Ⅰ. 2×2 분할표의 표본분포

2×2 분할표의 표본분포란 2×2 분할표를 구성하는 관찰치들을 생성하는 확률분포를 의미한다. 연구설계를 비타민 C와 감기자료처럼 미리 대조군과 처리군을 정해놓고 일정기간 동안 대상자를 추적(follow-up)하는 추적조사의 형태로 할 수도 있고, 폐암의 경우처럼 발병률이 낮은 경우에는 폐암 환자군과 정상군을 미리 정해놓고 흡연률을 조사하는 사례-대조(case-control) 연구도 있다. 또한 현황조사(cross-sectional study)의 경우는 전체 표본크기만 정해 놓기 때문에 2×2 분할표의 어느 주변합(marginal total)도 미리 결정되지는 않는다.

1. 이항분포의 곱의 분포

비타민 C를 투약하지 않은 대조군 140명과 비타민 C를 투약하는 처리군 139명이 결정되고 아직 감기에 대한 진단이 이루어지지 않은 단계에서는 다음의 확률변수를 정의할 수 있다.

$$\begin{aligned} X_1 &= \text{대조군 140명 중 감기 걸리는 사람의 숫자} \\ X_2 &= \text{처리군 139명 중 감기 걸리는 사람의 숫자} \end{aligned} \tag{2.1}$$

<표 1> 대조군과 처리군의 주변합이 미리 결정된 2×2 분할표

	감기 걸림	감기 안 걸림	합계
대조군	X_1	140-X_1	140
처리군	X_2	139-X_2	139
			279

대조군과 처리군 각각에서 감기 걸릴 확률을 각각 π_1과 π_2라고 하자. 그러면 X_1과 X_2는 각각 다음 식 (2.2)와 같은 이항분포를 이루게 된다.

$$\begin{aligned} X_1 &\sim B(140, \pi_1) \\ X_2 &\sim B(139, \pi_2) \end{aligned} \tag{2.2}$$

그런데 비타민 C와 감기자료의 경우에는 애초의 연구설계에서 대조군 140명과 처리군 139명은 서로 독립이 되도록 선정하였다. 따라서 대조군에 정의된 확률변수 X_1

과 처리군에 정의된 확률변수 X_2도 서로 독립을 이루게 된다.

두 확률변수 X_1과 X_2의 독립이라는 의미는 (X_1, X_2)의 결합밀도함수가 각각의 주변 밀도함수의 곱으로 표현되는 것을 뜻한다. 그러므로 식(2.2)의 X_1, X_2의 결합밀도함수도 2개의 이항분포의 곱 즉, 적이항분포로 나타나고 이를 기술하면 식 (2.3)과 같다.

$$\Pr(X_1 = x_1, X_2 = x_2) = \binom{140}{x_1}\pi_1^{140}(1-\pi_1)^{140-x_1} \times \binom{139}{x_2}\pi_2^{139}(1-\pi_2)^{140-x_2} \tag{2.3}$$

결국 <표 1>의 2×2 분할표는 행의 합이 미리 알려져 있으므로 X_1과 X_2만 결정되면 2×2 분할표가 관찰되는 셈이 된다. 이때 (X_1, X_2)의 결합확률밀도함수가 식 (2.3)의 적이항분포(product binomial distribution)가 되므로 식 (2.3)을 2×2 분할표의 표본분포(sampling distribution)라 일컫는다. 비타민 C와 감기자료는 일종의 추적조사의 형태라 할 수 있다.

일반적인 표현을 쓰면 어느 상태에 노출이 되었는지(exposed) 혹은 노출이 되지 않았는지(unexposed)의 여부가 질병 발생에 영향을 미치는가를 연구하고자 하는 연구설계를 추적조사 혹은 전향연구(前向硏究, prospective study)라 한다. 전향연구의 경우 2×2 분할표는 다음 <표 2>의 형태가 되며 n_1과 n_2는 사전에 미리 결정된다.

<표 2> 전향연구의 2×2 분할표

	질별에 걸림	질병에 안 걸림	
노출이 됨	X_1	$n_1 - X_1$	n_1
노출이 안 됨	X_2	$n_2 - X_2$	n_2
			n

암과 같은 희귀한 질병의 경우 전향연구를 통하여 암의 발병요인을 찾고자 하면 <표 2>에서 n_1과 n_2가 커야 될 뿐 아니라 추적조사 기간도 길어야 하므로 현실적으로 가능하지 않다. 이러한 경우는 질병에 걸린 사람과 걸리지 않은 사람, 즉 사례군과 대조군을 미리 결정하여 거꾸로 사례군의 발병원인을 찾아보는 이른바 후향연구(後向硏究, restropective study)로 설계할 수 있다. 따라서 후향연구의 경우는 2×2 분할표가 <표 3>과 같은 형태를 취하고 사례군과 대조군의 크기인 n_1, n_2는 사전에 결정된다.

<표 3> 후향연구의 2×2 분할표

	사례	대조
노출이 됨	X_1	X_2
노출이 안 됨	$n_1 - X_1$	$n_2 - X_2$
	n_1	n_2

후향연구에서는 설명변수와 반응변수의 시간적 순서가 바뀌어진 상태에서 거꾸로 추적을 해나가는 셈이다. 그러나 이 경우에도 오즈비(교차비)는 전향연구의 경우와 같기 때문에 오즈비를 이용하여 반응변수에 대한 설명변수의 효과를 추정할 수 있다.

2. 다항분포

다항분포는 이항분포를 확장한 것으로 확률벡터 $\tilde{X} = (X_1, \cdots, X_r)$이 n과 모수 $\tilde{p} = (p_1, \cdots, p_r),\ 0 \le p_i \le 1, \sum_{i=1}^{r} p_i = 1$을 갖는 다항분포를 이룬다는 것을 $\tilde{X} \sim M(n, \tilde{p})$로 표기하고 다음 식 (2.4)처럼 정의된다.

$$\Pr(\tilde{X} = \tilde{x}) = \binom{n}{x_1, \cdots, x_r} \prod_{i=1}^{r} P_i^{x_i} \tag{2.4}$$

$$x_i = 0, 1, \cdots, n, \sum_{i=1}^{r} X_i = n,\ 0 \le p_i \le 1, \sum_{i=1}^{r} p_i = 1$$

다항분포의 예로 공정한 주사위를 n회 던져 각 눈이 나오는 횟수를 $\tilde{X} = (X_1, \cdots, X_6)$로 정의하는 경우이다. $\tilde{X} \sim M(n, 1/6,\ 1/6, 1/6,\ 1/6, 1/6,\ 1/6)$을 따르게 된다.

식 (2.4)의 다항분포 확률밀도함수에서 $r = 2$인 경우가 바로 이항분포의 확률밀도함수임을 알 수 있다. 현황조사(cross-sectional study)의 경우는 조사비용과 주어진 精度(정도) 등의 제약으로 인하여 전체 표본 크기 n은 사전에 결정되는 경우가 대부분이다. 관심대상이 되는 요인이 A와 B로 2개가 있고 각 요인의 수준이 2개씩이라고 하자. 그러면 <표 4>와 같은 2×2 분할표를 구성한다.

<표 4> 현황조사의 2×2 분할표

A요인		B요인 b_1	B요인 b_2	
	a_1	X_1	X_{12}	X_{1+}
	a_2	X_2	X_{22}	X_{1+}
		X_{+1}	X_{+2}	n

<표 4>에서 a_1, a_2는 A요인의 각 수준을 나타내고 X_{ij}는 (i, j) 칸의 도수를 나타낸다. 그리고 X_{i+}는 두 번째 지수에 대하여 더하여진 합, 즉 $X_{i+} = \sum_j X_{ij}$를 나타내고 X_{+j}도 비슷하게 정의된다. 전체 표본에서 하나의 개체가 (i, j) 칸에 귀속될 확률을 p_{ij}로 표시하자. p_{ij}는 물론 $0 \le p_{ij} \le 1$와 $\sum_i \sum_j p_{ij} = 1$을 만족한다.

<표 4>에서 $(X_{11}, X_{12}, X_{21}, X_{22})$가 관찰되면 2×2 분할표가 생성되므로, 결국 $(X_{11}, X_{12}, X_{21}, X_{22})$의 확률분포가 앞의 2×2분할표의 표본분포가 되는 셈이다.

여기서 $(X_{11}, X_{12}, X_{21}, X_{22})$는 n과 $(p_{11}, p_{12}, p_{21}, p_{22})$를 모수로 하는 다항분포를 이루게 된다. 또한 전향연구나 후향연구와는 다르게 조사가 끝나기 전에는 어느 주변합도 결정되지 않는다.

Ⅱ. 로지스틱회귀계수 최대가능도 추정

로지스틱회귀모형에서 모수인 회귀계수를 추정하는 데 가장 널리 사용되는 방법은 최대가능도 방법이다. 제1장 4절에서 설명하였지만 가능도함수(likelihood function)는 n개 자료에 대한 결합확률함수(joint probability function)로 표시된다.

주어진 독립변수의 수준 x_i에서 종속변수 Y_i는 베르누이 확률분포에 따르기 때문에 확률함수는 다음과 같이 정의된다.

$$f(Y_i \mid x_i) = \pi(x_i)^{Y_i}[1-\pi(x_i)]^{1-Y_i}, \quad i = 1, \cdots, n \tag{1}$$

모든 자료들이 독립적일 때 결합확률함수는 개별 확률함수들의 곱으로 다음과 같이 표시된다.

$$f(Y_1, \cdots, Y_n|X) = \prod_{i=1}^{n} f(Y_i|x_i) = \prod_{i=1}^{n} \pi(x_i)^{Y_i}[1-\pi(x_i)]^{1-Y_i} \tag{2}$$

최대가능도법에서는 식(2)와 같은 결합확률함수 혹은 결합확률함수의 자연로그 변환한 것을 최대로 하는 모수의 값을 추정하게 된다. 두 가지 함수에서 같은 결과를 얻게 된다.

일반적으로 결합확률함수의 로그변환한 것을 최대화시키는 것이 결합확률함수 자체를 최대화시키는 것보다 수학적으로 편리하기 때문에 결합확률함수의 로그변환한 것을 주로 이용한다. 결합확률함수의 로그변환 식은 다음과 같이 표현된다.

$$\begin{aligned} &\log_e f(Y_1, \cdots, Y_n|X) \\ &= \log_e \prod_{i=1}^{n} \pi(x_i)^{Y_i}[1-\pi(x_i)]^{1-Y_i} \\ &= \sum_{i=1}^{n} Y_i \log_e \pi(x_i) + \sum_{i=1}^{n} (1-Y_i) \log_e [1-\pi(x_i)] \\ &= \sum_{i=1}^{n} \left[Y_i \log_e \left(\frac{\pi(x_i)}{1-\pi(x_i)} \right) \right] + \sum_{i=1}^{n} \log_e [1-\pi(x_i)] \end{aligned} \tag{3}$$

이제 식(3)을 독립변수가 하나인 단순로지스틱회귀모형에 적용하면 다음과 같다.

$$\begin{aligned} &[1-\pi(x_i)] = [1+\exp(\beta_0+\beta_1 X_i)]^{-1} \\ &\log_e \left(\frac{\pi(x_i)}{1-\pi(x_i)} \right) = \beta_0 + \beta_1 X_i \end{aligned} \tag{4}$$

식(4)를 식(3)에 대입하면 다음과 같다.

$$\log_e f(Y_1, \cdots, Y_n|X) = \log_e L(\beta_0, \beta_1) \tag{5}$$
$$= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log_e [1 + \exp(\beta_0 + \beta_1 X_i)]$$

식(5)를 최대로 하는 회귀계수 β_0와 β_1의 최대가능도 추정값을 구하기 위해서는 우선 자연대수가능도함수인 $\log_e L(\beta_0, \beta_1)$를 β_0와 β_1에 대해서 편미분하면 다음 식을 얻게 된다.

$$\frac{\partial \log_e L(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \tag{6}$$

$$\frac{\partial \log_e L(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} \frac{X_i \exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

β_0와 β_1에 대한 최대가능도 추정값 $\hat{\beta}_0$와 $\hat{\beta}_1$은 식(6)을 0으로 놓은 정규방정식의 해를 구하게 된다. 여기서 문제는 정규방정식이 모수에 대하여 선형이 아니고 비선형이기 때문에 $\hat{\beta}_0$와 $\hat{\beta}_1$를 직접적으로 구할 수 없고, Newton-Raphson 방법 혹은 피셔(Fisher)의 스코어링 방법(method of scoring) 등과 같은 반복적인(iterative) 추정 방법에 의하여 근사해를 구할 수밖에 없다. 반복적인 추정 절차는 특정한 수렴기준(convergence criteria)을 만족할 때까지 계속하게 된다.

Ⅲ. 가능도비(우도비) 검정 소개

로지스틱회귀모형의 유의성검정에 가장 널리 사용되는 가능도비검정(likelihood ratio test)에 대하여 설명한다. 모형의 유의성검정에 사용되는 첫 번째 가능도비 검정은 가능도비 카이제곱(likelihood ratio chi-square)검정이고, 두 번째 검정법은 이탈도(deviance)검정이다.

1. 가능도 카이제곱 검정

로지스틱 회귀모형에서 절편을 제외한 모든 회귀계수가 동시에 0이라는 귀무가설에 대한 가능도비 카이제곱 유의성검정은 다음과 같은 절차를 통해서 실시할 수 있다.

1) 가설설정

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1 : 최소한 하나의 $\beta_j \neq 0 \, (j = 1, 2, \cdots, k)$

귀무가설은 로지스틱회귀모형에 포함된 k개 독립변수들이 어떤 사건의 발생 여부에 전혀 정보를 제공하지 못한다는 의미이고, 대립가설은 적어도 하나 이상의 유의한 영향을 미치는 독립변수가 모형에 포함되어 있다는 의미이다.

2) 검정통계량

가능도비 카이제곱 검정통계량은 설정된 모형의 가능도와 모든 회귀계수가 0으로 제약된 모형의 가능도를 비교하는 것이다. 전자는 k개 독립변수를 포함한 로지스틱회귀모형에서 최대가능도 추정법에 의해서 구해진 가능도로 $L(\hat{\beta})$로 표시하자. 여기서 $\hat{\beta}$는 절편을 포함한 추정된 회귀계수 벡터이다.

$$L(\hat{\beta}) = \prod_{i=1}^{n} [\pi(x_i)]^{Y_i} [1 - \pi(x_i)]^{1 - Y_i} \tag{1}$$
$$= \prod_{i=1}^{n} \left(\frac{\exp(x_i' \hat{\beta})}{1 + \exp(x_i' \hat{\beta})} \right)^{Y_i} \left(\frac{1}{1 + \exp(x_i' \hat{\beta})} \right)^{1 - Y_i}$$

후자는 귀무가설이 참인 (모든 k개 회귀계수들이 동시에 0일 때) 제약된 모형에서 최대가능도 추정법에 의해서 구해진 가능도 $L(\hat{\beta_0})$로 표시하자. 제약된 모형은 절편만

포함하고 있기 때문에 가능도 $L(\hat{\beta}_0)$는 다음과 같이 구해진다.

$$L(\hat{\beta}_0) = \prod_{i=1}^{n} \left(\frac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)} \right)^{Y_i} \left(\frac{1}{1+\exp(\hat{\beta}_0)} \right)^{1-Y_i} \quad (2)$$

$$= \prod_{i=1}^{n} [\exp(\hat{\beta}_0)]^{Y_i} \left(\frac{1}{1+\exp(\hat{\beta}_0)} \right)$$

식(1)에서 가능도 $L(\hat{\beta})$는 확률의 곱으로 표시되기 때문에 0과 1 사이의 값을 갖게 된다. 만약 로지스틱회귀모형에 포함된 독립변수들의 종합적인 정보가 어떤 사건의 발생 여부에 영향을 미치는 정도가 커질수록 $L(\hat{\beta})$는 $L(\hat{\beta}_0)$에 비해서 커지게 될 것이고, 만약 모형에 포함된 독립변수들의 종합적인 정보가 어떤 사건의 발생 여부에 미치는 영향이 작을수록 두 가능도 사이의 차이는 작아지게 될 것이다. 따라서 모든 기울기 회귀계수들이 동시에 0인지 여부에 대한 검정하기 위해서는 두 가지 가능도의 크기를 비교하게 된다. 가능도비 카이제곱 검정통계량 LR은 두 가지 가능도의 비율에 자연대수를 취하여 아래와 같이 설정한다.

$$LR = -2\log_e \left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta})} \right) \quad (3)$$

$$= -2\log_e L(\hat{\beta}_0) + 2\log_e L(\hat{\beta})$$

로지스틱회귀모형에 포함된 독립변수의 종합정보가 유의할수록 가능도비 카이제곱 검정통계량 LR은 커지게 되고, 반대로 유의하지 않다면 LR은 0에 가까워지게 된다.

가능도비 카이제곱 검정통계량 LR은 귀무가설이 참일 때 자유도가 k인 카이제곱분포 χ_k^2에 따르게 된다. 현재 설정된 모형에서는 절편과 k개 회귀계수가 있기 때문에 자유도는 $(k+1)$이 되고, 절편만 포함한 제약된 모형에서는 자유는 1이 된다. 따라서 두 가지 모형에서 자유도의 차이는 k이기 때문에 카이제곱 검정통계량 LR의 자유도는 k가 된다. 자료에서 구해진 가능도비 카이제곱 검정통계량값을 LR^*로 표시하자.

3) 유의수준 설정과 임계값

유의수준이 α일 때 귀무가설의 기각여부를 결정하는 임계값은 자유도가 k인 카이제곱분포에서 상위 $100\times\alpha\%$에 해당되는 카이제곱값인 $\chi_{\alpha,k}^2$이 된다. 일반적으로 사용되는 유의수준 α의 값은 0.05, 0.01 등이다.

결론적으로 가능도비 카이제곱 검정통계량의 값 LR^*이 클수록 혹은 대응되는 P-값이 작을수록 모형에 포함된 독립변수들의 종합적인 정보가 어떤 사건이 발생될 확률을 예측할 수 있는 능력을 유의하게 향상시킬 수 있다고 판단할 수 있다.

2. 이탈도(deviance) 유의성 검정

이탈도를 이용한 검정은 설정된 로지스틱회귀모형이 완전정보를 가진 모형(포화모형)과 비교해서 유의한 차이가 있는지 여부를 평가하는 방법이다. 이탈도(deviance)는 설정된 모형과 연관된 오차를 반영하므로 종속변수에서 설명될 수 없는 변동의 유의성과 관계가 있다. 이탈도 검정은 다음 절차를 통해서 구할 수 있다.

1) 가설검정

$$H_0 : \pi(x_i) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \qquad H_1 : \pi(x_i) \neq \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$$

귀무가설은 k개 독립변수를 포함한 로지스틱회귀모형이 어떤 사건이 발생될 확률을 설명하는 데 충분한 정보를 제공한다는 의미이고, 대립가설은 현재 설정된 로지스틱회귀모형이 어떤 사건이 발생될 확률을 설명하는 데 불충분하다는 것이다.

2) 검정통계량

deviance 검정통계량은 설정된 모형의 가능도와 완전정보를 가진 모형(포화모형, 각각의 관측치에 대해 개별적인 모수를 갖는 모형)의 가능도를 비교하여 구하게 된다. 설정된 로지스틱회귀모형에서 구해진 가능도는 앞에서 $L(\hat{\beta})$로 표시하였다. 완전정보를 가진 모형은 각 관측값마다 하나의 개별적인 모수를 갖기 때문에 종속변수의 관측값을 완전하게 예측할 수 있는 모형을 의미한다. 완전정보를 가진 모형에 대한 가능도 $L(F)$는 아래와 같이 구해진다.

$$L(F) = \prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \tag{4}$$

가능도 $L(F)$를 계산하는 데 각 관측값과 관련된 하나의 모수 π_i를 사용하였고, 독립변수의 정보를 이용한 $\pi(x_i)$를 사용하지 않았다. 식(4)에서 π_i가 0 (혹은 1)일 때 Y_i도 0 (혹은 1)이 되기 때문에 가능도 $L(F)$의 값은 항상 1이 된다.

만약 설정된 로지스틱회귀모형의 예측력이 충분하다면, 가능도 $L(\hat{\beta})$는 1에 가까운 값을 갖게 되므로 완전정보를 가진 모형의 가능도 $L(F)$ 사이의 차이는 아주 작아지게 된다. 반대로 설정된 로지스틱회귀모형의 예측력이 미흡하게 되면 두 가능도의 차이는 커지게 될 것이다. deviance 검정통계량은 두 가지 가능도의 비율에 자연대수를 취하여 구해진다.

$$D = -2\log_e\left(\frac{L(\hat{\beta})}{L(F)}\right) \tag{5}$$

$$= -2\log_e L(\hat{\beta}) + 2\log_e L(F)$$

$$= -2\log_e L(\hat{\beta})$$

완전정보를 가진 모형에서 구한 가능도 $L(F)$는 항상 1이 되기 때문에 자연대수를 취한 가능도 $\log_e L(F)$는 0이 된다. 또한 가능도 $L(\hat{\beta})$는 확률의 곱으로 표시되므로 0과 1 사이의 범위를 가지기 때문에, 자연대수를 취한 가능도 $\log_e L(\hat{\beta})$의 범위는 $-\infty$과 0이 된다. 따라서 deviance 검정통계량 D의 범위는 0과 ∞사이가 된다.

deviance 검정통계량 D의 값이 작을수록 현재 설정된 로지스틱회귀모형의 적합도는 좋아지게 된다. deviance 검정통계량 D는 귀무가설이 참일 때 자유도가 $n-k-1$인 카이제곱분포 χ^2_{n-k-1}에 따르게 된다. 자유도를 살펴보면 완전정보를 가진 모형에서는 각 관측값마다 모수가 하나씩 있기 때문에 자유도는 n이 되고, 설정된 모형의 자유도는 모수 수인 $k+1$이 된다. 두 모형과 연관된 자유도의 차이는 $n-(k+1)$이기 때문에 deviance 검정통계량 D의 자유도는 $n-k-1$이다.

자료에서 구해진 deviance 검정통계량의 값을 D^*로 표시하자. deviance 검정통계량의 P-값은 자유도가 $n-k-1$인 카이제곱분포에서 검정통계량의 값 D^*보다 클 확률 $P(\chi^2_{n-k-1} > D^*)$을 의미하고, P-값이 클수록 귀무가설을 강하게 채택할 수 있는 근거가 된다.

결론적으로 deviance 검정통계량값 D^*가 작을수록 혹은 대응되는 P-값이 클수록 설정된 로지스틱회귀모형이 완전정보를 가진 모형과 비교해서 유의한 차이가 없다고 판단할 수 있을 것이다. McCullagh 와 Nelder(1989, pp. 120-122)에 의하면 표본 자료에서 독립변수들의 수준이 한 번씩만 관측이 되는 경우(독립변수가 연속형인 경우)에 deviance 검정방법은 모형의 적합도를 측정하는 척도로 적당하지 않다고 주장하였다.