양분포, 직사각형분포 등의 여러 명칭이 쓰이는데 구글에서의 웹페이지수는 다음 표 2.2와 같다. 그러므로 균일분포로 용어통일을 시도하는 것이 필요하다 생각한다.

3. 결론

포털사이트를 이용하여 통계분포의 웹페이지수를 알아 보고, 이 정보를 사회연결망분석에 사용하여 우리들의 일상생활에서 나타나는 확률분포들의 관계에 대하여 알아본 결과 기하분포, 지수분포, 균일분포, 극단값분포가 그 중요성에 비하여 기초통계학 교재에서 덜 중요하게 다루어지거나 언급이 되지 않았음을 알 수 있었다. 23개의 확률분포들을 포함한 더 많은 확률분포들을 대상으로 사회분석망을 이용한 확률분포들의 관계에 대한 연구가 추후 연구과제로서 가능할 것이다.

참고문헌

김상익, 김형문, 서한손, 안병진, 여성칠, 유규상, 이석구 (2010). <통계학의 이해와 응용>, 민영사, 서울.

김용학 (2004). <사회 연결망 이론>, 개정판, 박영사, 서울.

김우철, 김재주, 박병욱, 박성현, 송문섭, 이상열, 이영조, 전종우, 조신섭 (2010). <현대통계학>, 제 4개정판, 영지문화사, 서울.

김주한, 김흥기, 박래현, 박석윤, 배종호, 이낙영, 이석훈, 이민구, 이주호 (2009). <통계학 입문>, 정익사, 서울.

김진경, 박진호, 박헌진, 이재준, 전흥석, 황진수 (2010). <통계학>, 개정판, 자유아카데미, 서울.

손동원 (2002). <사회 네트워크 분석>, 경문사, 서울.

이외숙, 임용빈, 성내경, 소병수 (2002). <통계학 입문>, 경문사, 서울.

허명회 (2010). <R을 활용한 사회네트워크분석 입문>, 자유아카데미, 서울.

Brandes, U., Indlekofer, N. and Mader, M. (2011). Visualization methods for longitudinal social networks and stochastic actor-oriented modeling, *Social Networks*, to appear.

Gregson, J., Sowa, M. and Flynn, H. K. (2011). Evaluating form and function of regional partnerships: Applying social network analysis to the network for a healthy California, 2001–2007, *Journal of Nutrition Education and Behavior*, **43**, S67–S74.

Leemis, L. M. and McQueston, J. T. (2008). Univariate distribution relationships, *American Statistician*, **62**, 45–53.

Racherla, P. and Hu, C. (2010). A social network perspective of tourism research collaborations, *Annals of Tourism Research*, **37**, 1012–1034.

Sailer, K. and McCulloh, I. (2011). Social networks and spatial configuration-How office layouts drive social interaction, *Social Networks*, to appear.

Sternitzke, C., Bartkowski, A. and Schramm, R. (2008). Visualizing patent statistics by means of social network analysis tools, *World Patent Information*, **30**, 115–131.

A Study on the Frequency Structure of Probability Distributions Using Social Network Analysis

Dae-Heung Jang[1] · Seongbaek Yi[2]

[1]Department of Statistics, Pukyong National University
[2]Department of Statistics, Pukyong National University

(Received July 2011; accepted September 2011)

Abstract

Through social network analysis using portal site information, we study the relation of the probability distributions that appear in statistics textbooks with probability distributions that appear in daily life. Based on daily life, we discuss probability distributions that must be emphasized in frequent use.

Keywords: Probability distributions, portal sites, social network analysis.

[1]Corresponding author: Professor, Department of Statistics, Pukyong National University, 599-1 Daeyeon-dong, Nam-gu, Busan 08-737, Korea. E-mail: dhjang@pknu.ac.kr