J Vet Clin 26(5) : 401-407 (2009)

Journal of Veterinary Clinics
www.ksvc.or.kr

평균과 비율 비교

박선일[1] · 이영원*

강원대학교 수의학부대학 및 동물의학종합연구소, *충남대학교 수의과대학

(게재승인: 2009년 10월 13일)

Hypothesis Testing: Means and Proportions

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

School of Veterianry 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 : In the previous article in this series we introduced the basic concepts for statistical analysis. The present review introduces hypothesis testing for continuous and categorical data for readers of the veterinary science literature. For the analysis of continuous data, we explained t-test to compare a single mean with a hypothesized value and the difference between two means from two independent samples or between two means arising from paired samples. When the data are categorical variables, the x^2 test for association and homogeneity, Fisher's exact test and Yates' continuity correction for small samples, and test for trend, in which at least one of the variables is ordinal is described, together with the worked examples. McNemar test for correlated proportions is also discussed. The topics covered may provide a basic understanding of different approaches for analyzing clinical data.

Key words **:** t-test, x^2 test, Fisher's test, trend test, McNemar test.

서 론

관찰 혹은 실험으로 얻은 자료에 대하여 적절한 분석기법을 적용하지 않을 경우 사실과 다른 결과를 얻을 뿐만 아니라 결과를 잘못 해석함으로써 연구 결과의 정확성을 저해하는 심각한 문제가 발생할 수 있다. 자료의 특성에 부합하는 적절한 통계기법을 찾기 위해서는 변수의 척도, 비교집단의 개수, 표본의 독립성 여부 등 다양한 요인을 고려해야 한다. 명목 척도인 경우에는 빈도를 이용한 χ^2 검정이나 비율에 대한 z-검정, 등간척도 이상인 자료는 t 혹은 z 검정을 사용한다. 두 변수의 연관성은 상관분석, 두 개 이상의 평균을 동시에 비교하는 분산분석(analysis of variance), 교란변수(confounding variable)의 영향을 보정한 상태에서 치료효과의 차이를 검정하는 회귀분석(multiple regression analysis) 혹은 로짓분석(multiple logistic analysis) 등 다양한 기법을 사용할 수 있다(12). 한편 모수분석을 위한 가정을 충족하지 못할 경우 비모수 검정법을 사용한다. 이를테면 순위 척도 이상으로 관측된 두 독립표본의 평균비교를 위한 Wilcoxon rank-sum test (Mann-Whitney U test), 짝을 이룬 두 표본의 평균비교를 위한 Wilcoxon signed-rank test, 상관관계를 분석을 위한 Spearman's rho, 세 집단 이상의 표본평균을 비교하기 위해서는 Kruskal-Wallis 검정 등은 이러한 예다(2,3,13). 본 원고에서는 임상의학에서 흔히 접하는 평균(mean)과 비율(proportion)을 비교하는 기법을 설명하며, 두 집단 이상의 평균비교를 위한 분산분석은 추후에 다루기로 한다.

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

결 론

1. 평균 비교

1.1 단일 집단

예시: 임상의학에서 단일 표본의 평균을 가설상의 평균과 비교하는 예는 흔하지는 않지만 예를 들어 만성신부전으로 입원한 개를 대상으로 hemoglobin 농도를 측정한 결과가 Table 1과 같다고 하자. 모집단에서 헤모글로빈 농도의 평균

Table 1. Serum hemoglobin concentration (g/dl) in dogs with chronic renal failure

Data	12.3 13.1 12.8 10.5 11.2 10.2 10.8 10.9 11.8 10.4 13.2 10.6
	14.6 12.0 13.1 11.2 10.5 12.6 10.1 12.4
μ_0	16

$\bar{x} \pm SD = 11.72 \pm 1.26$, $SE = 0.28$, 95% CI = [11.1, 12.3], t-statistic = −15.3 ($p < 0.0001$)

이 16 g/dl 이라고 할 때 만성신부전 환자에서 측정된 헤모글로빈 농도가 모집단 평균과 차이가 있는지를 평가하고자 한다.

Table 1에서 보듯이 모집단 평균(μ_0 = 16)에 비하여 표본평균($\bar{x}$ = 11.72)이 상대적으로 낮은 값으로 나타났지만 모집단에서 무작위로 추출된 어느 한 표본에서 헤모글로빈 농도의 평균으로 정확히 16 g/dl을 얻을 가능성은 거의 없다. 따라서 이러한 차이가 표본의 무작위 변동에 기인한 우연에 의한 것인지 아니면 모집단과 표본평균이 실제로 차이가 있는지를 검정할 필요가 있다. 어느 설명이 맞는지를 확인하는 가장 좋은 방법은 신뢰구간을 계산하여 가설검정을 시도하는 것이다. 이 자료에 대하여 표준편차를 계산하면 1.26이고, 표본크기는 20두이므로 표준오차는 0.28이다. 유의수준 5%에서 자유도 19일 때 t 분포의 계수는 2.093이므로 신뢰구간은[11.1, 12.3]으로 계산된다. 이 구간은 20두의 표본이 모집단을 대표한다고 가정할 때 만성신부전으로 진단받은 개에서 헤모글로빈 농도 평균의 참값이 11.1에서 12.3 사이에 포함된다는 것을 95% 신뢰할 수 있는 구간으로 해석할 수 있다. 계산 결과 신뢰구간이 가설상의 모집단 평균 16을 포함하지 않으므로 표본평균은 모집단 평균과 다르다는 결론을 얻는다.

다음 단계에서는 관찰된 차이가 얼마나 우연히 나타날 수 있는지 확률을 추정하기 위하여 가설검정을 수행한다. 이 때 귀무가설(null hypothesis)은 만성신부전으로 진단받은 개의 헤모글로빈 농도의 평균은 모집단에서의 평균 농도와 동일하다는 것이며 두 평균의 차이가 클수록 우연에 의해 나타날 가능성은 희박하다는 것을 의미한다. 단일표본 평균을 검정하는데 사용되는 검정통계량은 두 집단의 평균 차이를 표본평균의 표준오차로 나누어 계산한다.

$$t = \frac{11.72 - 16}{0.28} = -15.3$$

계산된 검정통계량 -15.3은 표본에서 관찰된 헤모글로빈 농도의 평균이 가설상의 평균 보다 15.3SE 이하임을 의미한다. 이 값을 t 분포의 표에서 임계값(critical region)과 비교하여 이러한 차이가 우연에 의해 관찰될 유의확률로 $p < 0.0001$을 얻는다. 이 값은 표본에서 관찰된 헤모글로빈 농도의 평균과 모집단의 농도 평균 간의 차이가 우연에 의해 나타날 확률이 극히 낮기 때문에 두 집단 간 헤모글로빈 농도 평균에 차이가 있으며 전술한 방법과 동일한 결론을 얻는다. 유의확률은 평균 차이의 크기(size of difference)에 대한 정보를 제시해주는 것이 아니라 이러한 차이가 우연에 기인할 확률이다(8).

1.2 두 독립표본

가정검토: Student's t 분포는 두 독립표본 평균을 비교할 때 사용하는 기법이다. t 검정은 모수분석이므로 측정된 변수가 적어도 순위형 이상인 연속형 척도이어야 하며, 두 집단이 각각 정규분포에 근사하고 두 집단의 분산이 동일하다는 가정(등분산성)에 부합될 때 사용할 수 있다(2,11,13). 종속변수가 정규분포를 따르지 않으면 자료의 변형 (transformation)을 통하여 정규성을 검토하고 여전히 정규성을 충족하지 못하면 Mann-Whiney U 검정이나 범주화된 자료에 대하여 x^2 검정을 사용한다. 등분산성을 만족하지 못하는 경우에도 자료변환을 시도할 수 있다. 등분산성을 만족하지 못하는 경우에는 각 집단의 분산을 표본크기로 보정하여 t 통계량을 계산하는데 이를 Satterthwaite 검정 혹은 Welch 검정이라고 한다. 이 검정은 두 집단 간 표본크기가 다른 경우에도 사용할 수 있다. 두 표본이 독립적이지 않고 짝지어진 표본(paired, correlated, dependent)인 경우에는 paired t 검정을 사용한다. 예를 들어 동일한 개체로부터 특정한 약물을 처리하기 전과 후에 혈청효소 농도를 측정하였다면 처리 전후에 측정한 두 자료는 독립성 가정을 만족하지 못하고, 이 경우 두 집단 평균의 차이(difference)가 정규분포에 근사하다는 가정을 만족해야 한다. t 검정은 이러한 가정에서 어느 정도 벗어나도 큰 문제가 없기 때문에(이를 robust라고 함) 전술한 가정을 검토하지 않고 분석하기도 한다.

예시: 급성(n = 45) 혹은 만성(n = 55) 신부전으로 진단받은 100두의 개를 대상으로 혈청 칼슘농도를 측정한 결과가 Table 2와 같을 때 두 군간 칼슘농도에 차이가 있는지를 평가하고자 한다. 두 군간 칼슘농도의 차이는 1.6 mg/dl로 급성신부전에서 더 높은 것으로 나타났다. 혈청 칼슘농도 평균의 신뢰구간을 계산하면 표본크기가 크기 때문에 정규분포에 근사한다고 가정하면 급성신부전은[11.3, 12.3], 만성신부전의 경우[9.8, 10.6]으로 계산된다. 모집단의 참값이 포함될 두 신뢰구간이 서로 겹치지 않기 때문에 두 군간 혈청 칼슘농도가 다르다는 결론을 얻는다.

한편 관찰된 차이의 크기에 대한 정보는 유용하기 때문에 계산할 필요가 있다. 두 독립표본의 평균 차이에 대한 신뢰구간은 차이의 표준편차(SD for the mean difference, SD_{diff})와 합동표준오차(SE_{diff})로 계산된다(4).

$$SD_{diff} = \sqrt{\frac{(n_1 - 1) \times SD_1^2 + (n_2 - 1) \times SD_2^2}{n_1 + n_2 - 2}} = 1.54$$

$$SE_{diff} = SD_{diff} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.31$$

두 군간 혈청 칼슘농도의 평균 차이는 1.6 mg/dl이고 평균 차이의 신뢰구간은[0.99, 2.21]로 계산된다. 만일 두 군의 평균 칼슘농도에 차이가 없다면 평균 차이는 0에 근사할 것이므로 본 예에서 계산된 평균차이의 신뢰구간이 0을 포함하

Table 2. Serum calcium concentration (mg/dl) in dogs with renal failure (RF)

	Acute RF (n = 45)	Chronic RF (n = 55)
Mean ± SD	11.8 ± 1.7	10.2 ± 1.4
SE	0.25	0.19
95% CI	11.3, 12.3	9.8, 10.6

SD_{diff} = 1.54, SE_{diff} = 0.31, 95% CI of difference = [0.99, 2.21]
t-statistic = 5.16 ($p < 0.0001$)

지 않기 때문에 두 군간 칼슘 농도에 차이가 있다는 결론을 얻는다. 한편 가설검정에서 검정통계량은 5.16이고, 이 값을 t 분포 표의 값(자유도 = 99)과 비교하면 유의확률로 $p < 0.0001$을 얻는다. 이러한 결과는 표본에서 관찰된 평균 칼슘 농도와 모집단의 평균 농도 간의 차이가 우연에 의해 나타날 확률이 극히 낮기 때문에 두 집단 간 평균 칼슘 농도에 차이가 있다고 해석한다.

$$t = \frac{\overline{x_1} - \overline{x_2}}{SE_{diff}} = \frac{11.8 - 10.2}{0.31} = 5.16$$

1.3 두 종속표본

예시: 의학연구에서 짝지은 자료는 흔히 짝지은 사례-대조군 연구(matched case-control study), 반복측정(repeated measures study) 등의 상황에서 얻는다. 예를 들어 특정 약제를 투여하기 전과 후에 혈당(blood glucose) 농도를 측정한 결과가 Table 3과 같다고 하자. 투여 전과 후의 평균 혈당 농도가 각각 88, 94 mg/dl이라고 할 때 투여 전에 비하여 투여 후 혈당 농도는 약 6.8% 증가하였다. 이러한 차이가 우연에 의한 결과인지 약제의 효과에 기인한 것인지를 평가하고자 한다.

이 자료는 동일한 개체를 처리 전후로 반복하여 측정하였기 때문에 t 검정의 독립성 가정을 위반하므로 분석에서 의존성(dependency)을 고려하기 위하여 관찰치 쌍 간의 차이를 분석 한다. 이 때 귀무가설은 혈당 농도의 차이는 0이며, 두 혈당 농도차이의 평균을 가설상의 평균 0과 비교하므로 두 독립표본에 대한 t 검정의 특별한 경우로 볼 수 있다.

$$t = \frac{\bar{x}_{diff} - 0}{SE_{diff}} = \frac{\bar{x}_{diff}}{SE_{diff}} = 6/1.51 = 3.97$$

계산된 검정통계량 3.97은 유의수준 5%, 자유도 14에서 t 분포 표에 의해 $0.01 < p < 0.001$로 추정된다. 따라서 약제 처리 전과 후의 평균 혈당 농도는 우연에 의해 관찰된 가능성이 매우 낮기 때문에 약제의 혈당 증가는 약제의 효과에 기인한 것으로 해석할 수 있다. 한편 신뢰구간은[2.76, 9.24]로 계산되며 이 구간이 0을 포함하지 않기 때문에 처리 전후의 평균 혈당 농도는 다르다는 결론을 얻는다.

$6 \pm 2.145^{*}1.51 = [2.76,\ 9.24]$

1.4 두 집단 평균 비교를 위한 비모수분석

모수분석의 가정을 만족하지 못할 경우에는 모집단 분포의 정규성(normality)과 분산의 동질성(equal variance) 가정을 요구하지 않는 비모수 분석(nonparametric test)을 사용한다. 대부분의 비모수 분석은 원자료를 순위형(rank)이나 범주형(category) 형으로 전환하여 분석하기 때문에 정보의 손실과 분석의 정확성이 다소 저하된다. 모수기법에 비하여 검정력이 낮기 때문에 검정결과에 대하여 제2종 오류 즉 두 집단 간 평균 차이가 실제로 있지만 차이가 없는 것으로 잘못 판정할 오류의 가능성을 검토할 필요가 있다(8). 순위형으로 측정된 두 독립표본의 평균 차이를 검정하기 위하여 Mann-Whitney U 검정(Wilcoxon rank-sum 검정과 동일함)을 사용하며, 종속자료에 대해서는 Wilcoxon signed-rank 검정을 사용한다(2,3). 자세한 내용은 비모수분석에 관한 별도의 원고에서 다룬다.

Table 3. Glucose level (mg/dl) measured before and after treatment

	Before	After
Data	78 84 88 78 68 102 98 88 82 100 92 78 110 100 98 92 80 70 90 106 102 86	104 106 88 78 92 100 104 88
Mean ± SD	88 ± 11.78	94 ± 10.28

$\bar{x}_{diff} \pm SD = 6 \pm 5.86$, $SE_{diff} = 1.51$, 95% CI = [2.76, 9.24], t-statistic = 3.97 ($0.01 < p < 0.001$)

Table 4. Observed frequency by antimicrobial agents (n=400)

	Treatment result			
Agents	Smear +	Smear − Culture +	Smear − Culture −	Total
A	40	30	130	200
B	10	20	70	100
C	15	40	45	100
Total	65	90	245	400

2. 비율 비교

연구 상황: 의학연구에서 범주형(categorical) 자료는 이를테면 독성시험에서 폐사 여부, 특정 바이러스 감염 여부, PCR 검사결과, 연령별 항체 감염 두수, 수정란 발육 실험에서 배반포까지 발육한 난자수, 질병에 감염된 동물에 약물의 농도를 달리하여 투여할 때 회복여부, 빈혈 환자에서 PCV 수준별 생존여부 등은 전형적인 예다. 이러한 자료는 특정 범주에 대한 발생 빈도(frequency)나 도수(count)로 표현되며 두 변수 간의 연관성(association)을 분석할 목적으로 분할표(contingency table)로 요약한 후 χ^2 검정을 사용한다(1,15).

예시: 특정 세균 감염을 치료하기 위하여 사용되는 3종 항생제의 치료결과를 비교하는 실험에서 400두에 대한 치료결과를 3단계로 분류한 결과가 Table 4와 같다고 하자. 이 연구에서 연구자는 첫째, 항생제 종류별 도말 양성비율이 동일하다고 할 수 있는지 둘째, 항생제 종류와 치료결과 간 연관성(association)이 존재하는지에 관심을 갖는다. 전자의 가설에 대한 검정을 동질성 검정(homogeneity test), 후자의 가설에 대한 검정을 독립성 검정(independent test)이라고 하며, 두 검정은 귀무가설과 검정절차만 다를 뿐 실제로는 동일한 방법을 사용한다.

2.1 카이제곱 검정

원리: 두 변수의 연관성을 비교하기 위해서는 어느 한 변수의 범주별 비율이 다른 변수의 범주와 무관하게 동일할 때 기대되는 각 셀에서의 기대빈도(expected frequency)를 계산

Table 5. Expected frequency by antimicrobial agents (n=400)

	Treatment result			
Agents	Smear +	Smear − Culture +	Smear − Culture −	Total
A	32.5	45.0	122.5	200
B	16.25	22.5	61.25	100
C	16.25	22.5	61.25	100
Total	65	90	245	400

한다. 즉 기대빈도는 두 변수는 연관성이 없다는 귀무가설이 참일 때 기대되는 빈도를 의미한다. TAble 4에서 항생제 종류별 비율은 A=200/400=50%, B=100/400=25%, C=100/400=25%이다. 귀무가설이 참일 때 기대되는 환자 수는 항생제 A=65*(200/400), B=90*(100/400), C=245*(100/400)로 계산된다. 이와 동일한 방법으로 나머지 범주에 대하여 기대빈도를 계산하면 Table 5와 같이 정리된다.

항생제 A와 도말양성 셀에 대한 기대빈도는 $A = 65 \times (200/400)$ 혹은 $A = 200 \times 65/400$로 계산되므로 행 및 열의 합과 총 표본크기에 조건부로 기대빈도가 계산된다는 것을 알 수 있다. 연관성에 대한 검정은 관찰빈도와 기대빈도의 차이를 계산하게 되며, 차이가 크다는 것은 두 변수 간의 연관성이 존재함을 의미한다. c개의 행(column, c)과 r개의 열(row, r)을 갖는 분할표에서 i번째 행과 j번째 열의 관찰빈도와 기대빈도를 각각 O_{ij}, E_{ij}라 하면 검정통계량은 다음과 같다.

$$x^2 = \Sigma_{i=1}^{r}\Sigma_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

두 변수 간 연관성이 없다는 귀무가설에서 계산된 검정통계량은 자유도가 $(r-1)$ $(c-1)$인 카이제곱 분포에 근사한다. 이 자료에 대하여 패키지를 이용하여 계산하면 검정통계량은 29.14로 계산되고 이 값은 자유도 4일 때 임계값 18.47보다 크다($p < 0.001$). 따라서 두 변수 간 연관성이 없을 때 관찰된 빈도를 가질 확률은 0.001로 매우 낮으므로 항생제 종류와 치료반응은 연관성이 있다는 결론을 얻는다.

연속성 보정: 연관성에 대한 카이제곱 검정은 이산 확률(discrete probability)에 연속확률 분포를 근사시키는 방법으로 특히 소표본(small sample)에서는 카이제곱분포의 정규분포 근사성을 적용하면 카이제곱 통계량이 왜곡되어 나타날 수 있다. 따라서 2×2 분할표에서 소표본 자료에 대하여 보다 정확한 검정을 위한 방법으로 Yates가 제시한 연속성 보정(continuity correction)법을 사용한다(Yates 1934). 이 방법은 관찰빈도와 기대빈도의 차이를 0.5 만큼 줄임으로써 보수적인 통계량을 제공한다.

$$x^2 = \Sigma_{i=1}^{r}\Sigma_{j=1}^{c}\frac{(|O_{ij}-E_{ij}|-0.5)^2}{E_{ij}}$$

표본크기가 크면 x^2, Yates 보정법, Fisher의 정확확률 검정이 모두 동일한 결과를 보이지만 소표본에서는 Fisher 검정과 Yates 보정법이 보수적인 결과를 제공한다. 보수적인 결과람 함은 유의확률이 증가함으로써 두 변수 간 연관성이 있다는 결론을 얻을 가능성이 감소함을 의미한다.

Table 6. IMHA frequency by blood types in dogs (n=43)

Blood type	IMHA+	IMHA−	Total
DEA 7	2	3	5
Others	29	9	38
Total	31	12	43

* DEA, dog erythrocyte antigen, IMHA, immune-mediated hemolytic anemia

2.2 피셔의 정확 검정(Fisher's exact test)

연구 상황: 카이제곱 검정은 이항분포에 대한 정규분포 근사성을 이용한 분석기법이기 때문에 이러한 가정이 만족될 때 가장 적절하게 사용할 수 있다. 즉 행과 열의 개수가 두 개인 가장 단순한 형태의 2×2 분할표에 대한 카이제곱 분석에서는 표본크기가 20 보다 커야 하고 모든 셀의 기대빈도가 5 이상이어야 하는 가정을 필요로 한다(2,3). 두 개 이상의 수준을 갖는 자료에서 빈도를 비교할 경우에는 기대빈도가 5 이하인 셀이 전체 셀의 20%를 넘지 않아야 하며, 기대빈도가 1 이하인 셀이 없어야 한다. 만일 분할표의 기대빈도가 작으면 전술한 검정통계량 공식에서 분자의 값이 팽창되어(inflated) 결과적으로 통계량이 과대추정되는 문제가 초래되므로 이 경우 비모수검정의 일종인 초기하분포에 근거한 Fisher의 정확검정(Fisher's exact test)을 사용한다(6,7).

예시: 개에서 혈액형과 면역매개성용혈성빈혈(IMHA) 간의 관계에 대한 연구에서 Table 6과 같은 결과를 얻었다고 가정하자. 피셔의 정확 검정에서는 유의확률은 행과 열의 합이 고정되어 있을 때 관찰된 분할표를 얻을 확률과 이 보다 더 극단적인 분할표 (귀무가설이 참이 아니라는 증거)를 얻을 확률을 모두 합하여 계산한다. 행과 열이 고정되어 있을 때 관찰된 분할표 보다 더 극단적인 분할표는 Table 7과 같이 세가지 상황이 가능하다(9).

Table 7에서 Table (i), (ii) 및 (iii)을 관찰할 초기하확률은 다음과 같다.

$$\text{Table (i): } \frac{\binom{31}{2}\binom{12}{3}}{\binom{43}{5}} = 0.1063, \quad \text{Table (ii): } \frac{\binom{31}{1}\binom{12}{4}}{\binom{43}{5}} = 0.0159,$$

$$\text{Table (iii): } \frac{\binom{31}{0}\binom{12}{5}}{\binom{43}{5}} = 0.0008$$

단측검정에서 정확 유의확률은 $0.1063 + 0.0159 + 0.0008 = 0.1230$으로 계산되며 이 값은 5% 유의수준에서 귀무가설을 기각하지 못한다. 양측검정의 유의확률을 얻기 위하여 단측검정의 유의확률을 단순히 2배할 수도 있지만 Fisher의 정확

Table 7. Extreme contingency tables for calculation of Fisher's exact test

	Table (i)		Table (ii)		Table (iii)		
Blood type	IMHA+	IMHA–	IMHA+	IMHA–	IMHA+	IMHA–	Total
DEA 7	2	3	1	4	0	5	5
Others	29	9	30	8	31	7	38
Total	31	12	31	12	31	12	43

확률은 반드시 대칭을 보이는 것이 아니기 때문에 이 방법은 지나치게 보수적인 결과를 얻는다. 따라서 보다 정확한 값을 얻기 위해서는 관찰된 분할표의 확률과 같거나 더 작은 모든 확률을 더하여 계산하는 것이 바람직하다.

2.3 Cochran-Mantel-Haenszel (CMH) test

원리: CMH 검정은 층화표본에 근거한 독립표본 간 비율을 검정할 때 사용한다. 앞에서 설명한 카이제곱 검정과 피셔의 정확검정은 하나의 그룹 내에서 비교하였지만 CMH 검정은 여러 개의 그룹 즉 층화요인(stratification factor)이 있을 때 이러한 요인을 보정한 상태에서 처리 간 비율에 차이가 있는지를 검정할 때 유용하게 사용할 수 있다(2,3). 층화요인은 이원분산분석에서 블록요인(blocking factor)과 유사한 역할을 한다. 이를테면 두 약물의 치료율을 비교하는 연구가 여러 병원에서 평가한 것이라고 하면 이 때 병원은 하나의 층(stratum)이 된다. 이 경우 병원에 따라 치료율이 다를 수 있기 때문에 두 약물의 효과를 보다 정확하게 비교하기 위해서는 병원의 효과를 보정한 상태에서 치료율을 비교하는 것이 타당하다. 이와 같이 층화요인은 병원, 성별, 질병의 중증도, 지역 등 처리(약물)와 반응률(치료효과)에 영향을 미치는 어떠한 요인도 가능하다. 층의 개수 k ($k \geq 2$), j번째 층의 환자수 N_j, 그룹 1에서 총 환자 수 X_{j1}, 반응을 보인 환자 수 n_{j1}, 그룹 2에서 총 환자 수 X_{j2}, 반응을 보인 환자 수 n_{j2}라고 하면 CMH 검정통계량은 다음과 같이 계산되며 이 통계량은 귀무가설하에서 자유도 1을 갖는 카이제곱 분포에 근사한다.

$$\chi^2_{CMH} = \frac{[\Sigma_{j=1}^{k} NUM_j]^2}{\Sigma_{j=1}^{k} DEN_j}$$

$$NUM_j = \frac{X_{j1} \times n_{j2} - X_{j2} \times n_{j1}}{N_j}$$

$$DEN_j = \frac{n_{j1} \times n_{j2} \times (X_{j1} + X_{j2}) \times (N_j - X_{j1} - X_{j2})}{N_j^2 \times (N_j - 1)}$$

예시: 피부질환을 치료하는데 사용되는 기존의 약물(control)과 새로 개발된 약물(new)간의 치료율을 비교하는 연구를 4곳의 병원에서 평가한 결과(Table 8)를 분할표로 정리하여 검정통계량을 계산하면 Table 9와 같다.

층화요인인 병원의 효과를 보정한 상태에서 계산된 검정통계량 1.233은 유의수준 5%에서 임계값 3.841 보다 작으므로 두 약물 간 피부병 치료율에 차이가 없다는 결론을 얻는다(p = 0.2668). 층화요인을 보정하지 않은 상태에서 분석하면 카이제곱 검정통계량은 1.242이고 유의수준 5%에서 CMH 검정결과와 동일한 결과를 보인다(p = 0.2650). 그러나 이러한 결과는 우연히 일치한 것이므로 층화요인이 사용된 자료에 대해서는 반드시 이를 보정한 후 분석하는 것이 중요하다. CMH 검정은 역학연구에서 후향적 자료(retrospective data)를 분석하기 위하여 개발된 것으로 이를 변형한 다양한 검정통계량이 개발되어 있다. 예를 들면 코크란 카이제곱 검정(Cochran's χ^2 test)은 위의 공식에서 분모인 $N_j^2(N_j - 1)$ 를 N_j^3 으로 대치한 것으로 $k = 1$인 경우 카이제곱 검정통계량과 동일해진다. 분할표에서 기대빈도가 작은 셀이 많을 경우 연속성을 보정한 CMH 검정통계량(χ^2_{CMHadj}) 은 다음과 같다. 이 방법은 매우 보수적인 결과를 얻기 때문에 일반적으로 생

Table 8. Response frequency by study center

Study center	Group	Response	Non-response	Total (%)
1	New	22	10	32 (68.8)
	Control	10	7	17 (58.8)
	Sub-total	32	17	49
2	New	10	4	14 (71.4)
	Control	12	7	19 (63.2)
	Sub-total	22	11	33
3	New	14	7	21 (66.7)
	Control	9	6	15 (60.0)
	Sub-total	23	13	36
4	New	20	8	28 (71.4)
	Control	8	5	13 (61.5)
	Sub-total	28	13	41
Total	New	66	29	95 (69.5)
	Control	39	25	64 (60.9)

Table 9. Summary for CMH test statistic of Table 8

Study center	NUM_j	DEN_j
1	1.102	2.568
2	0.667	1.847
3	0.583	2.076
4	0.878	1.970
Total	3.230	8.461

$\chi^2_{CMH} = (3.230)^2/8.461 = 1.233$, Critical region = 3.841 ($\alpha = 0.05$, $df = 1$), $\chi^2_{CMH} = 1.233$

략하는 경우가 많다.

$$\chi^2_{CMHadj} = \frac{[|\Sigma_{j=1}^{k} NUM_j| - 0.5]^2}{\Sigma_{j=1}^{k} DEN_j}$$

동질성 검정: 처리(약물) 간 치료율의 차이가 층간에 일정하지 않다면 이는 상호작용(interaction)이 존재함을 의미한다. 예를 들어 기존약과 신약의 치료율이 층 1에서는 75%와 25%이고 다른 층에서는 25%와 75%인 경우 층을 무시하고 자료를 통합하면 상호작용 효과가 가리워지기 때문에(masking) 결과적으로 치료율이 상쇄되어 전체적으로 치료효과에 차이가 없다는 잘못된 결론을 얻을 수 있다. 상호작용이 존재한다는 것은 층간의 반응율이 동질하지 않다는 것을 의미하기 때문에(lack of homogeneity) 이 경우 각각의 층에 대하여 별도로 분석해야 한다. 동질성 검정은 흔히 Breslow-Day 검정을 사용하며 이 검정은 두 처리 간 치료율의 차이가 아니라 교차비(odds ratio)에 근거한 검정이다(2,3). 위의 자료에 대하여 분석하면 동질성에 대한 검정통계량은0.0324로 층간 치료율이 동질하다는 결론을 얻는다(p = 0.9985).

2.4 추세검정(trend test)

원리: 범주의 수준이 2개 이상인 분할표에서 행(response 변수)이나 열변수(group 변수)중 어느 하나가 순위형인 경우 다른 한 변수의 수준에 따른 반응 비율에 추세(trend) 패턴이 존재하는지를 검정할 수 있다(2,3).

예시: Table 10은 약물의 농도를 4개 수준으로 달리하여 실험동물의 폐사여부를 분류한 자료다. 이러한 연구에서 연구자는 약물의 농도와 폐사여부 간의 연관성을 관심을 갖기도 하지만 약물의 농도가 증가할수록 폐사율이 증가하는지 이를테면 양-반응관계에 관심을 둘 수도 있다. 추세검정은 직선의 기울기를 검정하기 위한 회귀분석(회귀분석 참고)을 사용하는 것과 유사하다. 자료 분석에서 반응변수인 폐사를 y라 하면 이 값은 1(생존) 아니면 2(폐사), group 변수인 약물 농도는 x = 1, 2, 3, 4로 지정한다. Cochran-Armitage 추세검정을 위한 검정통계량은 표본크기가 클 때 z 분포를 따르며 다음과 같이 계산한다.

$$\chi^2_{trend} = \frac{n[\Sigma_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})\,]^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}$$

추세검정에서 개별 셀의 크기는 중요하지 않지만 총 표본크기가 적어도 30이상일 때 적절하게 사용할 수 있다. Table 10의 자료에 대하여 추세검정을 수행하면 p = 0.0394(z = 2.063)으로 약물의 농도가 증가함에 따라 폐사율은 매우 유의하게 증가한다는 결론을 얻는다. 한편 이 자료에 대하여 연관성에 대한 카이제곱검정을 수행하면 p = 0.1909(χ^2 = 4.7516, df = 3)로 두 변수 간 연관성은 없는 것으로 해석할 수 있다.

한편 group 변수와 response 변수가 모두 순위형이고 어느 한 변수의 수준이 2개인 경우 Cochran-Armitage 추세검정 통계량은 Mantel-Haenszel 통계량에($N-1$)/N을 곱하면 (N = 총 표본크기) 동일한 값을 보인다. Table 10의 자료에서 $2.063^2 \times (152-1)/152 = 4.2171$로 SAS 출력에서 non-zero correlation의 검정통계량과 동일한 값을 보임을 알 수 있다. 따라서 group 변수가 순위형인 경우 non-zero correlation 통계량은 추세검정을 의미한다고 볼 수 있다.

앞에서는 group 변수가 순위형인 경우를 설명하였으며 이번에는 반응변수가 순위형인 경우를 가정하자(Table 11). 이 자료에서는 두 변수 간의 연관성 검정이외에 처리군에 따른 결과(반응)에 유의한 추세관계가 존재하는지를 검정할 수 있다. 이 자료에 대하여 연관성에 대한 카이제곱검정을 수행하면 p = 0.0029 (χ^2 = 11.71, df = 2)로 두 변수 간 유의한 연관성이 있고, 추세검정에서도 p = 0.0009 (χ^2 = 3.321, df = 1)로 매우 유의한 추세가 있는 것으로 분석된다.

2.5 McNemar 검정

원리: 예를 들어 동일한 시료에 두 종류의 진단검사를 적용하여 감염여부를 양성과 음성으로 판단하는 실험, 안구질환을 치료하는데 사용되는 2종류 약물의 효과를 비교하기 위하여 좌측 안구에 약물 A, 우측안구에는 약물 B를 투여하는 실험, 약물의 부작용을 평가하기 위하여 처치 전후의 ALP 농도를 측정하는 실험, 환자에서 채혈한 시료에 대하여 2명의 임상병리기사가 혈구계산판(hemocytometer)를 이용하여 백혈구수를 측정하여 정상과 비정상으로 판정하는 실험 등은 짝지은 자료의 예다. 이러한 실험에서는 두 표본 간 독립성이 유지되지 못하기 때문에 비율로 요약된 짝지은 자료에 대해서는 McNemar 검정을 사용한다(2,3,5,13).

예시: 86명의 환자를 대상으로 약물 처치 전후의 ALP 농도를 측정한 실험에서 ALP의 참고구간(reference range)에 근거하여 각 환자를 정상과 비정상으로 구분한 결과를 요약한 결과가 Table 12와 같다고 하자. 이 자료에서 66두는 처치 전후 모두 정상 혹은 비정상으로 판정되어 일치한 쌍

Table 10. Response summary by drug level

Level (mg)	Died	Survived	Total
5	5	35	40
10	6	29	35
15	10	28	38
20	12	27	39
Total	33	119	152

Table 11. Response by treatment groups (n=195)

Treatment group	Response category			Total
	Non-response	Incomplete recovery	Complete recovery	
Drug	16 (22.5%)	26 (36.6%)	68 (40.9%)	110
Placebo	24 (35.3%)	29 (38.2%)	32 (26.5%)	85
Total	40	55	100	195

(concordant pairs)이고 나머지 20쌍은 처치 전후의 결과가 다른 비일치쌍(discordant pairs)이다.

이 자료에서 연구자가 관심을 갖게 되는 비율에 대하여 처치 전 ALP 농도가 비정상인 비율을 P_1, 처치 후 ALP 농도가 비정상인 비율을 P_2라고 하면 귀무가설은 $P_1 = P_2$가 된다. 검정통계량은 총 비일치쌍 중에서 각 범주에 대한 비일치쌍 수(b, c)의 차이가 되며 이는 자유도가 1인 카이제곱분포에 근사한다.

$$\chi^2_M = \frac{(b-c)^2}{b+c}$$

Table 12의 자료에 대하여 계산하면 $\chi^2_M = 3.2\ [(14-6)^2/(14+6)]$이고 이 값은 임계값 3.84 보다 작으므로 처치 전후에 측정한 ALP 농도의 비정상 비율은 차이가 없다는 결론을 얻는다. 한편 연관성에 대한 카이제곱검정에서 설명한 방법과 마찬가지로 McNemar 검정통계량에 Yates' continuity correction을 적용할 수 있으며 자유도는 1이다. ALP 농도 자료에서 Yates 보정 통계량은 $\chi^2_M = 2.45\ [(|14-6|-1)^2/(14+6)]$로 동일한 결론을 얻는다. McNemar 검정과 신뢰구간 계산에 필요한 표본크기는 비일치쌍의 수가 적어도 10이상이어야 한다(10).

$$\chi^2_M = \frac{(|b-c|-1)^2}{b+c}$$

신뢰구간: 두 짝지어진 비율 자료에 대한 신뢰구간은 비일치쌍에 대한 두 비율의 차이(p_{diff})와 차이에 대한 표준오차(SE_{diff})로 다음과 같이 계산된다.

$$p_{diff} = \frac{b-c}{n}$$

$$SE_{diff} = \frac{\sqrt{(b+c)}}{n}$$

$$p_{diff} \pm z_{1-\alpha/2} \times SE_{diff}$$

Table 12의 자료에 대한 신뢰구간은[-0.009, 0.195]이고 이 구간이 0을 포함하므로 앞에서와 마찬가지로 두 비율 간 차이가 없다는 결론을 얻는다.

$$p_{diff} = \frac{b-c}{n} = \frac{14-6}{86} = 0.093$$

$$SE_{diff} = \frac{\sqrt{(b+c)}}{n} = \frac{\sqrt{14+6}}{86} = 0.052$$

$$0.093 \pm 1.96 \times 0.052 \Leftrightarrow [-0.009, 0.195]$$

요약하면 자료의 척도가 명목형이면서 두 집단의 관찰빈도에 대한 유의성을 검정할 목적으로 카이제곱 검정을 사용하며 두 표본이 독립적일 때 관찰빈도를 기대빈도와 비교하는 방법이다. 카이제곱 검정을 위한 기대빈도에 대한 가정이 충족되지 못할 경우 가능하다면 범주의 수준을 통합하여 수준의 개수를 줄이거나 Fisher의 정확검정(Fisher's exact test)이나 Yates 보정법을 사용하는 것이 바람직하다. 독립표본이 아니라 종속형 자료인 경우 반응수준이 이분형일 때 McNemar 검정을 사용하고, 범주의 수준이 3개 이상이면 McNemar 검정을 확장한 Cochran's Q 검정을 사용한다.

Table 12. Data summary for alkaline phosphatase (ALP)

Condition 1 (pre-treatment)	Condition 2 (post-treatment)		
	Normal	Abnormal	Total
Normal	60	14	74
Abnormal	6	6	12
Total	66	20	86

참 고 문 헌

1. Agresti A. Categorical data analysis. New York. NY: Wiley, 1990.
2. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall, CRC, 1997.
3. Armitage P, Berry G. Statistical Methods in Medical Research, 3rd edn. Blackwell Scientific Publications, Oxford, 1994.
4. Carlin JB, Doyle LW. Statistics for clinicians. 3: Basic concepts of statistical reasoning: Standard errors and confidence intervals. J Paediatr Child Health 2000; 36: 502-505.
5. Dwyer AJ. Matchmaking and McNemar in the comparison of diagnostic modalities. Radioology 1991; 178: 328-330.
6. Everitt BS. The analysis of contingency tables. 2nd ed. London, UK: Chapman & Hall, 1992.
7. Fisher RA. The logic of inductive inference. J Royal Stat Soc Ser A 1935; 98: 39-54.
8. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 1. hypothesis testing. Can Med Assoc J 1995; 152: 27-32.
9. Joe H. Extreme probabilities for contingency tables under row and colum independence with applications to Fisher's exact test. Comm Stat A Theory Methods 1988; 17: 3677-3685.
10. Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Oxford, UK: Blackwell Science Ltd, 2003.
11. Moser BK, Stevens GR. Homogeneity of variance in the two-sample means test. Am. Statistician 1992; 46: 19-21.
12. Mullner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. Ann Intern Med 2002; 136: 122-126.
13. Rosner B. Fundamentals of biostatistics. 4th ed. Boston, Mass: Duxburry, 1995.
14. Yates F. Contingency tables involving small numbers and the ch-square test. J Royal Stat Soc Ser B 1934; (supp 1): 2179-2235.
15. Zou KH, Fielding JR, Silverman SG, Tempany CM. Hypothesis testing I: proportions. Radiology 2003; 226: 609-613.