확률과 통계 3차 과제물

소프트웨어학과 2021041017 김규현

[문제 1]

두 확률변수 X와 Y의 결합확률분포가 다음과 같을 때, 다음 물음에 답하라.

$$f(x,y)=\begin{cases}\frac{6}{5}(x^2+y), & 0<x<1,\ 0<y<1\\ 0 \quad , & \text{그 외에서}\end{cases}$$

(1) 확률변수 X의 주변확률밀도함수를 구하라.

(2) 확률 $P(0<X<0.5,\ 0<Y<0.5)$을 구하라.

(3) 두 확률변수의 상관계수를 구하라.

(4) 두 확률변수는 독립인가? 그 이유는?

Sol)

(1) $f_X(x)=\int_0^1 f(x,y)\,dy=\int_0^1 \frac{6}{5}(x^2+y)\,dy$

$=\frac{6}{5}\left[x^2y+\frac{y^2}{2}\right]_0^1=\frac{6}{5}(x^2+0.5)=1.2x^2+0.6$

$\therefore f_X(x)=1.2x^2+0.6 \quad (0<x<1)$

(2) $P(0<X<0.5,\ 0<Y<0.5)$

$P=\int_0^{0.5}\int_0^{0.5} f(x,y)\,dx\,dy=\int_0^{0.5}\frac{6}{5}\left[\frac{x^3}{3}+yx\right]_0^{0.5}dy$

$=\int_0^{0.5}\frac{6}{5}\left[\frac{(0.5)^3}{3}+y(0.5)\right]dy=\int_0^{0.5}\frac{6}{5}\left[\frac{0.125}{3}+0.5y\right]dy$

$=0.1 \qquad \therefore P(0<X<0.5,\ 0<Y<0.5)=0.1$

(3) $\rho_{X,Y}=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ (상관계수)

$E(X)=\int_0^1 x f_X(x)\,dx=\int_0^1 x(1.2x^2+0.6)\,dx=0.6$

$E(Y)=\int_0^1 y f_Y(y)\,dy=0.6,\quad E(XY)=\int_0^1\int_0^1 xyf(x,y)\,dx\,dy=0.35$

$Var(X)=E(X^2)-E(X)^2=0.08,\quad Var(Y)=0.0733$

$Cov(X,Y)=E(XY)-E(X)E(Y)=0.35-(0.6)(0.6)=-0.01$

$\therefore \rho_{X,Y}=\frac{-0.01}{\sqrt{0.08\cdot 0.0733}}=-0.131$

(4) 독립이 성립할 경우: $f(x,y)=f_X(x)\cdot f_Y(y)$

그러나, $f(x,y)-f_X(x)\cdot f_Y(y)=-1.44x^2y+0.72x^2+0.48y-0.24$

$\neq 0 \quad \therefore$ 독립이 아니다.

[문제 2]

빨간색 공(R) 3개, 파란색 공(B) 4개, 흰색 공(W)이 5개 들어 있는 항아리에서 임의로 두 개의 공을 선택한다. 다음 물음에 답하라.

(1) 확률변수 X를 뽑힌 공 가운데 빨간색 공의 수, 확률변수 Y를 뽑힌 공 가운데 흰색 공의 수라고 할 때, 확률변수 X와 Y의 결합확률분포를 구하라.

(2) (1)의 결과를 이용하여, 확률변수 X와 확률변수 Y의 주변확률분포를 구하라.

(3) E(X), E(Y), E(XY)를 구하라.

(4) 확률변수 X, Y가 독립인가? 아니라면 공분산을 구하라.

Sol) (1) 전체 경우의 수: ${}_{12}C_2 = 66$

각 결합확률분포:

$$P(X=0, Y=2) = \frac{{}_3C_0 \cdot {}_5C_2}{{}_{12}C_2} = \frac{5}{33}$$

$$P(X=1, Y=1) = \frac{{}_3C_1 \cdot {}_5C_1}{{}_{12}C_2} = \frac{5}{22}$$

$$P(X=2, Y=0) = \frac{{}_3C_2 \cdot {}_5C_0}{{}_{12}C_2} = \frac{1}{22}$$

(2) 주변확률분포:

$$P_X(0) = P(X=0, Y=2) = \frac{5}{33}$$
$$P_X(1) = P(X=1, Y=1) = \frac{5}{22}$$
$$P_X(2) = P(X=2, Y=0) = \frac{1}{22}$$

$$P_Y(0) = P(X=2, Y=0) = \frac{1}{22}$$
$$P_Y(1) = P(X=1, Y=1) = \frac{5}{22}$$
$$P_Y(2) = P(X=0, Y=2) = \frac{5}{33}$$

(3) $$E(X) = \sum_x x \cdot P_X(x) = 0 \cdot \frac{5}{33} + 1 \cdot \frac{5}{22} + 2 \cdot \frac{1}{22} = \frac{7}{22}$$
$$E(Y) = \sum_y y \cdot P_Y(y) = 0 \cdot \frac{1}{22} + 1 \cdot \frac{5}{22} + 2 \cdot \frac{5}{33} = \frac{35}{66}$$

$$E(XY) = \sum_{x,y} x \cdot y \cdot P(X=x, Y=y)$$

$$= 0 \cdot 2 \cdot \frac{5}{33} + 1 \cdot 1 \cdot \frac{5}{22} + 2 \cdot 0 \cdot \frac{1}{22} = \frac{5}{22}$$

(4) "독립"일 경우: $P(X=x, Y=y) = P_X(x) \cdot P_Y(y)$가 성립되어야 한다.

$$Cov(X, Y) = E(XY) - E(X)E(Y) = \frac{5}{22} - \frac{7}{22} \cdot \frac{35}{66} = \frac{85}{1452}$$

∴ 두 확률변수는 독립이 아니다.

[문제 3]

정규분포 $N(30,\ 4)$를 따르는 모집단으로부터 크기 36인 확률표본을 추출하였다. 이때의 표본평균이 $\bar{X}$라고 할 때, 다음에 답하라.

(1) $P(\bar{X} \geq 30)$의 값은?

(2) $P(\bar{X} \geq a) = 0.05$를 만족하는 a의 값은?

(3) $P(-b \leq \bar{X} - 30 \leq b) = 0.95$를 만족하는 b의 값은?

Sol) 정규분포 $N(30, 4)$, $n = 36$

표본평균 $\bar{X} \sim N\left(\mu = 30,\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{3}\right)$

(1) $P(\bar{X} \geq 30) = P\left(Z \geq \frac{30-30}{\frac{1}{3}}\right) = P(Z \geq 0) = 0.5$

(2) $P(\bar{X} \geq a) = 0.05 \Rightarrow P(Z \geq z) = 0.05$

$P(Z \geq z) = 0.05$ 일 때, $z = 1.645$

$z = \frac{a-\mu}{\sigma_{\bar{x}}} \Rightarrow 1.645 = \frac{a-30}{\frac{1}{3}}$

$\therefore a = 30 + 1.645 \cdot \frac{1}{3} \approx 30.55$

(3) $P(-1.96 \leq Z \leq 1.96) = 0.95$

$Z = \frac{\bar{X}-30}{\sigma_{\bar{x}}} \Rightarrow \bar{X} - 30 = Z \cdot \sigma_{\bar{x}}$

$\therefore b = 1.96 \cdot \sigma_{\bar{x}} = 1.96 \times \frac{1}{3} \approx 0.65$

[문제 4]

성인 남자의 고혈압 환자 비율에 대한 조사를 하고자 한다. 신뢰수준 95%하에서 오차한계는 0.02를 넘지 않기 위해서, 다음 각 경우에 대해 얼마나 많은 표본을 추출해야 하는지를 결정하라.

(1) 실제 고혈압 보유 비율을 알 수 없는 경우

(2) 예년의 조사에 따라 고혈압 환자 비율이 0.07이라고 판단되는 경우

(3) (1)과 (2)의 결과에 따라 어떤 결론을 내릴 수 있는가?

Sol) Z : 신뢰수준에 대응하는 Z 값 (95% 신뢰수준, $Z = 1.96$)

P : 모비율 E: 허용오차 ($E = 0.02$)

(1) $P = 0.5$ 로 가정.

$$n = \frac{(1.96)^2 \cdot 0.5 \cdot (1-0.5)}{(0.02)^2} = \frac{3.8416 \cdot 0.25}{0.0004} = 2401$$

(2) $P = 0.07$ 이라 판단되는 경우.

$$n = \frac{(1.96)^2 \cdot 0.07 \cdot (1-0.07)}{(0.02)^2} = \frac{3.8416 \cdot 0.07 \cdot 0.93}{0.0004}$$

$$= \frac{3.8416 \cdot 0.0651}{0.0004} \approx 625$$

(3) 기존 모비율 정보를 활용하면 필요한 표본 크기를 크게 줄일 수 있다. 또한, 실무 상황에서는 예전 자료의 신뢰성이 높으면 625명으로 충분하지만 비율이 크게 변했을 가능성이 있다면 2401 명으로 조사하는 것이 안전하고 예산과 시간제약을 고려하여 결정해야 한다.

[문제 5]

어떤 TV 방송국에서 특집 프로그램을 방영하고 있는데 시청률이 30%라고 한다. 전국에서 900명을 임의로 추출하여 이 특집 프로그램을 시청하고 있는 비율을 추정하고자 한다.

(1) 표본자료의 비율로 전체 시청률을 추정하고자 할 때 적용되는 표본분포를 구하시오.

(2) 표본비율이 32% 이상일 확률을 구하시오. (비율의 표본분포, 이항분포의 정규 근사방법, 연속성 수정을 각각 이용하여 구하여라)

Sol) (1) 표본 비율 $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, $n = 900$

$p = 0.3$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3(1-0.3)}{900}} = \sqrt{\frac{0.21}{900}} = \sqrt{0.0002333} \approx 0.01526$$

$\therefore \hat{p} \sim N(0.3, 0.01526)$

(2) $\hat{p} \geq 0.32$ 일 확률?

비율의 표본분포 이용:

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.32 - 0.3}{0.01526} \approx 1.31$$

계속.

$P(Z \ge 1.31) = 1 - \Phi(1.31) \approx 1 - 0.9049 = 0.0951$

$\therefore P(\hat{p} \ge 0.32) = 0.0951$ (약 9.51%)

이항분포의 정규 근사 이용:

$X \sim B(n, p)$, $X \sim N(np, \sqrt{np(1-p)})$

$X \ge 0.32 \cdot 900 = 288$ 를 정규분포로 근사하여, 연속성 수정을 적용해

$X \ge 287.5$ 로 계산한다.

$Z = \frac{287.5 - 270}{13.75} \approx 1.27$ $P(Z \ge 1.27) = 1 - \Phi(1.27) \approx 1 - 0.8980$

$= 0.1020$

$\therefore P(X \ge 288) = 0.1020$ (약 10.2%)

연속성 수정을 적용 X:

마찬가지로 $X \ge 288$을 정규분포로 근사 (연속성 수정 X)

$Z = \frac{288 - 270}{13.75} \approx 1.31$

$P(Z \ge 1.31) = 1 - \Phi(1.31) \approx 0.0951$

$\therefore P(X \ge 288) = 0.0951$ (약 9.51%)

[문제 6]

베르누이 분포로부터 200개의 표본을 추출하여, 성공 71, 실패 129를 얻었다.

(1) 성공의 확률이 30% 보다 큰가를 알아보고자 한다. 가설을 설정하라.

(2) 영가설하에서의 검정통계량 값을 계산하라.

(3) 유의수준 5%에서 검정하고 결론을 설명하라.

(4) 유의확률(p-value)을 구하여 검정하고 결론을 설명하라.

(5) 성공의 확률이 30%와 다른가를 알아보고자 한다면, 가설을 설정하고 신뢰구간을 구하여 검정하라.

Sol) (1) 귀무가설 (H_0): $p = 0.3$ (성공확률 30%)

대립가설 (H_1): $p > 0.3$ (성공확률이 30% 보다 크다.)

(2) $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ ($\hat{p} = \frac{\text{성공 횟수}}{\text{전체 표본 수}} = \frac{71}{200} = 0.355$

$p_0 = 0.3$, $n = 200$)

계속.

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.3 \cdot 0.7}{200}} = \sqrt{\frac{0.21}{200}} = \sqrt{0.00105} \approx 0.0324$$

$$\therefore z = \frac{0.355 - 0.3}{0.0324} \approx 1.695 \quad (z:\ \text{검정통계량})$$

(3) 단측검정에서 유의수준 5%에 해당하는 임계값 $z_{0.05} = 1.645$ 이다.
계산된 z 값이 $z_{0.05}$ 보다 크므로 귀무가설을 기각한다.
∴ 성공확률이 30% 보다 크다고 판단할 수 있다.

(4) 유의확률(P-value)은 $P(z \geq 1.695)$로 계산:

$$P(z \geq 1.695) = 1 - \Phi(1.695) \approx 1 - 0.9545 = 0.0455$$

유의확률 $p \approx 0.0455$로, 유의수준 5% 보다 작으므로 귀무가설을 기각한다.

(5) $H_0 = 0.3$, $H_1 \neq 0.3$.
검정통계량: $z = 1.695$로 계산.
유의수준 5%에서 양측검정의 임계값: ± 1.96
→ 계산된 z 값이 1.96 보다 작으므로 귀무가설(H_0) 기각이 불가하다.
신뢰구간: $\hat{p} \pm z_{0.025} \cdot \sigma_{\hat{p}}$
$0.355 \pm 1.96 \cdot 0.0324 = 0.355 \pm 0.0635 = [0.2915, 0.4185]$
$p = 0.3$는 이 신뢰구간 안에 포함되므로 귀무가설(H_0) 기각이 불가하다.

∴ 성공확률이 30%와 다르다고 결론내릴 수 없다.