J Vet Clin 26(3) : 189-199 (2009)

Journal of Veterinary Clinics
www.ksvc.or.kr

자료 분석의 기초

박선일[1] · 이영원*

강원대학교 수의학부대학 및 동물의학종합연구소, *충남대학교 수의과대학

(게재승인 : 2009년 6월 18일)

An Introduction to Data Analysis

Son-Il Pak[1] and Young-Won Lee*

School of Veterinary Medicine and Institute of Veterinary Science, Kangwon National University, Chuncheon 200-701, Korea
**College of Veterinary Medicine, Chungnam National University, Daejeon 305-764, Korea*

Abstract : With the growing importance of evidence-based medicine, clinical or biomedical research relies critically on the validity and reliability of data, and the subsequent statistical inferences for medical decision-making may lead to valid conclusion. Despite widespread use of analytical techniques in papers published in the Journal of Veterinary Clinics statistical errors particularly in design of experiments, research methodology or data analysis methods are commonly encountered. These flaws often leading to misinterpretation of the data, thereby, subjected to inappropriate conclusions. This article is the first in a series of nontechnical introduction designed not to systemic review of medical statistics but intended to provide the journal readers with an understanding of common statistical concepts, including data scale, selection of appropriate statistical methods, descriptive statistics, data transformation, confidence interval, the principles of hypothesis testing, sampling distribution, and interpretation of results.

Key words **:** statistical method, data analysis.

서 론

임상의학에서는 환자의 건강이나 생명과 관련된 현상을 다루기 때문에 연구계획, 자료수집과 분석, 결론도출에 이르는 모든 과정이 과학적이고 논리적인 타당성과 자료의 신뢰성이 유지되어야 한다(1,12). 자료 분석은 자료에 대한 이해를 높이고 연구 결과를 가능한 객관적으로 해석함으로써 연구의 질적 향상을 추구할 수 있는 수단이다. 한국임상수의학회지에 발표된 논문에서 적용된 분석기법이 다양해지고 고급 통계기법을 활용하는 사례가 점차 증가하고 있고, 이러한 추세는 통계분석의 필요성에 대한 인식이 보편화되고 있음을 시사한다. 사용자 편의위주로 개발된 통계패키지의 보급이 분석기법의 저변확대에 상당부분 기여한 것이 사실이지만 분석원리에 대한 이해가 부족한 상태에서 패키지를 사용함으로써 분석에서의 타당성이 결여된 논문이 흔히 발견된다. 통계분석을 기술하는 방법이 저자별로 매우 다양하고 부적절한 실험계획을 사용하거나 자료 분석의 오류, 분석결과의 확대해석 등의 오류는 개별 논문의 질적 수준에도 문제가 되지만 나아가 국제수준의 학회지로 도약하는데 장애요인이 아닐 수 없다. 본 논문에서는 수의학 관련 연구자로 하여금 의학통계학에 대한 이해의 수준을 향상시키고자 자료 분석의 기초가 되는 원리를 소개한다.

[1]Corresponding author.
E-mail : paksi@kangwon.ac.kr

결 론

실험설계와 분석에서 고려할 사항

특정한 가설을 증명하기 위한 연구를 계획하고 수행할 때 고려해야할 요소는 매우 많다(12). 분석절차에 대한 이해를 돕기 위하여 사료첨가제의 효과를 평가하는 연구를 가정하자. 실험동물을 세 군으로 분류하여 제1군에는 첨가제 농도 5%, 제2군에는 농도 2%, 제3군에는 첨가제를 투여하지 않은 대조군으로 배정하고, 실험종료시점에서 각 군별 체중을 측정함으로써 사료첨가제의 효과 여부를 판단하게 된다. 이와 같은 실험에서 고려해야 할 사항은 다음과 같다.

첫째, 체중 증가에 기여하는 요인으로 사료를 제외한 모든 조건이 각 실험군에 동일하게 분포하고 있는지를 검토해야 한다(13). 모든 실험동물이 임상적으로 건강하다고 할지라도 개체의 유전적 특성 등에 기인한 고유한 차이로 두 군 간 차이가 우연히 발생할 수 있기 때문이다. 다시 말해 연구결

과로 얻은 군 간 증체율의 차이가 개체특성에 의한 차이인지 아니면 순전히 사료첨가제의 효과에 의한 차이인지를 명확히 하기 위해서는 사료첨가제의 효과에 영향을 미칠 수 있는 모든 요인에 대한 통제가 제대로 이루어졌는지를 검토하는 것이 필수적이다. 실험연구는 연구자가 통제할 수 있는 환경에서 수행되므로 결과에 영향을 미칠 것으로 생각되는 요인이 있다면 블록화 실험계획(block design)을 사용해야 한다. 둘째, 연구 종료시점에서 체중을 1회 측정할 것인지 아니면 일정한 시간 간격으로 체중을 반복하여 측정할 것인지를 결정해야 한다. 동일한 개체에서 체중을 2회 이상 반복측정을 실시하는 경우 분석방법이 달라지며 특히 요인이 2개 이상인 실험에서 반복측정이 이루어진 자료인 경우 두 요인의 상호작용(interaction) 효과를 측정할 수 있는 장점이 있다. 셋째, 실험결과는 체중(비율 혹은 평균)이므로 이러한 변수를 분석하는데 어떠한 분석기법을 사용할 것인지를 판단하는 것이다. 연구결과에 대한 신뢰도를 높이기 위해서는 연구상황에 가장 적절한 방법을 찾아야한다(1,7). 이를테면 약제의 효능을 평가하는 연구에서 약물 투여 전과 후의 수치를 비교하는 실험에 대해서는 짝지은(paired or matched) 분석기법을 사용하고, 두 독립적인 실험군의 평균치를 비교하는 실험이라면 독립(independent) 표본에 대한 분석기법을 사용해야 한다. 짝지은 자료와 그렇지 않은 자료의 분석기법이 다른 이유는 짝을 이룬 자료에서는 어느 한 개체가 대조군과 실험군의 역할을 하므로 자료의 변동성이 상대적으로 감소하기 때문이다. 또한 실험에서 고려한 요인의 개수, 요인 수준의 개수, 종속변수의 척도, 표본크기 등에 따라 분석기법은 달라질 수 있다. 넷째, 측정 자료가 정규분포를 만족하는지를 검토해야 한다. 대부분의 모수분석(parametric analysis)은 정규분포 가정을 근거로 하기 때문에 정규성(normality)을 만족하지 못하면 비모수분석(non-parametric analysis)을 고려해야 한다(4). 다섯째, 실험군간 차이가 있을 때 특정한 처리군 간의 차이를 검증하기 위하여 다중비교(multiple comparison)를 실시한다. 연구의 구체적인 목적이 모든 처리군 간 비교인지 아니면 대조군과 나머지군 간의 비교인지에 따라 다중비교 방법이 다르고 특정한 상황에서는 대비(contrast)를 이용하여 검정할 수 있다. 여섯째, 적정수준의 검정력으로 연구목적을 달성하는데 필요한 최소한의 표본크기를 결정해야 한다. 표본크기는 통계적 오류와 연관성이 있다(1,4,5). 표본크기가 부족하면 실제로 존재하는 중요한 차이를 검출하지 못할 가능성이 높아지며 반대로 표본크기가 너무 많으면 임상적으로 의미가 없는 차이도 통계학적으로 유의한 결과로 나타날 수 있기 때문이다. 일곱째, 가설검정의 형태를 단측(one-tailed test) 혹은 양측검정(two-tailed test)으로 수행할 것인지를 판단한다. 양측검정은 두 군간 차이의 존재유무만을 평가하는 것이고 단측검정은 이러한 차이가 방향성(direction)을 가지고 있는지를 검증하는 것으로 검정의 형태에 따라 귀무가설의 수용여부를 판단하는 기각역의 판단기준이 다르다. 유의성을 판단하는 기준으로 흔히 5%의 유의수준을 사용하지만 연구목적과 내용에 따라 10%, 1% 등으로 조정할 필요가 있다. 여덟째, 분석의 최종단계에서 처리군 간의 차이를 흔히 유의확률로 평가하는데 실험결과에서 나타난 두 군 간 차이에 대한 통계학적 유의성이 반드시 임상적으로 중요하다는 것을 의미하는 것은 아니다. 이를테면 두 군 간 증체율 1%의 차이가 설령 통계학적으로는 유의할지라도 임상적으로는 큰 차이가 없을 수 있기 때문이다.

자료의 특성

변수의 선택과 보정

변수(variable)는 연구자가 관찰하거나 측정한 어떤 특성(혈당 농도, 체중)을 의미하는 용어로 변수를 분류하는 방법은 매우 많다. 이를테면 기능적인 측면에서 독립변수 혹은 설명변수(independent, explanatory)와 종속변수 혹은 반응변수(dependent, response)로 구분할 수 있다. 독립변수는 연구자가 사전에 결정하거나 적절한 실험환경에서 통제할 수 있는 변수이고, 종속변수는 연구자가 실제로 측정하고자 하는 변수로 실험대상으로부터 각각의 독립변수에 대하여 얻는다. 대부분의 통계분석에서는 적어도 하나 이상의 독립변수와 종속변수를 갖는다. 예를 들어 수술 후 환자의 통증을 치료하는데 기존의 치료약제와 새로 개발한 약물의 효과를 비교하는 실험에서 환자가 보이는 증상의 완화정도(종속변수)는 어느 처리(독립변수, 실험군)를 받는지에 따라 다르게 나타날 것이다. 따라서 종속변수에 영향을 미칠 것으로 생각되는 혼란변수(confounding variable)를 실험을 수행하기 이전에 반드시 확인하고 이를 통제할 수 있는 연구방법으로 수행해야 올바른 결과를 얻을 수 있다. 예를 들어 통증을 완화시키는 약물의 효과를 평가하는 실험에서 통증의 강도가 매우 심한 환자를 기존 약제군에 할당하고 상대적으로 경미한 통증을 보이는 환자를 신약 처리군에 할당한다면 신약의 효과는 과대추정된다. 따라서 실험계획 단계에서 혼란변수인 통증의 분포가 실험군에 동일하게 분포하도록 환자를 배정하는 것이 중요하며, 분석단계에서도 이를 보정하는 방법을 사용해야 한다.

자료의 형태

의학에서 접하는 자료의 형태(척도, sale)는 혈청화학 측정치, 항체역가, PCR 검사 결과(양성 혹은 음성), 질병의 중증도에 대한 등급(grade), 적혈구내에 존재하는 기생충의 개수, 수정란 이식에서 성공률, 치료 후 생존시간, 약물투여 후 시간대별 약물의 농도 등 매우 다양하다. 수집된 자료에 대하여 어떤 분석기법을 적용할 것인지를 결정할 때 변수의 개수와 척도, 연구대상 집단의 개수, 표본 간 독립성 여부, 자료의 분포 등 매우 많은 요소를 검토해야 하며 이 중에서도 자료의 척도는 통계기법 선택을 위한 출발점이다(1,4,6).

명목(nominal) 척도는 가장 하위수준의 척도로 산술연산이 불가능한 척도다. 예를 들어 환자의 성별(gender), 품종 (breed), 고양이 혈액형(A, B, AB)은 명목척도에 해당하며 분석의 목적으로 A형을 0, B형을 1, AB형을 2로 코딩할 수 있지만 이러한 순서는 임의적인 것으로 어떠한 정보도 내포하고 있지 않다. 순위 혹은 서열(ordinal) 척도는 명목척도에 순서의

개념이 추가된 자료이지만 순위간의 차이가 불분명하기 때문에 정량적 계산을 할 수 없다. 이를테면 환자가 느끼는 통증의 수준을 등급으로 평가할 경우 에 대하여 통증이 없을 때 1, 경미할 때 2, 보통일 때 3, 중증일 때 4로 코딩한다면 4는 3보다 통증의 정도가 심하지만 이러한 차이가 2와 1의 차이와 동일하다고 할 수 없다. dipstick을 이용한 뇨분석 결과를 +, ++, +++ 등으로 평가한 자료는 전형적인 순위형 척도에 해당한다.

구간(interval) 및 비(ratio) 척도는 이웃하는 수치간의 차이가 일정하기 때문에 모든 산술적인 연산이 가능한 자료다. 구간척도의 경우 진정한 0의 값을 갖지 못한다는 점이 비척도와 다른 점이고 온도는 대표적인 예다. 자료를 수집할 때 가능한 상위척도를 사용하는 것이 바람직하다. 비척도는 필요시 순위나 명목척도로 변환이 가능하지만 명목척도나 순위척도는 상위의 척도로 변환이 불가능하기 때문이다. 또한 분석단계에서 상위척도를 하위척도로 변환하게 되면 정보의 손실이 초래되기 때문에 가능한 해당 척도에 부합하는 통계기법을 사용하는 것이 바람직하다(Table 1, 2).

자료요약

중앙집중성: 자료를 요약하는 중요한 속성의 하나인 중앙집중성(central tendency, location)은 개별 관찰치들이 평균적으로 분포하는 위치를 나타내는 수단으로(8,9,11,13) 자료의 척도에 따라 요약방법이 다르다. 명목척도는 최빈값으로 요약하는 반면 순위척도는 중위수(median, 관치를 오름차순으로 정렬할 때 중앙에 위하는 값으로 관찰치의 개수가 짝수이면 중앙에 위치하는 2개 관찰치의 평균값)를 사용한다. 자료의 분포가 대칭이고 이상점(outlier)이 없는 정규분포를 보이는 연속형 자료는 산술평균(mean)을 사용하지만 분포가 비대칭이거나 이상점이 있다면 중앙값을 사용해야 한다. 명목척도를 갖는 자료는 관찰빈도를 이용하여 백분율(%)을 사용하고 순위척도는 중위수나 백분율(%), 구간척도 이상의 자료는 평균이나 중위수를 사용한다. 산술평균은 이상점에 큰 영향을 받지만 중위수는 이와 무관하기 때문에 변동성이 매우 큰 자료를 요약할 때 유용하다.

산포성: 산포성(spread, dispersion, variability)은 개별 관찰치가 평균으로 근접해 있는 정도를 평가하는 수단으로 표

Table 1. Some commonly used statistical tests according to the type of data

Independent		Dependent		
		Continuous	Ordinal	Nominal
Continuous		Linear regression Multiple regression Multiple correlation ANCOVA	Spearman rho Kendall tau Kendall W	Cox regression Logistic regression Discriminant analysis
Nominal	Univariate	t-test	Mann-Whitney U Fisher Mantel-Haenszel χ^2 Cohen's Kappa	χ^2 test
	Multivariate	analysis of variance	Kruskall-Wallis χ^2 test Mantel-Haenszel χ^2	Log rank

ANCOVA=analysis of covariance

Table 2. Common statistical tests for independent or paired data

Type of data	Measures of location	Measures of variability	Statistical tests for independent groups of data	Statistical tests for paired groups of data
Nominal	Mode	None	2 groups: χ^2 ≥3 groups: χ^2	2 groups: McNemar ≥3 groups: Cochrane Q
Ordinal	Median, mode	Range, IQR	2 groups: WRS ≥3 groups: KW	2 groups: WSR ≥3 groups: Friedman 2-way ANOVA
Interval/ratio	Mean, median, mode	Range, IQR, SD	2 groups: t-test ≥3 groups: ANOVA	2 groups: paired t-test ≥3 groups: 2-way repeated ANOVA

IQR=interquartile range, SD=standard deviation, x2=chi-square test, WRS=Wilcoxon ran sum test, KW=Kruskal-Wallis test, ANOVA=analysis of variance, WSR=Wilcoxon signed rank test

Table 3. Applicability of measures of central tendency and variability by data scale or use

Scale or use	Central tendency			Variability			
	Mean	Median	Mode	Range	IQR	SD	SEM
≥ interval	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ordinal	No	Yes	Yes	Yes	Yes	No	No
Nominal	No	No	Yes	No	No	No	No
Affected by outlier	Yes	No	No				
Presenting sample variability				Yes	Yes	Yes	No
Using in statistical inference				No	No	Yes	Yes

IQR=interquartile range

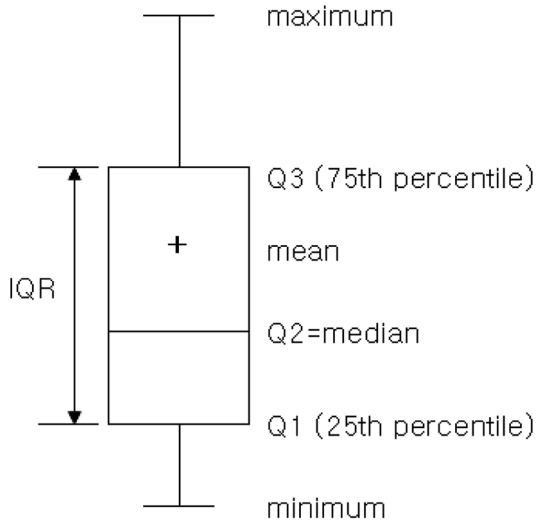

Fig 1. Box-whisker plot. IQR=interquartile range (the difference between 25% and 75% percentile rank).

준편차(standard deviation), 범위(range), 사분위 범위 (interquartile range, 25% 백분위수와 75% 백분위수의 차이) 등을 사용한다. 범위는 이상점이 있는 자료의 경우 두 극단값 사이에 위치하는 값에 대한 아무런 정보를 제공하지 못하기 때문에 적절한 수단이 아니다. 중위수는 전체 자료를 동일한 크기의 두개 군으로 구분하고, tertile은 세 개, quartile은 네 개, quintile은 다섯 개로 구분하여 요약하는 수단이다. 사분위범위는 자료를 4등분할 때 첫 1/4 등분(Q1, 25% 백분위수)과 3/4등분(Q3, 75% 백분위수)의 범위로 중앙에 관찰치의 50%가 포함된다는 것을 의미한다. 중위수와 마찬가지로 사분위범위는 이상점에 영향을 받지 않기 때문에 비대칭분포를 보이는 자료를 요약할 때 유용하다(8,9,11,13). 자료의 척도와 용도에 따른 중앙집중성과 산포성의 활용수단을 정리하면 Table 3과 같다. 정성자료는 흔히 막대그래프를 이용하여 요약하며 정량자료는 히스토그람(histogram)이나 box plot을 사용하며 최소값, Q1, 중위수, Q3, 최대값 등 다양한 통계량을 제공하기 때문에 매우 유용하다(Fig 1).

두 실험군의 평균이 동일하더라도 산포성이 다르면 두 자료는 전혀 다른 별개의 분포가 되기 때문에 자료의 요약통계량으로 중요한 의미를 갖는다. 예를 들어 Fig 2와 Table 4에서 두 집단의 평균은 25로 동일하지만 표준편차는 실험군 A에서는 10.8(중위수 23)이지만 B에서는 20.7(중위수 20)로 약 2배의 차이를 보인다. 특히 실험군 B의 관찰치 중 70이라는 이상점이 존재하여 정규분포에서 벗어날 경우에는 산술평균 보다는 중위수를 사용하는 것이 바람직하다.

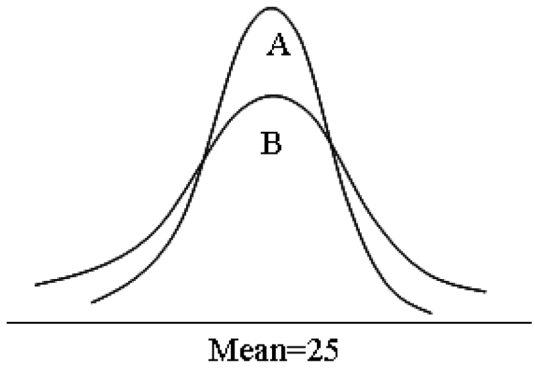

Fig 2. Two standard distributions with different standard deviation (SD).

Table 4. Hypothetical data from group A and B

	Group A	Group B
Data	12, 32, 45, 23, 18, 25, 20, 25	8, 15, 22, 20, 70, 26, 14, 25
Mean	25	25
SD	10.8	20.7
Median	23	20

모집단 신뢰구간: 정규분포 자료에 대하여 모집단의 평균과 표준편차(SD)를 안다면 신뢰구간(confidence interval)을 계산할 수 있다(표본분포 참고). 즉 정규분포를 만족하는 경우 자료의 95%는 평균으로부터 1.96SD에 위치한다. 예를 들어 평균이 100이고 표준편차가 20이라고 할 때 관찰치의 약 68%는 80과 120 사이에 위치함을 의미한다. 다른 예로 어느 모집단의 개로부터 혈청 AST(aspartate aminotransferase, U/L) 농도를 측정한 결과 평균 40과 표준편차 4를 얻었다면 AST 농도의 95% 신뢰구간은 $40 \pm 1.96 \times 4 = [32.2,\ 47.8]$로

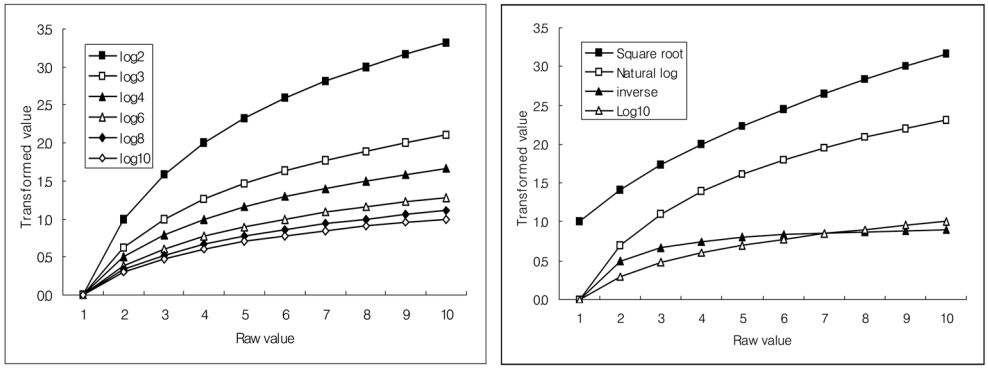

Fig 3. An example of data transformation.

계산된다. 즉 AST 농도의 95%는 32.2-47.8 사이에 위치한다. 왜곡분포를 보이는 자료에서 정규성을 검토하지 않고 신뢰구간을 계산하면 흔히 구간의 하한값(lower limit)이 0 이하의 비논리적인 값을 얻는 경우가 있는데 이는 자료의 특성을 고려하지 않고 계산하였기 때문이다. 이 경우 적절한 자료변환을 통하여 정규성을 만족하는 변환된 자료에 대하여 신뢰구간을 계산하고 이를 다시 원척도로 환산하여 구간을 제시해야 한다.

자료변환

자료는 대칭형이거나 분포의 양 극단으로 긴 꼬리를 갖는 비대칭형(asymmetrical) 분포를 보일 수 있다. 비대칭의 형태에 따라 오른쪽으로 긴 꼬리를 이룰 때 양성왜곡(positive, right skewness)이라고 하며 이는 대부분의 관찰치들이 작은 값을 갖지만 일부 관찰치는 매우 큰 값을 갖는다는 것을 의미한다. 양성왜곡을 보이는 자료에서 분포의 오른쪽에 위치하는 이상점들이 평균에 영향을 주기 때문에 중위수는 항상 평균 보다 더 작다. 한편 왼쪽으로 긴 꼬리를 이루는 음성왜곡(negative, left skewness) 자료에서 중위수는 평균 보다 항상 크다. 이와 같이 왜곡된 분포에 대한 요약통계량으로 평균은 대표값으로 적절하지 못하기 때문에 정규성을 만족시키기 위하여 자료변환(transformation)을 고려할 필요가 있다. 특히 도수로 측정되는 자료는 정규분포를 따르는 예가 매우 드물고 심지어 연속형으로 측정된 자료도 정규분포를 따르지 않는 경우도 많다. 따라서 정규분포를 만족하지 못하는 경우 자료변환을 통하여 정규분포로 변환하는 것이다. 자료변환은 종속변수의 선형성(linearity), 분산의 안정성(homoscedasticity, variance stabilization)을 유지할 목적으로 시도하며 특히 분산분석에서 실험조건하에서 관찰치가 상당히 왜곡되어 있거나 평균이 분산과 연관(correlated)되어 있는 경우에도 사용된다. 도수자료에 대한 제곱근 변환이나 비율자료에 대한 arcsine 변환은 전형적인 예다. 특히 양성왜곡을 보이는 자료를 제곱근(square root), 대수(logarithm) 혹은 역수(inverse) 등과 같은 자료변환을 시도하면 평균이 요약통계량으로 적절한 경우가 많다(Fig 3).

중금속에 폭로된 마우스에서 혈중 효소농도를 측정하여 히스토그램으로 정리한 자료가 Fig 4(A)와 같다고 할 때 이 자료는 분포의 우측에 이상점이 있는 양성왜곡분포(평균 21.8, 표준편차 12.1)를 보인다. 대수변환(Fig 4B)에서는 음성왜곡을 보이므로 적절한 변환이 아니고 제곱근 변환(Fig 4C)에서는 정규분포에 근사함을 알 수 있다 (Shapiro-Wilk, $p > 0.05$) (Table 5, Fig 5). 음성왜곡 자료는 양성왜곡 자료로 변환한 후 양성왜곡에 적용하는 변환 방법을 사용할 수 있다. 변환된 자료를 사용하여 분석을 수행하고 결과를 보고할 때에는 원자료의 척도로 환산해야 한다 (back transformation). 전술한 예에서 제곱근 변환된 자료의 평균이 4.4라면 원 척도로는 19.4가 된다. 대수 변환된 자료를 원 척도로 환산하기 위해서 antilog를 사용하며 이를 기하평균(geometric mean)이라고 한다.

모수분석과 비모수분석

통계적 추론에 사용되는 대부분의 모수기법은 모집단이 정규분포를 따른다는 가정에서 유도된 것이므로 자료가 정규분포, t 분포, F 분포, $\chi 2$ 분포(저자에 따라 비모수분석으로 분류함)와 같이 특정한 이론적 분포에 근사하다는 가정을 만족하면 모수기법을 사용할 수 있다. 표본분포의 특성과 관련하여 중심극한정리(central limit theorem)에 의하면 첫째, 정규분포를 따르는 모집단에서 선발된 표본평균은 정규분포를 따르고 둘째, 모집단이 정규분포하지 않더라도 표본크기(sample size, n)가 충분히 크면($n > 30$) 표본평균은 정규분포에 근사한다. 전술한 사료첨가제에 관한 예에서 각 실험군의 체중 자료가 정규분포 가정을 만족하면 z 혹은 t 검정을 사용할 수 있다. 정규분포를 만족하지 못한다면 원자료에 대하여 모수분석을 적용할 수 없고 자료변환 등과 같은 방법으로 정규성을 충족시킨 후에 사용하거나 분포에 대한 가정을 전제하지 않는 비모수분석을 사용하는 것이 바람직하다(3). 비모수분석은 특정한 분포를 전제로 하지 않고 단지 수집된 자료에 근거하여 분석하는 기법이다. 일반적으로 비모수분석에 비하여 모수분석의 검정력(statistical power)이 높기 때문에 연구자들이 선호하는 경향이 있다. 모수분석과 이에 상응하는 비모수분석을 정리하면 Table 6과 같다.

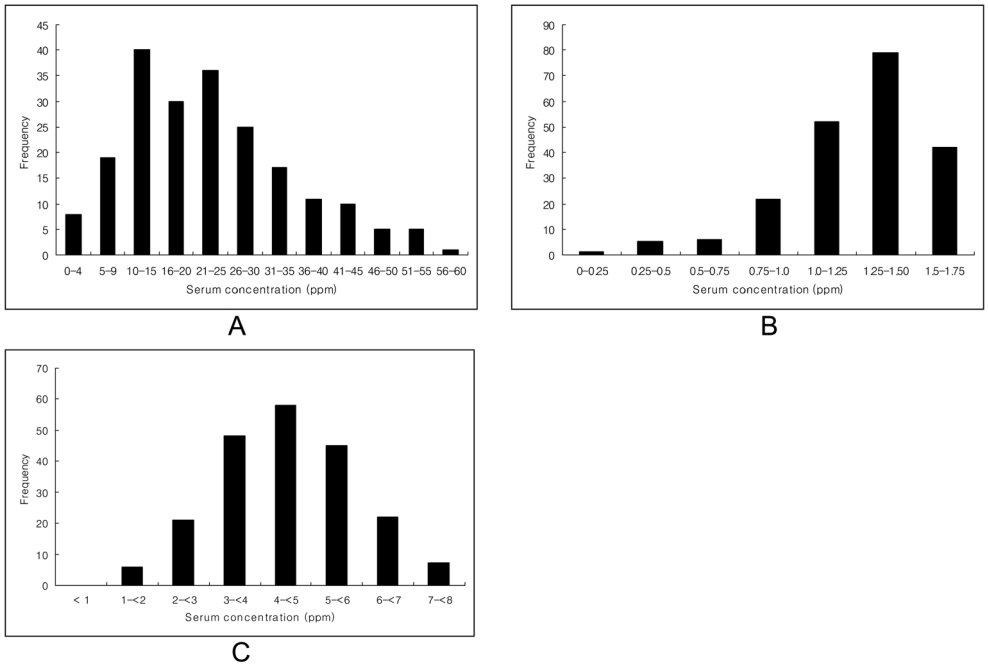

Fig 4. Transformation of data. A: Raw data, B: logarithmic transformation, C: square root transformation.

Table 5. Normality test of raw and transformed data

	Raw data (x)	1/x	Log(x)	Square root(x)
Skewness	0.668	-5.573	-1.130	-0.053
Normality (SW)	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$	$p = 0.4957$

Table 6. Parametric analysis and their corresponding non-parametric analysis

Parametric analysis	Nonparametric analysis
t-test	Wilcoxon rank-sum test
	Median test
	Kolmogorov-Smirnov test
	Wald-Wolfowitz test
Paired t-test	Wilcoxon signed-rank test
	McNemar test
One-way ANOVA	Kruskal-Wallis test
Two-way ANOVA	Friedman test
Pearson correlation	Spearman rank correlation

표본분포

모집단(population)은 연구자가 궁극적으로 관심을 두는 집단이지만 모집단을 전체를 조사하는 것은 불가능하기 때문에 모집단에서 일정한 크기의 실험대상을 무작위로 선발된 표본(sample)을 대상으로 실험하며 이러한 표본은 모집단의 특성을 대표하는 부분집합이 된다. 여기에서 무작위(random)라 함은 모든 통계분석의 전제조건인 확률표본을 의미하며 처리군 할당 측면에서도 각 실험단위가 특정한 처리(treatment, factor)를 받을 확률이 모두 동일하다는 것을 의미한다. 연구자의 주된 관심사는 표본 자체보다는 표본이 유래한 상위 모집단에 대한 추론이기 때문에 모집단을 대표하는 표본을 선발하는 것이 핵심이다. 이러한 확률표본으로부터 얻은 경험적 정보 즉 통계량(statistic)에 근거하여 이에 대응하는 모집단의 특성 즉 모수(parameter)를 올바르게 추정할 수 있는 것이다. 예를 들어 50두의 돼지에 대한 특정 사료

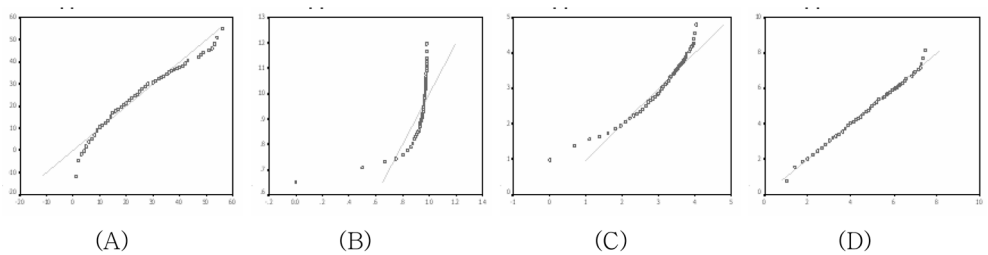

Fig 5. Quantile-quantile (QQ) plot.

'X'의 효과를 평가하는 연구에서 연구자는 'X'의 효과가 연구대상인 50두에 대한 효과에만 관심을 두는 것이 아니라 표본에서 얻은 결과에 근거하여 상위의 돼지 모집단 전체에서 사료 'X'의 효과가 있는지를 추정하는데 관심을 두기 때문에 확률표본이 모집단을 대표해야하는 특성은 매우 중요하다.

표본에서 얻은 통계량으로부터 모수를 추정할 때 언제나 오차가 개입되며 표본통계량과 모수와의 차이를 오차(error)라 한다. 이러한 오차는 표본오차(sampling error)와 비표본오차(non-sampling error)로 구분할 수 있는데 전자는 모집단의 일부분으로 선발한 표본을 대상으로 얻은 결과에 근거하여 모집단의 특성을 추정하기 때문에 발생하는 오차이기 때문에 전수조사에서는 표본오차가 발생하지 않는다. 한편 비표본오차는 표본추출과는 무관하게 발생하는 오차로 흔히 측정과 관련된 기술적인 문제 등에 기인하며 전수조사에서도 이러한 오차는 여전히 발생할 수 있다. 따라서 자료를 분석하기 위해서는 이러한 표본분포의 특성을 정확히 이해하는 것이 중요하다(13).

표준오차: 연구자가 실험대상으로 선정한 표본은 모집단에서 추출 가능한 모든 표본 중 하나의 표본에 불과하다. 모집단에서 동일한 표본크기를 갖는 모든 가능한 표본을 확률적으로 선발할 경우 각각의 표본들은 서로 다른 평균과 표준편차를 보이기 때문에 특정한 표본에서 얻은 통계량이 반드시 모집단의 특성을 대표한다고 할 수 없다. 따라서 표본 추정치에 내재된 변동성을 검토하는 것이 중요하다. 표준편차가 표본평균과 개별 관찰치 간의 편차를 측정하는 수단이듯이 표본평균들의 표준편차는 모집단 평균과 개별 표본평균 간의 편차를 측정하는 수단이다. 즉 표본평균들의 표준편차는 표본평균들의 변동성을 나타내는 지표이므로 표본의 표준편차와 구별하기 위하여 이를 표준오차(standard error of the mean, SE, SEM)라고 한다. 요약하면 표준편차는 단일 표본에 대하여 평균과 개별 관찰치 간의 변동성을 나타내는 지표이지만 표준오차는 단일 표본에서 모집단 평균의 참값을 추정할 때 추정치의 불확실성(uncertainty)에 대한 정밀도(precision)를 나타내는 용어로 표본평균들의 표준편차를 의미한다. 표준편차와 마찬가지로 표준오차가 크다는 것은 표본평균들의 변동성이 매우 크고 모집단 평균과는 상당한 차이가 있음을 의미한다. 표준오차는 모집단에서 개별 관찰치 간의 변동성과 표본크기에 좌우되며 다음과 같이 계산된다.

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{SD}{\sqrt{n}}$$

여기에서 σ는 모집단의 표준편차이고 *n*은 표본크기이다. 모집단의 표준편차가 알려진 경우가 거의 없기 때문에 표본의 표준편차를 편의적으로 사용하게 된다. 이 공식에서 보듯이 표준오차는 표본의 표준편차에 비하여 항상 작은 값으로 계산되는데 그 이유는 개별 관찰치 간의 변동성에 비하여 표본 평균들 간의 변동성이 더 작기 때문이다. 또한 표본크기가 증가하면 표준오차는 감소한다는 것을 알 수 있는데 이는 표본크기를 증가시키면 모집단의 특성을 보다 정확하게 추정할 수 있다는 의미로 해석할 수 있다(표본평균의 신뢰구간 참고). 따라서 연구목적을 달성할 수 있는 최소한의 표본크기를 대상으로 실험할 때 추정치의 정확도를 담보할 수 있다.

정규분포

정규분포의 가장 중요한 특성은 평균과 표준편차(SD)로 정의된다는 점이다(13). 즉 평균은 관찰치의 최고치를 보이는 위치를 결정하며 표준편차는 분포의 모양을 결정한다. 정규분포에서 자료의 68.3%는 평균으로부터 1SD(즉 평균 − SD와 평균 + SD 사이), 95.4%는 평균으로부터 2SD(즉 평균 − 2SD와 평균 + 2SD 사이), 99.7%는 평균으로부터 3SD (즉 평균 − 3SD와 평균 + 3SD 사이) 이내에 위치한다. 연구자들이 흔히 사용하는 95% 신뢰수준은 평균으로부터 1.96SD (즉 평균 − 1.96SD와 평균 + 1.96SD 사이)에 위치한다(Fig 6). 정규성을 검토하는 방법은 히스토그램이나 box-plot 등 그래프를 이용하여 분포가 종모양(bell-shaped)을 따르는지를 판단하는 방법, quantile-quantile(QQ) plot을 작성하여 관찰 자료가 이론상의 정규분포와 일치하는지를 판단하는 방법, Shapiro-Wilk, Kolmogorv-Smirnov, Anderson-Darling 등과 같은 검정법을 사용할 수 있다.

신뢰구간: 표본으로부터 계산된 모집단 평균 추정치를 점추정치(point estimate)라고 하며 신뢰구간(confidence interval)은 모집단의 모수가 포함될 구간을 표본으로부터 확률적으로 추정한 구간이다. 예를 들어 모집단의 평균을 95% 신뢰구간으로 제시하였다면 이는 표본을 동일한 방법으로 100회 반복 추출하여 신뢰구간을 계산할 때 100개의 신뢰구간 중 95개는 모평균을 포함한다는 것을 의미한다. 여기에서 신뢰구간이 모집단의 평균을 포함할 확률 95%를 구간추정의 신뢰수준(confidence level)이라고 한다(10,12,13).

표본평균들이 정규분포를 따른다면 표본평균의 신뢰구간 즉 모집단 평균의 참값이 알려져 있지는 않지만 표본으로부터 모집단 참값을 포함할 구간을 추정할 수 있다. 표본평균의 95% 신뢰구간은 평균 ± 1.96SE로 계산된다. 실제로 모집단 평균의 참값을 알지 못하기 때문에 계산된 신뢰구간의 어느 위치에든지 모집단 평균이 포함할 확률을 95% 수준에서 신뢰하는 것이며, 만일 99% 신뢰하기를 원한다면 평균 ± 2.58을 사용한다. 예컨대 특정 질병에 이환된 216두의 환자를 대상으로 혈청 알부민 농도를 측정한 결과 평균 34.46g/d*l*, 표준편차 5.84g/d*l*를 얻었다면 표준오차는 $5.84/\sqrt{216} = 0.397$ 으로 계산된다. 이는 모집단으로부터 동일한 크기의 표본을 반복하여 선발할 때 평균이 34.46g/d*l* 이고, 표본평균의 약 68% (1SE)는 [34.06 − 34.86]g/d*l* 사이에 있을 것으로 기대한다는 의미이다. 다른 예로 25두의 개를 대상으로 혈청 AST 농도를 측정한 결과 평균이 40이고 표준편차가 4라고 하면 SE = 0.8이므로 95% 신뢰구간은 $40 \pm 1.96 \times 0.8 = [38.4,\ 41.6]$로 계산된다. 이 구간은 모집단의 신뢰구간 [32.16, 47.84]에 비하여 상당히 좁은 구간으로, 표본크기를 100두로 증가시키면 95% 신뢰구간은 $40 \pm 1.96 \times 0.4 = [39.21,\ 40.78]$로 더욱 좁아진다. 예를 들어 어느 혈

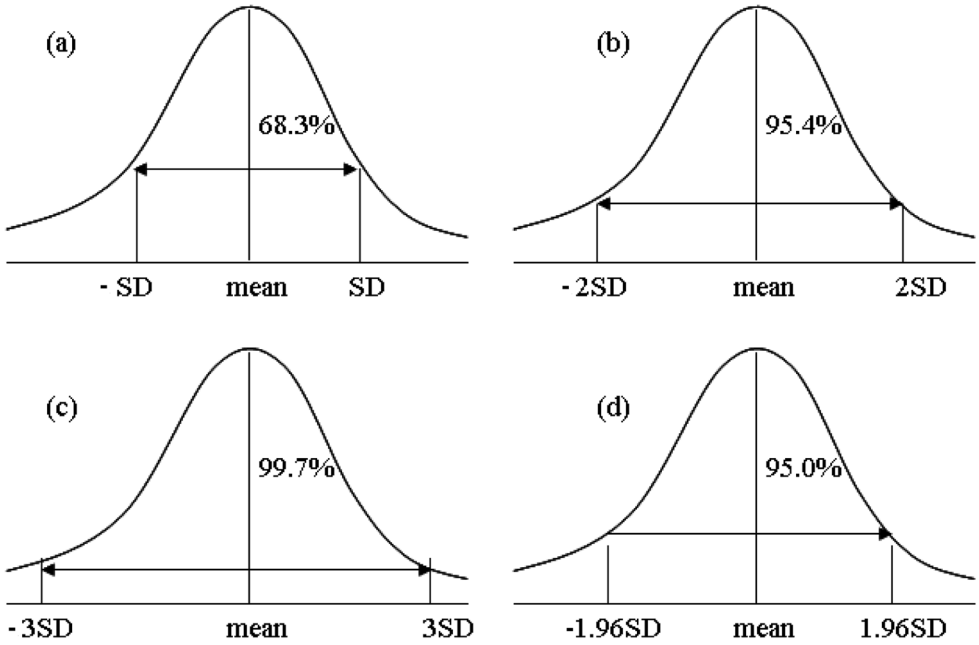

Fig 6. Normal distribution and percentage of area under the curve.

Table 7. Standard deviation (SD) and standard error (SE)

AST	12, 34, 25, 26, 31, 47, 42
Mean	31
SD & CI	SD = 11.6
	2SD: 31 ± 2*11.6 ↔ [7.8~54.2]
SE & CI	SE=4.4
	1.96SE: 31 ± 1.96*4.4 ↔ [22.4~39.6]

CI = Confidence interal

액화학 항목의 표준편차가 6인 어떤 집단에서 표본크기 10, 25, 100인 개체를 선발할 때 표준편차는 $6/\sqrt{10}$ = 1.9, $6/\sqrt{25}$ = 1.2, $6/\sqrt{100}$ = 0.6 으로 표본크기가 증가할수록 표준오차는 감소한다. 자료의 산포성이 클수록 신뢰구간이 넓어지며 이는 모집단의 특성(평균, 비율 등)을 추정함에 있어 표본이 정밀하게 추정하지 못함을 의미한다.

신뢰구간을 해석할 때 주의해야 할 것은 모든 신뢰구간은 표본이 선발된 해당 모집단과 관련하여 해석해야 한다는 점이다(13). 예를 들어 동물병원 'X'에 심장질환으로 내원한 환자를 대상으로 혈중 AST 농도를 측정하여 신뢰구간을 계산하였다면 동물병원 'X'에 심장질환으로 내원한 환자 모집단과 관련하여 해석하는 것이 타당하며 다른 병원에서 심장질환으로 내원한 환자로까지 확대해석하는 것은 바람직하지 못하다.

표준오차는 표준편차 보다 항상 작게 계산되므로 측정 자료의 신뢰성이 좋도록 보이기 위하여 기술통계량으로 표준오차를 사용하는 경우가 있는데 이는 잘못된 분석이다. 연구자가 기술통계량을 제시할 목적이라면 표준편차를 사용하고 추정이 목적일 경우에는 표준오차를 사용하는 것이 올바른 방법이다. 예를 들어 7두의 개를 대상으로 혈청 AST 농도를 측정한 결과 Table 7과 같다고 할 때 연구자는 이 자료에 대하여 AST 농도가 31 ± 23.2(SD) 범위를 보였다고 기술하거나 31 ± 8.6(SE)으로 기술하는 경우가 있다. 전자는 혈청 AST 농도의 기술통계량으로 관찰 자료의 범위에 대한 95% 신뢰구간(평균 ± 2SD)으로 7.8-54.2를 제시한 것이다. 반면에 후자는 표본평균의 범위를 신뢰구간 22.4-39.6로 제시한 것이므로 두 요약방법이 의미하는 바는 전혀 다르다. 다시 말해, 표준편차의 신뢰구간은 AST 농도가 정규분포를 가정할 때 관찰치의 95%는 평균 ± 2SD 이내에 있으므로 혈청 AST 농도의 범위는 7.8-54.2인 반면에, 표준오차의 신뢰구간은 7두의 표본이 선발된 상위 모집단에서 AST 평균 농도의 95% 신뢰구간이 22.4-39.6임을 의미한다. 자료의 기술통계량으로 후자의 방법을 사용하면 모집단에서 혈청 AST 농도의 범위가 매우 좁은 것으로 오해를 초래할 수 있기 때문에 표준편차와 표준오차는 분명히 구분하여 사용해야 한다.

t 분포

전술하였듯이 95% 신뢰구간은 표본평균들의 분포가 정규분포를 따르고 모집단의 표준편차가 표본의 표준편차에 근사하다는 가정이 성립할 때 가장 이상적으로 적용할 수 있다. 첫 번째 가정과 관련하여 표본크기가 충분히 큰 상황에서는 대부분 유효하게 사용할 수 있지만 모집단의 분포가 매우 심한 비정규분포를 보이고 표본크기도 매우 작을 때 즉 소 표본(small sample)인 경우 Student's t 분포를 사용한다(13). 이 분포는 대칭이고 단봉형이라는 점에서 정규분포와 유사하지만 분포의 중심이 낮고 양 극단 방향으로 긴 꼬리

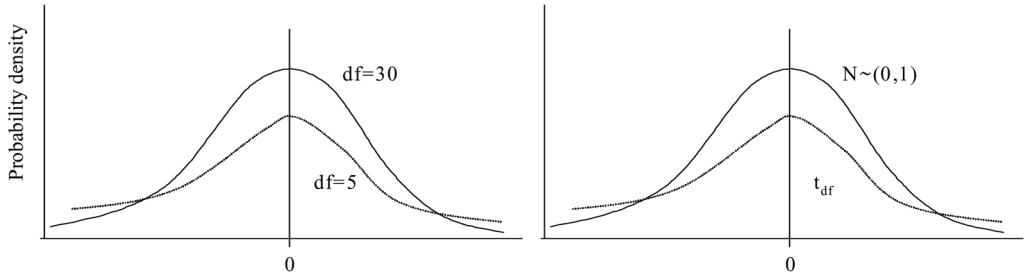

Fig 7. General t-distribution(left: various degrees of freedom, right: normal distribution and t-distribution).

Table 8. Comparison of 97.5 percentile of Student's t and normal distribution

Degrees of freedom(df)	$t_{df,\ 0.975}$	$z_{0.975}$	Degrees of freedom(df)	$t_{df,\ 0.975}$	$z_{0.975}$
4	2.776	1.96	10	2.228	1.96
20	2.086	1.96	24	2.064	1.96
30	2.042	1.96	60	2.000	1.96
200	1.972	1.96	∞	1.96	1.96

Table 9. Type of error in hypothesis testing

		True state of nature	
		No effect (Null hypothesis)	Effect present (Alternative hypothesis)
Conclusion of statistical analysis	No effect	Correct (level of confidence)	Wrong (type II error, β)
	Effect present	Wrong (type I error, α)	Correct (statistical power = 1 – β)

를 갖는다는 점이 다르며 분포의 정확한 모양은 자유도(표본크기-1)에 좌우된다(Fig 7). 표본크기가 클수록 t 분포의 신뢰도 계수는 1.96으로 점차 작아지기 때문에 결국 정규분포에 근사해진다. Table 8은 몇가지 자유도에 대하여 정규분포와 t 분포의 상위 2.5 퍼센타일 값을 비교한 것으로 표본크기가 작을수록 두 분포의 차이는 더 커짐을 알 수 있다.

신뢰구간: t 분포를 이용하여 신뢰구간을 계산하는 방법은 정규분포에서와 동일하다. 즉 정규분포에서 95% 신뢰구간을 계산할 때 표본평균들의 95%는 모집단 평균의 1.96SE에 위치한다는 특성에 따라 계산하였다. t 분포가 긴 꼬리를 갖는다는 것은 모든 가능한 표본평균들의 95%를 포함하기 위해서는 분포의 평균에서 꼬리방향으로 약간 더 넓게 계산된다는 것을 의미한다. 예를 들어 표본크기가 25인 개에서 AST 농도를 측정하여 평균 40, 표준편차 4을 얻은 경우 95% 신뢰구간은 t 분포 표에 의하여 평균 ± 2.064SE 사이에 위치한다. 자유도가 24이므로 t 분포에서 95% 신뢰구간은 [38.35, 41.65]으로 계산되며 이 구간은 정규분포에서 계산된 [38.4, 41.6]구간에 비하여 더 넓다는 것을 알 수 있다.

가설검정

검정의 오류: 가설검정에서 귀무가설을 수용(accept, fail to reject)하거나 기각(reject)할 때 오류가 발생한다(Table 9). 참인 귀무가설을 수용하거나 거짓 귀무가설을 기각하는 행위는 올바른 판정이다. 반면에 참인 귀무가설을 기각하거나 거짓 귀무가설을 수용하는 것은 잘못된 판정으로 전자를 제1종 오류(α error), 후자를 제2종 오류(β error)라고 한다. 예를 들어 사료첨가제에 관한 예에서 제1형 오류는 실제로 두 집단 간 증체율에 차이가 없음에도 불구하고 차이가 있는 것으로 판정하는 경우다(2,4,10). 제2형 오류는 두 집단 간 증체율의 차이가 실제로 존재함에도 불구하고 유의한 차이를 검출하는데 실패하는 경우다.

제1형 오류는 분석과 관련이 있으며 제2형 오류는 실험계획의 문제 특히 표본크기가 적정 수준 이하일 때 발생한다. 제1종 오류와 제2종 오류는 서로 반대로 작용하기 때문에 제1종 오류를 줄이면 제2종 오류는 증가한다(Fig 8). 따라서 오류를 최소화하는 방법으로 하나의 오류를 고정시킨 상태에서 다른 오류를 최소화 하는 방법을 사용하며 흔히 5%의 제1종 오류(신뢰수준 95%)와 20%의 제2종 오류(검정력 80%)를 사용한다. 이는 대부분의 연구에서 제2종 오류에 비하여 제1종 오류가 더 중요하기 때문에 낮게 설정한다는 의미이다(5,8). 예를 들어 제1형 오류가 증가하면 새로운 약제가 효과가 있는 것으로 판정될 가능성이 높고 이러한 결과

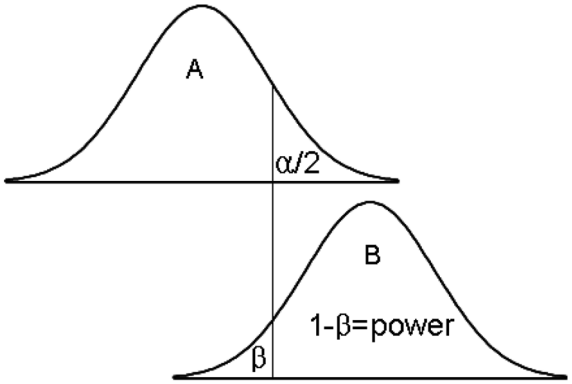

Fig 8. Relationship between α, β and 1 – β(power).

를 임상에 적용할 경우 잘못된 과학적 가설을 수용하게 되어 환자의 입장에서는 효과적인 치료를 받을 기회를 얻지 못하거나 약물의 부작용 등에 의한 문제가 초래된다. 반면에 제2형 오류가 증가하면 과학적 진실이 감추어지고 신약 개발에 투자한 비용이 낭비되는 문제는 있으나 환자에 미치는 영향은 없다.

유의수준과 유의확률: 유의수준(significance level)은 실제로 처리효과가 없음에도 불구하고 가설검정 결과 효과가 있다는 잘못된 결론을 얻는 오류의 크기를 어느 정도까지 허용할 것인지를 연구자가 설정하는 수준이다. 예를 들어 5%의 유의수준에서 검정한다는 것은 검정결과 귀무가설을 기각하고 대립가설을 수용할 때 이러한 결론이 거짓일 확률이 5% 정도까지는 허용하는 것으로 참인 귀무가설을 기각하는 제1종 오류의 최대 허용수준으로 정의된다. 연구자는 치료효과에 차이가 있다는 긍정적인 결과(positive result)를 제시하고 싶은 욕구가 있기 때문에 통계검정에서 그러한 결론을 내릴 때 범할 수 있는 α 오류의 수준을 연구 상황과 목적에 따라 적절히 결정할 필요가 있다.

귀무가설을 기각하는 유의수준의 최소값을 유의확률(significance probability)이라고 하며 흔히 p value로 제시한다. 이는 확률의 개념이므로 0에서 1까지의 값을 갖는다. $p = 0$이라는 것은 관찰된 차이가 우연에 기인하였을 가능성이 없으므로 두 군간 차이가 있으며, $p = 1$은 관찰된 차이가 순전히 우연에 의한 것이므로 두 군간 차이가 없음을 의미한다. 예를 들어 두 치료효과의 차이를 평균으로 검정한 결과 $p = 0.03$이라는 것은 치료효과에 차이가 없다는 귀무가설 하에서 연구결과로 얻은 치료효과의 차이가 우연히 나타날 확률이 3% 이하로 매우 낮기 때문에 이러한 차이는 우연에 기인한 것이 아니라 처리 즉 치료효과에 기인하였을 가능성이 높다는 것이다. 유의확률이 작을수록 귀무가설을 기각하는 증거의 강도가 높다는 것을 의미하며(2) 귀무가설이 참이라는 전제에서 표본통계량의 유의확률이 유의수준 α 보다 더 작기 때문에 귀무가설을 기각하는 것이다.

가설검정 결과를 제시하는 수단으로 유의확률 대신에 최근에는 신뢰구간을 사용하는 경우가 많다(7). 그 이유는 유의확률은 귀무가설을 반박하는 증거 혹은 두 군간 연관성의 강도를 측정하는 수단으로 처리효과의 크기에 대한 정보를 제공하지 못하지만 신뢰구간은 가설검정 결과뿐만 아니라 모수가 위치할 구간의 정밀도(precision) 혹은 정확도(accuracy)에 대한 정량적인 정보를 동시에 제공하기 때문이다.

유의한 결과가 반드시 임상적으로 중요하다는 것을 의미하는 것은 아니다. 예를 들어 새로운 치료제의 효과를 검정하는 연구에서 매우 작은 유의확률을 얻었을 때 치료제의 효과가 우연에 의한 결과일 가능성은 낮지만 임상적으로 활용할 것인지는 별개의 문제이다. 임상에 적용하기 위해서는 약제의 부작용, 비용 등 많은 요인을 고려해야 하기 때문이다. 또한 유의한 결과에 대해서는 유효크기(effect size)와 표본크기와의 관계를 검토해야 한다. 유효크기는 두 치료법 간 효과의 차이로 이 값이 클수록 유의확률은 작아져 귀무가설을 기각하기가 용이하며, 표본크기가 매우 커지면 치료효과의 차이가 미미하여도 유의한 결과를 얻을 수 있기 때문이다(10). 따라서 유의한 결과를 해석할 때 이러한 통계적 유의성(statistical significance)이 반드시 임상적 유의성(clinical significance)을 의미하는 것이 아니고, 유의확률이 크다는 것이 반드시 연관성이 없다는 것을 의미하는 것이 아니기 때문에 결과를 신중하게 기술해야 한다. 통계적으로 유의하지 않은 결과는 흔히 과소표본에 기인한 경우가 많기 때문에 검정력을 평가할 필요가 있다(4).

단측검정과 양측검정: 가설검정은 양측검정과 단측검정으로 구분할 수 있다. 유의수준 5%일 때 양측검정에서는 기각역이 표본분포의 위쪽과 아래쪽에 균등하게 2.5%씩 분포하는 반면 단측검정에서는 위 혹은 아래의 한 방향으로 5%가 분포한다. 양측검정이 흔히 사용되지만 연구 상황에 따라서는 단측검정을 사용하는 것이 더 적절한 경우가 있다. 예를 들어 기존 치료법과 새로운 치료법의 효과를 비교하는 연구에서 기존 치료법에 비하여 새로운 치료법의 효과가 더 크다는 것에 관심을 둘 수 있으며 이 경우 대립가설은 단측검정이 된다. 이와 같이 검정의 형태는 연구가설을 어떻게 설정하느냐에 따라 다르게 설정해야 한다(3). 이를테면 두 군의 치료효과가 차이가 있는지에 대한 차이성 시험에서는 흔히 양측검정을 사용하며, 새로운 치료제의 치료효과가 기존의 치료제의 효과에 비하여 나쁘지 않다는 것을 검정하는 비열등성 시험(non-inferiority testing)에서는 단측검정, 두 군에서 치료효과의 차이가 임상적으로 무시할 수 있는 동등성 인정한계 이내에 있는지를 검정하고 특히 새로운 치료제의 용량-반응관계를 제시하는 동등성 시험(equivalence testing)에서는 양측검정을 사용한다. 새로운 치료제가 기존의 치료제에 비하여 약제의 유효성이 더 우수하다는 것을 검정하는 우위성 검정(superiority testing)에서는 단측검정을 사용할 수 있다.

참 고 문 헌

1. Altman DG, Bland JM. Improving doctor's understanding of statistics. J R Stat Soc 1991; 154: 223-267.

2. Applegate KE, Crewson PE. An introduction to biostatistics. Radiology 2002; 225: 318-322.
3. Archibald CP, Lee HP. Sample size estimation for clinicians. Ann Acad Med Singapore 1995; 24: 328-332.
4. Cassidy LD. Basic concepts of statistical analysis for surgical research. J Surg Res 2005; 128: 199-206.
5. Evans RB. O'Connor A. Statistics and evidence-based veterinary medicine: Answers to 21 common statistical questions that arise from reading scientific manuscripts. Vet Clin North Am Small Anim Pract 2007; 37: 477-486.
6. Greenhalgh T. How to read a paper. Statistics for the non-statistician. I: Different types of data need different statistical tests. BMJ 1997; 315: 364-366.
7. Montiani-Ferreira F. Cardoso FF. Petersen-Jones S. Basic concepts in statistics for veterinary ophthalmologists. Vet Ophthalmol 2004; 7: 79-85.
8. Nick TG. Descriptive statistics. Methods Mol Biol 2007; 404: 33-52.
9. Nick TG, Williams JM, Barker JR. Descriptive and graphical strategies for assessing change: a case study of functional status in stroke patients. Top Health Inf Manage 1998; 18: 8-17.
10. Scales DC, Rubenfeld GD. Estimating sample size in critical care clinical trials. J Crit Care 2005; 20: 6-11.
11. Sonnad SS. Describing data: statistical and graphical methods. Radiology 2002; 225: 622-628.
12. Trevejo RT. A small animal clinician's guide to critical appraisal of the evidence in scientific literature. Vet Clin North Am Small Anim Pract 2007; 37: 463-475.
13. Whitley E, Ball J. Statistic review 2: samples and population. Crit Care 2002; 6: 143-148.