

■ 논 평

2001년 12월 게재 논문의 통계적 오류

박 용 규

가톨릭대학교 의과대학 의학통계학교실

서 론

2001년 12월 가정의학회지에 게재된 총 9편의 원저 중에는 다음과 같은 통계적 오류가 발견되었다.

- 잘못된 통계용어의 사용
- 주어진 자료의 분포에 적절하지 않은 분석방법을 적용

- 기대도수가 작은 경우에 카이제곱 검정의 사용
- 본문과 표에 제시된 숫자들의 불일치
- 분석 후 유의성에 대한 잘못된 결론제시
- 표본 추출방법의 문제

다음은 각 논문에 나타난 통계적인 문제점들을 제시한 것이다.

각 논문에서의 통계적 오류

1. 박섯별, 일개 직장에서 직장인을 대상으로 흡연, 음주, 스트레스와 관련된 건강에 대한 인식도 조사

1) 설문자료에 결측치가 많았던 것으로 보이지만, 표와 본문에 제시된 숫자들 중에 서로 일치하지 않는 부분이 있다. 표 3의 향후 건강에 대해서는 108명이 아니라 104명이 되어야 하고, 표 2의 흡연이유는 흡연자 120명뿐 아니라 전체대상자 중 결측치를 빼 것으로 간주되는 152명이 대상으로 되어있다.

2) 표 6의 스트레스와 건강간의 관계는 두 변수 모두 이항분포를 하므로 Pearson의 상관계수를 사용하

는 것은 적절하지 않다. 표 5와 같이 카이제곱 검정을 하거나, 연관성 척도(measure of association)를 제시하는 것이 좋을 것이다.¹⁾

2. 박효순 등, 중년의 성인에서 기억에 의한 5년 전 체중과 신장의 정확성

1) 연구결과 3의 내용은 표 2가 아니라 표 3에 제시되어있다.

2) 표 2의 신장의 차이에 - 기호를 사용했다라면 차이의 방향까지 알 수 있다.

3) 유의수준을 반드시 5%로 해야할 필요는 없지만, 1%보다는 더 많이 사용되고 있다. 결과서술부분에서는 체중의 회상오차와 현재 체중의 상관관계가 유의하지 않다고 옳게 결론을 내렸지만, 표 3의 주에서는 $P < 0.05$ 에서 유의하다고 잘못 제시하고 있다. (유의수준이 1%이므로)

3. 권오현 등, 정상 및 과체중인 성인 남성에서 허리 둘레/신장비에 따른 심혈관질환 위험인자의 평가

1) 통계용어 중 "비차비"는 "교차곱비"가 되어야 한다.²⁾

2) 연구방법에서는 총 대상자가 530명으로, 결과에는 516명으로 되어있다.

3) 표 2는 퍼센트를 비교한 것이므로, t 검정이 아닌 카이제곱 검정이나 z 검정이어야 한다.³⁾

4. 이상준 등, 개원가에서 혈압약을 복용중인 환자들의 시간경과에 따른 운동여부의 변화 및 주치의의 영향

1) 표 1에서 4가지의 빈도분석 중 기대도수가 5보다 작은 칸이 많이 보인다. 이 때에는 카이제곱 검정보다 Fisher의 exact test가 적절하다. 또한 분산분석에서 유의한 결과를 얻었을 경우 Scheffe의 다중비교까지 했으면 한다.⁴⁾

2) 그림 1과 2의 생존분석을 통해 시간에 따른 운동의 지속률과 비운동의 지속률을 각각 보이고자 하였다. 결과부분에 이에 대한 설명이 없이 그림의 주에만 분석결과를 제시하고 있어 자세히 알 수는 없지만, 각 대상군 115명과 78명 중 추적이 불가능한 경우가 없었다면, 굳이 생존분석을 할 필요가 없다고 본다. 생존분석의 특징은, 특히 Kaplan-Meier의 방법은 중도탈락(censored) 자료가 있는 경우에 사용된다는 것이며, 탈락이 없다면 단순히 해당시점까지의 사전발생수/전체대상자수가 해당기간의 생존율이 된다. 물론 탈락이 없더라도 생존분석을 통해 사건이 일어난 매 시점에서의 생존율을 계산할 수는 있지만, 정해진 몇 시점에서의 지속률을 표로 제시하는 것도 고려해 볼 수 있다.

3) 그림 1과 2의 세로축은 생존분석 프로그램의 출력결과를 그대로 나타내었다. 운동의 지속률과 비운동의 지속률로 표현을 바꾸어야 한다.

5. 강희성 등, 자궁경부 세포진검사서 Class II로 나타난 환자의 추적관찰 행태에 관련된 요인

1) 결과에서 수입과 학력간의 다중공선성에 관한 언급을 하고 있다. 다중공선성이란 다중회귀분석이나 다중 로지스틱 회귀분석과 같이 여러 개의 설명변수를 이용하여 하나의 반응값의 변화를 설명하고자 할 때, 설명변수들간의 상관관계가 지나치게 높아서 회귀계수의 추정에 문제를 일으키는 현상을 말한다. 저자들이 이를 언급한 것 자체가 잘못된 것이라고 할 수는 없지만, 해당 분석결과인 표 2를 보면, 분석모형에 수입과 학력만 있는 것이 아니라, 나이, 직업 등의 다른 설명변수가 함께 사용되었으므로, 유의한 두 변수간의 다중공선성에 대해서만 기술한 점은 적절하지

않다. 또한 이 부분은 연구방법부분에서 서술하는 것이 더 적절해 보인다.⁵⁾

6. 이도행 등, 진행된 암환자에서 수행능력상태와 임상증상 및 검사소견이 생존기간에 미치는 영향

1) 연구대상을 선정하는 방법으로 편의추출을 했다고 기술하였다. 편의표본추출이란 조사자의 편의(convenience)대로 표본을 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.⁶⁾

2) 표 2의 P값에 대한 주에 Kaplan-Meier method라고 되어있다. Kaplan-Meier 방법은 생존율 또는 median survival time을 계산하는 방법을 말하며, 집단별 생존율간에 차이가 있는지를 비교하는 방법은 Wilcoxon 또는 Log-rank test라고 해야한다.⁷⁾

7. 정통령 등, 체질량지수와 이상체중을 이용한 비만도 판정기준간의 일치도 비교 - 서울지역 일개 건강증진센터 수진자를 대상으로 -

1) 여러 가지 판정기준들간의 일치도를 비교할 때, 단순히 카파값만 제시하기 보다 카파값들의 크고 작음을 검정을 통해 밝혔으면 한다.

2) 이상체중을 이용한 판정기준의 일치도를 나타낸 표 5와 6에서, 가장 높은 일치도를 얻게되는 경계값을 구하는 방법이 궁급하다. 두 가지로 나누는 경우에는 ROC curve가 주로 사용된다.

8. 배상필 등, 개원 가정의의 평생의학교육에 대한 인식과 행태

1) 표 2에서 사용한 다중선형 회귀분석의 원래의 사용목적은 설명변수의 단위증가량에 따른 반응변수의 변화량을 회귀계수(기울기)로 나타내는데 있다. 이 표에서 저자들이 다중회귀분석을 한 목적은 다른 공변량들의 영향력을 조정된 후의 설명변수의 각 집단별 평균을 제시하고자 한 것으로 보인다. 물론 회

귀분석을 이러한 목적으로 사용할 수도 있지만, 저자들의 목적에 보다 부합되는 분석방법은 공분산분석이라고 할 수 있다. 공분산분석이나 다중회귀분석, 그리고 다원분산분석 등을 통틀어 통계학에서는 선형모형이라고 하며, 이들의 분석결과는 서로 동일하다. 따라서 각 분석의 원래 목적에 맞는 분석의 이름을 제시하는 것이 옳다.

2) 결과부분에 주당 학습시간에 관해 정규분포를 하지 않는 자료의 대표값인 중위수(median)를 굳이 제시한 의도가 무엇인지 궁금하다. 이와 관련된 분석(특히, 표 2)은 정규분포를 하는 자료의 대표값인 평균을 이용하고 있다.

3) 주당 학습시간의 중위수가 3시간이라는 것과 표 2의 조정된 평균값이 약 5.5시간이라는 점은, 자료가 상당히 한쪽으로 치우친 분포를 하거나 회귀분석을 통해 조정된 값이 잘못 제시되었다는 것을 말해 준다. 만약 치우친 분포를 한다면, 모수적 방법인 회귀분석이 적절하지 않다.

9. 이동진 등, 개원가에서의 전화 의료상담 및 의사-환자의 인식도 - 강릉시 지역 일차의료 개원의와 환자를 중심으로 -

1) 연구방법의 통계분석의 서술에서 "빈도의 비교는 Chi-square test를, 비모수적 빈도비교는 Fisher's exact test를 이용"이라는 표현은 "빈도의 비교는 Chi-square test와 Fisher's exact test를 이용"으로 되어야 한다. Fisher's exact test는 분할표를 구성하는 각 칸의 기대빈도가 5보다 작은 칸의 수가 전체의 25%를 넘을 때 사용하는 분석방법이다.

2) 표 5에 대한 Fisher's exact test의 p값이 0.05보다 작다는 결론은 잘못된 것이다. 실제 분석결과는 $p=0.636$ 이다.

결 론

연구대상자의 수가 많고 적음은 연구자가 내린 결과의 신빙성을 제시하는 간접적인 척도가 된다. 다시 말해 대상자수가 많다면 그 연구자는 보다 많은 정보를 근거로 하여 자신의 결론을 얻은 것이고, 대상자수가 적다면 연구의 결론은 충분하지 않은 증거를 기

반으로 내려진 것이라 할 수 있다. 따라서 가능한 많은 수의 연구대상자를 확보하는 것이 중요하지만, 대부분의 경우 현실적인 여건으로 인해 제한받지 않을 수 없다. 그러나 연구대상수가 적다고 해서 그 연구가 가치가 없다는 말은 아니다. 모든 연구가 반드시 통계적 검정을 거쳐야 하는 것도 아니고, 사례연구와 같이 몇 사람의 환자에 대한 보고자체가 대규모 관찰 연구보다 의학적으로 더욱 중요하게 취급될 수도 있기 때문이다.

많은 연구에서 표본수와 관련된 문제점은 최종적으로 확보된 대상자들이라 하더라도 측정하고자 하는 변수가 여러 개일 때 결측치가 발생한다는 것이다. 특히 설문조사의 경우 그 정도가 더 심한 경향이 있다. 이처럼 결측치가 많은 자료를 분석하여 그 결과를 표로 제시할 때 대상자수가 서로 일치하지 않는 것이 자주 발견된다. 더 심한 경우에는 연구자가 정의한 내용과 제시된 대상자가 서로 달라 독자들에게 상당한 혼란을 줄 때도 있다. 이러한 현상은 결측치에 대한 정보를 누락하거나, 자료처리과정에서 오류를 범한 경우에 자주 발생하므로 분석시 세심한 주의가 요구된다.

참 고 문 헌

1. Everitt BS. The analysis of contingency tables. London: Chapman and Hall;1977.
2. 한국통계학회. 통계용어사전. 서울: 자유아카데미, 1987.
3. 송혜향, 김동재. 통계학. 서울: 청문각, 1998.
4. Snedecor GW, Cochran WG. Statistical methods. 7th ed. Ames (IA): Iowa state univ. press;1980.
5. Draper NR, Smith H. Applied regression analysis. 3rd ed. New York (NY):Wiley;1998.
6. Cochran WG. Sampling techniques. 2nd ed. New York(NY): John Wiley and Sons, Inc.;1963.
7. 송혜향, 정갑도, 이원철. 생존분석. 서울: 청문각; 1996.
8. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York(NY): John Wiley and Sons, Inc.;1981.