가정의학회지 제23권 제4호 별책 2002
J Korean Acad Fam Med Vol. 23, No.4 Suppl April

■ 세 미 나

통계분석시 유의사항(사례중심)

박 용 규
가톨릭대학교 의과대학 의학통계학교실

초 록

이 강좌에서는 특히 논문을 심사하는 입장에서 이러한 통계적 표현의 적절성을 어떻게 판단할 것이지를 다루기로 한다. 또한 수많은 통계분석기법과 그것의 제시방법을 일일이 다루는 대신, 이미 게재된 논문의 몇 가지 예를 중심으로 논문 전반에 걸친 연구계획과 대상자 선정방법 그리고 결론의 유도과정이 적절한가에 대해서만 생각해 본다.

관찰 또는 실험연구의 결과를 이용하여 작성된 논문에는 통계적 표현이 많이 사용된다. 이 강좌에서는 특히 논문을 심사하는 입장에서 이러한 통계적 표현의 적절성을 어떻게 판단할 것이지를 다루기로 한다. 또한 수많은 통계분석기법과 그것의 제시방법을 일일이 다루는 대신, 이미 게재된 논문의 몇 가지 예를 중심으로 논문 전반에 걸친 연구계획과 대상자 선정방법 그리고 결론의 유도과정이 적절한가에 대해서만 생각해 본다.

1. 모집단

예 1.

문제점

1) 방법의 통계부분에 있는 "표본수가 적어 비모수적인 방법인 Kruskal-Wallis 검정을~"는, t 검정이나 분산분석이 자료가 정규분포에 따르고, 집단간 분산이 동일한 경우, 작은 표본에 적용되는 검정방법이므로, 적절하지 못한 표현이다. 즉, 표본수가 작다는 것만으로는 비모수 검정을 해야할 이유가 되지 못한다.

2) 통계적 검정은 표본에서 얻은 사실에 근거하여 모집단에 대한 주장을 할 때 필요하다. 또한 이 때의 표본추출방법은 랜덤 샘플링이 되어야 한다. 이 연구

1. 연구대상

본 연구는 우리 나라의 일차의료 수준과 보건지표를 평가하고 이 결과를 Starfield[2]가 미국, 캐나다, 네덜란드, 벨기에, 덴마크, 스페인, 스웨덴, 핀란드, 영국, 독일, 프랑스, 호주, 일본 등의 13개 OECD국가들에 대해 1990년대 중반 자료를 바탕으로 이미 시행한 연구 결과와 비교하여 현재 우리의 수준이 세계 속에서 어느 정도의 위치를 차지하고 있는지 살펴보았다.

3. 통계

일차의료 점수에 의해 분류된 세 군간의 비교는 표본수가 적어 비모수적인 방법인 Kruskall-Wall 검정을 이용하였고 통계처리는 SPSS 9.0 버전을 사용했다.

Table 6. Average ranking* of health indicators, health care expenditure, and satisfaction with health care system for countries grouped by primary care scores

	Highest	Middle	Lowest	P-value
Neonatal Mortality Rate	7.80	5.25	9.00	0.400
Postneonatal Mortality Rate †	4.60	5.50	12.00	0.010
Infant Mortality Rate	6.60	5.75	9.80	0.292
Age-standardized Death Rate †	9.40	3.00	9.20	0.039
Potential Years of Life Lost †	7.60	3.00	11.00	0.017
% of Low Birth weight Infants	4.80	7.25	10.40	0.105
Disability-adjusted Life Expectancy	8.40	3.50	9.80	0.058
Life expectancy at 1 year old (male) †	7.90	2.50	11.10	0.009
Life expectancy at 1 year old (female)	8.60	4.00	9.20	0.137
Life expectancy at 15 year old (male) †	8.30	2.50	10.70	0.012
Life expectancy at 15 year old (female)	8.50	4.00	9.30	0.134
Life expectancy at 40 year old (male) †	9.00	2.50	10.00	0.017
Life expectancy at 40 year old (female)	9.00	4.00	8.80	0.140
Life expectancy at 65 year old (male)	9.00	3.50	9.20	0.077
Life expectancy at 65 year old (female)	8.80	3.75	9.20	0.104
Cost of Health care per Capita	4.60	9.00	9.20	0.154
Satisfaction with Health Care System	7.50	5.00	9.50	0.232
Average Health Indicator Ranking †	8.60	2.75	10.60	0.023

* Best level of health indicator was ranked 1. thus, lower average ranks indicated better performance, when the rankings were tied, average ranking was assigned.
† $p<0.05$ by Kruskall-Wallis test.

에서는 만족도에 대해 "유럽연합 15개 회원국의 국민 일부를 대상으로 조사한 것"이라는 부분이 연구대상을 표본으로 인식하게 만든 것으로 보인다. 그러나 만족도 이외의 여러 보건지표와 의료비용에 관련된 내용들은 각각의 국가 전체에 대한 값들이므로, 표본자료로 볼 수 없다.

3) 표 6에 있는 상, 중, 하위 국가들은 14개 OECD 국가 전체를 분류한 것이므로, 모집단으로 간주해야 한다. 다시 말해 이 국가들은 많은 선진국들 중에서 랜덤하게 뽑혀진 일부 국가들이 아니다. 따라서 검정을 할 필요가 없으며, 단순히 평균순위만 제시하는 것만으로도 비교가 이루어진다.

2. 대상자의 선정

예 2.

방법: 1999년 충청남도 아산시에서 지역적 분포를 고려하여 초등학교 및 병설 유치원 13개를 임의표본추출하고, 자기기입식 설문지를 이용하여 비만에 관련된 식사와 생활 습관, 키, 몸무게를 조사하였다. 비만

Methods: We made 13 conventional samples of primary school and attached kindergarten located in Asan-city, ChungNam, in 1999. We surveyed height, weight, food habits and the live style

문제점

1) 한글 요약에는 임의표본추출, 영문 요약에는 conventional sample이라고 하였다. 임의표본추출이란 무작위추출이라는 표현과 함께 랜덤추출(random sampling)의 뜻으로 보통 사용된다.

2) 이 연구에서 실제 사용된 방법은, 지역적 분포를 고려하여 각 지역당 표본수를 비례적으로 할당한 후, 랜덤하게 뽑은 것으로 짐작된다. 이 경우에는 층화 랜덤추출(stratified random sampling)을 했다고 해야한다.

3) 만약 "conventional sample"을, 편의추출인 "convenience sampling"을 나타내기 위해 사용했다면 잘못된 것이다. 통계학에서는 이 용어가 사용되지 않는다

4) 편의표본추출이란 조사에 중요하다고 생각하는 표본을 조사자의 편의(convenience)대로 추출하는 방법으로, 비용이 적게 들고 시간이 절약된다는 장점이 있으나, 조사자가 접근하기 쉬운 대상자에 대해서만 조사가 이루어지는 비확률적 표본추출방법이므로, 조사자의 주관적 판단과 편견이 개입되기 쉽고, 모집단에 대한 대표성이 적어질 가능성이 높다는 단점이 있다.

예 3.

1. 조사 대상 및 방법

1979년 한국에 가정의학이 도입되었으며, 1989년부터 3년간의 전공의 과정을 마친 전문의가 배출되기 시작하였는데, 1989년 이후 가정의학전문의 자격을 취득한 2,075명 중 945명이 가정의학회를 통해 주소 확인이 가능하였고, 이 중에서 대학병원에서 근무하는 62명을 제외한 883명을 대상으로 하였다. 조사 방법은 1998년 7월~8월, 2개월간 우편을 통한 자기 기입식 1차 설문조사를 시행한 후, 다음 1개월간(9월) 불응답자에 대하여 2차 설문조사를 시행하였으며, 272통(1차 226명, 2차 46명)이 회수되어 회수율은 33.3%였다.

2. 일차의료에 대한 신념 및 수행의 수준

1) 일차 응답자와 이차 응답자의 비교(Table 2)

일차 응답자는 226명(83.1%), 이차 응답자는 46명(16.9%)이었다. 일차 응답자의 평균연령은 34.78(±2.97)세, 이차 응답자는 35.48(±2.82)세로 두 군간 연령의 유의한 차이는 없었다(p=0.56), 일차 응답자의 신념의 평균은 4.45(±0.41)점이고 이차 응답자는 4.41(±0.43)점으로 두 군간의 신념의 유의한 차이는 없었다(p=0.58). 일차 응답자의 수행의 평균은 3.64(±0.67)점이고 이차 응답자는 3.70(±0.60)으로 두 군간의 수행의 유의한 차이는 없었다.

따라서 본 연구의 결과가 정규 수련과정을 마친 가정의학과 전문의들을 대표한다고 추정할 수 있다.

문제점

1) 표 2의 1차 응답자와 2차 응답자간의 비교가 유의한 차이가 없었다는 결과를 근거로, 본 연구대상이 전체를 대표할 수 있다는 추정은 논리의 비약이라고 할 수 있다. 표본의 대표성은, 현실적으로 조사자체가 거의 불가능한 일이지만, 응답자와 비응답자간의 비교를 통해서 파악할 수 있는 문제다.

2) 이 연구에서의 모집단은 주소확인이 가능하고 대학병원에서 근무하지 않는 가정의학 전문의 883명이라고 할 수 있다. 연구자가 위와 같은 주장을 하게 된 이유는, 1차 설문조사에 불응하고 1개월 후 응답을 한 46명의 특성이, 조사하지 못한 611명과 동일할 것이라고 보았기 때문이다. 그러나 2차 설문조사에서 응답한 46명의 특성은, 최종 무응답자보다 1차 설문조사의 응답자에 더 가깝다고 보는 것이 타당하다.

3. 부정적 발견

예 4.

연구배경: 최근 국내에서 무분별하게 시행되고 있는 생혈액 검사의 임상적 유용성에 대해 알아보기 위해서, 환자군과 대조군에게 생혈액 검사를 시행하여, 생혈액 검사 상의 이상소견의 차이로 비교함으로써, 생혈액 검사의 임상적 유용성에 대해 평가하고자 하였다.

방법: 2000년 2월 11일부터 3월 8일까지 1개

대학병원에 입원하고 있던 환자 30명과 동 병원에 근무하고 있는 전공의, 연구강사, 직원을 포함한 대조군 30명에게 생혈액 검사를 시행하였다. 생혈액 검사의 비정상 소견 중에서 행혈액 검사에서 흔하게 관찰되는 rouleau formation, spicules (fibrin), protoplast 3가지 소견을 비교하였다.

결과: 환자군과 대조군의 비교에서 rouleau formation은 환자군에서는 3명을 제외한 27명에서 양성이었고, 대조군에서는 모두 양성의 결과를 볼 수 있었다. protoplast는 환자군에서 16명이 양성이었고, 대조군에서는 13명이 양성이었다. rouleau formation, spicule, protoplast의 유무에 따른 환자군과 대조군의 유의한 차이는 없었다.

결론: 생혈액 검사는 임상적인 가치를 가지기 힘들고, 임상적 가치를 가지기 위해서는 생혈액 검사의 결과에 영향을 미치는 요소들을 제거할 수 있는 방법을 제시해야 한다.(가정의학회지 2001; 22:70-77)

다. 분석 및 통계

환자군과 대조군에서 각 소견의 유무를 비교하기 위해서 무압박 채취혈액의 슬라이드에서 1점에서 8점까지를 양성이라고 하였고, 0점의 경우 음성이라고 판정하였다. rouleau formation, spicule의 결과는 환자, 대조군에 대하여 SAS 6.01를 이용한 Fisher's exact test를 시행하였고, protoplast는 환자, 대조군에 대하여 SAS 6.01를 이용한 Chi-Square test를 시행하였다. 수지 압박에 대한 영향을 알아보기 위해서, 각 소견의 출혈빈도를 무압박 채취혈액 슬라이드들과 압박 채취혈액 슬라이드들에서 SAS 6.01을 이용하여 paired t-test로 검증하였고, 두 군의 차이가 의미가 있을 경우, 수지 압박이 환자군과 대조군에서 출현빈도에 미치는 영향을 알아보기 위해서 SAS 6.01을 이용하여 t-test를 시행하였다.

문제점

1) 연구배경의 "생혈액 검사의 임상적 유용성에 대해 평가하고자 하였다"와 결과의 "유의한 차이는 없었다", 결론의 "생혈액 검사는 임상적 가치를 가지기 힘들고…"라는 표현에서는, 연구자의 의도가 생혈액 검사의 유용성을 밝히는데 있는지, 아니면 생혈액 검사의 유용성을 부정하기 위한 것인지 분명하지 않다.

2) 분석 및 통계에 있는 Fisher's exact test, chi-square test, paired t-test는, 환자군과 대조군이 서로 다르다는 것을 증명하기 위한 검정방법이다.

3) 유의수준을 5%로 하여 이러한 검정을 한 결과, 유의한 차이가 없었다면, 연구자는 두 집단간이 다르다는 주장하지 못할 뿐이다. 유의수준이란 제 1종의 오류, 즉 연구자의 차이가 있다는 주장이 거짓일 확률을 나타내므로, 차이가 없다는 주장이 어느 정도 믿을 수 있는지를 나타내는 척도가 될 수 없다.

4) 따라서 연구자가 "생혈액 검사는 임상적 가치를 가지기 힘들고…"라고 하였지만, 이렇게 판단하게된 근거는 제시하지 않았다고 볼 수 있다.

5) 이와 같은 부정적 발견에 대한 주장이 근거를 갖기 위해서는, 우리가 유의수준을 이용하여 "차이가 있다"는 주장의 근거를 제시하듯, "차이가 없다"는 주장의 근거를 β로 제시해야한다. 여기서 β는 제 2종의 오류, 즉 차이가 없다는 주장이 잘못일 확률을 나타낸다.

6) "유의한 차이가 없다"는 결론을 얻게 되는 경우는, 첫째 실제로 차이가 없거나, 둘째 차이를 증명하기에는 표본의 수가 너무 작기 때문일 수 있다. 이 것은 연구자가 차이가 없다는 것을 주장하기 위해, 고의적으로 표본의 수를 작게 할 수도 있다는 것을 의미하며, 이 때 유의수준은 연구자의 고의성을 막는데 아무런 역할을 하지 못한다.

7) "차이가 없다"는 것을 증명하기 위해서는, 우리가 통상적으로 사용하는 검정방법이 아닌, 동등성 검정(equivalence test)을 해야한다. 동등성 검정이란 임상적으로 중요하지 않다고 간주되는 만큼의 차이를 미리 정해놓고 (귀무가설), 집단간의 차이가 그 보다 더 큰지를 (대립가설) 검정하는 것을 말한다. 따라서 동등성 검정은 대부분 단측검정(one-tailed test)의 형태가 된다. 참고로 통상적인 유의성 검정에서는 "차이가 없다 또는 일정량만큼의 차이가 있다"가 귀무가설이 되고, 대립가설은 "차이가 있다 또는 일정 수준의 차이보다 더 큰 차이가 난다"로 둔다.

8) 평균을 비교하는 문제에서의 동등성 검정은 통상적인 검정방법과 별 차이가 없다. 그러나 비율을

비교할 때에는 검정방법이 서로 차이가 있으므로 주의를 해야한다.

9) 결국 연구자의 목적이 부정적 발견을 찾는데 있다면, 단순히 유의수준을 제시하는 것만으로는 설득력이 없게 된다. 따라서 동등성 검정을 하거나, 아니면 β 또는 검정력$(1-\beta)$으로 부정적 결론의 신빙성을 간접적으로나마 제시해야만 한다. 여기서 검정력이란 실제 차이가 있을 때, 그것을 발견해낼 수 있는 확률을 말한다.

10) 검정력을 구하는 방법을 이용하여, 유의한 차이를 발견하는데 필요한 표본의 수도 사전에 구할 수 있다.

참 고 문 헌

1. 대한가정의학회 간행위원회. 가정의를 위한 통계학 -통계적 오류의 예 -. 서울: 의학인쇄사;2001.
2. Lang TA, Secic M. How to report statistics in medicine - Annotated guidelines for authors, editors, and reviewers. Philadelphia(PA) : American College of Physicians;1997.