1.1. 범주형 자료분석이란

자연 과학, 사회 과학은 물론 의학관련 분야에까지 범주형 자료 분석은 널리 활용되고 있다. 기업의 부실 여부 판단, 새로운 의학 치료법에 대한 가치 평가, 사람들의 의견에 영향을 주는 요인들에 대한 평가 등 범주형 자료에 대한 분석의 필요성은 증가하고 있다. 그러나 범주형 자료분석은 다른 일반적인 분석과는 달리 모형이 다소 복잡하고 결과 해석이 용이하지 않아

1.1.1. 변수(variable)와 자료(data)

자료 수집의 대상이 되는 모집단의 특성을 변수(variable)라 하고 변수의 측정치를 관측치(observation)라 하며, 자료(data)는 이런 변수와 관측치로 이루어진 숫자 모임이다. 직장인의 식습관 중 아침 식사 여부, 점심 메뉴, 음주 횟수, 비만도에 관심이 있다면 이들 각각을 변수라 한다. 변수라는 의미는 각 측정치가 각 직장인마다 변하기 때문이다. 측정된 각 직장인들의 변수 측정치를 관측치라 한다. 각 변수를 열, 각 학생들의 측정치를 행으로 하여 만들어진 행렬을 자료 행렬(data matrix)이라 하고 이를 자료라 한다.
변수와 자료가 구별되기도 하지만 때로는 변수를 자료와 혼용하여 사용하기도 하는데 본 강의에서도 자료와 변수를 굳이 구별하지는 않겠다.

1.1.2. 자료 종류

변수의 형태에 따라 자료 분석 방법이 결정되므로 자료에 적합한 분석방법을 찾으려면 측정할 변수의 형태를 구별할 수 있어야 한다. 자료 분석에서의 변수 분류 방법은 측정할 수 있거나 셀 수 있는 측정형(measurable 혹은 numerical) 변수와 개체나 집단을 분류하는데 사용되는 분류형(categorical) 변수로 나누어진다. 측정형 변수는 양적(quantitative) 변수, 개체의 특성에 따라 집단을 분류하는 분류형 변수는 질적(qualitative) 변수라고 분류되기도 한다.

측정형(metric) 양적(quantitative)	측정 가능하거나 셀 수 있는 것에 대한 자료(변수)로 크기를 가지고 있다. (예) 키, 몸무게, 매출액, 나이, 교통량, 물가지수
비측정형(non-metric) 분류형(classified) **범주형(categorical)** 질적(qualitative)	개체를 분류하는데 사용되는 자료(변수) ✓순서형(ordinal): 순서가 있는 분류 (예) 학년, 소득 수준(상, 중, 하), 병의 단계 ✓명목형(nominal): 단지 분류만 (예) 성별, 거주지, 취업 유무, 병의 종류

ⓘ수리 통계에서는 변수를 나누는 경우 이산형(discrete), 연속형(continuous)으로 나누는데 이는 위의 자료 분류와는 다름에 유의하기 바란다.

ⓘ측정형 변수를 interval, ratio(Steven, 1951)로 나누기도 하지만 본 강의에서는 구별하지 않기로 한다. interval 은 크기를 가지고 있고 크기의 차이에 의해 상대적 비교가 가능한 경우(예: 온도, 지능 지수)이다. Ratio 는 interval 자료의 성질에 0 을 가지므로 값들의 비가 의미를 갖는 경우로 대부분의 측정형 변수(예:키, 몸무게, 소득)이다.

반응 변수와 설명 변수

통계 모형(인과 관계)에서는 영향을 주는 변수와 그 변수들에 의해 영향을 받는 변수가 존재한다. 영향을 받는 변수를 종속변수(dependent), 반응변수(response)라 하고 모형의 왼쪽에 위치하므로 Y 변수라고도 한다. 영향을 주는 변수들 독립변수(independent), 설명변수(explanatory)라 하며 모형의 오른쪽에 위치하며 X 변수라 한다.

일반적 통계 모형

$$Y = f(X_1, X_2, ..., X_p) + e$$

? 각 분석 방법에서 모형은...

범주형 자료분석

범주형 자료분석이란 반응 변수가 **하나이고 범주형인 통계 (인과)모형**을 분석하여 1)모형의 유의성과 2)설명변수의 유의성을 알아보는 방법이다.

○ **설명 변수가 하나이고 범주형인 경우**

교차분석(cross-tabulation) 혹은 분할표 분석(contingency table) 이용하기

○ **설명변수가 2 개 이상이고 모두 측정형 변수 혹은 측정형 변수와 분류형 변수 혼합**

Logistic Regression Model(로지스틱 회귀 모형) 혹은 Logit Model (로짓 모형): 반응 변수가 2 분류(binary, dichotomous)이거나 수준이 3 개 이상인 경우는 ordinal(순서형) 분류형 변수일 경우 사용된다.

ⓘ반응 변수의 수준이 3 개 이상이고 순서형인 경우에만 Logit 모형이라고 부르기도 한다.

○ **설명변수가 2 개 이상이고 모두 분류형 변수인 경우**

Log-Linear Model(로그 선형 모형)

ⓘ변수가 하나인 일변량 분석에서는 범주형 자료에 대한 분석 방법은 무엇인가? 숫자적 요약으로는 비율을 구하는 것이고 그래프적 요약으로는 파이 차트, 바 차트 등 다양한 그래프에 각 항목(수준)의 비율을 나타내면 된다.

ⓘ반응변수가 측정형이고 두 개 이상인 경우이고 독립 변수가 모두 범주형인 경우는 다변량 분산분석(Multivariate ANOVA), 독립 변수가 측정형, 범주형이 함께 있는 경우는 연립 방정식 회귀 모형 방법을 사용하면 된다.

1.2. 표본 추출 모형

범주형 자료 분석은 통계 모형에 의한 분석 방법이므로 반응 변수에 대한 확률적 분포에 대한 가정이 필요하다. 회귀분석이나 분산분석에서 종속변수(반응변수)에 대한 가정은 정규분포다. ➔ $Y = b_0 + b_1 x_1 + b_2 x_2 + ..b_p x_p + e$

범주형 자료분석에서는 반응변수에 대한 확률 모형으로 이항분포(binomial dist.)와 포아송 분포(Poisson dist.), 다항분포 (multinomial dist.) 가 중심 역할을 한다.

1.2.1. Poisson sampling

일정한 시간에 어떤 사건(event)이 발생하는 횟수에 대한 분포로 사용된다. 예를 들면 한남대학교 앞 도로 일주일 교통 사건 발생 건수, 하루 병원을 찾는 환자들의 수에 대한 분포가 Poisson 분포를 따른다.

Poisson 분포의 확률밀도 함수는

$$f(x) = \frac{\lambda^x \exp^{-\lambda}}{x!}, \quad x = 0,1,...$$

Poisson 분포의 평균 λ 이고 표준편차는 $\sqrt{\lambda}$ 이다.

분포의 특징

○ 한남대 도서관 분실 사고 발생 횟수가 포아송 분포를 따르고 평균 $\lambda = 2$ 회이면 3 주동안 사고 발생 횟수는 평균이 6 포아송 분포를 따른다. 그러면 3 주동안 사고가 전혀 발생하지 않을 확률은?

$$P(X = 0) = \frac{6^0 e^{-6}}{0!} = 0.002$$

○ 변수 $X_1, X_2, ..., X_n$ 이 서로 독립이고 각각 포아송 분포(λ_i)를 따르면 변수의 합 $\sum_{i=1}^{n} X_i \sim Poisson(\lambda_1 + \lambda_2 + ... + \lambda_n)$ 따른다.

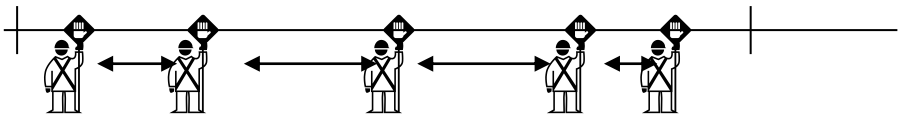
○ 포아송 분포는 평균 값(λ)이 커지면 표준편차($\sqrt{\lambda}$)도 증가한다. 일반적으로 평균이 커짐에 따라 (교통량 발생 평균이 커짐에 따라) 표준 편차가 증가하는 자료에 대한 모형화에 유용하다.

다른 분포와의 관계

○ Poisson 분포는 $n \to \infty$ 인 경우 정규분포($\mu = \lambda, \sigma = \sqrt{\lambda}$)에 근사한다.

○ Poisson 분포를 따르는 사건이 발생하는 사이 시간은 지수 분포(exponential)를 따른다.

지수 분포 확률밀도 함수는 $f(x) = \frac{1}{\beta} e^{-x/\beta}$ 이다. 평균과 표준편차는 모두 β 이다.



○ $X \sim \exp(\beta)$ 이고 $Y \sim Poisson(x/\beta)$ 인 경우 $\Pr(X \le x) = \Pr(Y \ge 1)$ 이다.

1.2.2. Binomial sampling

성공/실패 두 가지 결과만 발행하는 시행을 Bernoulli trial 이라 한다. 즉 동전을 던지는 실험에서 앞면/뒷면, 제품 검사에서 불량/정품이 나타나는 시행을 말한다. 베르누이 시행에서 성공(success) 확률이 p 인 경우 분포 함수는 다음과 같다.

$$f(x) = p^x (1-p)^{1-x}, \quad x = 0,1$$

이런 베르누이 시행을 n 번 하는 경우 성공 횟수 X 에 대한 분포가 이항 분포이다.

$$f(x) = P(X = x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0,1,\text{K}\ ,n$$

이항 분포의 평균은 p 이고 표준편차는 $\sqrt{npq}$ 이다.

다른 분포의 관계

○ $n \to \infty, p \to 0$ 이면 Poisson 분포($\lambda = np$)에 근사한다.

○ $n \to \infty$ 이면 정규분포($\mu = np, \sigma = \sqrt{npq}$)에 근사한다. (Normal Approximation to Binomial)

1.2.3. Multinomial sampling

시행의 결과가 2 개 이상인 경우, 예를 회사의 면접을 본 경우 합격, 불합격 뿐 아니라 보류라는 결과가 있는 경우 사건 발생 수는 다항 분포(multinomial) 분포를 따른다.

$$f(x_1, x_2, \mathrm{K}\ , x_n) = \frac{m!}{x_1! x_2! \mathrm{K}\ x_n!} p_1^{x1} p_2^{x2}\ \mathrm{K}\ p_n^{x_n}\ ,\ \sum x_i = m, \sum p_i = 1$$

1.2.4. 비율에 대한 추론 (일변량)

범주형 반응 변수에 대한 확률 모형으로 포아송 분포, 이항분포, 다항 분포를 고려하였으나 추정치, 표준 오차 추론에 있어서는 동일한 결과를 가지며 로지스틱 모형(logistic)이나 로그 선형 모형(log-linear)의 모수에 대한 추론에서도 동일한 결과를 가지므로 구별하여 사용하지는 않을 것이다.

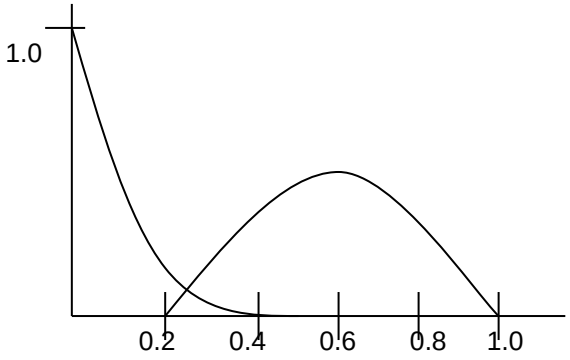
여기서는 이항 분포 모수 p 에 대한 maximum Likelihood Estimator(MLE: 최대 우도 추정량)을 구하는 방법을 살펴보자.

만약 제품의 불량률 p 를 모르는 제품에서 10 개를 임의로 추출하여 불량 여부를 조사하였더니 6 개였다. 이때 우도 함수(likelihood function)는

$$l(p \mid x) = f(x = 6 \mid p) = \binom{10}{6} p^6 (1-p)^4$$

우도 함수는? 표본으로부터 얻은 자료가 발생할 가능성이다.

불량 개수가 0 이라면 우도 함수는 $l(p \mid x) = f(x = 0 \mid p) = \binom{10}{0} p^0 (1-p)^{10}$



그러므로 0 개인 경우는 $\hat{p} = 0$, 6 개인 경우는 $\hat{p} = 0.6$. 일반화 하여 보자.

$$l(p \mid x) = f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$$

? 이항 분포(총 시행 회수: n)로부터 성공 개수가 X개 관측되었다면 이항 분포의 모수 p에 대한 MLE?