臨 床 藥 理 學 會 誌 第20卷 第2號 2012
J Korean Soc Clin Pharmacol Ther 2012;20(2):109-124

임상시험에서 사용되는 기본통계개념에 관한 고찰

[1]홍익대학교 과학기술대학 교양(수학)과,
[2]가톨릭대학교 의과대학 약리학교실 및 서울성모병원 임상약리과

최경미[1], 이종태[2], 전상일[2], 홍태곤[2], 백정기[2], 한승훈[2], 임동석[2]

=Abstract=

A Review of Fundamentals of Statistical Concepts in Clinical Trials

Kyungmee Choi[1], Jongtae Lee[2], Sangil Jeon[2], Taegon Hong[2], Jeongki Paek[2], Seunghoon Han[2], Dong-Seok Yim[2]

[1]Divison of Mathematics, College of Science and Technology, Hongik University at Sejong, Jochiwon, Chungnam, Republic of Korea, [2]Department of Pharmacology, College of Medicine, the Catholic University of Korea and Department of Clinical Pharmacology and Therapeutics, Seoul St. Mary's Hospital, Seoul, Republic of Korea

Statistical analysts engaged in typical clinical trials often have to confront a tight schedule to finish massive statistical analyses specified in a Standard Operation Procedure (SOP). Thus, statisticians or not, most analysts would want to reuse or slightly modify existing programs. Since even a slight misapplication of statistical methods or techniques can easily drive a whole conclusion to a wrong direction, analysts should arm themselves with well organized statistical concepts in advance. This paper will review basic statistical concepts related to typical clinical trials.

The number of variables and their measurement scales determine an appropriate method. Since most of the explanatory variables in clinical trials are designed beforehand, the main statistics we review for clinical trials include univariate data analysis, design of experiments, and categorical data analysis. Especially, if the response variable is binary or observations collected from a subject are correlated, the analysts should pay special attention to selecting an appropriate method. McNemar's test and multiple McNemar's test are respectively recommended for comparisons of proportions between correlated two samples or proportions among correlated multi-samples.

Key words: Measurement scale, Two-sample T-test, Crossover study, Chi-square test, Multiple McNemar's test

교신저자: 최경미
소 속: 홍익대학교 과학기술대학 교양(수학)과
주 소: 충청남도 연기군 조치원읍 신안리 산34 (우 339-701)
전화번호: 044-860-2237, 팩스: 044-860-2648, E-mail: kmchoi@hongik.ac.kr
접수일자: 2012. 06. 14. 수정일: 2012. 11. 02. 게재확정일: 2012. 11. 06.

서 론

생물학적 동등성 시험이나 약물의 제형 또는 용량, 투여경로 등을 변경하기 위한 1상 임상시험에서는 많은 통계분석법들이 사용된다. 이때 통계분석자는 짧은 기간 내에 많은 양의 자료를 표준작업지침서(Standard Operation Procedure)에 따라서 정확하게 분석해야 하기 때문에 기존의 통계처리 프로그램을 일부 수정하여 사용하는 경우가 많다.[1)] 하지만, 통계분석자가 임상시험에서 얻어진 자료의 실험계획법적 특성이나 자료의 특성을 정확하게 이해하지 못한 경우에는 통계처리 프로그램을 수정하는 과정에서 오류가 생길 수 있다.[1)] 이런 오류를 예방하기 위해서 자주 사용되는 실험계획법 용어 및 분석법 사이의 차이를 익히고 정리해두는 것이 좋다. 본 연구에서는 이와 같은 임상 시험에서 자주 사용되는 통계용어 및 분석법들을 검토하여 정리하며, 분석의 목적과 자료의 특성에 따라 어떤 분석법들이 적절한지를 실례를 통하여 살펴본다.

일반적인 생물학적 동등성 시험에서는 관찰연구(observational study)를 하기 보다는 연구 대상 요인과 효율적인 실험 크기를 미리 정한 후, 연구 대상자를 연구 대상 요인에 따라 구분되는 집단에 무작위로 배치함으로써 외부요인을 통제하는 실험계획법(experimental design)[2)]을 주로 사용한다. 그런데, 동일한 목적을 위한 실험일지라도 자료의 특성이 연속인지 이산인지, 각각의 자료가 독립인지 아닌지에 따라서 적용 가능한 통계방법이 달라지기 때문에 본 연구는 우선 자료의 특성을 결정짓는 척도에 관한 검토로부터 출발한다.

각 절은 수학에 근거한 통계적 이론 개발의 큰 줄거리를 따라 구성되었으며, 필요할 경우에는 SAS 프로시져에 대한 설명이 더해졌다. 우선, 자료가 가질 수 있는 척도를 정의하고 분류한다. 변수를 통계적 관점에서 반응변수와 설명변수로 나누어 설명하고, 특히 실험계획법에서 사용되는 설명변수의 정의와 특성을 살펴본다. 또한 임상시험에서 실제로 사용되는 반응변수와 설명변수들을 찾아보고 이들의 척도를 살펴본다. 또한 설명변수가 반응변수에 미치는 영향을 알아보기 위해서 각 척도 별로 사용 가능한 방법을 살펴본다. 둘째, 기술통계량, 점추정, 구간추정, 가설과 검정에 대하여 살펴본다. 셋째, 두 모집단의 평균의 동일성을 검정하는 T-검정에 대하여 살펴본다. 넷째, 임상시험에 주로 사용되는 실험계획법에 대하여 구체적으로 살펴본다. 여기에는 여러 모집단의 평균의 동일성을 검정하는 일원배치 분산분석법(one-way ANOVA)와 여러 개의 약 또는 처리를 다른 순서에 따라서 실험하는 교차설계법(crossover study) 등이 포함되며, 모집단의 분포에 대한 정규성 가정이 없는 경우 사용할 수 있는 비모수적 방법도 포함된다. 마지막으로 약의 이상반응(adverse event) 등과 같이 자료들이 빈도표로 정리될 수 있는 경우에 사용되는 χ^2-통계량에 대하여 알아본다. 또한 독립이 아닌 두 개 이상의 집단에서 비의 동일성을 검정하기 위한 맥네머의 검정(McNemar's test) 및 다중 맥네머의 검정(multiple McNemar's test)에 대하여도 알아본다.

자료의 척도와 통계량 분류

1. 자료의 척도

실험을 통하여 얻어지는 자료의 척도는 크게 명목척도(nominal scale)와 순서척도(ordinal scale), 구간척도(interval scale), 비척도(ratio scale)로 구

분되기도 하고,[3] 이산형(discrete) 또는 연속형(continuous)으로 구분되기도 한다.[4] 명목척도는 성별, 질병 여부, 흡연여부, 음주여부, 부작용 여부 등과 같이 범주형 자료(categorical data)를 표현할 때 사용되며, 값에 순서나 크기가 없고, 이산형으로 분류된다. 순서척도는 값 사이에 순서는 있지만 두 값 사이의 차이를 측정할 수 없는 경우를 나타내며, 리커트 척도(Likert scale)[5-7]나 의미차이 척도(semantic differential scale)를 포함한다. 예를 들어, 통증이나 이상반응의 정도를 표현할 때 '매우 적다', '조금 적다', '그저 그렇다', '조금 심하다', '매우 심하다' 등을 표현하는 5점 척도가 여기에 속한다. 이때, 각 통증을 나타내는 값들 사이에 순서는 있지만, '조금 적다'와 '매우 적다'의 차이를 측정할 수 없으며, 이는 '조금 심하다'와 '매우 심하다'의 차이와 같다고 할 수 없다. 순서척도 자료는 이산형으로 분류되어야 하나, 연속형으로 분류되어 처리되는 경우도 많으며 이는 아직까지 논란이 되고 있다.

명목척도나 순서척도와 달리, 두 값 사이의 크기를 측정할 수 있는 척도로는 구간척도와 비척도가 있으며, 이들은 모두 연속형으로 분류된다. 실제로 구간척도로 얻어지는 자료는 많지 않으나 가장 적절한 예로는 체온과 같은 온도를 들 수 있다. 온도에서 두 온도의 차이(39 ℃ - 38 ℃)를 측정할 수 있으며, 이것은 다른 두 온도 차이(38 ℃ - 37 ℃)와 같은 의미를 갖는다. 그러나 0°는 다른 값들의 기준으로 정해진 상대적인 값일 뿐이며, 두 온도의 비율은 아무런 의미를 갖지 못한다. 반면, 비척도에서는 두 값의 차이를 측정할 수 있을 뿐만 아니라, 두 값의 비율이 의미를 가지며, 0이 물리적으로 '없는' 상태를 나타낸다. 비척도의 예로는 키, 몸무게, 면적, 부피, 농도, AUC_{0-36h}, C_{max} 등이 있다.

2. 확률변수

통계 자료는 일변량 X 또는 이변량(X, Y) 쌍으로 측정이 된다. 수학에서는 X 또는 Y의 값이 변할 수 있기 때문에 이를 변수(variable)라고 정의하지만, 통계에서는 변수의 값들이 확률분포(probability density function)를 만들 수 있을 때에만 확률변수(random variable)로 정의될 수 있다.[8-9] 이변량의 경우 함수 f를 이용하여 X와 Y의 관계를 표현하면 $Y=f(X)+\varepsilon$가 된다. 여기서, $f(X)$는 관측치(observation)를 표현하는 확률변수 Y의 기대값(expected value)이고, ε은 오차항(error term)이다. 통계에서는 X를 독립변수(independent variable) 또는 설명변수(explanatory variable)라고 부르고, Y를 종속변수(dependent variable) 또는 반응변수(response variable)라고 부른다. 이때 Y와 ε은 확률변수이나 X는 실험설계 변수로서 실험설계자에 의해 실험 전에 그 값들이 미리 조정되거나 고정되기 때문에 일반적으로 확률변수가 될 수 없다. 예를 들어, 약의 종류, 약투여순서 등이 실험계획에 따라 미리 조정되는 설명변수이며, 약의 혈중농도나 AUC_{0-36h}, C_{max} 등은 실험설계변수에 따라 달라지는 반응변수가 된다. 기대값(Expected value)을 이용하여 표현해보면, $E[X]=X$, $E[\epsilon]=0$이므로 $E[Y]=f(X)$가 된다.

3. 통계량 분류

변수의 개수, 각 변수의 척도에 따라서 사용할 수 있는 통계량이 달라진다. 자료가 일변량일 경우에는 해당 확률분포의 중심이나 흩어진 정도를 추정하는 일변량분석법이 필요하며, 이변량(X, Y) 쌍인 경우에는 X와 Y의 관계를 추정하는 회귀분석,

실험계획법, 범주형자료 분석 등이 필요하다. 특히 X와 Y가 모두 연속척도이면 이들의 관계를 평면 위의 직선식으로 표현할 수 있으며, X와 Y의 관계를 추정하는 문제는 회귀분석(regression)을 이용하여 직선의 기울기와 절편을 추정하는 문제가 된다. 일반적으로 회귀분석에서(X, Y)는 관찰연구에 의하여 측정되는 값이다. X가 명목척도이고 Y가 연속척도인 경우에, X의 각 값이 집단을 나타내므로 집단 간 Y의 평균을 비교할 수 있다. 이때, 일반적으로 X의 각 값은 자료를 측정하기 전에 미리 설계 또는 계획되기 때문에 이 방법을 실험계획법(design of experiments)이라고 부른다. 특히 X가 두 개의 집단만을 표현하는 이항형(binary)척도이고 Y가 연속척도이면, X와 Y의 관계를 추정하는 문제는 두 모집단의 평균을 비교하는 문제가 되며 일변량분석으로 분류된다. X가 세 개 이상 값을 가지는 명목척도이고 Y가 연속척도이면 X와 Y의 관계를 추정하는 문제는 세 개 이상 모집단의 평균을 비교하는 문제가 되며, 이는 실험계획법 중 일원배치 분산분석법(one-way ANOVA)으로 분류된다. 두 개 이상의 X가 두 개 이상의 값을 가지는 명목척도이고, Y가 연속척도이면 X와 Y의 관계를 추정하는 문제는 일반적인 실험계획법으로 분류된다. X와 Y가 모두 명목척도일 경우에는 범주형자료분석이 적절하며, 자료는 빈도를 이용한 교차표(cross table) 등으로 정리될 수 있다. X또는 Y가 여러 개일 경우 또는 여러 개의 Y만 있을 경우를 다루는 문제를 다변량분석(multivariate analysis)라고 부른다. 특히, X가 시간을 나타내는 여러 개의 값을 가지고, Y가 각 개인 별 약의 혈중농도 등을 나타내는 연속척도 변수일 경우 이는 반복측정설계(repeated measures design)또는 피험자내설계(within-subjects design)이라 불리며, 검정하고자 하는 가설에 따라서는 다변량 분산분석법 등으로도 분석이 가능하다.[3)]

일반적인 임상시험에서는 비교하고자 하는 집단이 미리 설계되는 경우가 많기 때문에 실험계획법과 관련된 통계분석법이 더 자주 사용된다. 따라서 본 논문에서는 일변량분석과 실험계획법, 범주형 자료분석과 관련된 기본적인 통계개념을 살펴보며, 회귀분석이나 다변량분석은 생략한다.

일변량 분석

1. 기술통계량

기술통계량(descriptive statistics)은 임상시험을 통하여 얻어진 자료의 주요 특성을 정량적으로 정리하는 가장 기본적인 통계량이다.[4)] 자료의 중심집중경향(central tendency)을 표현하는 기술통계량으로는 평균(mean), 중앙값(median), 최빈값(mode)이 있는데, 평균은 연속형 자료에, 중앙값은 순서척도 자료에, 최빈값은 순서척도 또는 명목척도 자료에 적용될 수 있다. 자료의 산포(spread)를 표현하는 기술통계량으로는 분산(variance), 표준편차(standard deviation), 범위(range) 등이 있는데, 이들은 모두 연속형 자료에 적용될 수 있다. 평균이나 분산, 표준편차 등을 순서척도 자료에 적용하는 것은 아직까지 많은 논란이 되고 있다.[5-7)]

분포의 모양을 표현하기 위해서 사용되는 그래프로는 막대그래프(bar chart)와 히스토그램(histogram)이 있다. 막대그래프는 범주형 자료의 분포를 표현할 때 사용되므로, 수평축 값들 사이에는 순서가 없고 막대의 높이가 해당 자료의 빈도 또는 비율이 된다. 그러나 빈도는 자료가 커지면 같이 커질 수 있기 때문에 비율을 막대의 높이로 사용하는 것이 더 적절하다. 히스토그램은 연속형 자료의 분포를 표현할 때 사용되므로 수평축 값들 사이에 순서가 있으며

막대의 면적이 해당 구간에 속하는 자료의 비율이 되며, 총면적은 1이다. 히스토그램은 정규분포(normal distribution)의 확률밀도함수(probability density function)와 자주 비교되어, 자료가 정규분포를 따르는지 판단할 수 있는 근거가 된다. 이를 검정하는 방법으로는 샤피로(Sapiro)의 검정[3]이나 콜모고로프-스미어노프(Kolmogorov-Smirnov) 검정 등이 있다.[8] 순서척도를 갖는 자료들의 분포는 두 가지 방법으로 모두 표현 가능하나, 막대그래프가 더 자주 사용된다. 히스트그램 이외에도 연속형 자료의 분포를 표현하기 위해서 상자도표(box plot)이나 점도표(dot plot), 줄기-잎 그래프(stem-and-leaf graph) 등이 사용된다. 분포의 쏠림을 표현하기 위해서 왜도(skewness)가 사용되며, 분포의 꼬리의 두께를 표현하기 위해서 첨도(kurtosis)가 사용된다.

2. 추정 및 표본분포

모집단(population)의 모평균 μ를 추정하기 위하여 크기 n인 표본(sample)의 표본평균을 점추정치(point estimator)로 사용한다. 이때 모집단으로부터 무작위로 추출된 각각의 관측치(observation)는 독립이고 모평균 μ와 모분산 σ^2을 가지는 동일한 분포를 따른다고 가정한다. 표본크기 n이 클 때 모집단의 분포와 상관없이 표본평균 $\overline{X}$의 표본분포(sampling distribution)는 중심극한정리(central limit theorem)에 따라서 근사적으로 평균이 μ이고, 분산이 $\frac{\sigma^2}{n}$인 정규분포를 따른다.[4,8-11] 이때, 모분산 σ^2대신 표본분산 s^2을 사용할 경우, 표본평균 $\overline{X}$의 표본분포는 자유도(degree of freedom)가 $(n-1)$인 t분포를 따르게 되며, 이들은 다음과 같이 표현된다.

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \rightarrow N(0,1) \text{ 그리고 } \frac{\overline{X}-\mu}{\varepsilon/\sqrt{n}} \rightarrow t(n-1)$$

점추정치와 모평균의 차이를 편향(bias)이라고 부르며, 편향이 0인 점추정치를 일치추정치(consistent estimator)라 부른다.[8] 그러나 표본평균이 모평균과 동일할 확률은 이론적으로 0이므로, 표본평균만을 이용하여 모평균을 추정하기 보다는 모평균을 포함하는 구간을 추정하는 것이 더 안전하다.

모평균을 포함하는 구간을 만드는 과정을 구간추정(interval estimation)이라고 부른다. 100$(1-\alpha)$ % 신뢰구간은 모평균 μ를 포함할 확률이 $(1-\alpha)$인 구간을 의미하며, 모집단의 분포가 정규분포를 따르고 모분산 σ^2이 알려져 있을 때, 다음과 같이 얻어진다.

$$(\overline{X}-Z_{\frac{\alpha}{z}}\frac{\sigma}{\sqrt{n}},\ \overline{X}+Z_{\frac{\alpha}{z}}\frac{\sigma}{\sqrt{n}})$$

여기서 확률변수 Z가 표준정규분포 $N(0,1)$를 따를 때, $Z_{\frac{\alpha}{z}}$는 $P(Z>Z_{\frac{\alpha}{z}})=\frac{\alpha}{z}$로 정의된다. 가장 많이 사용되는 신뢰수준(confidence level) 100$(1-\alpha)$ %는 90 %, 95 %, 99 %이며, 해당하는 $Z_{0.05}$=1.645, $Z_{0.025}$=1.96, $Z_{0.005}$=2.58이다. 일반적으로 모분산 σ^2이 알려져 있지 않으므로 표본분포가 자유도 $(n-1)$인 T분포를 따르는 표본평균 $\overline{X}$를 이용하여 모평균 μ에 대한 100$(1-\alpha)$ % 신뢰구간은 다음과 같이 구한다.[4,8-11]

$$(\overline{X}-t_{\frac{\alpha}{2}}(n-1)\frac{3}{\sqrt{n}},\ \overline{X}+t_{\frac{\alpha}{2}}(n-1)\frac{3}{\sqrt{n}})$$

표본이 달라지면 신뢰구간도 달라지며, 신뢰구간은 실제 모평균 μ를 포함할 수도 있고, 포함하지 않을 수도 있다. 신뢰수준(confidence level) 100(1-0.05) %의 의미는 표본크기가 동일한 100개의 표본을 이용하여 100개의 95 % 신뢰구간을 만들면, 이중 5개는 실제 모평균 μ를 포함하지 않

을 수도 있다는 것을 의미한다.[4] 관측치가 질병 여부나 부작용 여부와 같이 0 또는 1의 값을 가지면 모평균 μ는 모비율 p가 되며, 표본평균 $\overline{X}$는 표본에 나타난 질병의 비율 또는 부작용의 비율 $\hat{p}$이 된다. 표본의 크기가 클 때 표본비율의 분포는 중심극한정리에 따라서 근사적으로 정규분포를 따른다.[4,8-11]

3. 가설과 검정

모평균 μ가 특정한 값 μ_0인지 아닌지를 판단하고자 하는 경우에는 가설검정(hypothesis test)을 실시하여야 한다.[4,9-11] 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)은 다음과 같다.

$H_0 : \mu = \mu_0$ vs. $H_0 : \mu \neq \mu_0$ (양측검정)

또는 $H_0 : \mu \leq \mu_0$ vs. $H_0 : \mu > \mu_0$ (단측검정)

또는 $H_0 : \mu \geq \mu_0$ vs. $H_0 : \mu < \mu_0$ (단측검정)

검정통계량은 귀무가설이 참이라는 가정 하에 얻어지는 표준화된 표본평균으로 다음과 같다.

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$ (σ가 알려진 경우)

$$T = \frac{\overline{X} - \mu_0}{\varepsilon / \sqrt{n}}$$ (σ가 알려지지 않은 경우)

검정통계량이 정해진 기각역에 속하면 '귀무가설을 기각한다'고 말한다.

이와 같은 통계적 결정을 내릴 때에는 귀무가설이 참일 때 귀무가설을 기각하는 제1종의 오류(type Ⅰ error)와 대립가설이 참일 때 대립가설을 기각하는 제2종의 오류(type Ⅱ error)가 발생할 수 있다. 제1종의 오류가 발생할 확률을 유의수준(significance level) α라고 부르고, 제2종의 오류가 발생할 확률을 β라고 부르며, $(1-\beta)$를 검정력(power of the test)이라 부른다. 기각역은 유의수준 α에 따라서 결정되며 이것은 $100(1-\alpha)$ % 신뢰구간과 여집합이다.[4,9-11] 따라서 신뢰구간이 μ_0를 포함하지 않으면 귀무가설을 기각한다. 표본크기가 커지거나 유의수준이 커지면, 신뢰구간의 길이가 짧아지므로, 귀무가설이 기각될 확률과 검정력이 커진다.[10-11]

신뢰구간이나 기각역 이외에 검정에 사용될 수 있는 값으로는 검정통계량에 대한 P-값(P-value)이 있는데, 이것은 귀무가설이 참이라는 가정 하에 얻어지는 검정통계량보다 더 극단적인 결과가 나올 수 있는 확률로 정의된다.[4,8-11] 따라서 P-값이 유의수준 α보다 작으면 검정통계량이 기각역에 속하므로 귀무가설을 기각할 수 있다. 통계패키지들은 통계분석가들이 사용할 유의수준 α를 알지 못하기 때문에, 자료로부터 계산되는 검정통계량에 대한 P-값을 자동으로 계산하여 돌려주므로 이를 각자가 정한 유의수준 α와 비교하여 가설검정을 하면 된다.

표본의 크기가 충분히 큰 경우에는 T분포가 Z분포에 근사적으로 접근한다. 이는 T와 Z의 평균이 같고, T의 분산이 Z의 분산보다 크지만 표본의 크기가 커지면 둘의 분산이 근사적으로 같아지기 때문이다. 특히 표본의 크기가 30보다 크면 R이나 Excel 등을 이용하여 두 값의 차이가 0.01 미만으로 떨어지는 것을 쉽게 확인할 수 있으므로, 이 경우에는 어느 것을 사용하나 일반적으로 사용되는 유의수준 0.1, 0.05, 0.01에서는 검정의 결과가 달라지지 않는다.

두 모집단에 대한 평균의 동일성 검정

임상시험에서 두 약이나 처리 사이의 평균 차이가 있는지를 알아보고자 할 때 가설은 $H_0 : \mu_1 = \mu_2$

vs. $H_A : \mu_1 \neq \mu_2$이다. 검정통계량은 모집단의 분포, 표본의 크기, 모분산 σ_1^2과 σ_2^2를 아는지 또는 모르는지에 따라서 달라지며 Table 1과 같이 정리될 수 있다. 일반적인 임상시험에서는 모집단의 크기가 충분히 크고, 분산이 알려진 경우가 흔하지 않기 때문에 르빈(Levene)의 검정을 이용하여 등분산성을 검정한 후, 등분산 또는 이분산 각각에 해당하는T-검정을 사용하여 평균의 동일성을 검정한다. 두 T-검정의 차이는 다른 분산을 사용함으로 인한 자유도의 차이로 결정된다. 표본의 크기에 대한 의견에는 조금씩 차이가 있으나, 소표본의 기준은 10 이상, 대표본의 기준은 30 이상이 일반적이다.[10-11)]

A제약회사 약의 임상시험에서 22명의 피험자에서 투여된 대조약과 시험약의 혈중농도를 비교해 보자. 약의 혈중농도는 일반적으로 로그정규분포를 따르는 것으로 알려져 있으며,[12)] 한 피험자가 대조약과 시험약을 모두 투여받기 때문에 피험자 효과를 교정할 수 있는 쌍체비교법(paired-T test)을 이용한다. 최대혈중농도(C_{max})에 로그변환을 취하면, 다음과 같은 C_{max}비의 로그변환이 된다.

$$d_i = \ln C_{max,T,i} - \ln C_{max,R,i} = \ln C_{max,T,i}/C_{max,R,i}$$

이를 분석하기 위한 SAS 프로그램으로는 PROC MEANS 또는 PROC UNIVARIATE이 있으며 후자는 다음과 같다.[13)]

```
PROC UNIVARIATE;
VAR DIFF;
RUN;
```

이 프로그램을 이용하여 두 처리의 차이 DIFF (d_i)의 평균에 대한 신뢰구간을 구하면 된다. 그런데, 일반적으로 신뢰수준은 통계프로그램 내부적으로 정해져 있지 않기 때문에 옵션으로 적어주지 않으면, 신뢰구간이 출력되어 나오지 않을 경우도 있다. 따라서 출력결과 중 평균(MEAN)과 표본오차(STDERR)를 이용하여 각자가 정한 신뢰수준에 근거하여 (MEAN-$t_{\frac{0.10}{2}}$ (22-1) × STDERR, MEAN+$t_{\frac{0.10}{2}}$(22-1) × STDERR)을 계산하면, 이 구간이 90 % 신뢰구간이 된다. 여기서 $t_{\frac{0.10}{2}}(22-1)$ =1.72이다. 이 차이에 대한 90 % 신뢰구간을 구한 후 지수변환을 하면, C_{max}비에 관한 신뢰구간을 (0.3560, 0.4394)와 같이 얻을 수 있다. 동일한 방법으로 AUC_{0-36h}비에 대한 90 % 신뢰구간을 (0.3538, 0.4112)로 구할 수 있다. 이 구간들이 1을 포함하지 않으므로 두 경우 모두에서 대조약과

Table 1. Two-sample location tests

test	assumptions				
	population	population variance	independence	sample size	sampling distribution
$H_0 : \mu_1 = \mu_2$	normal	known	independent	small	Z
$H_0 : \mu_1 = \mu_2$	normal	unknown	independent	small	T†
$H_0 : \mu_1 = \mu_2$	non-normal	unknown equivalence test‡	independent	large $n_1 \geq 30, n_2 \geq 30.$	T†
$H_0 : \mu_d = 0$	normal	unknown	dependent (paired)	small	T*
$H_0 : p_1 = p_2$	Bernouille	unknown	independent	large $n\hat{p} \geq 10, n\hat{q} \geq 10$	Z

* One-sample T-test with degree of freedom (n-1). † Two-sample T-test with degree of freedom (n_1+n_2-2) or min (n_1-1, n_2-1). ‡ Levene's Test for equality of variances.

시험약의 C_{max} 또는 AUC_{0-36h}가 동등하다는 귀무가설을 기각한다. Phoenix WinNonlin 6.2를 이용할 경우, 이 신뢰구간을 바로 얻을 수 있다.

Y자료들이 정규분포를 따르지 않을 경우, 비모수적 방법인 윌콕슨 순위합(Wilcoxon rank sum) 검정을 사용할 수 있다.[9] 하지만 이때에는 기존의 자료를 순위로 바꾸어 계산하는 것이 타당한지를 검토해야 한다. 예를 들어, 자료의 척도가 순서척도이더라도 실제로 나타나는 순서가 2 또는 3 종류로 매우 적으면 비모수적 방법보다는 범주형자료분석을 적용하는 것이 더 적절하다. 또한, 0/1을 가지는 이항형 Y에 대하여 그대로 T-검정을 하면, 이것은 비율검정이 된다. 윌콕슨 순위합 검정은 만-위트니(Mann-Whitney) 검정과 동등하다.

Table 1은 모집단의 분포과 표본의 독립성, 표본분산이 알려졌는지 여부, 모분산의 동질성, 표본의 크기 등에 따라 달리 써야 하는 두 모집단의 평균검정법들을 보여주고 있다. 가장 주의할 점은 분산이 알려져 있지 않으면 T-검정을 써야 하나, 이 경우 소표본이면 반드시 정규성이 만족되어야 한다는 것이다.

실험계획법에 의한 임상자료분석

1. 실험계획법의 변수들

1) 실험설계변수

이 실험설계에서는 대조약 또는 시험약을 구분하는 처리(treatment)가 실험설계변수 X가 된다. X는 0/1 또는 -1/1 등의 값을 가질 수 있고, 이것은 피험자를 두 집단으로 구분한다. 이와 같은 X값을 수준(level)이라고 부르며 이들은 실험설계단계에서 미리 결정되어야 한다. 이때 추정되는 X의 계수는 전체평균에 대비되는 두 집단의 평균효과가 되며, 두 집단이 고정되어 있어서 이것이 분포를 이루기 어렵기 때문에 상수로 취급하여 추정된다. 한편, 피험자로 인한 효과의 변동이 발생할 수 있기 때문에 피험자도 실험설계변수 X가 된다. 이때 추정되는 X의 계수는 전체평균에 대비되는 피험자의 평균효과를 나타낸다. 피험자가 충분히 큰 모집단에서 무작위로 추출된다고 가정하기 때문에 피험자 효과는 분포를 이룬다. 처리나 피험자와 같은 실험설계변수를 요인(factor)이라고 부르고, 분포를 가지지 않는 X의 계수(coefficient)를 고정효과(fixed effect)라 정의하며, 분포를 가지는 X의 계수(coefficient)를 변량효과(random effect)라 정의한다.[2] 즉, X의 값에 따라서 기울기와 절편이 달라지는 것이 몇 개의 값만으로 한정 또는 고정되는지, 아니면, 그 변동을 몇 개의 값으로 표현하기 보다는 분포를 사용하는 것이 더 적절한지를 결정하여 X를 고정효과 또는 변량효과로 구분한다.

두 종류의 약을 투약 후 피험자의 약의 혈중농도를 일정한 시간간격으로 반복 측정이 하는 임상시험을 생각해보자. 한 피험자 자료는 시간에 대하여 프로파일(profile)을 형성하며, 이 프로파일의 평균이나 기울기의 변동은 약의 종류에 따라서 생기기도 하나, 피험자 개인의 특성에 따라서도 조금씩 달라진다. 여기서 약의 종류에 따라서 달라지는 변동은 고정효과로 설명되며, 피험자 차이 때문에 발생하는 변동은 변량효과로 설명된다. 특히, 임상시험의 목적이 임상시험에 참여한 개개인의 피험자의 특성이 아닌 피험자가 속한 모집단의 특성이므로 피험자를 변량효과로 처리한다.

위와 같은 설명에도 불구하고, 고정효과에 비해서 변량효과를 이해하는 것이 훨씬 어렵다. 피험자처럼 충분히 큰 모집단으로부터 무작위로 추출되는 대표적인 변량요인으로는 시간이나 장소 등

이 있다. 예를 들어, 한 실험단위에 동일한 처리가 여러 시간대에 반복적으로 행해지면, 시간대별 날씨 등의 변화가 실험단위의 상태에 영향을 미칠 수 있다. 실험을 위해 결정되는 시간대는 무한히 큰 모집단에서 무작위로 추출될 수 있으므로 각 시간대 별 효과는 일반적으로 분포를 이룰 수 있기 때문에 변량인자로 처리될 수 있다. 그러나, 변량인자가 많아질수록 모형이 복잡해지므로, 변량효과가 미미하다고 판단되는 경우 이를 고정효과로 두고 모형을 만들어 분석한다. 또한 변량인자의 수준수가 적을 경우, 이를 고정효과로 처리하기도 한다.

2) 종속변수

실험계획법 모형에서 오차항은 독립이고 등분산을 갖는 정규분포를 따라야 한다. 따라서 실험계획법의 종속변수는 적어도 구간척도이어야 한다. 만약 종속변수가 순서척도일 경우에도 모수적(parametric) 실험계획법에 따른 분석법을 사용하는 경우도 있으나 아직까지 이에 대한 논란이 많다.[5-7] 이와 같이 종속변수가 연속척도를 갖지 못할 경우에는 비모수적(nonparametric) 방법을 사용하는 것이 좋으나, 이 경우에도 오차항의 독립성에 대한 가정은 지켜져야 한다.

3) 외생변수

실험계획 단계에서 설계하여 조정할 수 없는 변수를 공변량(covariate)이라고 부른다.[2,14] 임상시험과 관련된 실험계획법에서는 변수의 척도와 무관하게 성별, 키, 몸무게 등과 같이 실험설계로 조정할 수 없는 모든 변수들을 공변량이라고 둔다. 이와 같이 공변량을 정의할 경우, 모형에 따라서는 공변량이 고정효과를 갖는 요인으로 처리될 수도 있다. 회귀분석과 같은 일반적인 통계분석에서는 공변량을 일반적인 독립변수 또는 설명변수로 취급한다. 그러나 실험계획법에서는 성별, 키, 몸무게 등이 종속변수인 AUC_{0-36h}, C_{max} 등에 영향을 미치는 설명변수이기는 하나 실험자가 조정할 수 없는 변수들이라는 의미에서 이들을 따로 분류하여 외생변수라고 이름 붙여 사용하기도 한다.

2. 무작위 배정

실험단위(experimental unit)가 되는 한 피험자가 대조약과 시험약을 투여받는 임상시험에서는 피험자를 충분히 큰 모집단으로부터 무작위로 추출(random sampling)하기 어렵고, 표준작업지침서에서 정한 일정한 조건을 만족하는 지원자를 피험자로 정한다. 실험에서는 편향을 최소화하기 위해서 피험자를 대조약과 시험약에 무작위로 배정(random assignment or randomization)해야 하고, 실험순서 또한 무작위로 정해야 한다. 만약 전체적인 무작위 배정이 불가능한 경우에, 실험일, 실험장소 또는 시간적 차이를 두고 실시되는 반복 등은 블록요인이 되고, 랜덤으로 택한 드럼통, 로트(lot) 등은 집단요인이 된다.[15] 예를 들어, 두 종류의 약을 투약 후 피험자의 약의 혈중농도를 일정한 시간간격으로 반복 동일한 임상시험을 여러 임상센터에서 나누어 진행하는 경우에 임상시험센터는 집단요인으로 처리될 수 있다.

이외에도 가능하면 피험자가 처리의 종류를 알지 못하게 하는 눈가림(blind) 실험 또는 피험자와 실험자 모두 처리의 종류를 알지 못하게 하는 이중 눈가림(double-blind) 실험이 권장된다.

3. 일원배치 분산분석법

설명변수 X가 대조약, 시험약 1, 시험약 2 등과 같은 명목척도를 가질 경우에 설명변수 X는 실험집단을 표현하게 된다. 이때 반응변수 Y가 피험자에게서 측정된 혈중 약의 농도와 관련된 연속형 척도를 가지게 되면, X와 Y의 관계를 알아보는 문제는 집단 사이의 평균 차이를 알아보는 것과 같아진다. 3개 이상 집단에서 평균의 동일성을 검정하고자 할 때 일원배치 분산분석법(one-way ANOVA)[2,9,15]을 사용한다. 이 평균비교 방법을 굳이 분산분석(ANOVA)라고 부르는 이유는 평균 차이가 동일한 경우에도 분산이 작을 경우 그 차이가 더 커 보이고, 반대로 분산이 클 경우에는 그 차이가 더 작아 보이는 것처럼 분산에 따라서 평균의 상대적 차이가 달라지기 때문이다. 이 모형을 수식으로 표현해보면,

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim N(0, \sigma^2)$$

과 같다. 여기서 y_{ij}는 i번째 집단의 j번째 관측치이고, μ는 전체평균이다. α_i는 집단의 고정효과이며, ε_{ij}는 오차이다. 이때 α_i는 $y = X\beta + \varepsilon$ 선형모형에서 X에 0/-1/1 등의 값을 넣어 계산한 후 얻어지는 β의 다른 표현이며, 실험계획법에서는 X없이 계수 또는 효과(effect)만을 사용한 모형을 일반적으로 사용한다.

i번째 집단평균을 $\bar{y}_i$, 총평균을 $\bar{y}$라 두면, 분산분석 자료의 총변동 $(\bar{y}_{ij} - \bar{y})$은 집단 차이로 인하여 각 집단의 평균이 전체 평균으로부터 흩어진 정도를 나타내는 집단간 변동(between group variation) $(\bar{y}_i - \bar{y})$과 우연한 오차로 인하여 각 관측치가 집단 평균으로부터 흩어진 정도를 나타내는 집단내 변동(within group variation) $(\bar{y}_{ij} - \bar{y}_i)$의 합으로 나타난다. 즉,

$y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$이 되며, 이때 집단간 변동과 집단내 변동은 항상 직교하며 독립이다. 집단간 변동이 집단내 변동보다 클수록 집단 사이에 유의한 평균차이가 있다고 판단할 수 있는 근거가 커진다. 이 변동들의 제곱합(sum of squares)은 표본의 크기가 커질수록 커지기 때문에 이것을 그대로 쓰기 보다는 이 제곱합을 사용가능한 자료의 개수인 자유도로 나눈 평균을 사용하며, 이를 평균제곱합(mean square)라고 부른다. 이 두 평균제곱합의 비율로 얻어지는 통계량은 다음과 같이 정의된다.

F-검정 통계량=

$$\frac{\text{집단간평균분산(MStr)}}{\text{집단간평균분산(MSE)}} = \frac{\text{SStr}/(c-1)}{\text{SSE}/(n-c)}$$

여기서 c는 집단의 수이고 n은 총실험횟수이며, 이 통계량이 $F_\alpha(c-1, n-c)$ 보다 크면 귀무가설 'H_0: 모든 집단의 평균이 동일하다'를 기각한다. 일원배치 분산분석법 등의 실험에서 유의할 또 한 가지 사항은 피험자 등의 실험단위를 전체 실험에 무작위로 배치하는 일이다.

일반적인 일원배치 분산분석법은 자료들이 정규분포를 따르며 독립임을 가정한다. 그러나 자료가 정규분포를 따르지 않을 때에는 비모수적 방법인 크루스칼-왈리스(Kruskal-Wallis) 검정을 할 수 있다.[9] 이 경우에도 자료의 독립성은 가정되어야 한다. 집단의 수가 2일 때, 크루스칼-왈리스 검정은 윌콕슨 순위합 검정과 동등하다.

모든 집단에서 평균이 동일한지를 알아보는 일원배치 분산분석법과는 달리 다중비교법은 가능한 모든 쌍 별로 두 집단 간 평균 차이가 있는지를 일대일로 알아본다. 본페로니(Bonferonni)방법이 가장 기본적인 개념을 제공하며, 최소차이검정법(LSD)와 던컨(Duncan)방법 등이 자주 사용된다.[2]

4. 교차설계법

약동학 또는 약력학의 임상시험에서 가장 널리 쓰이는 실험계획법 중 하나가 교차설계법이다. 한 피험자는 하나의 실험 투여군(sequence)에 무작위로 배정되고, 정해진 시간(period)에 그 투여군에 해당하는 처리를 차례로 받게 된다.[9,16-20] 예를 들어, RT에 배정받은 피험자는 시기1에 대조약을, 시기 2에 시험약을 차례로 투여받는다. 2×2 교차설계법은 아래와 같다:

RT
TR.

다른 투여군에는 동일한 처리들이 다른 순서로 나열되어 있어서, 처리들 사이의 잔류효과(carryover effect)가 서로 상쇄되도록 하는 실험계획법이다. 이때, 다른 피험자들을 어느 투여군에 배정할지에 대한 것만을 무작위로 결정하고, 나머지 실험은 정해진 실험계획법의 순서에 따라서 해야한다.

2×2 교차설계법에서 고정효과를 갖는 요인은 대조약(R)과 시험약(T)를 나타내는 처리(treatment)와 피험자가 받는 처리의 순서를 나타내는 투여군(sequence), 실험이 실시되는 시기(period)이며, 변량효과를 갖는 요인은 블록효과를 유발하는 피험자(subject)가 된다. 이처럼 고정효과와 변량효과가 동시에 나타나는 실험계획법을 혼합효과모형(mixed-effect model)이라고 부르는데, 이것의 분산분석은 고정효과모형(fixed-effect model)보다 더 복잡하다. Grizzle이 제안한 모형은 다음과 같다.[19]

$$y_{ijk} = \mu + b_{ij} + \pi_k + \phi_m + \lambda_m + \varepsilon_{ijk}$$

여기서 μ는 전체평균이며, b_{ij}는 i번째 투여군에 속한 j번째 피험자를 나타내는 변량효과이며, $N(0, \sigma_b^2)$을 따른다. π_k는 k번째 시기를 나타내는 고정효과이며, ϕ_m는 m번째 약 또는 처리를 나타내는 고정효과이다. 또한 λ_m는 m번째 약 또는 처리의 잔류효과를 나타내는 고정효과이고, ε_{ijk}는 오차항이며, $N(0, \sigma_e^2)$을 따른다. 이 모형에서는 처리와 투여군*시기, 시기와 투여군*처리, 투여군과 처리*시기 효과들이 서로 교락(confounded)되어 있다.[20]

이러한 자료구조를 분석하기 위해서 사용되는 SAS 프로시져로는 GLM, MIXED 등이 사용될 수 있으며,[19] 좀더 일반적인 형태의 교차설계법으로 얻어진 자료의 분석을 위해서는 따로 SAS macro를 짜는 것이 좋다.[16] 교차설계법에서의 주된 관심사인 대조약과 시험약의 평균차이를 검정하는 가장 단순한 분석법으로는 각 피험자에서 (T-R)을 구하여, 단일표본 T-검정을 하는 것이다.[9]

예를 들어, 2×2 교차설계법에 따라서 얻어진 자료를 분석하는 SAS프로그램은 다음과 같다.[19]

```
PROC GLM;
CLASS SEQUENCE PATIENT PERIOD
TREATMENT;
MODEL RESULT= SEQUENCE PATIENT
(SEQENCE) PERIOD TREATMENT;
RANDOM PATIENT(SEQUENCE);
TEST H=SEQUENCE E=PATIENT
(SEQUENCE);
LSMEANS SEQUENCE TREATMENT;
RUN;
```

이때 피험자는 SUBJECT로 표현하는 것이 더 적합하나, SAS 내부에서 이를 내부변수로 이미 사용하고 있기 때문에, 이를 PATIENT 또는 ID 등으로 바꾸어 프로그램을 짜는 것이 좋다.

A 제약회사 약에 대한 임상시험 결과 얻어진 Table 2의 분산분석표를 살펴보자. 제곱합(sum of squares)은 해당 효과로 인하여 발생하는 집단간 변동의 제곱합을 의미한다. 앞서 설명한 바와 같이 이 값은 표본의 크기가 커질수록 커지기 때

Table 2. Analysis of variance for crossover studies with A corporation's drug

Ln (C_{max})					
Source	DF*	SS†	MS‡	F-stat	P-value**
Sequence	1	0.0388	0.0388	0.6516§	0.4290
Sequence×subject	20	1.1924	0.0596	1.4701	0.1981
Form	1	9.3830	9.3830	231.3650‖	<0.0001
Period	1	0.0122	0.0122	0.3019¶	0.5888
Error	20	0.8111	0.0406	-	-
Ln (AUC_{0-36h})					
Source	DF	SS	MS	F-stat	P-value
Sequence	1	0.0291	0.0291	0.5277§	0.4760
Sequence×subject	20	1.1046	0.0552	2.6723	0.0166
Form	1	10.1348	10.1348	490.3870‖	<0.0001
Period	1	0.0128	0.0128	0.6192¶	0.4406
Error	20	0.4133	0.0207	-	-

* Degree of Freedom. † Sum of Squares. ‡ Mean Square=Sum of Squares/Degree of Freedom. § F-stat=(MS Sequence)/(MS Sequence×subject). ‖ F-stat=(MS Form)/(MS Error). ¶ F-stat=(MS Period)/(MS Error).

문에 이것을 그대로 쓰기 보다는 이를 자유도로 나눈 평균제곱합(mean square)을 사용한다. 이를 계산하기 위하여 제3형 제곱합(type 3 SS or partial SS)이 가장 널리 사용되는데, 이것은 모든 요인들이 모형에 포함되어 있다는 가정 하에 검정하고자 하는 한 요인을 모형에서 제거할 때 생기는 제곱합의 변동이 유의한지를 자유도 (1, n-c)인 F-검정을 통하여 알아보는 방법이다.[13] 여기서 c는 집단의 수이다. ln(C_{max})의 분산분석을 위한 Table 2에서는 처리효과에 대한 P-값이 0.05보다 작으므로 처리효과가 유의하며, 대조약과 시험약 사이에 유의한 평균차이가 있다고 볼 수 있다. 마찬가지로, ln(AUC_{0-36h})의 분산분석표에서도 처리효과의 P-값이 0.05보다 작으므로 대조약과 시험약 사이에 유의한 평균차이가 있다고 볼 수 있다. 이때 투여군과 피험자의 교호작용 또한 유의하게 나타나지만, 이는 관심의 대상이 아니므로 무시한다.

범주형자료분석

1. 독립표본의 범주형자료분석

음주여부, 흡연여부, 카페인섭취여부와 같이 X와 Y가 이항형과 같은 명목척도를 갖는 범주형 자료인 경우에는 빈도를 교차표(cross table)로 정리하여 'X와 Y가 독립(independent)이다' 또는 'X와 Y사이에 관계(association)가 없다'는 등의 귀무가설에 대한 적합도 검정(Goodness-of-fit test)을 할 수 있다. 이때 검정통계량은 X와 Y가 독립이라는 가정 하에 얻어지는 기대빈도와 관측빈도의 차이를 이용하여 계산되는 피어슨의 카이제곱검정(Pearson's x^2-test)이 있으며, 이는 다음과 같이 얻어진다.[4,9,13]

$$x^2 = \sum_{\text{모든 셀}} \frac{(O-E)^2}{E}$$

여기서 O는 셀의 관측빈도이며, E는 셀의 기대빈도이다. 이 검정을 사용하기 위해서는 다음의 두 가지 가정이 만족되어야 한다. 첫째, 교차표의 각 셀(cell)에 속하는 자료가 독립이어야 한다. 즉, 한 실험 단위는 한 셀에만 속해야 한다. 둘째, 기대빈도가 5 이상이어야 한다.[4,13] 만약, 기대빈도가

5 이하인 셀이 하나라도 있을 경우에는 초기하분포(hyper geometric distribution)에 근거한 피셔통계량(Fisher's exact test)을 대신 사용하여야 한다.[13)]

50명의 피험자를 대상으로 실시된 Table 3의 B 제약회사 약의 임상시험 예에서 독립변수 X는 투여군을 표현하며, 종속변수 Y는 흡연여부, 음주여부, 카페인 섭취여부를 표현한다. 둘 다 YES/NO의 두 값만 갖는 명목척도이므로 범주형 자료분석을 실시하며, 이를 위한 SAS 프로그램은 다음과 같다.[13)]

```
PROC FREQ;
TABLES SEQUECNE*SMOKE/CHISQ;
TABLES SEQUECNE*DRINK/CHISQ;
TABLES SEQUECNE*CAFFEIN/CHISQ;
RUN;
```

이 프로그램은 다양한 종류의 적합도 검정 통계량을 출력결과로 돌려주는데, 이중에서 제일 앞에 나오는 'chi-square'가 피어슨의 카이제곱 검정통계량이고, 앞에서 다섯 번째 나오는 통계량이 피셔통계량이다. 흡연여부나 카페인섭취여부의 경우, 기대빈도가 5인 셀이 하나 이상 있으므로 피셔통계량을 사용하여 분석하는 것이 적당하고, 나머지 음주여부의 경우에는 피어슨의 카이제곱검정을 하는 것이 적절하다. 세 경우 모두, P-값이 0.05 이상이므로, 유의수준 0.05에서 투여군에 따른 흡연여부, 음주여부, 카페인섭취여부의 차이가 없다고 결론지을 수 있다.

2. 독립이 아닌 표본에서의 범주형자료분석

교차설계법으로 얻어진 자료에서 분석하고자 하는 반응변수가 약물부작용(AE: adverse event)과 같은 이항변수일 경우, 독립변수인 처리요인과 반응변수인 부작용 사이의 관계를 분석을 위해서는 범주형 자료분석법을 이용해야 한다. 이를 위한 빈도표는 다음과 같이 얻어진다.

AE		R	
		Yes	No
T	Yes	a	b
	No	c	d

이 자료는 한 피험자가 대조약과 시험약을 모두 투여받기 때문에, 쌍체자료(paired data)이며, (Yes, Yes), (Yes, No), (No, Yes), (No, No) 중 한 가지로 나타난다. 종속변수가 독립이 아닌 이항변수이므로 대조약과 시험약의 관계(association)를 검정하는 문제는 상관관계가 있는 비(correlated proportions)에 관한 검정 또는 독립이 아닌 두 집단에서 비의 동등성 검정으로 정의된다.[9,21-23)] 이때 한 피험자가 네 개의 셀(cell) 중 한 곳에만 나타나서 빈도가 서로 독립이 되도

Table 3. Goodness-of-fit test for categorical variables in B corporation's drug data

Category		Sequence		Total	P-value
Variable	Response	RT	TR		
Smoke	Yes	17	20	37	0.5202*
	No	8	5	13	
Drink	Yes	15	16	31	1.0000†
	No	10	9	19	
Caffeine	Yes	14	19	33	0.2321†
	No	11	6	17	

* Fisher's exact test. † Pearson's chi-square test.

록 빈도표를 정리하는 것이 매우 중요하며, 그렇지 못할 경우에는 여러 가지의 오류가 발생할 수 있다.[24] 귀무가설 H_0은 '주변확률밀도함수가 동일하다'는 것이며, 이것은 두 약물 또는 처리에서 나타나는 부작용의 비가 동일하다는 것과 같은 의미이다. 서로 독립이 아닌 두 표본에서 비의 동등성을 검정하는 통계방법으로는 맥네머(McNemar)의 검정이 있으며, 이는 다음과 같이 정의된다.[9,21-23]

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

B 제약회사의 약에 대한 임상시험에서 대조약과 시험약의 부작용(adverse event; AE) 비율이 동일한지를 알아보기 위해서 맥네머의 검정을 실시해보자. (AE in R, AE in T) 자료는 (Yes, Yes)가 2쌍, (Yes, No)가 1쌍, (No, Yes)가 2쌍, (No, No)가 45쌍이었다. 이때 사용되는 SAS의 PROC FREQ의 옵션 중 AGREE는 다음과 같이 사용될 수 있다.[13]

```
PROC FREQ DATA=AE;
TABLE AEinR*AEinT/AGREE;
RUN;
```

맥네머통계량 $Q = 0.3333$이고 해당하는 P-값 =0.5637이므로, 두 약의 부작용비율이 동일하다는 귀무가설을 기각하지 않는다. PROC FREQ의 CMH1옵션이 동일한 결과를 돌려주나, 만약 Table에서 subject를 생략하면, 동일한 피험자를 서로 다른 두 명으로 처리하는 오류가 발생하므로 매우 조심해야 한다. 이외에도 빈도표 중 0인 셀이 나타날 때 SAS를 이용하여 맥네머의 통계량을 사용 때에는 특별히 주의해야 한다.[1]

교차설계법을 이용하여 세 개의 약이나 처리를 비교하고자 할 때에는 다음과 같은 투여군을 사용한다:

R T S
R S T
T R S
T S R
S R T
S T R

반응변수가 약물부작용인 경우, 독립이 아닌 세 집단의 부작용의 비가 동일한지를 검정하여야 한다. 그런데, 실제 임상시험에서는 세 집단의 약물부작용 비가 동시에 동일한지 검정(simultaneous test)하기 보다는 각 쌍의 비율이 동일한지에 더 관심이 많기 때문에 일원배치 분산분석법보다는 다중비교법이 더 적절한 분석법이다.[23-24] 따라서 (AE in R, AE in S, AE in T) 자료를 (AE in R, AE in S), (AE in R, AE in T), (AE in S, AE in T)로 나누고, 각 쌍에서 맥네머 검정을 실시한다. 그런데, 세 쌍이 서로 긴밀히 연관되어 있어서 세 쌍의 맥네머 검정이 하나의 검정을 이루므로, 전체 유의수준(FWER: familywise error rate)가 FWER=0.05로 유지될 수 있도록 수정이 필요하다. 바꾸어 말하면, 각 쌍에 대한 유의수준을 0.05로 잡을 경우, 적어도 한 쌍에서 제1종의 오류가 발생할 확률은 $1-(1-0.05)^2 = 0.1426$으로 0.05보다는 훨씬 커진다. 따라서 본 연구에서는 가장 단순한 본페로니 수정[2]을 사용하며, 이때 각 쌍의 검정에서는 유의수준을 $\alpha = 0.05/3$로 둔다.

SAS에는 이와 관련된 프로시져가 따로 없기 때문에, 세 쌍의 자료에 대하여 맥네머검정을 한 후, 이를 유의수준 $\alpha = 0.05/3$과 비교하여 평균차이가 있는 쌍을 찾아내는 것이 좋다. 이와 관련된 SAS 프로시져는 다음과 같다.[13]

```
PROC FREQ DATA=AE;
TABLE AEinR*AEinT/AGREE;
TABLE AEinR*AEinS/AGREE;
```

```
TABLE AEinT*AEinS/AGREE;
RUN;
```

A제약회사 약에 대한 임상시험에서 1명의 피험자가 대조약, 공복 후 시험약, 고지방식 후 시험약 모두에 대해서 부작용을 나타냈으며, 2명의 피험자가 대조약에 대해서만 부작용을 나타냈고, 1명의 피험자가 고지방식 후 시험약에 대해서만 부작용을 나타냈다. 세 처리의 부작용비가 동일한지를 살펴보기 위하여 다중 맥네머검정을 실시해 보자. 이때 전체 전체유의수준이 0.05이고 3번의 쌍체비교를 하므로, 각 쌍의 비교에서의 유의수준은 0.05/3이 된다. (대조약, 공복 후 시험약)에 대한 맥네머통계량 Q=2이고 해당하는 P-값=0.1573이므로, 두 약의 부작용비율이 동일하다는 귀무가설을 기각하지 않는다. (대조약, 고지방식이 후 시험약)에 대한 맥네머통계량 Q=0.3333이고 해당하는 P-값=0.5637이므로, 두 약의 부작용비율이 동일하다는 귀무가설을 기각하지 않는다. (공복 후 시험약, 고지방식이 후 시험약)에 대한 맥네머통계량 Q=1이고 해당하는 P-값=0.3173이므로, 두 약의 부작용비율이 동일하다는 귀무가설을 기각하지 않는다.

결 론

임상시험 자료분석담당자는 임상시험 완료와 함께 정해진 시한까지 많은 자료를 한꺼번에 처리하고 보고서를 작성해야 하기 때문에 미리 분석방법을 준비해두고 있어야 하지만, 관련된 기본 통계개념의 양이 방대하기 때문에 이는 쉬운 일이 아니다. 그러나 단순히 통계방법을 이용하기 위한 조건과 결과만을 외우기 보다는 자료의 특성, 실험의 목적, 검정에 대한 논리적인 이해를 해두는 것이 가장 좋다. 본 논문은 이러한 어려움을 덜기 위하여, 임상시험에서 자주 사용되는 주요 통계개념을 수학적 논리에 따라서 정리하여 보았다. 특히, 자료의 특성이 연속인지 이산인지, 각각의 자료가 독립인지 아닌지에 따라서 적용가능한 통계방법이 달라지기 때문에 척도와 확률변수에 근거한 통례량의 분류를 이해해두는 것이 매우 유익하다.

임상시험에서는 한 피험자에게서 여러 번 자료가 측정되어 관측치들은 독립이 아닐 수 있기 때문에 특별한 주의가 필요하다. 또한 독립이 아닌 표본에서 반응변수가 연속이 아닌 경우에는 모수적방법을 쓸 것인지 아니면 비모수적 방법을 쓸 것인지 결정해야 한다. 그러나 이산인 경우이더라도 특별히 이항인 경우에는 비율검정이 더 적절할 수 있음을 잊지 말아야 한다.

임상시험과 관련된 통계방법들이 방대하기 때문에 본 논문의 제한된 지면에서 가능한 모든 통계방법과 해당 SAS 프로그램을 모두 고찰하기 어렵다. 그러나 해당하는 방법에 대한 통계용어와 SAS 프로시져의 이름이나 옵션에 대한 더 자세한 정보는 해당하는 참고문헌에서 찾아볼 수 있다.

참고문헌

1. Chen HL, Yang A. Common Pitfalls in SAS Statistical Analysis Macros in a Mass Production Environment. NESUG 2007; Statsitics and Data Analysis, http://www.nesug.org/proceedings/nesug07/sa/sa05.pdf [Online] (last visited on 8 May 2012).
2. Montgomery DC. Design and Analysis of Experiments. 8th ed, John Wiley & Sons, 2012.
3. Meyers LS, Gamst G, Guarino AJ. Applied multivariate research: Design and interpretation. Sage publications, 2006.

4. McClave JT, Sincich T. Statistics. 11th ed, Pearson Prentice Hall, 2009.
5. Jamieson S. Likert scales: How to (ab)use them. *Med Educ*, 2004;38:1212-1218.
6. University of Northern Iowa (UNI). SPSS techniques Series: Statistics on Likert Scale Surveys. http://www.uni.edu/its/support/article/604 [Online] (last visited on 31 May 2012).
7. Allen IE, Seaman CA. Likert Scales and Data Analyses. *Qual Prog*, 2007;40:64-65.
8. Mood AM, Graybill FA, Boes DC. Introduction to the theory of Statistics. 3rd ed, McGraw-Hill Higher Education, 1974.
9. Rosner B. Fundamentals of Biostatistics. 7th ed, Cengage Learning, 2000;666-673.
10. Simmons B, Bland MJ, Wojciechowski B. AP Statistics. Kaplan, Inc., 2010.
11. Sternstein M. AP Statistics. 5th ed, Barron's Educational Series Inc., 2010.
12. Patterson S, Jones B. Bioequivalence and Statistics in Clinical Pharmacology. Chapman and Hall/CRC, 2005;39-77.
13. SAS Institute Inc. SAS/STAT User's Guide. 1990.
14. Pinheiro JC, Bates DM. Mixed-Effects Models in S and S-PLUS. Springer, 2004.
15. 박성현. 현대실험계획법. 민영사, 2006.
16. Feng WW, Ding D. SAS@ application in 2×2 crossover clinical trial. http://www.lexjansen.com/pharmasug/2004/statisticspharmacokinetics/sp02.pdf [Online] (last visited on 8 May 2012).
17. Simpson PM, Hamer RM, Lensing S. Crossover studies off your list. http://www2.sas.com/proceedings/sugi24/Posters/p221-24.pdf [Online] (last visited on 8 May 2012).
18. Chinchilli VM, Esinhart JD. Design and analysis of intra-subject variability in cross-over experiments. *Stat Med*, 1996;15(15):1619-1634.
19. Brunelle,R. Review various methods to perform the analysis of a 2 treatment, 2 period crossover study. http://www.math.iupui.edu/~indyasa/crosover.pdf [Online] (last visited on 29 May 2012).
20. Dallal GE. The computer-aided analysis of crossover studies. http://www.jerrydallal.com/LHSP/crossovr.htm [Online] (last visited on 29 May 2012).
21. Nagelkerke NJD, Hart AAM, Oosting J. The two period binary response cross-over trial. *Biomed J*, 1986;28(7):863-869.
22. Consonni, G, Rocca, LL. Tests Based on Intrinsic Priors for the Equality of Two Correlated Proportions. *JASA*, 2008;103(483):1260-1269.
23. Westfall PH, Troendle JF, Pennello G. Multiple McNemar test. *Biometrics*, 2010;66(4):1185-1191.
24. Choi K. Common misapplications of independent sample analyses in SAS to correlated adverse events in crossover designs. Submitted to *Quantitative Bio-science*, 2012.