대한마취과학회지 2000; 39: 706~711

대한마취과학회지에 게재된 논문의 통계기법에 대한 고찰(1994－1998년)

한림대학교 의과대학 마취과학교실

안 원 식

= Abstract =

Statistical Methods in the Articles in the Korean Journal of Anesthesiology Published from 1994 to 1998

Wonsik Ahn, M.D.

Department of Anesthesiology, College of Medicine, Hallym University, Seoul, Korea

Background: The Korean Journal of Anesthesiology (KJA) was first published in 1968 containing only 16 articles. In 1998, the number is 291. However, the quantitative growth does not mean a qualitative growth. There are many aspects to improving quality. One of them is statistical accuracy. I have examined the statistical methods in our field and checked the accuracy of the methods. Then, I compared the results with the methods in the 1980s and examined what kinds of changes exist.

Methods: I reviewed all the articles except case reports and review articles in KJA published from 1994 to 1998. I focused on the methods of inferential statistics because those kinds of statistics were usually mentioned in the articles. It is based on the mentioned technique in the article to decide which inferential statistics are used, even though sometimes they are not accurate. I adopted the similar statistical error criteria selected by Ko.

Results: Basic statistical error, for example no statistics used even though statistical methods are needed, were dramatically reduced compared to the 1980s. It is increased to use the mean comparison methods correctly, but, some statistical methods are still misused frequently, for example Chi-square test, nonparametric analysis, multiple comparison methods and improperly adopted methods based on the variable scale.

Conclusion: Generally, based on my criteria statistical errors are reduced from about 75% in the 1980s to about 60% in the 1990s. **(Korean J Anesthesiol 2000; 39: 706~711)**

Key Words: Statistics: inferential statistics; statistical error.

서 론

오늘날 통계적 사고와 통계적 기법의 활용은 과학이란 용어를 붙여둔 모든 학문분야에서 널리 사용되고 있는 추세이다. 통계학이 추구하는 바가 우리가 직면한 혼잡스럽고 불확정적인 상황에서 합리적이고 과학적인 방법에 따라 숨겨져 있는 규칙성을 찾기 위한 최선의 의사결정이라고 볼 때 통계적 기법의 사용 증가는 당연한 결과라고 하겠다. 그러나 이러한 통계 내지 통계적 기법의 홍수를 면밀히 관찰해 보면 이용에 많은 문제점을 갖고 있는 것도 사실이다. 통계적 기법을 활용하는 논문에서 부적절한 통

논문접수일 : 2000년 10월 6일
책임저자 : 안원식, 서울시 영등포구 영등포동 94-195
한강성심병원 마취과, 우편번호: 150-030
Tel: 02-2639-5503, Fax: 02-2631-4387
E-mail: aws@plaza.snu.ac.kr

계적 기법의 사용은 연구논문의 성취도를 떨어뜨리거나 결정적인 오류를 초래하게 된다.[1]

이러한 오류에 빠지지 않고, 보다 효율적인 연구나 실험을 행하기 위해서 여러 가지 노력을 하고 있다. 이 노력 중의 하나가 통계적 적용의 적절성을 분석하여 보다 적절한 통계방법 사용을 유도하는 것이다. 현재까지 대한마취과학회지에서 통계적 기법의 오류에 대한 분석을 한 논문은 두 편이 있었는데, 한 편의 논문은 고홍 등이 발표한 기술 통계 기법, 추측 통계기법과 이에 대한 오류 분석이었고,[2] 이윤석 등이 발표한 다른 한 편은 적절한 검정력에 대한 논문이었다.[3] 본 연구의 목적은 이러한 통계적 기법의 사용이 고홍 등이 발표한 1993년 이후에 어떻게 변하였는가를 알아보아 보다 적절한 통계방법 사용을 유도하기 위함이다.

대상 및 방법

대한마취과학회지에 1994년부터 1998년까지(제 27권부터 제 35권까지) 발표된 원저 1,169편을 대상으로 하여, 통계적 처리가 필요 없는 논문이나 기술통계만 사용한 논문은 오류 여부를 검토하지 않고 빈도 수만 세었다. 또한, 사용된 여러 통계기법 중에서 연구 기획단계에서 적용되는 실험계획법, 표본추출법 등의 적절성 여부와 기술 통계의 적절성 검토는 이번 연구에서 제외하였고, 추론통계만을 대상으로 사용된 통계방법과 통계적 오류에 대한 검토를 하였다.

통계방법은 한 논문에서 여러 개의 방법을 사용하였을 경우 이를 모두 세어서 계산하였다. 즉, 한 개의 논문에서 5-6개의 통계방법이 사용되었을 경우 이것을 각각 빈도 수에 포함시켰다. 통계적 오류에 대한 기준은 여러 가지가 있겠으나 이전 연구 결과와 비교할 수 있게 고홍 등이[2] 적용한 통계적 오류의 기준 7개 항목을 대부분 채용하고 '변수의 척도에 부적절한 통계방법' 항목을 '다른 항목에 속하지 않는 부적절한 추론통계방법의 선정'으로 약간 확대하여 다음과 같이 정하였다.

1) 통계적 추론이 필요한 경우에 통계적 처리를 하지 않은 경우.

2) 통계방법의 제시 없이 추론적 결과를 유도한 경우.

3) 보정하지 않은 통계방법을 반복적으로 사용한 경우.

4) 변수의 독립성을 적절히 고려하지 않고 통계처리를 시행한 경우.

5) 불충분한 표본수에서 카이제곱 검정(chi-square test)을 시행한 경우.

6) 분산분석 후 적절한 다중비교가 고려되지 않은 경우.

7) 다른 항목에 속하지 않는 부적절한 추론통계방법의 선정.

오류의 횟수는 한 논문에서 여러 개의 오류가 있을 경우 이들의 빈도를 각각 세어 '통계적 오류 빈도 수'를 계산하였으나, 논문별 오류의 빈도를 계산할 때는 한 개로 처리하였다.

이윤석 등이[3] 지적한 검정력 기술여부는 제 1종 오류와 같은 정도의 중요성을 가지나 아직 연구자들이 이에 대한 인식이 매우 낮아 음성적 결과를 보인 논문들조차도 언급하지 않는 실정이므로 이번 연구에서는 제외하였다.

사용된 한글 통계 용어는 한국통계학회에서 간행한 통계학 용어집을 기준으로 하였다.[4]

결 과

1994년에서 1998년 사이의 원저 중에서 통계가 필요 없는 논문은 8편(0.7%)이었고, 기술통계만이 적용된 논문은 71편(6.1%)이었으며, 추론통계를 적용한 논문은 1,090편(93.2%)이었다(Table 1).

추론통계를 적용한 논문 중 사용된 통계방법은 t-검정, 분산분석으로 대표되는 평균치 비교방법이 제일 많이 사용되고 있으며, 이어서 카이제곱 검정, 비모수적 검정 등이 주로 사용되고 있다. 반면에 상관이나 회귀분석은 상대적으로 적게 사용되고 있다(Table 2). 기타로 분류된 것에는 로지스틱 회귀분석(logistic regression analysis), 판별분석(discrimination analysis), 일반화선형모형(generalized linear model) 등이 있었다.

추론통계를 적용한 논문의 오류를 종류별로 살펴보면, '적절한 다중비교가 없는 분산분석'이 26.9%로 가장 많았고, 이어서 '보정하지 않은 통계방법을 반복적으로 사용한 경우'가 18.6%로 많았다. 다음으로 '변수의 독립성을 적절히 고려하지 않고 통계 처리

Table 1. Statistical Methods Used in the Korean Journal of Anesthesiology Published from 1994 to 1998

Year	Inferential statistics	Descriptive statistics	No statistics	Total
1994	180 (90.0%)	19 (9.5%)	1 (0.5%)	200
1995	191 (92.7%)	14 (6.8%)	1 (0.5%)	206
1996	178 (93.2%)	9 (4.7%)	4 (2.1%)	191
1997	265 (94.3%)	15 (5.3%)	1 (0.4%)	281
1998	276 (94.8%)	14 (4.8%)	1 (0.3%)	291
Total	1,090 (93.2%)	71 (6.1%)	8 (0.7%)	1,169

Table 2. The Incidences of Inferential Statistical Methods Used in the Korean Journal of Anesthesiology Published from 1994 to 1998

	1994	1995	1996	1997	1998	Total
t-test*	44	48	34	58	56	240
Unpaired t-test	41	43	43	62	53	242
Paired t-test	53	44	39	55	50	241
ANOVA†	47	68	62	71	94	342
Repeated measures ANOVA	33	40	42	61	81	257
Correlation analysis	6	4	5	8	12	35
Linear regression analysis	8	12	13	23	23	79
Multiple regression analysis	4	4	2	2	2	14
Mann-Whitney rank-sum test	15	18	13	31	33	110
Wilcoxon signed rank test	5	6	13	19	15	58
Wilcoxon rank sum test	1	3	13	10	8	35
Friedman test	0	0	1	8	4	13
Kruskall-Wallis test	5	10	11	29	31	86
Chi-square test	23	38	41	62	59	223
Fisher exact test	5	11	10	29	32	87
Ridit analysis	1	1	0	2	2	6
Others	1	1	1	2	1	6
Total	292	351	343	532	556	2,074

*I divided the t-test and the unpaired t-test category because some authors used the term, t-test, in case that paired t-test was more appropriate. †ANOVA: analysis of variance

를 시행한 경우'가 7.8%, '불충분한 표본수에서 카이제곱 검정을 시행한 경우'가 6.9%를 차지하였다. 통계가 필요한 경우에도 통계처리를 하지 않거나(1.6%) 통계방법의 제시 없이 추론결과를 유도한 경우(1.2%)는 매우 낮았다. 위의 항목에 포함되지 않는 부적절한 추론통계방법의 선정의 예에는 변수의 척도에 부적절한 통계방법을 사용한 예, 비모수적 검정으로서 변수의 독립성을 고려하지 않은 예(Wilcoxon rank sum test와 Wilcoxon signed rank test)와 상관과 회귀를 혼동한 예 등이 있었다(Table 3).

오류의 빈도를 연도별로 살펴보면, 한 개 이상의 통계적 오류를 포함한 논문이 1994년에 132편(66.0%), 1995년에 144편(69.9%), 1996년에 102편(53.4%), 1997년에 161편(57.3%), 1998년에 159편(54.6%)으로 5년 평균이 약 60% 정도이었다(Table 4).

Table 3. The Classified Incidences of Statistical Errors in the Korean Journal of Anesthesiology Published from 1994 to 1998

Error	1994	1995	1996	1997	1998	Total
(1)	2	4	3	1	1	11 (1.5%)
(2)	1	3	2	0	2	8 (1.1%)
(3)	71	36	39	55	47	141 (18.6%)
(4)	8	11	9	15	16	59 (7.8%)
(5)	5	15	6	8	18	52 (6.9%)
(6)	28	40	24	50	62	204 (26.9%)
(7)	32	42	30	45	27	176 (23.2%)
Total	147	151	113	174	173	758

*The classification of Errors. (1) No statistics used even though statistical methods are needed, (2) Inferential results are induced without describing statistical method, (3) Repeatedly used t-test without correction, (4) T-test not considering variable independency, (5) Chi-square test with inappropriate sample size, (6) Analysis of Variance without appropriate multiple comparison, (7) Inappropriate inferential statistical method used which is not listed in the above items.

Table 4. The Incidences of Statistical Errors in the Korean Journal of Anesthesiology Published from 1994 to 1998

Year	Articles with error	Total article numbers	Percentage (%)
1994	132	200	66.0
1995	144	206	69.9
1996	102	191	53.4
1997	161	281	57.3
1998	159	291	54.6
Total	698	1,169	59.7

고 찰

신뢰할만한 연구결과를 얻기 위해서는 의학 연구자와 통계전문가가 임상시험연구의 준비단계에서부터 실험계획 및 연구의 운용, 분석은 물론 결과의 보고에 이르기까지의 전 과정에 걸쳐서 긴밀하게 협조해야 한다. 그러나 현실적으로는 의학분야와 통계학분야의 연구자들간의 공동연구가 제대로 이루어지지 못하고 있다. 의학 연구자들의 경우에는 의학적 전문지식은 많지만 통계지식의 기반이 단단하지 못하고, 반면에 통계 전공자들은 의학분야에서의 용어나 관례를 비롯하여 임상자료의 특성에 대한 이해가 부족하여, 서로 원활한 의사소통이나 효율적인 연구를 기대하기 어렵기 때문이라고 생각된다.[5] 그럼에도 불구하고 의학발전을 위해서 이러한 거리를 좁혀주는 상호간의 노력은 꼭 필요하다고 생각되며, 이 노력 중의 하나가 의학 논문에 있어서의 통계적 적용의 적절성을 분석하여 의학자들에게 통계 사용의 오류가 있음을 알리고 보다 적절한 통계방법 사용을 유도하는 것이다.

통계학의 여러 분야 중에서 마취과학 연구 논문을 쓸 때 주로 사용하는 통계적 기법으로는 표본(통계)조사론, 실험계획법, 기술통계기법, 추측통계기법을 들 수 있다. 논문을 쓸 때 상기한 4가지 분야 중에서 표본 조사론과 실험계획법 분야가 가장 먼저 검토되어져야 되는 분야이고 이곳에서 표본의 수를 포함한 적절한 설정이 이루어지면 많은 시간과 비용을 절약할 수 있고 보다 효율적인 논문이 될 수 있는 것이지만 현재 마취과에서는 주로 일원배치법 또는 이원배치법에 단순임의 추출법을 사용하여 대부분의 연구를 진행하며, 이러한 방법론적인 기술이 마취과 원저에서 구체적으로 언급되지 않고 있으므로 이번 연구에서는 제외하였다. 또한 기술통계는 오류의 검토 없이 빈도만을 조사하였고, 주로 추측통계기법의 사용상의 오류가 고홍 등이 논문을 발표한 이후 어떻게 달라졌는가를 살펴보았다.

통계가 사용되지 않은 논문은 1980년대에 약 3% 가량을 차지하였으나 1994년에서 1998년에는(이하, 90년대로 칭함) 0.7%로 상대적으로 줄어들었다. 또한, 기술통계만 사용한 경우도 18%에서 6.1%로 감소하였다.

추론통계기법의 사용빈도를 살펴보면 80년대와 90년대 모두 평균치 차이 검정방법(t-검정, 분산분석)이 가장 많이 사용되었다. 그러나, 80년대에 많은 비중을 차지하던 상관과 회귀분석은[2] 90년대에 들어와서 상대적으로 감소하였고, 반면에 카이제곱 검정, 비모수통계 등이 많은 빈도를 보였다. 80년대에 비해 카이제곱 검정과 비모수적 검정의 빈도가 늘어난 것은 실험 계획이 바뀌었다고 생각하기보다는 80년에도 이러한 기법이 적용되어야할 논문에서 다른 기법을 잘못 적용하였다가 90년대에 와서는 적절한 통계기법을 선정하였기 때문으로 생각된다.

추론통계기법의 오류를 종류별로 살펴보면 통계가 필요한 경우에도 통계처리를 하지 않은 오류는 90년대에 1.5%로 80년대의 약 10%에서 많은 감소를 보였다. 또한 통계방법의 제시 없이 추론결과를 유도하는 오류도 90년대에 1.1%로 80년대의 약 50%에서 매우 많은 감소를 보였다. 이러한 오류의 감소는, 어떤 결론을 도출할 때 객관적이고 과학적인 지지 수단으로 통계가 꼭 필요하다는 것을 연구자들이 인식한 결과로 생각된다. 그렇지만 아직도 몇몇 논문에서는 상기 두 가지의 오류를 범하고 있는데 이들은 주로 실험계획이 복잡하고 여러 항목을 측정하는 실험연구 논문에서 발생하였다.

대한마취과학회지에서 가장 많이 사용된 통계방법인 t-검정은, 두 개의 평균치에 대한 통계적 분석에서 많이 쓰이고 있는 방법으로, 여기에는 자료의 독립성 여부에 따라 unpaired t-test와 paired t-test가 있는데 자료가 상호 연관이 되어 있을 때는 paired t-test가 사용된다.[6,7] 만약 종속적인 관측치에 대하여 unpaired t-test를 수행하면 통계적으로 의미있는 것을 의미없다고 하게 되는 β-error가 커질 수 있어 임상적으로 유의한 좋은 결과가 사장될 가능성이 있어 주의해야한다. 또한, 3군 이상간의 비교에서 보정 없이 반복적으로 t-test를 사용하는 것은 의미 없는 것을 의미 있다고 하게되는 α-error가 증가되어 잘못되거나 위험한 결과가 실제로 적용될 위험이 있기 때문에 피해야 한다. 이런 경우에는 사후검정 또는 다중비교를 수행해야 하는데 Tukey, Duncan, Dunnett 등에 의한 방법들이 적용될 수 있다.[2] 보정하지 않은 통계방법을 반복적으로 사용한 경우는 90년대에 18.6%로 80년대의 약 30%보다 많은 감소를 보였다. 이러한 감소는 3군 이상의 비교에 분산분석을 적절하게 사용하는 빈도가 증가하면서 자연스럽게 감소하게 된 것으로 생각된다. 그리고, 변수의 독립성이 고려되지 않은 경우는 90년대에 7.8%로 80년대(4건)보다 많은 증가를 보였다. 이러한 증가의 이유는, 90년대에 실험 계획 중 검정력이 높은 짝지은(paired) 방법을 많이 선택하였으나 아직 통계적 검정법에 대한 인식이 뒷받침해 주지 못한 결과라고 생각된다.

독립변수와 종속변수가 모두 범주형일 때 사용하는 범주형 자료의 분석에는 카이제곱 검정, 로그선형 모형, 로지스틱 회귀분석, 관련도 등이 있는데, 마취과 영역에서는 카이제곱 검정을 주로 사용하고 있다.[8,9] 카이제곱 검정법이 적절히 적용되려면 대상군의 수가 50명 이상은 되어야 하며, 이론적으로 계산된 기대치가 최소한 5 이상은 되어야 한다. 이보다 작은 수의 표본으로는 Fisher의 직접 확률 계산법이 적용될 수 있다.[2] 불충분한 표본수의 카이제곱 검정을 시행한 오류는 90년대에 6.9%로 80년대의 6건보다 별 차이가 없었다. 차이가 없는 이유는 연구자들이 아직은 범주형 자료의 분석법을 적절하게 사용하지 못하는 것으로 생각된다.

분산 분석은 3군 이상간에서 평균치를 비교하는 통계기법으로 최근에 사용이 증가하였는데, 3군 이상간에서 분산 분석을 시행한 후 전체적으로 의미있는 결과가 나오면 사후검정(다중비교)을 실시하여 어느 두 군간의 차이에 의하여 전체적으로 의미 있는 결과가 나오게 되었는가에 대한 검정이 요구된다.[2] 이러한 다중비교를 하지 않은 오류는 90년대에 26.9%가 있었으며 80년대에는 약 10%가 있었다. 오류 증가 원인은 분산분석의 횟수가 80년대에 비해 90년대에 늘어났으나 아직 다중비교에 대한 인식이 부족하기 때문이라고 생각된다.

위의 항목에 포함되지 않는 부적절한 추론통계방법의 선정의 예 중에서 변수의 척도에 부적절한 통계방법을 사용한 예를 고찰해 보자면, 먼저 변수의 종류를 알아야 한다. 변수는 관측의 척도에 따라 범주형 자료와 연속형 자료로 구분하고, 기능에 따라 종속변수와 독립변수로 구분된다. 독립변수가 범주

형 자료이고 종속변수가 연속형 자료인 경우, 통계처리는 주로 평균치의 비교 방법인 t-검정 또는 분산분석을 시행하고 있다. 또한 독립변수와 종속변수가 모두 범주형인 경우 분할표의 분석을 사용하고, 두 변수 모두 연속형인 경우 회귀분석을 쓴다. 그 외에 독립변수가 연속형이고 종속변수가 범주형인 로지스틱 회귀분석이 있다.[8,9] 이러한 통계기법의 적용의 오류와 그밖에 다른 오류가 90년대에 23.2%나 되었고 80년대의 약 5%보다는 많은 증가를 보였다. 이러한 결과는 90년대에 들어와서 척도에 대한 인식이 낮아졌다기보다는 80년대에 주로 t-검정만 사용하다가 90년대에는 다른 적절한 검정방법을 추구하였으나 아직 정확한 검정법을 선택하지 못하여 일어난 결과라고 생각된다.

이번 연구에 제외되기는 하였지만 검정력의 기술은 음성적 결과('차이가 없다'로 표현됨)가 기술된 논문에서 꼭 필요하다.[3,10,11] 왜냐하면 이러한 결과가 어느 정도의 확률을 갖고 도출된 것인지를 알고 필요한 경우 반복실험을 할 수도 있기 때문이다. 역으로 설명하자면 차이가 있다는 양성적 결과에서 이것이 어느 정도의 제 1종 오류(α, P값)하에서 판단된 것이냐에 따라 반복실험을 결정하듯이, 음성적 결과에서는 제 2종 오류(β, 검정력 = $1-\beta$)가 어느 값이냐에 따라 반복실험 여부를 결정하게 된다. 그래서 실험의 기획 단계(결과가 어떻게 나올지 모르는 상태)에서 제 1종 오류와 제 2종 오류를 정하게 되고, 이러한 수치를 결과가 어느 방향으로 나오던 간에 기술해 주는 것이 타당하다고 생각되고, 특히 음성적 결과가 나올 때는 필수적으로 기재해야 된다고 생각한다.

통계적 오류를 포함한 논문 수에 따른 분석을 비교해보면 90년대에도 약 60%의 오류를 내재하고 있어 80년대의 약 75%에서 많은 감소를 보이고 있는 것이긴 해도 아직도 반수 이상이 오류를 포함하고 있는 것으로 나타났다.

결론적으로 추측통계기법의 사용상의 오류가 고홍 등이 논문을 발표한 1993년 이후 많은 질적인 향상이 이루어져서 통계처리가 거의 대부분의 논문에서 기술되었고, t-검정법이나 분산분석의 정확한 사용이 증가하였으나, 카이제곱 검정을 포함한 비모수적 검정법이나 분산분석 후 다중비교, 변수의 척도에 맞는 적절한 통계방법의 선택 등에는 아직 개선할 부분이 많은 것으로 나타났다.

참 고 문 헌

1. 최종후, 이재창: 학술논문과 통계적 기법. 서울, 자유아카데미. 1990, pp 1-7.
2. 고홍, 곽일용, 김광우, 함병문, 최익현: 대한마취과학회지에 게재된 논문의 통계적 분석에 관한 고찰(1981－1990년). 대한마취과학회지 1993; 26: 22-7.
3. 이윤석, 고 홍: 대한마취과학회지 논문 119편의 통계검정 음성적 결과에 대한 사후 검정력 조사. 대한마취과학회지 1999; 36: 286-92.
4. 한국통계학회: 통계학 용어집. 서울, 자유아카데미. 1997.
5. 박미라, 이재원: 임상시험 연구를 위한 통계적 방법. 서울, 자유아카데미. 1996, pp i-ii.
6. 안윤옥, 유근영, 박병주: 의학통계론. 2판. 서울, 서울대학교출판부. 1996, pp 59-86.
7. 유근영, 박병주, 김헌: 의약보건학을 위한 PC-SAS. 서울, 한울아카데미. 1995, pp 81-100.
8. 이태림, 김병수, 이용구: 범주형 자료분석. 서울, 한국방송대학교출판부. 1995, pp 1-9.
9. 유근영: 의학-보건학을 위한 범주형 자료분석론. 서울, 서울대학교출판부. 1996, pp 1-4.
10. Murphy KR, Myors B: Statistical Power Analysis. Mahwah, Lawrence erlbaum associates. 1998, pp 1-19.
11. Cohen J: Statistical power analysis for the behavioral sciences. 2nd ed. Mahwah, Lawrence erlbaum associates. 1988, pp 1-17.