카이제곱검정

정의

- 분포의 유도

1. X_1 ~ b(n,p_1) 일 때, 확률변수 Y = $\frac{x_1-np_1}{\sqrt{np_1(1-p_1)}} \rightarrow N(0,1)$ 이다.

1) 이제, 같이 X를 구성하면서 X_1의 여집합인 X_2를 생각하자.
X_2 = n - X_1 이고 $p_2 = 1 - p_1$ 이다.

2) 이제, $Y^2 = N(0,1)^2 \sim x^2(a)$ 이고, $Q_1 = Y^2$ 일때 Q_1은

(1) $Q_1 = \frac{(x_1-np_1)^2}{np_1(1-p_1)} = \frac{(x_1-np_1)^2}{np_1} + \frac{(x_1-np_1)^2}{(1-p_1)} = \frac{(x_1-np_1)^2}{np_1} + \frac{(x_1-np_1)^2}{np_2}$

(2) 이를 일반화하면

- $Q_{k-1} = \sum_{i=1}^{k} \frac{(x_i-np_i)^2}{np_i}$ 이고, 이는 $x^2(1)$를 따른다.

- 이 때, x_i = n- $\sum_{i=1}^{k} X_i$, p_k = 1- $\sum_{i=1}^{k} p_i$ 이다.

정의

- 분포의 유도

2) 이제, $p_1 \cdots p_n$을 어떤 확률실험을 n회 반복했을 때 각 X_i가 A_i에 속할 확률이라 가정

(1) 이 때,

- A_i는 이 확률실험에서 쓰이는 표본들의 부분 범위들을 뜻하고
- X_i는 실험의 실험값들이 각 A_i에 속하는 빈도수를 의미한다.

(2) 이 때 p_{i0}를 가설 확률이라 하고, $H_0 : p_1 = p_{10,} \cdots p_n = p_{n0}\ VS\ H_1 : p_i \neq p_{i0,}$ 라고 한다면

- H_0가 참일 때 확률변수 $Q_{k-1} = \sum_{i=1}^{k} \frac{(x_i - np_{i0})^2}{np_{i0}} \sim x^2(k-1)$ 이다.
- H_0가 참일 경우, $x_i = np_{i0}$ 이므로, 이 값은 작아야 하며, 이를 이용하여 검증을 할 수 있다.

	A_1	$\cdots$	A_k
빈도	X_i	$\cdots$	X_k

정의

- 최소카이제곱추정량

1. H_0하에서 p_i를 완전히 알 수 없는 경우가 존재한다. 이 때, 해당 p_i에 속하는 Y_i를 모수를 알 수 없는 pdf를 따르는 분포의 확률변수라 가정한다.
2. 즉, Y가 μ와 σ를 알 수 없는 정규분포 N(μ, σ) 에서 추출되었다고 하자.

1) 이 때, $p_i = \int_{A_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) dy$ 이고

2) 마찬가지로 X_i는 실험의 실험값들이 각 A_i에 속하는 빈도수를 의미한다고 할 때
$Q_{k-1}=\sum_{i=1}^{k} \frac{(x_i-np_{i0})^2}{np_{i0}}$

(1) 이 경우, μ, σ가 미지이므로 계산할 수 없다. 대신, 관찰된 $X_1 = x_1 \cdots X_k = x_k$
를 중심으로 Q_{k-1}을 최소화하는 μ, σ를 선택한다.

(2) 구체적으론 E(Q_{k-1}) = 0 으로 만드는 최대우도추정값 을 사용한다.

(3) 또한, 이 때 Q_{k-1}~ $x^2(k-3)$을 따른다.

정의

- 동질성에 대한 카이제곱 검정

1. 모수가 각각 n_j, p_{1j}, $p_{2j} \cdots p_{kj}$(j=1,2) 인 2개의 다항분포가 있다면

1) X_{1j}, $X_{2j} \cdots X_{kj}$이 그 대응 빈도이다. 이 때, 다항분포의 가법분포는

$$(1)\sum_{j=1}^{2}\sum_{i=1}^{k}\frac{(x_{ij}-n_j p_{ij})^2}{n_j p_{ij}} \sim x^2(2k-2)$$

2) 이 때, p_{ij}는 직접적으로 추정할 수 없으며, 최소카이제곱추정에 따라 MLE를 각 분포의 대리 모수로 활용하면, 그 때의 추정값은

(1) $p_{i1} = p_{i2}$ 라는 가설에 대하여 $p_{ij} = \frac{(X_{i1}+X_{i2})}{n_1+n_2}$ 이다.

(2) 이를 반영하여 가법분포를 다시 쓰면

(3) $\sum_{j=1}^{2}\sum_{i=1}^{k}\frac{(x_{ij}-n_j\frac{(X_{i1}+X_{i2})}{n_1+n_2})^2}{n_j\frac{(X_{i1}+X_{i2})}{n_1+n_2}}$ 이다.

	A_1	$\cdots$	A_k
B_1	X_{11}	$\cdots$	X_{1k}
$\cdots$	$\cdots$	$\cdots$	$\cdots$
B_k	X_{k1}	$\cdots$	X_{kk}

정의

- 분할표

1. 확률실험결과를 두 속성으로 분류한다고 하자(예를들면, 눈과 머리색)

1) $p_{ij} = p(A_j \cap B_j)$ 이고, 이 때 실험이 n회 반복된다고 쳤을 때 p_{ij}에 해당하는 빈도를 X_{ij} 라고 하자.

2) K=i·j 개만큼 존재하므로, 이 때 확률변수 Q_{ab-1}은

$$(1)\ Q_{ab-1} = \sum_{j=1}^{a}\sum_{i=1}^{b}\frac{(x_{ij}-n_j p_{ij})^2}{n_j p_{ij}} \sim x^2(ab-1)$$

(2) 이 때, 속성 A와 B의 독립성을 검증할 때, $H_0 : p(A_j \cap B_j) = p(A_j)\cdot p(B_j)$ 이므로

	밤색	…	초록색
눈	X_{11}	…	X_{1k}
머리	X_{k1}	…	X_{kk}

정의

- 분할표

2) 이 때, 속성 A와 B의 독립성을 검증할 때, $H_0 : p(A_j \cap B_j) = p(A_j) \cdot p(B_j)$ 이므로

(1) $p(A_i) = \sum_{j=1}^{b} p_{ij}$, $p(B_i) = \sum_{i=1}^{b} p_{ij}$

(2) 위 예시에서, i=머리색, j=눈색 이라면

- $p(A_{i=\text{밤색}}) = p_{\text{밤색,초록색}} + p_{\text{밤색,금색}} + \cdots$
- $p(A_{i=\text{금색}}) = p_{\text{금색,초록색}} + p_{\text{금색,금색}} + \cdots$

(3) 위를 일반화 하면

- $Q_{ab-1} = \sum_{j=1}^{a} \sum_{i=1}^{b} \frac{(x_{ij} - n_j p(A_j) \cdot p(B_j)^2}{n_j p(A_j) \cdot p(B_j)}$ 이다.