독립성 검정

정의

- 개요

1. X와 Y가 평균 μ_1, μ_2, 분산이 σ_1^2, σ_2^2이고 상관계수가 ρ인 이변량 정규분포를 따른다고 하자.
2. H_0 : 가설 $\rho = 0$ VS H_1 : $\rho \neq 0$를 검정하기 위한 우도비 검정을 정의하면

1) $\mathrm{L}(\theta; x_n, y_n) = f(x_1, y_1; \theta) f(x_2, y_2; \theta) \cdots f(x_n, y_n; \theta)$ 이고(단, $f(x_n, y_n; \theta)$은 이변량 정규분포의 PDF)

2) $\Lambda = \frac{\mathrm{L}(\rho=0; x_n, y_n)}{\mathrm{L}(\rho\neq 0; x_n, y_n)} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = R$이다.

3. 따라서 $\mathrm{P}[|\mathrm{R}| \geq \mathrm{c}]$에 대해서, 이를 알려진 분포로 변환하는 $\mathrm{g}(|\mathrm{R}|)$ 이 요구된다.

1) Y에서 iid인 확률표본 $Y_1 \cdots Y_n$을 선출하고 ,
이 때 X의 확률표본 $X_1 \cdots X_n$이 실현값 $x_1 \cdots x_n$을 각각 갖는다고 하자.

2) 이 가정하에서 $Y_i|(X_1 = x_1 \cdots X_n = x_n)$의 pdf를 구하면

(1) $\rho = 0$ 라는 가정 하에, Y_i와 X_i는 두 확률변수의 PDF의 단순결합으로 분리 가능하므로

(2) 이 조건부 PDF는 단지 Y_i의 PDF로만 도출된다.

정의

- 개요

3. 한편 Y_i도 iid임을 가정했으므로, 이 결합 pdf는 $\left[\frac{1}{\sqrt{2\pi}\sigma_y}\right]^{\frac{n}{2}} \exp\left(-\frac{\sum(y_i-\mu_y)^2}{2\sigma_y{}^2}\right)$ 이다.

1) 즉, $\rho = 0$일 경우 $Y_i|(X_1 = x_1 \cdots X_n = x_n) = Y_i$ 이다.
2) 이 때, $(X_1 = x_1 \cdots X_n = x_n)$이 주어졌을 때의 조건부 상관계수 R_c는

$$(1) R_c = \frac{\sum(X_i-\bar{X})\{[Y_i|(X_1=x_1\cdots X_n=x_n)]-\bar{Y}\}}{\sqrt{\sum(X_i-\bar{X})^2 \sum([Y_i|(X_1=x_1\cdots X_n=x_n)]-\bar{Y})^2}}$$ 이고,

$$(2) \rho = 0 \text{ 일 경우 } \frac{\sum(X_i-\bar{X})\{[Y_i|(X_1=x_1\cdots X_n=x_n)]-\bar{Y}\}}{\sqrt{\sum(X_i-\bar{X})^2 \sum([Y_i|(X_1=x_1\cdots X_n=x_n)]-\bar{Y})^2}} = \frac{\sum(X_i-\bar{x})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{x})^2 \sum(Y_i-\bar{Y})^2}}$$ 이다.

정의

- 개요

4. 이제, $Y_i|(X_1 = x_1 \cdots X_n = x_n)$ = Y_i를 통해 Y_i는 X_i와 확률적으로 무관한 독립 분포가 되었으며

1) $g(|R|) = \frac{R_C\sqrt{\sum(Y_i-\overline{Y})^2}}{\sqrt{\sum(x_i-\overline{x})^2}} = \frac{\sum(x_i-\overline{x})(Y_i-\overline{Y})}{\sum(X_i-\overline{x})^2} = \frac{\sum(x_i-\overline{x})Y_i}{\sum(X_i-\overline{x})^2}$ 를 도출할 수 있다.

2) 이 때, $\frac{\sum(x_i-\overline{x})Y_i}{\sum(X_i-\overline{x})^2}$ 는 회귀분석에서 파라미터 β의 불편추정량 $\hat{\beta}$와 같으며,

(1) 회귀분석에서 전개한 논리를 가져오면

(2) $\dfrac{R_C\sqrt{\sum(Y_i-\overline{Y})^2} \Big/ \sqrt{\sum(x_i-\overline{x})^2}}{R_C\sqrt{\sum\{Y_i-\overline{Y}-\left[R_C\sqrt{\sum(Y_j-\overline{Y})^2} \Big/ \sqrt{\sum(x_j-\overline{x})^2}\right](x_i-\overline{x})\}^2} \Big/ (n-2)\sqrt{\sum(x_j-\overline{x})^2}}$ ~ T(n-2) 이다.

정의

- 개요

4. 이제, $Y_i|(X_1 = x_1 \cdots X_n = x_n)$ = Y_i를 통해 Y_i는 X_i와 확률적으로 무관한 독립 분포가 되었으며

1) $g(|R|) = \frac{R_C\sqrt{\sum(Y_i-\overline{Y})^2}}{\sqrt{\sum(x_i-\overline{x})^2}} = \frac{\sum(x_i-\overline{x})(Y_i-\overline{Y})}{\sum(X_i-\overline{x})^2} = \frac{\sum(x_i-\overline{x})Y_i}{\sum(X_i-\overline{x})^2}$ 를 도출할 수 있다.

2) 이 때, $\frac{\sum(x_i-\overline{x})Y_i}{\sum(X_i-\overline{x})^2}$ 는 회귀분석에서 파라미터 β의 불편추정량 $\hat{\beta}$와 같으며,

(1) 회귀분석에서 전개한 논리를 가져오면

$$(2)\ \frac{R_C\sqrt{\sum(Y_i-\overline{Y})^2}\Big/\sqrt{\sum(x_i-\overline{x})^2}}{R_C\sqrt{\sum\left\{Y_i-\overline{Y}-\left[R_C\sqrt{\sum(Y_j-\overline{Y})^2}\Big/\sqrt{\sum(x_j-\overline{x})^2}\right](x_i-\overline{x})\right\}^2}\Big/(n-2)\sqrt{\sum(x_j-\overline{x})^2}} = \frac{R_C\sqrt{n-2}}{\sqrt{1-R_C{}^2}}$$

~ T(n-2) 이다.