大韓獸醫學會誌(1999) 第39卷 第6號
Korean J Vet Res(1999) 39(6) : 1187~1196

수의학회지 논문에 적용된 통계기법의 타당성 평가

박 선 일

서울대학교 수의과대학 내과학교실
(1999년 5월 13일 접수)

An assessment of statistical errors in articles in the Korean journal of veterinary research

Son-il Pak

Department of Internal Medicine, College of Veterinary Medicine, Seoul National University, Seoul 151-742, Korea
(Received May 13, 1999)

Abstract : The purpose of this study is to assess the suitability of the statistical techniques employed in papers published in the Korean Journal of Veterinary Research from March 1997 to March 1999 and it is hoped that the critical assessment may be of help to other researchers preparing their works for publication. Of the 246 original papers 94 were included in the analysis. Of 62 papers with the measure of central location and dispersion of data 34 (54.8%) used them correctly : 9 (39.1%) of 23 for *t*-test ; 1 (33.3%) of 3 for correlation analysis ; 7 (43.8%) of 16 for analysis of variance (ANOVA) ; 5 (62.5%) of 8 for chi-square test ; 44 (71%) of 62 for description of p-value. A number of papers employed ANOVA did not perform subsequent analysis of multiple comparison. Compared to the results of others, relatively higher proportion of papers in the present study was evaluated as appropriate analysis. The reason is that papers described insufficiently on the study design were not included, and evaluation items were restricted to the cases violated seriously inherent assumptions for each statistical technique. Statistical misuse or abuse appeared in the study is due to lack of knowledge on statistics and short of its importance for improvement the quality of paper. Because an inappropriate analysis can lead the readers to misunderstand on findings, observed statistical analyses must be valid, and correctly undertaken. It is suggested that more intensive statistical refereeing are needed, and the communication should be allowed for the controversial points.

Key words : statistics, statistical error, veterinary article.

Address reprint requests to Dr. Son-il Pak, Department of Internal Medicine, College of Veterinary Medicine, Seoul National University, Seoul, 151-742, Republic of Korea.

서 론

많은 연구자들은 과거에 경험하지 못한 새로운 현상을 발견하고 장기간 관찰하여 얻은 축적된 결과를 바탕으로 새로운 지식체계를 확립하려는 공통된 목적으로 연구를 진행한다. 특히 자연과학에서는 이를 실증적인 관점에서 규명하기 위하여 현장이나 실험실 연구를 통하여 관찰된 결과를 보다 객관적이고 신뢰성 있는 관계를 도출하게 되는데 이러한 모든 과정을 소위 과학적 접근법이라 한다. 관찰된 결과가 과학적으로 의미가 있는 현상으로 받아들여지기 까지에는 대다수의 연구자가 인정할 수 있는 연구방법과 분석을 근거로 해야 하는데 이 과정에 통계적 기법은 중요한 위치를 차지한다. 얻어진 결과를 단순히 나열하는 것은 기술적 연구의 한 수단이 되지만 문제의 적극적 해결에 관심을 두는 분석적 연구에서는 연구결과를 바라보는 시각에 따라 문제의 크기를 잘못 파악할 수 있어 경우에 따라서는 유사한 연구를 반복적으로 수행하는 번거로움이 초래되기도 한다.

수의학에 관한 대학교육이 이루어진지도 반세기가 지나면서 더불어 수많은 연구논문들이 국내외 학술지에 발표되었지만 그 중에는 매우 의미있는 실험을 하고서도 결과에 대한 분석을 적절하게 처리하지 못하여 논문의 질을 저하시키는 경우가 종종 발견되는 것이 사실이다. 학술지에 발표된 논문을 통계적으로 평가하려는 시도는 많이 있다. Schor와 Karten[1]은 1964년 1월에서 3월까지 3개월 동안 의학분야의 10개 학술지에 게재된 논문을 대상으로 분석한 결과 295편중 140편(47.5%)에서 통계적 오류가 있음을 지적하였다. White[2]는 1977년 7월부터 1978년 6월까지 1년간 영국정신의학회지에 게재된 168편의 논문을 분석한 결과 63편(45%)의 논문에서 통계적 오류를 범하였다고 보고하였다. 한편 국내의 상황도 큰 차이가 없는데 안과 고[3]는 1971년 국내에서 발간된 의학학술지에 게재된 92편의 연구논문을 분석한 결과 50편(54.3%)에서 통계적 적용이 불충분하다고 보고하였다. 하[4]는 1978년 서울대학교 보건대학원 학위논문 195편중 148편(76.0%)에서 하나 이상의 통계적 문제가 있었던 것으로 발표하였다. 현[5]은 1987년에서 1988년 동안 의학학술지에 발표된 75편을 분석한 결과 52편(69.3%)에서 한가지 이상의 오류가 있음을 지적하였다. 이 등[6]은 1980년 1월부터 1989년 12월까지 가정의학회지에 발표된 297편의 논문중 290편(97.6%)에서 하나 이상의 통계적 오류가 있음을 지적하였다. 최근 이와 이[7]는 1986 년부터 1995년까지 국내 간호학 연구논문 166편을 저자가 작성한 점검표에 준하여 조사한 결과 적절한 분석으로 평가받은 논문은 중앙집중성과 산포성의 경우 26%, t-검정 58~98%, 분산분석 11~100%, 상관분석 0~100%, 회귀분석 0~48%, 카이제곱검정 16~100% 등으로 나타났다.

모든 자료를 반드시 통계적 처리를 할 필요는 없지만 굳이 통계분석의 필요성을 말하자면 단순히 기술통계량을 나열하는 것보다는 얻어진 결과에 대한 객관적인 분석을 시도함으로써 신뢰성 있는 연구결과를 제시하여 독자로 하여금 이해의 폭을 넓히고 다양한 정보를 제공하기 위한 것이라 할 수 있다. 이러한 관점에서 저자는 수의학회지에 발표된 논문에서 사용된 다양한 통계적 기법에 대하여 그 적용의 타당성을 검토 및 평가함으로써 통계기법에 대한 올바른 이해를 정립하고자 본 연구를 시도하였다.

재료 및 방법

연구대상 논문의 선정 : 본 조사에 착수하기전 1995년을 전후로 각각 20편의 논문을 선정하여 검토한 결과 비교적 활용빈도가 높은 통계기법에서 조차 많은 오류가 발견되어 가장 최근에 발표된 논문을 중심으로 평가하였다. 연구대상 논문은 1997년부터 1999년 3월호 까지 대한수의학회지에 발표된 총 246편의 논문중 통계처리를 수행하여 분석의 대상에 해당하는 논문을 선정한 결과 97년 44편, 98년 38편, 99년 12편으로 총 94편이었다. 이들 논문에서 저자들이 사용한 통계기법의 활용빈도를 보면 중앙집중성과 산포성의 측정이 가장 많았으며 t-검정, 일원분산분석, 카이제곱분석, 회귀분석, 상관분석 등이었다. 한편 비모수검정으로는 유일하게 1편만이 Wilcoxon signed rank test를 사용하였고 기타 신뢰도 분석과 두 진단검사의 일치도를 평가하기 위하여 KAPPA 통계량을 사용한 논문이 각각 1편씩이었다.

분석내용 : 각각의 통계기법에 내재된 기본가정의 충족여부와 검정결과에 따른 해석, 표본의 크기, 자료의 특성 등 세부적인 항목으로 분류하여 분석하려고 하였으나 대부분의 논문에서 실험내용에 비하여 자료처리 부분에 대한 설명이 미약하여 충분히 그 내용을 파악할 수 없는 경우가 많았다. 이는 자연과학의 특성상 자료처리

의 필요성과 중요성에 대한 인식이 높지 않았을 것이라는 전제하에 부득이 Table 1과 같은 평가항목으로 한정하였다.

Table 1. Some selected methods and its corresponding items for statistical evaluation of the paper published in the Korean Journal of Veterinary Science, Vol 35 No 1, 1997~Vol 37 No 1, 1999

Statistical method	Item
Central tendency & dispersion	• Mean±standard deviation • Mean, median, mode, standard error, interquartile range • sample size
t-test	• Continuous random variable • Sample size • Independence • Normality
Correlation	• Data scale • Linearity • Sample size
Analysis of variance (ANOVA)	• Independence • Normality • Multiple comparison • Sample size
Chi-square test	• Sample size & expected frequency • Description of the type of the method employed • Independence
Others	• Description of statistical method • significance level • Sample size • Standard error

결 과

통계기법별 적용의 타당성을 평가한 결과는 Table 2와 같고 평가항목별로 살펴보면 다음과 같다.

중앙집중성과 산포성 : 중앙집중성과 산포성의 측정에 대하여 분석이 가능했던 총 62편의 논문중 34편(54.8%)이 올바르게 사용한 것으로 나타났다. 이를 세분하면 표준편차의 경우 12편중 8편(66.7%)이 적절하게 사용하였고 반면에 표준오차는 21편중 3편(14.3%)만이 올바르게 사용하였다. 또한 자료의 성격에 부합되는 평균과 표준편차를 동시에 정확하게 사용한 편수는 50%에 불과한 것으로 조사되었다.

***t*-검정** : *t*-검정을 이용한 총 23편의 논문중 9편(39.1%)만이 올바른 분석을 하였는데 부적절한 분석으로 평가된 이유로는 표본의 크기가 불충분하였거나 두 연관된 표본에 대하여 짝지어진 검정을 적용하지 않았기 때문이었다.

상관분석 : 상관분석을 이용한 3편의 논문중 2편은 표본의 크기와 변수들간의 선형성 가정과 해석상에 문제가 있는 것으로 판된되어 부적절한 분석으로 분류되었다.

분산분석 : 분산분석을 이용한 총 16편의 논문중 7편(43.8%)이 올바른 분석을 하였는데 그 이유로는 표본의 크기가 불충분한 자료에 대하여 비모수검정을 적용하지 않은 경우 동일한 개체를 대상으로 반복측정한 자료를 단순 분산분석한 경우가 대부분이었다. 한편 올바르게 분석한 논문중 분산분석에서 유의한 결과가 나온 경우 다중비교를 수행하지 않은 논문이 많았는데 이는 분산

Table 2. Results of evaluation by statistical methods

Statistical method	Results		No. papers evaluated
	Correct	Incorrect	
Central tendency & dispersion	34(54.8%)	28(45.2%)	62
t-test	9(39.1%)	14(60.9%)	23
Correlation	1(33.3%)	2(66.7%)	3
ANOVA	7(43.8%)	9(56.2%)	16
Chi-square test	5(62.5%)	3(37.5%)	8
Significance level	44(71.0%)	18(29.0%)	62

분석의 의도를 정확하게 파악하지 못한 상태에서 적용한 결과로 보아진다.

카이제곱검정 : 카이제곱검정를 사용하고 분석의 대상이 되는 8편의 논문중 5편(62.5%)이 올바르게 적용하였다. 이 검정과 관련하여 가장 흔한 문제는 기대도수가 5 이하인 셀이 20%를 초과하거나 기대도수가 1 이하인 셀이 있음에도 이를 무시한 경우 자료의 성격상 카이제곱분포를 적용하기 어려운 상황에서 정확확률검정이나 로그선형모형을 고려하지 않고 그대로 사용한 경우 혹은 동질성 검정을 독립성 검정으로 표현한 경우였다.

기타사항 : 유의수준이 적용된 총62편의 논문중 44편(71%)이 정확한 표현을 사용하였다. 몇가지 문제점으로 나타난 부분은 우선 분석결과의 해석과 관련된 것으로 분석내용과 그 결과에 대한 설명이 일치하지 않는 경우가 많았고 실험내용이나 연구결과의 해석으로 미루어볼 때 이원분산분석(two-way ANOVA)을 시도해야 하는 실험임에도 불구하고 일원 분산분석으로 분석하였거나 일원 분산분석을 시행하고도 결과의 해석에서는 이원 분산분석으로 유추한 논문도 있었다. 또한 유의한 결과에 대하여 p-value에 근거한 인과론적 해석을 하거나 확대해석하는 경우가 많았다. 분석방법과 관련된 오류로는 카이제곱 검정에 해당하는 자료를 분산분석으로 시도한 경우 짝지어진 자료에 대하여 단순 *t*-검정으로 분석하는 경우, 반복측정 분산분석(repeated ANOVA)에 해당하는 자료를 단순 분산분석을 적용한 경우도 있었다.

한편 실험연구에서 각 군에 할당된 표본의 수를 정확히 명시하지 않아 검정의 결과를 이해하는데 어려움을 준 논문도 다수 있었다. 분석방법에 대한 언급이 전혀 명시되어 있지 않음에도 불구하고 본문의 내용에는 "통계적으로 유의하였다"라는 표현을 사용하였거나 혹은 그 반대로 재료 및 방법에서는 "유의수준과 신뢰구간을 계산하였다"라는 표현이 있지만 본문의 결과에는 이러한 결과에 대한 언급이 전혀 없는 경우도 있었다. 분산분석에서 유의한 결과가 도출되어 귀무가설을 기각하는 경우 다중비교를 시행하게 되는데 여러가지 비교방법중에서 어떠한 방법을 사용하였는지에 대한 언급이 전혀 없는 경우도 있었다. 유의한 차이가 없는 결과에 대하여 $p < 0.05$로 표시하거나 특별한 이유없이 다양한 p값을 혼용하는 경우도 많았다.

고 찰

중앙집중성과 산포성 : 관찰치의 범위에서 가운데 위치하는 값을 흔히 평균(mean)이라 하며 통계학자들은 이를 중앙집중성(measure of central tendency, measure of location)이라고 한다. 평균은 정규분포 자료에 가장 적합한 통계량이며 평균 이외에도 비대칭분포나 순위척도로 측정된 질적자료(qualitative data)에 사용되는 중위수(median)와 명목자료에서 주로 사용되는 최빈수(mode)가 있다. 따라서 자료의 분포에 따라 적절한 방법을 사용해야 하며 이를 도식화 하면 아래의 그림과 같다(Fig 1). 결과를 제시할 때 자료에 대한 기술통계량으로 중앙집중성과 산포성(dispersion)을 함께 제시해야 하는데 그 이유는 평균은 동일하지만 표준편차가 상이한 경우 서로 다른 분포를 보일 수 있기 때문이다. 흔히 평균과 표준편차를 a±b 형태로 표현할 때 a에는 평균이나 중위수가 b에는 표준편차를 비롯한 4분위 편차, 표준오차 등 다양한 값이 가능하므로 어떠한 측정치를 사용하였는지를 반드시 밝혀야만 한다. 산포성이라 함은 관찰치가 퍼져있는 정도를 측정하는 수단으로 표준편차, 분산, 평균편차(mean deviation), 4분위편차(quartile deviation), 분위수(percentile) 등이 있으며 이들 역시 자료의 분포에 따라 적절하게 선

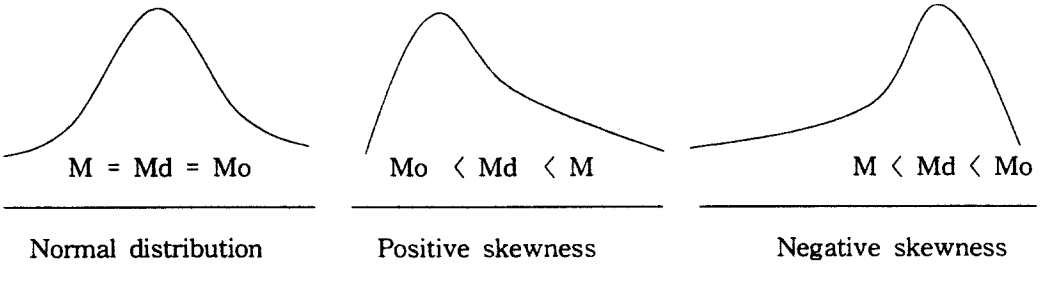

Fig 1. Measures of central tendency in various frequency distributions.
M : mean, Md : median, Mo : mode.

택해야 한다. 표준편차는 정규분포를 보이는 자료에 가장 적합하며, 중앙집중성을 중위수로 표현한 경우 산포성은 4분위 편차를 흔히 사용한다.

한편, 표준편차와 표준오차의 개념을 혼동하여 사용하는 논문이 평가대상 논문의 85.7%을 차지하였다. 표준편차는 자료의 산포성을 나타내는 척도인 반면 표준오차는 표본평균들의 분포에 대한 표준편차(standard error of the mean, SEM, SE)이므로 표본평균들의 정확도나 신뢰도를 나타내는 통계량이다. 표준오차는 모집단의 표준편차를 표본의 크기로 나누어 계산되기 때문에 표준편차에 비해서 작은 값을 갖는 특성이 있어 일부 연구자들은 자료의 신뢰성이 좋아 보이도록 조작하기 위하여 표준오차를 사용하는데 이는 분명한 오류다[8]. 예를 들어 혈액화학치의 표준편차가 6인 어떤 집단에서 크기가 10, 25, 100인 표본을 뽑는다고 할 때 기대되는 표준편차는 $\frac{6}{\sqrt{10}}=1.9;\ \frac{6}{\sqrt{25}}=1,2;\ \frac{6}{\sqrt{100}}=0.6$으로 표본의 크기가 증가할수록 낮은 값을 갖는다는 것을 알 수 있다. 또한 표준오차는 해석면에서 표준편차와 큰 차이가 있다. 예컨대 A라는 질병을 가진 216명을 대상으로 혈청 albumin을 측정한 결과 평균이 34.46g/dl, 표준편차가 5.84g/dl라고 하자. 이 경우 표준오차는 $\frac{5.84}{\sqrt{216}}=0.397$이 되는데 이는 동일한 크기의 표본을 반복하여 뽑을 때 평균이 34.46g/dl이고 표준편차는 0.397g/dl이 된다는 의미이다. 따라서 자료의 산포성을 나타내는 척도로서 표준편차를 사용해서는 안된다.

t-검정 : 이 기법은 여러가지 목적으로 사용되지만(예를 들어 상관계수가 0과 유의하게 다른가를 검정하는 경우) 가장 흔한 용도는 두 집단간 산술평균의 차이에 대한 가설을 검정하는 경우로 검정결과 두 집단간 표본평균이 매우 다르다면 연구자는 두 집단이 동일한 평균을 갖지 않는 것으로 결론을 내리게 된다. 모든 검정에서와 마찬가지로 *t*-검정을 올바르게 수행하기 위해서는 이와 관련된 기본가정을 충복시켜야 하는데 정규분포(normal distribution) 집단에서 독립적으로(independence) 추출된 연속확률변수(continuous random variable)로서 두 집단의 분산이 동일(equal variance) 해야 한다는 것이다. 여기에서 연속확률변수는 하위척도순으로 나열하는 경우 명목척도(nominal scale), 순위척도(ordinal scale), 등간격척도(interval scale), 비척도(ratio scale) 중에서 최소한 등간격척도 이상으로 측정된 자료를 말한다[9]. 정규분포 가정의 경우 실제로 수집되는 자료들이 정규분포를 만족하는 경우가 매우 드물지만 이 분포를 선호하는 이유는 중심극한 정리(central limit theorem)와 관련이 있는데 이는 모집단의 분포형태에 관계없이 표본의 크기가 클수록 정규분포에 근사(approximation)한다는 이론으로 모수를 추정하기 위하여 표본평균의 분포에 대하여 이 이론을 적용할 수 있기 때문이다[10]. 현실적으로 모집단을 전수조사하기 전에는 알 수 없는 모수를 추정하기 위하여 표본통계량을 사용한다고 하였는데 실제로 이러한 값들이 진정한 불편추정치(unbiased estimate)가 되는지에 대해서는 논란이 있지만 이 기법을 대신할 만한 방법이 증명되지 못하고 있는 실정이다[11]. 표본의 수가 적어도 25 이상으로 크고, 두 표본의 크기가 비슷하다면 정규분포에서 벗어난 자료라고 하더라도 *t*-검정이 robust 하다고 알려져 있다[10,12]. 이러한 측면에서 볼 때 수의학회지에 발표된 많은 논문들을 보면 3~7개 정도의 표본크기가 약 70%를 차지하고 있는데 이 정도의 크기는 작은 표본이며 정규분포 가정을 충족시키는 것이 불가능하므로 모수적 검정(parametric method) 보다는 비모수적 검정(non-parametric method)을 선택하는 것이 바람직하다. 세 번째로 등분산 가정이 필요한 이유는 *t*-검정의 통계량이 두 개의 독립적인 표본으로부터 계산되는 두 개의 표본분산이 동일하다는 것을 전제로 공통분산(common variance)으로 합병추정치(pooled estimate)를 사용하기 때문이므로 모집단의 분산이 동일하다는 가정이 어려운 경우에는 *t*-분포를 사용해서는 안된다. 또한 표본의 크기가 2배 정도 차이가 있는 경우 표본의 크기가 동일한 경우에 비하여 두 표본분산의 차이가 6% 증가하고, 표본의 차이가 크면 클수록 두 표본분산의 차이는 증가한다[11]. 따라서 표본의 크기가 작고 비대칭일 경우 비모수적 방법을 고려해야 한다. 마지막으로 독립성에 대한 가정은 하나의 관찰치와 다른 관찰치 간의 독립성은 물론이고, 한 표본 내에서도 관찰치간 독립성이 유지되어야 한다는 것이다. 이와 반대로 독립성이 유지되지 못한 자료를 연관된 표본(correlated sample) 혹은 짝지어진 표본(paired sample)이라 하며 이러한 자료에 대한 분석방법을 matched *t*-test, correlated *t*-test 혹은 paired *t*-test 등으로 부른다. 가장 흔한 예는 동일한 개체를 대상으로 시점을 달리하여 측정한 자료라든가 특정한 처치를 하기 전에 얻은 자료와 처치 후에 얻은 자료를 상호비교하는 경우이다. 독립

표본에 대한 *t*-검정에서의 자유도(degree of freedom)는 총관찰치에서 2를 빼어주지만 짝지어진 자료에는 짝지어진 전체 쌍에서 1을 빼면 된다.

상관분석 : 등간격 이상의 척도로 측정된 하나의 변수에 대한 특성을 파악하고자 할 때에는 중앙집중성과 산포성을 이용하지만 두 변수의 관계에 대하여 관심을 갖는 경우에는 상관분석을 사용한다. 두 변수간 관계의 크기를 상관계수(correlation coefficient, γ)라 하며 -1에서 +1까지의 값을 갖는다. 여기에서 부호는 상관성의 방향을, 숫자는 상관성의 강도를 의미한다. 검정결과 양의 상관관계(positive correlation)가 인정되면 두 변수간 직접적인 관련성이 있고, 음의 상관관계(negative correlation)가 인정되면 두 변수가 역상관성(inverse relationship)이 있다고 하며, 이 값이 클수록 양 혹 음의 방향으로 상관성이 매우 높다고 말할 수 있다. 상관관계에 대한 분석기법으로 모수적인 방법으로 Pearson product-moment correlation coefficient, 비모수적인 방법은 순위를 이용한 Spearman's rho가 있다.

상관분석에도 몇가지 가정이 있다. 앞서 언급한 바와 같이 측정한 자료의 척도에 따라 적용기법이 다르기 때문에 변수의 종류와 두 변수의 관계가 선형관계(linearity)를 보이는지의 여부이다. 선형관계가 아님에도 불구하고 단순 상관분석을 시행하게 되면 실제의 관련성 보다 약하게 나타나므로 진실을 왜곡시킬 수 있다. 또한 대응하는 변수값들의 분포와 관련된 문제로서 한 변수의 각 값에 대응하는 다른 변수의 값들은 서로 정규분포를 이루어야 한다는 것으로 이 가정이 충족되지 못하면 두 변수의 관계가 선형이라고 말할 수 없기 때문이다. 또한 표본의 크기가 작을수록 상관관계에 대한 편차가 심해지기 때문에 표본의 크기는 분석의 가정에 매우 중요하다. 임[13]에 의하면 실제로는 모집단들간의 상관관계가 전혀 없으나 표본의 크기에 따라 표본의 상관계수의 값이 다음과 같은 변화를 보인다고 보고하였다. 즉, 산출된 상관계수가 가지는 값들의 80%가 표본의 크기가 5일 경우 $-0.69 \le r \le 0.69$, 15일 경우 $-0.35 \le r \le 0.35$, 25일 경우 $-0.26 \le r \le 0.26$, 50일 경우 $-0.18 \le r \le 0.18$, 100일 경우 $-0.13 \le r \le 0.13$, 200일 경우 $-0.09 \le r \le 0.09$로 되어 표본의 크기가 작을 때는 실제로 상관이 없음에도 불구하고 상관계수가 높게 나타난다는 것이다. 따라서 무시할 수 있는 정도의 낮은 상관계수의 범위를 대략 $-0.20 \le r \le 0.20$ 으로 볼때 위의 기준에 근거하면 표본의 크기가 적어도 25에서 50 이상이라면 상관계수의 편차는 무시할 수 있으므로 분석이 가능하기 위해 요구되는 최소한의 표본크기는 25 이상이 되는 것이 바람직하다[7,14]. 몇가지 흔히 사용되는 상관계수를 살펴보면 $0.85 \le r \le 0.95$는 높은 양의 상관관계, $0.17 \le r \le 0.23$은 낮은 양의 상관관계, $-0.03 \le r \le 0.02$는 상관성이 없음, $-0.20 \le r \le -0.17$은 낮은 음의 상관관계, $-0.93 \le r \le -0.89$는 높은 음의 상관관계가 있다고 말한다[15].

한편 연구결과에 대한 해석상의 문제로 통계적 유의성과 실제적 유의성 간에 갈등이 빚어질 수 있다. 예를 들어 상관계수 0.08은 미미한 양의 상관계수로서 표본수가 100일 경우에는 유의수준 10%에서도 통계적으로 유의하지 않으나 표본수가 1000인 경우에는 유의수준 1%에서도 유의할 수 있다는 것이다. 따라서 표본수가 1000인 표본에서의 상관계수 0.08은 통계적으로는 유의하다고 하더라도 실제로 두 변수는 연관성이 없는 것으로 해석하는 것이 타당하다. 해석과 관련된 또 한가지 문제는 비록 상관계수가 높다고 하더라도 인과관계적 해석을 하는 등 결과를 확대하여 해석해서는 안되는데 그 이유는 혼란변수(confounding variable)가 개입되어 두 변수간 진정한 관계를 왜곡시키는 경우가 있기 때문이다. 상관분석은 어디까지나 변수간 관련성의 강도를 알고자 하는 것이지 원인과 결과의 관계를 규명하는 것이 아니다[7].

다중비교 : 두 집단의 모평균에 차이가 없다는 귀무가설($H_O : \mu_1 = \mu_2$, $H_A : \mu_1 \neq \mu_2$)은 *t*-검정을 사용하지만 세 집단 이상의 모평균에 차이가 없다는 귀무가설을 검정하는 경우에는 단순히 *t*-검정을 반복하는 접근방법은 잘못된 분석이다. 예를 들어 동일한 평균을 가지고 있는 모집단에서 5개의 표본을 임의로 추출하였다고 가정할 때 단순히 모든 가능한 두 개의 쌍에 대하여 별도로 *t*-검정을 사용하면 매우 심각한 문제가 발생한다. 즉, 이 경우 ${}_5C_2 = 10$개의 쌍이 만들어지므로 다음과 같은 *t*-검정을 10번 시행해야 한다.

$H_O : \mu_1 = \mu_2 = \cdots\cdots\cdots\cdots = \mu_{10}$

H_A : 모든 μ_j가 같은 것은 아니다(j = 1, 2,……10).

각 검정에서 $\alpha = 0.05$로 선택하면 각 검정에서 모평균에 차이가 없다는 가설을 기각하지 않을 확률은 0.95이다. 확률의 승법정리(multiplication law)에 의해 각 검정들이 서로 독립일 때 10개의 검정을 통하여 모든 모평균들이 서로 차이가 없다는 가설을 기각하지 않을 확률은 $0.95^{10} = 0.5987$이 된다. 따라서 차이가 없다는 가설중 적어

도 어느 하나를 기각할 확률은 1-0.5987 = 0.4013이 된다. 이 값이 의미하는 바는 최초에 설정한 $\alpha = 0.05$가 $\alpha = 0.4013$으로 증가(즉, $p = 0.4013$)되어 차이가 없다는 귀무가설을 기각할 확률을 40%나 증가시키는 결과를 초래하므로 분산분석을 사용하게 되는 것이다. 검정결과 귀무가설이 기각되는 경우 이는 "모든 k개의 모든 평균이 서로 다르다"는 것을 의미하는 것은 아니다. 집단간의 차이가 얼마나 큰 지 혹은 집단간 평균이 얼마나 떨어져 있는지는 다중비교를 수행하기 전에는 알 수 없다. 예를 들어 $k = 3$일 때 귀무가설 $H_O : \mu_1 = \mu_2 = \mu_3$가 기각되는 경우 이에 대한 대립가설중 $H_A : \mu_1 = \mu_2 \neq \mu_3$ 인지 아니면 $H_A : \mu_1 \neq \mu_2 \neq \mu_3$이 맞는 것인지 모른다. 결국 분산분석에서 유의한 차이가 없는 경우에는 더이상의 추가적인 분석이 필요없지만 반대로 귀무가설이 기각되면 즉, 처리간에 차이가 있다는 결론을 얻게 되면 "표본추출 변동(sampling variation)으로 기대되는 이상으로 적어도 하나의 군은 다른 군과 서로 다르다"라는 것이다. 다음 단계로 연구자는 어느 군들에서 서로 차이가 있는지에 관심을 갖게 되는데 이를 검정하는 방법을 소위 "다중비교(multiple comparison)"라 한다.

다중비교의 방법으로는 Fisher's LSD(least significant difference), Duncan's new multiple range test, Newman-Keuls test, Tukey's HSD(honestly significant difference) 및 Scheffe's multiple contrast 등이 있다. 한편 Dunnett's test법은 앞의 검정들과 약간 다른데[9,15] 특히 하나의 대조군(control)과 나머지 군들과 상호비교할 때 사용되는 방법으로 예를 들어 대조군을 포함하여 총 5개의 군이 있는 경우 대조군과 나머지 4개 군간에만 비교하는 것이다. 기본적인 5가지 방법중 후자로 갈수록 보수적인 검정법이다. 보수적인(conservative) 검정이라 함은 두 평균치가 매우 큰 차이가 있을 경우에 한해서 유의한 차이가 있는 것으로 판단해주는 검정으로 Scheffe 방법이 이에 해당되며 그 반대인 관대한(liberal) 검정은 약간의 차이만 있어도 유의한 차이가 있는 것으로 판단해주는 검정으로 LSD 검정이 여기에 해당된다. 따라서 보수적인 방법일수록 다른 검정에 비하여 p값이 상대적으로 크게 나온다. 검정의 종류에 따라 개별오류(comparison-wise error)와 모임오류(experiment-wise error)를 조정하여 주므로 적절한 방법을 선택하는 것이 좋다. 한편 비모수적인 방법으로는 평균값 대신 순위합(rank sum)을 이용한 Tukey-type test 및 중위수 검정법 등이 있다.

일반적으로 평균값에 대한 다중비교시 만족해야 하는 가정은 분산분석에서 적용되는 내용과 동일한데 즉, 정규분포(normality) 가정과 등분산(homogeneity of variance)에 대한 가정이다. 비록 이러한 가정에 벗어나는 경우에도 Tukey test는 적합성이 있는 것으로 사용되고 있는데(robustness) 제1형과 제2형 오류에 미치는 영향에 대하여는 잘 알려져 있지 않다. 두가지 가정중 특히 등분산에 대한 가정을 위반하는 경우(heteroscedasticity)에는 모수적인 다중비교방법을 사용하는 것은 매우 위험하다고 보고되어 있다[17,18]. 모든 다중비교에서는 집단간 표본의 크기가 달라도 사용할 수 있지만 동일한 표본인 경우 검정력(statistical power)과 적합성이 최대로 된다고 알려져 있다.

카이제곱 검정 : *t*-검정이나 분산분석과 같이 복수집단간 차이에 대한 검정이기는 하나 종속변수가 명목척도나 순위척도로 측정되었을 때 사용하는 비모수검정법의 하나로 의학잡지에서 가장 흔하게 볼 수 있는 검정의 하나다. 카이제곱 검정은 분석의 성격에 따라 적합성 검정(goodness-of-fit test), 독립성 검정(independent test) 및 동질성 검정(homogeneity test)으로 구분된다[19].

적합성 검정이란 어떤 주어진 표본이 있을 때 그 표본이 이항분포, 포아슨분포 혹은 정규분포 등과 같이 특정한 이론적인 분포와 일치하는지를 검정할 때 사용된다. 예를 들어 어떤 식물유전학자가 새로이 개발한 품종을 배양할 경우 황색과 녹색 잎사귀의 (표현형적) 발현율이 75% : 25%로 기대된다고 하자. 실제로 배양한 결과 84% : 16%로 관찰되었을 때 이러한 관찰결과가 귀무가설이 옳을 때($H_O : 75\% : 25\%$) 기대되는 빈도와 유의하게 다른지를 검정하는 경우가 적합성 검정의 예가 된다. 기대치(우연에 의해 기대되는 값)와 관찰치의 차이가 너무 커서 그것이 우연한 차이에 의한 것이라고 할 수 없는 경우 귀무가설을 기각하게 된다.

독립성 검정이라 함은 자료를 분류하는 두가지 기준들이 서로 독립적인가 하는 귀무가설을 검정하는 것으로 어떤 하나의 기준에 의한 분포가 다른 기준에 의한 분포에 무관하게 동일하다면 이때의 두가지 분류방법은 서로 독립이라고 말한다. 쉽게 말해 행변수(row variable)와 열변수(column vairable)간에 어떤 연관성이 있는지를 조사할 때 사용된다. 예를 들어 출하돈에서 폐렴병변의 정도와 계절이 독립적이라 하면 폐렴병변의 분포가 계절에 따라 동일한 비율로 분포할 것이라는 것을 의미한

다. 실제 검정에서 주로 분할표(contingency table)를 이용하여 검정하며 그 결과 독립이라는 귀무가설이 기각되면 두가지 분류기준간에 연관성이 있다는 결론을 내린다.

동질성 검정이 앞의 두가지 검정과 다른 점은 전자의 두 검정에서는 표본이 뽑혀진 후 두 가지 분류기준에 의해 분류되었다는 특성이 있다. 즉, 각 셀에 해당하는 관찰도수는 표본이 뽑혀진 후에 결정되므로 행과 열의 합계는 연구자가 조정할 수 있는 것이 아니고 우연한 결과라 할 수 있다. 그러나 간혹 연구자가 행과 열의 합계를 조정할 수 있는데 이 경우 하나의 분류기준에 따른 주변합(marginal probability)을 고정시킬 수가 있고 다른 기준에 따라 분류되는 것은 확률적(random)이 된다. 따라서 동질성 검정에서는 분류기준에 따른 그 표본들의 분포가 서로 동질한가를 검정하는 것이다. 예를 들면 개를 3개의 연령군(1년 미만, 1~3년, 3년 이상)에 따라 80, 90, 100마리를 선택하여 지난 1년간 기생충 감염여부를 조사하였다고 하자. 이 경우 각 모집단으로 특정한 숫자를 뽑았기 때문에 이는 행(혹은 열)의 합계를 고정시킨 효과를 가져온다. 검정결과 귀무가설을 기각하게 되면 각 모집단들은 (여기에서는 3개의 연령군) 기생충 감염에 관한한 동질하지 않다는 결론을 내리게 된다. 결국 동질성 검정은 각 표본들의 비율분포가 동일한 것인지를 검정하게 되므로 비율에 관한 검정이라고도 한다. 이 검정은 t-검정이나 분산분석과 유사하며 다만 카이제곱에서의 동질성 검정은 독립변수와 종속변수가 명목척도나 순위척도인 경우에도 사용할 수 있다는 점이다. 실제 검정에서 동질성 검정은 독립성 검정과 동일하지만 그 내용에서는 상이하기 때문에 정확하게 표현해주어야 한다.

카이제곱검정에서도 몇가지 검토해야할 가정이 있다. 첫째, 표본의 크기가 어느 정도 커야 하고 기대도수가 5 이상일 때 검정통계량의 x^2값이 이루는 분포가 실제 x^2값의 분포상의 분포에 접근한다고 알려져 있다[20]. 구체적으로 말하면 자유도가 1보다 큰 분할표상에서 기대도수가 5보다 작은 셀의 비율이 20% 보다 적고, 모든 셀의 기대도수가 1 이상인 경우에 카이제곱검정이 가능하다는 의미이다. Cochran[21]에 의하면 2×2 분할표에서 표본수가 40 이하이면서 기대도수가 5 미만인 경우 카이제곱검정을 사용하지 말고, 표본수가 40 이상일 경우에는 기대도수가 1이라 해도 무방하다고 하였다. 이러한 요건을 충족시키기 위해서는 표본의 크기가 최소한 20 혹은 30 이상이어야 한다[7]. 두번째로 관찰치들의 독립성과 관련한 문제로 각 표본들의 독립성이 위배되는 경우 짝지어진 자료에 대한 분석방법으로 McNemar 검정을 활용해야 한다[22].

유의수준 : 가설검정 과정에서 두 가지 형태의 오류가 있는데 옳은 귀무가설을 기각할 때의 제1형 오류(Type I error, α-error)와 틀린 귀무가설을 수용할 때의 제2형 오류(Type II error, β-error)가 있다. 두 가지 모두 오류이므로 α와 β 오류를 동시에 줄이는 것이 바람직 하지만 두 오류는 서로 반대로 작용하기 때문에 어느 한 쪽을 줄이면 다른 한 쪽이 증가하는 결과가 나타난다. 따라서 한 가지 전략은 α를 고정시킨 상태에서 β를 최소화 하는 방법을 사용하여 결과적으로 검정력(statistical power, 1-β)를 최대로 하는 방법을 선택하며 α = 0.01, 0.05, 0.1이 많이 사용된다. 흔히 95% 신뢰도란 용어를 사용하는데 이는 제시한 신뢰구간이 모평균을 포함할 확률이 95%이고 5%의 오차는 허용한다는 의미이다. 이때 α를 유의수준(significance level)이라 하며 옳은 귀무가설을 기각할 확률로 연구자가 기꺼이 감수할 수 있는 제1형 오류가 초래될 위험을 의미한다. 따라서 오차를 α라 하면 신뢰도는 (1-α) · 100%가 된다. α 수준은 연구내용과 목적에 따라 연구자가 스스로 결정할 문제이지만 가급적이면 일반적으로 많이 사용되는 기준으로 통일하는 것이 좋다. 특히 개인용 컴퓨터에서 통계 패키지를 사용하는 경우 정확한 점추정치를 계산하여 주기도 하는데 이러한 경우에도 세가지 유의수준에서 해석을 하는 것이 바람직할 것으로 사료된다.

p값(p-value)을 정의하자면 귀무가설이 옳을 때 기각역의 임계값(critical value)과 같거나 큰 검정통계량을 얻을 확률로, 계산된 확률의 크기가 유의수준 α 보다 작으면 (즉, 검정통계량의 값이 커질 것임) 귀무가설을 버리고 '통계적으로 유의하다(statistically significant)'라는 표현을 사용한다. 달리 표현하면 귀무가설이 맞을 때 실제 표본값 보다 더 극단적인 값을 얻게 될 확률로서 귀무가설을 기각할 증거의 강도를 측정하는 수단으로 이 값이 작을수록 귀무가설을 기각할 증거는 매우 강해진다. 얼마나 작은 p-value가 통계적으로 유의한 것인가 하는 것인데 원칙은 없고 $0.01 \le p \le 0.05$은 "significant", $0.001 \le p \le 0.01$은 "highly significant", $p < 0.001$은 "very highly significant", $p > 0.05$는 "statistically not significant"라는 표현을 사용한다.

기타 문제점 : 첫째, 표본의 크기가 작음에도 불구하고 모수검정을 고집하는 경향이 있는데 일반적으로 표본의 크기가 충분히 크지 않을 경우 비모수검정을 사용하는 것이 바람직하다. 학자에 따라서는 비모수검정 결과의 일반화에 따른 문제점 때문에 일차적으로 모수검정을 사용한 후 비모수검정을 사용할 것을 권하는 경우도 있지만 이 경우에도 표본의 크기가 어느 정도는 큰 경우라야 한다. 본 조사에 의하면 동물실험에서 3-7 정도의 표본에 대하여 평균과 표준편차를 사용하거나 모수검정을 시도하는 것은 매우 부적절한 분석이다. 또한 동일한 개체를 대상으로 처리(treatment) 전후에서 혈액화학치의 변화양상을 측정한다거나 혹은 처치후 경시적으로 측정한 연속변수형 자료를 *t*-검정이나 단순 분산분석으로 분석하는 것은 잘못이며 이러한 짝지어진 자료에 대해서는 반드시 짝지어진 *t*-검정이나 반복측정 분산분석을 사용해야 한다.

둘째, 제시한 자료가 분명히 정규분포 가정에 부합하지 않음에도 불구하고 원자료(raw data)를 log 변환(특히 체세포수 자료) 등과 같은 방법을 취하지 않고 그대로 사용하는 경우이다. 또한 왜곡(skewedness)이 심한 자료에 대하여 중앙집중성과 산포성의 측정으로 평균과 표준편차를 사용한 분석도 잘못이다[19].

셋째, 실험의 배치와 관련된 문제로 각 군에 할당된 표본의 크기를 명시하지 않은 경우가 많았다. 이는 독자로 하여금 연구결과를 해석하는데 어려움을 줄 뿐만 아니라 모수분석과 비모수분석의 판단기준이 될 수 있으므로 반드시 명시해야 한다. 또한 패키지에 따라 군당 할당된 표본의 크기가 상이한 경우 분산분석에서 사용되는 통계적 절차가 틀린데 예를 들어 SAS의 경우 PROC ANOVA 대신 PROC GLM을 사용해야 한다.

넷째, 진단검사(diagnostic test)의 일치도와 관련된 문제로 가양성(false positive)과 가음성(false negative)을 고려하지 않고 진양성(true positive)과 진음성(true negative)만을 고려하여 두 검사의 일치도를 계산하는 것은 잘못된 분석이다. 즉, 어떠한 진단검사도 100%의 민감도와 100%의 특이도를 갖지 못하기 때문에 가양성과 가음성 결과는 발생할 수 밖에 없으므로 올바른 일치도를 계산하기 위해서는 관찰된 일치율에서 우연히 일치할 수 있는 확률을 빼주어야 하며 이렇게 계산된 결과를 KAPPA 통계량이라 한다[23]. 이 통계량은 0에서 1 사이의 값을 갖고 0은 일치율이 없음을, 1은 완벽한 일치율이 있음을 의미하며 이 값이 클수록 두 검사의 일치도는 높다고 평가한다. 측정자료가 순위형인 경우에는 kappa 대신에 다른 방법을 사용해야 한다.

다섯째, 국내에서 사용되고 있는 자료분석용 통계 패키지는 SAS, SPSS, Epi Info, EGRET. MedCalc, Minitab, BMDP, StatView, GLIM, MacAnova, RATS, SigmaPlot, Statgraphics, S-Plus, XLSTAT, Epicure 및 기타 Fortran 언어로 작성된 프로그램 등 매우 다양하며 분석의 종류에 따라 패키지간 접근방식이 상이한 경우가 있기 때문에 본문에 인용할 때에는 사용한 패키지의 source, version 및 회사를 분명하게 밝히는 것이 좋다. 또한 악어와 관련된 문제로 *t*-test 혹은 Student's *t*-test, ANOVA 또는 F-test 등과 같이 정확하게 표현해야 하며 Duncan, Tukey, Scheffe 등의 첫 글자는 대문자로 표현하는 것이 바람직하다. 다중비교의 하나인 Dunnett's 검정을 *t*-test로 표현해서는 안된다. 또한 유의한 결과에 대하여 인과론적 해석을 하거나 확대해석해서는 안된다. 특히 범주형 자료에 대해서는 유일하게 카이제곱 통계량만이 사용되고 있는데 이에 대한 다양한 분석방법이 있으므로 연구자들의 적극적인 관심이 요구된다.

결 론

본 연구는 수의학회지에 발표된 논문에서 사용된 통계적 기법에 대하여 그 적용의 타당성을 검토하고 평가함으로써 통계기법에 대한 올바른 이해를 정립하고자 1997년부터 1999년 3월호 까지 대한수의학회지에 발표된 총 246편의 논문중 통계처리를 수행하여 분석의 대상에 해당되는 94편의 논문을 대상으로 분석하였다. 통계기법별 적용의 타당성을 평가한 결과는 다음과 같다.

중앙집중성과 산포성의 측정에 대하여 분석이 가능했던 총62편의 논문중 34편(54.8%), *t*-검정의 경우 전체 23편중 9편(39.1%), 상관분석의 경우 3편중 1편, 분산분석의 경우 총16편중 7편(43.8%), 카이제곱검정의 경우 8편중 5편(62.5%)이 올바르게 분석을 하였고 분산분석에서의 일차적으로 유의성이 있는 결과에 대하여 다중비교를 수행하지 않은 논문이 다수 있었다. 유의수준이 적용된 총62편의 논문중 44편(71%)이 정확한 표현을 사용하였다. 본 연구에서는 연구계획에 대한 충분한 설명이 부족한 논문을 제외하였고 또한 평가항목의 선정에서도 근본적인 문제가 있는 내용으로 한정하여 조사하

였기 때문에 다른 연구자들의 결과에 비하여 적절히 평가된 논문이 많았다. 본 연구를 통하여 제기된 통계적 오용과 남용은 근본적으로 통계학에 대한 충분한 사전 지식이 결여되어 나타난 문제로 판단되므로 필요하다면 전문가나 통계연구소를 방문하여 상의하는 것이 도움이 될 것으로 사료된다.

참 고 문 헌

1. Schor S, Karte I. Statistical evaluation of medical Journal manuscripts. *J Am Med Assoc*, 195:1123-1128, 1966.
2. White SJ. Statistical errors in papers in the British Journal of psychiatry. *Brit J Psychiat*, 135:336-342, 1979.
3. 안윤옥. 고응린. 자료처리 과정에 대한 통계적 검토 - 일부 의학잡지에 게대된 논문 예를 중심으로. 예방의학회지, 6:81-85, 1973.
4. 하헌선. 보건학 석사학위 논문에 대한 통계적 평가-서울대학교 보건대학원 석사학원 논문을 중심으로. 서울대학교 석사학위논문, 1984.
5. 현혜진. 보건학 관련 연구논문에 대한 통계기법 적용과 방법론 검토. 서울대학교 석사학위논문, 1990.
6. 이형기. 허봉렬, 안윤옥. 1980년대에 발표된 국내 의학연구 논문의 방법론 및 통계처리 기법의 타당성에 관한 평가 연구. 가정의학회지 12:46-67, 1991.
7. 이선미, 이승욱. 국내 간호학 연구논문에 활용된 통계기법의 타당성 평가 연구. 한국보건통계학회지, 23:42-64, 1998.
8. Brown GW. Standard deviation, standard error, which should we use?. *Am J Dis Child*, 136:937-941, 1982.
9. Zar JH. Biostatistical analysis. 3rd ed, Prentice Hall International Inc, 1996.
10. Rosner B. Fundamentals of biostatistics. *Duxburry press*, 1995.
11. Minium EW, King BM, Bear G. Statistica : reasoning in psychology and education, 3rd ed, John Wiley & Sons, Inc, 1993.
12. Sawilowsky SS, Balir RC. A more realistic look at the robustness an Type II error properties of the *t*-test to departures from population normality. *Psychol Bulletin*, 111:352-360, 1992.
13. 임인재. 통계방법, 박영사, 서울, 1991.
14. 김호정. 사회과학 통계분석, 삼영사, 서울, 1996.
15. Altman DG. Practical statistics for medical research. Chapman & Hall, New York, 1993.
16. Huck SW, Cormier WH, Bounds WG Jr. Reading statistics and research. Harper & Row, Publishers, New York, 1974.
17. Fry JC. Biological data analysis : a practical approach. IRL Press, 1993.
18. Hassard TH. Understanding biostatistics. Mosby, 1991.
19. Daniel W. Biostatistics : a foundation for analysis in health sciences. John Wiley & Sons, Inc, 1983.
20. Kirkwod BR. Essentials of medical statistics. Blackwell scientific publications, London, 1988.
21. Cochran WG. Some methods for strengthening the common x^2 tests. Biometrics, 10:417-451, 1954.
22. Knapp RG, Miller CM. Clinical epidemiology and biostatistics. Harwal publishing Co, Pennsylvania, 1992.
23. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics. 2nd ed, Appleton & Lange, 1994.